Malaysian Journal of Mathematical Sciences 3(1): 27-44 (2009)

A Case Study on Determination of House Selling Price Model Using Multiple Regression

H. J. Zainodin and G. Khuneswari
School of Science and Technology
Universiti Malaysia Sabah, Locked Bag No. 2073,
88999 Kota Kinabalu, Sabah, Malaysia
E-mail: zainodin@ums.edu.my

ABSTRACT

This research illustrated the procedure in selecting the best model in determining the selling price of house using multiple regression for the data set which was collected in Oxford, Ohio, in 1988. The five independent variables considered in this data set are: floor area (square feet), number of rooms, age of house (years), number of bedrooms and number of bathrooms. The multiple regression models were involved up to the fourth-order interaction and there were 80 possible models considered. To enhance the understanding of the whole concept in this work, multiple regression with eight selection criteria (8SC) had been explored and presented. In this work the process of getting the best model from the selected models had been illustrated. The progressive elimination of variables with the highest p-value (individual test) was employed to get the selected model. In conclusion the best model obtained in determining the house selling price was M73.15 (ie. 73^{rd} model).

Keywords: multiple regression, fourth-order interaction variables, eight selection criteria (8SC), progressive elimination of variables

INTRODUCTION

Regression analysis is the process of finding a mathematical model that best fits the data. Often sample data is used to investigate the relationship between two or more variables. The ultimate goal is to create a model that can be used to predict the value of a single variable. Multiple regression is the extension of simple regression. Usually, a model is simply called an equation. Model can be used to predict weather, the performance of the stock market, sales, profits, river levels and so on. Nikolopoulos *et al.* (2007) suggested that multiple linear regression is a common choice of method when a forecast is required and where data on several relevant independent variables are available. The technique has been used to produce forecasts in a wide range of areas and there is evidence that it is often used by companies to derive forecasts of demand from marketing variables and various macroeconomic measures.

Multiple regression has been effectively used in many business applications. For example, Evans and Olson (2003) studied the 2000 NFL data, it would be logical to suspect that the number of Games Won would depend not only on Yards Gained but also on the other variables like Takeaways, Giveaways, Yards Allowed and Points Scored.

Multiple linear regression is a popular method for producing forecasts when data on relevant independent variables are available. In this study, Nikolopoulos *et al.*, (2007) compared the accuracy of the technique in forecasting the impact on Greek TV audience shares of programmes showing sports events with forecasts produced by a simple bivariate regression model. Three different types of artificial neural network, three forms of nearest neighbour analysis and human judgment. The data used in this study is a television audience rating from 1996 to 2000 in Greece.

Nikolopoulos *et al.*, (2007) study shows that the multiple regressions models performed relatively badly as a forecasting tool and were outperformed by either conceptually simpler method like the bivariate regression model and nearest neighbour analysis. The multiple regression models were also outperformed badly compared to complex method like artificial neural method and forecasts based on human judgement. The relatively poor performance of multiple linear regression appears to result both from its tendency to over fit in sample data and its inability to handle complex non-linearities in the data. Forecasts based on a simple bivariate regression model, two types of artificial neural network and a simple nearest neighbour analysis shows higher accuracy than a multiple linear regression.

MULTIPLE REGRESSION

Multiple regression is a generalization of the simple linear regression analysis. Simple regression analysis could analyze a relationship between a dependent variable with a single independent variable. The same idea was used to analyze relationship between a dependent variable with two or more independent variables.

Several variables as $X_1, X_2, ..., X_k$ capable of providing a better prediction of the value Y where k is the number of variables (with K =k+1 is the number of parameters). Lind *et al*., (2005) defines the general multiple regression model as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

where,

Y_i is random variable representing the *i*th value of the dependent variable Y

$X_{1i}, X_{2i}, ..., X_{ki}$ are the *i*th value of independent variable for $i = 1, 2, ..., n$.

Basic assumptions of multiple regression models are made about the error terms u_i and the values of independent variables $X_1, X_2, ..., X_k$ as following (Kenkel, 1996):

a. *Normality*: For any value of the independent variables, the error term u_i is a normally distributed random variable.
b. *Zero mean*: For any set of values of the independent variables, $E(u_i) = 0$.
c. *Homoscedasticity*: The variance of u_i denoted as σ_u^2 is the same for all values of the independent variables.
d. *No serial correlation*: The error terms are independent of one another for $i \neq j$.
e. *Independence of u_i and X_{ji}:* The error terms u_i are independent of the values of the independent variables X_{ji}. The independent variables are either fixed numbers or random variables that are independent of the error terms. If the X_{ji}'s are random variables, then all inferences are carried out conditionally on the observed values of the X_{ji}'s.

