An Evaluation of Alternative Estimation Techniques and Functional Forms in Developing Statistical Appraisal Models

*Han Bin Kang**
*and Alan K. Reichert**

Abstract. **This paper examines the combined effects of multicollinearity, parameter stability, and alternative function forms in hedonic regression models. The results indicate that the significance and stability of the regression coefficients as well as prediction accuracy are sensitive to the choice of functional form and estimation technique. In certain respects nonlinear models proved to be more effective than linear models and ridge regression techniques were generally superior to OLS estimation. Since no single estimation technique or functional form was superior in all respects, the appraiser may have to choose between minimizing the average prediction error or maximizing prediction stability.**

Introduction

A number of articles describe the application of multiple regression analysis (MRA) techniques to real estate appraisal (Bruce and Sundell 1977; Mark and Campbell 1983). In general most researchers have found that while MRA is an appropriate tool for appraisal, the presence of multicollinearity makes it difficult to estimate the true hedonic price of individual property characteristics. Moreover, the forecasting ability of MRA techniques is seriously limited if the pattern of multicollinearity among the regression variables changes over time (Reichert and Moore 1986). Furthermore, many statistical appraisal models assume a linear or additive relationship between selling price and a predetermined set of housing characteristics. Empirical evidence suggests that a nonlinear relationship may be more appropriate, and that use of the linear form may introduce a misspecification problem in identifying the true relationship between selling price and housing features. On the other hand, researchers such as Quigley (1982) failed to identify the optimal functional form using common assumptions regarding utility theory.

The purpose of this study is to improve the accuracy and reliability of the statistical appraisal process by simultaneously addressing the issues of 1) multicollinearity, 2) the temporal stability of estimated parameters, and 3) alternative functional forms in the estimation of hedonic regression models for single-family homes. While each of these issues has individually been addressed in previous studies, the current paper is the first to consider all three issues simultaneously using a single set of data. The stability of the regression parameters has important implications for the appraisal process because hedonic prices serve as adjustment factors in the sales comparison approach. If the regression parameters or adjustment factors are not stable, appraisers cannot confidently use the same adjustment factors over time.

*Department of Finance and Law, Illinois State University, Normal, Illinois 61761.

This study examines how the temporal stability of regression parameters is affected by different estimation techniques and alternative functional forms.

Moore, Reichert and Cho (1984) and others have shown that ridge regression (RR) methods can be used to reduce the adverse effects of multicollinearity in linear models. But their study made no attempt to consider the nonlinear specification of the hedonic model in applying ridge regression techniques. This study focuses on identifying the best functional form and estimation technique in the presence of multicollinearity. To accomplish this, ridge regression results are presented along with ordinary least square (OLS) results using an identical sample of housing market transactions collected over a five-year period. Three alternative functional forms are considered: one linear and two nonlinear specifications. Conceptually, a nonlinear model is superior to a linear model because a nonlinear model of the Cobb-Douglas form allows for the possibility of diminishing marginal price effects.[1] Hence, this study provides a direct comparison of (1) OLS and RR techniques and (2) linear versus log-linear and semi-log models.

Theoretical and Empirical Models

Ridge Regression

The presence of extensive collinearity among regressor variables often generates highly unstable and illogical OLS results.[2] Consequently, to draw firm conclusions about specific coefficients, one must adjust for the presence of collinearity by altering either the underlying data and/or the standard regression procedure. One promising approach described by Belsley, Kuh and Welsch (1980) is ridge regression.[3] Ridge regression is a modified squared-error estimation technique. The ridge estimator:

$$\beta(k) = (X'X + kI)^{-1} X'Y \text{ for } 0 \leq k \leq 1 \quad (1)$$

employs the single ridge parameter k and is a biased estimator. The ridge estimator is similar to the ordinary least squares estimator, β, except that the main diagonal of the correlation matrix is augmented by a small positive quantity k, where k is an *index of bias*. While the ridge estimator, $\beta(k)$, is biased, its variance is generally smaller than the variance of the OLS estimator, making the ridge estimator less prone to yield estimates with improper signs or unstable magnitudes.

Anderson (1979) contends that if the pattern of collinearity among the explanatory variables does not change over time, then $\beta(k)$ and β become equally good estimates. Alternatively, if the pattern of collinearity changes over time, OLS estimators may not prove as accurate as those of ridge regression. In their study, Moore, Reichert and Cho (1984) found that the pattern of collinearity was in fact unstable over time and that ridge regression techniques provided more accurate and stable results than OLS over a five-year estimation period (1975-1979). The models in the current paper were estimated using the same set of data. This study applies linear and nonlinear ridge regression to the same data in hopes of achieving: 1) greater consistency in the size of the estimated hedonic prices over time, 2) a reduction in the incidence of improperly signed coefficients, and 3) improved forecast accuracy.

When estimating a ridge regression model, the researcher must select an "optimal" k-value that produces the maximum degree of orthogonality (independence) in the regressor set while introducing a minimal amount of bias in the estimation process. The authors initially employed the "ridge trace" approach which seeks to identify the minimum value for k that stabilizes the size of the regression coefficients for all three models (linear, log-linear and semi-log). This produced a minimum acceptable k-value of approximately 0.30 for the linear model and a $k = 0.20$ for the nonlinear models. (See Appendix A for the ridge trace for selected variables.) The trace approach is subjective by nature and leads to a stochastic determination of k since the plotted standardized coefficients are stochastic estimators.

Vinod (1976) suggests two alternative heuristic approaches to selecting k such that the reduction in variance outweighs the error introduced by the biasing factor, generating a reduced mean square error of the forecast. Vinod first recommends the use of the Index of Stability of Relative Magnitudes (ISRM).[4] The second approach is called the Numerical Largeness of More Significant Regression Coefficients (NLMS). In the present study both the ISRM and NLMS methods, like the trace approach, yielded an optimal k of 0.30 for the linear functional form and 0.20 for both nonlinear models.

Nonlinear Model

As previously mentioned, nonlinear hedonic regression models are superior to linear models on both practical as well as theoretical grounds. For example, on a pragmatic level, a log transformation often causes the distribution of the dependent variable and the estimated residuals to more closely approximate normality. At the theoretical level, nonlinear models are more appropriate when data interdependences (i.e., non-zero cross partial derivatives) exist among the regression set. Two nonlinear models are estimated. The first is a composite double-log and semi-log (i.e., Cobb-Douglas/exponential) model as indicated below:

Theoretical Model:

$$Y = (\exp^{\alpha_1 D_1 + \alpha_2 D_2 + \ldots + \alpha_m D_m}) \cdot X_1^{\beta_1} X_2^{\beta_2} \ldots X_p^{\beta_p}. \qquad (2)$$

Estimation Model:

$$\ln Y = \alpha_1 D_1 + \alpha_2 D_2 + \ldots + \alpha_m D_m + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \ldots + \beta_p \ln X_p. \qquad (3)$$

where

D_i = housing characteristics with a dummy variable specification such as the existence of a fireplace, central air conditioning, etc. ($i = 1 \ldots m$)

X_i = continuous housing characteristics such as square footage, age of house, etc. ($i = 1 \ldots p$)

Y = actual selling price.

α_i, β_i = represent the regression coefficients (hedonic prices) for the dummy and continuous variables respectively.

exp = the base of the natural logarithm.

The use of the double-log form with respect to the continuous variables allows for increasing, decreasing, or constant returns to scale since the parameter estimates are direct elasticity coefficients. For example, a $\beta < 1$ suggests a diminishing marginal price effect. The use of the semi-log form with respect to the dummy variables allows certain categorical housing characteristics to be absent without forcing the selling price to zero. More importantly, the anti-log of the parameter estimates for the dummy variables reflects the expected *proportionate* increase in price when a housing feature is present compared to a house not having the indicated feature.

The second nonlinear functional form is an adaptation of the first model and generates a semi-log estimation equation.

Theoretical Model:

$$\exp^{Y} = (\exp^{\alpha_1 D_1 + \alpha_2 D_2 + \dots + \alpha_m D_m}) \cdot X_1^{\beta_1} X_2^{\beta_2} \dots X_p^{\beta_p}. \quad (4)$$

Estimation Model:

$$Y = \alpha_1 D_1 + \alpha_2 D_2 + \dots + \alpha_m D_m + \beta_1 lnX_1 + \beta_2 lnX_2 \dots + \beta_p lnX_p. \quad (5)$$

As discussed before, a semi-log model allows for proportional rather than absolute dollar price effects, which is often more appropriate when dealing with a wide range of housing values. A more detailed discussion of alternative functional forms is provided by Colwell, Cannaday and Wu (1983).

The Database

The data for this study consisted of a random sample of 860 single-family home sales reported in the annual sales books of the Multiple Listing Association (MLA) of Fort Wayne, Indiana. This sample was drawn over the five-year period, 1975-1979, from the population of homes sold in the northeast quadrant of the city. The northeast quadrant represents a wide cross-section of socioeconomic lifestyles and residential dwellings, including older homes located near the downtown area and newer houses in the suburbs. These homes encompass the full spectrum of lot sizes, structural design, and household amenities.

