ASSESSMENT OF MODEL SPECIFICATION APPROPRIATENESS FOR THE ESTIMATION OF HEDONIC PRICE FUNCTION OF HOUSES

MARIOS FOKAS TSAMICHAS

APRIL 2022

UNIVERSITY OF SOUTHAMPTON

UNIVERSITY OF SOUTHAMPTON

DEPARTMENT OF ECONOMICS

SCHOOL OF ECONOMIC, SOCIAL & POLITICAL SCIENCES

DISSERTATION: RESEARCH TOPICS

ASSESSMENT OF MODEL SPECIFICATION APPROPRIATENESS FOR THE ESTIMATION OF HEDONIC PRICE FUNCTION OF HOUSES

Full name of author:

MARIOS FOKAS TSAMICHAS

Presented for B.Sc. (Social Sciences) in Economics

10 March 2022

or

28 April 2022

I declare that this dissertation is my own work, and that where material is obtained from published or unpublished work, this has been fully acknowledged in the references.

Signed: 30499704 Date: 16/03/2022

Contents

List of Tables

List of Abbreviations

OLS	***Ordinary Least Squares***
HPM	***Hedonic Price Model***
SHPM	***Spatial House Price Model***

Introduction

This paper aims to further investigate the relationship of various environmental, mobility and socioeconomic parameters with house prices in the area of Boston through spatial-econometric tools and techniques. More specifically, the suggested model specification sheds light on the appropriate model and functional form of the parameters as to better determine the determinants of the Hedonic Price Function of houses with subsequent goal to enhance the existing literature. This research question was chosen as enhance the on-going research around modelling house prices and the implicit markets that determine the demand and supply of houses and their various attributes with a spatial focus, as well as to address some of the inhibitors that implicate the accuracy of such models. In this paper it was found that alongside conventional socioeconomic parameters, various environmental, geographical and mobility related factors were found to contribute substantially on median house prices in the Metropolitan area of Boston. Finally, it was found that spatial dependencies are present within the specified model that implies potential reconsideration of the model.

Existing literature has indicated that one of the most prominent approaches that aid in the analysis and assessment of Quality of Life is that of measuring both the individual and bundled contribution of various attributes and amenities, that are often non-marketable, of which an asset comprised of (Rosen, 1974). More specifically the above approach can take the form of market/residence approach which essentially measures individual preferences through market behavior and is known as hedonic strategy and is most often illustrated through the Hedonic Price Model (HPM) (Montero & Fernandez-Aviles, 2014).

The complexity of the above hedonic strategy is increased when attention is paid to spatial dependencies. Nevertheless, taking spatial dependencies into consideration is of great significance according to Tobler's first law of geography where everything is related to everything, but neighboring or adjacent entities and assets are even more related than others that are distant (Tobler, 2016).

The above intuition’s validity in the case of hedonic strategies has been thoroughly researched, especially in the context of modelling house prices. More specifically, literature indicates that the omission of spatial parameters that potentially exhibit spatial effects and dependencies among a model’s variables may lead to estimators being inconsistent, inefficient, and inaccurate regardless of the sophistication of the used method. (Anselin, 1988).

Economic Framework

Hedonic Price Model

Prior to describing our model specification that aims to address the above mentioned research question it is important to establish the economic framework on which this paper builds upon. The main economic framework our model will try to further develop is the Hedonic Price Model and more specifically its spatial variation the Spatial House Pricing Model. The HPM methodology is most often used in calculating and assessing the contribution of the individual attributes of which a heterogeneous asset is comprised of to the asset’s bundled value. Moreover, under specific assumptions of perfect information and competition, this methodology’s ability to estimate the implicit prices of characteristics or assets that are considered non-market goods has been proven extremely useful when it comes to assessing the economic value of heterogeneous goods such as houses (Montero & Fernandez-Aviles, 2014).

In our case additional value is generated through HPM methodology due to the flexibility it offers as there are no specific requirements when formulating the functional relation between the individual attributes of a house and its price. This enables researchers to experiment freely while simultaneously producing noteworthy results.

In this paper we aim to formulate a specification of the hedonic price function that accurately and efficiently captures median house price as well as illustrates potential spatial dependencies.

Dataset Description & Variable Selection

The data from which we will be evaluating and interpreting the results of our research question are obtained from the cross-sectional dataset named Boston House Prices Dataset with file name “Boston corrected” which contains 506 census tracts in the Boston Standard Metropolitan Area in the 1970’s. The dataset contains cross-sectional data of 14 variables that are of structural, environmental, mobility/neighborhood and geographical nature. Nevertheless, the suggested model specification will only be containing 12 of them due to the inconsistency and uncertainty of the contribution of the two excluded variables. More specifically, our model will be comprised of CMEDV the median house prices, CRIM, ZN, INDUS, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO and LSTAT as seen in Table 1 (Table 1).

