International Journal of Geo-Information

MDPI

Article

Total Least Squares Estimation in Hedonic House Price Models

Wenxi Zhan [1,2], Yu Hu [1], Wenxian Zeng [1,*], Xing Fang [1], Xionghua Kang [1] and Dawei Li [1,2]

[1] School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Rd., Wuhan 430079, China
[2] Hubei Luojia Laboratory, Wuhan University, 129 Luoyu Rd., Wuhan 430079, China
* Correspondence: wxzeng@sgg.whu.edu.cn

Abstract: In real estate valuation using the Hedonic Price Model (HPM) estimated via Ordinary Least Squares (OLS) regression, subjectivity and measurement errors in the independent variables violate the Gauss–Markov theorem assumption of a non-random coefficient matrix, leading to biased parameter estimates and incorrect precision assessments. In this contribution, the Errors-in-Variables model equipped with Total Least Squares (TLS) estimation is proposed to address these issues. It fully considers random errors in both dependent and independent variables. An iterative algorithm is provided, and posterior accuracy estimates are provided to validate its effectiveness. Monte Carlo simulations demonstrate that TLS provides more accurate solutions than OLS, significantly improving the root mean square error by over 70%. Empirical experiments on datasets from Boston and Wuhan further confirm the superior performance of TLS, which consistently yields a higher coefficient of determination and a lower posterior variance factor, which shows its more substantial explanatory power for the data. Moreover, TLS shows comparable or slightly superior performance in terms of prediction accuracy. These results make it a compelling and practical method to enhance the HPM.

Keywords: hedonic price model; real estate; mass appraisal; total least squares; errors-in-variables

Citation: Zhan, W.; Hu, Y.; Zeng, W.; Fang, X.; Kang, X.; Li, D. Total Least Squares Estimation in Hedonic House Price Models. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 159. https://doi.org/10.3390/ijgi13050159

Academic Editors: Wolfgang Kainz and Jamal Jokar Arsanjani

Received: 10 March 2024
Revised: 28 April 2024
Accepted: 7 May 2024
Published: 8 May 2024

1. Introduction

Real estate holds a pivotal position in the national economy and household investment, necessitating a thorough analysis of the influencing factors and an accurate assessment of the value of real estate. It is widely recognized that the market approach is commonly applied in real estate valuation. The Hedonic Price Model (HPM) extends from the market approach [1,2], grounded in supply–demand theory, and uses regression analysis to relate characteristics to transaction prices. Studying the HPM in depth for parameter estimation and housing price prediction is crucial in facilitating well-informed decision-making, ensuring the integrity of real estate transactions, and conducting precise tax assessments [3].

Machine learning (ML) algorithms are increasingly used to analyze the housing market [4–6]. Much of the literature [7–9] recognizes the undeniable predictive power of ML algorithms; however, they also present a critical limitation due to their 'black box' nature (i.e., lack of model interpretability). It is difficult to discern the role that individual parameters play in value variation or to numerically define the causal relationships between prices and the characteristics of assets [10]. However, this is not the case with the HPM, which facilitates both parameter estimation and interpretation. Pérez-Rave et al. [11] point out that understanding today's complex housing market requires a thorough analysis of the relevant variables and the estimation of their significant impacts. Therefore, our work focuses on expanding the HPM to achieve a more comprehensive and thorough analysis of housing prices.

1.1. Hedonic Price Method

Currently, the research on the HPM can be delineated into two levels: practical studies and methodological exploration. For practical research, the HPM serves three main purposes. (1) It is used to construct quality-adjusted house price indexes [12,13] to broadly

track property price movements. Numerous countries and organizations have developed their own hedonic indexes, such as the hedonic house price index in the US Census Bureau and the Halifax and Nationwide indexes in the UK. (2) It is used to provide automated valuations (or general appraisals) of properties [14], which is also a critical step in property tax determination in some countries (e.g., the United States and Germany). (3) It is used to explain house price variations or determine the impact of certain characteristics on houses, revealing house price drivers and mirroring the real estate market development stages [2]. Housing price drivers can be roughly divided into two categories. The first category focuses on the physical characteristics, such as intrinsic characteristics [15,16] and building properties [17]. The second category examines the impact of public goods on housing prices, such as school quality [18,19], public transportation [20,21], and open spaces [22–24].

For methodological research, the classic HPM is estimated by Least Squares (LS) regression [25]. The observed dependent variable (transaction price) is expressed as the linear combination of independent variables, i.e., the implicit prices associated with the structural, neighborhood, and location characteristics of the real estate [26]. As the parameters of interest are estimated based on the formed mathematical model and a certain optimization criterion, the original method has been intensively extended in two aspects. (1) To enhance mathematical modeling, (i) the functional relation has been improved by applying the semiparametric model [27], Box–Cox model [28], and log–log model [29], which aim at improving both the goodness of fit and interpretability of the model; (ii) the stochastic model has been refined by considering the spatial effects via the Spatial Autoregressive (SAR) model [30–32], the Spatial Error Model (SEM) [33], and the Geographically Weighted Regression (GWR) [34–36]. (2) For the optimization criterion, (i) regularization (or penalized) methods, such as ridge regression and the Least Absolute Shrinkage and Selection Operator (LASSO), are utilized to overcome multicollinearity, which may be caused by the dependency of the independent variables [37–39]; (ii) the prior information can be incorporated to consider the advice of experts via Bayesian estimation [40,41] or inequality-restricted least squares [39] by using an informative prior distribution, where the hyperparameters are set according to expert knowledge of the characteristics of the model parameter; (iii) robust methods, such as the least median of squared residuals [42] and normalized interval regression [43], have been utilized to detect outliers and enhance the reliability.

1.2. Total Least Squares Estimation

Though previous research has explored the HPM from various perspectives, an aspect remains to be discussed. Real estate transactions lack transparency due to limited access to crucial details like actual transaction prices and real estate facilities on public websites [44,45]. Additionally, many property characteristics are qualitative and subjective, such as views or architectural styles [46,47], and measurement errors in recording key characteristics like the environmental quality of the house all contribute to potential "Errors-in-Variables" (EIV). However, previous studies often ignore the errors in the coefficient matrix, which can result in biased estimation and incorrect accuracy assessment. Anselin and Lozano-Gracia [48] explore EIV using instrumental variables in a two-step regression approach. Unlike this method, Total Least Squares (TLS) directly minimizes the sum of squared orthogonal distances from the data points to the model, thereby simultaneously addressing errors in both dependent and independent variables.

The terminology of TLS was first coined by Golub and van Loan [49], who also proposed the widely used algorithm based on Singular Value Decomposition (SVD). Nowadays, it finds application in the fields of signal processing [50,51], image processing [52,53], and applied geodesy [54–59], to name a few. Its aim is to minimize the sum of the squares of all random errors in the model. Though statistically appealing, it is more complicated to compute than LS due to the nonlinearity caused by the interaction of the random effects with the fixed effects of unknowns. The methods to obtain a TLS solution can be mainly

summarized into two categories: the SVD-based methods and the iterative methods. Inherited from SVD, the former method is numerically effective; however, it only works with some restricted structures of the stochastic model. For more details, one can refer to [50,60]. The latter regards it as a nonlinear constrained optimization problem and solves it via linearization and iteration, e.g., the Gauss–Newton method and Newton method [55,57,61,62]. In contrast to the former method, it is more general in terms of the model setup.

To give the reader an impression of the basic idea of TLS, we here consider a simple example published by Wooldridge [63] (p. 153). It tries to investigate the relationship between the logarithm of the house price and the logarithm of the distance from the house to an incinerator. Therefore, the model consists of two parameters, i.e., the intercept parameter and the distance parameter. We select the first 100 sampled points and fit the data with LS and TLS, respectively. The estimated results are shown in Figure 1, from which we can see that LS only adjusts the data in the vertical direction. In contrast, the direction of the residuals produced by TLS is orthogonal to the fitted line. The essential reason is that TLS considers the random errors in the distance and adjusts such quantities in the estimation. This is why TLS is called orthogonal regression in some studies.

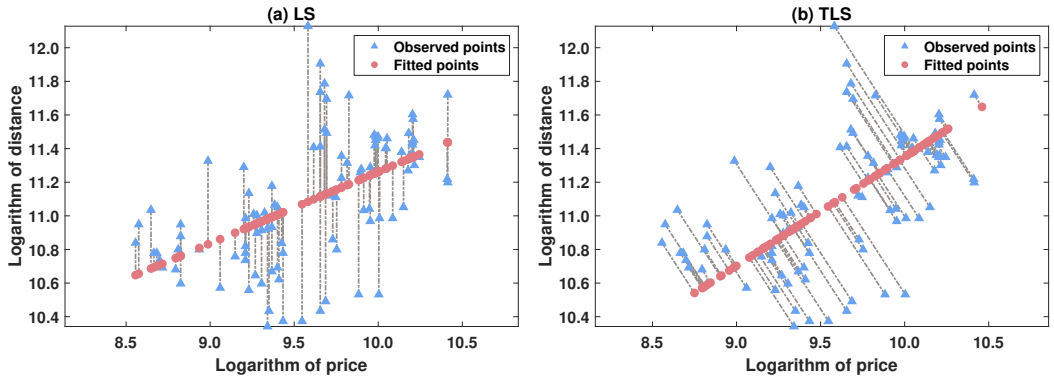

Figure 1. Fitting results for the house price data in [63]: (**a**) LS results; (**b**) TLS results.

The main objective of this contribution is first to apply TLS to hedonic price problems, taking into account errors in the dependent and independent variables, thereby enhancing the reasonable estimation of the hedonic parameters. Our approach leverages simulation experiments and two empirical datasets to comprehensively demonstrate its superiority in real estate evaluation.

1.3. Outline of the Paper

The rest of the paper is organized as follows. In Section 2, we first review LS estimation and then present TLS estimation in the HPM. In Section 3, Monte Carlo simulations are presented to show the advantages of our method. In Section 4, the Boston dataset is analyzed. In Section 5, we analyze a dataset collected in Wuhan. In Section 6, we discuss the three sets of experimental results. Finally, the conclusions are drawn in Section 7.

2. Method for the HPM

The HPM shows the relationship between a dependent variable (the house transaction price of the i-th sample) and independent variables (the selected m house characteristics for the i-th sample). The regression equation can be written as follows:

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} \quad i = 1, 2, \cdots, n, \tag{1}$$

where y_i is the dependent variable (i.e., transaction price), $x_{i,1}, \cdots, x_{im}$ are independent variables (i.e., house characteristics), β_0 is the unknown intercept parameter, and $\beta_1, \cdots, \beta_m$ are unknown hedonic parameters. The approximation symbol "$\approx$" is used here since such a relationship does not exactly hold for real collected data. The choice of estimation method

depends on how we specify the stochastic information of this equation. If only the random errors of the dependent variables (i.e., y_i) are considered, it becomes a linear model and LS is applied; if the random errors of the independent variables (i.e., $x_{i,1}, \cdots, x_{im}$) are additionally considered, it becomes an EIV model and TLS is applied. Next, these two methods are introduced.

