STATISTICAL ANALYSIS OF HOUSING PRICES IN PETALING DISTRICT USING LINEAR FUNCTIONAL MODEL

By

CHOONG WEI CHENG

A thesis submitted to the Department of Mathematics and Actuarial Sciences,
Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
in partial fulfillment of the requirements for the degree of
Master of Science
April 2018

ABSTRACT

STATISTICAL ANALYSIS OF HOUSING PRICES IN PETALING DISTRICT USING LINEAR FUNCTIONAL MODEL

Choong Wei Cheng

Despite numerous research efforts to develop on housing price estimation and prediction, there is still a limitation where not considering the explanatory variables that are subject to measurement errors which might cause an underestimation of estimator variances. The compensation of the data errors was introduced in linear functional modelling, and the explanatory variables are acting as functions in the modelling technique. In this study, a multiple un-replicated linear functional relationship model is derived where its maximum likelihood estimators are obtained from a single $p - 1$ dimensional fitted plane. Its properties of unbiasedness and consistency are investigated using Taylor approximation and Fisher information matrix respectively. The discussions of the significance test of partial coefficients and coefficient of determination of the proposed model are also included in this study. The developed model is applied to real estate with 41750 terrace housing transacted records for Petaling District over November 2008 to February 2016. The individual transacted housing price is correlated with eight housing attributes and a time factor. The housing attributes that included in this study are lot size, tenure type, time to expiry of lease term, terrace type, number of bedrooms, main building size, distance to the nearest shopping mall, and

distance to the nearest supermarket. The results obtained show that the fitting and predictive abilities of the proposed model are stronger as compared to multiple regression model when applied to the training and testing samples respectively as the coefficient of determination of the proposed model is close to one while its mean square error for the training and testing samples are both smaller compared to the results obtained using multiple regression model. In this study, the attributes that significantly contributed to housing prices are identified with some justifications based on previous studies while the performances of housing markets of the study cities are analysed using the proposed model and the results showed that the housing market in Sungai Buloh is relatively more volatile compared to other study cities. Besides, this study also included a comparison between the market price movements and the estimated prices of an "average" house in Petaling District using the proposed model and the results showed that, the "average" house was sold at estimated prices that are generally higher than the market's average prices from November 2008 to February 2016.

ACKNOWLEDGEMENTS

First of all, the author would like express his thanks and appreciation to Universiti Tunku Abdul Rahman for giving him the opportunity to further his education.

Secondly, the author would like to express his sincere gratitude to his supervisor, Dr Chang Yun Fah, and co-supervisor Dr Pan Wei Yeing for the immense guidance and support in the way of the successful realisation of this thesis. Their constructive suggestions and feedbacks have further improved this research and thesis.

Besides, the author would also like to express his deepest appreciation to the Jordan Lee & Jaafar (S) Pte Ltd who provided the housing data for this research.

The acknowledgement would be incomplete without expressing the author's sincere gratitude to his family members for their morale support.

Last but not least, the author would like to thank the thesis examiners who have spent their valuable time to read and give constructive comments on this thesis.

APPROVAL SHEET

This dissertation/thesis entitled "**STATISTICAL ANALYSIS OF HOUSING PRICES IN PETALING DISTRICT USING LINEAR FUNCTIONAL MODEL**" was prepared by CHOONG WEI CHENG and submitted as partial fulfilment of the requirements for the degree of Master of Science at Universiti Tunku Abdul Rahman.

Approved by:

_____
(Dr Chang Yun Fah)
Date:……..
Supervisor
Department of Mathematics and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

_____
(Dr Pan Wei Yeing)
Date:……..
Co-supervisor
Department of Mathematics and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

LEE KONG CHIAN FACULTY OF ENGINEERING AND SCIENCE

UNIVERSITI TUNKU ABDUL RAHMAN

Date: _____

SUBMISSION OF THESIS

It is hereby certified that ***CHOONG WEI CHENG*** (ID No: ***15UEM01738***) has completed this thesis entitled "***STATISTICAL ANALYSIS OF HOUSING PRICES IN PETALING DISTRICT USING LINEAR FUNCTIONAL MODEL***" under the supervision of Dr. Chang Yun Fah (Supervisor) from the Department of Mathematics and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, and Dr. Pan Wei Yeing (Co-Supervisor) from the Department of Mathematics and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science.

I understand that University will upload softcopy of my thesis in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

_____

CHOONG WEI CHENG

DECLARATION

I hereby declare that the dissertation is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

Name _____

Date _____

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACF	Autocorrelation function
AR	Autoregressive
ARIMA	Autoregressive integrated moving average
R^2	Coefficient of determination
CV	Coefficient of variation
CPPI	Commercial property price index
CI	Confidence interval
CPI	Consumer price index
CDF	Cumulative distribution function
FIM	Fisher information matrix
GIS	Geographical Information System
GDP	Gross domestic product
HPI	House price index
HHPI	Hypothetical House Price Index
KLSE	Kuala Lumpur Stock Exchange
MHPI	Malaysian House Price Index
MSE	Mean square error
MA	Moving average
MR	Multiple regression
M_PULFR	Multiple un-replicated linear functional relationship
NAPIC	National Property Information Centre
NID	Normally and independently distributed
NLA	Net leasable area
PACF	Partial autocorrelation function

PI	Prediction interval
RPGT	Real property gains tax
SSE	Sum of squared errors
SSR	Sum of squared residuals

CHAPTER 1

INTRODUCTION

1.1 Background

Affordability of housing has become a key challenge of the local authorities to provide a better living quality and liveability of many cities (Panagiotidis and Printzis, 2015). Housing is considered one of the most valuable assets and a fundamental portfolio of a household. Hence, affordable housing provides positive externalities in terms of social stability, public welfare and economic development of a city.

However, the significant price expansion of housing market in many countries over the past decades has made housing become unaffordable and generated a high entry level for younger generations (Yardney, 2015). According to the International Demographia (2017), the housing is considered affordable if the median house price is not more than three times of the median annual household income. As in Malaysia, for example, the median house price to median annual income ratio is 4.4 times (Ismail et al., 2015), which is at the seriously unaffordable level (Suhaida et al., 2011). House price index (HPI) in Malaysia averaged 4.07% from the year 1997 to the year 2016 with an all-time high of 44.50% in the first quarter of the year 2000. HPI in Malaysia reached 5.30% in the third quarter of the year 2016 and further inclined to 5.50% in the fourth quarter of the year 2016 (Trading Economics, 2017).

Miles (2008) admonished that increases in house prices will lead to higher house price volatility, a key factor of default and prepayment of housing loans. It creates a potential house price bubble that will endanger the stability of economic (Chen et al., 2013). Solving the housing problem could be a challenging task (Joint Center for Housing Studies, 2016) as misapplying policies or solutions can increase the economic inequality in a rising housing market (Wong, 2017).

The house price movements are affected by both macro-perspective (Hashim, 2010; Panagiotidis and Printzis, 2015) and micro-perspective (Osmadi et al., 2015) determinants. This study proposes a new functional model to investigate the relationship between housing prices and a set of micro-perspective determinants. This helps to understand the preference of home buyers for certain housing attributes in six sub-regions of Petaling District, Malaysia.

1.2 Problem Statements

There are some limitations found in previous studies that motivated this study. First and foremost, most of the past studies considered housing attributes as fixed values when modelling the housing prices. However, some attributes such as distance to the nearby amenities and housing age may subject to error. This may result in developing a housing price model that cannot truly explain the actual situation of the housing market behaviour. Furthermore, in reality, houses with similar housing attributes may be sold at different prices while same house price may get houses with different housing attributes. Therefore,

it is essential to study the interrelationship between housing prices and housing attributes. This has motivated us to the use of functional relationship model where the interrelationship between dependent and independent variables can be studied.

Secondly, house price predictions are highly affected by extreme values, thus the idea of using the average house price of a certain number of nearest houses is proposed in order to diminish or eliminate the effect of extreme prices and hence increase prediction accuracy. Thirdly, the small sample sizes used in the previous studies may not be able to reflect the housing market in the study areas.

Last but not least, Petaling District is one of the most developed areas in Malaysia; nonetheless, there is no research done on the housing market in this area. Therefore, it is important to conduct a study on the housing market in Petaling District.

1.3 Objectives

The main objective of this study is to develop a new functional relationship model that can better explain the behaviour of housing markets. In this study, we focus on the housing markets in Petaling District, Selangor, Malaysia. In order to achieve this objective, we divided the study into the following subtasks:

i. To identify the key determinants of housing prices in Petaling District and its six sub-regions.

ii. To develop a model that can be used to explain the housing market in Petaling District and its six sub-regions.

iii. To understand and compare the performance of the housing markets.

iv. To evaluate and compare the performance of the developed model with multiple regression model.

v. To forecast the average house price change in trend for the housing market in Petaling District using the idea of "nearest houses".

1.4 Layout of the Thesis

The first chapter gives a general introduction on the current condition of Malaysian housing market, motivation of the study, objectives of this study and ended with the thesis layout.

In Chapter Two, some preliminaries and reviews on housing studies are provided. Besides, some widely adopted models in housing studies and a brief review of functional models are also presented. In Chapter Three, a new multiple un-replicated linear functional relationship model is proposed, and all related derivations and proofs are presented. These included the proof of maximum likelihood estimators and some properties of the proposed model.

Chapter Four is devoted to the statistical analysis of housing prices and discussions of housing market behaviour in Petaling District. Besides, the fitting ability of the proposed model is investigated and compared with multiple regression (MR) model using the coefficient of determination (R^2) and mean square error (MSE). On the other hand, the predictive ability of the proposed model is evaluated in Chapter Five. This can be done by comparing the MSEs produced by the proposed and the MR models upon predictions.

Chapter Six provides a statistical analysis of average housing price change in Petaling District. This includes a comparison between actual and estimated selling prices of an "average house". This chapter also provides discussions of the housing bubble and future price movement for the study area. The last chapter gives a concluding remark and provides some recommendations for future studies.

CHAPTER 2

SOME PRELIMINARIES AND LITERATURE REVIEW

Numerous studies of the housing market in Malaysia and other countries have been conducted in this chapter. Most of the studies focus on the housing market modelling. Besides, a review of the functional models has also been done.

2.1 Studies of Malaysian Housing Market

Hedonic pricing models, derived from a regression analysis has been widely used to measure the influencing effect of tangible and intangible building features and other outside influencing factors on the overall transaction price (Monson, 2009; Giannoulakis et al., 2016). The word hedonic is derived from the Greek word "hedonikos", which means pleasure, and the hedonic pricing model was named after a school of economics that argues the aim of all economic activities to achieve the greatest possible satisfaction and thus can avoid any damages (Giannoulakis et al., 2016). The most commonly used hedonic model is "semi-log" function model (Baranzini et al., 2008) given by $\ln(P_i) = \alpha + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i$, where P_i is the housing price of i^{th} individual house with J housing attributes, $\boldsymbol{x}_i$, and α, β_j and ε_i are the intercept, coefficients and normally-distributed error term of the model.

The Malaysian House Price Index (MHPI) was first introduced by the National Property Information Centre (NAPIC) in 1997. MHPI is a transaction based index which is computed from Laspeyres weighted formula (Francis, 2004). MHPI was first using 1990 as the base year but it was rebased to the year 2000 in order to reflect the changes in buyers' preferences. The index measures the change in price which has been paid for an "average" house using the hedonic methodology, or more commonly known as regression analysis, with the aid of principal component and two-step cluster techniques. The "average" house was priced according to a set of fixed characteristics which comprised mostly of location, physical and legal characteristics (NAPIC, 2015).

In 1999, a research done by Tan (1999) studied and estimated the house price trend from macro-perspective (a study that using factors or indicators that are pertinent to economic performance) using a hedonic model which derived from multiple regression (MR) model (Freund et al., 2006). Tan (1999) concluded per capita income, total loans to housing, unemployment rate, and Kuala Lumpur Stock Exchange (KLSE) composite index are significant determinants of MHPI for the years of 1988 to 1997. He showed that unemployment rate contributed a positive effect to the housing prices which may due be to a 'too low' unemployment rate in the later period of the study. Besides, the existence of multicollinearity which resulted from the correlation between independent variables has made the explanation on the relationship between housing prices and housing attributes become less accurate (Matignon, 2007). In order to reduce the effect of multicollinearity, Tan (1999) adopted stepwise regression method.

Micro-perspective is a study that using factors that are pertinent to physical, structural or environmental housing characteristics. In 2002, Chau and Chin (2002) used hedonic price model which derived from the consumer theory proposed by Lancaster (1966) and the model proposed by Rosen (1974) to study the behaviour of the housing market in Penang from micro-perspective for the years of 1998 to 1999. Six variables, namely, actual floor area, floor level, distance from a central business district, proximity to large shopping centre, proximity to a premier school, and availability of facilities are included in their model and all these variables are significant in affecting housing prices. However, the study only considered 120 condominium units, and this is unable to reflect the behaviour of the housing market in Penang.

In 2012, Yusof (2012) examined the Malaysian housing prices from both the micro-perspective in which for the years of 1990, 1995, 2000, 2002, 2003 and 2004, and the macro-perspective in which for the years of 1990 to 2002, using a hedonic model which utilises MR analysis. Under the micro-perspective study, Yusof (2012) found that predominant factor varies by region. For example, the variations of housing prices in Kuala Lumpur are well explained by vocational factors while the variations of housing prices in Johor Bahru are well explained by utility-bearing characteristics. Under the macro-perspective study, he found that growth of gross domestic product (GDP) is the main macro-determinant of housing prices as it explained more than 80% of the variations of MHPI.

Another research done by Yusof and Ismail (2012) analysed the housing prices of 1500 double-storey terrace houses in Kuala Lumpur, Malaysia for the years of 2000 and 2007 using MR analysis. They found that locality is the most influential determinant of housing prices of double-storey terrace in Kuala Lumpur with about 50.3% and 63.0% of the variations in housing prices for the years of 2000 and 2007 respectively are explained by locality.

Besides, a group of researchers from the Central Bank of Malaysia had studied the effects of macroeconomic-factors, financial-factors, and government regulations and policies on housing prices using multivariate regression model over a period of study in which from the first quarter of 2001 to the second quarter of 2012 (Central Bank of Malaysia, 2013). In their models, the MHPI are regressed against 13 variables such as real GDP, inflation, base lending rate, and real property gains tax (RPGT), using ordinary least square method. According to their findings, economic growth, change in demography and inflation are significant determinants of housing prices.

Norshazwani et al. (2013) studied the housing market in three sub-districts of Kuala Lumpur from micro-perspective using a hedonic method. They involved 3980 of double-storey terrace transaction data in Kuala Lumpur over a period from the first quarter of 2005 to the second quarter of 2012 in their study. They concluded that all variables such as lot area, building area, number of bedrooms, number of bathrooms, housing age, distance to the nearest city centre, sub-district, and time dummy are significant in determining

the housing prices of 3980 double-storey terrace houses in Kuala Lumpur. Besides, they used time dummies obtained from their models to estimate Hypothetical House Price Index (HHPI) using time-variant method to serve as an alternative to MHPI. However, a contradiction arose from which the HHPI with the most similar pattern to MHPI does not come from the model with the highest R^2. In other words, time dummies may not be a good estimator for housing price index in this case.

In the same year, Ong and Chang (2013) adopted regression analysis to study the macro-determinants of MHPI which included inflation rate, GDP and income increment rate over a period of the year 2000 to the second quarter of the year 2012. They concluded that GDP is the only significant factor determining the MHPI. However, this study has serious drawbacks, i.e. the model constructed explains only 15.7% of the total variability of MHPI and the significant values (*p*-values) were treated as the coefficient of an individual factor when constructing the model. Another research done by Ong (2013) analysed the variations of MHPI over the years of 2001 to 2010 using regression method. Six variables namely, GDP, population, inflation rate, cost of construction, interest rate, and RPGT were studied and only GDP, population, and RPGT are found to be significant determinants of MHPI.

Kam et al. (2016) studied 250 double-storey houses in Mukim Rawang, Selangor for the year 2014, from micro-perspective using MR model. The study found that built-up area and shopping centre are predominating the housing prices of the terrace housing in Mukim Rawang under structural

attribute and locational attribute respectively. However, their model managed to explain only 66.8% of the total variability of the housing prices.

2.2 Studies of Housing Market in Other Countries

Gabriel (1984) performed econometric analyses on the effects of structural attributes and neighbourhoods on the prices of apartments in Beer Sheva, Israel using hedonic model. The structural attributes were comprised of a number of rooms, the presence of internal improvements, and whether the property is a ground-floor-apartment. A total of 78 observations for the year 1982 were grouped into three regions, i.e. Northeast, Western, and Southwest Beer Sheva. Some neighbourhood dummy variables were used to reflect the social and economic characteristics of a particular group of immigrants such as ethnic origin, the status of socio-economic, timing of developments, and etc. Gabriel (1984) claimed that, in Israel, ground-floor-apartments are not preferable due to security problems, privacy, and noise. This perception is reflected in the models for Northeast and Western but not Southwest Beer Sheva. Besides, he also concluded that number of rooms contributed significantly to the prices of apartments in Beer Sheva.

A hedonic pricing model developed by Clark and Herrin (2000) focused on the housing prices in Fresno County, California over the period of 1990 to 1994. A total of 6837 log housing prices from single-family residential were regressed against 41 independent variables. They further categorised these independent variables into seven structural attributes, 23 neighbourhood and year sold attributes, and 11 school attributes. Their findings showed that 80.4%

of the variations in the log of housing prices can be explained by their model. Besides, they concluded among a wide range of independent variables, the school attributes contributed significantly to the housing prices. The limitations of this research are the misinterpretation of the coefficients of the model and a size of 6837 data over a five-year period may be insufficient to reflect the housing market behaviour of a county.

Besides, Ismail and Macgregor (2005) studied the housing markets in Glasgow, Scotland using a hedonic model with the aid of Geographical Information System (GIS) and spatial statistics. They considered 2715 housing prices and 61 independent variables which mainly comprised of microeconomic determinants for the year 2002. Ismail and Macgregor (2005) claimed that multicollinearity, heteroscedasticity which is resulted from an unequal variances of dependent variable that corresponding to a set of independent variables (Hamid et al., 2000), and spatial autocorrelation where a single variable is correlated with itself in space over a set of regions (Reginald and Richard, 1980) are the main sources of problems in a housing price hedonic analysis. As such, GIS with spatial statistics information were adopted into the hedonic model to detect positive spatial autocorrelation. As a result, the adjusted R^2 is improved by 3.9% to 79.7% as compared to ordinary least square model which is only 75.8%.

Monson (2009) used a hedonic model to study the condominium price in South Boston, the office building in Peoria, Illinois, and multi-family condominium units in Reston, Virginia in the United States of America (USA).

Among 22 attributes considered, Monson (2009) concluded that private outdoor space, swimming pool, attached garage, extra storage space and security systems are dominant factors in contributing to the transaction price of condominiums in South Boston. For the office building in Peoria, 10 variables over 280 office properties were analysed and the results showed that total building square footage, real commercial property price index (CPPI), green technology, year renovated (refers to housing age where a particular house is considered new after renovation), and class of the building are statistically significant. There were 154 multi-family condominiums with five variables selected in Reston. The prediction of the price using the hedonic model is fairly accurate with an average difference of 10% from the actual transaction price.

Another study done by Cebula (2009) applied the hedonic model to 2888 single-family properties in the city of Savannah and the Savannah Historic Landmark District, Georgia, USA. Cebula (2009) correlated the housing price from the year 2000 to the year 2005 with 24 potential variables, in which they can be classified into interior physical characteristics, external physical characteristics, a spatial control variable for house, and other factors associated with a house. It is found that the transaction price of single-family properties in Savannah was positively correlated with the number of bedrooms, bathrooms, fireplaces, storeys in structure, size of a garage, square footage of finished living space, an exterior construction of brick or stucco, the existence of a deck, a private courtyard, a pool/hot-tub, the house conditions (whether is new or old), location of nearby parks or squares (whether is across or adjacent to the house), and on a dead end, a lake or a river. The study was also found

that the property price was negatively impacted by the distance to the nearest apartment complex or busy street.

In 2013, Chen et al. (2013) studied the Beijing housing market from macro-perspective using model developed by Coleman et al. (2008) which is based on vector error correction model (Pfaff, 2008) to investigate the existence of housing bubble from 1998 to 2010 using economic fundamental factors which included GDP, interest rates, inflation (or the consumer price index, CPI), and construction cost. Two models were developed to study the long-run trend and the short term dynamic of the house prices in Beijing with 83.55% and 80.20% of the variations of housing price index can be explained by the long-run and short-run models respectively. In their paper, the results showed that GDP and construction cost are significant determinants (with a significance level of 0.05) of HPI for the long-run while GDP and CPI are the significant determinants of the HPI for the short-run. Besides, the results also showed that housing bubble was likely present in Beijing from 2006 to 2007 and 2005 to 2007 for the short-run and long-run analyses respectively. There are some limitations in the paper where the authors misused the significance value when performing hypotheses testing and they explained the coefficients of the models in a wrong manner.

A new hedonic model which included the location and individual fixed effects was proposed by Jiang et al. (2014). This new hybrid approach is less prone to specification errors and with a greater computational efficiency. Jiang et al. (2014) fitted the model to private single-sale and repeat-sale properties in

Singapore between 1995 and 2014. The hybrid hedonic model slightly outperformed the Case-Shiller index (Case and Shiller, 1987; 1989) in predicting the price of single-sales homes out-of-sample cases, but less accurate for repeat-sales homes out-of-sample cases.

Giannoulakis et al. (2016) applied the hedonic model to evaluate the impact of the financial crisis in Greek on the residential property market. They considered 5000 comparable sale and valuation reports on residential property in Thessaloniki area from 2006 to 2016. The study concluded that structural factors which included age, size, floor, and quality of construction and view are significant and influential determinants of housing price in Thessaloniki area. It was also found that financial factors such as unemployment, construction activity, and CPI are significant factors of the Greek debt crisis which created a chaotic real estate environment.

2.3 A Review of Functional Models

Linear regression model has been widely used in studying the relationship between a continuous dependent variable and a set of independent variables. However, without considering the explanatory variables that are subject to measurement errors (as presented in studies of Chau and Chin, 2002; Yusof, 2012; Yusof and Ismail, 2012, and etc.) might cause an underestimation of estimator variances, moreover in many cases, the relationship between dependent and independent variables will become invisible as a result of random fluctuations associated with variables. As Fuller (1987) has pointed out, it is unrealistic if an independent variable can be measured exactly in all

situations. This might result in developing a poor model that cannot explain the actual situation of the study on housing market (low R^2 as presented in the studies of Ong and Chang, 2013; Kam et al., 2016 and etc.).

The functional model is introduced in accordance with the above issues. Adcock (1877) had first studied the problem using a functional model where both dependent and independent variables are subject to errors. Suppose that Y_i and X_i are unobservable dependent and independent variables respectively, where $Y_i = \alpha + X_i\beta$ and the corresponding random variables y_i and x_i that are observed with errors, $\varepsilon_i \sim NID\left(0, \sigma_1^2\right)$ and $\delta_i \sim NID\left(0, \sigma_2^2\right)$ respectively, such that

$$\left.\begin{aligned} y_i &= Y_i + \varepsilon_i \\ x_i &= X_i + \delta_i \end{aligned}\right\} \; i = 1,2,\cdots,n\,.$$

Pearson (1901) extended Adcock's work to multiple (principal component) relationship of a set of *p* independent variables with equal error variances. Pearson (1901) made use of the moment of inertia of *n* points, in which it yielded a set of $p-1$ dimensional fitted planes. In 1945, Reiersol (1945) included an instrumental variable called *Z* in the functional model to account for unexpected behaviours between variables *X* and *Y*. On the other hand, a discussion on multidimensional functional relationship with a single linear functional relationship was proposed by Sprent (1969) in 1969 where there is at least one (meaning one or more) independent linear relationship or replication with each showing a space of $p-1$ dimensions.

In 1984, Chan and Mak (1984) proposed a multivariate linear functional relationship model in which error variances and covariances are unnecessary to be homogenous. In 2002, James (2002) proposed a functional generalized linear model to handle functional independent variables which may be measured at differing time points and sample sizes. Caires and Wyatt (2003) introduced a linear functional relationship model with numerical approximation as a solution for its maximum likelihood estimation to compare two sets of circular data which are subjected to unobservable errors. Chang et al. (2010) generalized the un-replicated linear functional relationship model to multidimensional cases to assess the quality of JPEG compressed images.

The next chapter develops a functional relationship of dependent variable as a linear combination of p independent variables in a single $p-1$ dimensional fitted plane. The properties of the proposed model will be investigated and the coefficient of determination will be derived as a performance indicator.