All models were estimated using the inflation-adjusted selling price of single-family homes as the dependent variable. (It is necessary to adjust the selling prices for inflation because the objective of the study is to compare the stability of the regression coefficients over time.) The explanatory variables are locational and physical characteristics, and numerous household amenities that could likely influence the final selling price. The following twenty variables were used to describe each property.

Locational Variables To quantify the impact of location on market value, two dummy variables, *WDD*, and *ADDITION*, were considered. *WDD* was included under the premise that homes located on wooded lots and those immediately adjacent to forested areas could likely command, ceteris paribus, a greater price. Furthermore, local realtors were asked to categorize the quality of various additions and neighborhoods as being excellent (*AD*1), good (*AD*2), average (*AD*3), and poor (*AD*4). Thus, the model includes four locational dummy variables (*WDD, AD*1, *AD*2, *AD*4) with *AD*3 serving as the base of the *ADDITION* variable.

Structural Variables In an effort to capture the unique structural characteristics of a property, fourteen distinct variables were employed in the model. *AGE* of the house, measured in years, was included as a crude proxy of condition. Furthermore, a dummy variable, *NEW*, was included to reflect the value of a buyer-decorated or buyer-designed home plus the extra psychological advantages that a *new* house might provide the buyer. The categorical variable *BASEMENT* identified whether the home has a basement (either full or partial) or not. A variable describing the age (in years) of the present heating system (*AGHT*) was included. The physical dimensions of the living area, garage space, and lot size were all measured in square feet by the variables *SQFT*, *GRGE*, and *LTSZ*, respectively. Construction style was initially identified as either ranch, two-story, or other (tri-level, Cape Cod, duplex, etc.) by the categorical variable *STYLE*. *STYLO* is a dummy variable with 1 for other styles and 0 for ranch and two-story houses.[5] A similar variable, *SIDING*, indicated whether the exterior construction material was wood, aluminum, masonry, or some combination of these materials. Variables designed to measure the allocation of living space within the home included the total number of bedrooms (*BDRM*) and the total number of bathrooms (*BTHS*).

Household Amenities Dummy variables were employed in the model to capture the market value of various household amenities. For example, the variable *CTAR* designated whether or not the house had built-in central air conditioning. *FRPL* indicated the presence or absence of a built-in fireplace. A built-in dishwasher (*DSHW*) was viewed as a proxy variable for a new or remodeled kitchen.

Inflation As previously mentioned, to eliminate the influence of inflation during the five-year period, the actual selling price of the home was adjusted by dividing the selling price by the price index of new one-family houses (sold) published by the U.S. Department of Commerce, Bureau of the Census. A final regressor variable, *MNTH*, indicating the month of sale, was included to estimate the average monthly rate of property appreciation within a given year.

Statistical Results

Exhibits 1-3 report the OLS regression results for the linear and two nonlinear models. Exhibits 4-6 present comparable results using ridge regression. Exhibit 7 is a general summary of the OLS and ridge regression results with a variety of prediction accuracy statistics included. (Since the dependent variables have different values in two of the equations, R^2 alone is not the most appropriate measure of goodness of fit for comparison purposes.)

OLS Results

A comparison of the statistical significance for individual regression coefficients (Exhibits 1-3) indicates that the double-log nonlinear model yields a greater number of coefficients significant at the 5% level. As shown in Exhibit 7, out of 100 possibly significant regressor coefficients (5 yearly models times 20 variables per year), the linear model reported only 53 significant coefficients; the semi-log model generated 50 significant coefficients, while the double-log model reported 62 significant coefficients. Furthermore, the estimated results of the linear model shown in Exhibit 1 indicate that the lot size variable (*LTSZ*), for example, is insignificant for three out of five years. As shown in Exhibit 2, however, the double-log model reports that the variable is significant for all years. A comparison of coefficient stability measured by the coefficient of variation indicates a slight margin of superiority for

Exhibit 1
Results of Ordinary Least Squares Regression
(Linear Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
NEW	+	2861	4412	1624*	1769*	1467*	2427	1107	0.4561
		(2.04)	(3.42)	(1.31)	(1.68)	(1.48)			
AGE	-	-470.6	-341.3	-309.1	-327.7	-281.8	-346.1	65	0.1878
		(-7.36)	(-7.41)	(-6.90)	(-6.27)	(-6.06)			
SQFT	+	16.66	15.95	13.02	13.77	17.82	15.44	1.79	0.1193
		(7.43)	(8.24)	(7.61)	(8.07)	(11.32)			
GRGE	+	0.8040*	5.8872*	8.8948	9.9120	8.3729	6.77	3.26	0.4815
		(0.16)	(1.72)	(2.32)	(3.12)	(2.56)			
LTSZ	+	0.2102	0.2544	0.0223*	0.1190*	-0.0325*	0.1146	0.1084	0.9458
		(2.42)	(3.40)	(0.50)	(1.73)	(-0.45)			
WDD	+	3506	2331	3293	1691*	2636	2691	657	0.2441
		(3.22)	(2.09)	(2.67)	(1.73)	(3.78)			
SIDING									
MASON	+	4085*	329.0*	3145*	-734.6*	7406	2846	2883	1.013
		(1.95)	(0.17)	(1.71)	(-0.33)	(4.46)			
COMBO	+	1350*	-353.0*	1171*	741.4*	-646.5*	452.5	807	1.7814
		(1.37)	(-0.41)	(1.42)	(0.99)	(-0.85)			
ADDITION									
EXCELLENT (*AD*1)	+	5922	10152	8202	5000	6810	7217	1808	0.2505
		(4.09)	(6.35)	(6.03)	(3.15)	(4.74)			
GOOD (*AD*2)	+	1727*	2566	2574	2672	2672	2442	360	0.1474
		(1.50)	(2.32)	(2.56)	(2.44)	(2.65)			
POOR (*AD*4)	-	834.8*	-11.92*	-1204*	-2397*	-944.0*	-744.4	1096	1.4723
		(0.55)	(-0.01)	(-0.83)	(-1.92)	(-0.80)			

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 1 continued
Results of Ordinary Least Squares Regression
(Linear Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
BASEMENT	+	6309 (5.26)	5443 (5.48)	6888 (7.29)	7596 (7.52)	3122 (3.06)	5871	1545	0.2631
STYLE-OTHER	-	-8674 (-5.13)	24.11* (0.02)	-1337* (-0.92)	-3544 (-3.11)	-808.2* (-0.78)	-2868	3134	1.0927
BDRM	+	-790.7* (-0.72)	-494.3* (-0.62)	-497.2* (-0.64)	405.2* (0.51)	-183.0* (-0.23)	-312.0	407	1.3044
BTHS	+	3637 (2.46)	1512* (1.22)	2323 (2.34)	2943 (2.52)	3022 (2.98)	2687	720	0.2679
AGHT	-	112.8* (1.22)	-55.04* (-0.67)	-43.77* (-0.61)	29.72* (0.40)	-90.12* (-1.52)	-9.28	72.4	7.801
CTAR	+	2894 (3.18)	4058 (4.34)	1175 (1.96)	2566 (3.18)	1506 (2.16)	2440	1031	0.4225
FRPL	+	1424* (1.12)	668.4* (0.58)	1714* (1.82)	2156 (2.34)	124.5* (0.15)	1217	730	0.5998
DSHW	+	-1049* (-0.82)	1006* (0.81)	1313* (1.44)	-160.1* (-0.16)	1081* (1.23)	438.2	902	2.059
MNTH	+	324.4 (2.69)	312.8 (2.77)	620.7 (5.96)	352.0 (3.22)	27.10 (0.27)	327.4	188	0.5749
CONSTANT		5252	6487	7942	4046	5718	5888	1296	0.2201
N		121	155	191	198	195			
$\bar{R}^2$		.920	.927	.900	.899	.912	.912		
F-Value		70.25	99.29	86.47	88.94	101.21			
Average Errors (%)		NA	10.6	9.4	9.9	8.9	9.7		
C.V. of Errors (%)		NA	114.3	81.1	116.4	101.1	103.2		