In this study the variables are the selling price (Y) of a house to its characteristics such as square feet (X_1), number of rooms(X_2), number of bedrooms (X_3), age of the house (X_4) and number of bathrooms (X_5). The way to determine the possible models are shown in the Table 1 below:

TABLE 1: All possible models

Number of Variables	Individual	INTERACTION				
		First Order	Second Order	Third Order	Fourth Order	TOTAL
1	5	-	-	-	-	**5**
2	10	10	-	-	-	**20**
3	10	10	10	-	-	**30**
4	5	5	5	5	-	**20**
5	1	1	1	1	1	**5**
TOTAL	**31**	**26**	**16**	**6**	**1**	**80**

With five variables there are 80 models with interactions. SPSS is needed to choose the selected models from all possible models. The SPSS output will show the model summary table, ANOVA table and Coefficients table. The procedures in obtaining a selected model after the first multiple regression analysis run in SPSS are as below (Lind *et al.*, 2005):

i. Drop the independent variable with the highest p-value (only one variable will be dropped each time) and rerun the analysis with the remaining variables
ii. Conduct individual test on the new regression equation. If there are still regression coefficient that are not significant (p-value > α), drop the variable with the highest p-value again
iii. Repeat the steps above until the p-value of each variable are significant

After the procedure of obtain the selected model, the model selection criteria will be used to choose the best model. The measure of goodness of fit R^2 (coefficient of multiple determination), $\bar{R}^2$ (adjusted coefficient of multiple determination) and SSE (Sum of square Error) are the most commonly used criteria for model comparison. R^2 will clearly lie between 0 and 1. The closer the observed points are to the estimated straight line, the better the "fit", which means that SSE will be smaller and R^2 will be higher. $\bar{R}^2$ is a better measure of goodness of fit because its allows for the trade-off between increased R^2 and decreased degree of freedom. SSE is the unexplained variation because $\hat{u}_t$ is the effect of variables other than X_t that are not in the model. The R^2, $\bar{R}^2$ and SSE has weakness in selecting the best model. The R^2 did not consider the number of parameters included in the model and $\bar{R}^2$ is useful only to determine the fraction of the variation in Y explained by the Xs.

TABLE 2: Model Selection Criteria

EIGHT SELECTION CRITERIA (8SC) **K= number of estimated parameters, n=sample size, SSE=sum of square errors**			
AIC:	$\left(\frac{SSE}{n}\right)(e)^{(2K/n)}$	**RICE:**	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{2K}{n}\right)\right]^{-1}$
FPE:	$\left(\frac{SSE}{n}\right)\frac{n+K}{n-K}$	**SCHWARZ:**	$\left(\frac{SSE}{n}\right)n^{K/n}$
GCV:	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{K}{n}\right)\right]^{-2}$	**SGMASQ:**	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{K}{n}\right)\right]^{-1}$
HQ:	$\left(\frac{SSE}{n}\right)(\ln n)^{2K/n}$	**SHIBATA:**	$\left(\frac{SSE}{n}\right)\frac{n+2K}{n}$

Recently several criteria to choose the best model have been proposed. These criteria take the form of the sum of square error (SSE) multiplied by a penalty factor that depends on complexity of the model. A more complex model will reduce SSE but raise the penalty. A model with a lower value of a criterion statistics is judged to be preferable. The model selection criteria are finite prediction error (FPE), Akaike information criterion (AIC), Hannan and Quinn criterion (HQ criterion), SCHWARZ, SHIBATA, RICE, generalized cross validation (GCV) and sigma square (SGMASQ). Finite prediction error (FPE) and Akaike information criterion (AIC) was developed by Akaike (1970, 1974). HQ criterion was suggested by Hannan and Quinn in 1979. Golub *et al.* (1979) developed generalized cross validation (GCV). Other criteria are included SCHWARZ (Schwarz, 1978), SHIBATA (Shibata, 1981) and RICE (Rice, 1984). Table 2 shows the model selection criteria (Ramanathan, 2002).

ANALYSIS

The data used in this study is collected in Oxford, Ohio during 1988. In this study, we are relating the sales price (Y) of a house to its characteristics such as floor area in square feet (X_1), number of rooms (X_2), number of bedrooms (X_3), the age of the house (X_4) and number of bathrooms (X_5). We analyse what is the contribution of a specific attribute is determining the sales price. The data collected for each of 63 single-family residences sold during 1988 in Oxford, Ohio.

