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

CHOONG WEI CHENG[a], PAN WEI YEING[b] & CHANG YUN FAH[c]

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 as a single root of a nonlinear equation. Its properties of unbiasedness, consistency and coefficient of determination were investigated using Taylor approximation and Fisher information matrix. 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 (1987) has pointed out, it is unrealistic if an explanatory variable can be measured exactly in all situations. Adcock (1877) had first studied the problem using functional model where both response and explanatory variables are subject to errors. In 1984, Chan and Mak proposed a multivariate linear functional relationship model in which error variances and covariances are unnecessarily to be homogenous. In 2002, James 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 (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.

Multiple regression (MR) model is commonly used to study and analyze the Malaysian housing market (Yusof and Ismail, 2012; Ong and Chang, 2013; Kam et al., 2016). 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 (Matignon, 2007).

In this paper, we derive a multiple un-replicated linear functional relationship (M_PULFR) model where its maximum likelihood estimators are obtained as a single root of a nonlinear equation. 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, consistency, and coefficient of determination. 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

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

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,

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

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

1. $E(\varepsilon_i) = 0$ and $E(\boldsymbol{\delta}_i) = \boldsymbol{0}$,
2. $Cov(\varepsilon_i, \varepsilon_j) = 0$ and $Cov(\boldsymbol{\delta}_i, \boldsymbol{\delta}_j) = \boldsymbol{0}$ $\quad \forall i \neq j$,
3. $Cov(\varepsilon_i, \delta_{ik}) = 0 \quad \forall \boldsymbol{i}, k$ and
4. $\varepsilon_i \sim NID(0, \omega_{11})$ and $\boldsymbol{\delta}_i \sim NID(\boldsymbol{0}, \boldsymbol{\omega}_{22})$ where $\omega_{11} = \tau^2$, and $\boldsymbol{\omega}_{22} = \sigma^2 \mathbf{I}_p$ then $\boldsymbol{\omega} = \begin{pmatrix} \omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\omega}_{22} \end{pmatrix}$ where $\boldsymbol{\omega}_{21} = \boldsymbol{\omega}'_{12} = \boldsymbol{0}$.

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

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

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

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

$$f(\boldsymbol{x}_i, y_i) = \frac{1}{(2\pi)^{\frac{r}{2}} |\boldsymbol{\omega}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left\{ \left[(y_i - Y_i) \quad (\boldsymbol{x}_i - \boldsymbol{X}_i) \right] \boldsymbol{\omega}^{-1} \begin{pmatrix} y_i - Y_i \\ \left[\boldsymbol{x}_i - \boldsymbol{X}_i\right]' \end{pmatrix} \right\} \right] \tag{3}$$

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

$$L^* = -\ln K - \frac{n}{2}\ln\lambda - (p+1)n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\frac{1}{\lambda}\left(y_i - \alpha - \boldsymbol{\beta}'\boldsymbol{X}'_i\right)^2 + \left(\boldsymbol{x}_i - \boldsymbol{X}_i\right)\left(\boldsymbol{x}_i - \boldsymbol{X}_i\right)' \right] \tag{4}$$

where $K = (2\pi)^{\frac{m}{2}}$, $|\boldsymbol{\omega}| = |\omega_{11}\boldsymbol{\omega}_{22}| = \lambda\sigma^{2(p+1)}$ 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,

$$\hat{\alpha} = \bar{y} - \frac{1}{n}\left(\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\right)\hat{\boldsymbol{\beta}} \tag{5}$$

$$\hat{\boldsymbol{\beta}}' = \left(\sum_{i=1}^{n} y_i \hat{\boldsymbol{X}}_i - \hat{\alpha}\sum_{i=1}^{n} \hat{\boldsymbol{X}}_i\right)\left(\sum_{i=1}^{n} \hat{\boldsymbol{X}}_i' \hat{\boldsymbol{X}}_i\right)^{-1} \tag{6}$$

$$\hat{\boldsymbol{X}}_i = \left[\lambda \boldsymbol{x}_i + (y_i - \hat{\alpha})\hat{\boldsymbol{\beta}}'\right]\left(\lambda \mathbf{I} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'\right)^{-1} \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{(p+1)n}\sum_{i=1}^{n}\left[\left(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i\right)\left(\boldsymbol{x}_i - \hat{\boldsymbol{X}}_i\right)' + \frac{1}{\lambda}\left(y_i - \hat{\alpha} - \hat{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}\right)^2\right] \tag{8}$$

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

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

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

$$\therefore \hat{\alpha} = \bar{y} - \bar{\mathbf{x}}\hat{\boldsymbol{\beta}} \tag{9}$$

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

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

$$\lambda\sum_{i=1}^{n}\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)^2 + \left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - \lambda\right)\sum_{i=1}^{n}(y_i - \hat{\alpha})\boldsymbol{x}_i\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\sum_{i=1}^{n}(y_i - \hat{\alpha})^2 = 0 \tag{10}$$

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

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

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

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

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

$$\therefore \hat{\beta}_k = \frac{\left(S_{yy} - \lambda S_{x_k x_k}\right) + \sqrt{\left(S_{yy} - \lambda S_{x_k x_k}\right)^2 + 4\lambda S_{x_k y}^2}}{2S_{x_k y}} \tag{11}$$

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

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

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

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,

$$\mathrm{E}\left(\hat{\beta}_k\right) = \mathrm{E}\left(\theta_k\right) + \mathrm{E}\left(\sqrt{\theta_k^2 + \lambda}\right) \tag{12}$$