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 2
Results of Ordinary Least Squares Regression
(Log-Linear Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
NEW	+	-0.1422	-0.1535	-0.2387	-0.1587	-0.2749	-0.1936	0.0531	0.2742
		(-2.29)	(-2.37)	(-4.00)	(-2.56)	(-5.18)			
AGE	-	-0.1116	-0.1385	-0.0803	-0.0811	-0.0728	-0.0968	0.0247	0.2551
		(-7.43)	(-6.90)	(-6.49)	(-4.88)	(-4.73)			
SQFT	+	0.5794	0.6136	0.5214	0.4271	0.5360	0.5355	0.0632	0.1180
		(7.25)	(6.98)	(8.44)	(6.24)	(8.39)			
GRGE	+	0.0163	-0.0004*	0.0084*	0.0083	0.1081	0.0281	0.0403	1.4341
		(2.26)	(-0.15)	(0.17)	(2.70)	(2.96)			
LTSZ	+	0.1345	0.1347	0.0732	0.1073	0.0684	0.1036	0.0286	0.2760
		(4.75)	(4.55)	(3.77)	(4.90)	(3.03)			
WDD	+	0.0531	0.0393*	0.0558	0.0270*	0.0688	0.0488	0.0143	0.2930
		(2.12)	(1.29)	(1.97)	(1.05)	(3.79)			
SIDING									
MASON	+	0.0658*	-0.0298*	0.1038	0.0007*	0.1337	0.0548	0.0613	1.1186
		(1.30)	(-0.58)	(2.45)	(0.01)	(3.10)			
COMBO	+	0.0394*	-0.0090*	0.0511	0.0188*	-0.0068*	0.0187	0.0240	1.2834
		(1.65)	(-0.38)	(2.70)	(0.96)	(-0.35)			
ADDITION									
EXCELLENT (*AD*1)	+	0.1129	0.1417	0.0922	0.0829	0.1194	0.1098	0.0207	0.1885
		(3.33)	(3.41)	(2.90)	(1.99)	(3.24)			
GOOD (*AD*2)	+	0.0200*	0.0572*	0.0750	0.0400*	0.0500*	0.0484	0.0182	0.3760
		(0.73)	(1.90)	(3.25)	(1.40)	(1.89)			
POOR (*AD*4)	-	0.0710	0.0132*	-0.0457*	-0.0546*	-0.0358*	-0.0103	0.0469	4.55
		(1.99)	(0.38)	(-1.38)	(-1.68)	(-1.16)			

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 2 continued
Results of Ordinary Least Squares Regression
(Log-Linear Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
BASEMENT	+	0.0797	0.0749	0.0992	0.0932	0.0373	0.0768	0.0216	0.2812
		(3.01)	(2.80)	(4.70)	(3.80)	(1.48)			
STYLE-OTHER	-	-0.1448	-0.0017*	-0.0992	-0.0613	0.0077*	-0.0598	0.0578	0.966
		(-3.63)	(-0.05)	(-2.99)	(-2.04)	(0.29)			
BDRM	+	0.0550*	-0.1163*	0.0525*	0.1506	0.0701*	0.0423	0.0870	2.0567
		(0.67)	(-1.57)	(0.87)	(2.26)	(1.13)			
BTHS	+	0.1549	0.1278	0.1123	0.1605	1.1179	0.3346	0.3920	1.2282
		(2.73)	(2.30)	(2.73)	(3.17)	(2.66)			
AGHT	-	0.0443	0.0671	-0.0036*	0.0214*	-0.0206*	0.0217	0.0316	1.4562
		(4.17)	(3.71)	(-0.39)	(1.46)	(-1.75)			
CTAR	+	0.0377*	0.0878	0.0310	0.0716	0.0408	0.0537	0.0220	0.4097
		(1.76)	(3.46)	(2.21)	(3.40)	(2.26)			
FRPL	+	0.0720	0.0397*	0.0507	0.0764	0.0423*	0.0562	0.0151	0.2686
		(2.35)	(1.22)	(2.33)	(3.11)	(1.89)			
DSHW	+	-0.0264*	0.0197*	0.0259*	0.0149*	0.0306*	0.0309	0.0442	1.4304
		(-0.86)	(0.56)	(1.22)	(0.57)	(1.32)			
MNTH	+	0.0282	0.0281*	0.0765	0.0486	-0.0016*	0.0359	0.0258	0.7186
		(2.21)	(1.75)	(6.25)	(3.23)	(-0.12)			
CONSTANT		4.8174	4.8907	5.8128	6.0628	5.2962	5.3759	0.4932	0.0917
N		121	155	191	198	195			
$\bar{R}^2$		.920	.899	.903	.885	.897	.901		
F-Value		69.63	69.74	89.28	76.69	85.74			
Average Errors (%)		NA	11.9	12.3	10.0	9.3			
C.V. of Errors (%)		NA	78.2	73.2	87.5	87.6	81.6		

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 3
Results of Ordinary Least Squares Regression
(Semi-Log Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
NEW	+	-4464.6	-1991.3	-8402.1	-3665.3	-10568	-5818.3	3173.3	0.5454
		(-1.50)*	(-0.72)*	(-3.08)	(-1.35)*	(-4.36)			
AGE	-	-4153.7	-3832.4	-3214.9	-2415.5	-2995.8	-3322.5	615.12	0.1851
		(-5.75)	(-4.45)	(-5.67)	(-3.33)	(-4.26)			
SQFT	+	24158	24763	19043	18486	25707	22431	3039.6	0.1355
		(6.28)	(6.58)	(6.73)	(6.19)	(8.81)			
GRGE	+	-26.047	-53.242	-249.18	149.66	4721.9	908.62	1910.8	2.1030
		(-0.07)*	(-0.43)*	(-0.76)*	(1.12)*	(2.84)			
LTSZ	+	4383.1	5537.5	1963.4	2964.9	425.93	3055.0	1791.0	0.5863
		(3.22)	(4.37)	(2.22)	(3.10)	(0.41)*			
WDD	+	3517.4	2488.2	3488.9	2561.5	3784.6	3168.1	535.77	0.1691
		(2.92)	(1.90)*	(2.67)	(2.29)	(4.57)			
SIDING									
MASON	+	3214.0	192.22	2158.5	-1689.2	5572.4	1889.6	2493.1	1.3194
		(1.32)*	(0.09)*	(1.11)*	(-0.64)*	(2.83)			
COMBO	+	1322.7	-335.07	1697.7	1196.9	-559.91	664.46	925.47	1.3928
		(1.15)*	(-0.33)*	(1.98)	(1.40)*	(-0.63)*			
ADDITION									
EXCELLENT (*AD1*)	+	7898.9	12709	8353.3	6489.6	8503.6	8790.9	2084.1	0.2371
		(4.85)	(7.14)	(5.79)	(3.58)	(5.06)			
GOOD (*AD2*)	+	1430.9	3246.3	3009.2	2684.0	2317.0	2537.5	635.54	0.2505
		(1.09)*	(2.52)	(2.87)	(2.15)	(1.92)*			
POOR (*AD4*)	-	1621.3	-280.92	-2352.2	-3013.8	-882.10	-981.54	1630.0	1.6607
		(0.94)*	(-0.19)*	(-1.54)*	(-2.12)	(-0.62)*			

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 3 continued
Results of Ordinary Least Squares Regression
(Semi-Log Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
BASEMENT	+	5478.4 (4.30)	5303.8 (4.62)	6495.3 (6.79)	6500.7 (6.07)	1904.8 (1.65)*	5136.6	1690.9	0.3292
STYLE-OTHER	-	-8405.4 (-4.38)	-873.50 (-0.61)*	-1963.8 (-1.35)*	-3080.6 (-2.35)	-590.23 (-0.48)*	-2982.7	2850.5	0.9557
BDRM	+	2014.6 (0.51)*	-2402.3 (-0.76)*	1438.0 (0.52)*	4273.7 (1.47)*	2384.5 (0.84)*	1541.7	2189.2	1.4200
BTHS	+	6107.7 (2.23)	3702.4 (1.55)*	5389.4 (2.91)	7121.4 (3.23)	3869.5 (1.91)*	5238.1	1308.2	0.2497
AGHT	-	1466.4 (2.87)	1356.4 (1.75)*	124.37 (0.29)*	551.06 (0.86)*	-656.1 (-1.22)*	568.43	790.47	1.3906
CTAR	+	2325.9 (2.26)	4647.9 (4.28)	1632.4 (2.54)	3310.0 (3.60)	2019.5 (2.45)	2787.1	1083.6	0.3888
FRPL	+	1665.5 (1.13)*	362.39 (0.26)*	1580.8 (1.58)*	1737.2 (1.62)*	274.27 (0.2)*	1124.0	660.30	0.5875
DSHW	+	-746.36 (-0.51)*	766.37 (0.51)*	1378.6 (1.41)*	-99.42 (-0.09)*	654.97 (0.62)*	390.83	737.57	1.8872
MNTH	+	847.43 (1.38)*	1510.6 (2.20)	3386.5 (6.06)	1378.0 (2.10)	-404.39 (0.65)*	1343.6	1225.1	0.9118
CONSTANT		-182978	-193543	-125284	-134626	-179992	-163285	27741	0.1699
N		121	155	191	198	195			
$\bar{R}^2$		0.894	0.900	0.884	0.870	0.875	0.885		
F-Value		51.82	70.40	73.01	65.17	69.05			
Average Errors (%)		NA	13.2	12.4	10.3	9.3	11.6		
C.V. of Errors (%)		NA	71.1	77.2	103.8	10.5	86.8		