TABLE 3: A correlation table for sales price and its characteristics.

		sales_price	X_1	X_2	X_3	X_4	X_5
sales_price	Pearson Correlation	1	.785(**)	.580(**)	.512(**)	-.289(*)	.651(**)
	Sig. (2-tailed)		.000	.000	.000	.021	.000
X_1	Pearson Correlation	.785(**)	1	.711(**)	.754(**)	-.109	.628(**)
	Sig. (2-tailed)	.000		.000	.000	.395	.000
X_2	Pearson Correlation	.580(**)	.711(**)	1	.722(**)	.170	.402(**)
	Sig. (2-tailed)	.000	.000		.000	.183	.001
X_3	Pearson Correlation	.512(**)	.754(**)	.722(**)	1	.017	.352(**)
	Sig. (2-tailed)	.000	.000	.000		.893	.005
X_4	Pearson Correlation	-.289(*)	-.109	.170	.017	1	-.409(**)
	Sig. (2-tailed)	.021	.395	.183	.893		.001
X_5	Pearson Correlation	.651(**)	.628(**)	.402(**)	.352(**)	-.409(**)	1
	Sig. (2-tailed)	.000	.000	.001	.005	.001	

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 3 shows the relationship between selling price of a house and its characteristics such as floor area in square feet, number of rooms, number of bedrooms, the age of the house and number of bathrooms. There is a significant positive relationship (correlation coefficient) between selling price and square feet, indicating that selling price increase as the square feet increases ($r = 0.785$, p-value < 0.0001). There is a significant positive relationship between selling price and number of rooms, that the selling price increase as the number of rooms increase ($r = 0.580$, p-value < 0.001). The relationship between selling price and number of bedrooms is significant and positive relationship ($r = 0.512$, p-value < 0.001). Besides that there is a significant negative relationship between sales price and the age of the house, indicate that selling price decreases as the age of the house increase ($r = -0.289$, p-value < 0.001). Selling price and number of bathrooms has a significant positive relationship where selling price increases as the number of bathrooms increase ($r = 0.651$, p-value < 0.001). The relationship between independent variables (X_1, X_2, X_3, X_4 and X_5) shows that there is no multicollinearity.

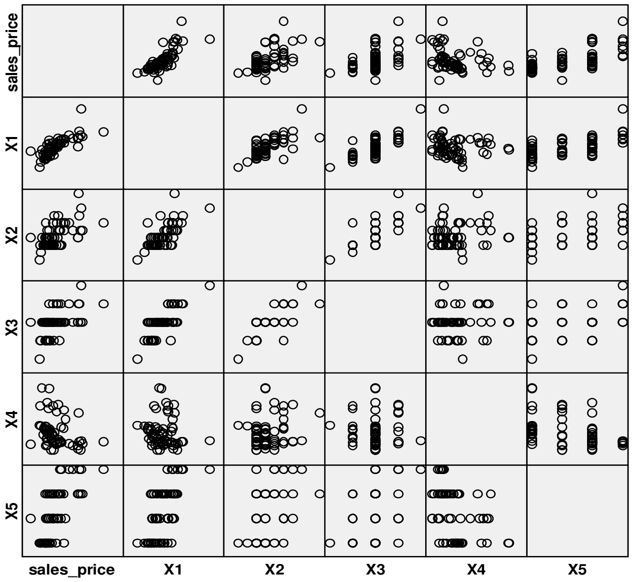

Figure 1: The matrix scatter plot of selling price, floor area in square feet(X_1), number of rooms (X_2), number of bedrooms (X_3), the age of the house (X_4) and number of bathrooms (X_5).

All the possible models are subjected to individual test (based on p-value). For illustration purpose, consider model M67 where Table 4 shows the p-value for each variable of the model. As can be seen from Table 4, each variable has p-value higher than 0.05 which means that the corresponding independent variable is not significant. Hence, by omitting the variable with highest p-value that is variable X_3 (p-value =0.934) and rerun the analysis with remaining variables. The resulting p-value after eliminating variable X_3 is shown in Table 5.

TABLE 4**:** The p-values and coefficient of variables in M67

Variables	Unstandardized Coefficients		Standardized Coefficients	t	p-value
	B	Std. Error	Beta		
(Constant)	-8.934	129.343		-0.069	0.945
X_3	3.533	42.723	0.067	0.083	0.934
X_4	2.711	3.681	1.773	0.736	0.465
X_5	37.513	69.879	0.545	0.537	0.594
X_{34}	-.648	1.218	-1.403	-0.532	0.597
X_{35}	4.421	22.360	0.279	0.198	0.844
X_{45}	-2.496	2.355	-2.212	-1.060	0.294
X_{345}	0.601	0.759	1.910	0.792	0.432

