

PRACTICAL ECONOMETRIC ANALYSIS FOR ASSESSMENT OF REAL PROPERTY USING THE SAS® SYSTEM ON PERSONAL COMPUTERS

Shah Amal and Robert Weselowski
City of Winnipeg
Winnipeg, Manitoba

ABSTRACT

This study investigates the possibilities and potential of applying econometric (hedonic price) analysis for mass appraisal of Real property using the **PC SAS System**. There has been a tendency in the last decade by Real Property Taxation organizations to custom build or subscribe to customized mainframe modeling software. These systems are costly and are proving to be increasingly restrictive with the rapid changes in computer hardware and software technology. This research focuses on the capability of the **PC SAS System** in handling and computing over 53,000 records (residential property sales) with various property characteristics. The practical application of econometric models along with the relevant regression diagnostics as it relates to mass property valuation is also highlighted. This paper clearly demonstrates the power, speed, reliability, flexibility and cost efficiency of the **PC SAS System** as a Real property valuation and research tool.

INTRODUCTION

Property tax is a major source of local government revenue in Canada and the U.S.A. Local government services such as public works, public health, schools, parks, fire and police protection depend upon revenue generated from the property tax system. Netzer(1966) remarks that the property tax system is a paradox that is here to stay and will continue to play a major role in local government affairs. The property tax system is based on the 'ad Valorem' concept, that is the amount of tax paid is dependent on the value of property owned. This idea while simple in concept, has proved to be a complex task to implement (Netzer,1966; IAAO,1991). The mass appraisal of properties for assessment purposes is a major undertaking for many civic governments. The administration of a fair and equitable property tax system depends upon accurate property appraisal.

Recent advances in computer technology, both hardware and software, have transformed the assessment process into a modern mass appraisal system based on market data.

Econometric concepts could play a greater role in property valuations based on market data with these rapid changes in technology.

OBJECTIVE

This study investigates the possibilities and potential of applying econometric (hedonic price) analysis to residential property assessment using the **PC SAS System**. There has been a tendency in the last decade by Real Property Taxation organizations to custom build or subscribe to customized mainframe modeling software. These systems are costly and are

proving to be increasingly restrictive with rapid changes in computer hardware and software technology. Econometric modeling for the mass appraisal of properties both for valuation and research is not popular in Canada (Rosiers,1991). In part this could be due to the high costs involved with regression computations in the mainframe environment. This research project also focuses on the latest developments in the computer industry. Exciting technological advancements have been taking place in the Personal Computer industry. The software developers have kept abreast with these advancements. Complex computations with huge datasets are no longer the monopoly of large mainframe computer systems. It is the intent of this paper to demonstrate that econometric analysis of huge datasets such as those of mass appraisal systems, could be undertaken in the Personal Computer environment using the **PC SAS System**. It is envisioned that the Personal Computer environment is efficient as well as cost effective. In the literature dealing with the application of econometric models in property appraisal, there is very little reference to **Regression Diagnostics**. The use of simple (**Influence Statistics**) as well as **Specific Diagnostics**, which is readily available in the **SAS System** would be explored.

REVIEW OF LITERATURE

The Property Appraisal process is nothing but applied economic analysis. Transactions in real estate business subscribes to the rules of price theory. The real estate market place is a dynamic process where prices fluctuate around an equilibrium point dictated by such factors like political, social and enterprise factors (Graaskamp, 1981). Where property sales are adequate, appraisers try to quantify these determinants in an attempt to explain the real estate market place. Hass(1920) reported the first application of regression analysis in property appraisal. His article did not generate much interest due to the large number of computations involved with the regression technique. Griliches(1968) was one of the first to introduce Hedonic Price Indexes. Hedonic Price is conceptually the regression of price on various explanatory variables that represent its qualitative determinants. In recent years, with the advancements in data processing, there is a renewed interest in the application of regression analysis for property valuation (IAAO, 1991).

HEDONIC PRICING VALUATION

Hedonic pricing valuation is based on the equation,

$$p = f(x)$$

where, P = Price of Property

x = Vector of characteristics, X1, X2, Xn which describes the property.