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 4
Results of Ridge Regression
(Linear Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
NEW	+	4029	4191	1761	1973	1516	2694	1166	0.4328
		(4.06)	(4.23)	(2.16)	(2.89)	(2.09)			
AGE	-	-238.0	-208.8	-198.7	-178.6	-153.0	-195.4	28.58	0.1463
		(-7.29)	(-7.58)	(-7.74)	(-6.93)	(-7.16)			
SQFT	+	8.918	8.341	8.008	8.213	10.20	8.736	0.7920	0.0907
		(10.62)	(12.86)	(12.45)	(13.20)	(16.69)			
GRGE	+	7.948	7.955	10.17	10.87	10.35	9.459	1.252	0.1324
		(2.28)	(3.32)	(4.00)	(4.89)	(4.32)			
LTSZ	+	0.1374*	0.2237	0.0586*	0.1364	0.0467*	0.1206	0.0640	0.5307
		(1.94)	(3.99)	(1.89)	(2.60)	(0.85)			
WDD	+	2280	1675*	2662	2004	2089	2142	325.4	0.1519
		(2.67)	(1.83)	(2.79)	(2.73)	(3.81)			
SIDING									
MASON	+	2688*	1672*	2182*	-437.9*	6763	2573	2349	0.9129
		(1.55)	(1.14)	(1.55)	(-0.26)	(5.18)			
COMBO	+	1394*	626.9*	1441	1098	174.9*	947.0	482.8	0.5098
		(1.93)	(0.99)	(2.60)	(1.97)	(0.32)			
ADDITION									
EXCELLENT (*AD1*)	+	6218	9194	7020	5611	7123	7033	1214	0.1726
		(5.92)	(8.73)	(7.80)	(4.85)	(6.55)			
GOOD (*AD2*)	+	908.3*	1663	2139	2296	2387	1879	545.6	0.2904
		(1.03)	(2.18)	(3.02)	(2.82)	(2.92)			
POOR (*AD4*)	-	222.8*	-380.3*	-1546*	-2129	-1699*	-1106	881.4	0.7969
		(0.20)	(-0.38)	(-1.36)	(-2.15)	(-1.80)			

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 4 continued
Results of Ridge Regression
(Linear Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
BASEMENT	+	3488 (4.39)	3868 (5.47)	5460 (8.92)	5421 (7.85)	2177 (3.35)	4083	1243	0.3044
STYLE-OTHER	-	-4596 (-3.50)	105.2* (0.11)	-1530* (-1.46)	-1902 (-2.17)	233.61* (0.28)	-1538	1751	1.1385
BDRM	+	1970 (3.14)	1350 (2.66)	1224 (2.68)	1968 (4.27)	2244 (4.84)	1751	394.1	0.2251
BTHS	+	4532 (7.32)	3684 (7.23)	3365 (7.06)	3720 (7.85)	4025 (8.77)	3865	393.5	0.1018
AGHT	-	-42.04* (-0.64)	-95.66* (-1.80)	-78.98* (-1.70)	-55.71* (-1.25)	-140.3 (-3.98)	-82.54	34.30	0.4156
CTAR	+	2395 (3.31)	4022 (5.97)	1384 (3.00)	2673 (4.71)	1854 (3.46)	2466	896.0	0.3633
FRPL	+	2178 (2.71)	2052 (3.12)	2505 (4.28)	2849 (4.79)	1388 (2.41)	2194	489.0	0.2229
DSHW	+	1505 (1.97)	2331 (3.53)	1994 (3.46)	1317 (2.18)	1861 (3.24)	1802	358.8	0.1991
MNTH	+	123.8* (1.31)	176.2* (1.94)	488.4 (6.00)	260.6 (3.01)	13.13* (0.16)	212.4	159.6	0.7514
CONSTANT		3759.9	6677.9	6619.6	3693.3	3365.8	4823.3	1497	0.3104
N		121	155	188	191	195			
$\bar{R}^2$		.892	.909	.883	.882	.890	.891		
F-Value		50.56	77.58	72.92	74.88	79.18			
Average Errors (%)		NA	10.1	9.2	9.3	9.3	9.5		
C.V. of Errors (%)		NA	72.8	81.6	120.4	98.6	93.4		

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 5
Results of Ridge Regression
(Log-Linear Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
NEW	+	0.0198*	-0.0013*	-0.0336*	-0.0098*	-0.0564	-0.0163	0.0264	1.6196
		(0.73)	(-0.04)	(-1.55)	(-0.52)	(-2.71)			
AGE	-	-0.0416	-0.0388	-0.0308	-0.0231	-0.0239	-0.0316	0.0075	0.2373
		(-7.88)	(-7.35)	(-7.40)	(-6.81)	(-6.88)			
SQFT	+	0.3497	0.3371	0.3394	0.3197	0.3574	0.3407	0.0128	0.0376
		(8.83)	(9.41)	(11.65)	(10.39)	(11.97)			
GRGE	+	0.0182	0.0019*	0.0118*	0.0085	0.1242	0.0329	0.0459	1.3951
		(2.79)	(0.73)	(1.94)	(3.35)	(4.36)			
LTSZ	+	0.1078	0.1227	0.0772	0.1003	0.0626	0.0941	0.0215	0.2285
		(4.51)	(4.93)	(5.14)	(5.49)	(3.40)			
WDD	+	0.0380*	0.0291*	0.0436*	0.0352*	0.0456	0.0383	0.0059	0.1540
		(1.71)	(1.07)	(1.79)	(1.69)	(3.03)			
SIDING									
MASON	+	0.0491*	0.0046*	0.0680*	-0.0261*	0.1333	0.0458	0.0548	1.1965
		(1.09)	(0.11)	(1.88)	(-0.54)	(3.69)			
COMBO	+	0.0441	0.0215*	0.0495	0.0295*	0.0080*	0.0305	0.0151	0.4951
		(2.29)	(1.11)	(3.41)	(1.86)	(0.52)			
ADDITION									
EXCELLENT (*AD1*)	+	0.1208	0.1563	0.1148	0.1007	0.1458	0.1277	0.0204	0.1597
		(4.32)	(4.82)	(4.84)	(3.04)	(4.83)			
GOOD (*AD2*)	+	0.0170*	0.0488	0.0618	0.0454	0.0559	0.0458	0.0155	0.3384
		(0.72)	(2.07)	(3.36)	(1.95)	(2.48)			
POOR (*AD4*)	-	0.0420*	0.0003*	-0.0562	-0.0571	-0.0390*	-0.0215	0.0385	1.7907
		(1.40)	(0.0009)	(-1.95)	(-2.06)	(-1.49)			

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 5 continued
Results of Ridge Regression
(Log-Linear Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
BASEMENT	+	0.0307* (1.45)	0.0320* (1.51)	0.0828 (5.25)	0.0615 (3.18)	0.0236* (1.28)	0.0461	0.0225	0.4881
STYLE-OTHER	-	-0.0877 (-2.54)	-0.0129* (-0.47)	-0.0838 (-3.12)	-0.0386* (-1.55)	0.0210* (0.92)	-0.0404	0.0416	1.0297
BDRM	+	0.2222 (3.98)	0.0956* (1.85)	0.1504 (3.66)	0.2090 (4.75)	0.1943 (4.55)	0.1743	0.0462	0.2651
BTHS	+	0.1937 (6.17)	0.1862 (6.19)	0.1628 (6.87)	0.1767 (6.90)	0.1586 (6.66)	0.1756	0.0134	0.0763
AGHT	-	0.0105* (1.78)	0.0038* (0.62)	-0.0057* (-1.24)	-0.0034* (-0.85)	-0.0195 (-4.61)	-0.0029	0.0101	3.4828
CTAR	+	0.0459 (2.41)	0.0909 (4.40)	0.0350 (2.95)	0.0685 (4.20)	0.0494 (3.32)	0.0579	0.0197	0.3402
FRPL	+	0.0841 (3.81)	0.0659 (3.11)	0.0733 (4.66)	0.0875 (4.98)	0.0556 (3.33)	0.0733	0.0117	0.1596
DSHW	+	0.0403* (1.90)	0.0840 (3.92)	0.0522 (3.38)	0.0462 (2.56)	0.0558 (3.35)	0.0557	0.0151	0.2711
MNTH	+	0.0101* (0.89)	0.0149* (1.06)	0.0638 (6.06)	0.0427 (3.37)	0.0003* (0.03)	0.0264	0.0234	0.8864
CONSTANT		6.4286	6.6132	6.8364	6.7345	6.2459	6.5717	0.2121	0.0323
N		121	155	191	198	195			
$\bar{R}^2$		.893	.874	.886	.871	.880	.881		
F-Value		50.57	54.21	74.75	67.77	72.31			
Average Errors (%)		NA	13.0	11.8	10.8	9.8	11.4		
C.V. of Errors (%)		NA	81.8	68.0	83.2	89.3	80.6		