2.1. Least Squares Regression

In most cases, we only consider the errors in the dependent variables y_i. Collecting the equations of all n sampled points yields the well-known linear Gauss–Markov model

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \mathbf{e}_\mathbf{y} \quad \mathbf{e}_\mathbf{y} \sim (\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{y} = \sigma^2 \mathbf{Q}_\mathbf{y}) \tag{2}$$

with

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \quad \mathbf{e}_\mathbf{y} = \begin{bmatrix} e_{y1} \\ e_{y2} \\ \vdots \\ e_{yn} \end{bmatrix}, \tag{3}$$

where $\mathbf{y}$ is the n-vector of observations; $\mathbf{A}$ is the design matrix of order $n \times (m+1)$ with rank $(m+1)$; $\mathbf{e}_\mathbf{y}$ is the n-vector of random errors; $\boldsymbol{\beta}$ is the $(m+1)$-vector of the unknown parameters to be estimated; $\boldsymbol{\Sigma}_\mathbf{y}$ is the symmetric positive definite covariance matrix of order $n \times n$; $\mathbf{Q}_\mathbf{y}$ is the cofactor matrix; and σ^2 is the (unknown) variance factor (VF).

By minimizing $\mathbf{e}_\mathbf{y}^\mathsf{T} \mathbf{Q}_\mathbf{y}^{-1} \mathbf{e}_\mathbf{y}$, the LS estimator reads

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = (\mathbf{A}^\mathsf{T} \mathbf{Q}_\mathbf{y}^{-1} \mathbf{A})^{-1} \mathbf{A}^\mathsf{T} \mathbf{Q}_\mathbf{y}^{-1} \mathbf{y}. \tag{4}$$

The residual vector reads

$$\hat{\mathbf{e}}_{\mathbf{y},\text{LS}} = \mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}} = \left(\mathbf{I}_n - (\mathbf{A}^\mathsf{T} \mathbf{Q}_\mathbf{y}^{-1} \mathbf{A})^{-1} \mathbf{A}^\mathsf{T} \mathbf{Q}_\mathbf{y}^{-1}\right) \mathbf{y}. \tag{5}$$

Utilizing the error propagation law yields the cofactor matrix of the LS estimator, i.e.,

$$\mathbf{Q}_{\text{LS}} = (\mathbf{A}^\mathsf{T} \mathbf{Q}_\mathbf{y}^{-1} \mathbf{A})^{-1}. \tag{6}$$

The (weighted) Sum of Squared Errors (SSE) reads

$$\text{SSE}_{\text{LS}} = \hat{\mathbf{e}}_{\mathbf{y},\text{LS}}^\mathsf{T} \mathbf{Q}_\mathbf{y}^{-1} \hat{\mathbf{e}}_{\mathbf{y},\text{LS}}. \tag{7}$$

Since the degree of freedom of the model is $(n - m - 1)$, the a posteriori square root of VF can be evaluated as

$$\hat{\sigma}_{\text{LS}} = \sqrt{\text{SSE}_{\text{LS}}/(n - m - 1)}. \tag{8}$$

Combining (6) and (8), the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{LS}}$ reads

$$\widehat{\boldsymbol{\Sigma}}_{\text{LS}} = \hat{\sigma}_{\text{LS}}^2 \cdot \mathbf{Q}_{\text{LS}} = \hat{\sigma}_{\text{LS}}^2 (\mathbf{A}^\mathsf{T} \mathbf{Q}_\mathbf{y}^{-1} \mathbf{A})^{-1}. \tag{9}$$

If the cofactor matrix is chosen as $\mathbf{Q}_\mathbf{y} = \mathbf{I}_n$, all formulations degrade into those for the Ordinary Least Squares (OLS) regression. However, the general covariance matrix $\boldsymbol{\Sigma}$ enables us to consider various model setups, such as the spatial correlations [64].

2.2. Total Least Squares Estimation

In LS estimation, the design matrix $\mathbf{A}$ is assumed to be non-stochastic (i.e., errorless). However, due to subjectivity (e.g., the quality of decoration and architectural style) and measurement uncertainty (e.g., the commercial service level of the house), some dependent variables cannot be exactly evaluated. The ignorance of the stochasticity of the design matrix can cause the LS regression results to be less accurate. Therefore, we address the

issue of endogeneity from the specific perspective of EIV by introducing a random matrix $\mathbf{E}$ of order $n \times (m+1)$ to the linear model (2), i.e.,

$$\mathbf{y} = (\mathbf{A} - \mathbf{E})\boldsymbol{\beta} + \mathbf{e}_\mathbf{y}, \tag{10}$$

with

$$\mathbf{e} = \begin{bmatrix} \mathbf{e}_\mathbf{y} \\ \mathbf{e}_A \end{bmatrix} \sim \left(\mathbf{0}, \boldsymbol{\Sigma} = \sigma^2 \mathbf{Q}\right) \qquad \mathbf{Q} = \begin{bmatrix} \mathbf{Q}_\mathbf{y} & \mathbf{0} \\ \mathbf{0} & \kappa^2 \mathbf{Q}_A \end{bmatrix}, \tag{11}$$

where $\mathbf{e}_A = \mathrm{vec}(\mathbf{E})$ is the $(nm+n)$-vector of random errors; $\mathrm{vec}(\cdot)$ represents the vectorization operator that stacks the columns of the argument; $\mathbf{Q}_A$ is the symmetric positive semi-definite matrix of order $(nm+n) \times (nm+n)$; $\kappa^2 = \sigma_A^2/\sigma^2$ is the VF ratio; and σ_A^2 represents the VF of the design matrix. In practice, the cofactor matrix $\mathbf{Q}_A$ is usually determined by considering factors such as the reliability of the data sources, the collection methods, and the nature of the variables themselves.

Since we have the relationship $\mathbf{E}\boldsymbol{\beta} = (\boldsymbol{\beta}^\mathsf{T} \otimes \mathbf{I}_n)\mathbf{e}_A$ [65], the model (10) can be reformulated as

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{e} \quad \mathbf{e} \sim (\mathbf{0}, \boldsymbol{\Sigma} = \sigma^2 \mathbf{Q}), \tag{12}$$

where $\otimes$ represents the Kronecker operator and $\mathbf{B} = [\mathbf{I}_n \quad -(\boldsymbol{\beta}^\mathsf{T} \otimes \mathbf{I}_n)]$. Based on this, the general TLS objective reads

$$\min \mathbf{e}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{e} \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{e}. \tag{13}$$

Note that if the dependent parameters are fixed (or non-stochastic), the corresponding blocks of $\mathbf{Q}$ are zero matrices, which leads to its singularity, i.e., $\mathrm{rank}(\mathbf{Q}) < (n+1)(m+1)$. Therefore, strictly speaking, the regular inverse $\mathbf{Q}^{-1}$ does not exist since we at least have the intercept parameter β_0. However, we still use $\mathbf{Q}^{-1}$ to establish the objective since the singularity caused by such a structure does not affect the final estimate. A similar treatment can be found in, e.g., [57]. Unlike the traditional TLS method based on SVD, we can see that the structure of the model has been considered by forming the covariance matrix $\boldsymbol{\Sigma}$, which is automatically kept in the estimation.

With the Lagrangian method, we can obtain the iterative solution forms. For simplicity, the derivations are placed in the Appendix A. The TLS estimator reads

$$\hat{\boldsymbol{\beta}}_{\text{TLS}} = (\widehat{\mathbf{A}}^\mathsf{T} \widehat{\mathbf{Q}}_\mathbf{B}^{-1} \widehat{\mathbf{A}})^{-1} \widehat{\mathbf{A}}^\mathsf{T} \widehat{\mathbf{Q}}_\mathbf{B}^{-1} (\mathbf{y} - \widehat{\mathbf{E}}\hat{\boldsymbol{\beta}}), \tag{14}$$

where $\widehat{\mathbf{A}} = \mathbf{A} - \widehat{\mathbf{E}}$ and $\widehat{\mathbf{Q}}_\mathbf{B} = \widehat{\mathbf{B}}\mathbf{Q}\widehat{\mathbf{B}}^\mathsf{T}$. The residual vector is

$$\hat{\mathbf{e}} = \mathbf{Q}\hat{\mathbf{B}}^\mathsf{T} \widehat{\mathbf{Q}}_\mathbf{B}^{-1} (\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}}). \tag{15}$$

Reshaping the residual vector (15), we can obtain

$$\begin{aligned} [\,\widehat{\mathbf{E}} \quad \hat{\mathbf{e}}_\mathbf{y}\,] &= \mathrm{vec}_{n,m+2}^{-1} \hat{\mathbf{e}} \\ &= \left((\mathrm{vec}\mathbf{I}_{m+2})^\mathsf{T} \otimes \mathbf{I}_n\right)(\mathbf{I}_{m+2} \otimes \hat{\mathbf{e}}), \end{aligned} \tag{16}$$

where $\mathrm{vec}_{n,m+2}^{-1}(\cdot)$ is the inverse operator of $\mathrm{vec}(\cdot)$, i.e., restructuring an $n \times (m+2)$ matrix from an $(nm+2n)$-vector.

Since the residuals are determined by the unknowns, the final estimate should be obtained by iterating these expressions. The steps of the TLS estimation can be summarized as Algorithm 1. Here, $\|\cdot\|_2^2 = (\cdot)^\mathsf{T}(\cdot)$ is the squared Euclidean norm and ε is a very small positive constant and is chosen as 10^{-8} in this paper. In this paper, all computations are implemented with `MATLAB`. For the computational aspects of TLS, one can refer to Fang [57,62] for a Newton-type iteration.

Algorithm 1 Total Least Squares Estimation

Require: $\mathbf{y}$, $\mathbf{A}$ and $\mathbf{\Sigma}$

Obtain the initial values $\boldsymbol{\beta}^0$ by the LS estimation and set $i = 0$
Set $\mathbf{E}_0 \leftarrow \mathbf{0}$
repeat
  Construct $\mathbf{B}_0 \leftarrow [\mathbf{I}_n \quad -(\boldsymbol{\beta}_0^{\mathsf{T}} \otimes \mathbf{I}_n)]$
  $\hat{\boldsymbol{\beta}} \leftarrow (\mathbf{A}_0^{\mathsf{T}}\mathbf{Q}_{\mathbf{B}0}^{-1}\mathbf{A}_0)^{-1}\mathbf{A}_0^{\mathsf{T}}\mathbf{Q}_{\mathbf{B}0}^{-1}(\mathbf{y} - \mathbf{E}_0\boldsymbol{\beta}_0)$
  $\hat{\mathbf{e}} \leftarrow \mathbf{Q}\mathbf{B}_0^{\mathsf{T}}\mathbf{Q}_{\mathbf{B}0}^{-1}(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}})$
  $[\,\widehat{\mathbf{E}} \quad \hat{\mathbf{e}}_{\mathbf{y}}\,] \leftarrow \mathrm{vec}_{n,m+2}^{-1}\hat{\mathbf{e}}$
  Update the coefficient matrix $\mathbf{A}_0 \leftarrow \mathbf{A} - \widehat{\mathbf{E}}$
  Update the residual matrix $\mathbf{E}_0 \leftarrow \widehat{\mathbf{E}}$
  Update the parameter vector $\boldsymbol{\beta}_0 \leftarrow \hat{\boldsymbol{\beta}}$
  Update $i = i + 1$
until $\|\delta\| < \varepsilon$ (δ is the difference in parameter between two successive iterations)
Calculate $\hat{\sigma}_{\text{TLS}}$ and $\widehat{\mathbf{\Sigma}}_{\text{TLS}}$

Next, we consider how to perform the precision assessment in the TLS estimation. According to [66], the EIV model can be formed as

$$\mathbf{y}_c = \overline{\mathbf{A}}\boldsymbol{\beta} + \mathbf{e}_c \quad \mathbf{e}_c \sim (\mathbf{0}, \mathbf{\Sigma}_{\mathbf{B}} = \mathbf{B}\mathbf{\Sigma}\mathbf{B}^{\mathsf{T}}), \tag{17}$$

where $\mathbf{y}_c = \mathbf{y} - \mathbf{E}\boldsymbol{\beta}$, $\overline{\mathbf{A}} = \mathbf{A} - \mathbf{E}$ and $\mathbf{e}_c = \mathbf{e}_{\mathbf{y}} - \mathbf{E}\boldsymbol{\beta}$. Therefore, we can have the element as

$$e_{ci} = e_{yi} - \mathbf{e}_{\mathbf{a},i}^{\mathsf{T}}\boldsymbol{\beta}, \tag{18}$$

where e_{ci} and e_{yi} are the i-th elements of $\mathbf{e}_c$ and $\mathbf{e}_y$, respectively; and $\mathbf{e}_{\mathbf{a}_i}^{\mathsf{T}}$ is the i-th row vector of $\mathbf{E}$. From the definition, we cannot immediately judge which one of $|e_{ci}|$ and $|e_{yi}|$ is greater as the sign of the product $\mathbf{e}_{\mathbf{a},i}^{\mathsf{T}}\boldsymbol{\beta}$ cannot be determined. This is why $\mathbf{e}_c$ is called the "total error" by [66] and has been employed for statistical inference, such as hypothesis testing.