Table 1 - Variable Selection

Variables used in model specifications

	Functional form	**Definition**
Dependent		
CMEDV	LogCMEDV	Median price for owner-occupied houses (in 1000$)
Independent		
ZN	ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	INDUS	Proportion of non-retail acres per town
CHAS	CHAS	Charles River dummy variable (=1 if tract bounds the Charles River, =0 if else)
NOX	NOX^2	Nitrogen Oxides concentration (parts per 10 million)
RM	RM^2	Average number of rooms per owner-occupied house
DIS	LogDIS	Weighted mean of distances to five Boston employment centers
RAD	LogRAD	Index of accessibility to radial highways
TAX	TAX	Full-value property-tax rate per 10,000$
PTRATIO	PTRATIO	Pupil-Teacher ratio by town
LSTAT	LSTAT	Lower status of population (in percentage %)
B	B	Black proportion of population

Variable Selection

In this paper's suggested specification, the included variables will be the independent variable CMEDV, and the dependent variables DIS, RAD, CHAS, INDUS, NOX, ZN, RM, TAX, PTRATIO, LSTAT, CRIM, B whose descriptive statistics can be seen below (Table 2).

Table 2 - Descriptive Statistics of Selected Variables

Variable	***Obs***	***Mean***	***Std. Dev.***
logCMEDV	506	1.317892	.1773115
ZN	506	11.36364	23.32245
INDUS	506	11.13678	6.860353
CHAS	506	0.06917	0.253994
NOX^2	506	0.3210877	0.1392125
RM^2	506	39.98932	9.079531
LogDIS	506	0.5159559	0.2343221
LogRAD	506	0.8111149	0.3799353
TAX	506	408.2372	168.5371
PTRATIO	506	18.45553	2.164946
LSTAT	506	12.65306	7.141062
B	506	356.674	91.29486

Econometric Analysis

This section of the dissertation is comprised of (i) the investigation of the model specification (ii) benefits and implications arising from this specification, (iii) analysis and interpretation of results and finally (iv) the investigation of the existence of spatial correlations and dependencies.

OLS Regression Model

The suggested model or housing price equation is constructed through OLS multiple linear regression methodology. Moreover, some of the abovementioned variables may need to be transformed as to contribute for a better fit of our model. Prior to the transformation a Shapiro-Wilks (Table 3) test was conducted to evaluate the normality of the above specified variables. This test indicated that all variables were not normally distributed and needed potential transformation. Nevertheless, it can be observed from the comparison of the two OLS models (Table 3) that the model was subject to a slight increase of the R squared when the variables CMEDV, DIS, RAD, NOX and RM were transformed into logCMEDV, logDIS, logRAD, NOX^2 and RM^2 (equation 1).

Table 3- Shapiro-Wilks test for Normality

Variable	***Obs***	***W***	***V***	***Z***	***Prob>z***
CMEDV	506	0.91979	27.274	7.951	0.00000
ZN	506	0.87065	43.981	9.100	0.00000
INDUS	506	0.91690	28.255	8.036	0.00000
CHAS	506	0.94891	17.373	6.866	0.00000
NOX	506	0.94352	19.206	7.108	0.00000
RM	506	0.96087	13.305	6.225	0.00000
DIS	506	0.90366	32.757	8.392	0.00000
RAD	506	0.72197	94.537	10.941	0.00000

TAX	506	0.84029	54.307	9.608	0.00000
PTRATIO	506	0.92629	25.064	7.748	0.00000
LSTAT	506	0.93691	21.451	7.374	0.00000
B	506	0.50336	168.871	12.336	0.00000

OLS specification:

CMEDV=a0+a1*INDUS+a2*ZN+a3*CHAS+a4*NOX+a5*RM+a6*DIS+a7*RAD+a8*TAX+a9*PTRATIO+a10*LSTAT+a11*B+ε (equation 1)

Before interpreting the coefficients and the contribution of each variable to the median house price (CMEDV) it is of great significance to take into consideration both the benefits of the chosen methodology and specification as well as their limitations. The selection of OLS multiple linear regression model as the appropriate methodology can be derived by addressing whether Gauss Markov's assumptions of OLS hold for our data (Hallin, 2006). However, the OLS specification has its limitations in terms of addressing misspecifications, endogeneity, normality of variables and omitted variable bias, some of which could potentially be addresses by the box-cox transformation (Osborne, 2010).

logCMEDV=a0+a1*INDUS+a2*ZN+a3*CHAS+a4*NOX2+a5*RM2

Spatial Regression Models

In this sub-section of our model's econometric analysis, we will be investigating the hypothesis that spatial dependencies are existent withing our model. Intuitively due to the nature neighborhood, geographic, accessibility and environmental nature of various specified variables some spatial correlations are expected to be present.