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 6
Results of Ridge Regression
(Semi-Log Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
NEW	+	1579.6	1976.4	-887.30	394.80	-2080.0	196.70	1515.1	7.7026
		(1.26)*	(1.50)*	(-0.91)*	(0.49)*	(-2.20)			
AGE	-	-1565.1	-1206.4	-1230.3	-767.20	-961.3	-1146.1	269.81	2.3647
		(-6.45)	(-5.56)	(-6.52)	(-5.26)	(-6.09)			
SQFT	+	14841	14135	13082	13417	16463	14388	1201.9	0.0835
		(8.14)	(9.60)	(9.90)	(10.16)	(12.13)			
GRGE	+	160.18	28.155	-13.62	194.63	5184.7	1110.8	2038.4	1.8351
		(0.53)*	(0.27)*	(-0.05)*	(1.77)*	(4.00)			
LTSZ	+	3164.2	4786.5	2277.6	2839.2	726.7	2758.8	1314.8	0.4766
		(2.88)	(4.67)	(3.34)	(3.62)	(0.87)*			
WDD	+	2589.0	1922.7	2830.4	2615.9	2645.5	2520.7	310.75	0.1233
		(2.53)	(1.72)*	(2.56)	(2.93)	(3.86)			
SIDING									
MASON	+	3214.0	1533.6	1299.2	-1820.8	5767.6	1998.7	2489.0	1.2453
		(1.17)*	(0.85)*	(0.79)*	(0.88)*	(3.51)			
COMBO	+	1475.9	690.41	1877.4	1438.1	91.00	1114.6	639.49	0.5737
		(1.66)*	(0.86)*	(2.85)*	(2.11)*	(0.13)*			
ADDITION									
EXCELLENT (*AD*1)	+	7586.8	11507	8183.4	6595.4	9024.5	8579.4	1664.2	0.1940
		(5.90)	(8.63)	(7.61)	(4.64)	(6.58)			
GOOD (*AD*2)	+	977.82	2290.4	2476.2	2508.0	2468.3	2144.1	588.12	0.2743
		(0.91)*	(2.36)	(2.96)	(2.52)	(2.41)			
POOR (*AD*4)	-	743.02	-624.65	-2627.4	-2933.3	-1361.1	-1360.7	1345.0	0.9885
		(0.54)*	(-0.50)*	(-2.01)	(-2.46)	(-1.14)*			

*Indicates coefficient *not* statistically significant at 5% level.

Exhibit 6 continued
Results of Ridge Regression
(Semi-Log Model)

Variable	Expected Sign	1975	1976	1977	1978	1979	Mean	Standard Deviation	Coefficient of Variation
BASEMENT	+	3322.6 (3.42)	3719.7 (4.25)	5336.8 (7.45)	5142.3 (6.20)	1500.2 (1.80)*	3804.3	1391.9	0.3659
STYLE-OTHER	-	-5559.8 (-3.50)	-704.9 (-0.62)*	-2148.2 (-1.76)*	-1952.9 (-1.83)*	263.74 (-0.25)*	-2020.4	1974.6	0.9773
BDRM	+	8455.5 (3.29)	4508.4 (2.12)	5478.8 (2.94)	7288.4 (3.86)	8079.8 (4.16)	6762.2	1523.6	0.2253
BTHS	+	7925.7 (5.49)	6676.4 (5.40)	6753.5 (6.28)	7262.9 (6.61)	6187.4 (5.72)	6961.2	590.64	0.0848
AGHT	-	238.48 (0.88)*	-199.00 (-0.80)*	-148.99 (-0.71)*	-190.91 (-1.10)*	-724.1 (-3.76)	-204.90	306.43	1.4955
CTAR	+	2448.8 (2.79)	4637.2 (5.46)	1816.4 (3.38)	3186.4 (4.55)	2314.0 (3.42)	2880.56	981.75	0.3408
FRPL	+	2536.4 (2.49)	1951.3 (2.24)	2506.7 (3.51)	2551.2 (3.38)	1221.2 (1.61)*	2153.4	517.61	0.2404
DSHW	+	1759.6 (1.80)*	2731.9 (3.10)	2233.1 (3.19)	1265.2 (1.63)*	1793.8 (2.36)	1956.7	494.14	0.2525
MNTH	+	193.96 (0.37)*	907.69 (1.57)*	2793.3 (5.84)	1155.3 (2.12)	-196.79 (0.38)*	970.69	1032.1	1.0633
CONSTANT		-116395	-122032	-95143	-103409	-132171	-113830	13176	0.1158
N		121	155	191	198	195			
$\bar{R}^2$		0.871	0.885	0.869	0.856	0.856	0.867		
F-Value		41.51	59.97	63.72	59.70	58.53			
Average Errors (%)		NA	13.6	10.9	10.5	10.6	11.4		
C.V. of Errors (%)		NA	73.6	76.9	96.3	95.5	85.6		

*Indicates coefficient *not* statistically significant at 5% level.

EXHIBIT 7
Coefficient Stability and Predicting Accuracy for the 1975-79 Estimation Period

	Exhibit 1	Exhibit 2	Exhibit 3	Exhibit 4	Exhibit 5	Exhibit 6
Estimation method	OLS	OLS	OLS	RR	RR	RR
Functional form	Linear	Log-Linear	Semi-Log	Linear	Log-Linear	Semi-Log
# of significant coefficients (maximum of 100)	53	62	50	74	64	62
# of correctly signed coefficients (maximum of 100)	86	82	80	96	90	92
Average C.V. for all regression coefficients (%)	107.4	94.9	83.5	39.5	73.3	104.5
Average $\bar{R}^2$ for all years (1975-79)	.912	.901	.885	.891	.881	.867
Average prediction error[a]	9.7	10.9	11.6	9.5	11.4	11.4
Average C.V. of prediction errors (%)[b]	103.2	81.6	86.8	93.4	80.6	85.6
Chow-Test *F*-Value[c]	6.39	6.30	6.13	4.65	5.05	5.20

[a]Average absolute percentage error (1975-79)

[b]Average coefficient of variation of absolute percentage errors (1975-79)

[c]The critical *F*-values at the 1% and 5% levels of significance for a variable with 21 and 806 degrees of freedom are approximately 1.94 and 1.60, respectively.

the double-log model. (The coefficient of variation is the standard deviation of the yearly regression coefficients divided by their mean.) For example, for 8 out of the 20 regressor variables, the double-log model generated the smallest coefficient of variation. This is particularly true for the lot size variable (*LTSZ*) where the coefficient of variation for *LTSZ* for the double-log model is only half the size of the coefficients reported in the other two models (.28 versus .95 for the linear model and .28 versus .59 for the semi-log model). On the other hand, the double-log model generated only marginal improvement in the temporal stability of the square footage variable (*SQFT*).

Ridge Regression Results

The empirical results using the ridge regression approach are reported in Exhibits 4-6 and summarized in Exhibit 7. As shown in Exhibit 7, ridge regression generated a larger number of correctly signed coefficients than OLS for every functional form. A comparison of the linear OLS and the linear ridge regression results (Exhibit 1 versus Exhibit 4) shows a dramatic improvement in the level of statistical significance in the ridge regression coefficients: the number of significant coefficients over the five-year period increased from 53 to 74. Coefficient stability in the linear ridge model was also superior to the linear OLS method, with 17 out of the 20 variables in the ridge model showing a smaller coefficient of variation. In many cases, the improvement in stability was quite pronounced, as can be seen by the results for the number of bedrooms (*BDRM*), age of the heating system (*AGHT*) and the proxy for a remodeled kitchen (*DSHW*). Furthermore, the mean of the coefficients of variation is 107.4% for the linear OLS model compared to only 39.5% for the linear ridge regression model.

Little improvement in the number of significant variables is shown in Exhibits 2 and 5 for the double-log ridge model compared to the double-log OLS results (62 versus 64). On the other hand, the double-log ridge regression model compared to the double-log linear model produced smaller coefficients of variation in 14 out of 20 cases. As shown in Exhibit 7, the average coefficient of variation is 94.9% for the double-log OLS model and 73.3% for the double-log ridge model. A comparison of the total number of significant variables using ridge regression indicates no improvement for the double-log ridge model compared to the linear ridge model (Exhibit 4 versus Exhibit 5). However, the double-log ridge model shows more significant and stable coefficients for several critical continuous variables such as *LTSZ* and *SQFT*.

In the semi-log functional form, the ridge regression model generated 62 statistically significant coefficients compared to only 50 for the OLS semi-log model (Exhibit 3 versus Exhibit 6). Furthermore, for 14 out of the 20 regressor variables the semi-log ridge regression model generated a smaller coefficient of variation than the semi-log OLS model. When the double-log ridge regression model is compared to the semi-log ridge regression model (Exhibit 5 versus Exhibit 6), the double-log specification resulted in more significant and stable coefficients.

The preceding discussion is based upon examination of the results for *individual* regressor variables. The *F*-test developed by Chow (1960) was employed to more formally test the stability of the annual regression models overtime. Acceptance of the null hypothesis would suggest that the yearly regression models are drawn from the same underlying population and hence the annual data could be pooled to estimate a single equation. Rejection of the null hypothesis would call for the estimation of separate annual models. The Chow-test results presented in Exhibit 7 indicate that the null hypothesis should be rejected at all reasonable levels of statistical significance for each of the six models. The implication is that annual models are appropriate and the researcher should expect to see the hedonic price (regression coefficients) change over time.

While these results provide evidence to support the log-linear specification, a more powerful and conceptually more appropriate test of the optimal functional form should be based on a variety of functional forms in an unconstrained multivariate framework such as Box-Cox transformations.

Box-Cox Functional Form Approach

Since hedonic price equations represent reduced-form models reflecting both demand and supply factors, the optimal functional form is difficult to identify on purely theoretical grounds (Halvorsen and Pollakowski 1981 and Quigley 1982). For the three functional forms studied in this paper, the adjusted coefficient of determination ($\bar{R}^2$) cannot be used to choose the best specification of the model because the dependent variable is not the same for all models. The dependent variable is the natural log of selling price for the log-linear model and the actual selling price for other two models. The Box-Cox transformation permits us to choose the best functional form empirically in an unconstrained multivariate framework. The Box-Cox transformation was performed to test whether the log-linear model truly represents the most appropriate functional form.