TABLE 5**:** The p-values and coefficient after eliminating variable X_3

Variables	Unstandardized Coefficients		Standardized Coefficients	t	p-value
	B	Std. Error	Beta		
(Constant)	1.634	19.776		0.083	0.934
X_4	2.435	1.551	1.593	1.570	0.122
X_5	32.073	23.372	0.466	1.372	0.175
X_{34}	-0.556	0.489	-1.204	-1.137	0.260
X_{35}	6.211	5.535	0.392	1.122	0.267
X_{45}	-2.337	1.356	-2.071	-1.724	0.090
X_{345}	0.549	0.416	1.744	1.321	0.192

TABLE 6: The p-values and coefficient after eliminate variable X_{35}

Variables	Unstandardized Coefficients		Standardized Coefficients	t	p-value
	B	Std. Error	Beta		
(Constant)	-5.787	18.680		-0.310	0.758
X_4	3.308	1.345	2.164	2.459	0.017
X_5	55.489	10.551	0.807	5.259	0.000
X_{34}	-0.800	0.439	-1.732	-1.824	0.073
X_{45}	-3.396	0.976	-3.010	-3.479	0.001
X_{345}	0.863	0.308	2.742	2.805	0.007

From Table 5, the variables in the new regression equation are not significant because all the variables had p-value larger than 0.05. The variable X_{35} (p-value =0.267) omitted from the model and rerun the analysis with the remaining variables. The new set of p-values after eliminating variable X_{35} is shown in Table 6. As can be seen from Table 6, the variable X_{34} is not significant (p-value > 0.05), X_{34} is omitted from the model and rerun the analysis. The p-values after eliminating variable X_{34} are shown in Table 7.

TABLE 7**:** The coefficient after eliminate variable X_{34}

Variables	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**p-value**
	B	**Std. Error**	**Beta**		
(Constant)	-7.329	19.031		-0.385	0.702
X_4	1.019	0.493	0.666	2.067	0.043
X_5	56.892	10.732	0.827	5.301	0.000
X_{45}	-1.825	0.468	-1.617	-3.899	0.000
X_{345}	0.323	0.086	1.027	3.771	0.000

The Table 7 shows that all the remaining independent variables are significant where the p-value of each variable is less than 0.05. Thus, after the 3 variables had been omitted a selected model is obtained i.e. model M67.3 where $Y = -7.329 + 1.019X_4 + 56.892X_5 - 1.825X_{45} + 0.323X_{345}$.

Similar procedures are carried to all possible models systematically. At the end of the procedure, altogether there are 47 selected models obtained and their summary is shown in Table 8. For each selected model, find the value of each criterion mentioned in Table 2 and corresponding values are shown in Table 9.

Majority of the criteria shown in Table 8 indicates that model M73.15 is the best model.

TABLE 8: The summary for selected models

Selected Model	Summary	K =k+1	SSE
M1	M1	**2**	30998.9710
M2	M2	**2**	53609.0350
M3	M3	**2**	59601.8130
M4	M4	**2**	74012.6580
M5	M5	**2**	46527.2500
M8	M8	**3**	27603.1960
M9	M9	**3**	27654.2200
M11	M11	**3**	41096.4610
M12	M12	**3**	36779.0590
M13	M13	**3**	52416.2600
M14	M14	**3**	39148.0960
M24	M24	**4**	34185.3180
M34	M34.1	**3**	27328.9650
M35	M35.1 => M35.2	**2**	28355.1480
M36	M36.1 => M36.2	**2**	52105.2640
M37	M37.1	**3**	40397.9520
M38	M38.1 => M38.2	**2**	35395.2180
M39	M39.1	**3**	51824.6620
M40	M40.1 => M40.2	**2**	36751.5540
M43	M43.1 => M43.2	**4**	24045.0190
M44	M44.1 => M44.2 => M44.3 => M44.4	**3**	27524.7090
M46	M46.1 => M46.2 => M46.3	**6**	23804.6840
M48	M48.1 => M48.2 => M48.3 => M48.4	**3**	39493.5960
M50	M50.1 => M50.2 => M50.3 => M50.4	**3**	31466.8550
M52	M52.1 => M52.2 => M52.3 => M52.4 => M52.5 => M52.6	**5**	22116.7010
M53	M53.1 => M53.2 => M53.3 => M53.4 => M53.5 => M53.6 => M53.7	**4**	26426.4150
M54	M54.1 => M54.2 => M54.3 => M54.4 => M54.5 => M54.6 => M54.7	**4**	24110.5790
M57	M57.1 =>… => M57.10	**6**	21774.7130
M58	M58.1 => M58.2 => M58.3 => M58.4	**4**	26735.4780
M59	M59.1 => M59.2 => M59.3 => M59.4	**4**	25591.7630
M60	M60.1 => M60.2 => M60.3	**5**	24703.7830
M62	M62.1 => M62.2 => M62.3 => M62.4	**4**	25499.6560
M63	M63.1 => M63.2 => M63.3 => M63.4	**4**	25837.9670
M66	M66.1 => M66.2 => M66.3 => M66.4 => M66.5	**3**	32497.3660
M67	M67.1 => M67.2 => M67.3	**5**	35837.7440
M68	M68.1 =>… => M68.8	**7**	19300.7880
M69	M69.1 =>… => M69.9	**6**	21734.8560
M70	M70.1 =>… => M70.10	**5**	22732.4740