CHAPTER 3

THE PROPOSED MODEL – M_PULFR MODEL

From Section 2.1, it is observed that multiple regression (MR) model is commonly used to study and analyse the Malaysian housing market. The main limitation of MR model is the statistical and inferential problems of multicollinearity which can cause the interpretation of the linear relationship between independent variables and dependent variable becomes nearly impossible (Matignon, 2007).

In this chapter, a multiple un-replicated linear functional relationship (M_PULFR) model is derived where its maximum likelihood estimators are obtained from a single $p-1$ dimensional fitted plane. The unbiasedness and consistency properties of M_PULFR model is discussed using Taylor approximation (Hox, 2002) and Fisher information (Yan and Su, 2009), respectively. The coefficient of determination and confidence interval of M_PULFR model will also be discussed.

3.1 Derivation of M_PULFR Model

Suppose that Y_i is an unobservable dependent variable and vector $\boldsymbol{X}_i = \left(X_{i1}, X_{i2}, \cdots, X_{ip}\right)$ be a vector consisting of p unobservable independent variables such that,

$$Y_i = \alpha + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots X_{ip}\beta_p = \alpha + \boldsymbol{X}_i\boldsymbol{\beta}\,,\ i = 1,2,\cdots,n\,, \tag{3.1}$$

where α is an intercept, each β_k is the coefficient of the linear function and we let vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$. The corresponding random variables y_i and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ are observed with errors, ε_i and $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{ip})$ such that,

$$\left.\begin{aligned} y_i &= Y_i + \varepsilon_i \\ \boldsymbol{x}_i &= \boldsymbol{X}_i + \boldsymbol{\delta}_i \end{aligned}\right\} \; i = 1,2,\cdots,n, \tag{3.2}$$

and both error vectors are assumed to be mutually independent and normally distributed with the following properties:

1. $E(\varepsilon_i) = 0$ and $E(\boldsymbol{\delta}_i) = \mathbf{0}$,
2. $Cov(\varepsilon_i, \varepsilon_j) = 0$ and $Cov(\boldsymbol{\delta}_i, \boldsymbol{\delta}_j) = \mathbf{0}\ \forall i \neq j$,
3. $Cov(\varepsilon_i, \delta_{ik}) = 0\ \forall i, k$ and
4. $\varepsilon_i \sim NID(0, \omega_{11})$ and $\boldsymbol{\delta}_i \sim NID(\mathbf{0}, \boldsymbol{\omega}_{22})$ where we let

$$\boldsymbol{\omega} = \begin{pmatrix} \omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\omega}_{22} \end{pmatrix} = \begin{pmatrix} \tau^2 & \mathbf{0} \\ \mathbf{0} & \sigma^2 \mathbf{I}_p \end{pmatrix}.$$

Result 1:

Given the MPULFR model defined by Equations (3.1) and (3.2), the maximum likelihood estimators of α, β_k, $\boldsymbol{X}_i$, and σ are,

$$\hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}},$$

$$\hat{\beta}_k = \frac{\left(S_{yy} - \lambda S_{x_k x_k}\right) + \sqrt{\left(S_{yy} - \lambda S_{x_k x_k}\right)^2 + 4\lambda S_{x_k y}^2}}{2S_{x_k y}}, \; k = 1, 2, \cdots, p,$$

$$\hat{\boldsymbol{X}}_i = \left[\lambda \boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda \mathrm{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}, \; i = 1, 2, \cdots, n,$$

$$\hat{\sigma}^2 = \frac{SSE}{(p+1)n},$$

where $S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$, $S_{x_kx_k}=\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2$, $S_{x_ky}=\sum_{i=1}^{n}(y_i-\bar{y})x_{ik}$, the residual sum of square, $SSE=\sum_{i=1}^{n}\left[(\boldsymbol{x}_i-\hat{\boldsymbol{X}}_i)(\boldsymbol{x}_i-\hat{\boldsymbol{X}}_i)'+\frac{1}{\lambda}\left(y_i-\hat{\alpha}-\hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)^2\right]$ and $\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'$ is a reversible $p\times p$ symmetric matrix and λ is a positive ratio of error variances.

Proof:

The joint density function of $(x_{i1},x_{i2},\cdots,x_{ip},y_i)$ or equivalently, $(\boldsymbol{x}_i,y_i)$ is

$$f(\boldsymbol{x}_i,y_i)=\frac{1}{(2\pi)^{\frac{r}{2}}|\boldsymbol{\omega}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\left\{\begin{bmatrix}(y_i-Y_i) & (\boldsymbol{x}_i-\boldsymbol{X}_i)\end{bmatrix}\boldsymbol{\omega}^{-1}\begin{pmatrix}y_i-Y_i\\ [\boldsymbol{x}_i-\boldsymbol{X}_i]'\end{pmatrix}\right\}\right], \tag{3.3}$$

where $r=p+1$, $E(\boldsymbol{x}_i)=E(\boldsymbol{X}_i+\boldsymbol{\delta}_i)=\boldsymbol{X}_i$ and $E(y_i)=E(Y_i+\varepsilon_i)=Y_i$. The likelihood function is given by,

$$L=\prod_{i=1}^{n}f(\boldsymbol{x}_i,y_i)=\prod_{i=1}^{n}\frac{1}{(2\pi)^{\frac{r}{2}}|\boldsymbol{\omega}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\left\{\begin{bmatrix}(y_i-Y_i) & (\boldsymbol{x}_i-\boldsymbol{X}_i)\end{bmatrix}\boldsymbol{\omega}^{-1}\begin{pmatrix}y_i-Y_i\\ [\boldsymbol{x}_i-\boldsymbol{X}_i]'\end{pmatrix}\right\}\right]$$

$$=\frac{1}{(2\pi)^{\frac{rn}{2}}|\boldsymbol{\omega}|^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left\{\begin{bmatrix}(y_i-Y_i) & (\boldsymbol{x}_i-\boldsymbol{X}_i)\end{bmatrix}\begin{pmatrix}\omega_{11}^{-1} & 0\\ 0 & \boldsymbol{\omega}_{22}^{-1}\end{pmatrix}\begin{pmatrix}y_i-Y_i\\ [\boldsymbol{x}_i-\boldsymbol{X}_i]'\end{pmatrix}\right\}\right]$$

$$=\frac{1}{(2\pi)^{\frac{rn}{2}}|\boldsymbol{\omega}|^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left\{(y_i-Y_i)\omega_{11}^{-1}(y_i-Y_i)+(\boldsymbol{x}_i-\boldsymbol{X}_i)\boldsymbol{\omega}_{22}^{-1}[\boldsymbol{x}_i-\boldsymbol{X}_i]'\right\}\right],$$

and the log-likelihood function is,

$$L^*=-\ln(2\pi)^{\frac{rn}{2}}-\frac{n}{2}\ln|\boldsymbol{\omega}|-\frac{1}{2}\sum_{i=1}^{n}\left[(y_i-Y_i)^2\omega_{11}^{-1}+(\boldsymbol{x}_i-\boldsymbol{X}_i)\boldsymbol{\omega}_{22}^{-1}[\boldsymbol{x}_i-\boldsymbol{X}_i]'\right]. \tag{3.4}$$

Since $Y_i=\alpha+\boldsymbol{X}_i\boldsymbol{\beta}$, $\omega_{11}=\tau^2$, and $\boldsymbol{\omega}_{22}=\sigma^2\mathbf{I}_p$, Equation (3.4) becomes,

$$L^*=-\ln K-\frac{n}{2}\ln|\boldsymbol{\omega}|-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{1}{\tau^2}(y_i-\alpha-\boldsymbol{X}_i\boldsymbol{\beta})^2+(\boldsymbol{x}_i-\boldsymbol{X}_i)\frac{1}{\sigma^2}[\boldsymbol{x}_i-\boldsymbol{X}_i]'\right],$$

where $K=(2\pi)^{\frac{rn}{2}}$ and for simplicity, we let $\tau^2=\lambda\sigma^2$, where λ is a positive constant, then $|\boldsymbol{\omega}|=\lambda\sigma^{2(p+1)}$, thus,

$$L^*=K^*-(p+1)n\ln\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\frac{1}{\lambda}(y_i-\alpha-\boldsymbol{\beta}'\boldsymbol{X}_i')^2+(\boldsymbol{x}_i-\boldsymbol{X}_i)(\boldsymbol{x}_i-\boldsymbol{X}_i)'\right], \quad (3.5)$$

where $K^*=-\ln K-\frac{n}{2}\ln\lambda$ and $\boldsymbol{\beta}'\boldsymbol{X}_i'=\boldsymbol{X}_i\boldsymbol{\beta}$.

Hence, differentiate Equation (3.5) with respect to α and equate the result to zero,

$$\frac{\partial L^*}{\partial\alpha}=-\frac{1}{2\lambda\hat{\sigma}^2}\sum_{i=1}^{n}2\left(y_i-\hat{\alpha}-\hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)(-1)=0$$

$$\sum_{i=1}^{n}y_i-n\hat{\alpha}-\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}=0$$

$$\hat{\alpha}=\bar{y}-\frac{1}{n}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\hat{\boldsymbol{\beta}}. \quad (3.6)$$

Similarly, differentiate Equation (3.5) with respect to $\boldsymbol{\beta}$, $\boldsymbol{X}_i$ and σ respectively, we will obtain

$$\frac{\partial L^*}{\partial\boldsymbol{\beta}'}=-\frac{1}{2\lambda\hat{\sigma}^2}\sum_{i=1}^{n}2\left(y_i-\hat{\alpha}-\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{X}}_i'\right)\left(-\hat{\boldsymbol{X}}_i\right)=\mathbf{0}$$

$$\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i=\sum_{i=1}^{n}y_i\hat{\boldsymbol{X}}_i-\hat{\alpha}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i$$

$$\hat{\boldsymbol{\beta}}'=\left(\sum_{i=1}^{n}y_i\hat{\boldsymbol{X}}_i-\hat{\alpha}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1}, \quad (3.7)$$

$$\frac{\partial L^*}{\partial\boldsymbol{X}_i}=-\frac{1}{2\hat{\sigma}^2}\left[\frac{2}{\lambda}\left(y_i-\hat{\alpha}-\hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)\left(-\hat{\boldsymbol{\beta}}'\right)+2\left(\boldsymbol{x}_i-\hat{\boldsymbol{X}}_i\right)(-1)\right]=\mathbf{0}$$

$$\hat{\boldsymbol{X}}_i\left(\lambda\mathbf{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)=\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'$$

$$\hat{\boldsymbol{X}}_i=\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda\mathbf{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}, \quad (3.8)$$

and,

$$\frac{\partial L^*}{\partial \sigma} = -\frac{(p+1)n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}\left[\left(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i\right)\left(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i\right)' + \frac{1}{\lambda}\left(y_i - \hat{\alpha} - \hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)^2 \right] = 0$$

$$\hat{\sigma}^2 = \frac{1}{(p+1)n}\sum_{i=1}^{n}\left[\left(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i\right)\left(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i\right)' + \frac{1}{\lambda}\left(y_i - \hat{\alpha} - \hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)^2 \right]. \tag{3.9}$$

To estimate $\hat{\alpha}$, substitute Equation (3.8) into Equation (3.6) and get,

$$\hat{\alpha} = \bar{y} - \frac{1}{n}\left\{\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\right\}\hat{\boldsymbol{\beta}}$$

$$\hat{\alpha}\hat{\boldsymbol{\beta}}' = \bar{y}\hat{\boldsymbol{\beta}}' - \frac{1}{n}\left\{\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\right\}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'$$

$$\hat{\alpha}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} = \bar{y}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} - \frac{1}{n}\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}$$

$$\hat{\alpha}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right) = \bar{y}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right) - \frac{1}{n}\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$

$$\lambda\hat{\alpha}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} + \hat{\alpha}\hat{\boldsymbol{\beta}}' = \lambda\bar{y}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} + \bar{y}\hat{\boldsymbol{\beta}}' - \frac{\lambda}{n}\sum_{i=1}^{n}\boldsymbol{x}_i - \frac{1}{n}\sum_{i=1}^{n}y_i\hat{\boldsymbol{\beta}}' + \frac{1}{n}\sum_{i=1}^{n}\hat{\alpha}\hat{\boldsymbol{\beta}}'$$

$$\hat{\alpha}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} = \bar{y}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} - \bar{\boldsymbol{x}}$$

$$\therefore \hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}, \text{ where } \bar{\boldsymbol{x}} = \left(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p\right). \tag{3.10}$$

To estimate $\hat{\boldsymbol{\beta}}$, substitute Equation (3.8) into Equation (3.7) and rearrange the outcome,

$$\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\left\{\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\right\}'\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}$$
$$= \sum_{i=1}^{n}(y_i - \hat{\alpha})\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}.$$

For any matrices $\mathbf{A}$, $\mathbf{B}$, and symmetric matrix $\mathbf{C}$, we have $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ and $\mathbf{C}' = \mathbf{C}$. By using these properties of transpose:

$$\hat{\boldsymbol{\beta}}'\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i' + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}\right]\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$
$$= \sum_{i=1}^{n}(y_i - \hat{\alpha})\left[\lambda\boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right] \qquad \because \left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} \text{ is a symmetric matrix}$$

$$\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\left(\lambda\mathbf{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i'+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}\right]\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$
$$=\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i-\hat{\alpha})\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$

and for a square matrix $\mathbf{A}$ and symmetric matrix $\mathbf{B}$, $\mathbf{AB}=\mathbf{BA}$, then,

$$\left(\lambda\mathbf{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i'+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}\right]\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$
$$=\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i-\hat{\alpha})\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$

$$\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i'+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}\right]\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$
$$=\lambda\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i-\hat{\alpha})\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]+\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i-\hat{\alpha})\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$

$$\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\left[\lambda\boldsymbol{x}_i'+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}\right]\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$
$$=\lambda\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i-\hat{\alpha})\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$
$$+\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i-\hat{\alpha})\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$

$$\lambda\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{x}_i+\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\left[\boldsymbol{x}_i'(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]$$
$$=\lambda\sum_{i=1}^{n}(y_i-\hat{\alpha})\boldsymbol{x}_i+\sum_{i=1}^{n}(y_i-\hat{\alpha})^2\hat{\boldsymbol{\beta}}'$$

$$\lambda\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{x}_i\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\boldsymbol{x}_i'(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$
$$=\lambda\sum_{i=1}^{n}(y_i-\hat{\alpha})\boldsymbol{x}_i\hat{\boldsymbol{\beta}}+\sum_{i=1}^{n}(y_i-\hat{\alpha})^2\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$

$$\therefore\ \lambda\sum_{i=1}^{n}\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)^2+\left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}-\lambda\right)\sum_{i=1}^{n}(y_i-\hat{\alpha})\boldsymbol{x}_i\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i-\hat{\alpha})^2=0\,. \qquad (3.11)$$

When substitute Equation (3.10) into Equation (3.11), we will get

$$\lambda\sum_{i=1}^{n}\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)^2+\left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}-\lambda\right)\sum_{i=1}^{n}\left(y_i-\bar{y}+\bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}\right)\boldsymbol{x}_i\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}\left(y_i-\bar{y}+\bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}\right)^2=0$$

$$\lambda\sum_{i=1}^{n}\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)^2+\left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}-\lambda\right)\sum_{i=1}^{n}(y_i-\bar{y})\boldsymbol{x}_i\hat{\boldsymbol{\beta}}-\lambda n\left(\bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}\right)^2-\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i-\bar{y})^2=0$$

and this equation can be written as,

$$\lambda\sum_{i=1}^{n}\left(\sum_{j=1}^{p}x_{ij}\hat{\beta}_j\right)^2+\left(\sum_{j=1}^{p}\hat{\beta}_j^2-\lambda\right)\sum_{i=1}^{n}\left[(y_i-\bar{y})\sum_{j=1}^{p}x_{ij}\hat{\beta}_j\right]-\lambda n\left(\sum_{j=1}^{p}\bar{x}_j\hat{\beta}_j\right)^2$$
$$-\sum_{j=1}^{p}\hat{\beta}_j^2\sum_{i=1}^{n}(y_i-\bar{y})^2=0.$$

To solve for $\hat{\beta}_k$, the coefficient of the k^{th} independent variable,

$$\lambda\hat{\beta}_k^2\sum_{i=1}^{n}x_{ik}^2+\left(\hat{\beta}_k^2-\lambda\right)\sum_{i=1}^{n}(y_i-\bar{y})x_{ik}\hat{\beta}_k-\lambda n\hat{\beta}_k^2\bar{x}_k^2-\hat{\beta}_k^2\sum_{i=1}^{n}(y_i-\bar{y})^2=0$$

$$S_{x_ky}\hat{\beta}_k^2+\left(\lambda S_{x_kx_k}-S_{yy}\right)\hat{\beta}_k-\lambda S_{x_ky}=0$$

$$\therefore\hat{\beta}_k=\frac{\left(S_{yy}-\lambda S_{x_kx_k}\right)+\sqrt{\left(S_{yy}-\lambda S_{x_kx_k}\right)^2+4\lambda S_{x_ky}^2}}{2S_{x_ky}},\qquad(3.12)$$

where $S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$, $S_{x_kx_k}=\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2$, and $S_{x_ky}=\sum_{i=1}^{n}(y_i-\bar{y})x_{ik}$.

It is observed that, the estimated individual beta, $\hat{\beta}_k$ does not depend on other independent variables other than its respective independent variables. This implies that the estimated $\hat{\beta}_k$ for MPULFR model is not affected by multicollinearity which resulted from the correlation between independent variables (Matignon, 2007).

To estimate $\hat{\sigma}^2$, we substitute Equation (3.8) into Equation (3.9) and get,

$$\hat{\sigma}^2=\frac{1}{(p+1)n}\sum_{i=1}^{n}[\mathbf{ab}+c]$$

where

$$\boldsymbol{a}=\boldsymbol{b}'=\boldsymbol{x}_i-\left[\lambda\boldsymbol{x}_i+(y_i-\hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda\mathrm{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}=-\left(y_i-\hat{\alpha}-\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{\beta}}'\left(\lambda\mathrm{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1},$$

and

$$
\begin{aligned}
c &= \frac{1}{\lambda}\left\{ y_i - \hat{\alpha} - \left[\left(\lambda \boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}' \right) \left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \right] \hat{\boldsymbol{\beta}} \right\}^2 \\
&= \frac{1}{\lambda}\left\{ \left[(y_i - \hat{\alpha}) \left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \right)^{-1} \hat{\boldsymbol{\beta}}' \left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right) - \left(\lambda \boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}' \right) \right] \left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \hat{\boldsymbol{\beta}} \right\}^2 \\
&= \frac{1}{\lambda}\left\{ \left[\lambda (y_i - \hat{\alpha}) \left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \right)^{-1} \hat{\boldsymbol{\beta}}' - \lambda \boldsymbol{x}_i \right] \left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \hat{\boldsymbol{\beta}} \right\}^2 \\
&= \lambda \left(y_i - \hat{\alpha} - \lambda \boldsymbol{x}_i \hat{\boldsymbol{\beta}} \right)^2 \left[\hat{\boldsymbol{\beta}}' \left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \hat{\boldsymbol{\beta}} \right]^2,
\end{aligned}
$$

then,

$$
\begin{aligned}
\therefore \hat{\sigma}^2 = \frac{1}{(p+1)n} \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \boldsymbol{x}_i \hat{\boldsymbol{\beta}} \right)^2 \Big\{ & \hat{\boldsymbol{\beta}}' \left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \hat{\boldsymbol{\beta}} \\
& + \lambda \left[\hat{\boldsymbol{\beta}}' \left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \left(\lambda \mathrm{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \right)^{-1} \hat{\boldsymbol{\beta}} \right]^2 \Big\}. \qquad (3.13)
\end{aligned}
$$

3.2 Properties of Parameters of M_PULFR Model

The probability distribution of the error terms, ε_i and $\boldsymbol{\delta}_i$ in Equation (3.2) are not known in general, so, is a must to study the properties of the estimators for α and β, and to ensure that the estimated parameters $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ fulfil certain properties such as unbiasedness and consistency.

3.2.1 Unbiased Estimators

Result 2:

The maximum likelihood estimators of α and $\boldsymbol{\beta}$ are approximate unbiased estimators,

$$
\mathrm{E}(\hat{\alpha}) = \alpha \qquad \text{and} \qquad \mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.
$$

Proof:

Let $\theta_k = \dfrac{S_{yy} - \lambda S_{x_k x_k}}{2S_{x_k y}}$, we can rewrite Equation (3.12) as $\hat{\beta}_k = \theta_k + \sqrt{\theta_k^2 + \lambda}$.

The expected value of $\hat{\beta}_k$ is,

$$\mathrm{E}\left(\hat{\beta}_k\right) = \mathrm{E}\left(\theta_k\right) + \mathrm{E}\left(\sqrt{\theta_k^2 + \lambda}\right). \tag{3.14}$$

To solve Equation (3.14), we used the first order of Taylor approximations for the mean of $\theta_k\left(x_{ik}, y_i\right)$. Given that

$$\theta_k\left(x_{ik}, y_i\right) = \theta_k\left(X_{ik} + \delta_{ik}, Y_i + \varepsilon_i\right) = \theta_k\left(X_{ik}, Y_i\right) + \delta'_{ik}\left.\frac{\partial \theta_k}{\partial x_{ik}}\right|_{x_{ik}=X_{ik}} + \varepsilon'_i\left.\frac{\partial \theta_k}{\partial y_i}\right|_{y_i=Y_i}, \tag{3.15}$$

where the partial derivatives are evaluated at the mean $\left(X_{ik}, Y_i\right)$ and the Equation (3.15) will be valid if and only if the error variances, $\sigma_{\boldsymbol{\delta}}^2$ and σ_{ε}^2 are small. Since

$$\mathrm{E}\left(\delta'_{ik}\left.\frac{\partial \theta_k}{\partial x_{ik}}\right|_{x_{ik}=X_{ik}}\right) = \sum_{k=1}^{p}\left[\left.\frac{\partial \theta_k}{\partial x_{ik}}\right|_{x_{ik}=X_{ik}} \mathrm{E}\left(\delta_{ik}\right)\right] = 0,$$

Then $\mathrm{E}\left(\boldsymbol{\delta}_i\right) = \mathrm{E}\left(\mathbf{0}\right) \rightarrow \mathrm{E}\left(\delta_{ik}\right) = 0$.

Similarly, $\mathrm{E}\left(\varepsilon'_i\left.\dfrac{\partial \theta_k}{\partial y_i}\right|_{y_i=Y_i}\right) = 0$, therefore, the expected value of Equation (3.15) is,

$$\mathrm{E}\left[\theta_k\left(x_{ik}, y_i\right)\right] = \mathrm{E}\left[\theta_k\left(X_{ik}, Y_i\right)\right] = \theta_k\left(X_{ik}, Y_i\right) = \frac{S_{YY} - \lambda S_{X_k X_k}}{2S_{X_k Y}}, \tag{3.16}$$

where $\theta_k\left(X_{ik}, Y_i\right)$ is a fixed value.