The Box-Cox procedure employs a transformation of both the dependent and independent variables of the following general form:

$$\frac{V^{\lambda_L}-1}{\lambda_L} \quad ; \quad \frac{V^{\lambda_R}-1}{\lambda_R}$$

where V represents the untransformed regression variables and λ_L, λ_R represents the "functional form specification coefficients" for the left-hand side (dependent) variable and right-hand side (independent) variable, respectively.[6] Since it can be shown that $\lim_{\lambda \to 0} [(V^{\lambda} - 1) / \lambda] = \ln V$, the double-log functional form is regarded as a special case with $\lambda_L = \lambda_R = 0$. The semi-log form is also a special case with $\lambda_L = 1$ and $\lambda_R = 0$. When both λ_L and λ_R are equal to one, the linear functional form is specified. While it is possible to obtain maximum likelihood estimates for the models directly, a more common technique is to obtain the concentrated likelihood function that depends upon the values of λ_L and λ_R and then conduct a grid search over reasonable values of λ_L and λ_R to obtain the global maximum likelihood estimates for λ_L and λ_R (see Judge 1980). A confidence region can then be developed around the maximum likelihood estimates for λ_L and λ_R using the χ^2 distribution since, for large samples, twice the difference in the logarithmic likelihood between the null and alternative hypothesis follows a χ^2 distribution (Halvorsen and Pollakowski; Judge). Thus a $[100\ (1\text{-}\alpha)]$ percent confidence region includes all functional form parameter estimates (λ_L and λ_R) which conform to the following inequality: optimal max log likelihood (λ_L^*, λ_R^*) - max log likelihood $(\lambda_L, \lambda_R) < \frac{1}{2}\, \chi^2_{(2)}(\alpha)$ where λ_L^* and λ_R^* represent the estimated optimal maximum log likelihood parameters. Two degrees of freedom are used because the test involves two constrained parameters (λ_L and λ_R).

As mentioned above, the procedure to choose the best functional form requires a grid search for optimal values of λ_L and λ_R that maximize the logarithmic likelihood function (Colwell and Cannaday 1985). Data for 1975 were used to conduct the grid search with values of λ_L and λ_R ranging from -1.0 to 1.0. Since the Box-Cox transformation cannot be performed for dummy variables and other independent variables with zero values, these variables were not transformed.[7] To simplify the analysis, λ_R was restricted to be equal for all continuous independent variables. λ_L was allowed to assume its own value, independent of λ_R, to identify the optimal combination of λ_L and λ_R. The optimal maximum value of the log likelihood was obtained at $\lambda_L = 0.2$ and $\lambda_R = 0.2$. The values of the log likelihood for each of the three functional forms are given below.

Functional Form	λ_L	λ_R	Log Likelihood	Difference From Maximum		½ χ^2 (5%)[a]
1. Linear	1	1	-1305.07	15.87[b]	>	2.995
2. Semi-Log (Log on right side)	1	0	-1318.36	29.16	>	2.995
3. Log-Linear (double-log)	0	0	-1289.80	0.60	<	2.995
Optimal Maximum	0.2	0.2	-1289.20			

[a]χ^2 Value consistent with a 95% confidence region.

[b]For example, 15.87 = -1289.20 (optimal max log likelihood) - (-1305.07) (linear log likelihood).

We can test whether each model is significantly different from the optimal model where $\lambda_L = 0.2$ and $\lambda_R = 0.2$. The test can be performed by constructing a confidence region such that [maximum log likelihood (0.2, 0.2) - maximum log likelihood (λ_L, λ_R)] $< \frac{1}{2} \chi^2_{(2)}(\alpha)$, where α is the level of statistical significance employed in the test. If the log likelihood value associated with λ_L and λ_R satisfies the inequality, we accept the null hypothesis that the model is not significantly different from the optimal model. The χ^2 value at the 5% significance level with two degrees of freedom is 2.995. Hence, the null hypothesis is accepted if the difference between the estimated log likelihood and the optimal maximum log likelihood is less than 2.995. As indicated in the table above the null hypothesis is rejected for all models except the log-linear model. That is, only the log-linear model yields a log likelihood that is not significantly different from the optimal maximum value of the log likelihood. This result strongly supports our initial conclusion that the log-linear model is the more appropriate functional form.

Forecast Accuracy and Reliability

In addition to the degree of temporal stability and level of statistical significance associated with the regression coefficients, appraisers are concerned with the average accuracy and stability of their statistical forecasts. To test each model's ability to forecast as much as a year into the future, the regression coefficients estimated using data for a given year were used to forecast the value of homes sold during the following year. The price forecasts made using the previous year's regression coefficients were then compared to their actual selling prices.[8] Four annual forecast comparisons (1976-1979) were possible. The average absolute percentage error and the average coefficient of variation (C.V.) for each of the six models are presented in Exhibit 7, along with summary statistics relating to coefficient stability.

Using the average absolute percentage prediction error as our measure of accuracy, the linear ridge regression model resulted in the smallest average error (9.5%) closely followed by the linear OLS model (9.7%). In terms of prediction stability, the linear ridge model yielded a somewhat smaller C.V. than the linear OLS model (93.4% vs. 103.2%). The next most accurate model is the log-linear OLS specification (10.9%), which also produced one of the lowest coefficients of variation (81.6%). As previously mentioned, the linear ridge regression model generated the largest number of statistically significant and properly signed regression coefficients. Furthermore, as indicated in Exhibit 7, the average C.V. for all the regression coefficients was by far the smallest for the linear ridge model (39.5%).

On the other hand, the reliability of the forecast, as measured by the average C.V. of the *prediction errors*, was better with ridge regression than with OLS for every functional form. Furthermore, the log-linear specification proved to generate more stable forecasts in both the OLS and ridge regression models. The log-linear model also generated the largest number of significant coefficients among the three OLS specifications.

Conclusions, Implications and Suggestions for Further Research

This study indicates that the significance and stability of the regression coefficients as well as average prediction accuracy are sensitive to the choice of the specific functional form and estimation technique. The results suggest that appraisers may be forced to choose between minimizing the average prediction error or maximizing prediction stability. Thus, there is no single estimation technique or functional form that is consistently superior in all aspects.

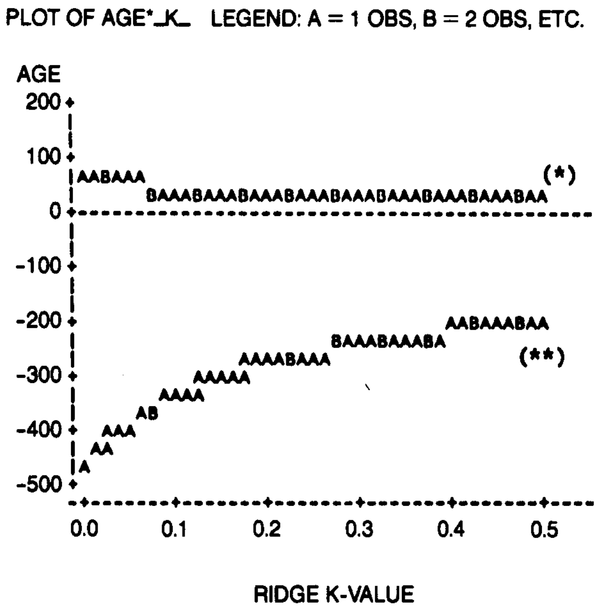
That is, the intended purpose of the appraisal may be a critical factor in selecting the optimal functional form and estimation technique. For example, if the primary purpose of the analysis is for tax assessment, assessors may likely be more concerned with prediction stability given their traditional emphasis on tax equity.

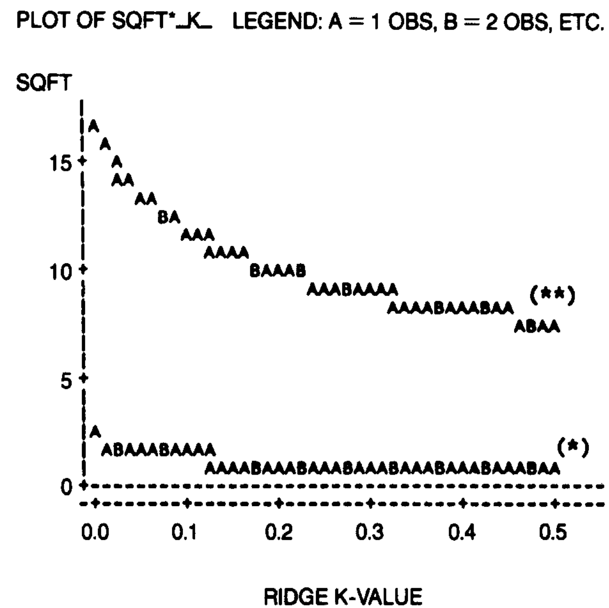
On the other hand, the individual fee appraiser deals with a single property at a given time and may be more concerned with prediction accuracy. The appraiser must carefully identify his own unique forecasting needs. To illustrate, while this paper finds that ridge regression techniques are superior to OLS in terms of forecasting *stability*, the OLS model may be a good choice for forecasting accuracy. In other words, the log-linear ridge regression model may be more appropriate for the tax assessor dealing with equity issues covering a large number of properties and having to forecast tax assessments perhaps several years into the future.