There are some excellent reviews on the application of hedonic pricing valuation in the literature (Miller,1982; Church,1975; Ball,1973). In a practical sense, estimation of Hedonic Price equations is normally achieved using Multiple Regression Analysis (MRA). The availability of Statistical and Econometric procedures in the **SAS System** has made this technique easily accessible to the non academic world, including assessment administration.

REGRESSION DIAGNOSTICS

There has been concerns, however, that real estate data typically violates the underlying assumptions of econometric models (Gipe,1975). In particular the problem of multicollinearity is especially highlighted as a major problem area in the application of regression techniques to real estate data (IAAO,1974). There is however very little reference to multicollinearity diagnostics in particular and other regression diagnostics in general in publications dealing with the application of MRA to appraisal data. The **SAS System** provides the latest techniques in regression diagnostics.

NON SPECIFIC DIAGNOSTICS

Non specific diagnostics can be very useful in detecting data outliers, especially in smaller datasets. In recent years, there has been much interest in the field of Influence Statistics. **Belsley, Welch and Kuh** (1980) have developed a number of excellent **Influence Statistics**. These techniques have been incorporated as routine options in the **Regression Procedure of SAS/STAT**, 1991. There is no reference to Influence Statistics as a diagnostic tool in Appraisal literature.

A close examination of residuals is an important step in detecting outliers in a dataset. Residual analysis can also be helpful in detecting specification errors.

STUDENTIZED RESIDUALS

Studentized residuals are obtained by dividing the residuals by their Standard Errors. The Studentized Residuals are t distributed. **Belsley, Kuh and Welsch** (1980) suggest that, when error degrees of freedom exceed 10, it would be unusual for Studentized Residuals to exceed 2.5. Those that exceed this limit can be classified as having abnormally large residuals. The SAS output conveniently flags down these values by assigning five stars!

INFLUENCE STATISTICS

Another way of looking at outliers is to follow the changes in the various statistical estimates when a certain observation is not used in the regression analysis. This procedure helps to trace the potential influence of a particular observation. Some of these procedures are **RSTUDENT**, **HAT MATRIX DIAGONALS**, **COV RATIO**, **DFFIT** and **DFBETAS**.

MORE SPECIFIC DIAGNOSTICS

NON NORMALITY

The **UNIVARIATE PROCEDURE** in base SAS* (1991) is an excellent tool to detect non normality.

MULTICOLLINEARITY

The following is a brief discussion on the various techniques that can be employed to detect multicollinearity.

1. A very high R^2 with very few statistically significant t ratios. (Gujarati,1988).
2. High pairwise correlations among regressors.
3. **VARIANCE INFLATION FACTOR** (Belsley, Kuh and Welsch,1980). The VIF (for the i th independent variable) is defined as $1/(1-R^2)$, where R^2 is the coefficient of determination. This is a very useful procedure in defining which variable may be involved in the multicollinearity. Myers(1990) suggests VIF values exceeding 10 may be indicative of multicollinearity.
4. **EIGENVALUES** of zero are indicative of exact collinearities. A high degree of linear dependency is indicated by very small eigenvalues.
5. **CONDITION NUMBER** is defined as the ratio of the largest to the smallest eigenvalue. It is generally considered that a value exceeding 30 is indicative of multicollinearity (Belsley, Kuh and Welsch,1980).
6. **VARIANCE PROPORTIONS (VAR PROP)**. Belsley, Kuh and Welsch(1980) suggest that for **CONDITION NUMBER** greater than 30, the variance proportions of each independent variable should be checked. The variables with variance proportions greater than 0.5 are considered candidates of multicollinearity. The detail methodology for the variance proportion technique can be found in Belsley, Kuh and Welsch(1980).

HETEROSCEDASTICITY

A number of tests are available for detecting heteroscedasticity. Detail discussion on these tests can be found in standard econometric texts. Some of these tests are Bartlett's Chi Square test, Hartley's F test, Goldfeld-Quandt test, Breusch-Pagan test and White's Heteroscedasticity test. Among these tests only the White's Heteroscedasticity test (White,1980) is available in the SAS software. When heteroscedasticity is present the ACOV option of the SAS regression procedure provides a consistent estimate of the covariance matrix. If the data is from a homoscedastic sample a correct covariance matrix is produced.