Taking it as the model with a fixed coefficient matrix, we can develop the LS ensemble formulations for TLS, such as precision assessment and hypothesis testing. Therefore, we can have the cofactor matrix as

$$\mathbf{Q}_{\text{TLS}} = (\widehat{\mathbf{A}}^{\mathsf{T}}\widehat{\mathbf{Q}}_{\mathbf{B}}^{-1}\widehat{\mathbf{A}})^{-1}. \tag{19}$$

Similarly, the SSE reads

$$\begin{aligned} \mathrm{SSE}_{\text{TLS}} &= \hat{\mathbf{e}}_c^{\mathsf{T}}\widehat{\mathbf{Q}}_{\mathbf{B}}^{-1}\hat{\mathbf{e}}_c \\ &= (\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}}_{\text{TLS}})^{\mathsf{T}}\widehat{\mathbf{Q}}_{\mathbf{B}}^{-1}(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}}_{\text{TLS}}), \end{aligned} \tag{20}$$

which leads to the posterior square root of the VF

$$\hat{\sigma}_{\text{TLS}} = \sqrt{\mathrm{SSE}_{\text{TLS}}/(n - m - 1)} \tag{21}$$

and the posterior covariance matrix of the parameters

$$\widehat{\mathbf{\Sigma}}_{\text{TLS}} = \hat{\sigma}_{\text{TLS}}^2 \cdot \mathbf{Q}_{\text{TLS}} = \hat{\sigma}_{\text{TLS}}^2(\widehat{\mathbf{A}}^{\mathsf{T}}\widehat{\mathbf{Q}}_{\mathbf{B}}^{-1}\widehat{\mathbf{A}})^{-1}. \tag{22}$$

2.3. Coefficient of Determination

The coefficient of determination (R^2) is a popular measure of the goodness of fit for linear models. To assess the explanatory power of the effects in $\beta_1, \cdots, \beta_m$, the null model can be formed as

$$\mathbf{y} = \mathbf{1}_n\beta_0 + \mathbf{e}_\mathbf{y} \quad \mathbf{e}_\mathbf{y} \sim (\mathbf{0}, \sigma^2\mathbf{Q}_\mathbf{y}) \tag{23}$$

by dropping all these effects in the relation (2) or (10). By performing LS estimation, we have

$$\mathrm{SSE}_{\mathbf{1}_n} = (\mathbf{C}\mathbf{y})^\mathsf{T}\mathbf{Q}_\mathbf{y}^{-1}(\mathbf{C}\mathbf{y}) = \mathbf{y}^\mathsf{T}\mathbf{Q}_\mathbf{y}^{-1}\mathbf{C}\mathbf{y} \tag{24}$$

with the projector $\mathbf{C} = \mathbf{I}_n - \mathbf{1}_n(\mathbf{1}_n^\mathsf{T}\mathbf{Q}_\mathbf{y}^{-1}\mathbf{1}_n)^{-1}\mathbf{1}_n^\mathsf{T}\mathbf{Q}_\mathbf{y}^{-1}$, which is also called the (weighted) total sum of squares.

Therefore, we can have the coefficient of determination as

$$R_i^2 = 1 - \mathrm{SSE}_i/\mathrm{SSE}_{\mathbf{1}_n} \quad i = \mathrm{LS}, \mathrm{TLS}. \tag{25}$$

From the definition, R^2 ranges from 0 to 1 and indicates how much the dependent variable's variability (quantified by statistical measures like variance or the standard deviation) can be explained by the independent variables. For example, $R^2 = 0.66$ suggests that 66% of the variation in the dependent variable is captured or explained by the model, and the remaining 34% of the variation is caused by factors not included or inherent randomness.

However, the coefficient (25) automatically increases if an extra independent variable is added to the model. To address such a limitation, its adjusted version is formed as

$$R^2_{\mathrm{adj},i} = 1 - \frac{\mathrm{SSE}_i/(n-m-1)}{\mathrm{SSE}_{\mathbf{1}_n}/(n-1)} = 1 - \frac{(1-R_i^2)(n-1)}{(n-m-1)} \quad i = \mathrm{LS}, \mathrm{TLS} \tag{26}$$

by considering the degree of freedom of the model, which ensures that the inclusion of additional variables is justifiably reflected in the overall explanatory power of the model. The adjusted one R^2_{adj} is usually less than the original one R^2.

Further, the F-test statistic can be expressed with R^2 as

$$F_i = \frac{R_i^2/m}{(1-R_i^2)(n-m-1)} \quad i = \mathrm{LS}, \mathrm{TLS}, \tag{27}$$

which evaluates the overall significance of the independent variables within the HPM when estimating house prices. If the result is significant, it underscores the collective impact of these variables on the housing values, affirming the statistical significance of the HPM analysis; a non-significant outcome, however, suggests that the model lacks statistical efficacy.

3. Monte Carlo Simulations

TLS is statistically superior to LS in estimating the EIV model. To show such an improvement qualitatively in the hedonic analysis, Monte Carlo simulations are designed. In Xu et al. [67], it is shown that the bias of using LS under the EIV model depends on the parameter magnitudes and the covariance matrix of the random errors (or the noise level in the homogenous cases). Therefore, in order to investigate the superiority of TLS over LS in the hedonic analysis, we have to generate data from a real house pricing example. Then, the magnitudes of the parameters and the structure of the coefficient matrix can be close to practical situations. Based on this, we select a real example reported by Wooldridge Wooldridge [63] (p. 211). To ensure the accuracy of the simulations, we employ the TLS method to correct the noisy data and then take the corrected data as the ground truth in the following experiments.

To show the superiority of TLS in hedonic house pricing analysis, we design two experiments: the first is for parameter estimation and the second is for prediction. The dataset consists of $n = 88$ observations and the regression equation reads

$$\log(\text{price}) = \beta_0 + \beta_1 \log(\text{lotsize}) + \beta_2 \log(\text{sqrfit}) + \beta_3 \log(\text{bdrms}) + \beta_4 \log(\text{colonial}), \tag{28}$$

where "price" is the house price in \$1000; "lotsize" is the size of the lot in square feet; "sqrfit" is the size of the house in square feet; "bdrms" is the number of bedrooms; and "colonial" is an indicator variable that equals 1 if the house is of a colonial style and equals 0 otherwise.

As "bdrms" and "colonial" are fixed, we assume that the cofactor matrices take the form $\mathbf{Q} = \mathbf{I}_n$ and $\mathbf{Q}_A = \mathbf{v} \otimes \mathbf{I}_n$ with

$$\mathbf{v} = \text{diag}\{0, 1, 1, 0, 0\}, \tag{29}$$

where $\text{diag}\{\cdot\}$ represents the operator that constructs a diagonal matrix according to its argument. Since the Monte Carlo simulations are conducted based on the ground truths, we conduct the TLS estimation for the whole system by setting $\sigma^2 = \sigma_A^2 = 0.10^2$, and we regard the estimates as $\overline{\mathbf{y}}$ and $\overline{\mathbf{A}}$.

3.1. Parameter Estimation

The purpose of this experiment is to verify that TLS can provide a more accurate parameter estimator than OLS. We set $\sigma = 0.10$ and varied σ_A from 0.01 to 0.20 with increments of 0.01, conducting a total of 20 experiments to show such an improvement at different noise levels. In order to show the statistical performance, the experiment with a specific noise level was replicated 500 times [68].

The steps of the simulation are listed below.

1. Set σ_A and form the covariance matrix $\mathbf{\Sigma}$;
2. Conduct 500 replicated trials and, in each trial,
   (a) Generate the noise $\mathbf{e}$ from the normal distribution $\mathbb{N}(\mathbf{0}, \mathbf{\Sigma})$;
   (b) Reconstruct $\mathbf{e_y}$ and $\mathbf{E}$ from $\mathbf{e}$, and then form $\mathbf{y} = \overline{\mathbf{y}} + \mathbf{e_y}$ and $\mathbf{A} = \overline{\mathbf{A}} + \mathbf{E}$;
   (c) Perform the estimations to obtain $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{TLS}}$;
   (d) Record the discrepancy vectors $\boldsymbol{\epsilon}_{\text{OLS}} = \hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}$ and $\boldsymbol{\epsilon}_{\text{TLS}} = \hat{\boldsymbol{\beta}}_{\text{TLS}} - \boldsymbol{\beta}$.
3. Compute the root mean square error (RMSE) for each parameter for these two schemes. Taking β_i ($i = 0, \cdots, 4$), for example, we have

$$\text{RMSE}\{\beta_i\} = \sqrt{\frac{\sum_{j=1}^{500} \epsilon_{i,j}^2}{500}} \quad i = 0, \cdots, 4.$$

4. Compute the sum of the RMSEs of five parameters for OLS and TLS, respectively.

The computed RMSEs are demonstrated in Figure 2, from which we can see the following.

1. For β_0, β_2, and β_3, the RMSEs of TLS are much smaller than those of OLS. In addition, the improvement becomes more significant as σ_A increases. With $\sigma_A = 0.20$, the improvement ratios (the percentage reduction in the RMSE of the TLS relative to the OLS) for β_0, β_2, and β_3 are 73.72%, 73.39%, and 55.31%, respectively.
2. For β_1 and β_4, the RMSEs of TLS are comparable to those of OLS. However, we can see that the magnitudes of the RMSEs of these two parameters are much smaller than those of the other three parameters (particularly β_0, the intercept). In terms of the sum of the RMSEs, TLS significantly outperforms OLS, achieving a 71.22% improvement with $\sigma_A = 0.20$.
3. In the setting of the covariance matrix, we assume the coefficients corresponding to β_0, β_3, and β_4 to be errorless. However, we can see that the estimates of these parameters are still significantly influenced. Therefore, although we have set some variables to be non-stochastic, the corresponding parameters are also affected in the estimation. More specifically, in the EIV model, all parameters can be biased if LS is applied. For the analytical bias of LS (or approximately the difference between LS and TLS), one can refer to [67].