In order to verify the presence of spatial correlations two spatial regression models will be used, the spatial lag model that considers dependence in the explanatory variable CMEDV of a spatial unit, a house in our case, and its corresponding neighboring units (equation 2). The second model is the spatial error model that takes into considerations the spatial dependence in the error term of a house and its corresponding neighboring houses (equation 3). Additionally, Moran's I will be initially calculated as to measure the spatial autocorrelation of our model according to the given Spatial Weights Matrix (Saputro, 2019). While negative values will indicate that neighboring houses' attributes will have increasingly dissimilar values compared to ones further away (Anselin, 2002).

$$spatial\ lag\ model:\ y = \rho * Wy + X * \beta + \varepsilon \quad (equation\ 2)$$

$$spatial\ error\ model:\ y = X * \beta + u, \qquad where\ u = \lambda * Wu + \varepsilon \quad (equation\ 3)$$

Results

OLS regression

From the figures below that have been produced through the OLS regression of the specified model evaluation and interpretation of the variables' contribution to the median house price can be conducted (Table 4). It can be observed that all coefficients are highly statistically significant at 99% confidence level except for ZN whose coefficient is statistically significant at 90% confidence level and INDUS that is statistically significant. More specifically, an increase of 1 unit of INDUS does not contribute towards the CMEDV, an increase of 1 unit of ZN will increase CMEDV by 0.03044% , an increase of 1 unit of CHAS will increase logCMEDV by 5.12% , an increase of 1 unit of NOX^2 will decrease logCMEDV by 27.2%, an increase of 1 unit of RM^2 will increase logCMEDV by 0.37% , an increase of 1% of logDIS will decrease logCMEDV by 0.19%, an increase of 1% of logRAD will increase logCMEDV by 0.069%, an increase of TAX by 1 unit will decrease log CMEDV by 0.03%, an increase of PTRATIO by 1 unit will decrease logCMEDV by 1.43%, an increase of LSTAT by 1 unit will decrease logCMEDV by 1.36, an increase of 1 unit of B will increase logCMEDV by 0.021% and finally the constant has a fixed contribution f o1.744 towards logCMEDV(Table 4). From the above it can be deduced that most the variables in our specification model contribute towards the median house price and can be considered some of its determinants. As to reinforce the validity of the above specification, studies such as that of Rubinfeld & Harrison have found similar evidence with slight variations in their variable selection (Harrison & Rubinfeld , 1978).

Table 4 - OLS and Spatial Model Regression Results

Variable	**OLS equation**	**Spatial Lag Model**	**Spatial Error Model**
Dependent Variable	LogCMEDV	LogCMEDV	LogCMEDV
Independent Variables		**Coefficients**	
ZN	.0003044* (.0001717)	.0002729 (.0001742)	.0002667 (.0001737)
INDUS	.0011767 (.0007954)	.0008583 (.0007809)	.0009797 (.0007862)
CHAS	.0512394*** (.0164679)	.0544948*** (.0166714)	.05418*** (.0167937)
NOX^2	-.2718397*** (.058646)	-.2731919*** (0575579)	-.2735144*** (.057893)
LSTAT	-.0136794*** (.0014347)	-.0135065*** (.0013962)	-.0136238*** (.001408)
B	.000209*** (.0000667)	.0002357*** (.0000648)	.0002283*** (.0000656)
LogDIS	-.1858507*** (.0409457)	-.104989** (.0516451)	-.1300561*** (.0488374)
LogRAD	.0685649*** (.0162577)	.0618996*** (.0158965)	.0642389*** (.015957)
TAX	-.0002582*** (.0000432)	-.0002577*** (.0000419)	-.0002601*** (.0000423)
RM^2	.0036929*** (.0007973)	.0035822*** (.0007947)	.0036356*** (.0007951)
PTRATIO	-.0143149*** (.0016806)	-.0135728*** (.0016635)	-.0139402*** (.0016687)
Constant	1.744348***	1.693834***	1.712053***

	(.0832228)	(.0841658)	(.083818)
	Additional Statistics		
Rho	-	.0001777*** (.0000603)	-
Lambda	-	-	.0000878** (.0000434)
R2	0.7786	0.782	0.773
Log Likelihood	-	542.65379	540.87601

Note: *p<0.1; **p<0.05; ***p<0.01 and the values in the parentheses are Robust Standard Errors

Spatial Regression Models

Through calculating Global Moran's I in STATA for each of the specified variables the results were positive and highly statistically significant, alongside with their positive z-scores (Table 5). Meaning that the null hypothesis of the test that each investigated variable is randomly distributed among the spatial units or houses in Boston is rejected and that the spatial distribution of both the high and low values are spatially clustered (Kondo, 2021). Thus, spatial dependencies are present and spatial regressions are appropriate to be conducted (Lee, 2017).