Now let $\vartheta_k(x_{ik}, y_i) = \sqrt{\theta_k^2(x_{ik}, y_i) + \lambda}$ for the second term of Equation (3.14). This implies that $\frac{\partial \vartheta_k}{\partial x_{ik}} = (\theta_k^2 + \lambda)^{-\frac{1}{2}} \theta_k \frac{\partial \theta_k}{\partial x_{ik}}$ and by using the first order of Taylor approximations,

$$\mathrm{E}[\vartheta_k(x_{ik}, y_i)] = \vartheta_k(X_{ik}, Y_i) = \sqrt{\theta_k^2(X_{ik}, Y_i) + \lambda} = \sqrt{\left(\frac{S_{YY} - \lambda S_{X_k X_k}}{2S_{X_k Y}}\right)^2 + \lambda}\,. \qquad (3.17)$$

When substitute Equation (3.16) and Equation (3.17) into Equation (3.14), we will obtain,

$$\mathrm{E}(\hat{\beta}_k) = \frac{(S_{YY} - \lambda S_{X_k X_k}) + \sqrt{(S_{YY} - \lambda S_{X_k X_k})^2 + 4\lambda S_{X_k Y}^2}}{2S_{X_k Y}}. \qquad (3.18)$$

When we rewrite S_{YY} and $S_{X_k Y}$ in term of $S_{X_k X_k}$, we have

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\alpha + \boldsymbol{X}_i\boldsymbol{\beta} - \alpha - \bar{\boldsymbol{X}}\boldsymbol{\beta})^2 = \sum_{i=1}^{n} (\boldsymbol{X}_i\boldsymbol{\beta})^2 - n(\bar{\boldsymbol{X}}\boldsymbol{\beta})^2$$

$$\therefore S_{YY}\big|_{x_{ik}=X_{ik}} = \left(\sum_{i=1}^{n} X_{ik}^2 - n\bar{X}_k^2\right)\beta_k^2 = \beta_k^2 S_{X_k X_k}$$

and,

$$S_{\boldsymbol{X}Y} = \sum_{i=1}^{n} \boldsymbol{X}_i' Y_i - n\bar{\boldsymbol{X}}'\bar{Y} = \sum_{i=1}^{n} \boldsymbol{X}_i'(\alpha + \boldsymbol{X}_i\boldsymbol{\beta}) - n\bar{\boldsymbol{X}}'(\alpha + \bar{\boldsymbol{X}}\boldsymbol{\beta}) = \left(\sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{X}_i - n\bar{\boldsymbol{X}}'\bar{\boldsymbol{X}}\right)\boldsymbol{\beta}$$

$$\therefore S_{X_k Y}\big|_{x_{ik}=X_{ik}} = \left(\sum_{i=1}^{n} X_{ik}^2 - n\bar{X}_k^2\right)\beta_k = \beta_k S_{X_k X_k}\,.$$

Thus, Equation (3.18) becomes,

$$\mathrm{E}(\hat{\beta}_k) = \frac{(\beta_k^2 S_{X_k X_k} - \lambda S_{X_k X_k}) + \sqrt{(\beta_k^2 S_{X_k X_k} - \lambda S_{X_k X_k})^2 + 4\lambda(\beta_k S_{X_k X_k})^2}}{2\beta_k S_{X_k X_k}} = \beta_k\,.$$

From Equation (3.10), $\hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}$, then,

$$\mathrm{E}(\hat{\alpha}) = \mathrm{E}(\bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}) = \bar{y} - \bar{\boldsymbol{x}}\boldsymbol{\beta} = \alpha\,.$$

Therefore, both $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are approximate unbiased estimators of α and $\boldsymbol{\beta}$ respectively.

3.2.2 Consistent Estimators

Result 3:

Given the MPULFR model, $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are consistent maximum likelihood estimators of α and $\boldsymbol{\beta}$ respectively.

Proof:

To prove that $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are consistent estimators of α and $\boldsymbol{\beta}$ respectively, we first have to obtain their variances. The Fisher information matrix of parameters $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ is used to obtain the variance and covariance of $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$. The second order partial derivatives for the log-likelihood function and their negative expected values are,

$$\frac{\partial^2 L^*}{\partial \alpha^2} = -\frac{n}{\lambda\sigma^2} \quad \text{hence} \quad \mathrm{E}\left(-\frac{\partial^2 L^*}{\partial \alpha^2}\right) = \frac{n}{\lambda\sigma^2},$$

$$\frac{\partial^2 L^*}{\partial \alpha \partial \boldsymbol{\beta}} = -\frac{1}{\lambda\sigma^2}\sum_{i=1}^{n} \boldsymbol{X}_i' \quad \text{hence} \quad \mathrm{E}\left(-\frac{\partial^2 L^*}{\partial \alpha \partial \boldsymbol{\beta}}\right) = \frac{1}{\lambda\sigma^2}\sum_{i=1}^{n} \boldsymbol{X}_i',$$

$$\frac{\partial^2 L^*}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}} = -\frac{1}{\lambda\sigma^2}\sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{X}_i \quad \text{hence} \quad \mathrm{E}\left(-\frac{\partial^2 L^*}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}}\right) = \frac{1}{\lambda\sigma^2}\sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{X}_i,$$

$$\frac{\partial^2 L^*}{\partial \boldsymbol{\beta}' \partial \alpha} = -\frac{1}{\lambda\sigma^2}\sum_{i=1}^{n} \boldsymbol{X}_i \quad \text{hence} \quad \mathrm{E}\left(-\frac{\partial^2 L^*}{\partial \boldsymbol{\beta}' \partial \alpha^2}\right) = \frac{1}{\lambda\sigma^2}\sum_{i=1}^{n} \boldsymbol{X}_i,$$

then,

$$\mathbf{F}=\begin{bmatrix} \dfrac{n}{\lambda\hat{\sigma}^2} & \dfrac{1}{\lambda\hat{\sigma}^2}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i \\ \dfrac{1}{\lambda\hat{\sigma}^2}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i' & \dfrac{1}{\lambda\hat{\sigma}^2}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i \end{bmatrix}=\begin{bmatrix} A & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix},$$

where $A=\dfrac{n}{\lambda\hat{\sigma}^2}$ is a 1×1 matrix, $\boldsymbol{B}=\dfrac{1}{\lambda\hat{\sigma}^2}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i$ is a $1\times p$ matrix, $\boldsymbol{C}=\boldsymbol{B}'=\dfrac{1}{\lambda\hat{\sigma}^2}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'$ is a $p\times1$ matrix, and $\boldsymbol{D}=\dfrac{1}{\lambda\hat{\sigma}^2}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i$ is a $p\times p$ matrix are the negative expected values of the second partial derivatives for the log-likelihood function. The inverse of $\mathbf{F}$ is

$$\mathbf{F}^{-1}=\begin{bmatrix} \left(A-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & -A^{-1}\boldsymbol{B}\left(\boldsymbol{D}-\boldsymbol{C}A^{-1}\boldsymbol{B}\right)^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\left(A-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & \left(\boldsymbol{D}-\boldsymbol{C}A^{-1}\boldsymbol{B}\right)^{-1} \end{bmatrix}=\begin{bmatrix} \hat{\mathrm{V}}\mathrm{ar}(\hat{\alpha}) & \hat{\mathrm{C}}\mathrm{ov}\left(\hat{\alpha},\hat{\boldsymbol{\beta}}\right) \\ \hat{\mathrm{C}}\mathrm{ov}\left(\hat{\alpha},\hat{\boldsymbol{\beta}}\right) & \hat{\mathrm{V}}\mathrm{ar}\left(\hat{\boldsymbol{\beta}}\right) \end{bmatrix},$$

thus, the variance and covariance of $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{\alpha})=\lambda\hat{\sigma}^2\left[n-\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right)\right]^{-1}, \tag{3.19}$$

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\boldsymbol{\beta}}\right)=\lambda\hat{\sigma}^2\left[\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i-\frac{1}{n}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\right]^{-1}, \tag{3.20}$$

$$\hat{\mathrm{C}}\mathrm{ov}\left(\hat{\alpha},\hat{\boldsymbol{\beta}}\right)=-\lambda\hat{\sigma}^2\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\left[n-\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right]^{-1}. \tag{3.21}$$

Now, we apply the following mathematical theorems to prove that $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are consistent estimators of α and $\boldsymbol{\beta}$ respectively.

Definition 3.1:

An estimator of θ, $\hat{\theta}_n$ with random sample of size *n*, is said to be consistent if $\hat{\theta}_n$ converges to θ as *n* approaches infinity. Mathematically,

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\hat{\theta}_n - \theta\right| > \gamma\right) = 0, \forall \gamma > 0 .$$

Theorem 3.1:

According to *Chebyshev's Inequality*, if a random variable, X with a finite mean, $\mu = \mathrm{E}(X)$ and a finite non-zero variance, $\sigma^2 = \mathrm{Var}(X)$, then

$$\mathrm{P}\left(|X - \mu| \geq \varphi\right) \leq \frac{\mathrm{Var}(X)}{\varphi^2}, \ \forall \varphi > 0 .$$

Theorem 3.2:

An unbiased estimator of θ, $\hat{\theta}_n$ is a consistent estimator of θ if $\lim_{n\to\infty} \mathrm{Var}\left(\hat{\theta}_n\right) = 0$.

From Result 2, we have $\mathrm{E}(\hat{\alpha}) = \alpha$ and $\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$. By adopting Theorem 3.1, we can see that,

$$\mathrm{P}\left(|\hat{\alpha} - \alpha| \geq \varphi\right) \leq \frac{\mathrm{Var}(\hat{\alpha})}{\varphi^2}, \forall \varphi > 0 . \qquad (3.22)$$

Without the loss of generality, the equality inside the probability in Equation (3.22) is being removed and hence compared with Definition 3.1, then,

$$\lim_{n\to\infty} \mathrm{P}\left(|\hat{\alpha} - \alpha| > \varphi\right) \leq \lim_{n\to\infty} \frac{\mathrm{Var}(\hat{\alpha})}{\varphi^2} = 0 \quad \Rightarrow \quad \lim_{n\to\infty} \mathrm{Var}(\hat{\alpha}) = 0, \quad \forall \varphi > 0 .$$

To prove that $\hat{\alpha}$ is a consistent estimator of α, we have to show that the variance of $\hat{\alpha}$ approaches zero as n approaches infinity. From Equation (3.19),

$$\lim_{n\to\infty} \hat{\mathrm{Var}}(\hat{\alpha}) = \lim_{n\to\infty} \lambda \hat{\sigma}^2 \left[n - \left(\sum_{i=1}^{n} \hat{\boldsymbol{X}}_i \right) \left(\sum_{i=1}^{n} \hat{\boldsymbol{X}}_i' \hat{\boldsymbol{X}}_i \right) \left(\sum_{i=1}^{n} \hat{\boldsymbol{X}}_i' \right) \right]^{-1} = \lambda(0)(0) = 0 ,$$

where

$$\lim_{n\to\infty}\hat{\sigma}^2 = \lim_{n\to\infty}\frac{1}{(p+1)n}\sum_{i=1}^{n}\left(y_i-\hat{\alpha}-\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)^2\left\{\hat{\boldsymbol{\beta}}'\left(\lambda\mathrm{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda\mathrm{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\hat{\boldsymbol{\beta}}\right.$$

$$\left.+\lambda\left[\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda\mathrm{I}+\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\hat{\boldsymbol{\beta}}\right]^2\right\}$$

$$=0.$$

Similarly, for $\hat{\boldsymbol{\beta}}$, we have

$$\lim_{n\to\infty}\hat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)=\lim_{n\to\infty}\lambda\hat{\sigma}^2\left[\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i-\frac{1}{n}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\right]^{-1}=\lambda(0)\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1}=0.$$

When $\lim_{n\to\infty}\hat{\mathrm{Var}}(\hat{\alpha})=\lim_{n\to\infty}\hat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)=0$, we can claim that both $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are consistent estimators of α and $\boldsymbol{\beta}$ respectively.

3.3 Significance Test of Partial Coefficients of M_PULFR Model

Statistical hypothesis testing is adopted to determine whether a particular independent variable has statistically significant to contribute useful information to the dependent variable.

The null and alternative hypotheses are, $\mathrm{H}_0:\beta_k=0$ and $\mathrm{H}_1:\beta_k\neq 0$ respectively, for $k=1,2,\cdots,p$.

The test statistics under the null hypothesis is given by:

$$t_0=\frac{\hat{\beta}_k}{\mathrm{se}\left(\hat{\beta}_k\right)}=\frac{\hat{\beta}_k}{\sqrt{\mathrm{Var}\left(\hat{\beta}_k\right)}}\sim t_{\propto;n-p-1}, \tag{3.23}$$

where $\propto$ is the significance level, n is the sample size, and p is the number of independent variables.

H_0 is rejected if $|t_0|>t_{\frac{\propto}{2};n-p-1}$ or p-value $>\propto$, where p-value $=2\times\left[1-\mathrm{CDF}(t_0)\right]$.

If H_0 is not rejected, this implies that the particular independent variable, x_k

does not contribute significantly in determining the dependent variable. In other words, this independent variable can be eliminated from the model.

3.4 Coefficient of Determination of M_PULFR Model

Consider Equations (3.1) and (3.2), we can rewrite the random variable, y_i as

$$y_i = \alpha + \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + (\varepsilon_i - \boldsymbol{\delta}_i\boldsymbol{\beta}) = \alpha + \mathbf{x}_i\boldsymbol{\beta} + E_i, \tag{3.24}$$

where the error of the model, $E_i = \varepsilon_i - \boldsymbol{\delta}_i\boldsymbol{\beta} = y_i - \alpha - \mathbf{x}_i\boldsymbol{\beta}, i = 1,2,3,\cdots,n$.

Given that $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimators of α and $\boldsymbol{\beta}$ respectively, by using the idea of least square estimation, the residual of the model, $E_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \mathbf{x}_i\hat{\boldsymbol{\beta}}, i = 1,2,\cdots,n$.

The coefficient of determination can be defined as

$$R^2 = \frac{SSR}{S_{yy}} = 1 - \frac{SSE}{S_{yy}}, \tag{3.25}$$

where $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, and the residual sum of squares (see Result 1) as

$$SSE = \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right)^2 \left\{\hat{\boldsymbol{\beta}}'\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\hat{\boldsymbol{\beta}} + \lambda\left[\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\hat{\boldsymbol{\beta}}\right]^2\right\},$$

$$SSE = \left\{\hat{\boldsymbol{\beta}}'\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\hat{\boldsymbol{\beta}} + \lambda\left[\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\hat{\boldsymbol{\beta}}\right]^2\right\}\sum_{i=1}^{n}E_i^2. \tag{3.26}$$

3.5 Confidence and Prediction Intervals of M$_P$ULFR Model

Recall that $\mathrm{E}(y_i)=Y_i$ and $Y_i=\alpha+\boldsymbol{X}_i\boldsymbol{\beta}$, $i=1,2,\cdots,n$. Let y_{rk} be unknown dependent variable with a set of independent variables, $\boldsymbol{x}_{rk}$, then y_{rk} can be estimated by $\hat{\mathrm{E}}(y_{rk})=\hat{Y}_{rk}$ where $\hat{Y}_{rk}=\hat{\alpha}+\boldsymbol{X}_{rk}\hat{\boldsymbol{\beta}}$ and hence, the variance of $\hat{\mathrm{E}}(y_{rk})$ can be determined as,

$$\begin{aligned}\mathrm{Var}\left[\hat{\mathrm{E}}(y_{rk})\right] &= \mathrm{Var}\left(\hat{\alpha}+\boldsymbol{X}_{rk}\hat{\boldsymbol{\beta}}\right)=\mathrm{Var}(\hat{\alpha})+\mathrm{Var}\left(\boldsymbol{X}_{rk}\hat{\boldsymbol{\beta}}\right)-2\,\mathrm{Cov}\left(\hat{\alpha},\boldsymbol{X}_{rk}\hat{\boldsymbol{\beta}}\right)\\ &= \mathrm{Var}(\hat{\alpha})+\boldsymbol{X}_{rk}\,\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{X}'_{rk}-2\boldsymbol{X}_{rk}\,\mathrm{Cov}\left(\hat{\alpha},\hat{\boldsymbol{\beta}}\right)\\ &= \lambda\hat{\sigma}^2\left\{\left[1+2\boldsymbol{X}_{rk}\left(\sum_{i=1}^{n}\hat{\mathbf{X}}'_i\hat{\mathbf{X}}_i\right)^{-1}\sum_{i=1}^{n}\hat{\mathbf{X}}'_i\right]\left[n-\sum_{i=1}^{n}\hat{\mathbf{X}}_i\left(\sum_{i=1}^{n}\hat{\mathbf{X}}'_i\hat{\mathbf{X}}_i\right)^{-1}\sum_{i=1}^{n}\hat{\mathbf{X}}'_i\right]^{-1}\right.\\ &\quad\left.+\boldsymbol{X}_{rk}\left[\sum_{i=1}^{n}\hat{\mathbf{X}}'_i\hat{\mathbf{X}}_i-\frac{1}{n}\left(\sum_{i=1}^{n}\hat{\mathbf{X}}'_i\right)\left(\sum_{i=1}^{n}\hat{\mathbf{X}}_i\right)\right]^{-1}\boldsymbol{X}'_{rk}\right\}. \qquad (3.27)\end{aligned}$$

The $(1-\propto)100\%$ confidence interval (CI) for the mean response, $\mathrm{E}(y_{rk})$ at a specified value of $\boldsymbol{x}_{rk}$ is,

$$\hat{\mathrm{E}}(y_{rk})-t_{\frac{\propto}{2};n-p-1}\,\mathrm{se}\left[\hat{\mathrm{E}}(y_{rk})\right]\le\mathrm{E}(y_{rk})\le\hat{\mathrm{E}}(y_{rk})+t_{\frac{\propto}{2};n-p-1}\,\mathrm{se}\left[\hat{\mathrm{E}}(y_{rk})\right], \qquad (3.28)$$

where $\mathrm{se}\left[\hat{\mathrm{E}}(y_{rk})\right]=\sqrt{\mathrm{Var}\left[\hat{\mathrm{E}}(y_{rk})\right]}$, n is the sample size, and p is the number of independent variables.

For the prediction interval (PI), since additional uncertainty is attributed to the random error, ε_{rk} when predicting y_{rk} , thus the PI for Y_{rk} is wider as compared to CI for $\mathrm{E}(y_{rk})$ and hence we have,

$$\mathrm{Var}\left(\hat{Y}_{rk}\right)=\lambda\hat{\sigma}^2\left\{1+\left[\left(1+2\boldsymbol{X}_{rk}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right)\left(n-\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right)^{-1}\right.\right.$$

$$\left.\left.+\boldsymbol{X}_{rk}\left[\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i-\frac{1}{n}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\right]^{-1}\boldsymbol{X}_{rk}'\right]\right\}$$

$$=\lambda\hat{\sigma}^2+\mathrm{Var}\left[\hat{\mathrm{E}}\left(y_{uk}\right)\right]. \tag{3.29}$$

Therefore, the $(1-\propto)100\%$ PI for Y_{rk} at a specified value of $\boldsymbol{x}_{rk}$ is given by,

$$\hat{Y}_{rk}-t_{\frac{\propto}{2};n-p-1}\,\mathrm{se}\left(\hat{Y}_{rk}\right)\le Y_{rk}\le\hat{Y}_{rk}+t_{\frac{\propto}{2};n-p-1}\,\mathrm{se}\left(\hat{Y}_{rk}\right), \tag{3.30}$$

where $\mathrm{se}\left(\hat{Y}_{uk}\right)=\sqrt{\lambda\hat{\sigma}^2+\mathrm{Var}\left[\hat{\mathrm{E}}\left(y_{uk}\right)\right]}$, n is the sample size, and p is the number of independent variables.

CHAPTER 4

STATISTICAL ANALYSIS OF THE HOUSING MARKET BEHAVIOUR IN PETALING DISTRICT

In this chapter, the proposed multiple un-replicated linear functional relationship (M_PULFR) model is used to study the Malaysian housing market in Petaling District from micro-perspective. The estimated parameters from M_PULFR model and multiple regression (MR) models were compared while the independent variables (attributes) that significantly contributed to the dependent variable (housing price) were identified and some justifications from the previous studies were also given. Besides, the performances of M_PULFR and MR models in predicting the housing prices from the testing sample set were investigated using mean square error (MSE). All statistical hypotheses were conducted at 0.05 level of significance.

4.1 Data Collections and Descriptions

This study focuses on the resale market (secondary market) of terrace houses in Petaling District, one of the most populated regions in Malaysia. Petaling District is located in Selangor state and adjacent to Kuala Lumpur (see Figure 4.1). It is divided into six sub-regions under three administrative zones. Shah Alam and Sungai Buloh are under the governance of Shah Alam City Council; Subang Jaya, Puchong, and Seri Kembangan are under the

governance of Subang Jaya City Council; and Petaling Jaya is under the governance of Petaling Jaya City Council (Wikipedia, 2017a).

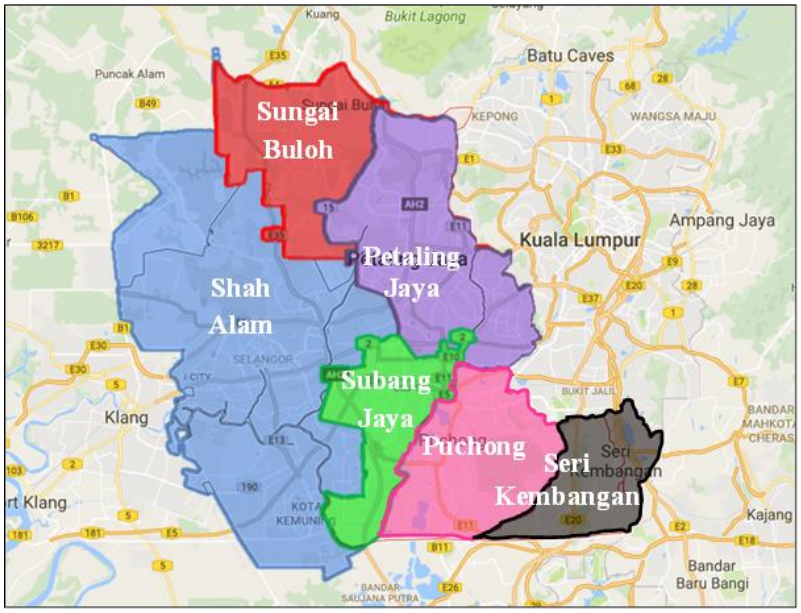

Figure 4.1: Administrative Map of Petaling District

The main data source was collected from Jordan Lee & Jaafar (S) Pte Ltd, a registered valuer and real estate agent based in Klang Valley, and it consists of 44331 terrace house transacted records from November 2008 to February 2016. A total of 1167 (2.63%) duplicated records and 1414 (3.19%) records with missing values or non-rectifiable errors were removed. Hence, the remaining 41750 records were used and grouped according to the six sub-regions based on the address given where 9643 cases from Shah Alam, 9341 cases from Puchong, 8741 cases from Petaling Jaya, 7956 cases from Subang Jaya, 5477 cases from Seri Kembangan, and 592 cases from Sungai Buloh.

The relevant information extracted from the data are transacted house prices, lot sizes, tenure types (freehold or leasehold), expiry dates of lease

terms, terrace types, numbers of bedrooms, main building sizes, and transaction date. Additional information from Google Maps such as distance to the nearest shopping mall and distance to the nearest supermarket were also included based on the property's address. Table 4.1 summarised the variables used in this study.

Table 4.1: Descriptions of Variables Used

Variable	Abbreviation	Description
y	Price	Individual housing price (RM'000)
x_1	Lot size	Lot size (m^2)
x_2	Tenure	Tenure type (this study used "0" to represent freehold and "1" to represent leasehold)
x_3	Expiry	Years to expiry of lease term (assuming 200 years for freehold)
x_4	Terrace	Terrace type (total number of floors, example: single, 1.5, double, and et cetera)
x_5	Bedroom	Number of bedrooms
x_6	Built up	Main building size (m^2)
x_7	Mall	Distance to nearest shopping mall (km)
x_8	Market	Distance to nearest supermarket (km)
x_9	Time	Transaction date (in month) to differentiate the transaction time of repeat-sales houses

According to United States of America before Federal Trade Commission (1998), a supermarket is defined as a full-line retail grocery store that supplies a wide variety of foods and grocery items which include bread and dairy products, chilled and frozen food, beverages, fresh and processed meats and poultry, fresh fruits and vegetables, canned products, the staple food, and some non-food products such as detergents, toiletries, paper goods, other household items, and health and beauty aids. There are a few well-known

supermarkets in Petaling District such as Giant Supermarket, Tesco, and NSK Supermarket.

On the other hand, according to the International Council of Shopping Centres (2015), an Asia-Pacific shopping mall is defined as a group of retail and other commercial establishments (bars, restaurants, private offices, and et cetera) which is planned, developed and managed intentionally as a single property which comprised of commercial multi-branded rental units and common areas. In general, the net leasable area (NLA) of a shopping mall is capped at 20,000 square feet but this excluded areas whose primary purpose is not retail under the context of a mixed-use project, for example, adjoined office and hotel lobby. In Petaling District, the well-known shopping malls include One-Utama, Sunway Pyramid, The Curve and IOI Mall.

The transacted housing prices are regressed on the above-mentioned independent variables using M_PULFR and MR models. In this study, seven sets of models have been developed, where six sets were used to study the terrace housing market for each sub-region and the remaining one set was used to study the overall performance of the terrace housing market in Petaling District. For each model, the data were randomly divided into 70% training set and 30% testing set. The training set was used to train the model, and the testing set was used to validate the performance of the trained model. All values of the independent variables were normalised before constructing the models using the following formula:

$$\frac{x_{ij} - \min\left(x_j\right)}{\max\left(x_j\right) - \min\left(x_j\right)}.$$

4.2 Comparisons of the Results Obtained from Training Samples

In this study, upon model training, insignificant attributes determined by MR model in the first run were removed and the second run was then performed in order to give a 'two peas in a pod' comparison.

4.2.1 Shah Alam

From Table 4.2, M_PULFR model shows that all housing attributes are significant determinants of the housing prices in Shah Alam while tenure type is not significant (*p*-values = 0.0533) by MR model. M_PULFR model produces positive relationships between housing prices and all housing attributes, except for tenure type where 0 indicates freehold and 1 indicates leasehold property.