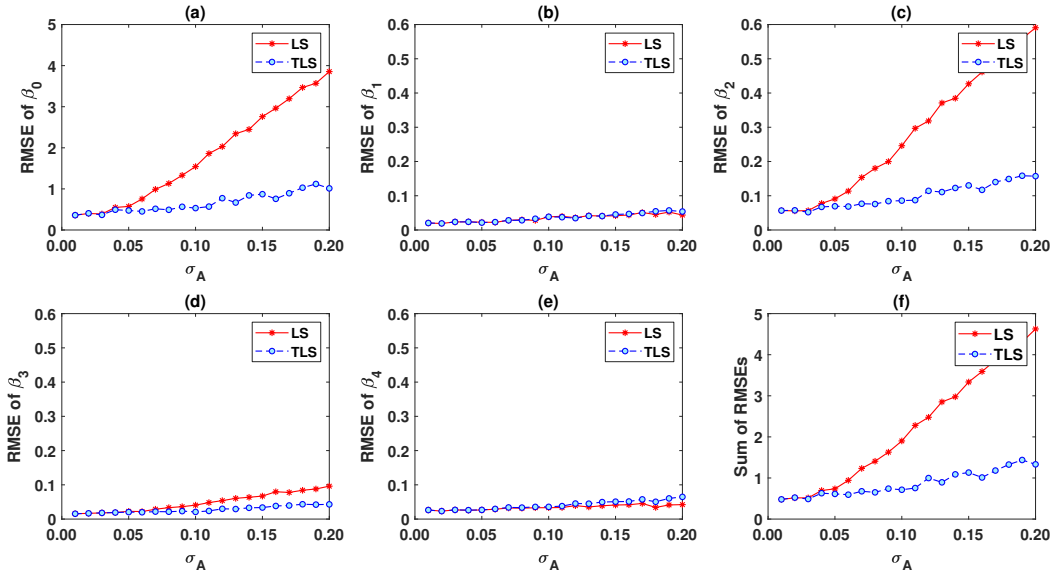

Figure 2. Results of RMSEs in the simulation of parameter estimation: (**a**–**f**) correspond to six parameters.

3.2. Price Prediction

This part is designed to compare the performance of OLS and TLS in terms of prediction. For such a purpose, we partition the observations (both $\overline{\mathbf{y}}$ and $\overline{\mathbf{A}}$) into two parts, i.e.,

$$\overline{\mathbf{y}} = \begin{bmatrix} \overline{\mathbf{y}}_1 \\ \overline{\mathbf{y}}_2 \end{bmatrix} \qquad \overline{\mathbf{A}} = \begin{bmatrix} \overline{\mathbf{A}}_1 \\ \overline{\mathbf{A}}_2 \end{bmatrix},$$

where $\overline{\mathbf{y}}_1$ and $\overline{\mathbf{y}}_2$ are 50- and 38-dimensional vectors; and $\overline{\mathbf{A}}_1$ and $\overline{\mathbf{A}}_2$ are of order 50×5 and 38×5, respectively. Specifically, the first 50 observations are the training set ($\overline{\mathbf{y}}_1$ and $\overline{\mathbf{A}}_1$), and the remaining 38 observations are the validation set ($\overline{\mathbf{y}}_2$ and $\overline{\mathbf{A}}_2$). In the simulations, the cofactor matrices for the training set ($\mathbf{\Sigma}_1$) and the validation set ($\mathbf{\Sigma}_2$) can be constructed similarly to $\mathbf{\Sigma}$: $\mathbf{\Sigma}_1$ is formed with $\mathbf{Q}_{\mathbf{y}_1} = \mathbf{I}_{50}$ and $\mathbf{Q}_{\mathbf{A}_1} = \mathbf{v} \otimes \mathbf{I}_{50}$; $\mathbf{\Sigma}_2$ is formed with $\mathbf{Q}_{\mathbf{y}_2} = \mathbf{I}_{38}$ and $\mathbf{Q}_{\mathbf{A}_2} = \mathbf{v} \otimes \mathbf{I}_{38}$.

With $\sigma = 0.10$, we consider two experiments by setting $\sigma_A = 0.01$ and $\sigma_A = 0.05$, respectively. In each experiment, the following steps are conducted 1000 times.

1. Generate noise $\mathbf{e}_1$ from the normal distribution $\mathbb{N}(\mathbf{0}, \mathbf{\Sigma}_1)$.
2. Reconstruct $\mathbf{e}_{\mathbf{y}_1}$ and $\mathbf{E}_1$ from $\mathbf{e}_1$, and then form $\mathbf{y}_1 = \overline{\mathbf{y}}_1 + \mathbf{e}_{\mathbf{y}_1}$, $\mathbf{A}_1 = \overline{\mathbf{A}}_1 + \mathbf{E}_1$.
3. Perform the estimations to obtain $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{TLS}}$.
4. Repeat the predictions 500 times via the following:
   (a) Generate noise $\mathbf{e}_2$ from the normal distribution $\mathbb{N}(\mathbf{0}, \mathbf{\Sigma}_2)$;
   (b) Reconstruct $\mathbf{e}_{\mathbf{y}_2}$ and $\mathbf{E}_2$ from $\mathbf{e}_2$, and then form $\mathbf{y}_2 = \overline{\mathbf{y}}_2 + \mathbf{e}_{\mathbf{y}_2}$, $\mathbf{A}_2 = \overline{\mathbf{A}}_2 + \mathbf{E}_2$;
   (c) Compute the prediction discrepancy norms $\tau_{\text{OLS}} = \|\mathbf{y}_2 - \mathbf{A}_2\hat{\boldsymbol{\beta}}_{\text{OLS}}\|$ and $\tau_{\text{TLS}} = \|\mathbf{y}_2 - \mathbf{A}_2\hat{\boldsymbol{\beta}}_{\text{TLS}}\|$.
5. Record the ratio of the number of times that TLS has a smaller norm in these 500 predictions.

The results are shown in Figure 3, from which we can see the following. (1) The points (ratios) are more densely distributed between 0.5 and 1.0 (i.e., TLS outperforms OLS). More specifically, the ratios of $\tau_{\text{TLS}} < \tau_{\text{OLS}}$ in these two cases are 75.50% and 61.90%, respectively. (2) In the case of $\mathbf{A}_2$, with larger uncertainty (i.e., larger σ_A), the ratio is smaller. This phenomenon will be analytically discussed in Section 5.

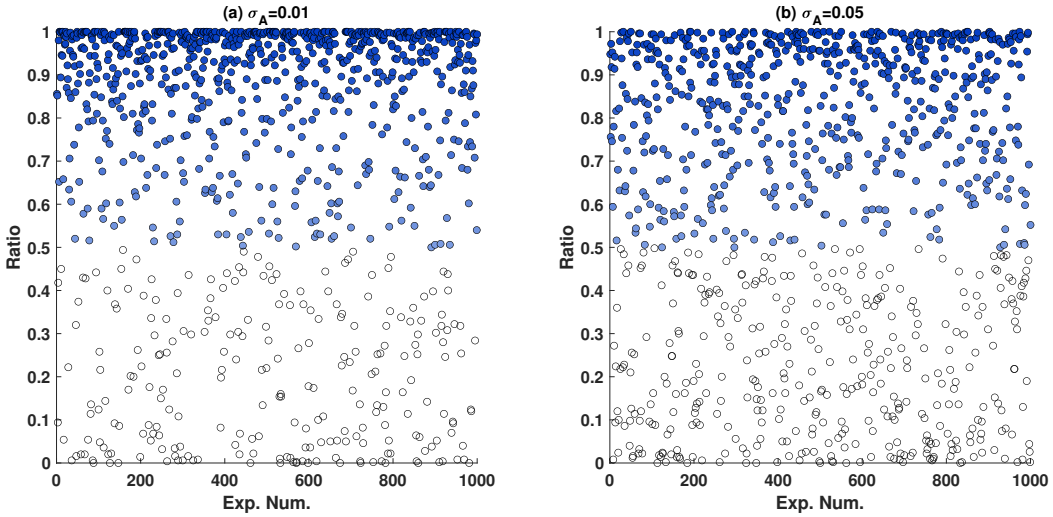

Figure 3. Ratios by which TLS has a smaller prediction discrepancy norm than OLS in 1000 repeated trials.

4. Boston Dataset Analysis

In this section, the Boston house price dataset is analyzed, which was initially presented by [69] in their hedonic analysis of the demand for clean air. It is popular and has been used in many studies, such as robust estimation [70,71], residual normality analysis [72], and non-parametric estimation [73,74]. The original data of $n = 506$ census tracts were published by [75] and found to contain several incorrectly coded observations. In our experiment, the corrected dataset provided by [76] is utilized. The descriptions of the dependent and $m = 13$ independent variables in the dataset are listed in Table 1.

Table 1. Definitions of dependent and independent variables.

Variable	Definitions
Dependent variable	
LMV	logarithm of median price for owner-occupied houses in each census tract
Independent variable	
CRIM	per capita crime rate by town
ZN	proportion of a town's residential land zoned for lots over 25,000 square feet
INDUS	proportion of nonretail business acres per town
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOXSQ	nitrogen oxide concentration (parts per hundred million) squared
RMSQ	average number of rooms per dwelling squared
AGE	proportion of owner-occupied units built prior to 1940
DIS	logarithm of weighted distances to five Boston employment centers
RAD	logarithm of index of accessibility to radial highways
TAX	full-value property tax rate per 10,000
PTRATIO	pupil–teacher ratio by tract
B	$1000(\mathrm{Bk} - 0.63)^2$, where Bk is the proportion of black residents
LSTAT	logarithm of the proportion of the population that is of lower status

To implement TLS, we have to first determine the cofactor matrix $\mathbf{Q_A}$. The values of LSTAT and CRIM are likely to have large uncertainty as the socioeconomic indicators are often determined based on sample surveys, which are susceptible to sampling and recording biases. In contrast, the variable CHAS is deemed almost fixed, primarily because it is based on clear geographical features with minimal variability and subjectivity. The uncertainty of the remaining variables, which is caused by factors such as outdated data or limitations in measurement methods, is assumed to be the same. Given the uncertainty in the variables, we subjectively set $\mathbf{Q_y} = \mathbf{I}_n$ and $\mathbf{Q_A} = \mathbf{v} \otimes \mathbf{I}_n$ with

$$\mathbf{v} = \text{diag}\{0, 1, 0.1, 0.1, 0, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 1\}.$$

The OLS method and the TLS method are applied and compared in the following three cases: (1) $\sigma^2 = 0.10^2$, $\sigma_A^2 = 0.02^2$; (2) $\sigma^2 = 0.10^2$, $\sigma_A^2 = 0.06^2$; (3) $\sigma^2 = 0.10^2$, $\sigma_A^2 = 0.10^2$.

The parameter estimates, the corresponding t-statistic values (in parentheses), and other performance indicators are reported in Table 2. For the overall fitting results, with these two different model setup assumptions, the following can be observed.

1. The parameter estimates differ. The norms $\|\hat{\boldsymbol{\beta}}_{\text{OLS}} - \hat{\boldsymbol{\beta}}_{\text{TLS}}\|_2$ in the three cases are 0.0151, 0.1548, and 0.5659, which show that the difference between OLS and TLS becomes significant as the noise level of the design matrix increases.
2. The significance analysis of the parameters differs. For the first two cases, OLS and TLS identify the same significant parameters. However, for the third case, TLS regards AGE while OLS does not.
3. TLS fits the data better than OLS. For R^2 and R^2_{adj}, TLS produces a higher value than OLS, indicating stronger explanatory power for the observed data; for VF, TLS produces a lower value than OLS, indicating a closer fit to the observed data; for the F-test statistic, indicative of the overall significance of the regression, TLS produces a higher value than OLS, reinforcing the evidence of a statistically sounder model. It is worth mentioning that the effects of EIV on the VF have been systematically investigated by [77]. He shows that OLS always overestimates the VF, which verifies our conclusion.