AUTOCORRELATION

Several methods are available to test the assumption of autocorrelation. Most common among these is the Durbin-Watson test. This statistic provides a measure of association between successive values of the error term. As a rule of thumb values of D close to 2 means the errors are uncorrelated (Durbin and Watson,1950).

METHODOLOGY

Residential sales datasets were obtained from the City of Winnipeg Assessment Department. The software used for analysis were base SAS*, SAS/STAT and SAS/ETS. These software were chosen because of their extensive statistical and econometric procedures, as well as their excellent programming capabilities. The computer used is a 486 Personal Computer.

Five years (1987-1991) of single family residential sales along with core fields that describe property characteristics were obtained. The total dataset comprised over 53,000 observations. Data management required a fair amount of programming and data transfer facilities.

Traditionally, in assessment administration, the value of land is first determined through vacant land sales. The portion of sales attributable to the building itself is extracted from the total sale by subtracting the estimated land value. The same procedure was used in this analysis. Therefore the sales data field consisted of that portion of the sales that is attributed to the building alone.

The dataset consisted of the following variables (fields).

YEAR is the year of sale
GROSS_AR is the gross area of the building in square feet
RM is the number of rooms
GAR is a binary variable for garage
BATH_ROO is number of bathrooms
LOC is location (a subjective numeric score)
AGE is effective age of the building
AREA_COD is a area geographic stratification
TYPE_COD is type of building

Analysis was done on the total dataset as well as subsets of the total dataset. Subsets were created as samples of homogenous groups of data based on geographic stratification and building type codes. The following is a brief description of the datasets.

DATASET_A

This dataset comprised sales of 1991 for the geographic demarcation of newer St. Vital (code A46). The properties in this area code (A46) are all in the same age group (about 8 years). The properties in this dataset comprises a homogenous group. This dataset was further subdivided by building type codes. These subsets are as follows.

A46_15 for bungalows
A46_18 for split levels
A46_19 for bilevels

This small subsets were primarily created to test the **INFLUENCE STATISTICS** diagnostics. The dependent variable is sale prices while the independent variables were gross area, number of rooms, number of bathrooms and garage. Age was left out as the houses are all almost of the same age group. Basement was excluded as almost all the houses in this area code have basements. Location was left out because of its subjective scoring.

DATASET_B

This dataset comprised sales from all years with the same data structure as above.

B46_15 (bungalow sales from 1987-1991)
B46_18 (split level sales from 1987-1991)
B46_19 (bilevel sales from 1987-1991)

COMBINED DATASET

This dataset represents the combined (total) dataset of over 53,000 observations. This dataset was created primarily to test regression computation capabilities in the Personal Computer environment. Another aim was to check the stability of the coefficients in this huge heterogeneous dataset.

MODEL SPECIFICATION

LINEAR and **LOGLINEAR** models were used. An **AUTOREGRESSIVE** model on one of the datasets was also used.

LINEAR MODEL

$$\hat{Y} = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \dots + \beta_n(X_n) +$$

LOG LINEAR MODEL

$$\text{Log}(Y) = \beta_0 + \beta_1(\text{Log}(X_1)) + \beta_2(\text{Log}(X_2)) + \dots + \beta_n(\text{Log}(X_n)) +$$

RESULTS AND DISCUSSION

The results are first presented in the sequence of the datasets. Results from all the datasets are then discussed in the general context.

DATASET_A

For all the three sub datasets, cutoff points for the Influence Statistics were established. The individual observations that exceeded these cutoff points were flagged down and their respective values noted.

SUBSET A46_15 (Table 1)

STUDENT RESIDUAL (S.RES) and the STUDENT (RSTU) did not reveal any influential points. The HAT DIAGONALS (HAT D) for observations 1, 36 and 40 exceed the limit only by a fraction. However observation 72 may be considered as having high leverage. Almost the same observation is made using the COV RATIO (COV R) criteria. Only observation 72 seems to be a real influential point. Using the DFFITS (DFF) criterion, observation number 84 is a high influential data point. None of the DFBETAS (DFB) exceeded the cutoff point. From these influence statistics it can be clearly seen that observations 72 and 84 exert the greatest influence on the dataset.

SUBSET A46_18 (Table 2)

Table 1. Summary of Influence Statistics for dataset A46_15 (bungalows).