Both M_PULFR and MR models show that lot size, main building size, and time to expiry to the lease term have a positive impact on housing prices. Besides longer tenure, buyers in Shah Alam are willing to pay higher for a larger lot size and main building size which are consistent with the studies from Pashardes and Savva (2009), and Owusu-ansah (2012).

In the study of Ooi et al. (2014), freehold housings are preferable compared to leasehold housings. This finding is further supported by M_PULFR model, but MR model shows a positive relationship between housing prices and leasehold housings. The discrepancy between MR model and the housing market behaviour may be caused by the existence of multicollinearity which resulted from a high negative correlation (–0.995, see Table 4.3) between

tenure type and time to expiry of lease term in MR model that affected the estimation of the coefficients of the model (Matignon, 2007).

Table 4.2: Estimated Parameters and Performance Measures for the Housing Study of Shah Alam

	First Run				Second Run			
	M_PULFR		**MR**		**M_PULFR**		**MR**	
Attribute	beta	*p*-value	beta	*p*-value	beta	*p*-value	beta	*p*-value
Constant	-13120.67	-	-230.78	1.6E-06	-13798.40	-	-139.11	8.06E-72
Lot size	24775.06	0.0000	4242.50	0.0000	24775.06	0.0000	4233.28	0.0000
Tenure	-1487.71	0.0000	76.28	0.0533*	-	-	-	-
Expiry	1829.54	0.0000	235.85	3.02E-06	1829.54	0.0000	138.71	1.6E-190
Terrace	10727.04	0.0000	-113.90	9.15E-08	10727.04	0.0000	-109.30	2.48E-07
Bedroom	4746.85	0.0000	-69.35	5.51E-06	4746.85	0.0000	-63.92	2.03E-05
Built up	3007.68	0.0000	1395.17	0.0000	3007.68	0.0000	1396.89	0.0000
Mall	9972.66	0.0000	-174.10	2.84E-66	9972.66	0.0000	-168.55	1.36E-67
Market	7397.08	0.0000	63.79	2.12E-14	7397.08	0.0000	66.27	9.38E-16
Time	2675.77	0.0000	373.72	0.0000	2675.77	0.0000	371.21	0.0000
MSE	1.38E-05		18058.46		5.76E-06		18068.47	
R^2	0.9999998		0.8184		0.9999998		0.8183	

*insignificant at significance level of 0.05

From Table 4.2, MR model shows that terrace type and the number of bedrooms are negatively related to housing prices which are different from M_PULFR model. These outputs from MR model contradict with the housing market behaviour where the more level of floors and bedrooms will result in higher prices.

Table 4.3: Correlation between Housing Attributes

Attribute	Lot size	Tenure	Expiry	Terrace	Bedroom	Built up	Mall	Market	Time
Lot size	1.000								
Tenure	-0.162	1.000							
Expiry	0.167	**-0.995**	1.000						
Terrace	0.058	-0.151	0.175	1.000					
Bedroom	0.240	-0.281	0.309	0.383	1.000				
Built up	0.456	-0.359	0.384	0.504	0.596	1.000			
Mall	0.149	-0.201	0.236	0.139	0.192	0.302	1.000		
Market	-0.002	-0.112	0.138	0.166	0.202	0.209	0.334	1.000	
Time	0.020	-0.043	0.031	0.061	0.127	0.112	0.018	0.147	1.000

An interesting observation from Shah Alam's housing market is that M_PULFR produces a positive relationship between housing prices and the distance to the nearest shopping mall and supermarket. This implies that house buyers from Shah Alam prefer to stay further away from shopping mall and supermarket. On the other hand, MR model shows that the distance to the nearest supermarket is positively related to housing prices while the distance to the nearest shopping mall has a negative impact. Although the effect of nearby amenities on housing prices is ambiguity (Rosiers et al., 1996), it is believed that M_PULFR model can better explain the relationship between housing prices and the distance to the nearest shopping mall and supermarket. This unusual housing market behaviour in Shah Alam is partly caused by town planning and residents' race, and will be revisited and discussed in Section 4.3.

From Table 4.2, both M_PULFR and MR models show that lot size is the predominating factor in affecting housing prices in Shah Alam as it contributes the highest weight to the housing prices in Shah Alam.

It is also seen in Table 4.2, the proposed M_PULFR model has a better fitting ability as compared to MR model. For the training sample, the MSE for M_PULFR model for the first and second runs are 1.38E-05 and 5.76E-06 respectively while the R^2 for both runs are close to 1.0. In comparison with M_PULFR model, the MSE for MR model in the first and second runs are 18058.46 and 18068.47 respectively while the R^2 for the first and second runs are close to 0.82.

It can be observed that the estimated coefficients (betas) of M_PULFR model are the same for both runs even though one of the attributes, tenure type, is being removed. This indicates that the estimation of individual beta does not depend on other attributes other than its respective attribute (this can also be seen in Equation (3.12)). Therefore, multicollinearity which resulted from the correlation between independent variables gives no influence on the estimation of the coefficients of M_PULFR model.

4.2.2 Puchong, Subang Jaya and Sungai Buloh

M_PULFR model resulted in the same set of significance housing attributes and type of relationship with housing prices for Puchong (see Table 4.4), Subang Jaya (see Table 4.5) and Sungai Buloh (see Table 4.6). These outputs are the same as Shah Alam except for the distance to the nearest

shopping mall. In Shah Alam, the housing prices have a positive relationship with the distance to shopping mall, but it has a negative relationship in the regions of Puchong, Subang Jaya and Sungai Buloh. This indicates that house buyers from these areas are more willing to invest in houses that are nearer to shopping mall but trying to stay away from a supermarket. It is believed that the repulsion effect of a supermarket in these areas may due to the problems of traffic congestions and noise or air pollution as mentioned by Tse and Love (2000) in a different study.

Table 4.4: Estimated Parameters and Performance Measures for the Housing Study of Puchong

	First Run				Second Run			
	M_pULFR		**MR**		**M_pULFR**		**MR**	
Attribute	beta	*p*-value	beta	*p*-value	beta	*p*-value	beta	*p*-value
Constant	-4127.02	-	-285.74	4.11E-06	-2354.56	-	-287.34	3.63E-06
Lot size	7457.99	0.0000	1075.49	0.0000	7457.99	0.0000	1075.37	0.0000
Tenure	-4291.02	0.0000	128.48	9.24E-03	-4291.02	0.0000	127.28	9.92E-03
Expiry	4892.61	0.0000	282.37	9.57E-06	4892.61	0.0000	280.24	1.11E-05
Terrace	8773.64	0.0000	-80.70	7.18E-04	8773.64	0.0000	-89.28	1.29E-04
Bedroom	4274.03	0.0000	-20.94	0.0904*	-	-	-	-
Built up	4179.32	0.0000	1678.66	0.0000	4179.32	0.0000	1662.33	3.45E-83
Mall	-18731.48	0.0000	-120.26	1.02E-80	-18731.48	0.0000	-121.44	2.72E-04
Market	12261.81	0.0000	-30.05	2.38E-04	12261.81	0.0000	-29.77	0.0000
Time	2177.06	0.0000	454.05	0.0000	2177.06	0.0000	453.99	3.63E-06
MSE	8.06E-08		16206.46		2.77E-08		16213.58	
R^2	0.9999993		0.7721		0.9999993		0.7720	

*insignificant at significance level of 0.05

Based on M_PULFR model, the distance to the nearest shopping mall is the most influential determinant of housing prices in Puchong. A similar remark was also noted in Kam et al. (2016). However, the most influential determinant in Subang Jaya is the terrace type, and Sungai Buloh is the distance to the nearest supermarket.

Table 4.5: Estimated Parameters and Performance Measures for the Housing Study of Subang Jaya

	First Run				Second Run			
	M_PULFR		**MR**		**M_PULFR**		**MR**	
Attribute	beta	*p*-value	beta	*p*-value	beta	*p*-value	beta	*p*-value
Constant	-18373.48	-	-391.00	0.0406	-7507.76	-	-125.317	2.6E-44
Lot size	11579.74	0.0000	1886.40	0.0000	11579.74	0.0000	1872.803	0.0000
Tenure	-11245.92	0.0000	110.24	0.5309*	-	-	-	-
Expiry	12352.46	0.0000	264.15	0.1664*	-	-	-	-
Terrace	18214.99	0.0000	338.86	4.57E-25	18214.99	0.0000	263.1489	4.73E-15
Bedroom	9869.07	0.0000	75.43	4.98E-05	9869.07	0.0000	61.0253	1.44E-03
Built up	3610.81	0.0000	1004.98	0.0000	3610.81	0.0000	1037.45	0.0000
Mall	-14776.55	0.0000	-168.48	3.02E-50	-14776.55	0.0000	-146.423	1.08E-36
Market	13421.37	0.0000	94.46	1.52E-13	13421.37	0.0000	87.72058	2.55E-11
Time	1792.68	0.0000	563.98	0.0000	1792.68	0.0000	560.573	0.0000
MSE	9.14E-05		14913.70		3.50E-06		15967.41	
R^2	0.9999996		0.8121		0.9999999		0.7988	

*insignificant at significance level of 0.05

Table 4.6: Estimated Parameters and Performance Measures for the Housing Study of Sungai Buloh

	First Run				Second Run			
	M_PULFR		**MR**		**M_PULFR**		**MR**	
Attribute	beta	*p*-value	beta	*p*-value	beta	*p*-value	beta	*p*-value
Constant	-3249.39	-	98.47	0.4902	-2714.12	-	-14.86	0.1926
Lot size	3403.91	0.0000	462.04	3.1E-21	3403.91	0.0000	424.12	6.75E-18
Tenure	-755.68	0.0000	-64.23	0.5260*	-	-	-	-
Expiry	968.11	0.0000	40.50	0.7417*	-	-	-	-
Terrace	2292.34	0.0000	-46.05	0.0949*	-	-	-	-
Bedroom	3629.91	0.0000	0.64	0.9823*	-	-	-	-
Built up	1306.05	0.0000	685.58	2.77E-46	1306.05	0.0000	856.52	1.9E-109
Mall	-2916.83	0.0000	-85.04	0.0706*	-	-	-	-
Market	5844.27	0.0000	-113.54	7.91E-03	5844.27	0.0000	-88.50	0.0190
Time	4311.73	0.0000	305.90	1.24E-54	4311.73	0.0000	289.58	3.22E-49
MSE	6.82E-08		8819.52		8.99E-10		9739.49	
R^2	0.9999988		0.8483		0.9999994		0.8324	

*insignificant at significance level of 0.05

The results obtained from MR model are not consistent in these three areas. In Puchong, the MR model shows that the number of bedrooms (*p*-value = 0.0904) is not a significant determinant, while terrace type and the distance to the nearest shopping mall or supermarket have negative relationships with the housing prices. All other housing attributes have positive relationships in Puchong. In contrast, MR model produces insignificant results for terrace type (*p*-value = 0.5309) and years to the expiry of the lease term (*p*-value = 0.1664) in Subang Jaya. The distance to the nearest shopping mall is the only attribute

that has a significant negative relationship with housing prices in Subang Jaya, while all other attributes have positive relationships.

In Sungai Buloh, MR model produces insignificant attributes, that are tenure type (*p*-value = 0.5260), year to expiry of the lease term (*p*-value = 0.7417), terrace type (*p*-value = 0.0949), number of bedrooms (*p*-value = 0.9823) and the distance to the nearest shopping mall (*p*-value = 0.0706). Both lot size and main building size have significant positive relationships and the distance to the nearest supermarket has significant negative relationship with the housing prices in Sungai Buloh.

Again, M_PULFR model shows a better fitting ability compared to MR model in Puchong, Subang Jaya and Sungai Buloh with a smaller MSE and higher R^2.

4.2.3 Petaling Jaya

Table 4.7 shows that all housing attributes are significant determinants of the housing prices in Petaling Jaya in both models. The M_PULFR model remains outperformed MR model in terms of fitting ability as shown by MSE and R^2 values.

M_PULFR model shows the same outcomes as obtained under Puchong, Subang Jaya and Sungai Buloh except for the distance to the nearest supermarket. The distance to the nearest supermarket has a negative impact on housing prices implies that buyers in Petaling Jaya preferred houses that

provide good accessibility and convenience which are nearby supermarket and shopping mall. This is different from other sub-regions where the house buyers only in favour with a shorter distance to shopping mall. A possible explanation for this phenomenon is that most of the supermarkets in Petaling Jaya are operating in the shopping mall, and hence there is no difference for the residents to go to shopping mall or supermarket.

Table 4.7: Estimated Parameters and Performance Measures for the Housing Study of Petaling Jaya

	M_PULFR		**MR**	
Attribute	beta	*p*-value	beta	*p*-value
Constant	-3201.36	-	-417.13	5.45E-13
Lot size	9264.85	0.0000	1037.80	6.5E-241
Tenure	-2094.88	0.0000	143.76	2.08E-03
Expiry	2621.80	0.0000	340.39	2.2E-08
Terrace	7739.19	0.0000	435.73	4.2E-38
Bedroom	8216.49	0.0000	50.76	0.0417
Built up	6637.34	0.0000	1811.63	2.4E-272
Mall	-9766.77	0.0000	-285.56	1.49E-78
Market	-14091.05	0.0000	-112.54	3.73E-13
Time	3365.03	0.0000	576.33	0.0000
MSE	1.84E-07		40856.55	
R^2	0.9999997		0.7171	

*insignificant at significance level of 0.05

In the study of Ooi et al. (2014), freehold housings are preferable compared to leasehold housings. This finding is further supported by M_PULFR model, but MR model shows a positive relationship between housing prices

and leasehold housings. The shorter distance to the nearest supermarket is the most influential determinant of housing prices in Petaling Jaya using M_PULFR model.

4.2.4 Seri Kembangan

According to M_PULFR model, house buyers from Seri Kembangan are in favour of houses that are nearer to the shopping mall, but not the supermarket (see Table 4.8). This observation is consistent with those from Puchong, Subang Jaya and Sungai Buloh. MR model produces the same results for the distance to the nearest shopping mall and supermarket. The distance to the nearer shopping mall is the most important determinant of housing prices in Seri Kembangan for M_PULFR model but the predominating factor in MR model is the lot size.

The main differences between house buyers from Seri Kembangan and other sub-regions are the tenure type and time to expiry of the lease term. M_PULFR model shows that all other sub-regions have a positive relationship between housing prices and year to the expiry of the lease term. However, it is observed that this attribute has a negative relationship in Seri Kembangan, which in turn yields a preference of shorter year to expiry of the lease term.

Table 4.8: Estimated Parameters and Performance Measures for the Housing Study of Seri Kembangan

	First Run				Second Run			
	M_PULFR		**MR**		**M_PULFR**		**MR**	
Attribute	beta	*p*-value	beta	*p*-value	beta	*p*-value	beta	*p*-value
Constant	23068.01	-	-373.75	2.39E-30	25063.68	-	-376.31	5.08E-31
Lot size	10511.04	0.0000	884.09	2E-140	10511.04	0.0000	883.118	2.5E-140
Tenure	8575.89	0.0000	288.68	3.09E-23	8575.89	0.0000	287.5743	4.01E-23
Expiry	-25149.91	0.0000	349.01	9.13E-24	-25149.91	0.0000	347.5682	1.2E-23
Terrace	2434.93	0.0000	-56.49	1.1E-08	2434.93	0.0000	-56.6737	9.76E-09
Bedroom	3989.68	0.0000	-10.33	0.4253*	-	-	-	-
Built up	1998.76	0.0000	661.59	0.0000	1998.76	0.0000	656.2413	0.0000
Mall	-54273.64	0.0000	-26.91	1.66E-04	-54273.64	0.0000	-27.0161	1.55E-04
Market	5545.67	0.0000	62.48	2.77E-25	5545.67	0.0000	63.64464	1.06E-27
Time	1020.14	0.0000	311.30	0.0000	1020.14	0.0000	311.3766	0.0000
MSE	7.34E-04		6542.79		2.73E-04		6543.87	
R^2	0.9999951		0.7260		0.9999950		0.7260	

*insignificant at significance level of 0.05

These observations contradict with Ooi et al. (2014) and indicate that the house buyers in Seri Kembangan are willing to pay higher for leasehold housings. The reason for the differences arisen will be discussed in Section 4.3.

Again, the results from MR model failed to reflect the dynamic behaviour of the housing market in Seri Kembangan, possibly due to the multicollinearity problems of the parameters' estimation. The MR model only explained 72.6% of the total variability of the housing prices in Seri Kembangan, while M_PULFR model recorded a R^2 close to one.

4.2.5 Petaling District

The previous sub-sections identified significant housing attributes for each sub-region within Petaling District. It is also interesting to investigate if the significant housing attributes changed at the district level. The estimated parameters for M_PULFR and MR models for the whole Petaling District is presented in Table 4.9.

Table 4.9: Estimated Parameters and Performance Measures for the Housing Study of Petaling District

	First Run				Second Run			
	M_PULFR		**MR**		**M_PULFR**		**MR**	
Attribute	beta	*p*-value	beta	*p*-value	beta	*p*-value	beta	*p*-value
Constant	10660.79	-	-363.46	1.85E-36	13077.15	-	-357.22	2.78E-36
Lot size	32433.17	0.0000	4107.73	0.0000	32433.17	0.0000	4123.04	0.0000
Tenure	-2510.63	0.0000	167.15	3.67E-14	-2510.63	0.0000	160.86	7.58E-14
Expiry	3232.74	0.0000	374.33	2.86E-36	3232.74	0.0000	365.70	1.65E-36
Terrace	10711.51	0.0000	-16.66	0.2036*	-	-	-	-
Bedroom	9968.89	0.0000	-119.98	1.18E-29	9968.89	0.0000	-121.03	2.56E-30
Built up	5077.58	0.0000	1897.75	0.0000	5077.58	0.0000	1888.94	0.0000
Mall	-11291.13	0.0000	-249.59	0.0000	-11291.13	0.0000	-249.14	0.0000
Market	-74566.31	0.0000	-93.11	1.13E-60	-74566.31	0.0000	-93.43	3.38E-61
Time	2752.30	0.0000	460.77	0.0000	2752.30	0.0000	460.77	0.0000
MSE	1.94E-05		27102.99		5.01E-04		27104.49	
R^2	0.9999997		0.7268		0.9999997		0.7268	

*insignificant at significance level of 0.05

M_PULFR model remains outperformed the MR model with much smaller MSE and R^2 close to 1.0. M_PULRF is also more consistent where it

produces the same set of significant housing attributes in the first and second runs for each sub-region and at the district level. In general, house buyers in Petaling District are more likely to spend more in exchange for housings with the larger lot and building sizes, freehold tenure, newer houses, more floors, more bedrooms, and good accessibility to shopping mall and supermarket. On the other hand, the terrace type is not significant under MR model. The model also reveals that the housing prices will be increased if the house is leasehold and has fewer bedrooms. This observation is contradicted by the previous studies in Babawale and Adewunmi (2011), Owusu-ansah (2012), and Pashardes and Savva (2009).

M_PULFR model shows that distance to the nearest supermarket is the most influential determinant of housing prices in Petaling District where a shorter distance will result in a higher price. In contrast, MR model shows that lot size is the predominating factor in affecting housing prices in Petaling District where a larger size will result in a higher price. It is observed that a negative relationship between housing prices and distance to the nearest supermarket in Petaling District. This is in line with the outcome from Petaling Jaya. In fact, the housing prices in Petaling Jaya are generally higher as compared to other sub-regions. In the year 2015 for example, an average transacted housing price in Petaling Jaya is about RM 850,000 but the average range of transacted housing price in other sub-regions is about RM 390,000 to RM 780,000, and about RM 500,000 for the entire Petaling District. So, it is believed that the effect of this attribute on housing prices in Petaling Jaya giving a dominating impact in the model of Petaling District.

4.3 Discussion of the Housing Market Behaviour in Petaling District

Table 4.10 shows that there are some differences between the housing market behaviour in Petaling District and its sub-regions. The housing market behaviours in Petaling District can be categorised into four types; that are

i. Type I: House buyers in Shah Alam preferred houses that are far away from shopping mall and supermarket.
ii. Type II: House buyers in Puchong, Subang and Sungai Buloh preferred houses that nearby shopping mall but not the supermarket.
iii. Type III: House buyers in Seri Kembangan preferred houses that nearby shopping mall but not the supermarket. Besides, they also preferred leasehold houses.
iv. Type IV: House buyers in Petaling Jaya (and the whole Petaling District) preferred houses that are nearby both the shopping mall and supermarket.

Table 4.10: Estimated Parameters ($\hat{\alpha}$ and $\hat{\beta}$) for M_PULFR Model for Petaling District and Its Six Sub-Regions

Attribute	Petaling District	Shah Alam	Puchong	Petaling Jaya	Subang Jaya	Seri Kembangan	Sungai Buloh
Constant	10660.79	-13120.67	-4127.02	-3201.36	-18373.48	23068.01	-3249.39
Lot size	32433.17	24775.06	7457.99	9264.85	11579.74	10511.04	3403.91
Tenure	-2510.63	-1487.71	-4291.02	-2094.88	-11245.92	8575.89*	-755.68
Expiry	3232.74	1829.54	4892.61	2621.80	12352.46	-25149.91*	968.11
Terrace	10711.51	10727.04	8773.64	7739.19	18214.99	2434.93	2292.34
Bedroom	9968.89	4746.85	4274.03	8216.49	9869.07	3989.68	3629.91
Built up	5077.58	3007.68	4179.32	6637.34	3610.81	1998.76	1306.05
Mall	-11291.13	9972.66**	-18731.48	-9766.77	-14776.55	-54273.64	-2916.83
Market	-74566.31	7397.08**	12261.81**	-14091.05	13421.37**	5545.67**	5844.27**
Time	2752.30	2675.77	2177.06	3365.03	1792.68	1020.14	4311.73

*deviated from previous studies and different from Petaling District **different from Petaling District

The following sub-sections will provide some justifications on the housing market behaviours in Shah Alam and Seri Kembangan.

4.3.1 Housing Market Behaviour in Shah Alam

In Section 4.2.1, it was highlighted that house buyers in Shah Alam are less likely to spend more money to buy housings that are nearby shopping mall and supermarket. This may due to the earlier development prior to the year 2000 in Shah Alam, as its development was solely based on the unique identity of a Malay city with no entertainment outlets (City Declaration by Shah Alam City Council as cited in Wikipedia, 2017b).

In order to investigate the “unique identity of a Malay city” in Shah Alam, further analysis has been done on the data set from Shah Alam. The data set was subdivided into two groups, a group with cases where the majority of house buyers or residents in a particular residential garden (usually called “Taman” or section in Malaysia) are Malay and the other with cases where the majority are non-Malay (refer to Table 4.11).