Table 2. Estimation results produced by OLS and TLS in three cases (coefficients and test statistics marked with *** and ** are significant at the 99% and 95% levels, respectively).

	OLS	**TLS**		
		$\sigma_A = 0.02$	$\sigma_A = 0.06$	$\sigma_A = 0.10$
Parameter Estimates				
CONSTANT	2.83601 ***	2.83801 ***	2.86409 ***	2.9968 ***
	(19.22)	(19.24)	(19.33)	(19.45)
CRIM	−0.01177 ***	−0.01177 ***	−0.01176 ***	−0.01187 ***
	(−9.59)	(−9.59)	(−9.54)	(−9.27)
ZN	0.00009	0.00009	0.0007	0.0003
	(0.18)	(0.11)	(0.13)	(0.05)
INDUS	0.00018	0.00026	0.00099	0.00293
	(0.08)	(0.11)	(2.74)	(1.20)
CHAS	0.09213 ***	0.09188 ***	0.09011 ***	0.08887 ***
	(2.81)	(2.81)	(2.74)	(2.60)
NOXSQ	−0.63724 ***	−0.65158 ***	−0.78409 ***	−1.16814 ***
	(−5.71)	(−5.84)	(−6.98)	(−9.81)
RMSQ	0.00626 ***	0.00610 ***	0.00480 ***	0.00219
	(4.83)	(4.71)	(3.69)	(1.61)
AGE	0.00007	0.00012	0.00049	0.00130 **
	(0.14)	(0.22)	(0.94)	(2.39)
DIS	−0.19784 ***	−0.19912 ***	−0.21104 ***	−0.24723 ***
	(−6.01)	(−6.05)	(−6.38)	(−7.18)
RAD	0.08957 ***	0.08999 ***	0.09372 ***	0.10329 ***
	(4.75)	(4.77)	(4.94)	(5.25)
TAX	−0.00042 ***	−0.00042 ***	−0.00040 ***	−0.0035 ***
	(−3.46)	(−3.45)	(−3.28)	(−2.76)
PTRATIO	−0.02960 ***	−0.02975 ***	−0.03124 ***	−0.03592 ***
	(−5.99)	(−6.02)	(−6.29)	(−6.95)
B	0.00036 ***	0.00036 ***	0.00032 ***	0.00024 ***
	(3.55)	(3.51)	(3.15)	(2.30)
LSTAT	−0.37489 ***	−0.37901 ***	−0.41227 ***	−0.47404 ***
	(−15.20)	(−15.37)	(−16.60)	(−18.13)

Table 2. *Cont.*

	OLS	TLS		
		$\sigma_A = 0.02$	$\sigma_A = 0.06$	$\sigma_A = 0.10$
Performance Indicators				
R^2	0.8108	0.8122	0.8240	0.8496
R^2_{adj}	0.8054	0.8072	0.8194	0.8456
$\hat{\sigma}$	1.7994	1.7926	1.7351	1.6041
F-test statistic	162.1439	163.6525	177.2467	213.7848

Next, we further analyze the results for different parameters. (i) In all three cases, CRIM, DIS, RAD, TAX, and PTRATIO are regarded as significant. It shows a stable and robust relationship between them and the other dependent variables. (ii) For NOXSQ, as we vary σ_A, both the coefficient value and t-statistic dramatically change. This suggests that NOXSQ is a relatively sensitive parameter. It coincides with the analysis in previous work [69] since such a model is formed mainly to investigate the effects of the nitrogen oxide concentration. (iii) For AGE, the significance analysis results are completely different from these model assumptions. This result is consistent with the conclusion of [78], from which we can infer the following reason: older houses are often located in areas closer to urban centers, which may be associated with lower status. Therefore, AGE as a variable demonstrates complex interaction effects, and its significance is subject to change as σ_A increases. TLS and LS yield different significant analysis results because TLS accounts for errors in all variables, leading to different parameter estimates and posterior variances compared to LS. This discrepancy impacts the statistical testing results, as shown in our previous study [67].

The squared residuals are depicted in Figure 4, from which we can see that (1) compared with OLS, TLS has a smaller mean value of squared residuals. Specifically, as increases, the average squared residual of TLS decreases significantly. In the case of σ_A, the mean squared residual of TLS is about 41.94% lower than that of OLS. (2) The squared residuals of TLS are more concentrated. In the interval of residual squared values from 0 to 0.10, the OLS and the TLS with different σ_A contain 470 (91.70%), 471 (93.08%), 475 (93.87%), and 491 (97.04%) points, respectively. (3) OLS produces some extreme residuals; in contrast, TLS, particularly with $\sigma_A = 0.06$ and $\sigma_A = 0.10$, appears to constrain these extreme residuals more effectively. This indicates that TLS is more appropriate to handle these extreme observed values. (4) It can be seen from the shape of the violin plot in the TLS model that as σ_A increases, the kernel density estimate (KDE) shows that most of the squared residuals are more concentrated around the median, indicating less variability and more stable model performance.

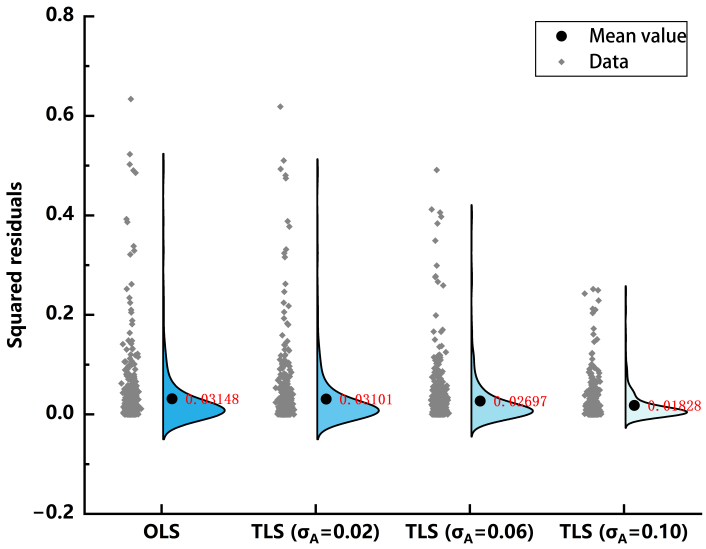

Figure 4. Distributions of squared residuals and the corresponding half violin plots for OLS and TLS in three cases.

5. Practical Tests and Analysis

In this section, "Guanshan Boulevard" of Wuhan is utilized as the study area. The data were manually collected from unrestricted public domains. In addition to parameter estimation, we also analyzed their performance in house price prediction.

5.1. Study Area and Data Source

The studied real estate market is on Guanshan Boulevard in Wuhan, China, whose area is roughly 4.11 square kilometers; see Figure 5. As the employment population grows, the real estate market experiences a boost, creating a mutually reinforcing cycle of development.

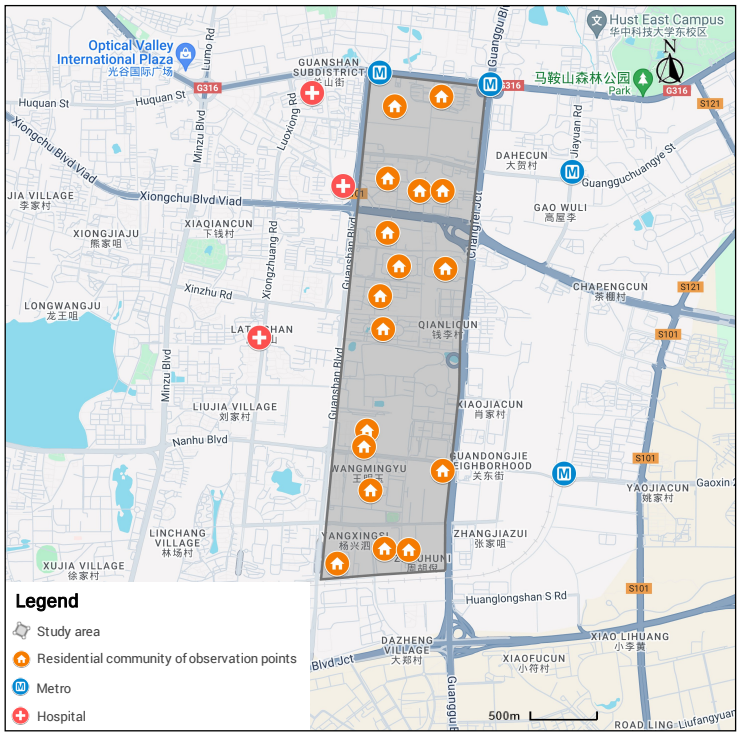

Figure 5. Study area in Guanshan Boulevard.

The listing price of the house and some of the house characteristics are obtained from the Lianjia website (http://www.lianjia.com/, accessed on 23 November 2023) and the AMAP website (http://www.amap.com/, accessed on 23 November 2023). To avoid potential biases caused by real estate market segmentation, the selected residential samples are all commercial projects. Low-rent housing provided by the government as social welfare is not included in the study. In addition, our study only includes closed high-rise residential buildings, excluding villas and bungalows located in the study area. Records with missing data for any feature were removed from the dataset. All data points were recorded from June 2022 to November 2023.

5.2. Data Preparation

For the independent variables, we initially collected a series of variables to capture the residential characteristics, referencing mainstream variables [1,2,79,80] and adapting them to the local real estate market.

- Structural attributes. We select management fees, the ratio of elevators to residents, the ratio of parking spaces to residents, the total number of functional rooms, the living room orientation, the building type, the housing year, the green space rate, and the building's floor area ratio.

- Neighborhood attributes. To account for the educational level, we compile diverse data points (number, distance, and quality) for kindergartens, primary schools, and middle schools. For medical services, we assess the distance to the nearest tertiary hospital. For commercial services, we evaluate the availability of nearby supermarkets, malls, and other amenities. For the level of leisure, we count the parks and attractions within a 3 kilometer radius of the residential community.
- Locational attributes. We only select the logarithm of the distance (m) to the nearest metro station and bus station. This is because all house samples are within a small area, and their external location factors, such as the distance to the Wuhan Central Business District (CBD) or distance to large landscapes (East Lake, etc.), do not show significant changes.

For the dependent variables (i.e., the house price), we have the following three preprocessing steps. (1) Unit price calculation. House prices are preprocessed by calculating the unit price in yuan per square meter from the total listing price and building area. (2) Floor-level standardization. To mitigate the nonlinear effects of the floor levels, the prices are standardized across different floors using a correction coefficient, following the local guidelines specific to floor-level adjustments. (3) Transaction date adjustment. The transaction dates are adjusted using the average price change rate, converting each transaction's unit price into the valuation time point's value.

During the exploratory data analysis, we evaluate the significance of each independent variable using p-values and address potential collinearity by calculating variance inflation factors. We meticulously screen the housing characteristic variables pertinent to our study area. In the end, a refined set of $m = 10$ characteristic variables is selected to construct our HPM; see Table 3.

Table 3. Definitions of variables in the example of the Wuhan dataset.

