SIGNIFICANT INDICATORS OF LOW COST HOUSING DEMAND: COMPARISON BETWEEN RESULTS OBTAINED FROM PRINCIPAL COMPONENT ANALYSIS, BACK ELIMINATION AND REGRESSION ANALYSIS METHODS

Noor Yasmin Binti Zainun[1], Aftab Hameed Memon[2], Mohd Firhan Anuar[1]
[1]Faculty of Civil and Environmental Engineering,
Universiti Tun Hussein Onn Malaysia

[2]Civil Engineering Department,
Quaid-e-Awam University of Engineering, Science and Technology, Pakistan

Corresponding E-mail : aftabm78@hotmail.com

Abstract

This paper has reported comparison between Principal Component Analysis (PCA), Back Elimination Method (BEM) and Regression Method. These techniques were applied by using statistical software package SPSS 13.0. For the purpose of comparison, all the methods were tested on nine prime indicators of low cost housing demand which include population growth, birth rate, mortality rate, inflation rate, unemployment rate, GDP (gross domestic product), housing stock, household income and poverty rate. Data for the indicators was obtained from ministry of housing for low cost housing demand in Gombak District. From analysis it was found that PCA method had identified three significant indicators for low cost housing demand that is GDP/Capita in Selangor, housing stock and mortality baby rate. BEM had identified four significant indicators that is inflation rate, GDP/Capita in Selangor, Poverty Rate and Housing Stock. While, regression method identified only one significant indicator that is poverty rate. From these findings it can be concluded that BEM is the best method in determining significant indicators as compared to PCA and regression method. These finding will help the researcher in adopting suitable method for determining significant indicators in any field.

Keywords: *principal component analysis, back elimination method, regression method, low cost housing demand, Indicator of low cost housing, Malaysia*

1.0 Introduction

Housing is a basic social need and provision of adequate, quality and affordable housing for all income groups is a national imperative and it is one of the main aspects of urban problems which are directly linked to the economy. Thus, it is compulsory that housing availability be ensured for all classes of people including people with middle and low income. The Malaysian government's policy on low-cost houses scheme mainly to address one of the essential needs of the lower income bracket of the population, that is to own houses (National Housing Policy, 2012). From the early period of independence until the presentation of Tenth Malaysian Plan, there are several of policies affecting the low-cost houses which have been introduced and implemented. However, still a significant amount of people with low income could not get low cost house as the availability of houses is higher than the demand. This demand of the low cost housing is increasing day by day. In order to ensure that the availability of the houses is adequate, it is very important to forecast the demand of the houses. The demand of the houses depends on various indicators which are very imperative to determine. Hence, it is very crucial to determine significant indicator or variable of low cost housing demand.

In order to determine significant indicators different researchers have used different approaches such as Importance index, average index as well as multivariate methods which include Principal Component Analysis (PCA), Back Elimination Method (BEM) and Regression method. Among the variety of methods for analysis, it is imperative to select the appropriate method for obtaining accurate and precise results. Hence, this study has focused on comparing various statistical methods in determining significant indicators. However, this study is limited to compare three multivariate statistical methods only which are PCA, BEM and Regression.

Principal component analysis is a mathematical procedure that transforms a set of correlated indicators into a smaller set (reduction of dimensionality) of uncorrelated indicators called principal components. PCA requires no assumptions about the population from which the data are sampled. It is a way identifying pattern in the data. The main advantages of using PCA is that once we have found these patterns in the data and we compress the data by using the number of dimensions without much loss information (Zainun, 2004; Yeung and Walter 2000).

Back elimination method (BEM) is a procedure that starts with the full model and removes one indicator at a time without adding indicators. One includes all possible regressed indicators, and attempts to eliminate them from the model one at a time until no removal occurs. Since, backward elimination method only seeks to remove indicators from the model; the indicator with the smallest incremental contribution to the regression is tested at each step to determine whether it can be eliminated from the model. The applications of BEM are mostly used for geometry theorem proving and decidability result (Yahya, 2002).

Regression is a collection of statistical techniques that serve as a basis for drawing inferences about relationships among interrelated indicators (Golberg and Cho, 2004). Main purposes of regression analysis are for data description, Interpretation and Inference. Researchers have used regression method for several purposes such as predicting nitrogen oxide concentrations with auto regressive modeled disturbances (Inoue *et al*, 1986), health care (Kooreman, 1994; Juras and Brooks 1993), and the insurance industry (Cummins *et al.* 1999; Carr 1997).

Though, all three methods have been used by previous researchers to determine their significant indicators but there is lack of studies in giving clear picture that among these which method could gives the most accurate result. It is very important to know in order to select the appropriate method for conducting as research as precise significant indicators will give precise results of the study. Hence, this study is carried out to compare all three methods using t-series data for indicators of low cost housing in Gombak city of Malaysia.

2.0 Review of Literature

2.1.1 Indicators of Low Cost Housing

The demand of low cost housing depends on several factors. According to Abdul Karim (1995) population growth can give pressure to demand on social services such as school, housing and hospital development. Studies in Thailand, Singapore and United Kingdom show that population size has an influence towards the increment of housing demand. For example, population size in Thailand gives rise to significant influence towards housing demand but in Singapore, population growth does not give high influence towards residential construction demand (Goh, 1998). According to Goh (1998), there is a close relationship among three factors, which are population, the construction activities and the housing stocks and vice versa, the slow growth of population slows down the construction activities and decreases the housing stocks. Besides that, Goh also stated there are seven indicators to forecast residential construction demand in Singapore these include: (1) building tender price index; (2) bank lending; (3) population; (4) housing stock; (5) National savings; (6) gross fixed capital formation; and (7) unemployment level.

Yahya and Abd Majid (2002) used more indicators to forecast demand on low cost housing compare others. The indicators considered are: (1) population growth; (2) birth rate; (3) average mortality baby: (4) unemployment rate; (5) inflation rate; (6) Gross Domestic Product (GDP); (7) poverty rate; (8) income rate; and (9) housing stock. According to Sirat et. al (1999) the factors that can influence demand on low-cost housing can be divided into seven indicators: (1) demographic factors; (2) income factor; (3) ability factor; (4) profit to own house; (5) loan facilities factor; (6) speculation factors; and (7) government policy towards housing ownership, also plays an important role. In Malaysia, there are nine prime indicators of low cost housing which include (1) population growth; (2) birth rate; (3) child mortality rate; (4) inflation rate; (5) income rate; (6) housing stock; (7) GDP rate; (8) unemployment rate; and (9) poverty rate (Zainun et. al. 2014)

2.1.2 Principal Component Analysis

Principal Component Analysis (PCA) is a method for producing a small number of constructed indicators derived from the larger number of indicators. These derived indicators are uncorrelated and these reduced number of indicators help to understand the underlying structure of the data. Principal Component Analysis does not have an underlying statistical model. It is a mathematical technique used in other statistical analyses driven by different models such as factor analysis. PCA is very useful in cases where the size of the data set becomes unwieldy as working with fewer dimensions makes it easier to visualize the data and identify interesting patterns. Various researchers have applied PCA for different types of researcher such as Cavalli-Sforza (2000) used PCA for genetic mapping. In PCA dimension reduction is the creation of indices from survey or experimental data. Survey researchers often use many different questions to get at one particular property or characteristics of the survey respondent. For example, Ofir and Simonson (2001) used a battery of 18 questions developed by Cacioppo, Petty, and Kao (1984) to get at each subject's "need for cognition" (example, the extent to which the subject enjoys and engages in thinking and problem solving).