Table 4.11: Population Distribution by Race at Different Residential Gardens in Shah Alam

Area	Malay	Non-Malay	Total	% of Malay	Area	Malay	Non-Malay	Total	% of Malay
Alam Budiman	79	24	103	77%	Seksyen 26	1	3	4	25%
Bandar Nusa Rhu	78	21	99	79%	Seksyen 27	257	151	408	63%
Bdr Pinggiran Subang	29	117	146	20%	Seksyen 28	109	59	168	65%
Bukit Bandaraya U11	96	5	101	95%	Seksyen 3	3	0	3	100%
Bukit Jelutong	575	505	1080	53%	Seksyen 33	6	3	9	67%
Cahaya Alam	107	4	111	96%	Seksyen 4	78	11	89	88%
Damai	5	95	100	5%	Seksyen 6	38	2	40	95%
Denai Alam	228	321	549	42%	Seksyen 7	376	42	418	90%
Elmina Gardens	3	25	28	11%	Seksyen 8	129	7	136	95%
Impian	64	708	772	8%	Seksyen 9	40	1	41	98%
Indah	21	370	391	5%	Setia Eco-Park	0	7	7	0%
Aman Suria	4	27	31	13%	Subang Bestari	160	41	201	80%
Duta Villa	0	9	9	0%	Subang Pelangi	21	94	115	18%
Subang Permata	14	6	20	70%	Subang Sejahtera	5	3	8	63%
U1/84(C)	11	15	26	42%	Sunway Alam Suria	59	29	88	67%
H'-Glenmarie Ind' Park	8	23	31	26%	Sunway Kayangan	86	61	147	59%
Baru Hicom	0	1	1	0%	Batu Tiga	54	8	62	87%
Bukit Lanchung	2	0	2	100%	Bukit Saga	44	37	81	54%
Padang Jawa	8	0	8	100%	Bukit Sandaran	46	4	50	92%
Perindustrian Temasya	11	151	162	7%	Bukit Subang	125	149	274	46%
Laman Glenmarie	19	59	78	24%	Desa Subang	26	21	47	55%
Mutiara Subang	32	51	83	39%	Ladang Jaya	2	14	16	13%
Alam Nusantara	60	446	506	12%	Mutiara S'bang	14	21	35	40%
Subang Impian	43	23	66	65%	Taman Nusa Subang	33	40	73	45%
Seksyen 10	45	3	48	94%	Paya Jaras Permai	48	10	58	83%
Seksyen 11	82	6	88	93%	Puteri Subang	18	37	55	33%
Seksyen 13	163	12	175	93%	Setia Warisan	39	0	39	100%
Seksyen 16	16	2	18	89%	Sri Buloh	57	25	82	70%
Seksyen 17	188	36	224	84%	Sri Muda	1	1	2	50%
Seksyen 18	242	34	276	88%	Subang Baru	60	25	85	71%
Seksyen 19	313	23	336	93%	Subang Intan	3	28	31	10%
Seksyen 2	42	5	47	89%	Subang Murni	48	51	99	48%
Seksyen 20	155	23	178	87%	Subang Perdana	74	52	126	59%
Seksyen 23	7	4	11	64%	TTDI Jaya	298	64	362	82%
Seksyen 24	173	13	186	93%	Cahaya Heights (SPK)	47	47	94	50%
					Total	**5328**	**4315**	**9643**	**55%**

The grouped models are shown as follows:

Malay Group:

$$Y_i = -290892.84 + 18479.55X_{i1} - 1596.44X_{i2} + 1992.62X_{i3} + 7145.34X_{i4} + 4656.08X_{i5} + 3286.71X_{i6} + 3428.41X_{i7} + 1331807.39X_{i8} + 2785.56X_{i9}$$

and,

Non-Malay Group:

$$Y_i = 5048.58 + 27092.75X_{i1} - 2719.88X_{i2} + 3819.29X_{i3} + 45120.92X_{i4} + 8924.69X_{i5} + 4337.91X_{i6} - 94786.98X_{i7} + 10620.97X_{i8} + 2450.94X_{i9}$$

where $X_{ip} = x_{ip} - \delta_{ip}$, $p = 1, 2, \cdots, 9$.

From the two grouped models, buyers in the areas where the majority of the buyers are Malay would be more likely to spend more money for a house that is far from shopping malls while buyers in the areas where the majority of the buyers are non-Malay would prefer a house that is nearby shopping malls. However, this housing market behaviour does not occur naturally, and this can be explained by considering the geographical position of the houses in Shah Alam (see Figure 4.2). It can be seen from Figure 4.2 that, the structure of the development in the town centre and central area is similar to the concentric zone theory where housing price increases as the distance from corporate area increases (Balchin et al., 1995). The town centre and the central area were developed before other areas, and these areas are mainly populated by the Malays. In addition, most of the shopping malls in Shah Alam were built at the corporate area. It can be seen from Figure 4.2, in the Town Centre and the Central Area, the bubbles (represent housing prices) around black crosses (represent shopping malls) are relatively small (lower housing prices)

compared to those that are farer. In contrast, on the west-side of Shah Alam, known as Setia Alam which is highly populated by non-Malay (refer Table 4.11 for the residential gardens Impian, Indah and Damai), the bubbles that around the cross are relatively big (higher housing prices). These indicate that the housing market behaviour is actually influenced by the structure of the earlier development of Shah Alam which results in a positive relationship between housing prices and distance to the nearest shopping malls in the regions where the majority of the buyers are Malay.

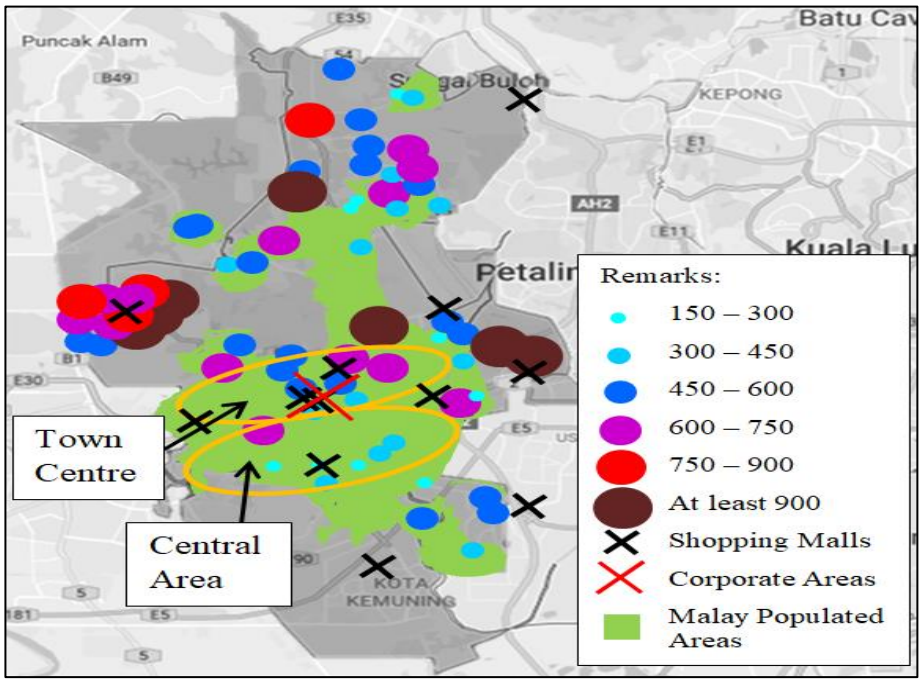

Figure 4.2: Average Housing Prices (RM'000) by Residential Garden in Shah Alam in 2015 (Source: Partially from Unit DEGIS Pejabat Daerah & Tanah Petaling, 2014 and Partially from Google Maps)

4.3.2 Housing Market Behaviour in Seri Kembangan

M_PULFR model shows that the leasehold housings are preferable in Seri Kembangan compared to freehold housings. In order to explain this behaviour, we consider the geographical position of the transacted houses in Seri Kembangan as shown in Figure 4.3. It is observed that leasehold housings are mainly located industrial area or business centre (labelled as ① and ③) as compared to freehold housings (labelled as ②). The spearman correlation between tenure type (freehold or leasehold) and area (residential area or industrial/business centre) is 0.68. Areas ① and ③ provide more work opportunities which contribute an attractive force to the housings that located in these areas. Therefore, the prices of the leasehold housings are higher as compared to freehold housings.

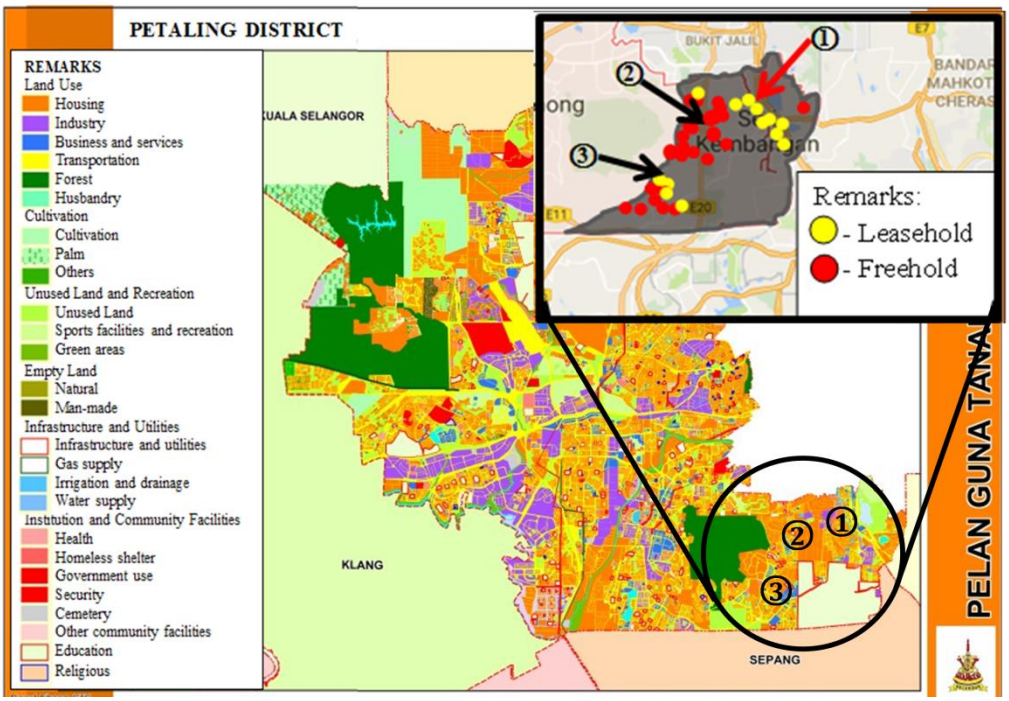

Figure 4.3: Land Use of Petaling District (Source: Unit DEGIS Pejabat Daerah & Tanah Petaling, 2014) and Geographical Position of the Houses in Seri Kembangan (Source: Google Maps)

CHAPTER 5

PREDICTIONS OF HOUSING PRICES IN PETALING DISTRICT

Previous chapter studies and analyses the housing market behaviour in Petaling District using training sample sets and we have seen that the proposed M_PULFR model outperformed MR model in estimating parameters with higher consistency for the training sample. In this chapter, we will evaluate the performances and the robustness of the M_PULFR model by comparing the predicted housing prices in Petaling District using testing sample sets.

5.1 Reference Value upon Prediction for M_PULFR Model

From Result 1 in Chapter Three, it is observed that the estimation of the acquired housing price, Y_a is based on the observed value of the acquired housing price, y_a and its acquired housing attributes $\boldsymbol{x}_a$. However, the price of an acquired house is not known upon prediction. In other words, M_PULFR model requires a reference housing price, $\tilde{y}_m$ (as an estimation of y_a) when predicting the price of an acquired house. Define the Euclidean distance between the acquired housing attributes and the *i*-th transacted house attributes from the same residential area as

$$d_i = \sqrt{(x_{a1} - x_{i1})^2 + (x_{a2} - x_{i2})^2 + \cdots + (x_{ap} - x_{ip})^2}, \quad i = 1, 2, \cdots, n.$$

Thus, the reference housing price is obtained by averaging the transacted (historical) housing prices of *h* nearest houses with the most similar housing attributes (smallest d_i)

$$\tilde{y}_m = \frac{1}{h}\sum_{i=1}^{h} y_{\min_h(d_i)}, \tag{5.1}$$

where $y_{\min_h(d_i)}$ is the price of the houses with *h* smallest d_i values.

5.2 Comparisons of the Results Obtained from Testing Samples

This section evaluates the prediction accuracy of M_PULFR and MR models when applied to testing sample sets using MSE. The attributes used in each model are based on the significance tests done in Chapter Four.

In terms of prediction accuracy, M_PULFR model remains outperformed MR model in all cases as the MSE values produced by M_PULFR model are smaller than those produced by MR model for Petaling District and its sub-regions. From Table 5.1 and Table 5.2, for Petaling District, we can see that M_PULFR model produced an MSE of 19034.6 which is smaller than the MSE produced by MR model which is 27199.7. The same situation goes for all sub-regions. Besides, the percentage of smaller prediction error for M_PULFR model is higher compared to MR model for all regions. About 20% to 27% of the predicted housing prices using M_PULFR model are less than 5% of differences compared to the actual prices, which is a sign of better performance compared to MR models which only cover from about 11% to 22%. This implies that, the housing prices which predicted by M_PULFR model are closer to the actual prices compared to MR model.

Table 5.1: Performance Measures of MR Model for Petaling District and Its Sub-Regions

Error of prediction	Petaling District	Shah Alam	Puchong	Petaling Jaya	Subang Jaya	Seri Kembangan	Sungai Buloh
<5% difference	14.66%	17.49%	15.17%	16.29%	22.20%	18.50%	11.24%
<10% difference	28.09%	33.74%	29.73%	31.08%	42.10%	35.48%	23.60%
<30% difference	67.98%	74.80%	72.66%	70.37%	86.34%	78.51%	64.61%
MSE							
For difference <30%	9795.9	7884.3	6499.9	15483.8	8158.1	3098.3	5228.4
Whole testing sample	27199.7	21956.4	18649.4	38256.4	15348.6	6844.8	21417.8

Table 5.2: Performance Measures of M_PULFR Model for Petaling District and Its Sub-Regions

Error of prediction	Petaling District	Shah Alam	Puchong	Petaling Jaya	Subang Jaya	Seri Kembangan	Sungai Buloh
<5% difference	22.85%	26.13%	27.12%	20.71%	25.72%	25.81%	24.72%
<10% difference	42.12%	45.70%	47.97%	40.16%	50.36%	47.17%	40.45%
<30% difference	83.25%	85.62%	88.01%	82.07%	90.53%	87.10%	79.78%
MSE							
For difference <30%	6962.2	5603.9	4796.7	11916.7	6317.0	2383.9	3576.0
Whole testing sample	19034.6	13281.5	10753.7	28421.7	13985.6	6027.6	18509.6
Smallest *h* to produce better result than MR model	1	1	1	2	3	3	3
***h* nearest houses to achieve minimum MSE**	5	5	4	4	5	5	7

On the other hand, in terms of the consistency of the prediction accuracy, M_PULFR model remains outperformed MR model in all cases. This can be seen from Table 5.1 and Table 5.2 where M_PULFR model consistently predicted more than 20%, 40%, and 80% of the predicted housing prices with prediction errors less than 5%, 10%, and 30% of differences respectively for all

sub-regions except for Sungai Buloh which is 79.78%. These statistics remain unchanged when it comes to the district level. However, from Table 5.1, it is noticed that the level of the prediction accuracy of MR model dropped when comes to district level where MR model shows only 67.98% of the predicted housing prices are with prediction errors less than 30% of differences. This implies that MR model is less flexible and not robust to geographical attributes as its prediction accuracy is affected when the study area becomes larger (from sub-region level to district level) where the housing price variations become higher.

From Table 5.1 and 5.2, it can be seen that there is a significant reduction in MSE values for those cases with less than 30% difference from actual housing prices for both models. In Petaling District, (MR, M_PULFR) model predicted that (67.98%, 83.25%) of the cases have prediction error of less than 30% between predicted and actual housing prices with MSE of (9795.9, 6962.2) which is much lower than the MSE (27199.7, 19034.6) for the entire testing sample. This means that a sizable portion of MSE that is (17403.8, 12072.4) or equivalently (63.99%, 63.42%) in Petaling District is caused by the remaining (32.02%, 16.75%) of the cases. The same situations happened to the sub-regions for M_PULFR and MR models. This implies that, when the predictions are accurate, the values predicted might close to actual values. However, when the predictions are less accurate, the values predicted might far different from the actual values.

The smallest *h* in Table 5.2 indicates the numbers of houses with the most similar housing attributes required by M_PULFR model in order to achieve a better prediction as compared to MR model. For example, in Petaling District, M_PULFR model achieved a better result (smaller MSE) compared to MR model when the housing prices of the most similar houses ($h=1$) are served as the reference prices. On the other hand, in Petaling Jaya, M_PULFR model failed to generate a smaller MSE value compared to MR model when $h=1$ is used. However, the prediction of M_PULFR model is improved and outperformed MR model when $h=2$ is used.

The *h* nearest houses showed in Table 5.2 are the numbers of houses with the most similar housing attributes needed to achieve the best predictions. This *h* nearest houses will be used to produce the smallest MSE for M_PULFR model. In Petaling District, the MSE is minimised when $h=5$ is used while in Petaling Jaya, the best prediction is achieved when $h=4$. Refer to Appendix A for the detailed prediction accuracy of M_PULFR model using different *h* values. In general, M_PULFR model will achieve better results than MR model for Petaling District and all sub-regions when $h=3$ similar housings with the most similar attributes were used.

5.3 Discussion of Housing Market Volatility

It can be observed from Table 5.2 that Petaling Jaya produces the largest MSE value for the testing sample when M_PULFR model is used, while Seri Kembangan produces the smallest MSE value. This indicates that there is a high variation in the housing prices in Petaling Jaya while a small variation in

the housing prices in Seri Kembangan. However, this information is not comparable to the housing price levels or the averages of the housing prices are neither generally equal in all sub-regions nor in the entire Petaling District.

The coefficient of variation (CV) is a typical statistical measure used to analyse and compare the relative variability of two or more different data sets with different means (Groebner et al., 2010). However, without computing CV, the variability of the housing prices in different housing markets can also be analysed and compared using *h* nearest houses in M_PULFR model. Table 5.2 shows that M_PULFR model required five houses with the most similar attributes to produce minimum MSE (best prediction) for the housing prices in Petaling District, Shah Alam, Subang Jaya, and Seri Kembangan. On the other hand, M_PULFR model required seven houses with the most similar attributes to produce minimum MSE for the housing prices in Sungai Buloh. In other words, in Petaling District, Shah Alam, Subang Jaya, and Seri Kembangan, there exists a house with closest possible housing price for every five houses with the most similar attributes. However, in Sungai Buloh, there exists a house with closest possible housing price for every seven houses with the most similar attributes. This implies that the housing market in Sungai Buloh is relatively more volatile as compared to other sub-regions in Petaling District.

In can also be observed from Table 5.2 that Subang Jaya has the highest prediction accuracy, i.e. 90.53% of the predicted housing prices are within 30% of prediction errors. This indicates that the predictive ability of M_PULFR model is relatively better in predicting the housing prices in Subang Jaya as

compared to Petaling District and other sub-regions. This may also imply that the housing prices in Subang Jaya are comparatively more predictable or foreseeable.

5.4 Investigation of the Effect of λ on M_PULFR Model

The ratio of the error variances, $\lambda = 1$ (refer to Result 1, Section 3.2) was used throughout this study, i.e. both dependent and independent variables have the same errors' size. However, this assumption may not be held in the real situation where the housing prices and the housing attributes usually have different error sizes. Therefore, this section will investigate the effect of different λ values based on the data from Petaling District. We fixed $h = 5$ in this investigation because it produced the best prediction outputs in Petaling District for M_PULFR model.

Table 5.3 compares the outputs obtained from $\lambda = 0.5, 1.0, 1.5$. It is observed that the magnitude and sign (positive or negative relationship) of the estimated coefficient are almost the same for different λ values while it produces the smallest MSE value when $\lambda = 1$. For the testing sample, the differences for the percentages of the number of houses with prediction errors which are <5%, <10% and <30% are negligible when λ value is deviated from 1.0. It is also observed that the differences between the MSEs produced when using different λ values are insignificant. Thus it can be concluded that the assumption of $\lambda = 1$ is still acceptable even if there is a mild deviation in the actual data set.

Table 5.3: Performance Measures of M_PULFR Model for Petaling District using $\lambda = 0.5, 1.0, 1.5$

λ	**0.5**	**1.0**	**1.5**
	Training Sample		
Constant	10660.79	10660.79	10660.78
Lot size	32433.17	32433.17	32433.17
Tenure	-2510.63	-2510.63	-2510.63
Expiry	3232.74	3232.74	3232.74
Terrace	10711.51	10711.51	10711.51
Bedroom	9968.89	9968.89	9968.89
Built up	5077.58	5077.58	5077.58
Mall	-11291.13	-11291.13	-11291.13
Market	-74566.32	-74566.31	-74566.30
Time	2752.30	2752.30	2752.30
MSE	8.15E-04	1.94E-05	1.57E-04
R^2	0.9999997	0.9999997	0.9999997
	Testing Sample		
<5% difference	22.86%	22.85%	22.86%
<10% difference	42.13%	42.12%	42.10%
<30% difference	83.27%	83.25%	83.29%
MSE			
For difference <30%	6963.60	6962.23	6968.77
Whole testing sample	19034.63	19034.64	19034.62

5.5 Summary

In chapter four, the results obtained have proven that the proposed M_PULFR model is better in terms of fitting ability. In this chapter, the results showed that M_PULFR model remains outperformed MR model as M_PULFR model shows a stronger predictive ability where it produced lower MSE but higher R^2 as compared to MR model.

Another strength showed by M_PULFR model is the model robustness where M_PULFR model consistently predicted more than 20%, 40%, and 80% (except Sungai Buloh which is only 79.78% but near to 80%) of the observations in the testing samples whose error of predictions is less than 5%, 10%, and 30% respectively, regardless of predictions by region or by the district. In contrast, MR model only shows these statistics for the testing sample for the cases from Subang Jaya.

Besides, M_PULFR model can also be used to study and compare the volatility of the housing prices in different sub-regions which cannot be done by using MR model. The results showed that the housing markets in Puchong and Petaling Jaya are relatively less volatile as compared to the housing market in Sungai Buloh.

Last but not least, from section 5.4, the results showed that the effect of λ on M_PULFR model is insignificant as it does not provide an intensive impact on the estimations and predictions of housing prices.

CHAPTER 6

ANALYSIS OF AVERAGE HOUSE PRICE CHANGE

Previous chapters discuss the performance measures of the proposed multiple un-replicated linear functional relationship (M_PULFR) and multiple regression (MR) models. The results showed that M_PULFR model outperformed MR model in terms of fitting and predictive abilities. This chapter provides a comparison between the average market price movements and the estimated price of an "average" house in Petaling District from November 2008 to February 2016. This comparison shows whether the estimated prices of an "average" house is fairly priced compared to market's average prices. In this chapter, ARIMA model is used to study and predict the future trend of the prices of an "average" house in Petaling District.

6.1 Data Transformation Using M_PULFR Model

The proposed multiple un-replicated linear functional relationship (M_PULFR) model is used to study the changes in the price of an "average" house over the period of study from November 2008 to February 2016, with a total of 88 quarters. The cleaned data were transformed into time series basis using the methodology modified from NAPIC,

$Y_t = \hat{\alpha}_t + \hat{\boldsymbol{X}}_t\hat{\boldsymbol{\beta}}_t,\ t = 1,2,\cdots,n,$

where Y_t is expected price of the 'average' house with attributes $\boldsymbol{x}_0$ at time t,

$\hat{\alpha}_t$ is intercept and $\hat{\boldsymbol{\beta}}_t$ are coefficients of the linear function at time t,

$$\hat{\boldsymbol{X}}_t = \left[\lambda \boldsymbol{x}_0 + (\tilde{y}_t - \hat{\alpha}_t)\hat{\boldsymbol{\beta}}_t'\right]\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}_t \hat{\boldsymbol{\beta}}_t'\right)^{-1},$$

$\boldsymbol{x}_0$ are the attributes of the 'average' house at a base year, and

$\tilde{y}_t$ is the average price of optimum h nearest houses.

6.2 Models for Time Series Analysis

This chapter also provides an analysis and a prediction of the future housing price trend for an "average" house in Petaling District using ARIMA model.

In an autoregressive (AR) model, a linear combination of past values of a particular variable is used to forecast the value of the variable. In other words, the variable is regressed against itself under autoregressive processes (Hyndman and Athanasopoulos, 2013a). On the other hand, a moving average (MA) model uses past forecast errors instead of past values of the variable in a regression (Hyndman and Athanasopoulos, 2013b).

The combination of AR and MA processes will result in an autoregressive moving average (ARMA) process which models a stationary time series data (Chatfield, 2003). However, in many cases, time series data may not be stationary. Therefore, to fit a stationary data, "differencing" is applied to the time series data in order to remove non-stationary sources of variation (Chatfield, 2003). This integrated ARMA model is called autoregressive integrated moving average (ARIMA) model and is used to fit the differenced time series data. ARIMA(p , d , q) model is given by,

$$W_t = \alpha_1 W_{t-1} + \cdots + \alpha_p W_{t-p} + Z_t + \cdots + \beta_q Z_{t-q},$$

where $W_t = (1-B)^d X_t$, with a degree of differencing d,

B is a backward shift operator,

X_t is an autoregressive process of order p,

Z_t is a moving average process of order q.

6.3 A Study of the Average House Price Change using MpULFR Model

This section discusses the housing price movements of an "average" house in November 2008 (base-year-house) using MpULFR model. MpULFR model makes use of pure price movements of h nearest houses with the most similar attributes as the base-year-house at each time point instead of taking the average price of all houses at each time point. This new approach can reduce the impact of extreme cases (either a particular house was sold at a price that is too high or too low). In this study, $h = 5$ was used to estimate the reference prices as $h = 5$ resulted in the best prediction of MpULFR model in Petaling District.