Variable	Variable Definition and Measurement Method	Mean	Std.	Sign
Dependent variable				
PRICE	Logarithm of preprocessed price (yuan)	10.0727	0.2319	\
Structural attributes				
NROOMS	Total number of functional rooms	6.9450	1.4220	+
BUILDINGTYPE	Building types, including tower blocks (=1), slab blocks (=3), and a combination of the two (=2)	1.9450	0.6196	+
FEE	Property management fees (yuan/Mon·m^2)	2.5029	1.1969	+
RPARKING	Ratio of the number of parking spaces to the number of residential units	0.9289	0.6180	+
RGREENING	$100(G - 0.30)^2$, where G is the rate of green space in residential areas	0.4747	0.6229	+
PSCHOOL	Score based on the number, quality, and distance of primary schools around the house, from the Lianjia website	8.5980	0.2848	+
MSCHOOL	Logarithm of distance (m) to the nearest middle school	7.1029	0.3433	−
DHOSPITAL	Logarithm of distance (m) to the nearest hospital	7.1623	0.6802	Unknown
COMMERCIAL	Score based on the quantity and quality of supermarkets, shopping malls, and other facilities near residential areas, from the AMAP website	2.2235	0.4443	+
DISTANCE	Logarithm of distance (m) to the nearest metro station	6.6057	0.6557	−

5.3. Parameter Estimation

In this part, we have a total of $n = 200$ house transaction sample points. We use the actual values (i.e., numerical data) for some variables, rather than converting them into categories or levels (i.e., categorical data), to avoid subjectivity when dividing levels. Despite this, the characteristic variables still exhibit varying levels of noise, which is attributed to the limitations and ambiguities in publicly available information. For instance, the "distance to the nearest hospital" may not be individually measured for each property but is instead represented by the average distance from the entire neighborhood, which

fails to accurately reflect the attributes of individual properties. Based on a comprehensive analysis of the accuracy of the obtained data, we set $\mathbf{Q_A} = \mathbf{v} \otimes \mathbf{I}_n$ with

$$\mathbf{v} = \text{diag}\{0, 0.1, 0.1, 0.2, 0.3, 0.2, 0.1, 0.4, 0.4, 0.1, 0.4\}.$$

The results for parameter estimation are shown in Table 4. At the 99% confidence level, the critical F-distribution value for the corresponding degrees of freedom is 2.41594. The F-test statistic values for the OLS and TLS methods in this HPM are 82.0777 and 87.7251, respectively, far exceeding this threshold and thereby demonstrating the effectiveness of the HPM. From the results of the remaining performance indicators, the TLS method produces a higher R^2 and R^2_{adj} and a lower VF than the OLS method. It indicates that TLS better fits the model, which is consistent with the conclusion drawn for the Boston dataset.

Table 4. Estimation results produced by OLS and TLS in the real data (coefficients and test statistics marked with ***, **, and * are statistically significant at the 99%, 95%, and 90% levels, respectively).

	OLS	**TLS**
Parameter Estimates		
NROOMS	0.0223 ***	0.0223 ***
BUILDINGTYPE	0.0599 ***	0.0592 ***
FEE	0.1439 ***	0.1412 ***
RPARKING	0.0196	0.0226
RGREENING	0.0488 ***	0.0506 ***
PSCHOOL	0.3832 ***	0.4111 ***
MSCHOOL	−0.0804	−0.1053**
DHOSPITAL	0.2486 ***	0.2894 ***
COMMERCIAL	0.0809 ***	0.0780 **
DISTANCE	−0.0642 *	−0.0892 ***
Constant	5.1399 ***	4.9615 ***
Performance Indicators		
R^2	0.8128	0.8227
R^2_{adj}	0.8029	0.8134
$\hat{\sigma}$	1.0294	1.0017
F-test statistic	82.0777	87.7251

For the estimates of the parameters, in this study area, (1) these two methods regard the following parameters as significant: PSCHOOL, DHOSPITAL, FEE, COMMERCIAL, RPARKING, BUILDINGTYPE, RGREENING, and NROOMS. This suggests that the importance of these factors in the real estate market is generally recognized. On the contrary, the impact of RPARKING on housing prices, which is usually considered important, is relatively small in this area. The reason may be that residents prefer to park their vehicles in areas with no parking fees, such as on the streets outside the community. This preference leads to a reduced demand for parking spaces within the community among residents, thereby diminishing the influence of onsite parking facilities on property values in this region. (2) However, MSCHOOL and DISTANCE show different results in the two models: (i) MSCHOOL is significant in TLS (−0.1053 **) but not in OLS (−0.0804). The premium on housing prices due to educational resources is primarily a result of the "Nearby Enrollment" policy (i.e., students attend schools based on their residential location). Therefore, homes in districts with high-quality education are particularly favored by parents. This is especially true for primary schools, where enrollment strictly depends on the residential address. For middle schools, while many well-known ones require entrance exams, reducing the impact of the location, the need to shorten children's commuting times and boost the chances of entering a top-tier school at the next educational level still results in a premium for housing near middle schools. Thus, both primary and middle schools (i.e., the compulsory education stage) exhibit a significant school district effect, directly contributing to the rise in housing prices. This aligns with the findings of some Chinese scholars [19,81].

(ii) For DISTANCE, TLS indicates greater significance (-0.0892 ***) than OLS (-0.0642 *). Despite our study area being relatively small and having generally good traffic, this does not diminish the significant impact of accessibility on housing prices, even in areas with well-developed transportation.

The different methods always yield a different understanding of the determinants of real estate values. In this experiment, by considering the randomness of the dependent variables, we can see that TLS gains more insights than OLS in factors like MSCHOOL and DISTANCE.

5.4. Price Prediction

Besides analyzing the influence on property prices, we further collected 90 points $y_{p,i}$ together with the dependent variables contained in $\mathbf{a}_{p,i}$ ($i = 1, \cdots, 80$) to test the performance in prediction. Based on the previously estimated parameters (i.e., Table 4), the predicted prices can be obtained as $\tilde{y}_{p,i} = \mathbf{a}_{p,i}^{\mathsf{T}}\hat{\boldsymbol{\beta}}$. By comparing the predicted values with the observed values $y_{p,i}$, we can calculate the relative error

$$\mathrm{RE}_i = (\tilde{y}_{p,i} - y_{p,i})/y_{p,i} \qquad i = 1, \cdots, 80.$$

The summary statistics of the relative errors are presented in Table 5. In addition, the frequency histograms (together with the KDE) and the boxplots of the relative errors are shown in Figure 6.

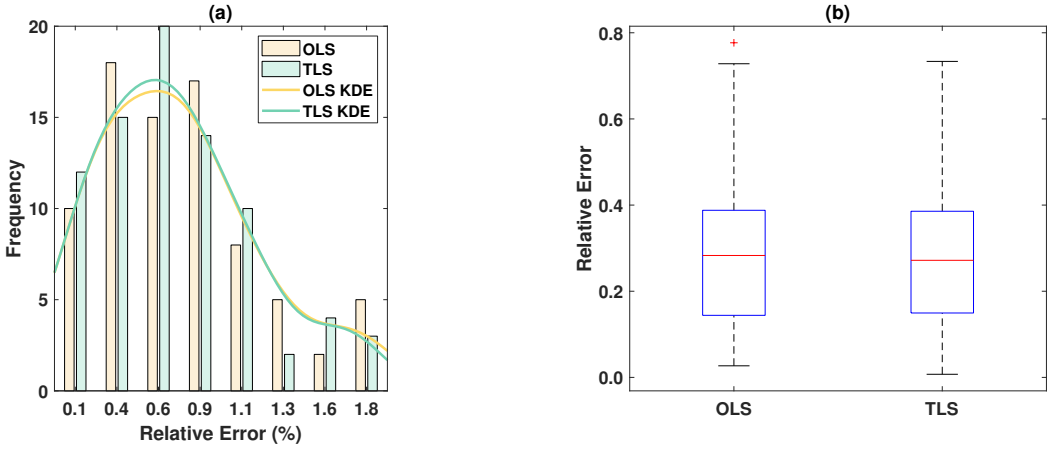

Figure 6. The relative errors: (**a**) the frequency histograms (together with the KDE); (**b**) the boxplots.

From these results, we can see that (1) in Table 5, most quantities corresponding to TLS are slightly smaller than those of OLS. To be more specific, for 62.50% of the sampled points, TLS provides a more accurate predicted value. (2) For OLS, the relative errors range from 0.07% to 1.94%, while those of TLS range from 0.02% to 1.83%, which shows a marginally narrower range. (3) Both methods show similar patterns in the RE distribution, primarily concentrated in the lower error ranges. However, we can also see that the frequency of TLS is smaller than that of OLS with a higher relative error (i.e., the last hist in the histogram), suggesting that TLS could offer greater robustness in certain scenarios.

Therefore, TLS slightly outperforms (or is at least comparable to) OLS in terms of prediction accuracy. Let us then attempt to analyze such a phenomenon from a theoretical perspective. We denote the estimation discrepancy as $\boldsymbol{\epsilon} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$; then, we have the prediction discrepancy

$$\begin{aligned} \epsilon_{p,i} &= \tilde{y}_{p,i} - y_{p,i} \\ &= \mathbf{a}_{p,i}^{\mathsf{T}}\hat{\boldsymbol{\beta}} - (\mathbf{a}_{p,i} - \mathbf{e}_{p,ai})^{\mathsf{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\epsilon}) - e_{p,i} \\ &= \mathbf{a}_{p,i}^{\mathsf{T}}\boldsymbol{\epsilon} + \mathbf{e}_{p,ai}^{\mathsf{T}}\boldsymbol{\beta} - e_{p,i}, \end{aligned} \tag{30}$$

where $\mathbf{e}_{p,ai}$ and $e_{p,i}$ are random errors corresponding to $\mathbf{a}_{p,i}$ and $y_{p,i}$, respectively, i.e., $y_p - e_{p,i} = (\mathbf{a}_{p,i} - \mathbf{e}_{p,ai})^\mathsf{T}\boldsymbol{\beta}$. Provided that the training observations are statistically independent of the validation observations, taking the expectation of $\epsilon_{p,i}$ yields the prediction bias

$$\mathbb{E}\{\epsilon_{p,i}\} = \bar{\mathbf{a}}_{p,i}^\mathsf{T}\mathbb{E}\{\boldsymbol{\epsilon}\} \tag{31}$$

From such an expression, we can see that although TLS has a smaller estimation bias norm (i.e., $\|\mathbb{E}\{\boldsymbol{\epsilon}\}\|_2$), it may not be guaranteed to produce a smaller prediction bias because of the combination $\bar{\mathbf{a}}_{p,i}^\mathsf{T}$. In addition, in one single experiment, the prediction discrepancy additionally relies on the uncertainty of the dependent variable $\mathbf{a}_{p,i}$ and the uncertainty of $y_{p,i}$. Therefore, although TLS greatly outperforms OLS in parameter estimation, its advantage in prediction is marginal, even in simulations.

Table 5. The summary statistics of the relative errors.

Statistic	RE (OLS)	RE (TLS)
Mean	0.74%	0.72%
STD	0.46%	0.44%
Min	0.07%	0.02%
25% quantile	0.36%	0.37%
50% quantile	0.71%	0.70%
75% quantile	0.97%	0.96%
Max	1.94%	1.83%

6. Discussion

In the simulations, the noise levels were systematically explored by designing experiments with a fixed σ and incrementally increasing σ_A. TLS demonstrates stronger explanatory power and a closer fit to the observed data. The RMSE is significantly reduced, with improvements exceeding 70% for the sum of the RMSEs at $\sigma_A = 0.20$. Notably, even the estimates of non-stochastic parameters are influenced by the randomness of other variables. A comparison of the TLS and LS results confirms the following.