2.1.3 Back Elimination Method

Back elimination begins with the full model and sequentially eliminates from the model the least important indicator. The importance of an indicator is judge by the size of the *t* (or equivalent *F*) statistic for dropping the indicator from the model, i.e., the statistic for testing whether the corresponding regression coefficient is 0. After the indicator with the smallest absolute t statistics is dropped, the model is refitted and the t statistic is recalculated. Again the indicator with the smallest absolute t statistics is dropped. The process ends when all of the absolute values of the t statistics are greater than some predetermined level. The predetermined level can be a fixed number for all steps or it can change depending on a steps. When allowing it to change depending on the steps, set up the process so that it stops when all of the P values are below a fixed level. Back elimination may sometime break down from beginning if the full model cannot be fitted (Fahrmeir & Frost, 1992). For example: in case of nonexistence of estimates, which is usually negligible for classical linear models, becomes much more serious for some non-normal models involving a large number of parameters, in particular for models with multi categorical indicators, back elimination is broken down while forward elimination is still applicable.

2.1.4 Regression Method

Regression analysis is the method to finding the best straight line relationship to explain how the variation in an outcome indicator, Y. It depends on the variation in a predictor indicator, X. Once the relationship has been estimated the following equation can be used as $Y = b0 + b1X$ to predict the value of the outcome indicator for different values of the explanatory indicator. For example, if age is a predictor for the outcome of treatment, then the regression equation would enable us to predict the outcome of treatment for a person of a particular age. Of course this is only useful if most of the

variation in the outcome indicator is explained by the variation in the explanatory indicator. In many situations the outcome will depend on more than one explanatory indicator. This leads to use of multiple regression, in which the dependent indicator is predicted by a linear combination of the possible explanatory indicators is to leads the multiple regression. For example, it is known that the male peak expiratory flow rate (PEFR) depends on both age and height, so that the regression equation will be as PEFR = b0 + b1 x age x b2 x height. In this relation, the values b0, b1, b2 are called the regression coefficients and are estimated from the study data by a mathematical process called least squares (Altman 1991). Step wise regression process extracts several possible explanatory indicators in the data set while in analysis only one indicator can be considered at one time. The one that explains most variation in the dependent indicator will be added to the model at each step. The process will stop when the addition of an extra indicator will make no significant improvement in the amount of variation explained. The amount of variation in the dependent indicator that is accounted for by variation in the predictor indicators is measured by the value of the coefficient of determination, often called R2 adjusted. The closer this is to 1 the better, because if R2 adjusted is 1 then the regression model is accounting for all the variation in the outcome variable (Altman 1991, Campbell & Machin 1993). Main purpose of regression analysis is to investigate or refute a relationship among indicators and to interpret that it can give a summary or an interpretation through the fitted model to obtain an interpolation or calibration curve. Regression also uses to develop or improve the theoretical model or method which should be chosen to extend and generalize it to other sets of data.

3.0 Methodology

Qualitative mode of research method was adopted in this study which focused on data collection for nine independent indicators of low cost housing demand were identified from previous studies. These indicator include population growth, birth rate, average mortality baby rate, unemployment rate, inflation rate, gross domestic product (GDP), poverty rate, income and housing stock (Yahya and Abd. Majid 2002, Chander 1977, Yang and Packer 1997, Yahya 2002, Zainun et. al. 2014). For statistical tests, data against each indicator/variable was collected from ministry of housing. This data was limited for Gombak district showing the monthly record of low cost housing demand for a total of 5 years duration as shown in table 1 where V1 represents population growth, V1 represents birth rate, V1 represents average mortality baby rate, V1 represents unemployment rate, V1 represents inflation rate, V1 represents gross domestic product (GDP), V1 represents V1 represents poverty rate, V1 represents income and V1 represents housing stock. Data consist of 58 time series data where it will be used to analyze using SPSS 13.0 software.

Table 1: Statistical Data for Low Cost Housing Demand in District of Gombak

V1	V2	V3	V4	V5	V6	V7	V8	V9	DEMAND	MONTH
463.4	29.7	6.4	2	3.5	9263	2.14	3.12	8705	250	Feb-96
465.2	29.65	6.35	2.5	3.55	9681.5	2.19	3.21	8852.5	311	Mar-96
466.1	29.63	6.33	2.75	3.58	9890.75	2.22	3.26	8926.25	314	Apr-96
467	29.6	6.3	3	3.6	10100	2.25	3.31	9000	226	May-96
468.5	29.55	6.25	3.6	3.38	10250	2.23	3.32	9000	217	Jun-96
469.25	29.53	6.23	3.9	3.26	10325	2.21	3.32	9000	327	Jul-96
470	29.5	6.2	4.2	3.15	10400	2.2	3.32	9000	240	Aug-96
471.5	29.5	6.18	4.7	2.95	10350	2.17	3.31	9250	140	Sep-96
472.25	29.5	6.16	4.95	2.85	10325	2.16	3.31	9375	151	Oct-96
473	29.5	6.15	5.2	2.75	10300	2.15	3.3	9500	172	Nov-96