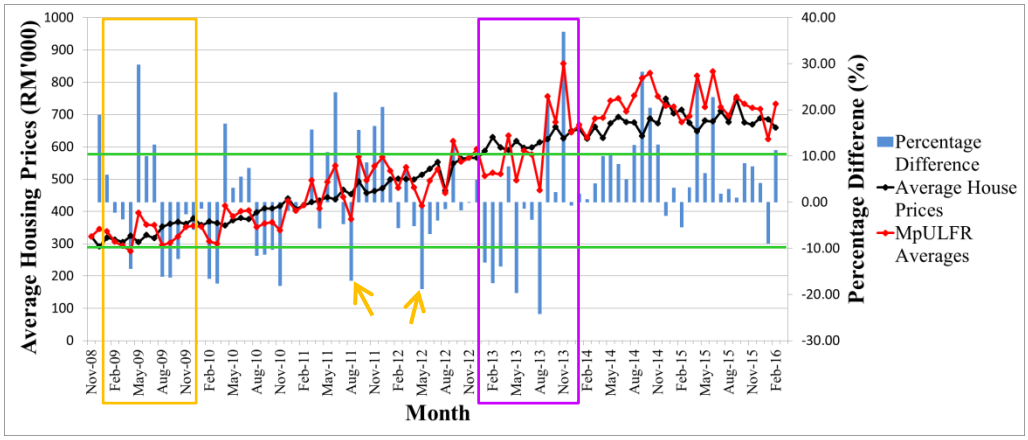

Figure 6.1: Estimated and Actual Average Housing Prices in Petaling District from November 2008 to February 2016

Figure 6.1 represents the average housing prices from actual data, the estimated average housing prices using M_PULFR model and the difference of average housing prices (in percentage) in Petaling District over a period from November 2008 to February 2016. It is observed that the prediction errors between estimated prices and actual prices have generally fluctuated within 10% of differences over the study period except for the year 2009 (framed by an orange rectangle) and 2013 (framed by a purple rectangle) where the prediction errors exceeded 10% of differences in general. It is believed that the performance of M_PULFR model in these two periods, i.e. 2009 and 2013 are affected by the market instability due to the external factors such as global financial crisis and reinforcement of new policies where consumers' responses are uncertain as the housing markets are neither predictable nor foreseeable. This implies that the estimations from M_PULFR model are reliable when the economy is free from any form of crises and no enforcement of new rules and regulations.

NAPIC (2015) reported an average drop in housing prices in Petaling District in 2009 and 2013. The global financial crisis which resulted from the failure of Lehman Brothers Bank, one of the US investment banks, in the fourth quarter of the year 2008 caused a significant contraction in real GDP in the first quarter of 2009 in Malaysia. Malaysian economy contracted by 1.7% in 2009 (Bank Negara Malaysia, 2009). According to NAPIC (2015), the MHPI for Petaling District dropped from 129.0 in 2008 to 126.1 in 2009. From Figure 6.1, it is observed that the base-year-house was sold at a lower price under-estimation from M_PULFR model compared to the average market price

in the year 2009, and this circumstance persisted until the first quarter of 2011. M_PULFR model shows that the negative impact of the global financial crisis on the housing prices in Petaling District had persisted for about two years.

Besides, NAPIC (2015) reported that the MHPI for Petaling District decreased from 204.3 in 2012 to 204.0 in 2013. This is in line with the estimations produced by M_PULFR model. It can be seen from Figure 6.1 that on average, the base-year-house was generally sold at estimated prices that are lower than the market prices. This might due to the reinforcement of stricter lending guidelines introduced by the Central Bank of Malaysia (Bank Negara Malaysia, 2012) and a heavier RPGT, from 10% in 2012 (Lembaga Hasil Dalam Negeri Malaysia, 2012) to 15% in 2013 (Lembaga Hasil Dalam Negeri Malaysia, 2013) for the houses that resold within two years from purchase, imposed by Inland Revenue Board of Malaysia with the intention to discourage speculative activity in the property market (Bank Negara Malaysia, 2012). This effect did not persist for a long duration as M_PULFR model shows that the base-year-house was sold at an estimated price higher than the market price in September 2013 and this circumstance persisted since then where the base-year-house was sold at estimated prices that are generally higher than the market prices. Although the reinforcement of policy on RPGT did not give a persistence effect on housing prices, it does cause a reduction in housing transaction (Starproperty.my, 2017). In 2014, the RPGT was further increased to 30% for the houses that resold within three years from purchase (Lembaga Hasil Dalam Negeri Malaysia, 2014) and this has caused a downward trend on the transaction amount which can be seen from Figure 6.2.

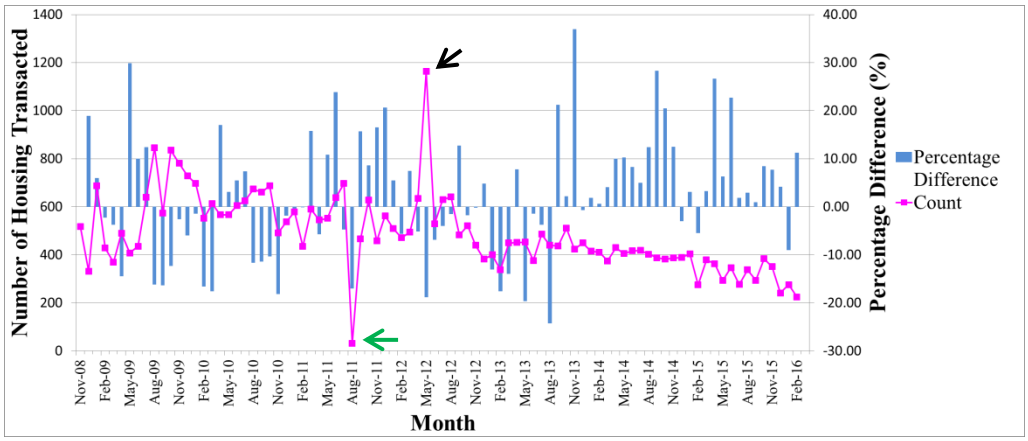

Figure 6.2: Percentage Difference between Estimated and Actual Average Housing Prices and Number of Housing Transacted in Petaling District from November 2008 to February 2016

It can be seen from Figure 6.1, MPULFR model shows that the base-year-house was sold at estimated prices that are generally higher than the market prices in 2011 and fluctuated within 10% of differences in 2012 except for August 2011 and May 2012 which are 15% lower as compared to the market prices (pointed by arrows in Figure 6.1). The estimations produced by MPULFR model for August 2011 and May 2012 may be less accurate as the number of observations for August 2011 is limited (less than 50 which can be seen from Figure 6.2, pointed by a green arrow). Besides, there are no observations for May 2012. Therefore, in order to obtain the estimation for this time point, the observations for April 2012 and June 2012 were grouped and served as the housing data for May 2012 (pointed by a black arrow). This may affect the accuracy of the estimation for May 2012.

6.4 Is there a Housing Bubble in Petaling District?

From Figure 6.1, MPULFR model shows that the base-year-house was sold at estimated prices that are generally higher than the market prices since

September 2013. Is this an indication of a housing bubble in the housing market in Petaling District? Since there are no commonly accepted methodologies for identifying the boom and burst in housing markets (Agnello and Schuknecht, 2009), thus, in order to study and investigate the housing bubble, some definitions of the housing bubble from previous studies have to be quoted. Firstly, according to Adalid and Detken (2007), a boom of asset prices is defined as a period in which real asset prices are 10% higher than an estimated trend. Secondly, according to Lind (2009), bubble exits only if the housing prices increased drastically for a certain period of time and then fell drastically.

In this study, the burst of housing bubbles is defined by adopting the standard definition of "recession", where "recession" is defined as the negative growth of GDP for two consecutive quarters (six months) or longer (Arnold, 2008), with the aid of the above-quoted definitions. A housing bubble is defined as a period in which houses are sold at estimated prices that are 10% higher than the market's average prices for two consecutive quarters or more. In contrast, the burst of housing bubbles is defined as a period in which houses are sold at estimated prices that are 10% lower than the market's average prices for two consecutive quarters or more. By using these definitions, there are no indications of bubbles or a burst (refer to Figure 6.1) in the housing market in Petaling District.

6.5 Prediction of House Price Movements using ARIMA

This section provides a comparison of the predictions of the estimated and average housing prices in Petaling District over March to August 2016 using ARIMA model.

It can be seen from Figure 6.3 and Figure 6.4 that the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots suggest an MR(1) model and an AR(2) model respectively for the market's average housing prices in Petaling District. For the estimated prices of the "average" house in Petaling District, Figure 6.5 and Figure 6.6 show that the ACF and PACF plots suggest an MR(1) model and an AR(3) model respectively for these data.

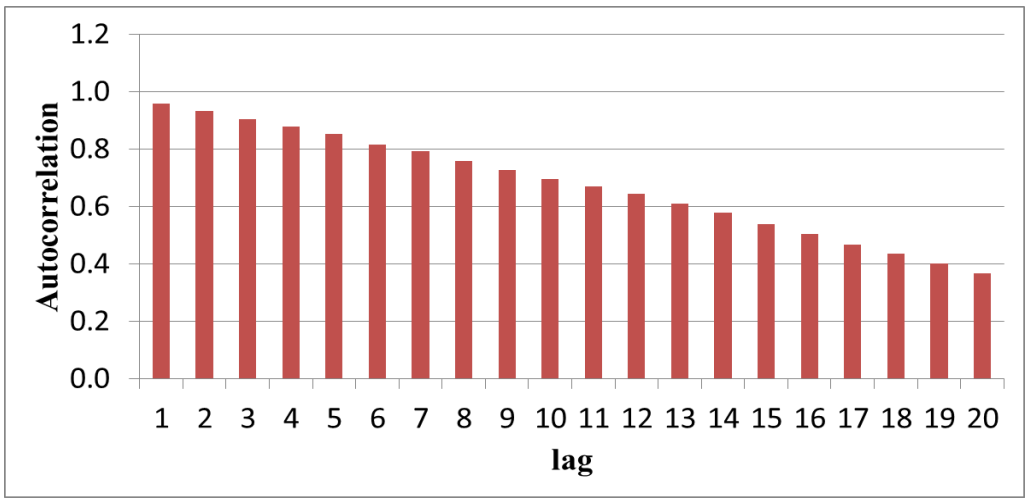

Figure 6.3: Autocorrelation Function for Average Housing Prices in Petaling District from November 2008 to February 2016

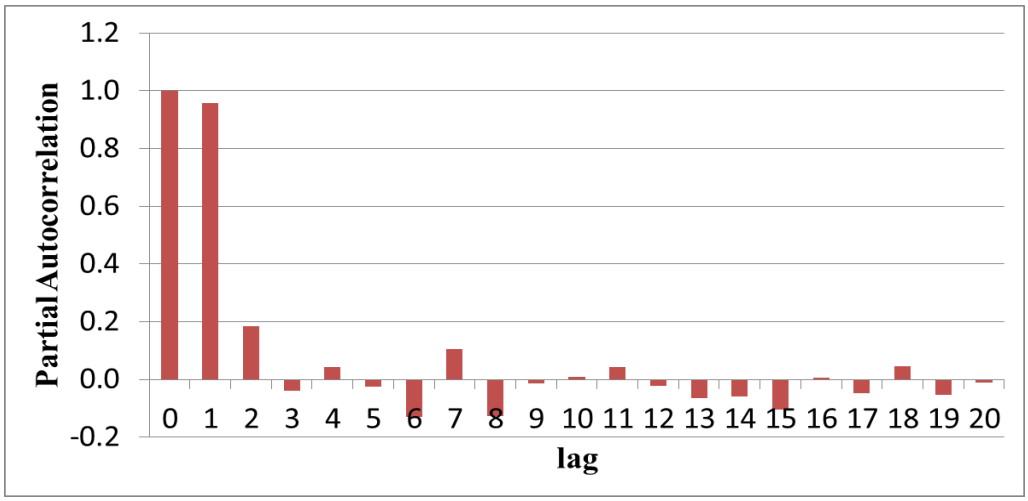

Figure 6.4: Partial Autocorrelation Function for Average Housing Prices in Petaling District from November 2008 to February 2016

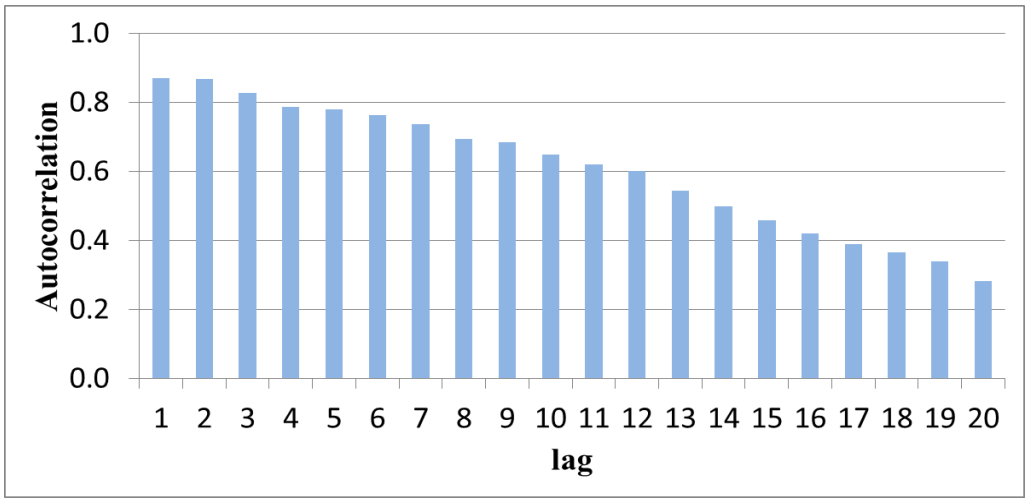

Figure 6.5: Autocorrelation Function for the Prices of "Average" House Estimated by M_PULFR Model from November 2008 to February 2016

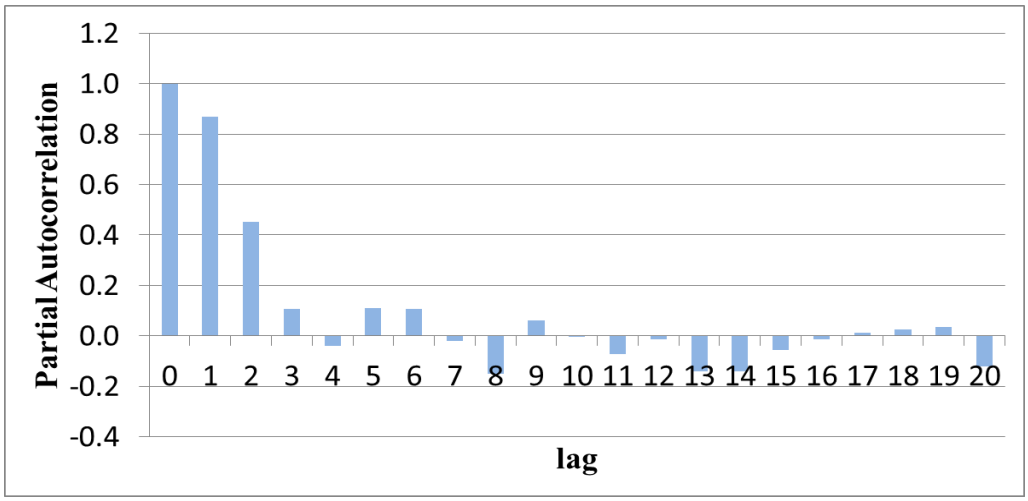

Figure 6.6: Partial Autocorrelation Function for the Prices of "Average" House Estimated by MPULFR Model from November 2008 to February 2016

In this study, ARIMA(1, 1, 2) and ARIMA(1, 1, 3) were used to study and predict the price movements of the market's average housing prices and the estimated prices of the "average" house respectively. First-order differencing, $d = 1$, was applied in this study as the statistical studies show that first-order differencing is sufficient to make a series stationary (Chatfield, 2003).

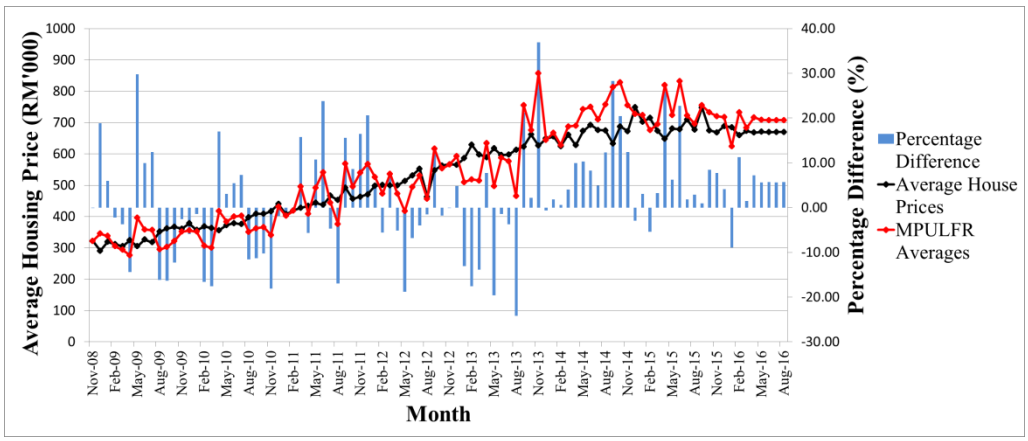

Figure 6.7: Percentage Difference between Prediction of Estimated and Average Housing Prices in Petaling District from March to August 2016

It can be seen from Figure 6.7 that there are no indications of bubbles or a burst in the housing market in Petaling District for the following six months, March to August 2016. However, the “average” house is predicted to be sold at estimated prices that are higher than the predicted market’s average prices for the period from March to August 2016.

CHAPTER 7

CONCLUDING REMARKS

Housing price modelling is not only can help to prepare for the worst scenario, but it does provide useful information for investment-decision-making. If one can understand housing market well, they will be able to know the right time to enter the housing market and earn a profit. Besides, buyers could determine if a house is overpriced with the housing prices model. At the same time, investors can avoid buying an unproductive house that is miss-built or with unfavourable attributes which may cause the house to remain vacant or unsold.

7.1 Research Conclusion

7.1.1 Development of M_PULFR and its theoretical properties

A new functional model for analysing the relationship between a dependent variable (housing price) and a set of independent variables (attributes) was proposed. The parameters in the proposed M_PULFR model were estimated using maximum likelihood estimator. The ratio of error variances, λ is assumed to be known and set equal to one. However, further study was done and the results showed insignificant difference in terms of accuracy despite different λ values were used. The properties of the estimators such as unbiasedness and consistency were investigated using Taylor approximations (refer to Section 3.2.1) and Fisher information matrix (refer to

Section 3.2.2). The coefficient of determination was also developed to assess the performance of the model. The coefficient of determination (R^2) was also derived as a performance indicator for the model. The estimation of the M_PULFR model's parameters is not affected by the multicollinearity effect that caused by the correlation between independent variables (refer to Section 4.2.1), which in turn provides better and consistent results in housing prices modelling.

7.1.2 Application in Housing Market

The proposed M_PULFR model is the first functional relationship model that applied to the study of the housing market. There are several contributions or achievements of this study in solving the problems of the housing market.

i. Using larger actual transacted data set

   The literature review chapter showed that most of the housing market studies in Malaysia relied on the survey data or small transacted housing data. These studies also limited to a smaller housing area such as a town or a condominium block. In this study, we considered one of the most economically developed and populated Petaling District and its sub-regions in Malaysia. The dataset we used consists of 41750 actual transacted housing records from November 2008 until February 2016. There were also comparisons between the district level and its sub-regions.

ii. Better fitting and predictive abilities

The proposed M_PULFR model showed an advantage of producing more consistent and justifiable results for housing market data as compared to the MR model (refer to Section 4.2). In the model validation part, M_PULFR model provided a predictive ability with higher accuracy, regardless of overall MSE or the smallest 5, 10 and 30 percentiles of the predicted housing price difference (refer to Tables 5.1 and 5.2). In additions, a reliable performance in model consistency is observed in M_PULFR model as M_PULFR models always outperformed compared to MR models in Petaling District and its sub-regions (refer to Section 5.2). Therefore, the proposed M_PULFR model is effective forecasting of individual housing prices, which is not only benefited those potential house buyers but also benefited real estate properties investors and policy-makers. In other words, they can use the proposed model to predict the market price of a particular house and compare with the offer price to assist them to know the current market condition in which they are able to know whether or not the house is worth to buy (sell) or invest.

iii. Non-uniformity of housing market behaviour in Petaling District

Through the analysis of the M_PULFR model results, we found that the housing market behaviour in Petaling District is not uniform (refer to Section 4.3). There are four types of housing

market behaviour in Petaling District, represented by Shah Alam, Petaling Jaya, Seri Kembangan and other sub-regions. The two main housing attributes that differentiating the housing market behaviours are the distance to the nearest shopping mall and supermarket. An additional housing attribute that differentiates Seri Kembangan from other regions is the tenure type.

iv. Under-priced of housing market in Petaling District

It is a common perception that the terrace houses in Malaysia are over-priced and unaffordable over the past decades. However, our study in Chapter Six indicated that the housing prices in Petaling District were under-priced since the fourth quarter of 2013. The "average" house in Petaling District was sold at the estimated prices using M_PULFR model that is generally higher than the market averages after September 2013 (refer to Figure 6.1).

7.2 Areas of Further Research

7.2.1 Estimation of Reference Housing Price

The prediction of housing price using M_PULFR model does not only depend on the acquired house attributes, but also the acquired house price (refer Section 5.1). This drawback of the proposed M_PULFR model is solved by calculating a reference house price using the average transacted housing price that has the most similar attributes in Chapter Five. In addition, a preliminary study on reference price using median house price of h nearest

houses using Subang Jaya's data showed that there is no significant difference in terms of accuracy compared to using average house price (refer to Appendix C). This indicates that the average house price used has reduced the effect of extreme prices. However, other statistics could be considered in future study to improve the accuracy of the proposed model in predicting housing prices.

7.2.2 Study of the M_PULFR Model with Autoregressive Errors

In this study, we assumed that the housing prices and housing attributes have uncorrelated observations (spatial data without autoregressive errors). However, it is possible that the observed values of housing prices or any of the housing attributes are correlated from one to another. Hence, the study of M_PULFR model can be extended to a times series based functional relationship model with the errors independence assumption is relaxed and the models in Equation (3.1) and Equation (3.2) can be rewritten as

$$y_t - \phi_1 Y_t - \phi_2 Y_{t-1} - \cdots - \phi_{p+1} Y_{t-p} = \theta_t$$

$$x_{tk} - \varphi_1 X_{tk} - \varphi_2 X_{(t-1)k} - \cdots - \varphi_{q+1} X_{(t-q)k} = \vartheta_{tk}$$

where θ_t and ϑ_{tk} are white noise process. The time series based functional relationship model follows the *p*-order autoregressive model and *q*-order autoregressive model respectively.

7.2.3 Consideration of Fixed Effects of Attributes on Prices

In Section 4.3.2, it is observed that house buyers in Seri Kembangan prefer leasehold houses as these houses are mainly located at areas nearby industrial area or business centre. These unevenly distributed houses may result

in developing a bias model. Therefore, a fixed effect for tenure type on housing prices could be added to the regression model.

REFERENCES

Adalid, R. and Detken C., 2007. Liquidity shocks and Asset Price Boom/Bust Cycles. European Central Bank: Working paper series, no. 732.

Adcock, R. J., 1877. Note on the method of least squares. *Analyst*, 4, pp. 183 – 184.

Agnello, L. and Schuknecht, L., 2009. Booms and Busts in housing markets: Determinants and implications. European Central Bank: Working paper series, no. 1071.

Arnold, R. A., 2007. *Economics*. United States of America: Cengage Learning.

Babawale, G. K., and Adewunmi, Y, 2011. The impact of neighbourhood churches on house prices. *Journal of Sustainable Development*, 4, pp. 246 – 253.

Balchin, P. N., Bull, G. H., and Hieve, J. L., 1995. *Urban land economics and public policy*, 5th ed. Houndmills: Palgrave.

Bank Negara Malaysia, 2009. *Economic developments in 2009* (Annual report 2009). Malaysia: Bank Negara Malaysia.

Bank Negara Malaysia, 2012. *Monetary policy in 2012* (Annual report 2012). Malaysia: Bank Negara Malaysia.

Baranzini, A., Ramirez, J., Schaerer, C., and Thalmann, P., 2008. *Hedonic methods in housing markets: pricing environmental amenities and segregation*. New York: Springer Science.

Caires S. and Wyatt, L. R., 2003. A linear functional relationship model for circular data with an application to the assessment of ocean wave measurements. *Journal of Agricultural, Biological, and Environmental Statistics*, 8, pp. 153 – 169.

Case, K. E. and Shiller, R.J., 1987. Prices of single-family homes since 1970: New indexes for four cities. *New England Economic Review*, pp. 45 – 56.