- For parameter estimation, TLS consistently achieves a higher R^2 and R^2_{adj}, a lower VF, and a higher F-test statistic in the analysis of both the Boston and Wuhan datasets. This performance demonstrates that TLS has stronger explanatory power and a closer fit to the observed data. Furthermore, TLS also aligns more closely with the findings from previous studies [19,69,78,81]. Importantly, TLS effectively bounds extreme data points, enhancing the reliability of the estimates. Moreover, TLS highlights the importance of factors such as educational resources for middle schools and the proximity to metro stations, which OLS tends to underestimate in the Wuhan dataset.
- For price prediction, the performance advantage of TLS over OLS diminishes with increasing uncertainty (i.e., larger σ_A) in the simulations. This performance is also evident in the Wuhan dataset, in which TLS outperforms OLS in 62.50% of the observations, and most statistics of the relative errors are slightly better. We consider that the limited advantage is believed to stem from the additional prediction discrepancies that depend on the uncertainties of the dependent and independent variables.

In the real examples, similar conclusions can be drawn. One may note the minor increase in R^2 observed when applying TLS compared to OLS, which can be attributed to three main factors. (1) The similarity of the magnitude of R^2 does not imply the similarity of the estimation results. From the definition of R^2, we can see that it only considers the estimated parameters, while the uncertainty of the data (or the difference in the posterior precision) is completely ignored. In [67], it is shown that LS tends to be optimistic about the precision assessment. Thus, in the real example, we can see that although the improvement in R^2 seems to be marginal, the significance analysis results are very different. (2) A higher R^2 does not always indicate better performance. By assuming a higher noise level in the

coefficient matrix, the improvement ratio is expected to increase. However, this might not yield meaningful results and could potentially result in overfitting. (3) The enhancements in TLS over LS are significantly influenced by the noise levels in the coefficient matrix. Despite the minimal changes in R^2, substantial gains in other metrics, such as the RMSE and variance factor, are evident. Collectively, these points demonstrate that the advancements provided by TLS are both substantive and beneficial.

7. Conclusions and Outlook

This article introduces an advanced approach to analyzing housing prices using TLS estimation within the HPM. It comprehensively addresses errors in the dependent and independent variables, making it suitable for real estate data characterized by measurement inaccuracies and subjectivity. In addition, the Gauss–Newton-type iterative algorithm is derived, and the posterior precision assessment is given.

In both the simulated and real examples, the application of TLS in the HPM is shown to enhance the explanatory power and accuracy of the model. Our work enriches the HPM framework; it not only facilitates a thorough analysis of the relevant variables but more accurately assesses their significant impacts. This helps us to navigate the complexities of the real estate market and make more informed decisions.

The formulation presented in this paper has the potential to consider correlations. The consideration of spatial effects (such as spatial autocorrelation and heterogeneity) is necessary for spatial data analysis [82–84]. In our formulation, we will next take into account some assumptions about the covariance matrix $\mathbf{\Sigma}$ to consider the spatial correlations among the sampled points and even the cross-correlations between the dependent and independent variables. This will be part of our future research, aimed to refine our method.

Author Contributions: Wenxi Zhan: conceptualization, data curation, investigation, software, writing—original draft. Yu Hu: methodology. Wenxian Zeng: formal analysis, funding acquisition. Xing Fang: supervision, writing—review and editing. Xionghua Kang: validation. Dawei Li: validation, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (42174049; 42274007) and the Special Fund of Hubei Luojia Laboratory (No. 220100003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The Lagrangian of TLS reads

$$\Phi(\eta, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathbf{e}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{e} + \boldsymbol{\lambda}^{\mathsf{T}}(\mathbf{y} - \mathbf{A}\boldsymbol{\beta} - \mathbf{B}\mathbf{e}), \tag{A1}$$

where $\boldsymbol{\lambda}$ is an n-vector of Lagrange multipliers.

The Euler–Lagrange conditions read

$$\frac{1}{2}\frac{\partial \Phi}{\partial \boldsymbol{\beta}}\bigg|_{\hat{\boldsymbol{\beta}},\hat{\mathbf{e}},\hat{\boldsymbol{\lambda}}} = -\mathbf{A}^{\mathsf{T}}\hat{\boldsymbol{\lambda}} + \hat{\mathbf{E}}^{\mathsf{T}}\hat{\boldsymbol{\lambda}} = \mathbf{0} \tag{A2}$$

$$\frac{1}{2}\frac{\partial \Phi}{\partial \mathbf{e}}\bigg|_{\hat{\boldsymbol{\beta}},\hat{\mathbf{e}},\hat{\boldsymbol{\lambda}}} = \mathbf{Q}^{-1}\hat{\mathbf{e}} - \hat{\mathbf{B}}^{\mathsf{T}}\hat{\boldsymbol{\lambda}} = \mathbf{0} \tag{A3}$$

$$\frac{1}{2}\frac{\partial \Phi}{\partial \boldsymbol{\lambda}}\bigg|_{\hat{\boldsymbol{\beta}},\hat{\mathbf{e}}} = \mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}} - \hat{\mathbf{B}}\hat{\mathbf{e}} = \mathbf{0}. \tag{A4}$$

From (A3), we can obtain the residual vector

$$\hat{\mathbf{e}} = \mathbf{Q}\hat{\mathbf{B}}^{\mathsf{T}}\hat{\boldsymbol{\lambda}}. \tag{A5}$$

Reshaping the residual vector (A5), we can obtain

$$\begin{aligned}[\,\widehat{\mathbf{E}} \quad \hat{\mathbf{e}}_{\mathbf{y}}\,] &= \mathrm{vec}^{-1}_{n,m+2}\hat{\mathbf{e}} \\ &= \Big((\mathrm{vec}\mathbf{I}_{m+2})^{\mathsf{T}} \otimes \mathbf{I}_n\Big)(\mathbf{I}_{m+2} \otimes \hat{\mathbf{e}}),\end{aligned} \tag{A6}$$

where $\mathrm{vec}^{-1}_{n,m+2}(\cdot)$ is the inverse operator of $\mathrm{vec}(\cdot)$, i.e., restructuring an $n \times (m+2)$ matrix from an $(nm+2n)$-vector.

Inserting (A5) into (A4) yields

$$\hat{\boldsymbol{\lambda}} = \widehat{\mathbf{Q}}_{\mathbf{B}}^{-1}(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}}), \tag{A7}$$

where $\hat{\mathbf{Q}}_{\mathbf{B}} = \hat{\mathbf{B}}\mathbf{Q}\hat{\mathbf{B}}^{\mathsf{T}}$. Backsubstituting it into (A2) yields

$$\hat{\boldsymbol{\beta}}_{\mathrm{TLS}} = (\widehat{\mathbf{A}}^{\mathsf{T}}\widehat{\mathbf{Q}}_{\mathbf{B}}^{-1}\widehat{\mathbf{A}})^{-1}\widehat{\mathbf{A}}^{\mathsf{T}}\widehat{\mathbf{Q}}_{\mathbf{B}}^{-1}(\mathbf{y} - \widehat{\mathbf{E}}\hat{\boldsymbol{\beta}}), \tag{A8}$$

where $\widehat{\mathbf{A}} = \mathbf{A} - \widehat{\mathbf{E}}$.

References

1. Wen, H.; Lu, J.; Lin, L. An improved method of real estate evaluation based on Hedonic price model. In Proceedings of the 2004 IEEE International Engineering Management Conference (IEEE Cat. No. 04CH37574), Singapore, 18–21 October 2004; Volume 3, pp. 1329–1332.
2. Khoshnoud, M.; Sirmans, G.S.; Zietz, E.N. The Evolution of Hedonic Pricing Models. *J. Real Estate Lit.* **2023**, *31*, 1–47. [CrossRef]
3. Geerts, M.; De Weerdt, J. A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 200. [CrossRef]
4. Pai, P.F.; Wang, W.C. Using machine learning models and actual transaction data for predicting real estate prices. *Appl. Sci.* **2020**, *10*, 5832. [CrossRef]
5. Zulkifley, N.H.; Rahman, S.A.; Ubaidullah, N.H.; Ibrahim, I. House Price Prediction using a Machine Learning Model: A Survey of Literature. *Int. J. Mod. Educ. Comput. Sci.* **2020**, *12*, 46–54. [CrossRef]
6. Kang, Y.; Zhang, F.; Peng, W.; Gao, S.; Rao, J.; Duarte, F.; Ratti, C. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy* **2021**, *111*, 104919. [CrossRef]
7. Mullainathan, S.; Spiess, J. Machine learning: An applied econometric approach. *J. Econ. Perspect.* **2017**, *31*, 87–106. [CrossRef]
8. Baldominos, A.; Blanco, I.; Moreno, A.J.; Iturrarte, R.; Bernárdez, Ó.; Afonso, C. Identifying real estate opportunities using machine learning. *Appl. Sci.* **2018**, *8*, 2321. [CrossRef]
9. Del Giudice, V.; De Paola, P.; Forte, F.; Manganelli, B. Real estate appraisals with Bayesian approach and Markov chain hybrid Monte Carlo method: An application to a central urban area of Naples. *Sustainability* **2017**, *9*, 2138. [CrossRef]
10. Yacim, J.A.; Boshoff, D.G.B. Impact of artificial neural networks training algorithms on accurate prediction of property values. *J. Real Estate Res.* **2018**, *40*, 375–418. [CrossRef]
11. Pérez-Rave, J.I.; Correa-Morales, J.C.; González-Echavarría, F. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *J. Prop. Res.* **2019**, *36*, 59–96. [CrossRef]
12. Goh, Y.M.; Costello, G.; Schwann, G. Accuracy and robustness of house price index methods. *Hous. Stud.* **2012**, *27*, 643–666. [CrossRef]
13. Hill, R.J. Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. *J. Econ. Surv.* **2013**, *27*, 879–914. [CrossRef]
14. Glumac, A.Š.; Hemida, H.; Höffer, R. Wind energy potential above a high-rise building influenced by neighboring buildings: An experimental investigation. *J. Wind Eng. Ind. Aerodyn.* **2018**, *175*, 32–42. [CrossRef]
15. Kohlhase, J.E. The impact of toxic waste sites on housing values. *J. Urban Econ.* **1991**, *30*, 1–26. [CrossRef]
16. Garrod, G.D.; Willis, K.G. Valuing goods' characteristics: An application of the hedonic price method to environmental attributes. *J. Environ. Manag.* **1992**, *34*, 59–76. [CrossRef]
17. Goodman, A.C.; Thibodeau, T.G. Age-related heteroskedasticity in hedonic house price equations. *J. Hous. Res.* **1995**, *6*, 25–42.
18. Clark, S.C. Work/family border theory: A new theory of work/family balance. *Hum. Relations* **2000**, *53*, 747–770. [CrossRef]
19. Zhang, J.; Li, H.; Lin, J.; Zheng, W.; Li, H.; Chen, Z. Meta-analysis of the relationship between high quality basic education resources and housing prices. *Land Use Policy* **2020**, *99*, 104843. [CrossRef]
20. Seo, K.; Golub, A.; Kuby, M. Combined impacts of highways and light rail transit on residential property values: A spatial hedonic price model for Phoenix, Arizona. *J. Transp. Geogr.* **2014**, *41*, 53–62. [CrossRef]