475.25	29.55	6.13	5.6	2.73	10306.8	2.12	3.28	10250	226	Dec-96
476.38	29.58	6.11	5.8	2.71	10310.2	2.11	3.27	10625	240	Jan-97
477.5	29.6	6.1	6	2.7	10313.7	2.09	3.27	11000	100	Feb-97
479	29.65	6.05	6.2	2.95	10256.8	2.09	3.28	13000	202	Mar-97
479.75	29.68	6.03	6.3	3.08	10228.4	2.08	3.29	14000	202	Apr-97
480.5	29.7	6	6.4	3.2	10200	2.08	3.3	15000	167	May-97
484.25	29.75	6.05	6.5	3.68	10100	2.07	3.31	17750	183	Jun-97
486.13	29.78	6.08	6.55	3.91	10050	2.07	3.32	19125	296	Jul-97
488	29.8	6.1	6.6	4.15	10000	2.06	3.32	20500	104	Aug-97
491	29.9	6.18	6.6	4.58	9900	2.04	3.35	23000	200	Sep-97
492.5	29.95	6.22	6.6	4.79	9850	2.06	3.37	24250	311	Oct-97
494	30	6.2	6.6	5	9800	2.05	3.38	25500	299	Nov-97
V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**DEMAND**	**MONTH**
496.55	30.05	6.23	6.5	5.15	9824.55	2.05	3.4	26836	151	Dec-97
497.83	30.08	6.22	6.45	5.23	9836.83	2.04	3.41	27504	128	Jan-98
499.1	30.1	6.2	6.4	5.3	9849.1	2.04	3.42	28172	70	Feb-98
501.05	30.15	6.2	6.15	5.15	10124.5	2.04	3.46	27586	128	Mar-98
502.03	30.18	6.21	6.03	5.08	10262.3	2.03	3.47	27293	154	Apr-98
503	30.2	6.21	5.9	5	10400	2.03	3.49	27000	150	May-98
504.25	30.25	6.21	5.5	4.55	10800	2.03	3.54	25000	130	Jun-98
504.88	30.28	6.21	5.3	4.33	11000	2.02	3.56	24000	128	Jul-98
505.5	30.3	6.21	5.1	4.1	11200	2.02	3.58	23000	191	Aug-98
507.25	30.15	6.2	4.65	3.75	11650	2.02	3.62	21000	137	Sep-98
508.13	30.08	6.2	4.43	3.53	11875	2.01	3.63	20000	126	Oct-98
509	30	6.2	4.2	3.4	12100	2.01	3.65	19000	116	Nov-98
511.65	29.8	6.2	3.95	3.1	12434	2.01	3.68	17631	156	Dec-98
512.98	29.7	6.2	3.83	2.95	12601	2	3.69	16946.5	189	Jan-99
514.3	29.6	6.2	3.7	2.8	12768	2	3.7	16262	70	Feb-99
518.4	29.59	6.19	3.55	2.6	12834	2	3.71	15881	132	Mar-99
520.45	29.58	6.19	3.48	2.5	12867	1.99	3.71	15690.5	172	Apr-99
522.5	29.57	3.4	6.19	2.4	12900	1.99	3.71	15500	215	May-99
527.25	29.58	6.18	3.38	2.25	13000	1.99	3.72	15625	200	Jun-99
529.63	29.58	6.17	3.36	2.17	13050	1.98	3.72	15687.5	255	Jul-99
532	29.58	6.17	3.35	2.1	13100	1.98	3.72	15750	330	Aug-99
537.5	29.59	6.17	3.35	2	13150	1.97	3.72	16125	321	Sep-99
540.25	29.59	6.17	3.35	1.95	13175	1.97	3.72	16312.5	172	Oct-99
543	29.59	6.17	3.35	1.9	13200	1.96	3.72	16500	161	Nov-99
548.2	29.6	6.19	3.33	1.75	13356.9	1.96	3.73	16731	172	Dec-99
550.8	29.6	6.19	3.31	1.68	13435.4	1.95	3.73	16846.5	70	Jan-00
553.4	29.6	6.2	3.3	1.6	13513.8	1.95	3.73	16962	50	Feb-00
549.7	29.61	6.2	3.4	1.7	13306.9	1.96	3.71	16906	189	Mar-00

547.85	29.62	6.2	3.45	1.75	13203.5	1.96	3.69	16878	200	Apr-00
546	29.62	6.21	3.5	1.8	13100	1.97	3.68	16850	98	May-00
539.5	29.64	6.21	3.6	2	13000	1.98	3.64	16725	178	Jun-00
536.25	29.64	6.21	3.65	2.1	12950	1.98	3.62	16662.5	296	Jul-00
533	29.65	6.21	3.7	2.2	12900	1.99	3.6	16600	344	Aug-00
524	29.67	6.21	3.85	2.45	12350	2.00	3.56	16500	281	Sep-00
519.5	29.68	6.21	3.93	2.58	12075	2.00	3.53	16450	147	Oct-00
515	29.69	6.22	4	2.7	11800	2.02	3.51	16400	179	Nov-00

4.0 Data analysis and Findings

4.1 Principal Component Analysis (PCA)

From the analysis, the determinant of the correlation matrix for the data |R| was found as $1.30x10^{-10}$ which is very close to zero. This indicates that linear dependencies exist among the indicators. Therefore, PCA can be performed. The data are multivariate normal because of all indicators are uncorrelated, then testing the hypothesis that the population correlation matrix is equal to the identify matrix. In this study there are nine indicators and 58 data therefore, p = 9 and N = 58, Thus;

$$-a.\ln(v) = -(N - 1 - (2p + 5)/6\ln(|R|)$$

$$= -(58 - 1 - (2x9 + 5)/6\ln(1.3\ x\ 10^{-10})$$

$$= 1210.259$$

Therefore, value for the test statistic for these data is 1210.352 and the critical point of the chi-square distribution with p (p-1) = 36. For degrees of freedom, α= 0.001, the critical point is 67.92 (Lee 1997). Clearly the test hypothesis will be rejected at the 0.001 significant level because 1210.259 > 67.92. Variance extracted for the tested data with PCA is as shown in Table 2.

Table 2: Total Variance Extracted

Component	Initial Eigen values		
	Total	**% of variance**	**Cumulative %**
1	4.451	49.458	49.458
2	2.857	31.744	81.202
3	1.379	15.317	96.520
4	0.153	1.701	98.220
5	0.095	1.055	99.276
6	0.039	0.430	99.706
7	0.023	0.260	99.966
8	0.003	0.031	99.998

9	0.000	0.002	100.000

Table 2 shows that the first principal component give the highest number of Eigen value with 4.451 consist of 49.458% of the total variation while the second principal component give 2.857 Eigen value with 31.744% of the total variation. Principal component (PC) three give 1.379 Eigen values that consist of 15.317% of the total variation. PC four has Eigen value as 0.153 which constructs 1.701% of the total variation. PC five has Eigen value of 0.095 that consists 1.055% of the total variation. PC six contains 0.43% of the total variation with 0.039 Eigen value. PC seven contributes to 0.26% of the total variation with 0.023 Eigen value. PC eight has 0.031% of the total variation with 0.003 Eigen value while PC nine has zero Eigen value that consist of 0.002% of the total variation. Scree plot for the results obtained from PCA is shown in figure 1.