Table 5- Global Moran's I

Variables	***Global Moran's I***
logCMEDV	1.1e+12*** (51.962)
ZN	6.9e+11*** (32.183)
INDUS	1.5e+12*** (71.502)
CHAS	3.7e+11*** (17.191)
NOX^2	1.6e+12*** (74.957)
RM^2	5.1e+11*** (24.049)
LogDIS	2.0e+12*** (95.171)
LogRAD	1.3e+12*** (61.058)
TAX	1.6e+12*** (76.267)
PTRATIO	5.7e+11*** (26.571)
LSTAT	1.2e+12*** (56.670)
B	6.7e+11*** (31.372)

Note: *p<0.1; **p<0.05; ***p<0.01 and the values in the parentheses are z-scores

From conducting a spatial lag model on the same variables as in the previous OLS econometric specifications a positive and highly statistically significant rho value of 0.0001777 was derived (Table 4).

Additionally, all the variables' coefficients are highly statistically significant at 99% confidence level besides LogDIS that is statistically significant at 95% confidence level and ZN and INDUS that are statistically insignificant similarly to the OLS model. The positive rho value indicates the rejection of the null hypothesis that there is no spatial correlation and subsequently that high prices of houses can be observed among neighboring houses and similarly for lower price houses. Furthermore, a value of rho other than zero also implies that OLS is biased, and inconsistent, thus spatial regression models are more appropriate to address this research question (Yamagata & Seya, 2020). Finally, from conducting a spatial error model of the chosen specification a lambda value of 0.0000878 was produced that is statistically significant at 95% confidence level. Similarly, all of the predictors were highly statistically significant at 99% confidence level except for ZN and INDUS that are statistically insignificant in all three regression models that have been formulated (Table 4). The positive lambda value verifies the presence of spatial correlation between the errors and implies that OLS is unbiased and consistent however its standard errors and coefficients are inefficient (Dubin, 1992).

Conclusion

Overall, from the abovementioned findings it has become evident that the usage of model methodologies and the construction of a model specification to estimate the Hedonic Price Function of houses is extremely complex and sophisticated, due to phenomena such as Multicollinearity, Endogeneity, Heteroskedasticity, model misspecification, omitted variable bias and the presence of spatial dependencies. Nevertheless, through this research paper light was shed on some of the above and issues by conducting substantial econometric analysis of both the OLS specifications as well as through two Spatial Regression Models. Moreover, this paper contributed significantly to assessing various model's appropriateness that aimed to accurately evaluate the contribution of various socioeconomic, structural, and environmental house attributes towards the median house price. More specifically, it was found that while the OLS estimation is much simpler and easy to interpret it was biased and didn’t accurately reflect the contributions to the median house prices due to the presence of spatial correlation. Consequently, it can be deduced that Spatial Regression Models can pose potentially a better fit to address the research question by accurately decomposing one of the main determinants of growth and quality of life, Real Estate, and house prices into various heterogeneous attributes that have unobserved economic value.

Appendix

STATA commands

OLS models

- ***Step 1: general model***

Line 1: reg CMEDV INDUS ZN CHAS NOX RM DIS RAD TAX PTRATIO LSTAT B

- ***Step 2: hettest option after step 1 that* performs various versions of the Breusch-Pagan (1979) and Cook-Weisberg (1983) tests to measure linear heteroskedasticity.**

Line 1: *Line 1: reg CMEDV INDUS ZN CHAS NOX RM DIS RAD TAX PTRATIO LSTAT B*

Line 2: hettest

- ***Step 3: general model with robust option to address heteroskedasticity:***

Line 3: reg CMEDV INDUS ZN CHAS NOX RM DIS RAD TAX PTRATIO LSTAT B, robust

- ***Step 4: specification model:***

Line 4: reg logCMEDV INDUS ZN CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B

- ***Step 5: hettest option after step 4 that* performs various versions of the Breusch-Pagan (1979) and Cook-Weisberg (1983) tests to measure linear heteroskedasticity.**

Line 4: reg logCMEDV INDUS ZN CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B

Line 5: hettest

- ***Step 6: specification model with robust option to address heteroskedasticity:***

Line 6: reg logCMEDV INDUS ZN CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, robust

- ***Step 7: vif command after step 6 as to calculate the variance inflation factors for the independent variables as to address multicollinearity***

Line 6: reg logCMEDV INDUS ZN CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, robust

Line 7: vif

Spatial Regression Models:

- ***Step 1: import Spatial Weights Matrix after being normalised manually on excel***

Line 9: spatwmat using "C:\normalised_wmatrix_stata.dta", name(aweights)

- ***Step 2: calculate Global Moran's I to find potential autocorrelation or spatial dependency of variables***

Line 10: spatgsa logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights) moran

- ***Step 3: perform a spatial lag model***

Line 10: spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights) eigenval(aweights) model(lag)

- ***Step 3: use robust option after step 3 to address heteroskedasticity***

Line 11: spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights) eigenval(aweights) model(lag) robust

- ***Step 4: perform a spatial error model***

Line 12: spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights) eigenval(aweights) model(error)