21. Blake, J. *Family Size and Achievement*; University of California Press: Berkeley, CA, USA, 2022; Volume 3.
22. Sander, H.A.; Polasky, S. The value of views and open space: Estimates from a hedonic pricing model for Ramsey County, Minnesota, USA. *Land Use Policy* **2009**, *26*, 837–845. [CrossRef]
23. Wu, C.; Ye, X.; Du, Q.; Luo, P. Spatial effects of accessibility to parks on housing prices in Shenzhen, China. *Habitat Int.* **2017**, *63*, 45–54. [CrossRef]
24. Wu, C.; Ren, F.; Hu, W.; Du, Q. Multiscale geographically and temporally weighted regression: Exploring the spatiotemporal determinants of housing prices. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 489–511. [CrossRef]
25. Rosen, S. Hedonic prices and implicit markets: Product differentiation in pure competition. *J. Political Econ.* **1974**, *82*, 34–55. [CrossRef]
26. Sirmans, S.; Macpherson, D.; Zietz, E. The composition of hedonic pricing models. *J. Real Estate Lit.* **2005**, *13*, 1–44. [CrossRef]
27. Malpezzi, S. Hedonic pricing models: A selective and applied review. *Hous. Econ. Public Policy* **2003**, *1*, 67–89.
28. Curto, R.; Fregonara, E.; Semeraro, P. Listing behaviour in the Italian real estate market. *Int. J. Hous. Mark. Anal.* **2015**, *8*, 97–117. [CrossRef]
29. Clapp, J.M. A new test for equitable real estate tax assessment. *J. Real Estate Financ. Econ.* **1990**, *3*, 233–249. [CrossRef]
30. Wilhelmsson, M. Spatial models in real estate economics. *Housing Theory Soc.* **2002**, *19*, 92–101. [CrossRef]
31. Wu, B.; Li, R.; Huang, B. A geographically and temporally weighted autoregressive model with application to housing prices. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1186–1204. [CrossRef]
32. Wang, W.C.; Chang, Y.J.; Wang, H.C. An application of the spatial autocorrelation method on the change of real estate prices in Taitung City. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 249. [CrossRef]
33. Haider, M.; Miller, E.J. Effects of transportation infrastructure and location on residential real estate values: Application of spatial autoregressive techniques. *Transp. Res. Rec.* **2000**, *1722*, 1–8. [CrossRef]
34. Lu, B.; Brunsdon, C.; Charlton, M.; Harris, P. Geographically weighted regression with parameter-specific distance metrics. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 982–998. [CrossRef]
35. Cellmer, R.; Cichulska, A.; Bełej, M. Spatial analysis of housing prices and market activity with the geographically weighted regression. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 380. [CrossRef]
36. Tomal, M. Modelling housing rents using spatial autoregressive geographically weighted regression: A case study in Cracow, Poland. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 346. [CrossRef]
37. Ahn, J.J.; Byun, H.W.; Oh, K.J.; Kim, T.Y. Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Syst. Appl.* **2012**, *39*, 8369–8379. [CrossRef]
38. Li, S.; Whalley, J.; Zhao, X. Housing price and household savings rates: Evidence from China. *J. Chin. Econ. Bus. Stud.* **2013**, *11*, 197–217. [CrossRef]
39. Doszyń, M. Algorithm of real estate mass appraisal with inequality restricted least squares (IRLS) estimation. *J. Eur. Real Estate Res.* **2020**, *13*, 161–179. [CrossRef]
40. Berg, N. A simple Bayesian procedure for sample size determination in an audit of property value appraisals. *Real Estate Econ.* **2006**, *34*, 133–155. [CrossRef]
41. Wheeler, D.C.; Páez, A.; Waller, L.A.; Spinney, J. Housing Sub-markets and Hedonic Price Analysis: A Bayesian Approach. *Sustain. J. Rec.* **2007**, *9*, 2138.
42. Morano, P.; Tajani, F. Bare ownership evaluation. Hedonic price model vs. artificial neural network. *Int. J. Bus. Intell. Data Min.* **2013**, *8*, 340–362. [CrossRef]
43. Wang, C.; Li, J.; Guo, P. The normalized interval regression model with outlier detection and its real-world application to house pricing problems. *Fuzzy Sets Syst.* **2015**, *274*, 109–123. [CrossRef]
44. Mason, P.; Pryce, G. Controlling for transactions bias in regional house price indices. *Hous. Stud.* **2011**, *26*, 639–660. [CrossRef]
45. Doszyń, M. Prior information in econometric real estate appraisal: A mixed estimation procedure. *J. Eur. Real Estate Res.* **2021**, *14*, 349–361. [CrossRef]
46. Powe, N.A.; Garrod, G.; Willis, K. Valuation of urban amenities using an hedonic price model. *J. Prop. Res.* **1995**, *12*, 137–147. [CrossRef]
47. Li, J.; Chiang, Y.H. What pushes up China's real estate price? *Int. J. Hous. Mark. Anal.* **2012**, *5*, 161–176. [CrossRef]
48. Anselin, L.; Lozano-Gracia, N. Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empir. Econ.* **2008**, *34*, 5–34. [CrossRef]
49. Golub, G.H.; van Loan, C.F. An Analysis of the Total Least Squares Problem. *SIAM J. Numer. Anal.* **1980**, *17*, 883–893. [CrossRef]
50. Markovsky, I.; Van Huffel, S. Overview of total least-squares methods. *Signal Process.* **2007**, *87*, 2283–2302. [CrossRef]
51. Strutz, T. *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 1.
52. Chen, W.; Chen, M.; Zhou, J. Adaptively regularized constrained total least-squares image restoration. *IEEE Trans. Image Process.* **2000**, *9*, 588–596. [CrossRef]
53. Hirakawa, K.; Parks, T.W. Image denoising using total least squares. *IEEE Trans. Image Process.* **2006**, *15*, 2730–2742. [CrossRef]
54. Fang, X. A structured and constrained total least-squares solution with cross-covariances. *Stud. Geophys. Geod.* **2014**, *58*, 1–16. [CrossRef]
55. Fang, X. On non-combinatorial weighted total least squares with inequality constraints. *J. Geod.* **2014**, *88*, 805–816. [CrossRef]

56. Fang, X. A total least squares solution for geodetic datum transformations. *Acta Geod. Geophys.* **2014**, *49*, 189–207. [CrossRef]
57. Fang, X. Weighted total least-squares with constraints: A universal formula for geodetic symmetrical transformations. *J. Geod.* **2015**, *89*, 459–469. [CrossRef]
58. Hu, Y.; Fang, X.; Kutterer, H. Center strategies for universal transformations: Modified iteration policy and two alternative models. *GPS Solut.* **2023**, *27*, 92. [CrossRef]
59. Hu, Y.; Fang, X.; Zeng, W.; Kutterer, H. Multiframe Transformation with Variance Component Estimation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3302322. [CrossRef]
60. Van Huffel, S.; Vandewalle, J. *The Total Least Squares Problem: Computational Aspects and Analysis*; SIAM: Philadelphia, PA, USA, 1991.
61. Fang, X. Weighted Total Least Squares Solutions for Applications in Geodesy. Ph.D. Thesis, Gottfried Wilhelm Leibniz Universität Hannover, Hannover, Germany, 2011.
62. Fang, X. Weighted total least squares: Necessary and sufficient conditions, fixed and random parameters. *J. Geod.* **2013**, *87*, 733–749. [CrossRef]
63. Wooldridge, J.M. *Introductory Econometrics: A Modern Approach*, 6th ed.; South-Western: London, UK, 2015.
64. Ver Hoef, J.M.; Peterson, E.E.; Hooten, M.B.; Hanks, E.M.; Fortin, M.J. Spatial autoregressive models for statistical inference from ecological data. *Ecol. Monogr.* **2018**, *88*, 36–59. [CrossRef]
65. Magnus, J.R.; Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*; John Wiley & Sons: Hoboken, NJ, USA, 2019.
66. Amiri-Simkooei, A.; Jazaeri, S. Weighted total least squares formulated by standard least squares theory. *J. Geod. Sci.* **2012**, *2*, 113–124. [CrossRef]
67. Xu, P.; Liu, J.; Zeng, W.; Shen, Y. Effects of errors-in-variables on weighted least squares estimation. *J. Geod.* **2014**, *88*, 705–716. [CrossRef]
68. Box, M.J. Bias in Nonlinear Estimation. *J. R. Stat. Soc. Ser. B Methodol.* **1971**, *33*, 171–201. [CrossRef]
69. Harrison, D.; Rubinfeld, D.L. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* **1978**, *5*, 81–102. [CrossRef]
70. Krasker, W.S.; Kuh, E.; Welsch, R.E. Estimation for dirty data and flawed models. *Handb. Econom.* **1983**, *1*, 651–698.
71. Subramanian, S.; Carson, R.T. Robust regression in the presence of heteroskedasticity. *Adv. Econom.* **1988**, *7*, 85–138.
72. Lange, N.; Ryan, L. Assessing normality in random effects models. *Ann. Stat.* **1989**, *17*, 624–642. [CrossRef]
73. Pace, R.K. Nonparametric methods with applications to hedonic models. *J. Real Estate Financ. Econ.* **1993**, *7*, 185–204. [CrossRef]
74. Mason, C.; Quigley, J.M. Non-parametric hedonic housing prices. *Hous. Stud.* **1996**, *11*, 373–385. [CrossRef]
75. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons: Hoboken, NJ, USA, 1980.
76. Gilley, O.W.; Pace, R.K. On the Harrison and Rubinfeld data. *J. Environ. Econ. Manag.* **1996**, *31*, 403–405. [CrossRef]
77. Xu, P. The effect of errors-in-variables on variance component estimation. *J. Geod.* **2016**, *90*, 681–701. [CrossRef]
78. Simlai, P. Estimation of variance of housing prices using spatial conditional heteroskedasticity (SARCH) model with an application to Boston housing price data. *Q. Rev. Econ. Financ.* **2014**, *54*, 17–30. [CrossRef]
79. Ali, G.; Bashir, M.K.; Ali, H. Housing valuation of different towns using the hedonic model: A case of Faisalabad city, Pakistan. *Habitat Int.* **2015**, *50*, 240–249.
80. Poudyal, N.C.; Hodges, D.G.; Merrett, C.D. A hedonic analysis of the demand for and benefits of urban recreation parks. *Land Use Policy* **2009**, *26*, 975–983. [CrossRef]
81. Wen, H.; Zhang, Y.; Zhang, L. Do educational facilities affect housing price? An empirical study in Hangzhou, China. *Habitat Int.* **2014**, *42*, 155–163. [CrossRef]
82. Bao, H.; Wan, A. Improved estimators of hedonic housing price models. *J. Real Estate Res.* **2007**, *29*, 267–302. [CrossRef]
83. Mueller, J.M.; Loomis, J.B. Spatial dependence in hedonic property models: Do different corrections for spatial dependence result in economically significant differences in estimated implicit prices? *J. Agric. Resour. Econ.* **2008**, *33*, 212–231.
84. Helbich, M.; Böcker, L.; Dijst, M. Geographic heterogeneity in cycling under various weather conditions: Evidence from Greater Rotterdam. *J. Transp. Geogr.* **2014**, *38*, 38–47. [CrossRef]