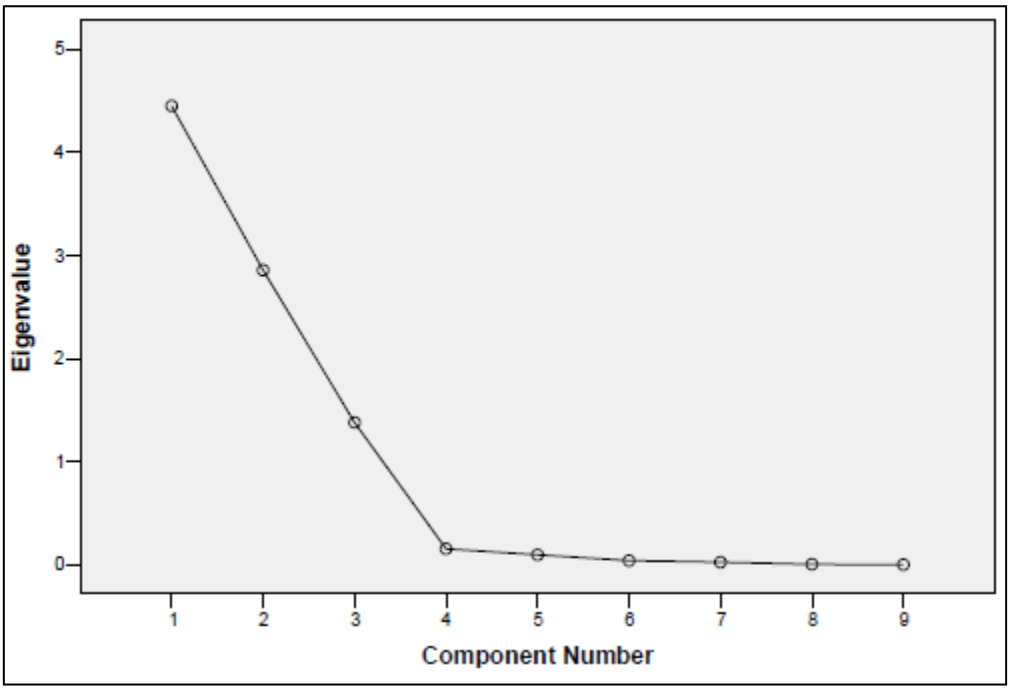


Figure 1: Scree Plot of

From the scree plot in Figure 1, Eigen values for principal component (PC) four to nine are close to zero. Since Eigen values for PC one to three are greater than one, total variation for the three PC is 96.5% and others Eigen values are close enough to zero that they can be ignored. Therefore three PC are used for the analysis. Hence, for further analysis only 3 components are considered and the component score co-efficient matrix for these 3 components is shown in Table 3.

Table 3: Component Score Coefficient Matrix

Variables	**Component**		
	1	**2**	**3**
Population Growth	0.204	0.131	0.001
Birth Rate	-0.086	0.291	0.208
Mortality Baby Rate	0.010	-0.054	0.708
Unemployment Rate	-0.153	0.182	-0.359

Inflation Rate	-0.181	0.180	0.182
GDP/Capita In Selangor	0.223	0.019	-0.020
Household Income Rate	-0.168	-0.214	0.100
Poverty Rate	0.202	0.125	0.055
Housing Stock	-0.019	0.344	0.072

According to Johnson (1998), the number of component is to be equal to the number of Eigen value of **R**, which is 1. Therefore, the significant indicators for each component are with the value of component score coefficient matrix nearest to 1. The other indicators are still considered but they give less effect compared to the significant indicators. From 3 it can be perceived that for PC1 the indicator that has value of component scores that nearest to 1 is GDP/Capita in Selangor (0.223). So it will be significant indicator for component 1. The indicator that has value nearest to 1 for component 2 is housing stock (0.344). So it will be significant indicator for component 2. The indicator that has value nearest to 1 for component 3 is mortality baby rate (0.708). So it will be significant indicator for component 3. These results are summarized in table 4.

Table 4: Summary of PCA findings

Indicators	**Component**		
	1	**2**	**3**
Population Growth	-	-	-
Birth Rate	-	-	-
Mortality Baby Rate	-	-	0.708
Unemployment Rate	-	-	-
Inflation Rate	-	-	-
GDP/Capita In Selangor	0.223	-	-
Household Income Rate	-	-	-
Poverty Rate	-	-	-
Housing Stock	-	0.344	-

From Table 4 it is seen that significant indicators for low cost housing demand in Gombak district using PCA method are;

a. GDP/Capita In Selangor;
b. Housing Stock; and
c. Mortality baby rate.

4.2 Back Elimination Method (BEM)

In back elimination method, initially all indicators are considered for analysis and then BEM will eliminate the indicators that gives less effect compared than others. Independent indicators used in this method are as population growth (PGROWTH), birth rate (BRATE), mortality baby rate

(MBRATE), unemployment rate (UNEMRATE), inflation rate (INFLARATE), GDP/Capita in Selangor (GDPC), household income rate (HHOLDRATE), poverty rate (POVRATE) and housing stock (HSSTOCK). Results obtained from back elimination method for the indicators of low cost housing are shown in Table 5.

Table 5: Coefficients for Back Elimination Method

Model	Indicators	Unstandardized Coefficients		Standardized Coefficients	T	Sig
		B	Std. Error	Beta		
1	(CONSTANT)	-6129.7915	8362.132		-0.303	0.763
	PGROWTH	10.648	11.713	3.855	0.909	0.368
	BRATE	59.222	172.710	0.186	0.343	0.733
	MBRATE	-302.186	447.550	-0.274	-0.675	0.503
	UNEMRATE	33.681	68.734	0.604	0.490	0.626
	INFLARATE	419.554	349.305	6.063	1.201	0.236
	GDPC	0.187	0.102	3.498	1.830	0.074
	HHOLDRATE	-519.112	1016.800	-0.552	-0.511	0.612
	POVRATE	-1130.570	507.779	-2.785	-2.227	0.031
	HSSTOCK	-0.067	0.076	-5.053	-0.878	0.384
2	(CONSTANT)	70.912	3457.943		0.021	0.984
	PGROWTH	7.436	6.967	2.692	1.067	0.291
	MBRATE	-257.251	424.064	-0.234	-0.607	0.547
	UNEMRATE	16.619	46.990	0.298	0.354	0.725
	INFLARATE	326.381	217.520	4.716	1.500	0.140
	GDPC	0.183	0.101	3.426	1.819	0.075
	HHOLDRATE	-299.277	782.077	-0.318	-0.383	0.704
	POVRATE	-1099.366	495.039	-2.708	-2.221	0.031
	HSSTOCK	-0.046	0.045	-3.462	-1.026	0.310
3	(CONSTANT)	1156.682	1577.462		0.733	0.467
	PGROWTH	5.774	5.098	2.090	1.132	0.263
	MBRATE	-329.489	368.369	-0.299	-0.894	0.375