TABLE 8 (continued): The summary for selected models

Selected Model	Summary	K =k+1	SSE
M71	M71.1 =>… => M71.10	**5**	24178.5540
M72	M72.1 =>… => M72.11	**4**	29244.9580
M73	M73.1 =>… => M73.15	**11**	15073.4450
M74	M74.1 =>… => M74.10	**7**	19300.7880
M75	M75.1 =>… => M75.11	**5**	21565.1040
M76	M76.1 =>… => M76.9	**7**	20962.8220
M77	M77.1 =>… => M77.12	**4**	25499.6560
M79	M79.1 =>… => M79.22	**9**	16634.4070
M80	M80.1 =>… => M80.19	**13**	14840.3610

TABLE 9: The corresponding selection criteria value for the selected models

Selected Model	R^2	Adj R^2	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M1	0.616	0.610	524.301	524.313	524.841	538.520	525.406	561.214	508.180	523.288
M2	0.336	0.325	906.717	906.736	907.651	931.307	908.628	970.553	878.837	904.965
M3	0.262	0.250	1008.076	1008.097	1009.114	1035.415	1010.200	1079.048	977.079	1006.128
M4	0.084	0.069	1251.814	1251.840	1253.103	1285.763	1254.452	1339.946	1213.322	1249.395
M5	0.424	0.415	786.939	786.956	787.750	808.281	788.597	842.343	762.742	785.418
M8	0.658	0.647	481.926	481.961	483.056	501.663	484.267	533.706	460.053	479.874
M9	0.658	0.646	482.817	482.851	483.949	502.590	485.162	534.692	460.904	480.761
M11	0.491	0.474	717.506	717.557	719.188	746.891	720.991	794.597	684.941	714.451
M12	0.545	0.529	642.128	642.174	643.634	668.426	645.247	711.120	612.984	639.394
M13	0.351	0.329	915.139	915.205	917.285	952.618	919.584	1013.464	873.604	911.243
M14	0.515	0.499	683.489	683.538	685.092	711.481	686.809	756.925	652.468	680.579
M24	0.577	0.555	616.095	616.200	618.694	649.965	621.551	705.900	579.412	611.529
M34.1	0.662	0.650	477.138	477.172	478.257	496.679	479.456	528.403	455.483	475.107
M35.2	0.649	0.643	479.585	479.595	480.079	492.591	480.596	513.350	464.838	478.658
M36.2	0.355	0.344	881.283	881.302	882.191	905.183	883.140	943.329	854.185	879.580
M37.1	0.500	0.483	705.310	705.361	706.964	734.196	708.736	781.091	673.299	702.308
M38.2	0.562	0.555	598.657	598.670	599.274	614.893	599.919	640.805	580.249	597.501
M39.1	0.358	0.337	904.810	904.875	906.932	941.866	909.205	1002.026	863.744	900.958
M40.2	0.545	0.538	621.598	621.611	622.238	638.456	622.908	665.361	602.484	620.397
M43.2	0.702	0.687	433.344	433.418	435.173	457.168	437.182	496.510	407.543	430.133
M44.4	0.659	0.648	480.556	480.590	481.682	500.237	482.890	532.188	458.745	478.510
M46.3	0.705	0.679	457.135	457.400	461.587	495.346	466.759	560.645	417.626	449.824
M48.4	0.511	0.495	689.521	689.571	691.138	717.760	692.870	763.605	658.227	686.586
M50.4	0.610	0.597	549.382	549.421	550.670	571.881	552.050	608.409	524.448	547.043
M52.6	0.726	0.707	411.448	411.586	414.195	439.915	417.296	487.736	381.322	406.782

TABLE 9 (continued): The corresponding selection criteria value for the selected models