A caveat is in order here. The appraiser must keep in mind the limitations of his data and his ability to effectively explain and justify his procedures to others. The more complicated the model the more difficult it may be to interpret. For example, ridge regression requires the appraiser to estimate the optimal biasing factor (k) and Baynesian regression requires the identification of appropriate statistical priors.

In addition, if the appraiser is using small data sets he may want to consider the use of rank regression techniques (Cronan, Epley and Perry 1986). On the other hand, if ample data is available the appraiser may want to consider a segmented regression approach that subdivides the properties by price range or area/income level. Other possible techniques might include adaptive feedback and Baynesian regression. Adaptive feedback uses the forecasting residuals to refine the regression coefficients to maximize within-sample forecast accuracy. Baynesian regression uses information regarding the prior distribution of the estimated coefficients and allows for direct nonlinear Gauss/Newton estimation techniques. Furthermore, other nonlinear functional forms should be explored in the future. The purpose of the current paper was to evaluate three alternative functional forms. The future researcher may want to identify the optimal functional form in a more general sense by estimating the Box-Tidwell transformation in an unconstrained format that would allow each independent variable to have a unique lambda and hence its own unique functional form.

APPENDIX A
(Ridge Trace for Selected Variables)

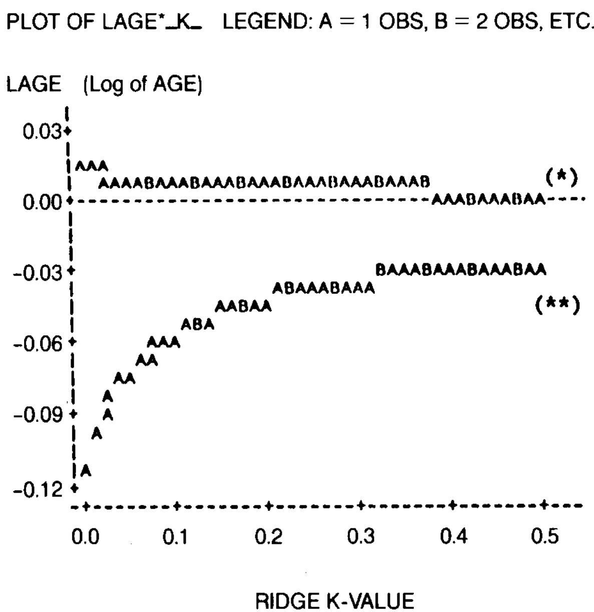


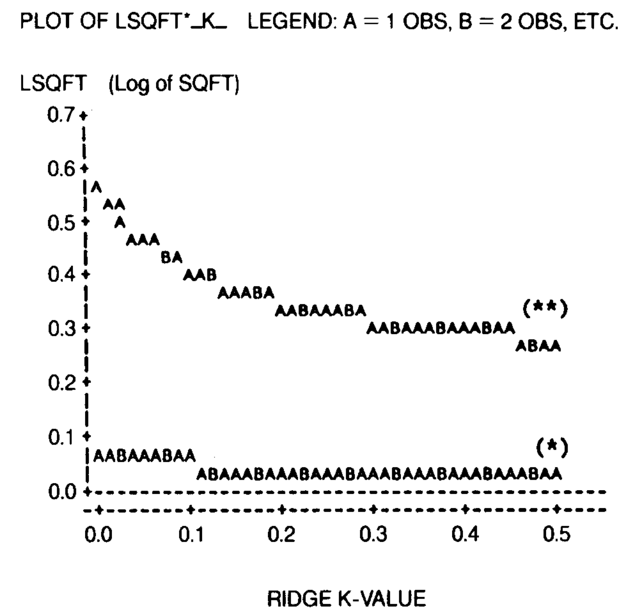


*standard errors of coefficients

**ridge regression coefficients

APPENDIX A continued
(Ridge Trace for Selected Variables)





*standard errors of coefficients

**ridge regression coefficients

APPENDIX B
(Results of ISRM and NLMS)

Table A
Index of Stability of Relative Magnitudes (ISRM)
(Linear)

K	75	76	77	78	79
0.00	18.04	18.35	13.79	16.92	16.41
0.05	8.82	8.37	7.59	8.48	8.54
0.10	5.25	4.91	4.81	5.03	5.08
0.15	3.47	3.28	3.30	3.29	3.30
0.20	2.45	2.37	2.37	2.29	2.30
0.25	1.83	1.81	1.77	1.68	1.69
<u>0.30</u>*	1.42	1.45	1.36	1.29	1.31
0.35	1.16	1.20	1.07	1.03	1.06
0.40	0.99	1.04	0.88	0.86	0.90
0.45	0.87	0.92	0.74	0.76	0.80
0.50	0.81	0.85	0.64	0.70	0.73

*Optimal *k*-value

Table B
Index of Stability of Relative Magnitudes (ISRM)
(Log-Linear)

K	75	76	77	78	79
0.00	28.62	41.02	30.81	52.32	46.77
0.05	10.95	11.74	11.81	13.29	11.90
0.10	5.87	5.68	6.08	5.73	5.51
0.15	3.66	3.48	3.64	3.21	3.28
<u>0.20</u>*	2.52	2.47	2.42	2.17	2.30
0.25	1.87	1.94	1.76	1.70	1.81
0.30	1.49	1.65	1.38	1.49	1.55
0.35	1.27	1.49	1.17	1.40	1.41
0.40	1.14	1.40	1.05	1.38	1.34
0.45	1.07	1.36	0.99	1.40	1.32
0.50	1.05	1.36	0.97	1.44	1.32

*Optional *k*-value

Table C
Numerical Largeness of More Significant Regression Coefficient (NLMS)
(Linear)

K	75	76	77	78	79
0.00	0.93165	0.92145	0.92932	0.93031	0.96299
0.05	0.96811	0.96471	0.96400	0.96442	0.97939
0.10	0.98493	0.98459	0.98157	0.98213	0.98947
0.15	0.99330	0.99396	0.99117	0.99162	0.99506
0.20	0.99734	0.99799	0.99616	0.99639	0.99795
0.25	0.99902	0.99918	0.99841	0.99852	0.99922
0.30*	0.99939	0.99886	0.99906	0.99908	0.99955
0.35	0.99902	0.99775	0.99878	0.99873	0.99934
0.40	0.99826	0.99625	0.99798	0.99804	0.99885
0.45	0.99730	0.99460	0.99693	0.99705	0.99821
0.50	0.99625	0.99294	0.99577	0.99595	0.99751

*Optimal *k*-value

Table D
Numerical Largeness of More Significant Regression Coefficient (NLMS)
(Log-Linear)

K	75	76	77	78	79
0.00	0.93028	0.91144	0.88952	0.77599	0.89048
0.05	0.97309	0.96859	0.94895	0.94096	0.96597
0.10	0.98894	0.98901	0.97811	0.98107	0.98768
0.15	0.99488	0.99500	0.99098	0.99044	0.99380
0.20*	0.99675	0.99575	0.99582	0.99168	0.99419
0.25	0.99675	0.99458	0.99688	0.99028	0.99247
0.30	0.99592	0.99280	0.99624	0.98802	0.99015
0.35	0.99477	0.99078	0.99491	0.98545	0.98784
0.40	0.99353	0.98885	0.99338	0.98303	0.98577
0.45	0.99234	0.98710	0.99187	0.98087	0.98401
0.50	0.99124	0.98556	0.99039	0.97901	0.98256

*Optimal *k*-value

Notes

1 When estimating a nonlinear model of the Cobb-Douglas form, logarithmic transformations of variables on both sides of the equation permit the researcher to use standard linear regression techniques. The log transformation also reduces the adverse impact of heteroscedasticity.

2 In a strictly theoretical sense collinearity does not bias the estimated regression coefficients but it does increase their variance. That is, if the regression coefficients generated from a large number of samples drawn from the same population were averaged together, the *mean* value of the coefficients would equal the true population parameter, even in the face of severe multicollinearity. But from a practical perspective the researcher is normally dealing with a single estimation sample where the increased variance is likely to generate unstable, statistically insignificant, and possibly illogically signed regression coefficients.

3 One of the earliest discussions of using ridge regression to handle multicollinearity (MC) dates back to Hoerl and Kennard (1970). Over time a variety of other methods have been proposed to handle

the problems associated with MC. One approach frequently employed involves deleting the collinear variables. This is often not a true corrective measure since the researcher is tempted to respond to a symptom of the problem (i.e., low significance levels or inappropriate signs) and consequently may delete a variable that conceptually belongs in the model. This last condition may introduce a misspecification problem into the model. On the other hand, Belsley, Kuh and Welsch (1980) argue for the introduction of *additional* data to improve the statistical properties of the data set. Unfortunately, the researcher has no assurance that the new data, if available, would prove to be statistically independent. Morton (1977) suggests the use of factor analysis to screen real estate data prior to the use of OLS. Using factor analysis the ill effects of MC are reduced since only one variable from each factor or group of related variables is permitted to enter into the regression. While factor analysis sounds appealing, it introduces estimation problems of its own. For example, selecting only one variable to represent a group of similar variables makes it impossible to identify individual hedonic prices for each housing characteristic and may introduce specification errors similar to those encountered using variable deletion techniques.