	INFLARATE	275.359	161.366	3.979	1.706	0.094
	GDPC	0.185	0.100	3.461	1.857	0.069
	HHOLDRATE	-135.509	624.726	-0.144	-0.217	0.829
	POVRATE	-1133.924	481.036	-2.793	-2.357	0.022
	HSSTOCK	-0.034	0.029	-2.560	-1.169	0.248
4	(CONSTANT)	952.606	1254.293		0.759	0.451
	PGROWTH	5.139	4.136	1.860	1.728	0.220
	MBRATE	-294.151	327.283	-0.267	-0.899	0.373
	INFLARATE	253.173	123.637	3.659	2.048	0.046
	GDPC	0.191	0.094	3.582	2.032	0.047
	POVRATE	-1149.611	471.104	-2.831	-2.440	0.018
	HSSTOCK	-0.029	0.019	-2.203	-1.539	0.130
5	(CONSTANT)	596.583	1188.002		0.502	0.618
	PGROWTH	1.998	2.208	0.723	0.905	0.370
	INFLARATE	154.706	57.194	2.236	2.705	0.009
	GDPC	0.179	0.093	3.354	1.927	0.059
	POVRATE	-1060.225	459.635	-2.611	-2.307	0.025
	HSSTOCK	-0.014	0.008	-1.049	-1.759	0.102
6	(CONSTANT)	1544.043	560.744		2.754	0.008
	INFLARATE	135.607	53.065	1.960	2.555	0.014
	GDPC	0.220	0.081	4.116	2.707	0.009
	POVRATE	-1190.314	435.822	-2.932	-2.731	0.009
	HSSTOCK	-0.007	0.004	-0.542	-1.873	0.067

From the Table 5, it can be seen that there are 6 model regression used in this method. It means that it has 6 step of analysis. Step 1 is for the process to develop the model and is represented as:

$$\text{Ln(demand)} = \beta_0 + \beta_1 {}^{*}\text{PGROWTH} + \beta_2 {}^{*}\text{BRATE} + \beta_3 {}^{*}\text{MBRATE} + \beta_4 {}^{*}\text{UNEMRATE} + \beta_5 {}^{*}\text{INFLARATE} + \beta_6 {}^{*}\text{GDPC} + \beta_7 {}^{*}\text{HHOLDRATE} + \beta_8 {}^{*}\text{POVRATE} + \beta_9 {}^{*}\text{HSSTOCK}$$

The t statistics is used to test Null hypothesis which is defined as:

Null hypothesis: One of the independent indicators (smallest t) that not affect the demand.

H_0: $\beta_0 = 0$; H_0: $\beta_1 = 0$; H_0: $\beta_2 = 0$; H_0: $\beta_3 = 0$; H_0: $\beta_4 = 0$; H_0: $\beta_5 = 0$; H_0: $\beta_6 = 0$; H_0: $\beta_7 = 0$; H_0: $\beta_8 = 0$; $B_9 = 0$

Alternative hypothesis: All the independent indicators that not affect the demand

H_0: $\beta_0 \neq 0$; H_0: $\beta_1 \neq 0$; H_0: $\beta_2 \neq 0$; H_0: $\beta_3 \neq 0$; H_0: $\beta_4 \neq 0$; H_0: $\beta_5 \neq 0$; H_0: $\beta_6 \neq 0$; H_0: $\beta_7 \neq 0$; H_0: $\beta_8 \neq 0$; $B_9 \neq 0$

Analysis from the computer shows that the t value for the model 1 that shows from the equation below;

Ln(demand) = -6129.7915 + 10.648 *PGROWTH + 59.222 *BRATE + (-302.186) *MBRATE
(3.855) (0.186) (-0.274)

+ 33.681 *UNEMRATE + 419.554 *INFLARATE + 0.187 *GDPC
(0.604) (6.063) (3.498)

+ (-519.112) *HHOLDRATE + (-1130.570) *POVRATE
(-0.552) (-2.785)

+ (-0.067) *HSSTOCK
(-5.053)

From nine of independent indicators, BRATE or birth rate have the smallest t absolute where as for the value for $t_{a/2}^{n-k}$ with $\alpha = 0.1$, $n = 58$ and $k = 10$, the value is as $t_{0.05}^{48}$=1.677 (Lee 1997). Therefore, BRATE has $|t| < 1.677$, so null hypothesis is true and H_0: $\beta_3 \neq 0$ can be ignored. The significant value for BRATE is 0.733 that is more than $\alpha = 0.1$. Because of that BRATE will be eliminated. The other indicators will maintain although the t is less than $t_{a/2}^{n-k}$ as shown in Figure 4.2. Then the back elimination method will do the regression analysis for model 2 with the staying independent indicators that is PGROWTH, MBRATE, UNEMRATE, INFLARATE, GDPC, HHOLDRATE, POVRATE and HSSTOCK. For Model 2, with the same step with, $\alpha = 0.1$, $n = 58$ and $k = 10$, so the $t_{0.05}^{49} = 1.674$. The independent indicator with the smallest t test comes from UNEMRATE with $0.354 < t_{0.05}^{49}$. So UNEMRATE will be eliminated. Similarly model 3 with independent indicators PGROWTH, MBRATE, INFLARATE, GDPC, HHOLDRATE, POVRATE and HSSTOCK will be analyzed and the same process will be repeated. The process will be stopped until there are no t value of indicator are $\leq t_{0.05}^{48} = 1.677$. Overall results of Back elimination run for 6 models are presented in Table 6.