Selected Model	R^2	Adj R^2	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M53.7	0.673	0.656	476.262	476.344	478.272	502.445	480.480	545.684	447.905	472.733
M54.7	0.702	0.686	434.526	434.600	436.359	458.414	438.374	497.864	408.654	431.305
M57.10	0.730	0.707	418.152	418.395	422.224	453.105	426.955	512.835	382.013	411.465
M58.4	0.669	0.652	481.832	481.915	483.865	508.321	486.100	552.066	453.144	478.261
M59.4	0.683	0.667	461.220	461.299	463.166	486.576	465.305	528.450	433.759	457.802
M60.3	0.694	0.673	459.577	459.731	462.645	491.373	466.109	544.788	425.927	454.365
M62.4	0.684	0.668	459.560	459.639	461.499	484.825	463.630	526.548	432.198	456.154
M63.4	0.680	0.664	465.657	465.737	467.622	491.257	469.781	533.533	437.932	462.206
M66.5	0.598	0.584	567.373	567.414	568.704	590.610	570.129	628.334	541.623	564.958
M67.3	0.556	0.526	666.708	666.931	671.159	712.835	676.184	790.324	617.892	659.147
M68.8	0.761	0.735	382.599	382.952	387.739	420.164	393.894	485.469	344.657	374.442
M69.9	0.731	0.707	417.387	417.629	421.452	452.275	426.174	511.896	381.313	410.712
M70.10	0.719	0.699	422.904	423.045	425.727	452.163	428.915	501.315	391.939	418.108
M71.10	0.701	0.680	449.806	449.957	452.809	480.926	456.199	533.205	416.872	444.705
M72.11	0.638	0.620	527.059	527.149	529.282	556.034	531.727	603.885	495.677	523.152
M73.15	0.813	0.778	339.258	340.487	351.193	393.049	367.645	493.222	289.874	322.813
M74.10	0.761	0.735	382.599	382.952	387.739	420.164	393.894	485.469	344.657	374.442
M75.11	0.733	0.715	401.187	401.321	403.865	428.943	406.889	475.571	371.812	396.637
M76.9	0.740	0.713	415.546	415.929	421.128	456.345	427.813	527.273	374.336	406.686
M77.12	0.684	0.668	459.560	459.639	461.499	484.825	463.630	526.548	432.198	456.154
M79.22	0.794	0.764	351.359	352.051	359.385	396.320	369.653	477.216	308.045	339.478
M80.19	0.816	0.772	355.907	358.053	373.977	423.521	401.091	553.855	296.807	332.777

Result of the individual test of the Model M73.15 are shown in Table 10 (all the p-values < 0.05) and the corresponding result global test is shown in Table 11.

TABLE 10: The final coefficients of model M73.15

Variables	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**p-value**
	B	**Std. Error**	**Beta**		
(Constant)	101.891	26.906		3.787	0.000
x2	-26.829	7.454	-1.162	-3.599	0.001
x4	-2.615	0.733	-1.710	-3.565	0.001
x12	0.041	0.009	6.538	4.581	0.000
x15	-0.074	0.025	-3.255	-2.990	0.004

TABLE 10 (continued): The final coefficients of model M73.15

Variables	Unstandardized Coefficients		Standardized Coefficients	t	p-value
	B	Std. Error	Beta		
x45	3.128	0.990	2.772	3.158	0.003
x123	-0.009	0.002	-6.966	-4.412	0.000
x135	0.028	0.008	5.378	3.654	0.001
x145	-0.001	0.000	-2.082	-3.043	0.004
x234	0.155	0.040	2.802	3.911	0.000
x345	-0.558	0.258	-1.772	-2.162	0.035

Thus, the best model is M73.15 where

$$Y = 101.891 - 26.829X_2 - 2.615X_4 + 0.41X_{12} - 0.017X_{15}$$
$$+ 3.128X_{45} - 0.009X_{123} + 0.028X_{135} - 0.001X_{145} + 0.155X_{234} - 0.558X_{345}$$

The house selling price will decreases 26.829 times when number of rooms (X_2) increases by 1 unit. For variable X_4, the house selling price decrease 2.615 times when age of the house (X_4) increases by 1 unit. When the interaction effect between square feet (X_1) and X_2 increase 1 unit, the house selling price increase by 0.41 times. The constant shows that the starting house sales price is predicted as 101.891.

TABLE 11: The ANOVA table of global test for model M73.15

Source of variations	Sum of Squares	df	Mean Square	F	p-value
Regression	65701.989	10	6570.199	22.666	0.000
Residual	15073.445	52	289.874		
Total	80775.434	62			

For a clear view, the house selling price increase when square feet of house interact with number of rooms and also when age interact with bathrooms. When square feet, number of bedrooms and number of bathrooms interact together the house selling price will increase. House selling price also will increase when number of bedrooms, age and number of bathrooms interact together.