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

$$\theta_k\left(x_{ik}, y_i\right) = \theta_k\left(X_{ik} + \delta_{ik}, Y_i + \varepsilon_i\right) = \theta_k\left(X_{ik}, Y_i\right) + \delta'_{ik} \left.\frac{\partial \theta_k}{\partial x_{ik}}\right|_{x_{ik} = X_{ik}} + \varepsilon'_i \left.\frac{\partial \theta_k}{\partial y_i}\right|_{y_i = Y_i} \tag{13}$$

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

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

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

$$\mathrm{E}\left[\theta_k\left(x_{ik}, y_i\right)\right] = \theta_k\left(X_{ik}, Y_i\right) = \frac{S_{YY} - \lambda S_{X_k X_k}}{2S_{X_k Y}} \tag{14}$$

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

$$\mathrm{E}\left[\vartheta_k\left(x_{ik}, y_i\right)\right] = \vartheta_k\left(X_{ik}, Y_i\right) = \sqrt{\theta_k^2\left(X_{ik}, Y_i\right) + \lambda} = \sqrt{\left(\frac{S_{YY} - \lambda S_{X_k X_k}}{2S_{X_k Y}}\right)^2 + \lambda} \tag{15}$$

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

$$\mathrm{E}\left(\hat{\beta}_k\right) = \frac{\left(S_{YY} - \lambda S_{X_k X_k}\right) + \sqrt{\left(S_{YY} - \lambda S_{X_k X_k}\right)^2 + 4\lambda S_{X_k Y}^2}}{2S_{X_k Y}} \tag{16}$$

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

$$\mathrm{E}\left(\hat{\alpha}\right) = \mathrm{E}\left(\bar{y} - \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}\right) = \bar{y} - \bar{\mathbf{x}}\boldsymbol{\beta} = \alpha$$

Result 3: Given the 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,

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

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 $p\times 1$ 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

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

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

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

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

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

$$= 0$$

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

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

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

From Equation (8), $\hat{\sigma}^2 = \frac{SSE}{(p+1)n}$ 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,

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

and the coefficient of determination can be defined as

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

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

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 1 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 1. Results obtained from M$_P$ULFR Model and MR Model

Attributes	M$_P$ULFR Model		MR Model	
	Beta value	**p-value**	**Beta value**	**p-value**
Constant	-3201.36	-	-417.13	5.45E-13
Lot Size	9264.85	0.0000	1037.80	6.5E-241
Tenure Type	-2094.88	0.0000	143.76	2.08E-08
Time to Expiry of Lease Term	2621.80	0.0000	340.39	2.20E-08
Terraced House Type	7739.19	0.0000	435.73	4.20E-38
Number of Bedrooms	8216.49	0.0000	50.76	0.0417
Main Building Size	6637.34	0.0000	1811.63	2.4E-272
Distance to Nearest Shopping Mall	-9766.77	0.0000	-285.56	1.49E-78
Distance to Nearest supermarket	-14091.05	0.0000	-112.54	3.73E-13
Transaction Date	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 (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 contradiction may due to the existence of multicollinearity in MR model and affects the estimation of MR model.

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. (1996) 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 1 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. A. M. Yusof, and S. Ismail, *Multiple regressions in analysing house price variations* (Communications of The IBIM, 2012).
2. A. Owusu-ansah, A. *Examination of the determinants of housing values in urban Ghana and implications for policy makers* (African Real Estate Society Conference, Accra, Ghana, 2012), pp. 1–16.
3. F. D. Rosiers, A. Lagana, M. Thériault, and M. Beaudoin, *Shopping centres and house values: an empirical investigation* (Journal of Property Valuation and Investment, 1996), vol. 14, pp. 41–62.
4. G. M. James, *Generalized linear models with functional predictors* (Journal of the Royal Statistical Society, 2002), pp. 411–432.
5. J. T. L. Ooi, T. T. T. Le, and N. J. Lee, *The impact of construction quality on house prices* (Journal of Housing Economics, 2014), pp. 126–138.

6. K. J. Kam, S. Y. Chuah, T. S. Lim, and F. L. Ang, *Modelling of property market: the structural and locational attributes towards Malaysian properties* (Pacific Rim Property Research Journal, 2016), pp. 203–216.
7. N. N. Chan and T. K. Mak, *Heteroscedastic errors in a linear functional relationship* (Biometrika, 1984), vol. 71, pp. 212–215.
8. P. Pashardes, and C. S. Savva, *Factors affecting house prices in Cyprus: 1988-2008*. (Cyprus Economic Policy Review, 2009), pp. 3–25.
9. R. J. Adcock, *Note on the method of least squares* (Analyst, 1877), vol. 4, pp. 183–184.
10. R. Matignon, "Chapter 2 Explore Nodes," in *Data mining using SAS enterprise miner* (Wiley-Interscience, 2007), pp. 100.
11. S. Caires and L. R. Wyatt, *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, 2003), vol. 8, pp. 153–169.
12. T. S. Ong, and Y. S. Chang, *Macroeconomic determinants of Malaysian housing market* (ORIC Publications, 2013). vol. 1, pp. 119–127.
13. Y. F. Chang, M. R. Omar and A. R. A. B. Syed, *Multidimensional Unreplicated Linear Functional Relationship Model with single slope and its coefficient of determination* (WSEAS Transaction on Mathematics, 2010), vol. 9, pp. 295–313.
14. W. A. Fuller, *Measurement error models* (John Wiley, New York, 1987).