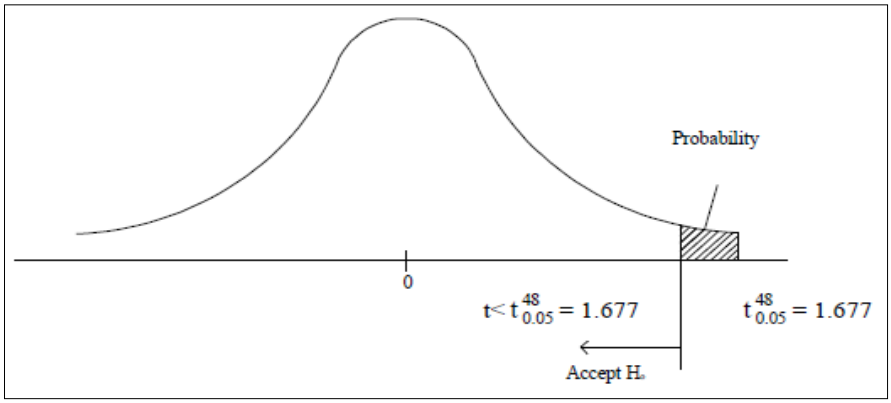


Figure 2: Point of t elimination $t_{a/2}^{n-k}$

Table 6: Result of ANOVA Analysis for Back Elimination Method

Model		Sum Of Squares	df	Mean Square	F	Sig.
1	Regression	80964.048	9	8996.005	1.838	0.085(a)
	Residual	234962.573	48	4895.054		
	Total	315926.621	57			
2	Regression	80388.490	8	10048.561	2.090	0.055(b)
	Residual	235538.130	49	4806.901		
	Total	315926.621	57			
3	Regression	79787.213	7	11398.173	2.413	0.033(c)
	Residual	236139.408	50	4722.788		
	Total	315926.621	57			
4	Regression	79565.008	6	13260.835	2.861	0.018(d)
	Residual	236361.613	51	4634.541		
	Total	315926.621	57			
5	Regression	75821.304	5	15164.261	3.284	0.012(e)
	Residual	240105.317	52	4617.410		
	Total	315926.621	57			
6	Regression	72401.069	4	18010.267	3.914	0.007(f)
	Residual	243885.552	53	4601.614		
	Total	315926.621	57			

Statistic F is used to test the Model 1 where the hypothesis is as:

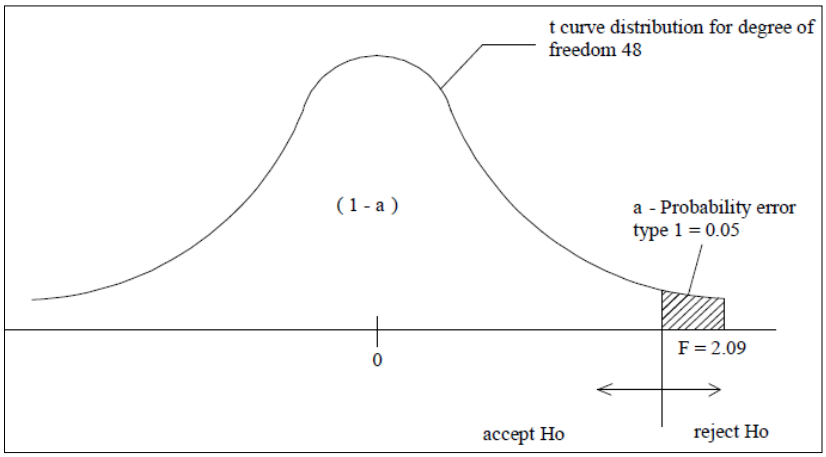
H_0: $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, B_9 = 0$

H_1: at least of the $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, B_9 \neq 0$

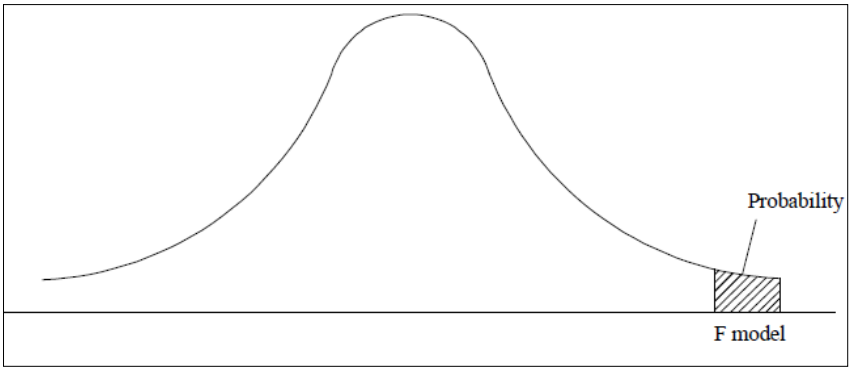
Analysis of F model as shown in Table 5 is giving 1.838 and the probability is 0.085. The value of $F_{(0.05)}^{(9.48)}$ for $\alpha = 0.5$, n = 58 and k = 10, $F_{(0.05)}^{(9.48)} = 2.09$ (Bowerman and O'Connell1992). Therefore, model 1 is $F < F_{(a)}^{(k-1,n-k)}$, so the null hypothesis is true and H_1 is rejected. The significant value is more than 0.05, therefore, all the independent value in Model 1 do not have effect to the demand. Then back elimination method will perform F analysis for Model 2 which include PGROWTH, MBRATE, UNEMRATE, INFLARATE, GDPC, HHOLDRATE, POVRATE and HSSTOCK. This process will be continued until F analysis for model 6 is performed. F value of Model 6 is 3.914> $F_{(\alpha/2)}^{(k-1,n-k)}$, so the alternative hypothesis is true and H_0 can be ignored. The significant value is 0.007 is less than 0.05. From these results of F analysis, it is found that there are four indicators which give effect to the demand of low cost housing that are INFLARATE, GDPC, POVRATE and HSSTOCK. Overall summary of back elimination model is presented in Table 7 and point of elimination is shown in Figure 3.

Table 7: Model Summary for Back Elimination Method

Model	R	R^2	$\breve{R}^2$	Std. Error of The Estimate	R Square Change
1	0.506(a)	0.526	0.117	69.96466	0.526
2	0.504(b)	0.254	0.133	69.33182	-0.002
3	0.503(c)	0.253	0.148	68.72254	-0.002
4	0.502(d)	0.252	0.164	68.07747	-0.001
5	0.490(e)	0.240	0.167	67.95153	-0.012
6	0.478(f)	0.228	0.170	67.83520	-0.012