- ***Step 5: se robust option after step 4 to address heteroskedasticity***

Line 13: spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights) eigenval(aweights) model(error) robust

```
1   reg CMEDV ZN INDUS CHAS NOX RM DIS RAD TAX PTRATIO LSTAT B
2   hettest
3   reg CMEDV ZN INDUS CHAS NOX RM DIS RAD TAX PTRATIO LSTAT B, robust
4   reg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B
5   hettest
6   reg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, robust
7   vif
8
9   spatwmat using "C:\normalised_wmatrix_stata.dta", name(aweights)
10  spatgsa logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights)
    moran
11  spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights)
    eigenval(aweights) model(lag)
12  spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights)
    eigenval(aweights) model(lag) robust
13  spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B  , weights(aweights)
     eigenval(aweights) model(error)
14  spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B  , weights(aweights)
     eigenval(aweights) model(error) robust
```

Stata Output

```
. su logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B
```

Variable	Obs	Mean	Std. Dev.	Min	Max
logCMEDV	506	1.317892	.1773115	.69897	1.69897
ZN	506	11.36364	23.32245	0	100
INDUS	506	11.13678	6.860353	.46	27.74
CHAS	506	.06917	.253994	0	1
NOX2	506	.3210877	.1392125	.148225	.758641
RM2	506	39.98932	9.079531	12.68072	77.0884
LOGDIS	506	.5159559	.2343221	.0529247	1.083735
LOGRAD	506	.8111149	.3799353	0	1.380211
TAX	506	408.2372	168.5371	187	711
PTRATIO	506	18.45553	2.164946	12.6	22
LSTAT	506	12.65306	7.141062	1.73	37.97
B	506	356.674	91.29486	.32	396.9

```
. reg CMEDV ZN INDUS CHAS NOX RM DIS RAD TAX PTRATIO LSTAT B
```

Source	SS	df	MS
Model	31461.6423	11	2860.1493
Residual	11116.0964	494	22.5022195
Total	42577.7387	505	84.3123539

Number of obs	=	506
F(11, 494)	=	127.11
Prob > F	=	0.0000
R-squared	=	0.7389
Adj R-squared	=	0.7331
Root MSE	=	4.7437

CMEDV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ZN	.0438034	.0135703	3.23	0.001	.0171407	.070466
INDUS	.0306284	.061432	0.50	0.618	-.0900718	.1513286
CHAS	2.832453	.8590969	3.30	0.001	1.144519	4.520387
NOX	-16.97366	3.67492	-4.62	0.000	-24.19406	-9.753257
RM	3.827777	.4086112	9.37	0.000	3.024946	4.630607
DIS	-1.428409	.1892867	-7.55	0.000	-1.800316	-1.056503
RAD	.2463649	.0637231	3.87	0.000	.1211632	.3715666
TAX	-.0125569	.0037573	-3.34	0.001	-.0199392	-.0051747
PTRATIO	-.914722	.1303528	-7.02	0.000	-1.170836	-.6586076
LSTAT	-.5551201	.0470144	-11.81	0.000	-.647493	-.4627472
B	.0102531	.002661	3.85	0.000	.0050247	.0154814
_cons	35.29619	5.072972	6.96	0.000	25.32893	45.26345

```
. hettest

3reusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of CMEDV

         chi2(1)      =    12.71
         Prob > chi2  =   0.0004
```

```
. reg CMEDV ZN INDUS CHAS NOX RM DIS RAD TAX PTRATIO LSTAT B, robust

Linear regression                               Number of obs     =        506
                                                F(11, 494)        =     107.45
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7389
                                                Root MSE          =     4.7437
```

CMEDV	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
ZN	.0438034	.0130744	3.35	0.001	.0181151	.0694917
INDUS	.0306284	.049769	0.62	0.539	-.0671566	.1284134
CHAS	2.832453	1.292389	2.19	0.029	.2931956	5.37171
NOX	-16.97366	3.369956	-5.04	0.000	-23.59487	-10.35244
RM	3.827777	.7946215	4.82	0.000	2.266522	5.389031
DIS	-1.428409	.2183403	-6.54	0.000	-1.8574	-.9994192
RAD	.2463649	.0556737	4.43	0.000	.1369784	.3557514
TAX	-.0125569	.0026789	-4.69	0.000	-.0178205	-.0072934
PTRATIO	-.914722	.1094715	-8.36	0.000	-1.129809	-.6996348
LSTAT	-.5551201	.0873843	-6.35	0.000	-.7268108	-.3834294
B	.0102531	.0027547	3.72	0.000	.0048407	.0156654
_cons	35.29619	7.733126	4.56	0.000	20.10231	50.49006

```
. reg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B
```