Case, K. E. and Shiller, R.J., 1989. The efficiency of the market for single-family homes. *The American Economic Review*, pp. 125 – 137.

Cebula, R. J., 2009. The hedonic pricing model applied to the housing market of the city of Savannah and its Savannah historic landmark district. *The Review of Regional Studies*, 39, pp. 9 – 22.

Chan N. N. and Mak, T. K., 1984. Heteroscedastic errors in a linear functional relationship. *Biometrika*, 71, pp. 212 – 215.

Chang, Y. F., Omar M. R. and Syed, A. R. A. B., 2010. Multidimensional Unreplicated Linear Functional Relationship Model with single slope and its coefficient of determination. *WSEAS Transaction on Mathematics*, 9, pp. 295 – 313.

Chatfield, C., 2003. *Some time-series models*, 6th ed. London: Chapman and Hall.

Chau, K. W., and Chin, T. L., 2002. A critical review of literature on the hedonic price model and its application to the housing market in Penang. *Housing Science and Its Application*, 27, pp. 145 – 165.

Chen, D. R., Gan, C., Hu, B., and Cohen, D. A., 2013. An empirical analysis of house price bubble: A case study of Beijing housing market. *Research in Applied Economics*, 5, pp. 77 – 97.

Clark, D. E. and Herrin, W. E., 2000. The Impact of public school attributes on home sale price in California. *Growth and Change*, 31, pp. 385 – 407.

Demographia, 2017. *13th annual international housing affordability survey: 2017: rating middle-income housing affordability*.

Central Bank of Malaysia, 2013. *Financial stability and payment systems report 2012: Developments in the housing market and implications on financial stability*.

Coleman, M., LaCour-Little, M., and Vandell, K. D., 2008. Subprime lending and the housing bubble; Tail wags dog?. *Journal of Housing Economics*, 17, pp. 272 – 290.

Francis, A, 2004. *Business mathematics and statistics*, 6th ed. Canada: Cengage Learning EMEA.

Freund, R. J., Wilson, W. J., and Sa, P., 2006. *Regression Analysis: Statistical modeling of a response variable*, 2nd ed. United Sates of America: Elsevier.

Fuller, W. A., 1987. *Measurement error models*. New York: John Wiley.

Giannoulakis, S., Karanikolas, N., Xifilidou, A. and Perchanidis, L., 2016. The impact of the financial crisis on residential property market of Greece. Christchurch: Recovery from Disaster: FIG Working Week 2016.

Gabriel, S., 1984. A note on housing market segmentation in an Israeli development town. *Urban Studies*, 21, pp. 189 – 194.

Groebner, D. F., Shannon, P. W., Fry, P. C., and Smith, K. D., 2010. *Business Statistics*, 8th ed. Prentice Hall: Pearson.

Hamid, S., Lawler, K. A., and Katos, A. V., 2000. *Econometrics: a practical approach*. London: Routledge.

Hardy, M. A., 1993. *Regression with dummy variables*. Newbury Park: SAGE publications.

Hashim, Z.A., 2010. House price and affordability in housing in Malaysia. *Akademika*, 78, pp. 37 – 46.

Hyndman, R. J., and Athanasopoulos, G., 2013a, *Forecasting: principles and practice: Section 8.3 Autoregressive models* [online]. Available at: https://www.otexts.org/fpp/8/3 [Accessed: 25 April 2017].

Hyndman, R. J., and Athanasopoulos, G., 2013b, *Forecasting: principles and practice: Section 8.4 Moving average models* [online]. Available at: https://www.otexts.org/fpp/8/4 [Accessed: 25 April 2017].

Hox, J. J., 2002. *Multilevel analysis: techniques and applications*. London: Lawrence Erlbaum Associates.

International Council of Shopping Centers, 2015, *Asia-Pacific shopping centre classification* [online]. Available at: https://www.icsc.org/uploads/research/general/Asia-Pacific_Shopping_Centre_Classification_Standard.pdf [Accessed: 27 August 2017].

Ismail, S., and Macgregor, B. D., 2005. *Hedonic modelling of housing markets using geographical information system (GIS) and spatial statistics: A case study of Glasgow, Scotland*. PhD thesis, University of Aberdeen, Scotland.

Ismail, S., Jalil, I. N. and Megat Muzafar, P. M., 2015. *Making housing affordable*. Kuala Lumpur: Khazanah Research Institute.

James, G. M., 2002. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society*, 64, pp. 411 – 432.

Jiang, L., Phillips, R. C. B. and Yu, J., 2014. A new hedonic regression for real estate prices applied to the Singapore residential market. *Yale University: Cowles Foundation Discussion Paper, No. 1969.*

Joint Center for Housing Studies., 2016. *The state of the nation's housing 2016: housing challenges*. Cambridge: Harvard University.

Kam, K. J., Chuah, S. Y., Lim, T. S., and Ang, F. L., 2016. Modelling of property market: the structural and locational attributes towards

Malaysian properties. *Pacific Rim Property Research Journal*, 22, pp. 203 – 216.

Kendall, M. G., 1951. Regression, structure and functional relationship – Part I. *Biometrika*, 38, pp. 11 – 25.

Lancaster, K. J., 1966. A new approach to consumer theory. *Journal of Political Economy*, 74, pp. 132 – 157.

Lembaga Hasil Dalam Negeri Malaysia, 2012. *Tax Brochure 2012: Real Property Gains Tax (RPGT)*.

Lembaga Hasil Dalam Negeri Malaysia, 2013. *Tax Brochure 2013: Real Property Gains Tax (RPGT)*.

Lembaga Hasil Dalam Negeri Malaysia, 2014. *Tax Brochure 2014: Real Property Gains Tax (RPGT)*.

Lind, H., 2009. Price bubbles in housing markets: Concept, theory and indicators. *International Journal of Housing Markets and Analysis*, 2, pp. 78 – 90.

Matignon, R., 2007. Explore Nodes. In: *Data mining using SAS enterprise miner*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Miles, W., 2008. Volatility clustering in U.S. home prices. *Journal of Real Estate Research*, 30, pp. 73 – 90.

Monson, M., 2009. Valuation using hedonic pricing models. *Cornell Real Estate Review*, 7, pp. 62 – 73.

NAPIC, 2015. *The Malaysian house price index: Explanatory notes*.

Norshazwani, A. R., Mohd, L., and Abdul, J. O., 2013. The revisited of Malaysian house price index. *Proceedings International Conference of Technology Management, Business and Entrepreneurship 2012*, 18 – 19 December 2012 Melaka, Malaysia. Melaka: ICTMBE UTHM, pp. 656 – 667.

Ong, T. S., 2013. Factors affecting the price of housing in Malaysia. *Journal of Emerging Issues in Economics, Finance and Banking*, 1, pp. 414 – 429.

Ong, T. S., and Chang, Y. S., 2013. Macroeconomic determinants of Malaysian housing market. *Human and Social Science Research*, 1, pp. 119 – 127.

Ooi, J. T. L., Le, T. T. T., and Lee, N. J., 2014. The impact of construction quality on house prices. *Journal of Housing Economics*, 26, pp. 126 – 138.

Osmadi, A., Mustafa Kamal, E., Hassan, H. and Abdul Fattah, H., 2015. Exploring the elements of housing price in Malaysia. *Canada: Asian Social Science*, 11, pp. 26 – 38.

Owusu-ansah, A., 2012. Examination of the determinants of housing values in Urban Ghana and implications for policy makers. *Journal of African Real Estate Research*, 2, pp. 58 – 85.

Panagiotidis, T. and Printzis, P., 2015. On the macroeconomic determinants of the housing market in Greece: A VECM approach. *Hellenic Observatory European Institute, GreeSE paper no.* 88.

Pashardes, P., and Savva, C. S., 2009. Factors affecting house prices in Cyprus: 1988 – 2008. *Cyprus Economic Policy Review*, 3, pp. 3 – 25.

Pearson, K., 1901. On lines and planes of closet fit to systems of points in space. *Philosophical Magazine*, 2, pp. 559 – 572.

Pfaff, B., 2008. *Analysis of integrated and cointegrated time series with R*. New York: Springer Science.

Reginald, W. T., and Richard, J. H., 1980. *Modelling in geography: a mathematical approach*. New Jersey: Barnes & Noble Books.

Reiersol, O., 1945. Confluence analysis by means of instrumental sets of variables. *Ark. Mat. Astronomique Fysics*, 32, pp. 1 – 119.

Rencher, A. C., 2002. *Methods of multivariate analysis*, 2nd ed. Canada: John Wiley and Sons, Inc.

Rosen, S., 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82, pp. 35 – 55.

Rosiers, F. Des, Lagana, A., Thériault, M., and Beaudoin, M., 1996. Shopping centres and house values: an empirical investigation. *Journal of Property Valuation and Investment*, 14, pp. 41 – 62.

Sprent, P., 1969. *Model in regression and related topics*. London: Methuen & Co. Ltd.

Starproperty.my., 2017, *Has Malaysia's housing bubble burst in 2013?* [online]. Available at: http://www.starproperty.my/index.php/articles/events/pundits-property-market/ [Accessed: 16 June 2017].

Suhaida, M. S. et al., 2011. Housing affordability: A conceptual overview for house price index. *Procedia Engineering*, 20, pp. 346 – 353.

Tan, Y. K., 1999. An hedonic model for house prices in Malaysia. *Proceedings of the PRRES Conference 1999*, 26 – 30 January 1999 Kuala Lumpur, Malaysia. Kuala Lumpur: Pacific Rim Real Estate Society, pp. 1 – 15.

Trading Economics, 2017, *Malaysia house price index* [online]. Available at: http://www.tradingeconomics.com/malaysia/housing-index [Accessed: 21 May 2017].

Tse, R. Y. C., and Love, P. E. D., 2000. Measuring residential property values in Hong Kong. *Property Management*, 18, pp. 366 – 374.

Unit DEGIS Pejabat Daerah & Tanah Petaling, 2014. Pelan gunatanah semasa Daerah Petaling [Current land use of Petaling District].

United States of America before Federal Trade Commission, 1998, *Complaint: docket no. C-3838* [online]. Available at: https://www.ftc.gov/sites/default/files/documents/cases/1998/12/9810134cmp.htm [Accessed: 27 August 2017].

Wikipedia, 2017a, *Petaling District* [online]. Available at: https://en.wikipedia.org/wiki/Petaling_District [Accessed: 8 March 2017].

Wikipedia, 2017b, *Shah Alam* [online]. Available at: https://en.wikipedia.org/wiki/Shah_Alam [Accessed: 10 May 2017].

Wong, R., 2017, *Turning Hong Kong's housing challenge into a housing solution* [online]. Available at: http://www.scmp.com/business/article/2066745/turning-hong-kongs-housing-challenge-housing-solution [Accessed: 18 April 2017].

Yan, X., and Su, X. G., 2009. *Linear regression analysis: theory and computing*. Singapore: World Scientific.

Yardney, M., 2015, *What stops young people from getting on the property ladder… and 12 ways to start climbing* [online]. Available at: http://www.smartcompany.com.au/finance/what-stops-young-people-from-getting-on-the-property-ladder-and-12-ways-to-start-climbing/ [Accessed: 10 May 2017].

Yusof, A. M., 2012. Malaysian housing investment information price modelling. *1st NAPREC Conference*, 21 October 2008, pp. 1 – 30.

Yusof, A. M., and Ismail, S., 2012. Multiple regressions in analysing house price variations. *Communications of The IBIM*, 2012, doi: 10.5171/2012.383101.

APPENDICES

Appendix A1 Performance Measures of M_PULFR and MR Models for Shah Alam

Error of prediction	**MR**	**M_PULFR**							
<5% difference	17.49%	27.13%	26.03%	26.10%	26.13%	26.13%	26.37%	25.51%	25.34%
<10% difference	33.74%	44.94%	44.38%	45.25%	46.15%	45.70%	46.32%	46.63%	46.49%
<30% difference	74.80%	81.40%	84.20%	85.55%	85.69%	85.62%	86.10%	86.48%	86.07%
MSE									
For difference <30%	7884.3	5357.2	5531.2	5505.9	5600.7	5603.9	5764.8	5628.8	5754.4
Whole testing sample	21956.4	17619.8	14516.2	13693.7	13476.9	13281.5	13371.9	13356.8	13507.9
h nearest	N/A	1	2	3	4	5(best)	6	7	8

Appendix A2 Performance Measures of M_PULFR and MR Models for Puchong

Error of prediction	**MR**	**M_PULFR**							
<5% difference	15.17%	24.77%	25.34%	27.55%	27.12%	26.59%	26.34%	27.05%	26.27%
<10% difference	29.73%	44.68%	46.72%	47.82%	47.97%	47.82%	48.04%	48.50%	47.25%
<30% difference	72.66%	83.08%	86.37%	87.47%	88.01%	88.08%	88.01%	88.47%	88.65%
MSE									
For difference <30%	6499.9	4570.0	4510.0	4735.2	4796.7	4866.9	5051.2	5050.6	4988.6
Whole testing sample	18649.4	14599.0	12256.5	10959.8	10753.7	10787.8	10927.7	11051.5	11248.9
h nearest	N/A	1	2	3	4(best)	5	6	7	8

Appendix A3 Performance Measures of M_PULFR and MR Models for Petaling Jaya

Error of prediction	**MR**	**M_PULFR**							
<5% difference	16.29%	20.21%	20.40%	22.46%	20.71%	21.28%	20.79%	21.40%	21.09%
<10% difference	31.08%	37.26%	38.29%	41.15%	40.16%	40.69%	40.27%	40.69%	39.24%
<30% difference	70.37%	75.13%	80.59%	81.88%	82.07%	82.23%	82.61%	82.04%	81.62%
MSE									
For difference <30%	15483.8	11196.3	11846.7	11628.9	11916.7	12400.7	12579.3	12387.4	12551.0
Whole testing sample	38256.4	41330.8	31548.5	29714.9	28421.7	28730.0	28954.4	28998.5	29353.5
h nearest	N/A	1	2	3	4(best)	5	6	7	8

Appendix A4 Performance Measures of M_PULFR and MR Models for Subang Jaya

Error of prediction	**MR**	**M_PULFR**							
<5% difference	22.20%	22.12%	25.64%	27.61%	26.02%	25.72%	26.27%	26.10%	26.18%
<10% difference	42.10%	44.32%	48.22%	49.14%	49.52%	50.36%	50.48%	51.07%	51.49%
<30% difference	86.34%	85.30%	88.52%	90.36%	90.62%	90.53%	90.74%	90.66%	90.78%
MSE									
For difference <30%	8158.1	5575.5	6486.0	6339.6	6546.5	6317.0	6155.9	6087.1	6313.1
Whole testing sample	15348.6	18369.3	15032.4	14149.0	14170.1	13985.6	14016.5	14074.2	14289.2
h nearest	N/A	1	2	3	4	5(best)	6	7	8

Appendix A5 Performance Measures of M_PULFR and MR Models for Seri Kembangan

Error of prediction	**MR**	**M_PULFR**							
<5% difference	18.50%	22.40%	23.31%	25.75%	25.75%	25.81%	25.87%	25.14%	25.08%
<10% difference	35.48%	41.27%	44.55%	46.32%	46.44%	47.17%	46.50%	46.20%	47.35%
<30% difference	78.51%	80.46%	84.11%	85.70%	87.04%	87.10%	86.79%	87.04%	86.73%
MSE									
For difference <30%	3098.3	2509.5	2473.1	2465.5	2454.7	2383.9	2277.8	2284.0	2377.6
Whole testing sample	6844.8	8696.6	7069.8	6422.9	6144.4	6027.6	6108.2	6116.6	6177.3
h nearest	N/A	1	2	3	4	5(best)	6	7	8

Appendix A6 Performance Measures of M_PULFR and MR Models for Sungai Buloh

Error of prediction	**MR**	**M_PULFR**							
<5% difference	11.24%	19.66%	19.66%	20.22%	25.28%	24.72%	28.65%	24.72%	19.10%
<10% difference	23.60%	34.27%	36.52%	39.89%	42.70%	37.64%	41.01%	40.45%	43.26%
<30% difference	64.61%	75.84%	82.58%	79.78%	80.90%	78.65%	79.78%	79.78%	78.09%
MSE									
For difference <30%	5228.4	4492.4	5799.2	5744.7	4652.7	4572.1	3963.5	3576.0	5168.9
Whole testing sample	21417.8	24392.3	21299.2	19806.5	18974.2	18643.0	18797.1	18509.6	18631.3
h nearest	N/A	1	2	3	4	5	6	7(best)	8

Appendix A7 Performance Measures of M_PULFR and MR Models for Petaling District

Error of prediction	**MR**	**M_PULFR**							
<5% difference	14.66%	22.61%	22.85%	23.23%	23.07%	22.85%	22.74%	21.94%	22.30%
<10% difference	28.09%	40.28%	41.76%	42.44%	42.48%	42.12%	42.13%	41.50%	41.80%
<30% difference	67.98%	78.80%	81.76%	82.81%	83.41%	83.25%	83.10%	83.14%	82.87%
MSE									
For difference <30%	9795.9	6810.7	7016.6	6891.1	6865.0	6962.2	7060.0	7051.8	7171.7
Whole testing sample	27199.7	24629.3	20584.1	19792.1	19085.3	19034.6	19069.2	19186.0	19434.9
h nearest	N/A	1	2	3	4	5(best)	6	7	8

Appendix B Comparisons of the Performance Measures of M_PULFR Model Using Mean and Median Prices of *h* Nearest Houses for Subang Jaya

Error of prediction	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
<5% difference	27.61%	27.27%	25.56%	27.27%	25.85%	25.72%	25.97%	27.52%
<10% difference	48.89%	49.06%	49.22%	50.10%	50.19%	51.07%	49.60%	51.24%
<30% difference	89.74%	89.07%	90.07%	89.99%	90.16%	89.86%	90.24%	90.03%
MSE								
For difference <30%	6162.27	5755.44	6324.20	6441.77	6324.24	6164.27	6248.27	6347.14
Whole testing sample	14624.11	15415.93	14680.28	15258.21	14712.36	15697.11	15007.92	16040.38
h nearest	3		4		5		6	

Appendix C Publication (Under Review)

Multiple Un-replicated Linear Functional Relationship Model and Its Application in Real Estate

Choong Wei Cheng[1,a)], Pan Wei Yeing[1,b)] and Chang Yun Fah[1,c)]

[1]*Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Sungai Long Campus, Jalan Sungai Long, Bandar Sungai Long, Cheras 43000, Kajang, Selangor, Malaysia*

[a)]choongwc2@1utar.my, [b)]panwy@utar.edu.my, [c)]changyf@utar.edu.my

Abstract. In this paper, a multiple un-replicated linear functional relationship model is derived where its maximum likelihood estimators are obtained from a single $p - 1$ dimensional fitted plane. Its properties of unbiasedness and consistency were investigated using Taylor approximation and Fisher information matrix respectively. Simulations were conducted to investigate the effect of different sizes of error variances and sample sizes. The developed model is applied to real estate with housing data from Petaling Jaya, Selangor state. The results obtained show that the fitting and predictive abilities of the proposed model are stronger as compared to multiple regression model when applied to the training and testing samples respectively.

Keywords: Functional model; Multiple Un-replicated Linear Functional Relationship; Petaling Jaya; Real estate

1.0 INTRODUCTION

Linear regression model has been widely used in studying the relationship between a continuous response variable and a set of explanatory variables. However, in many cases, the relationship will become invisible as a result of random fluctuations associated between variables. As Fuller[1] has pointed out, it is unrealistic if an explanatory variable can be measured exactly in all situations. Adcock[2] had first studied the problem using functional model where both response and explanatory variables are subject to errors. In 1984, Chan and Mak[3] proposed a multivariate linear functional relationship model in which error variances and covariances are unnecessarily to be homogenous. In 2002, James[4] proposed a functional generalized linear model to handle functional explanatory variables which may be measured at differing time points and sample sizes. Caires and Wyatt[5] introduced a linear functional relationship model with numerical approximation as a solution for its maximum likelihood estimation to compare two sets of circular data which are subjected to unobservable errors. Chang et al.[6] generalized the un-replicated linear functional relationship model to multidimensional cases to assess the quality of JPEG compressed images.

Multiple regression (MR) model is commonly used to study and analyze the Malaysian housing market[7,8,9]. The main limitation of MR model is the statistical and inferential problems of multicollinearity which can cause the interpretation of the linear relationship between explanatory variables (attributes) and response variable (housing price) becomes nearly impossible[10].

In this paper, we derive a multiple un-replicated linear functional relationship (M_PULFR) model where its maximum likelihood estimators are obtained from a single $p - 1$ dimensional fitted plane. M_PULFR model can overcome the limitation of MR model as multicollinearity gives no influence to M_PULFR model. We also investigate properties of these estimators such as unbiasedness and consistency. The proposed model is then applied to Petaling Jaya's

housing market and the results obtained are compared with MR model to evaluate the relevance of its application.

2.0 MULTIPLE UN-REPLICATED LINEAR FUNCTIONAL RELATIONSHIP (M_PULFR) MODEL

Suppose that Y_i is an unobservable value of dependent variable and $\boldsymbol{X}_i = \left(X_{i1}, X_{i2}, \cdots, X_{ip}\right)$ are p unobservable values of independent variables. We defined the M_PULFR model as

$$Y_i = \alpha + \boldsymbol{X}_i\boldsymbol{\beta},\ i = 1,2,\cdots,n \tag{3.31}$$

where α is intercept and $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \cdots, \beta_p\right)'$ are coefficients of the linear function. The two corresponding random variables y_i and $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \cdots, x_{ip}\right)$ are observed with errors, ε_i and $\boldsymbol{\delta}_i = \left(\delta_{i1}, \delta_{i2}, \cdots, \delta_{ip}\right)$ such that,

$$\left.\begin{aligned} y_i &= Y_i + \varepsilon_i \\ \boldsymbol{x}_i &= \boldsymbol{X}_i + \boldsymbol{\delta}_i \end{aligned}\right\} i = 1,2,\cdots,n \tag{3.32}$$

Both error vectors are assumed to be mutually independent and normally distributed with the following properties,

$E\left(\varepsilon_i\right) = 0$ and $E\left(\boldsymbol{\delta}_i\right) = \boldsymbol{0}$,

$Cov\left(\varepsilon_i, \varepsilon_j\right)$ and $Cov\left(\boldsymbol{\delta}_i, \boldsymbol{\delta}_j\right) = \boldsymbol{0}$ $\quad \forall i \neq j$,

$Cov\left(\varepsilon_i, \delta_{ik}\right) = 0$ $\quad \forall i, k$ and

$\varepsilon_i \sim NID\left(0, \omega_{11}\right)$ and $\boldsymbol{\delta}_i \sim NID\left(\boldsymbol{0}, \boldsymbol{\omega}_{22}\right)$ where $\omega_{11} = \tau^2$, and $\boldsymbol{\omega}_{22} = \sigma^2\mathbf{I}_p$ then

$$\boldsymbol{\omega} = \begin{pmatrix} \omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\omega}_{22} \end{pmatrix} \text{ where } \boldsymbol{\omega}_{12} = \boldsymbol{\omega}_{21}' = \boldsymbol{0}.$$

Result 1: Given the M_PULFR model defined by Equations (1) and (2), the maximum likelihood estimators of α and β_k are,

$$\hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}$$

$$\hat{\beta}_k = \frac{\left(S_{yy} - \lambda S_{x_k x_k}\right) + \sqrt{\left(S_{yy} - \lambda S_{x_k x_k}\right)^2 + 4\lambda S_{x_k y}^2}}{2S_{x_k y}}$$

Proof:

The joint density function of $\left(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i\right)$ or equivalently, $\left(\boldsymbol{x}_i, y_i\right)$ is

$$f\left(\boldsymbol{x}_i, y_i\right) = \frac{1}{\left(2\pi\right)^{\frac{r}{2}}\left|\boldsymbol{\omega}\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left\{\left[\left(y_i - Y_i\right) \quad \left(\boldsymbol{x}_i - \boldsymbol{X}_i\right)\right]\boldsymbol{\omega}^{-1}\begin{pmatrix} y_i - Y_i \\ \left[\boldsymbol{x}_i - \boldsymbol{X}_i\right]' \end{pmatrix}\right\}\right] \tag{3.33}$$

where $r = p + 1$, $E\left(\boldsymbol{x}_i\right) = E\left(\boldsymbol{X}_i + \boldsymbol{\delta}_i\right) = \boldsymbol{X}_i$ and $E\left(y_i\right) = E\left(Y_i + \boldsymbol{\varepsilon}_i\right) = Y_i$. For simplicity, let $\tau^2 = \lambda\sigma^2$, where λ is a positive constant then the log-likelihood function is,

$$L^* = -\ln K - \frac{n}{2}\ln\lambda - \left(p+1\right)n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\frac{1}{\lambda}\left(y_i - \alpha - \boldsymbol{\beta}'\boldsymbol{X}_i'\right)^2 + \left(\boldsymbol{x}_i - \boldsymbol{X}_i\right)\left(\boldsymbol{x}_i - \boldsymbol{X}_i\right)'\right] \tag{3.34}$$

where $K = \left(2\pi\right)^{\frac{m}{2}}$, $\left|\boldsymbol{\omega}\right| = \left|\omega_{11}\boldsymbol{\omega}_{22}\right| = \lambda\sigma^{2(p+1)}$ and $\boldsymbol{X}_i\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{X}_i'$.