(a) Pint of F Elimination



(b) Value of model probability

Figure 3: Point of $F_{(\alpha/2)}^{(k-1,n-k)}$elimination and the value of significant probability for F modelFrom table 7, it can be noted that R^2 and s values have highest value if there is more independent indicator but $\breve{R}^2$ will be small. Model 1 with 9 independent indicators have the highest $\breve{R}^2$value (=0.526) but wit the small $\breve{R}^2$ (=0.117). The best characteristic of model is it has the highest R^2 and $\breve{R}^2$and also the smallest s value. The highest $\breve{R}^2$mean that the changing in indicator is effective to give an effect in demand. Also, from table 7, it can be concluded that the independent indicators for model 4, 5 and 6 are fulfilling these criteria where Model 4 has independent indicators MBRATE with

the value is $|t = 0.899| < t_{(\propto/2)}^{(n-k)} = 1.677$ and the significant value is 0.373 which is more than α=0.1. The relationship between the independent indicators can be referred from Model 5. The removable of the independent indicators MBRATE will cause the PGROWTH indicators that have $|t|$ >1.677 in Model 4; will be not significant in Model 5 with $|t| < 1.6777$ Therefore, Model 6 with independent indicators INFLARATE, GDPC, POVRATE and HSSTOCK having$|t|$ >1.677 will be considered as significant indicators. Thus, it can be concluded that the significant indicators for low cost housing demand in Gombak district using BEM are;

a. Inflation rate (INFLARATE)
b. GDP/Capita in Selangor (GDPC)
c. Poverty rate (POVRATE)
d. Housing stock (HSSTOCK)

4.3 Regression Method

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative indicators so that one indicator can be predicted from the other, or other. The method for regression that being used is a enter method. Enter method is a method that only doing an analysis in one model which is adopted in this study. All independent indicators being used in this analysis are population growth (PGROWTH), birth rate (BRATE), mortality baby rate (MBRATE), unemployment rate (UNEMRATE), inflation rate (INFLARATE), GDP/Capita in Selangor (GDPC), household income rate (HHOLDRATE), poverty rate (POVRATE) and housing stock (HSSTOCK). Co-efficient values for indicators obtained from regression method are presented in Table 8.

Table 8: Coefficients for Regression Method

Model	Indicators	Unstandardized Coefficients		Standardized Coefficients	t	Sig
		B	Std.Error	Beta		
1	(CONSTANT)	-6129.7915	8362.132		-0.303	0.763
	PGROWTH	10.648	11.713	3.855	0.909	0.368
	BRATE	59.222	172.710	0.186	0.343	0.733
	MBRATE	-302.186	447.550	-0.274	-0.675	0.503
	UNEMRATE	33.681	68.734	0.604	0.490	0.626
	INFLARATE	419.554	349.305	6.063	1.201	0.236
	GDPC	0.187	0.102	3.498	1.830	0.074
	HHOLDRATE	-519.112	1016.800	-0.552	-0.511	0.612
	POVRATE	-1130.570	507.779	-2.785	-2.227	0.031
	HSSTOCK	-0.067	0.076	-5.053	-0.878	0.384

As indicated in Table 8, there is 1 model regression used in this method. It means that it have 1 step of analysis. Step 1 is for the process to develop the model which is as:

Ln(demand) = $\beta_0 + \beta_1$ *PGROWTH + β_2 *BRATE + β_3 *MBRATE + β_4 *UNEMRATE

$+ \beta_5 {}^{*}\text{INFLARATE} + \beta_6 {}^{*}\text{GDPC} + \beta_7 {}^{*}\text{HHOLDRATE} + \beta_8 {}^{*}\text{POVRATE}$
$+ \beta_9 {}^{*}\text{HSSTOCK}$

The t statistics is use to testing for the;

Null hypothesis: One of the independent indicators (smallest t) that not affect the demand.

H_0: $\beta_0 = 0$; H_0: $\beta_1 = 0$; H_0: $\beta_2 = 0$; H_0: $\beta_3 = 0$; H_0: $\beta_4 = 0$; H_0: $\beta_5 = 0$; H_0: $\beta_6 = 0$; H_0: $\beta_7 = 0$; H_0: $\beta_8 = 0$; $B_9 = 0$

Alternative hypothesis: All the independent indicators that not affect the demand

H_0: $\beta_0 \neq 0$; H_0: $\beta_1 \neq 0$; H_0: $\beta_2 \neq 0$; H_0: $\beta_3 \neq 0$; H_0: $\beta_4 \neq 0$; H_0: $\beta_5 \neq 0$; H_0: $\beta_6 \neq 0$; H_0: $\beta_7 \neq 0$; H_0: $\beta_8 \neq 0$; $B_9 \neq 0$

Based on t value results obtained from analysis, model equation is as below:

$$\text{Ln(demand)} = -6129.7915 + \underset{(3.855)}{10.648} {}^{*}\text{PGROWTH} + \underset{(0.186)}{59.222} {}^{*}\text{BRATE} + \underset{(-0.274)}{(-302.186)} {}^{*}\text{MBRATE}$$

$$+ \underset{(0.604)}{33.681} {}^{*}\text{UNEMRATE} + \underset{(6.063)}{419.554} {}^{*}\text{INFLARATE} + \underset{(3.498)}{0.187} {}^{*}\text{GDPC}$$

$$+ \underset{(-0.552)}{(-519.112)} {}^{*}\text{HHOLDRATE} + \underset{(-2.785)}{(-1130.570)} {}^{*}\text{POVRATE} + \underset{(-5.053)}{(-0.067)} {}^{*}\text{HSSTOCK}$$

These results indicate that BRATE or birth rate has the smallest t absolute. Although the fit has improved somewhat only the poverty rate (POVRATE) is significant. When the regression model is nested that is when one model contains a proper subsets of the parameter to another, model F-statistics can use to test the significance of the improvement fit. The idea is exactly the same as the overall test model significance. In this case the test fit for general model in which all parameter were set to zero. Then test the fit of a more general model (denoted by an *f* for *full*) against restricted model (denoted by an *r* for *restricted*) in which only some of the parameters are set equal to zero. Results of ANOVA analysis for F-test and regression test summary is shown in Table 9 and Table 10 respectively.

Table 9: Result of ANOVA Analysis for Regression Method

Model		Sum Of Squares	df	Mean Square	F	Sig.
1	Regression	80964.048	2	8996.005	1.838	0.085(a)
	Residual	234962.573	48	4895.054		
	Total	315926.621	50			

Table 10: Model Summary for Regression Method

Model	R	R^2	$\tilde{R}^2$	Std. Error of The Estimate	R Square Change
1	0.506(a)	0.526	0.117	69.96466	0.526