Source	SS	df	MS
Model	12.3612825	11	1.12375296
Residual	3.51559008	494	.007116579
Total	15.8768726	505	.031439352

```
Number of obs     =        506
F(11, 494)        =     157.91
Prob > F          =     0.0000
R-squared         =     0.7786
Adj R-squared     =     0.7736
Root MSE          =     .08436
```

logCMEDV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ZN	.0003044	.0002285	1.33	0.183	-.0001446	.0007533
INDUS	.0011767	.0010875	1.08	0.280	-.0009599	.0033133
CHAS	.0501266	.0152919	3.28	0.001	.0200815	.0801718
NOX2	-.2671024	.0513615	-5.20	0.000	-.3680164	-.1661884
RM2	.0036929	.0005555	6.65	0.000	.0026015	.0047844
LOGDIS	-.1858507	.0319466	-5.82	0.000	-.2486186	-.1230828
LOGRAD	.0685649	.0200974	3.41	0.001	.0290779	.1080518
TAX	-.0002582	.0000563	-4.59	0.000	-.0003688	-.0001476
PTRATIO	-.0143149	.0023084	-6.20	0.000	-.0188504	-.0097794
LSTAT	-.0136794	.0008349	-16.38	0.000	-.0153199	-.0120389
B	.000209	.0000471	4.43	0.000	.0001164	.0003017
_cons	1.744348	.0660438	26.41	0.000	1.614587	1.87411

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of logCMEDV

         chi2(1)      =    97.17
         Prob > chi2  =   0.0000
```

```
. reg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, robust

Linear regression                               Number of obs     =        506
                                                F(11, 494)        =     164.76
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7786
                                                Root MSE          =     .08436
```

logCMEDV	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
ZN	.0003044	.0001717	1.77	0.077	-.0000329	.0006417
INDUS	.0011767	.0007954	1.48	0.140	-.000386	.0027394
CHAS	.0501266	.0164679	3.04	0.002	.0177709	.0824823
NOX2	-.2671024	.058646	-4.55	0.000	-.3823288	-.1518761
RM2	.0036929	.0007973	4.63	0.000	.0021264	.0052594
LOGDIS	-.1858507	.0409457	-4.54	0.000	-.2663	-.1054014
LOGRAD	.0685649	.0162577	4.22	0.000	.0366221	.1005076
TAX	-.0002582	.0000432	-5.98	0.000	-.000343	-.0001734
PTRATIO	-.0143149	.0016806	-8.52	0.000	-.0176169	-.011013
LSTAT	-.0136794	.0014347	-9.53	0.000	-.0164982	-.0108605
B	.000209	.0000667	3.13	0.002	.000078	.0003401
_cons	1.744348	.0832228	20.96	0.000	1.580834	1.907863

```
. vif
```

Variable	VIF	1/VIF
TAX	6.39	0.156504
LOGRAD	4.14	0.241702
LOGDIS	3.98	0.251481
INDUS	3.95	0.253201
NOX2	3.63	0.275643
LSTAT	2.52	0.396411
ZN	2.02	0.496239
RM2	1.81	0.553960
PTRATIO	1.77	0.564237
B	1.31	0.760718
CHAS	1.07	0.934139
Mean VIF	2.96	

```
. spatwmat using "C:\normalised_wmatrix_stata.dta", name(aweights)

The following matrix has been created:

1. Imported non-binary weights matrix aweights
   Dimension: 506x506

. spatgsa logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights) moran
```

Measures of global spatial autocorrelation

Weights matrix

Name: **aweights**
Type: **Imported (non-binary)**
Row-standardized: **No**

Moran's I

Variables	I	E(I)	sd(I)	z	p-value*
logCMEDV	1.1e+12	-0.002	2.1e+10	51.962	0.000
ZN	6.9e+11	-0.002	2.1e+10	32.183	0.000
INDUS	1.5e+12	-0.002	2.1e+10	71.502	0.000
CHAS	3.7e+11	-0.002	2.1e+10	17.191	0.000
NOX2	1.6e+12	-0.002	2.1e+10	74.957	0.000
RM2	5.1e+11	-0.002	2.1e+10	24.049	0.000
LOGDIS	2.0e+12	-0.002	2.1e+10	95.171	0.000
LOGRAD	1.3e+12	-0.002	2.1e+10	61.058	0.000
TAX	1.6e+12	-0.002	2.1e+10	76.267	0.000
PTRATIO	5.7e+11	-0.002	2.1e+10	26.571	0.000
LSTAT	1.2e+12	-0.002	2.1e+10	56.670	0.000
B	6.7e+11	-0.002	2.1e+10	31.372	0.000

*1-tail test

```
. spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights) eigenval(aweights) model(lag)

initial:       log likelihood =  539.18512
rescale:       log likelihood =  539.18512
rescale eq:    log likelihood =  539.18512
Iteration 0:   log likelihood =  539.18512
Iteration 1:   log likelihood =  542.61025
Iteration 2:   log likelihood =  542.65377
Iteration 3:   log likelihood =  542.65379
```

```
Weights matrix
 Name: aweights
 Type: Imported (non-binary)
 Row-standardized: No