There are 15 variables omitted from the model M73. A Wald Test is carried out to the final model (Ramanathan, 2002) where the restricted model (M73.15) is the selected model and unrestricted model is the initial possible model (M73).

The unrestricted model (Possible Model):

U:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_{12}X_{12}+\beta_{13}X_{13}+\beta_{14}X_{14}+\beta_{15}X_{15}+\beta_{23}X_{23}+\beta_{24}X_{24}+\beta_{25}X_{25}+\beta_{34}X_{34}+\beta_{35}X_{35}$$
$$+\beta_{45}X_{45}+\beta_{123}X_{123}+\beta_{124}X_{124}+\beta_{125}X_{125}+\beta_{134}X_{134}+\beta_{135}X_{135}+\beta_{145}X_{145}+\beta_{234}X_{234}+\beta_{235}X_{235}+\beta_{245}X_{245}+\beta_{345}X_{345}+u$$

The restricted model (Selected Model):

R:

$$Y=\beta_0+\beta_2X_2+\beta_4X_4+\beta_{12}X_{12}+\beta_{15}X_{15}+\beta_{45}X_{45}+\beta_{123}X_{123}+\beta_{135}X_{135}+\beta_{145}X_{145}+\beta_{234}X_{234}+\beta_{345}X_{345}+v\ .$$

The hypothesis:

H_0:
$$\beta_1=\beta_3=\beta_5=\beta_{13}=\beta_{14}=\beta_{23}=\beta_{24}=\beta_{25}=\beta_{34}=\beta_{35}=\beta_{124}=\beta_{125}=\beta_{134}=\beta_{235}=\beta_{245}=0$$

H_1: At least one β_s is nonzero

Decision:

$$F_{cal}=\frac{(SSE_R-SSE_U)/(DF_R-DF_U)}{SSE_U/DF_U}$$

$$=\frac{(SSE_R-SSE_U)/(K-m)}{SSE_U/(n-K)}$$

$$= 0.23468$$

F_{table} =F(21, 36, 5%) =1.92. Since F_{calc} is less than F_{table}, H_0 is accepted. Thus, this is justified (Lind *et al.*, 2005). The similar procedure of Wald Test is carried out for all other selected models and same results are obtained.

Based on the best model, the predicted Y was determined. Using the residuals obtained, randomness test is carried out. Both randomness test and residual scatter plot indicates that the residuals are random and independent. That means the model M73.15 is the best model to describe the house selling price in Ohio and it's ready to be used for further analysis.

DISCUSSION & CONCLUSION

To minimize the effects of bias, SPSS exclude temporarily variables that contribute to multicollinearity (when there exists a high level of correlation between some of the independent variables). Multicollinearity has some effects in finding the final model. Thus, careful selection/treatment should be taken at initial stage. Since there exist effect of higher order interaction, polynomial of higher order interaction should be included in the possible models. Other variables such as number of garage, location of the house and other relevant characteristics should be considered for future study.

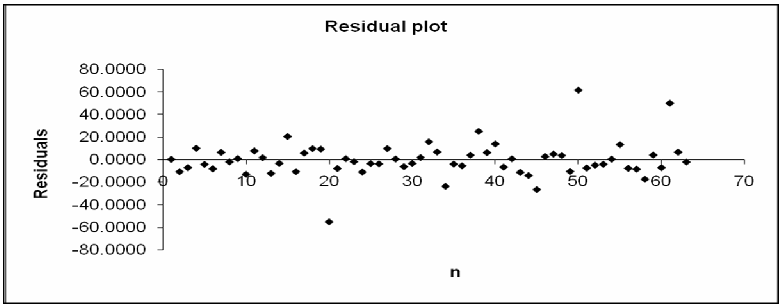

Figure 2: The residuals for model M73.15

Based on the observations of model M73.15, the p-values of main variables and removed variables in each step are summarized in Table 12 and Figure 3, where the p-value decreases to a value less than 0.05 to the corresponding removed variables. At the same time the p-value for the main independent variables converge to less than 0.05. This indicates that the corresponding variables contribute distinctly to the selling price. The study shows that model M73.15 is the best model to describe the house selling price in Ohio. Now the house selling price model is ready for forecasting to make a logical decision to determine the house selling price. The randomness test shows that model M73.15 has random and independent observation residuals. Model M73.15 also shows that there exists interaction effect. The floor area and number of rooms interact together.