Zellner (1971) argues in favor of using Baynesian estimation procedures to handle ill-conditioned data where the researcher possesses subjective prior information regarding the parameters to be estimated. Earlier, Theil and Goldberger (1960) advanced a related technique called mixed-estimation that incorporates auxiliary information into the ill-conditioned data matrix. Unfortunately, both of these techniques require extensive additional information which is generally not available. Finally, Mark (1983) advocates the use of principle component regression (PCR) which has much in common with factor analysis, with one important difference. Using PCR the researcher retains the entire set of variables in the model, with individual components constructed in a mutually orthogonal manner. While PCR is a biased estimator, there exists the potential for a reduction in the mean square error similar to ridge regression. In actual practice when PCR is applied to real estate data it consistently fails to provide superior results relative to OLS. (See Mark 1983, and Moore, Reichert & Cho, 1986.)

In a recent paper Reichert and Moore (1986) explored the use of latent root regression as a means of reducing the negative impact of "non-predictive MC," but concluded that since the majority of collinearity in real estate appraisal models is predictive in nature, the use of latent root regression techniques will have limited benefit. Thus, it seems evident from the brief literature review presented above that no single approach to handling MC is optimal in every case and that each technique represents a more or less ad hoc approach to handling a very difficult data problem.

[4] ISRM represents a measure of the lack of orthogonality in the regressor set and is related to the size of the eigenvalues of the matrix of regressor correlation coefficients. The optimal value for k is that value which minimizes ISRM. The NLMS technique assumes that for orthogonal regressors the larger the absolute value of the standardized regression coefficients the more statistically significant the coefficients. For each possible value of k, NLMS represents the value of the correlation coefficient between the absolute values of the regression coefficents and their t-statistics. The optimal k is the biasing factor that maximizes the size of this correlation coefficient. In addition to the Vinod citation, the reader may want to consult Vandell and Zerbst (1984) for another empirical application of these two techniques. (See Appendix B for selected test results.)

[5] Initial regression results indicated that the differences between ranch and two-story houses were statistically insignificant. Thus, in the model results presented in Exhibits 1-6 these two styles were combined into one category, with ranch/two-story serving as the base. The number of dummy variables used to represent a given housing characteristic (e.g., style) is equal to the number of unique categories for the specific housing characteristic, minus one for the base characteristic.

[6] In the most general formulation of the BC/BT transformation it is possible to have a unique lambda for each of the right-hand side variables. To simplify the current analysis, all the appropriate independent variables were treated in a similar fashion with one common lambda (λ_R) being estimated. However, the values for λ_L and λ_R are not constrained and are free to take different values.

[7] Since the Box-Cox transformation cannot be performed on dummy variables and other independent variables with zero values, the test results reported in this paper represent only a partial test of the optimal functional form. Any distortion this might cause in the results is minimized by the fact that the nonlinear relationships in the model relate most directly to the continuous variables, such as square footage, lot size, etc. Thus, the exclusion of the dummy variables probably does not constitute a major problem.

[8] The log of selling price is forecasted for the log-linear model. The antilog of the predicted log selling price generates the dollar value of the predicted selling price.

References

[1] J.E. Anderson. A Comparison of the Predictive Ability of OLS and Ridge Coefficients in Property Valuation. *Assessors Journal* 16 (December 1981), 187-93

[2] ——. Ridge Regression: A New Regression Technique Useful to Appraisers and Assessors. *The Real Estate Appraiser and Analyst* (November/December 1979), 48-51.

[3] D. Belsley, E. Kuh and R. Welsch. *Regression Diagnostics*. New York: John Wiley & Sons, 1980.

[4] G.E.P. Box and S.R. Cox. An Analyses of Transformation. *Journal of the Royal Statistical Society* Series B, 26 (1964), 211-43.

[5] —— and P.W. Tidwell. Transformation of the Independent Variables. *Technometrics* (November 1962), 531-50.

[6] R.W. Bruce and D.J. Sundell. Multiple Regression Analysis: History and Applications in the Appraisal Profession. *The Real Estate Appraiser* (January/February 1977), 37-44.

[7] T.P. Brennan, R.E. Cannaday and P.F. Colwell. Office Rent in the Chicago CBD. *Journal of the American Real Estate and Urban Economics Association* 12 (Fall 1984), 243-60.

[8] J.K. Brueckner and P.F. Colwell. A Spatial Model of Housing Attributes: Theory and Evidence. *Land Economics* 59 (February 1983), 58-69.

[9] R.E. Cannaday and Han Bin Kang. Estimation of Market Rent for Office Space. *The Real Estate Appraiser and Analyst* 50:2 (Summer 1984), 67-72.

[10] P.F. Colwell and R.E. Cannaday. Trade-offs in the Office Market. Working paper, University of Illinois, September 1985.

[11] —— and Chunchi Wu. The Analytical Foundations of Adjustment Grid Methods. *Journal of the American Real Estate and Urban Economics Association* (Spring 1983), 11-29.

[12] T.P. Cronan, D.R. Epley and L.G. Perry. The Use of Rank Transformation and Multiple Regression Analysis in Estimating Residential Property Values with a Small Sample. *The Journal of Real Estate Research* 1 (Fall 1986), 19-31.

[13] R.G. Edmonds. A Theoretical Basis for Hedonic Regression: A Research Primer. *Journal of the American Real Estate and Urban Economics Association* 12:1 (Spring 1984). 72-85.

[14] R.F. Gunst and R.L. Mason. Biased Estimation in Regression: An Evaluation Using Mean Squared Error. *Journal of the American Statistical Association* 72 (September 1977), 616-28.

[15] R. Halvorsen and H.O. Pollakowski. Choice of Functional Form for Hedonic Price Equation. *Journal of Urban Economics* 10 (July 1981), 37-49

[16] D.M. Hawkins. Relations Between Ridge Regression and Eigenanalysis of the Augmented Correlation Matrix. *Technometrics* 17 (November 1975), 477-80.

[17] C.T. Janssen and P.P. Talwar. Ridge Estimators in House Valuation Models. *Interfaces* (August 1982), 120-27.

[18] George G. Judge. *The Theory and Practice of Econometrics*. 1980, 308-11.

[19] J.G. Kowalski and P.F. Colwell. Market Versus Assessed Values of Industrial Land. Working paper, September 1985.

[20] J.H. Mark. An Empirical Examination of the Stability of Housing Price Equations Over Time. *Journal of the American Real Estate and Urban Economics Association* 11:3 (Fall 1983), 397-415.

[21] —— and G. Campbell. Multiple Regression Analysis and Mass Assessment: A Review of the Issues. Lincoln Institute of Land Policy, 1983.

[22] J.S. Moore, A.K. Reichert and C. Cho. Analyzing the Temporal Stability of Appraisal Model Coefficients an Application of Ridge Regression Techniques. *Journal of the American Real Estate and Urban Economics Association* (Spring 1984). 50-71.

[23] T.G. Morton. Factor Analyses, Multi-Collinearity and Regression Appraisal Models. *The Appraisal Journal* (October 1977), 578-88.

[24] J.M. Quigley. Nonlinear Budget Constraints and Consumer Demand: An Application to Public Programs for Residential Housing. *Journal of Urban Economics* 12 (1982), 177-201.

[25] A.K. Reichert and J. Moore. Using Latent Root Regression to Identify Nonpredictive Collinearity in Statistical Appraisal Models. *Journal of the American Real Estate and Urban Economics Association* 14:1 (Spring 1986), 136-52.

[26] S. Rosen. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy* 82:1 (January/February 1974), 34-55.

[27] S. Sharma and W.L. James. Latent Root Regression: An Alternate Procedure for Estimating Parameters in the Presence of Multicollinearity. *Journal of Marketing Research* 18 (May 1981), 154-61.

[28] J.J. Spitzer. A Primer on Box-Cox Transformation. *The Review of Economics and Statistics* 64 (May 1982), 307-13.
[29] H. Theil and A. Goldberger. On Pure and Mixed Statistical Estimation in Economics. *International Economic Review* 2 (1960), 65-78.
[30] R.A. Thisted and C.N. Morris. Theoretical Results for Adaptive Ordinary Ridge Regression Estimators. *Technical Report* No. 94. University of Chicago, 1980, 21.
[31] K.D. Vandell and R.H. Zerbst. Estimation of the Effect of School Desegregation Plans on Housing Value Over Time. *Journal of the American Real Estate and Urban Economics Association* 12:2 (Summer 1984), 109-35.
[32] H.D. Vinod. Application of New Ridge Regression Methods to a Study of Bell System Scale Economics. *Journal of the American Statistical Association* 71 (December 1976), 835-41.
[33] A. Zellner. *An Introduction to Baynesian Inference in Econometrics*. New York: John Wiley & Sons, 1971.