Spatial lag model                              Number of obs    =       506
                                               Variance ratio   =     0.782
                                               Squared corr.    =     0.782
Log likelihood = 542.65379                     Sigma            =      0.08
```

logCMEDV	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logCMEDV						
ZN	.0002729	.0002246	1.22	0.224	-.0001673	.0007131
INDUS	.0008583	.0010743	0.80	0.424	-.0012472	.0029638
CHAS	.0544948	.0151016	3.61	0.000	.0248962	.0840933
NOX2	-.2731919	.0504649	-5.41	0.000	-.3721013	-.1742825
RM2	.0035822	.0005469	6.55	0.000	.0025104	.004654
LOGDIS	-.104989	.0440383	-2.38	0.017	-.1913025	-.0186756
LOGRAD	.0618996	.0198895	3.11	0.002	.0229168	.1008823
TAX	-.0002577	.0000553	-4.66	0.000	-.000366	-.0001494
PTRATIO	-.0135728	.0022834	-5.94	0.000	-.0180482	-.0090975
LSTAT	-.0135065	.0008222	-16.43	0.000	-.0151179	-.0118951
B	.0002357	.0000474	4.97	0.000	.0001428	.0003285
_cons	1.693834	.0676389	25.04	0.000	1.561264	1.826404
rho	.0001777	.000068	2.61	0.009	.0000445	.0003109

```
Wald test of rho=0:                  chi2(1) =   6.838 (0.009)
Likelihood ratio test of rho=0:      chi2(1) =   6.793 (0.009)
Lagrange multiplier test of rho=0:   chi2(1) =   7.028 (0.008)

Acceptable range for rho: -2.868 < rho < 0.075
```

```
. spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B, weights(aweights) eigenval(aweights) model(lag) robust

initial:       log pseudolikelihood =  539.18512
rescale:       log pseudolikelihood =  539.18512
rescale eq:    log pseudolikelihood =  539.18512
Iteration 0:   log pseudolikelihood =  539.18512
Iteration 1:   log pseudolikelihood =  542.61025
Iteration 2:   log pseudolikelihood =  542.65377
Iteration 3:   log pseudolikelihood =  542.65379

Weights matrix
 Name: aweights
 Type: Imported (non-binary)
 Row-standardized: No

Spatial lag model                                Number of obs    =        506
                                                 Variance ratio   =      0.782
                                                 Squared corr.    =      0.782
Log likelihood = 542.65379                       Sigma            =       0.08
```

logCMEDV	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
logCMEDV						
ZN	.0002729	.0001742	1.57	0.117	-.0000686	.0006144
INDUS	.0008583	.0007809	1.10	0.272	-.0006722	.0023889
CHAS	.0544948	.0166714	3.27	0.001	.0218195	.0871701
NOX2	-.2731919	.0575579	-4.75	0.000	-.3860033	-.1603805
RM2	.0035822	.0007947	4.51	0.000	.0020247	.0051397
LOGDIS	-.104989	.0516451	-2.03	0.042	-.2062117	-.0037664
LOGRAD	.0618996	.0158965	3.89	0.000	.0307431	.0930561
TAX	-.0002577	.0000419	-6.15	0.000	-.0003398	-.0001756
PTRATIO	-.0135728	.0016635	-8.16	0.000	-.0168333	-.0103124
LSTAT	-.0135065	.0013962	-9.67	0.000	-.016243	-.01077
B	.0002357	.0000648	3.64	0.000	.0001087	.0003626
_cons	1.693834	.0841658	20.12	0.000	1.528872	1.858796
rho	.0001777	.0000603	2.95	0.003	.0000596	.0002958

```
Wald test of rho=0:                     chi2(1) =   8.699 (0.003)
Lagrange multiplier test of rho=0:      chi2(1) =   7.028 (0.008)
```

```
. spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B  , weights(aweights) eigenval(aweights) model(error)

initial:       log likelihood =  539.18512
rescale:       log likelihood =  539.18512
rescale eq:    log likelihood =  539.18512
Iteration 0:   log likelihood =  539.18512
Iteration 1:   log likelihood =  540.85952
Iteration 2:   log likelihood =  540.87601
Iteration 3:   log likelihood =  540.87601


Weights matrix
 Name: aweights
 Type: Imported (non-binary)
 Row-standardized: No


Spatial error model                               Number of obs   =        506
                                                  Variance ratio  =      0.864
                                                  Squared corr.   =      0.773
Log likelihood = 540.87601                        Sigma           =       0.08