TABLE 12: The summary of p-value for model M73.15

Step	X_1	X_2	X_3	X_4	X_5	Removing	p-value
0	0.310	0.225	0.834	0.075	0.471	X_{235}	0.892
1	0.779	0.540	0.693	0.547	0.842	X_{124}	0.960
2	0.687	0.514	0.691	0.515	0.842	X_{13}	0.897
3	0.500	0.409	0.805	0.429	-	X_5	0.923
4	0.456	0.406	0.815	0.312	-	X_{25}	0.906
5	0.441	0.233	0.712	0.308	-	X_{24}	0.772
6	0.469	0.012	0.577	0.109	-	X134	0.805
7	0.475	0.010	0.612	0.103	-	X_{34}	0.688
8	0.426	0.007	-	0.088	-	X_3	0.755
9	0.353	0.006	-	0.082	-	X_{245}	0.596
10	0.430	0.004	-	0.093	-	X_{14}	0.668
11	0.497	0.003	-	0.004	-	X_{125}	0.572
12	-	0.002	-	0.003	-	X_1	0.697
13	-	0.001	-	0.001	-	X_{35}	0.347
14	-	0.001	-	0.001	-	X_{23}	0.374
15	-	0.001	-	0.001	-	-	<0.05

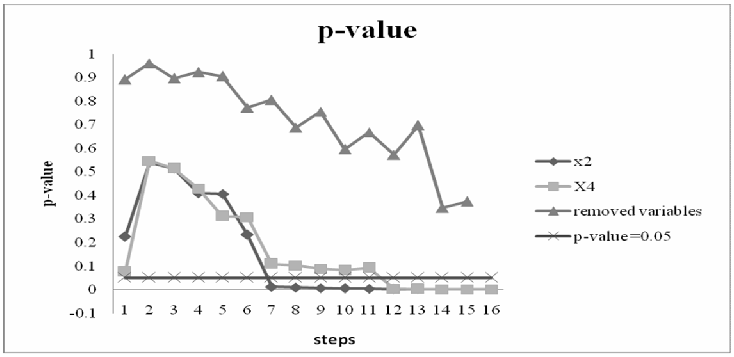

Figure 3: The convergence of p-value for model M73.15

This model shows that the variables like, the floor area, number of bedrooms and number of bathrooms does not have a direct effect on the selling price of a house. These variables cannot act as a single-effect variable. The number of rooms and age of the house can have a direct effect in determining the house selling price. But when the number of rooms or age increase, the house selling price decreases.

This model also shows that, to determine a house selling price, the variables should interact with each other. Based on the best model, it can be concluded that to determine a house selling price, one should consider house characteristics like floor area, number of rooms, number of bedrooms, age of the house and number of bathrooms. Besides these variables, a person's willingness/readiness to buy a house, income status, and the facilities around the housing areas can also affect the house selling price.

As can be seen from the above finding and elaborate discussion, best multiple regression could successfully be obtained where several single independent variables and higher order interactions had been included in the initial models. Thus, in a similar fashion, local data with *unbiased details* on house sale or related data can also be applied. Hence, different model is identified with different set of independent variables and interaction variables.

REFERENCES

Akaike, H. 1970. Statistical Predictor Identification. *Annals Instit. Stat. Math*. **22:** 203-217.

Akaike, H. 1974. A New Look at Statistical Model Identification. *IEEE Trans. Auto Control* **19**: 716-723.

Evans, J., R. and Olson, D., L. 2003. *Statistics, Data Analysis and Decision Modeling*. 2nd edition. New Jersey: Prentice Hall.

Golub, G.H., Heath, M. and Wahba, G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technomtrics* **21**: 215-223.

Hannan, E., J. and Quinn, B. 1979. The Determination of the Order of an Autoregression. *J. Royal Stat. Society*, **41**(B): 190-195.

Kenkel, J. L.1996. *Introductory Statistics for Management and Economics*. 4th edition. New York: Duxbury Press, Wadsworth Publishing Company.

Lind, D. A., Marchal, W. G. and Mason, R. D. 2005. *Statistical Techniques in Business & Economics.* 11th edition. New York: McGraw Inc.

Nikolopoulos, K., Goodwin, P., Patelis, A. and Assimakopoulos, V. 2007. Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. *European Journal of Operational Research*. **180**: 354-368.

Ramanathan, R. 2002. *Introductory Econometrics with Application*. 5th edition. Ohio: South Western, Thomson learning Ohio.

Rice, J. 1984. Bandwidth Choice for Nonparametric Kernel Regression. *Annals of Stat*. **12**: 1215-1230.

Runyon, R. P., Coleman, K. A. and Pittenger, D. J. 2000. *Fundamental of Behavioral Statistics*. 9th edition. New York: McGraw Inc.

Schwarz, G. 1978. Estimating the Dimension of a Model. *Annals of Stat.* **6**: 461-464.

Shibata, R. 1981. An Optimal Selection of Regression Variables. *Biometrika* **68**: 45-54.