	S. RES	RSTU	HAT D	COV R	DFF	DFB
CUTOFF 2.5		2	0.11	1+.16	0.47	0.20
POINT	O V	O V	O V	O V	O V	
	N	N	1 0.12 31	1.19 36	0.49	N
			36 0.12 28	1.17 39	0.53	
			60 0.12 40	1.19 84	0.74	
			72 0.19 72	1.27 85	0.51	

Table 2. Summary of Influence Statistics for dataset A46_18 (split/s).

	S. RES	RSTU	HAT D	COV R	DFF	DFB
CUTOFF 2.5		2	0.23	1+.34	0.67	0.30
POINT	O V	O V	O V	O V	O V	
	4 2.75	4 3.03	4 0.40	5 1.35	1 0.85	N
				8 0.59	2 1.01	
				31 1.51	4 2.99	
				32 1.43	6 0.78	
				36 1.35	8 0.69	

O = Observation
V = Value
N = None

Using the same Influence Statistics as above, observation 4 is the most outstanding (influential) observation in this data subset.

SUBSET A46_19

From the various Influence Statistics, observations 19, 31 and 48 have consistently exceeded the cutoff points. These observations are therefore considered having the greatest influence on this data subset.

The purpose of this analysis was to demonstrate a practical method of flagging down sales that are not representative in an area. This method can be especially useful when appraiser analyze smaller homogenous property groups. The properties that were flagged down could be looked at more closely and other appraisal techniques could be applied.

Table 3. Summary of Influence Statistics for dataset A46_19 (bilevels).

	S. RES		RSTU		HAT D		COV R		DFF		DFB
CUTOFF POINT	2.5		2		0.19		1+.28		0.61		0.27
	O	V	O	V	O	V	O	V	O	V	
1	2.60	19	2.71	N			3	1.30	1	0.88	N
31	2.6	31	2.77				19	0.57	19	0.65	
48	2.8	48	3.05				31	0.58	31	0.93	
							34	0.69	47	0.76	
							48	0.70	48	2.24	
									51	1.02	

REGRESSION ANALYSIS

SUBSET A46_15

The R^2 for the analysis of this subset was 0.78. The coefficients for all the independent variables were highly significant. None of the assumptions were violated except for autocorrelation. The DW statistic indicated a high positive autocorrelation.

SUBSET A46_18

The R^2 was 0.64. Two independent variables (rooms and bath rooms) were not significant. The DW statistic indicated a positive autocorrelation. The rest of the regression assumptions were not violated.

SUBSET A46_19

The R^2 square for this subset was higher (0.87) than the other two subsets in this data category. The DW statistic of 1.83 suggests a lack of autocorrelation.

Table 4. Summary of Analysis for Dataset A

RESULT SUMMARY DATASET A46_15 (Bungalow)

R^2 0.78
Significance of Coefficients ($P < 0.01$) ALL SIGNIFICANT

Multicollinearity
VIF NO VIOLATION
EIGENVALUES NO VIOLATION
CONDITION INDEX NO VIOLATION

Heteroscedasticity (W) NO VIOLATION

Autocorrelation (DW) POSITIVE AUTOCORRELATION

DATASET A46_18 (Splits)

R^2 0.74

Significance of Coefficients ($P < 0.01$) ALL SIGN. except RM & BATH_ROO

Multicollinearity
VIF NO VIOLATION
EIGENVALUES NO VIOLATION
CONDITION INDEX NO VIOLATION

Heteroscedasticity (W) NO VIOLATION

Autocorrelation (DW) POSITIVE AUTOCORRELATION

DATASET A46_19 (BI_LEVELS)

R^2 0.87

Significance of Coefficients ($P < 0.01$) ALL SIGN. except RM

Multicollinearity
VIF NO VIOLATION
EIGENVALUES NO VIOLATION
CONDITION INDEX NO VIOLATION

Heteroscedasticity (W) NO VIOLATION

Autocorrelation (DW) POSITIVE AUTOCORRELATION

DATASET_B

This dataset comprised the same data structure as the above except that it includes sales from all the years (1987-1991). No Influence Statistics was performed due to the large number of sales. In addition to the regression diagnostic tests conducted on the previous dataset, nonnormality of residuals was tested using the Chi Square Test.