------------------------------------------------------------------------------
    logCMEDV |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
logCMEDV     |
          ZN |   .0002667   .0002265     1.18   0.239    -.0001773    .0007107
       INDUS |   .0009797   .0010771     0.91   0.363    -.0011314    .0030908
        CHAS |     .05418   .0152564     3.55   0.000     .0242779     .084082
        NOX2 |  -.2735144   .0507218    -5.39   0.000    -.3729273   -.1741014
         RM2 |   .0036356   .0005487     6.63   0.000     .0025602     .004711
      LOGDIS |  -.1300561   .0441375    -2.95   0.003     -.216564   -.0435482
      LOGRAD |   .0642389   .0199738     3.22   0.001      .025091    .1033868
         TAX |  -.0002601   .0000555    -4.68   0.000     -.000369   -.0001513
     PTRATIO |  -.0139402   .0022888    -6.09   0.000    -.0184262   -.0094542
       LSTAT |  -.0136238   .0008248   -16.52   0.000    -.0152404   -.0120072
           B |   .0002283   .0000477     4.79   0.000     .0001349    .0003217
       _cons |   1.712053   .0675702    25.34   0.000     1.579618    1.844489
-------------+----------------------------------------------------------------
      lambda |   .0000878   .0000491     1.79   0.074    -8.53e-06    .0001841
------------------------------------------------------------------------------
Wald test of lambda=0:                 chi2(1) =    3.191 (0.074)
Likelihood ratio test of lambda=0:     chi2(1) =    3.237 (0.072)
Lagrange multiplier test of lambda=0:  chi2(1) =   96.894 (0.000)

Acceptable range for lambda: -2.868 < lambda < 0.075
```

```
. spatreg logCMEDV ZN INDUS CHAS NOX2 RM2 LOGDIS LOGRAD TAX PTRATIO LSTAT B  , weights(aweights) eigenval(aweights) model(error) robust

initial:       log pseudolikelihood =  539.18512
rescale:       log pseudolikelihood =  539.18512
rescale eq:    log pseudolikelihood =  539.18512
Iteration 0:   log pseudolikelihood =  539.18512
Iteration 1:   log pseudolikelihood =  540.85952
Iteration 2:   log pseudolikelihood =  540.87601
Iteration 3:   log pseudolikelihood =  540.87601

Weights matrix
 Name: aweights
 Type: Imported (non-binary)
 Row-standardized: No

Spatial error model                               Number of obs    =       506
                                                  Variance ratio   =     0.864
                                                  Squared corr.    =     0.773
Log likelihood = 540.87601                        Sigma            =      0.08
```

logCMEDV	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
logCMEDV						
ZN	.0002667	.0001737	1.54	0.125	-.0000737	.0006071
INDUS	.0009797	.0007862	1.25	0.213	-.0005613	.0025207
CHAS	.05418	.0167937	3.23	0.001	.021265	.0870949
NOX2	-.2735144	.057893	-4.72	0.000	-.3869826	-.1600461
RM2	.0036356	.0007951	4.57	0.000	.0020772	.005194
LOGDIS	-.1300561	.0488374	-2.66	0.008	-.2257757	-.0343365
LOGRAD	.0642389	.015957	4.03	0.000	.0329638	.095514
TAX	-.0002601	.0000423	-6.15	0.000	-.0003431	-.0001772
PTRATIO	-.0139402	.0016687	-8.35	0.000	-.0172108	-.0106696
LSTAT	-.0136238	.001408	-9.68	0.000	-.0163834	-.0108643
B	.0002283	.0000656	3.48	0.000	.0000998	.0003568
_cons	1.712053	.083818	20.43	0.000	1.547773	1.876334
lambda	.0000878	.0000434	2.02	0.043	2.70e-06	.0001728

```
Wald test of lambda=0:                   chi2(1) =   4.090 (0.043)
Lagrange multiplier test of lambda=0:    chi2(1) =  96.894 (0.000)

Acceptable range for lambda: -2.868 < lambda < 0.075
```

Bibliography

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Dordrecht: Springer.

Anselin, L. (2002). Under the hood Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 247-267.

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 433-452.

Hallin, M. (2006). Gauss–Markov Theorem in Statistics. *Encyclopedia of Environmetrics*.

Harrison, D., & Rubinfeld , D. L. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manage*, 81-84.

Kondo, K. (2021, July 9). *@inproceedings{Kondo2018TestingFG,.* Retrieved from Semantic Scholar: http://fmwww.bc.edu/RePEc/bocode/m/moransi.pdf

Lee, S.-I. (2017). Correlation and Spatial Autocorrelation. In S. Shekhar, H. Xiong, & X. Zhou, *Encyclopedia of GIS* (pp. 360-365). Springer.

Montero, J., & Fernandez-Aviles, G. (2014). Hedonic Price Model. *Encyclopedia of Quality of Life and Well-Being Research.*

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation: Vol. 15 , Article 12.*

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 34-55.

Saputro, D. R. (2019). Spatial autoregressive with a spatial. *ournal of Physics*, 1-3.

Tobler, W. R. (2016). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 234-240.

Yamagata, Y., & Seya, H. (2020). Chapter Five - Spatial econometric models. In Y. Yamagata , & H. Seya, *Spatial Analysis Using Big Data* (pp. 115-119). Elsevier.