Let $SSE_f \sum_i (y_i - \hat{y}_i^f)^2$ denotes the sum squared error of the full model and let $SSE_r \sum_i (y_i - \hat{y}_i^r)^2$ denotes the sum squared error of the restricted model. The *F*-test for this comparison is given by

$$F = \frac{(SSE_r - SSE_f)/(df_r - df_f)}{SSE_f/df_r}$$

Where df_f and df_r are the numbers of degrees of freedom associated with the full model and restricted model. Because 1-R^2 is directly proportional to SSE, the same test in terms of the regression as follows:

$$F = \frac{(R_f^2 - R_r^2)/(df_r - df_f)}{(1 - R_f^2)/df}$$

When the restricted model contains only an intercept term (i.e., all other model indicators are set equal to zero) then $R_r^2 = 0$. As revealed from table 10, R^2 of the general model is 0.526 (with 48 degrees of freedom) and the R^2 of the restricted model is 0.117 (with 50 degrees of freedom). The model comparison test is given by

$$F = \frac{(0.526 - 0.117)/2}{(1 - 0.526)/48} = 20.71$$

The critical value for an *F*-statistic on (2, 48) degree of freedom is 3.22 at the 0.05 level of significance (Lee 1997). Hence, it can be conclude that the improvement is fit from adding the effect of POVRATE to the model is significant (3.22<20.71) and accepted with better-fitting above described model. Also, the result show that the significant indicators for low cost housing demand in Gombak district using regression method is;

i) Poverty rate (POVRATE)

5.0 Summary

This study analyzed data regarding 9 indicators of low cost housing demand obtained from ministry of housing for Gombak district. Analysis was carried out through 3 different statistical approached for comparing those methods in determining significant indicator. Those methods included PCA, BEM and Regression. From analysis, it was found that BEM is the best method to find the significant indicators as it gives the more number of significant indicators compared to PCA and regression method. Overall results obtained from analysis for all three methods are summarized in Table 11.

Table 11: Significant indicators using PCA BEM and Regression Method

No	Method		
	PCA	**BEM**	**Regression**
I	GDP	Inflation rate	Poverty rate
Ii	Housing stock	GDP rate	-
Iii	Mortality baby rate	Poverty rate	-
Iv	-	Housing stock	-

Table 11 highlights that PCA method had identified 3 significant indicators for low cost housing demand in Gombak district while BEM had identified 4 significant indicators and regression

method had identified only 1 significant indicator. These findings will help the researchers in selecting the suitable method for studies where the aim of study is to determine significant indicator/factor.

References

Abdul Karim, M. R. (1995). *Housing for the Urban Lower Income Group*. United Nations: Public Administration and Finance, New York.

Altman D.G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall, London.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The Efficient Assessment of Need for Cognition, *Journal of Personality Assessment*, 48(3):

Campbell M.J., & Machin D. (1993). *Medical Statistics a Commonsense Approach*. 2nd edn. Wiley, London.

Chander, R. (1977). *Housing Needs versus Effective Demand in Malaysia*. Kuala Lumpur: Department of Statistics Malaysia

Carr, R., & Cummins, J. D. (1997). Strategic choices, firm efficiency and competitiveness in life insurance industry cited in Hitt, L. M. "the impact of information technology management practices on the performance of life insurance companies" in *Changes in the Life Insurance Industry: Efficiency, Technology and Risk* edited by David Cummins, Anthony Santomero

Cavalli-Sforza, L.L. (2000). *Genes, Peoples, and Languages*. North Point Press, New York. ISBN 0-86547-529-6

Cummins, J.D., Tennyson, S., & Weiss, M.A. (1999). Consolidation and efficiency in the US life insurance industry. *Journal of Banking and Finance*, 23: 325-357.

Department of Statistics, Malaysia. (2007). *Population and Housing Census, Malaysia 2007 (2007 CENSUS)*. Putrajaya: Department of Statistics, Malaysia

Fahrmeir, L., & Frost, H. (1992). On Stepwise Variable Selection in Generalized Linear Regression and Time Series Models. *Computational statistics quarterly*, 7:137-137.

Goh, B.H. (1998). *Forecasting Residential Construction Demand in Singapore. A Comparative Study of the Accuracy of Time Series, Regression and Artificial Neural Network Technique*. Engineering Construction and Architectural Management, 5(3):261-275.

Golberg, M.A., & Cho, H.A. (2004). *Introduction to Regression Analysis*. Ashurst, Ashurst, Southampton, SO40

7AA, UK: *WIT* Press.

Inoue, S., Inoue, H., Hiroyoshi, T., Matsubara, H., & Yamanaka, T. (1986). Developmental variation and amino acid sequences of cytochromes c of the fruit fly Drosophila melanogaster and the flesh fly Boettcherisca peregrina. *J. Biochem., Tokyo,* 100(4), 955-965

Johnson, D.E. (1998). *Applied Multivariate Methods for Data Analysts*. United States of America: Brooks/Cole Publishing Company

Juras, P.E., & Brooks, C.A. (1993). Supporting Operational Decision Making. *Health Care Supervisor*, 12(2): 25-31

Kooreman, P. (1994). Data envelopment analysis and parametric frontier estimation: complementary tools. *Journal of Health Economics*, 13(3): 345-346

National Housing Department (2012). *Quarterly Statistical Report* (April-June), Ministry Of Housing and Local Government.

Ofir, C., & Simonson, I. (2001). In Search of Negative Customer Feedback: The Effect of Expecting to Evaluate on Satisfaction Evaluations. *Journal of Marketing Research,* 38(2): 170-182

Sirat, M., Che Mamat, A.F., Abd Aziz, A. R., Rahim, A., Salleh, H., & Hj. Yaakob, U., (1999). *Low-cost Housing in Urban Industrial Centres of Malaysia: Issue and Challengers*. Penerbit Universiti Sains Malaysia, Kedah.

Yahya, K. (2002). *Penilaian Terhadap Keupayaan Model Rangkaian Neural BagiMeramal Permintaan Rumah Kos Rendah Kawasan Bandar Di Negeri Selangor*. Ms Eng (Construction Management), Final Year Thesis. Universiti Teknologi Malaysia

Yahya, K., & and Abd. Majid, M.Z. (2002). Comparative Study on Forecasting Demand on Low-cost House in Urban Areas Using Artificial Neural Networks And ARIMA Model. *First International Conference on Construction in the 21st Century (CITC2002),* USA: 687 – 694.

Yang, & Packer (1997). Applying Artificial Neural Networks to UK Construction Demand Forecaasting (Private

Sector). *RCIS Construction and Building Research Conference (COBRA '97),* University of Portsmouth: 1 – 16.

Yeung, K.Y., & Ruzzo, W.L. (2000). *An empirical study on Principal Component Analysis for clustering gene expression data*. Technical Report UW-CSE-2000-11-03 November, 2000

Zainun, N. Y. (2004). *Computerized Forecasting Model Based On Artificial Neural Networks For Low Cost Housing Demand In Urban Area*. Ms Eng (Construction Management), Final Year Thesis, Universiti Teknologi Malaysia.

Zainun, N.Y., Roslan, N., & Memon, A. H. (2014). Assessing Low-Cost Housing Demand in Melaka: PLS-SEM
Approach. *Advanced Materials Research*, 838-841: 3156-3162