SUBSET B46_15 (Table 5)

The R^2 was 0.86. All the independent variables were highly significant. The residuals were not normally distributed. Heteroscedasticity was detected by both Bartlett's Chi Square

test and Hartley's F test. Autocorrelation was detected by both the Durbin-Watson and t tests. Multicollinearity was not a problem. Analyzing the data using a Log Linear model did not change the results or interpretation dramatically.

SUBSET B46_18 (Table 5)

The R^2 for this subset was 0.66. All the independent variables were highly significant. The residuals were not normally distributed. Heteroscedasticity nor multicollinearity were detected. Positive autocorrelation was a problem. The same conclusions were made using a log linear model.

Table 5. Summary of Analysis for Dataset B

Data - B46-15 (Bungalows) $R^2 = 0.8640$

RESIDUAL DIAG. TEST	VIOLATED	CRITICAL VALUE	DF	TEST
Non-Normality	yes	$\chi^2=7.85$	3	$^2=11.19$
Autocorrelation	yes	t=1.96	394	t=9.09
Heteroskedasticity B	yes	$\chi^2=7.85$	3	$^2=8.47$
H	yes	F=2.60	125	F=5.99
W	no			
Multicollinearity C.I.	no			
V.I.F.	no			

DATA -B46-18 (Split) $R^2 = 0.65$

Non-Normality	yes	$\chi^2=7.85$ t=1.96	3 283	$^2=16.85$ t=6.28
Autocorrelation	yes			DW=1.30
Heteroskedasticity B	no	$\chi^2=7.85$		$^2=0.36$
H	no	F=2.60		F=1.19
Multicollinearity C.I.	no			
V.I.F.	no			

DATA -B46-19 (BILEVELS) $R^2 = 0.81$

Non-Normality	yes	$\chi^2=7.85$	3	$^2=21.55$ t=8.96
Autocorrelation	yes	t=1.96	369	DW=1.15
Heteroskedasticity B	no	$\chi^2=7.85$	3	$^2=7.67$
H	yes	F=2.60	125	F=8.24
W	no			
Multicollinearity C.I.	no			
V.I.F.	no			

Legend: B is Bartlett's Test
H is Hartley's Test
W is White's Test
C.I. is Condition Index

SUBSET B46_19 (Table 5)

The R^2 was 0.81. The coefficients for all the independent variables were highly significant. The residuals were normally distributed. Positive autocorrelation was detected. Multicollinearity was not a problem. Bartlett's Chi Square test showed homoscedasticity. Hartley's F Test indicated heteroscedasticity. The same conclusions were made using a Log Linear model.

COMBINED DATASET

Analysis of the total dataset using LINEAR, LOGLINEAR and AUTOREGRESSIVE models did not change the pattern of the results obtained from the subsets.

CONCLUSIONS

This analysis conclusively demonstrates the applicability of econometric models for residential property valuation and research. Residential sales prices have been effectively partitioned into the various explanatory attributes that physically describe the properties. Only core characteristics that explain sales prices are considered in this research. This is certainly a limitation in terms of a thorough analysis. Nevertheless, for all the datasets considered, the amount of variation explained by the hedonic model was reasonably high. The R^2 was consistent with other research reports on hedonic price modeling of residential properties for mass appraisal purposes.

Majority of the coefficients were highly significant ($P>0.01$). The estimates of the coefficients appear fairly stable across datasets. The estimates from the linear model appears quite consistent with evidence obtained from the other traditional appraisal methods. The signs of the coefficients are in agreement with real estate expectations and corroborate other research publications.

The practical application of regression diagnostics in the context of hedonic price modeling is demonstrated in this research. Influence Statistics as a diagnostic procedure has been shown to be a very useful initial diagnostic tool for data screening. The techniques described are especially useful in mass appraisal valuation process, as all kinds of sales are entered in the assessment database.

The practical application of the different regression diagnostic techniques is demonstrated. Multicollinearity does not seem to be a major problem as reported in other studies. This in part could be due to the limited number of explanatory variables in the model.

Heteroscedasticity does not seem to be a major concern. White's (1984) test was consistently negative for all the datasets. However, Bartlett's Chi Square test and Hartley's F test did indicate non constant variances in some datasets.