Hence, differentiate Equation (4) with respect to α, $\boldsymbol{\beta}$, $\boldsymbol{X}_i$ and σ, and equate them to zero will yield,

$$\hat{\alpha} = \bar{y} - \frac{1}{n}\left(\sum_{i=1}^{n} \hat{\boldsymbol{X}}_i\right)\hat{\boldsymbol{\beta}} \tag{3.35}$$

$$\hat{\boldsymbol{\beta}}' = \left(\sum_{i=1}^{n} y_i \hat{\boldsymbol{X}}_i - \hat{\alpha}\sum_{i=1}^{n} \hat{\boldsymbol{X}}_i\right)\left(\sum_{i=1}^{n} \hat{\boldsymbol{X}}_i' \hat{\boldsymbol{X}}_i\right)^{-1} \tag{3.36}$$

$$\hat{\boldsymbol{X}}_i = \left[\lambda \boldsymbol{x}_i + \left(y_i - \hat{\alpha}\right)\hat{\boldsymbol{\beta}}'\right]\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} \tag{3.37}$$

$$\hat{\sigma}^2 = \frac{1}{(p+1)n}\sum_{i=1}^{n}\left[\left(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i\right)\left(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i\right)' + \frac{1}{\lambda}\left(y_i - \hat{\alpha} - \hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)^2\right] \tag{3.38}$$

To estimate $\hat{\alpha}$, substitute Equation (7) into Equation (5) and get,

$$\hat{\alpha} = \bar{y} - \frac{1}{n}\left\{\sum_{i=1}^{n}\left[\lambda \boldsymbol{x}_i + \left(y_i - \hat{\alpha}\right)\hat{\boldsymbol{\beta}}'\right]\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\right\}\hat{\boldsymbol{\beta}}$$

$$\hat{\alpha}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right) = \bar{y}\hat{\boldsymbol{\beta}}'\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right) - \frac{1}{n}\sum_{i=1}^{n}\left[\lambda \boldsymbol{x}_i + \left(y_i - \hat{\alpha}\right)\hat{\boldsymbol{\beta}}'\right]$$

$$\therefore \hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}} \tag{3.39}$$

To estimate $\hat{\boldsymbol{\beta}}$, substitute Equation (7) into Equation (6) and rearrange will get,

$$\hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}\left\{\left[\lambda \boldsymbol{x}_i + \left(y_i - \hat{\alpha}\right)\hat{\boldsymbol{\beta}}'\right]\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}\right\}'\left[\lambda \boldsymbol{x}_i + \left(y_i - \hat{\alpha}\right)\hat{\boldsymbol{\beta}}'\right]\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} = \sum_{i=1}^{n}\left(y_i - \hat{\alpha}\right)\left[\lambda \boldsymbol{x}_i + \left(y_i - \hat{\alpha}\right)\hat{\boldsymbol{\beta}}'\right]\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1}$$

$$\lambda\sum_{i=1}^{n}\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)^2 + \left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - \lambda\right)\sum_{i=1}^{n}\left(y_i - \hat{\alpha}\right)\boldsymbol{x}_i\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}\left(y_i - \hat{\alpha}\right)^2 = 0 \tag{3.40}$$

Then, substitute Equation (9) into Equation (10) and get,

$$\lambda\sum_{i=1}^{n}\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)^2 + \left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - \lambda\right)\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\boldsymbol{x}_i\hat{\boldsymbol{\beta}} - \lambda n\left(\bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}\right)^2 - \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 0$$

$$\lambda\sum_{i=1}^{n}\left(\sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right)^2 + \left(\sum_{j=1}^{p}\hat{\beta}_j^2 - \lambda\right)\sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right)\sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right] - \lambda n\left(\sum_{j=1}^{p}\bar{x}_j\hat{\beta}_j\right)^2 - \sum_{j=1}^{p}\hat{\beta}_j^2\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 0$$

To solve for $\hat{\beta}_k$,

$$\lambda\hat{\beta}_k^2\sum_{i=1}^{n} x_{ik}^2 + \left(\hat{\beta}_k^2 - \lambda\right)\sum_{i=1}^{n}\left(y_i - \bar{y}\right)x_{ik}\hat{\beta}_k - \lambda n\hat{\beta}_k^2\bar{x}_k^2 - \hat{\beta}_k^2\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 0$$

$$\therefore \hat{\beta}_k = \frac{\left(S_{yy} - \lambda S_{x_k x_k}\right) + \sqrt{\left(S_{yy} - \lambda S_{x_k x_k}\right)^2 + 4\lambda S_{x_k y}^2}}{2S_{x_k y}} \tag{3.41}$$

where $S_{yy} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$, $S_{x_k x_k} = \sum_{i=1}^{n}\left(x_{ik} - \bar{x}_k\right)^2$, and

$S_{x_k y} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)x_{ik} = \sum_{i=1}^{n} x_{ik}y_i - n\bar{x}_k\bar{y}$.

Result 2: The maximum likelihood estimators of α and $\boldsymbol{\beta}$ are approximate unbiased estimators,

$$\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} \qquad \text{and} \qquad \mathrm{E}\left(\hat{\alpha}\right) = \alpha$$

Proof:

Rewrite Equation (11),

$$\hat{\beta}_k = \theta_k + \sqrt{\theta_k^2 + \lambda}$$

where $\theta_k = \dfrac{S_{yy} - \lambda S_{x_k x_k}}{2S_{x_k y}}$, and hence, the expected value of $\hat{\beta}_k$ is,

$$E\left(\hat{\beta}_k\right) = E\left(\theta_k\right) + E\left(\sqrt{\theta_k^2 + \lambda}\right) \tag{3.42}$$

We used first order of Taylor approximations for the mean of $\theta_k\left(x_{ik}, y_i\right)$.

$$\theta_k\left(x_{ik}, y_i\right) = \theta_k\left(X_{ik} + \delta_{ik}, Y_i + \varepsilon_i\right) = \theta_k\left(X_{ik}, Y_i\right) + \delta'_{ik}\left.\frac{\partial \theta_k}{\partial x_{ik}}\right|_{x_{ik}=X_{ik}} + \varepsilon'_i\left.\frac{\partial \theta_k}{\partial y_i}\right|_{y_i=Y_i} \tag{3.43}$$

where the partial derivatives are evaluated at the mean $\left(X_{ik}, Y_i\right)$and the Equation (13) will be valid if and only if the error variances, $\sigma_{\boldsymbol{\delta}}^2$ and σ_{ε}^2 are small. Since

$$\mathrm{E}\left(\delta'_{ik}\left.\frac{\partial \theta_k}{\partial x_{ik}}\right|_{x_{ik}=X_{ik}}\right) = \sum_{k=1}^{p}\left[\left.\frac{\partial \theta_k}{\partial x_{ik}}\right|_{x_{ik}=X_{ik}} \mathrm{E}\left(\delta_{ik}\right)\right] = 0 \because \mathrm{E}\left(\boldsymbol{\delta}_i\right) = \mathrm{E}\left(\mathbf{0}\right) \rightarrow \mathrm{E}\left(\delta_{ik}\right) = 0$$

Similarly, $\mathrm{E}\left(\varepsilon'_i\left.\dfrac{\partial \theta_k}{\partial y_i}\right|_{y_i=Y_i}\right) = 0$. Therefore, Equation (13) becomes,

$$\mathrm{E}\left[\theta_k\left(x_{ik}, y_i\right)\right] = \theta_k\left(X_{ik}, Y_i\right) = \frac{S_{YY} - \lambda S_{X_k X_k}}{2S_{X_k Y}} \tag{3.44}$$

Now let $\vartheta_k\left(x_{ik}, y_i\right) = \sqrt{\theta_k^2\left(x_{ik}, y_i\right) + \lambda}$ for the second term of Equation (12). This implies, $\dfrac{\partial \theta_k}{\partial x_{ik}} = \left(\theta_k^2 + \lambda\right)^{-\frac{1}{2}}\theta_k \dfrac{\partial \theta_k}{\partial x_{ik}}$ and

$$\mathrm{E}\left[\vartheta_k\left(x_{ik}, y_i\right)\right] = \vartheta_k\left(X_{ik}, Y_i\right) = \sqrt{\theta_k^2\left(X_{ik}, Y_i\right) + \lambda} = \sqrt{\left(\frac{S_{YY} - \lambda S_{X_k X_k}}{2S_{X_k Y}}\right)^2 + \lambda} \tag{3.45}$$

Substitute Equation (14) and Equation (15) into Equation (12) will obtain,

$$\mathrm{E}\left(\hat{\beta}_k\right) = \frac{\left(S_{YY} - \lambda S_{X_k X_k}\right) + \sqrt{\left(S_{YY} - \lambda S_{X_k X_k}\right)^2 + 4\lambda S_{X_k Y}^2}}{2S_{X_k Y}} \tag{3.46}$$

Then rewrite S_{YY} and $S_{X_k Y}$ in term of $S_{X_k X_k}$, we have

$$S_{YY}\Big|_{x_{ik}=X_{ik}} = \left(\sum_{i=1}^{n} X_{ik}^2 - n\bar{X}_k^2\right)\beta_k^2 = \beta_k^2 S_{X_k X_k}$$

and,

$$S_{X_k Y}\Big|_{x_{ik}=X_{ik}} = \left(\sum_{i=1}^{n} X_{ik}^2 - n\bar{X}_k^2\right)\beta_k = \beta_k S_{X_k X_k}.$$

Then, Equation (16) becomes,

$$\mathrm{E}\left(\hat{\beta}_k\right) = \frac{\left(\beta_k^2 S_{X_k X_k} - \lambda S_{X_k X_k}\right) + \sqrt{\left(\beta_k^2 S_{X_k X_k} - \lambda S_{X_k X_k}\right)^2 + 4\lambda\left(\beta_k S_{X_k X_k}\right)^2}}{2\beta_k S_{X_k X_k}} = \beta_k$$

And from Equation (9), $\hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}$, then, $\mathrm{E}\left(\hat{\alpha}\right) = \mathrm{E}\left(\bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}\right) = \bar{y} - \bar{\boldsymbol{x}}\boldsymbol{\beta} = \alpha$.

Result 3: Given the $\mathrm{M_P ULFR}$ model, $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are consistent maximum likelihood estimators of α and $\boldsymbol{\beta}$ respectively.

Proof:

The Fisher Information Matrix (FIM) of parameters $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ is used to obtain the variance and covariance of $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$. Thus, the estimated Fisher Information Matrix (FIM) for $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ is as followed,

$$\mathbf{F} = \begin{bmatrix} \dfrac{n}{\lambda\hat{\sigma}^2} & \dfrac{1}{\lambda\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i \\ \dfrac{1}{\lambda\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i' & \dfrac{1}{\lambda\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i \end{bmatrix} = \begin{bmatrix} \mathrm{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$$

where $\mathrm{A} = \dfrac{n}{\lambda\hat{\sigma}^2}$ is a 1×1 matrix, $\mathbf{B} = \dfrac{1}{\lambda\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i$ is a $1\times p$ matrix, $\mathbf{C} = \mathbf{B}' = \dfrac{1}{\lambda\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'$ is a $p\times 1$ matrix, and $\mathbf{D} = \dfrac{1}{\lambda\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i$ is a $p\times p$ matrix are the negative expected values of the second partial derivatives for the log-likelihood function. The inverse of $\mathbf{F}$ is

$$\mathbf{F}^{-1} = \begin{bmatrix} \left(\mathrm{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & -\mathrm{A}^{-1}\mathbf{B}\left(\mathbf{D}-\mathbf{C}\mathrm{A}^{-1}\mathbf{B}\right)^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\left(\mathrm{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & \left(\mathbf{D}-\mathbf{C}\mathrm{A}^{-1}\mathbf{B}\right)^{-1} \end{bmatrix}$$

Thus, the variance and covariance of $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are $\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\alpha}\right) = \left(\mathrm{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}$, $\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{D}-\mathbf{C}\mathrm{A}^{-1}\mathbf{B}\right)^{-1}$, and $\hat{\mathrm{C}}\mathrm{ov}\left(\hat{\alpha},\hat{\boldsymbol{\beta}}\right) = -\mathbf{D}^{-1}\mathbf{C}\left(\mathrm{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}$. To show that the estimators are consistent, we have to show that the variances approach zero as n approaches infinity.

then,

$$\begin{aligned}
\lim_{n\to\infty}\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\boldsymbol{\beta}}\right) &= \lim_{n\to\infty}\lambda\hat{\sigma}^2\left[\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i - \frac{1}{n}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\right]^{-1} \\
&= \lambda\lim_{n\to\infty}\frac{1}{(p+1)n}\sum_{i=1}^{n}\left[\left(\boldsymbol{x}_i-\hat{\boldsymbol{X}}_i\right)\left(\boldsymbol{x}_i-\hat{\boldsymbol{X}}_i\right)' + \frac{1}{\lambda}\left(y_i-\hat{\alpha}-\hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)^2\right]\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1} \\
&= \lambda(0)\sum_{i=1}^{n}\left[\left(\boldsymbol{x}_i-\hat{\boldsymbol{X}}_i\right)\left(\boldsymbol{x}_i-\hat{\boldsymbol{X}}_i\right)' + \frac{1}{\lambda}\left(y_i-\hat{\alpha}-\hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)^2\right]\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1} \\
&= \mathbf{0}
\end{aligned}$$

Similarly, for $\hat{\alpha}$, we have

$$\lim_{n\to\infty}\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\alpha}\right) = \lim_{n\to\infty}\lambda\hat{\sigma}^2\left[n-\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i\right)^{-1}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i'\right)\right]^{-1} = 0$$

Therefore, both $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are consistent estimators of α and $\boldsymbol{\beta}$ respectively.

2.1 Coefficient of Determination

Consider Equations (1) and (2) and rewrite as

$$y_i = \alpha + \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i = \alpha + \boldsymbol{x}_i\boldsymbol{\beta} + \left(\varepsilon_i - \boldsymbol{\delta}_i\boldsymbol{\beta}\right) = \alpha + \boldsymbol{x}_i\boldsymbol{\beta} + E_i$$

where $E_i = \varepsilon_i - \boldsymbol{\delta}_i\boldsymbol{\beta} = y_i - \alpha - \boldsymbol{x}_i\boldsymbol{\beta}$, $i = 1,2,\cdots,n$, is the errors of the model.

Given $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are maximum likelihood estimators of α and $\boldsymbol{\beta}$ respectively, by using the idea of least square estimation, $E_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$, $i = 1,2,\cdots,n$, is the residuals of the model.

From Equation (8), $\hat{\sigma}^2 = \dfrac{SSE}{(p+1)n}$ where $SSE = \sum_{i=1}^{n}\left[(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i)(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i)' + \frac{1}{\lambda}(y_i - \hat{\alpha} - \hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}})^2\right]$, simplify using Result 1 will get,

$$SSE = \left\{\hat{\boldsymbol{\beta}}'(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}')^{-1}(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}')^{-1}\hat{\boldsymbol{\beta}} + \lambda\left[\hat{\boldsymbol{\beta}}'(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}')^{-1}(\lambda\mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}')^{-1}\hat{\boldsymbol{\beta}}\right]^2\right\}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}})^2$$

and the coefficient of determination can be defined as

$$R^2 = \frac{SSR}{S_{yy}} = 1 - \frac{SSE}{S_{yy}}$$

where $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

2.2 Simulation Study

In this study, $n = 100, 1000, 10000$ of random variables with different sizes of random errors are generated for 100000 times each in order to investigate the fitting ability of the proposed M$_P$ULFR model. The results where we set $\sigma^2 = \tau^2$ are shown in Table 1.

TABLE 1. Average coefficient of determination (R^2) of random variables with different variances at different *n*

n	Actual R^2 (without errors)	$\sigma^2 = 0.5$		$\sigma^2 = 1.0$		$\sigma^2 = 1.5$	
		M$_P$ULFR Model	MR Model	M$_P$ULFR Model	MR Model	M$_P$ULFR Model	MR Model
100	0.9991029	0.9999989	0.2440	0.9999955	0.0811	0.9999816	0.0428
1000	0.9991142	0.9999989	0.2441	0.9999956	0.0753	0.9999903	0.0354
10000	0.9991040	0.9999989	0.2438	0.9999957	0.0747	0.9999903	0.0347

It is observed that the average R^2 produced by M$_P$ULFR model are close to 1.0 for different σ^2 and *n* values and theses average R^2 are close to the actual R^2 where error terms are not added. This implies that M$_P$ULFR model poses a better fitting ability compared to MR model when both response and explanatory variables are subject to error. It is somehow interesting to investigate if M$_P$ULFR model remains outperformed MR model when applied to real data.

3.0 APPLICATION OF M$_P$ULFR MODEL IN REAL ESTATE

In this study, we utilized a cleansed data of 8741 terraced housing actual transaction records over the period of November 2008 to February 2016, from Petaling Jaya city, Selangor. These data were randomly divided into 70% training set and 30% testing set. The training set was used to train the model, and the testing set was used to validate the performance of the trained models.

The transacted housing price is regressed on nine explanatory variables using M$_P$ULFR and MR models. The explanatory variables are lot sizes (m^2), tenure types (0 for freehold and 1 for leasehold), time to expiry of lease term (assuming 200 years for freehold), terraced house types (floor numbers), number of bedrooms, main building sizes (m^2), distances to the nearest shopping mall (km), distances to the nearest supermarket (km), and transaction dates (in month) to serve as time adjustor factor. The performance of M$_P$ULFR and MR models were compared using mean square error (MSE) and coefficient of determination (R^2) obtained from the training and testing sets.

Take note that in M$_P$ULFR model, a reference housing price from houses with similar attributes is required to predict a new house price. This reference housing price is defined as the average house price of *h* nearest houses with similar attributes. In this study, we found that $h = 4$ resulted in the best performance of M$_P$ULFR model with minimum MSE.

Table 2 shows the estimated parameters for M$_P$ULFR and MR models and their performance measures. The small *p*-values (typically ≤ 0.05) imply that all variables used in this study are significant determinants of the housing prices in Petaling Jaya.

TABLE 2. Results obtained from M_PULFR Model and MR Model

Attributes	M_PULFR Model		MR Model	
	Beta value	**p-value**	**Beta value**	**p-value**
Constant	-3201.36	-	-417.13	5.45E-13
Lot Size (x_1)	9264.85	0.0000	1037.80	6.5E-241
Tenure Type (x_2)	-2094.88	0.0000	143.76	2.08E-08
Time to Expiry of Lease Term (x_3)	2621.80	0.0000	340.39	2.20E-08
Terraced House Type (x_4)	7739.19	0.0000	435.73	4.20E-38
Number of Bedrooms (x_5)	8216.49	0.0000	50.76	0.0417
Main Building Size (x_6)	6637.34	0.0000	1811.63	2.4E-272
Distance to Nearest Shopping Mall (x_7)	-9766.77	0.0000	-285.56	1.49E-78
Distance to Nearest supermarket (x_8)	-14091.05	0.0000	-112.54	3.73E-13
Transaction Date (x_9)	3365.03	0.0000	576.33	0.0000
R^2	0.9999997		0.7171	
MSE of Training Sample	1.84E-07		40856.55	
MSE of Testing Sample	28421.70		38256.35	

Both models show that lot sizes and main building sizes have a positive impact on housing prices. Buyers are willing to pay more for a larger lot and main building sizes which is also indicated in the studies from Pashardes and Savva[11] and Owusu-ansah[12]. In the study of Ooi et al.[13], freehold housings are preferable compared to leasehold housings. This finding is further supported by M_PULFR model but MR model shows a positive relationship between housing prices and leasehold housings. The contradiction may due to the existence of multicollinearity which resulted from the high correlation (see Table 3) between tenure type and time to expiry of lease term in MR model and affects the estimation of MR model[10]. In contrast, the estimated individual beta for M_PULFR model does not depend on other independent variables other than its respective independent variable which can be seen from Equation 11. This implies that the estimated beta for M_PULFR model is not affected by multicollinearity.

TABLE 3. Correlation between housing attributes

Attributes	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	1.000								
x_2	-0.132	1.000							
x_3	0.133	**-0.991**	1.000						
x_4	0.044	-0.082	0.137	1.000					
x_5	0.218	-0.176	0.198	0.293	1.000				
x_6	0.328	-0.311	0.352	0.674	0.468	1.000			
x_7	-0.090	0.310	-0.262	0.101	-0.083	-0.017	1.000		
x_8	-0.061	0.202	-0.168	0.073	-0.033	-0.090	0.327	1.000	
x_9	-0.023	0.034	-0.048	-0.024	0.061	-0.039	0.042	0.031	1.000

M_PULFR and MR models show that house buyers prefer a house with longer length of residential lease, and they willing to pay more to own a house with more bedrooms. It is also observed that the distance to the nearest amenities such as shopping mall and supermarket have negative impact to the housing prices in Petaling Jaya. This can be interpreted as the house buyers in Petaling Jaya are more willing to invest in the houses that have better accessibility and convenience. However, as Rosiers et al.[14] has pointed out, the impact of the distance to nearest amenities on housing prices is ambiguity where these attributes have contributed either repulsion or attraction effect.

It is also seen in Table 2 that the proposed M_PULFR model has a better fitting and prediction ability as compared to MR model. For the training sample, the MSE and R^2 for M_PULFR model are 1.84E-07 and close to 1.0 respectively. This is much better than the MR model where its MSE is 40856.55 and R^2 is 0.7171. Besides, M_PULFR produces smaller MSE value for testing sample as compared to MR model. This indicates that M_PULFR model is able to predict the housing prices with higher accuracy.

4.0 CONCLUDING REMARKS

In this study, we propose a new functional model for analyzing the relationship between a response variable and a set of explanatory variables. The parameters in the proposed M_PULFR model were estimated using maximum likelihood estimator assuming the ratio of error variances is known. The properties of the estimators such as unbiasedness and consistency are investigated using Taylor approximations and Fisher information matrix. The coefficient of determination was also developed to assess the performance of the model.

The proposed M_PULFR model was then applied to housing market using actual transaction data from Petaling Jaya and the results show that it has stronger fitting and predictive abilities compared to multiple regression model. The results from M_PULFR are more justifiable and interpretable because its parameters estimations are not affected by the multicollinearity of explanatory variables. However, further study is needed to develop the reliability of the proposed model by using housing data from different regions.

ACKNOWLEDGMENTS

We would like to thank the Jordan Lee & Jaafar (S) Sdn Bhd who provided the housing data from Petaling Jaya for this study.

REFERENCES

1. W. A. Fuller, *Measurement Error Models* (John Wiley, New York, 1987).
2. R. J. Adcock, The Analyst **4**, 183–184 (1877).
3. N. N. Chan and T. K. Mak, Biometrika **71**, 212–215 (1984).
4. G. M. James, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**, 411–432 (2002).
5. S. Caires and L. R. Wyatt, Journal of Agricultural, Biological, and Environmental Statistics **8**, 153–169 (2003).
6. Y. F. Chang, M. R. Omar and A. R. A. B. Syed, WSEAS Transaction on Mathematics **9**, 295–313 (2010).
   A. M. Yusof, and S. Ismail, Communications of the IBIM, (2012).
7. T. S. Ong, and Y. S. Chang, ORIC Publications **1**, 119–127 (2013).
8. K. J. Kam, S. Y. Chuah, T. S. Lim, and F. L. Ang, Pacific Rim Property Research Journal **22,** 203–216 (2016).
9. R. Matignon, *Data mining using SAS enterprise miner* (Wiley-Interscience, Hoboken, N.J., 2007), p. 100.
10. P. Pashardes, and C. S. Savva, Cyprus Economic Policy Review, 3–25 (2009).
    A. Owusu-ansah, in (African Real Estate Society Conference, Accra, Ghana, 2012), pp. 1–16.
11. J. T. L. Ooi, T. T. T. Le, and N. J. Lee, Journal of Housing Economics **26**, 126–138 (2014).
12. F. D. Rosiers, A. Lagana, M. Thériault, and M. Beaudoin, Journal of Property Valuation and Investment **14**, 41–62 (1996).