Positive autocorrelation was consistently observed in almost all the datasets. The Combined Dataset was reanalysed using AUTO REGRESSIVE (GLS) procedure.

Log linear models were applied to some of the datasets. Contrary to other research findings, the results or conclusions did not alter significantly with the use of the log linear models in this study.

The results of this research are in general agreement with those of Sulock (1986). The Linear Multiple Regression Model provided estimates that are easy to explain to those involved in residential mass appraisal business. Subsets of the total dataset were created to group homogeneous properties. A simple and effective way of stratifying data is by geographical demarcation and by

building type codes. Sample comparisons of predicted and actual values for a heterogeneous and a homogenous dataset were made. Predictions for the stratified group were much more accurate than for the heterogeneous dataset.

This research project has clearly and conclusively demonstrated the practical applicability of hedonic price models as a cost effective valuation and research tool for mass appraisal of properties. The techniques discussed in this paper can be an excellent complement to the other traditional methods of assessment. Further research is warranted to develop working models that would include all the relevant explanatory variables. This paper clearly demonstrates the power, speed, reliability, flexibility and cost efficiency of the PC SAS System as a Real Property mass appraisal and research tool.

ACKNOWLEDGMENT

Help in the preparation of the manuscript by Jeannine Gelineau is greatly appreciated.

SAS, SAS/STAT and SAS/ETS are registered trademarks of SAS Institute Inc. in the USA and other countries. * indicates USA registration.

REFERENCES

1. BALL, M.J. (1973). Recent empirical work on the determination of relative house prices. *Urban Studies*. 9:213-34.
2. BELSLEY, D.A; KUHE and WELCH R.E.(1980). *Regression Diagnostics*. New York, John Wiley & Sons.
3. CHURCH, A.M.(1975). An econometric model for appraising. *American Real Estate and Urban Economic Association Journal* 3:17-31.
4. COOK, R.D. (1977). Influential observations in Linear Regression. *J. Ame. Stat. Assoc.* 74:169-174
5. COOK, R.D and WEISBERG, S (1962). *Residuals and Influence in Regression*. London, Chapman & Hall.
6. DURBIN, J and WATSON, G.S.(1951). Testing for serial correlation in Least Square Regression. *Biometrika* 37:409-428
7. Des ROSIERS, F(1991). RESIVALUE: An hedonic residentil price model for the Quebec region. *Property TAX Journal* 10:2.
8. GUJARATI, D.N. (1988). *Basic Econometrics* (2nd Ed.) McGraw Hill Book Comp. New York.
9. GRAASKAMP, A (1981). A framework for real estate development. Washington D.C. Urban Land Institute.
10. GRILICHES, Z. (1968). Hedonic Price Indexes for automobiles. An econometric analysis of quality change. *Readings in Economic Statistics & Econometrics*. A. Zellner (Ed.) Little Brown. Boston.
11. HAAS, C.G. (1920). Prices as a basis for farm land appraisal. *Minnesota Agricultural Exp. Sta. Tech. Bull.* No. 9.
12. GIPE, G.W. (1975) Understanding Multiple REgression Analysis. *Assessors Journal*. 10:1-3
13. IAAO (1990). *Property Appraisal and Assessment Administration*. The International Association of Assessing Officers. Chicago.
14. IAAO (1974). *The application of Multiple Regression in Assessment Administration*. IAAO Studies in Property Taxation. Chicago.
15. JOHNSTON, J. (1984). *Econometric Methods*. 3rd Ed. McGraw Hill Comp. New York. 1984.
16. MILLER, N.G. (1982). Residential property, hedonic price models. A Review. *Research in Real Estate*.
17. MYERS, R.H. (1990). *Classical and Modern REgression with application* (2nd. Ed.). Boston. PWS & Kent Pub. Comp. Inc.
18. NETZER, D. (1966). *Economics of the Property Tax*. Washington, D.C. Brookings Institute.
19. PINDYCK, R.S. and RUBINFELD, D.L. (1981) . *Econometric Models and Econometric Forecasts* (2nd. Ed.) McGraw Hill Book Comp. New York.
20. SAS (1991). Version 6.03. Cary. North Carolina.
21. WHITE, HALBERT (1980). A Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrics* 48:817-830.