

# Homework 2: Using Spatial Lag, Spatial Error and Geographically Weighted Regression to Predict Median House Values in Philadelphia Block Groups

Yiming Cao, Sujan Kakumanu, Angel Sanaa Rutherford

2025-10-30

```
Reading layer `RegressionData' from data source
  `C:\Users\Angel\Documents\GitHub\MUSA-5000\HW 1\Lecture 1 - RegressionData.shp'
  using driver `ESRI Shapefile'
Simple feature collection with 1720 features and 13 fields
Geometry type: POLYGON
Dimension:      XY
Bounding box:   xmin: 2660605 ymin: 207610.6 xmax: 2750171 ymax: 304858.8
CRS:            NA
```

```
Warning in poly2nb(Regression_shpData, row.names = Regression_shpData$POLY_ID): neighbour ob
if this sub-graph count seems unexpected, try increasing the snap argument.
```

```
Neighbour list object:
Number of regions: 1720
Number of nonzero links: 10526
Percentage nonzero weights: 0.3558004
Average number of links: 6.119767
2 disjoint connected subgraphs
Link number distribution:
```

```
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  18  27
  4  16  52 175 348 493 344 177  62  28  10   4   2   1   1   1   1   1
4 least connected regions:
441 708 1391 1665 with 1 link
1 most connected region:
1636 with 27 links
```

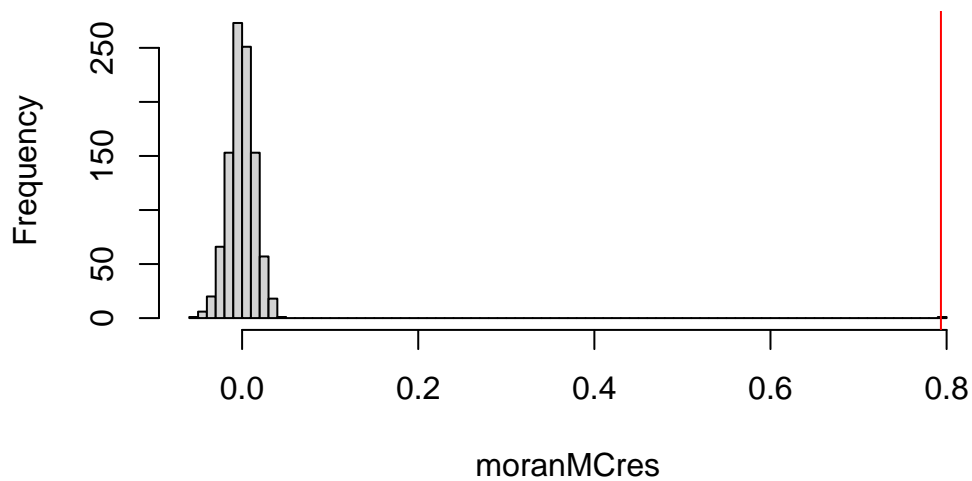
```
[1] 0.793565
```

Monte-Carlo simulation of Moran I

```
data: Regression_shpData$LNMEDHVAL
weights: queenlist
number of simulations + 1: 1000
```

```
statistic = 0.79356, observed rank = 1000, p-value <
0.000000000000000022
alternative hypothesis: two.sided
```

### Histogram of moranMCres



	Ii	E.Ii	Var.Ii	Z.Ii	Pr(z != E(Ii))
1	5.35196819	-0.003049833231	1.3051455983	4.687394	0.0000027670601
2	4.41225942	-0.002216601273	0.7590492452	5.066922	0.0000004043007
3	3.50068095	-0.003049833231	0.7444928827	4.060696	0.0000489266898
4	2.44447746	-0.000843880799	0.2410048919	4.981074	0.0000006323230
5	1.88349103	-0.001094174334	0.6259107138	2.382098	0.0172143256212
6	0.09949306	-0.000001607927	0.0009208032	3.278811	0.0010424535943

Call:

```
lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +  
    LNNBELPOV, data = Regression_shpData)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.25817	-0.20391	0.03822	0.21743	2.24345

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.1137781	0.0465318	238.843	< 0.0000000000000002 ***
PCTVACANT	-0.0191563	0.0009779	-19.590	< 0.0000000000000002 ***
PCTSINGLES	0.0029770	0.0007032	4.234	0.0000242 ***
PCTBACHMOR	0.0209095	0.0005432	38.494	< 0.0000000000000002 ***
LNNBELPOV	-0.0789035	0.0084567	-9.330	< 0.0000000000000002 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

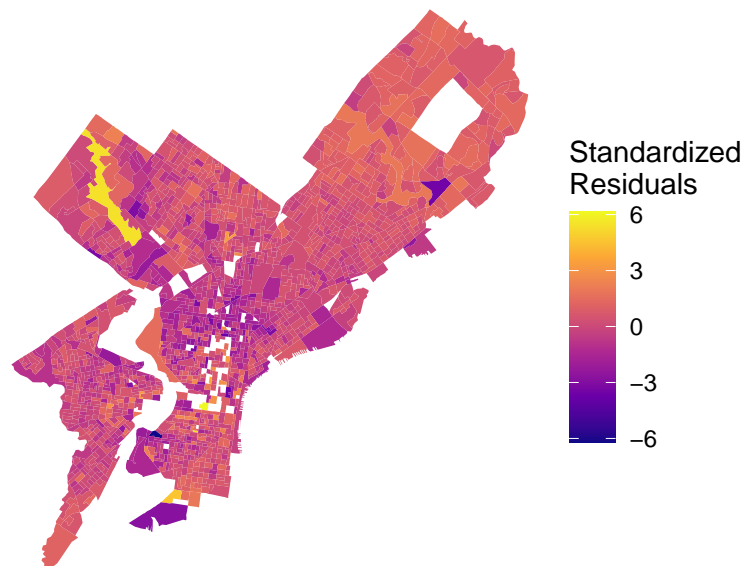
Residual standard error: 0.3665 on 1715 degrees of freedom

Multiple R-squared: 0.6623, Adjusted R-squared: 0.6615

F-statistic: 840.9 on 4 and 1715 DF, p-value: < 0.00000000000000022

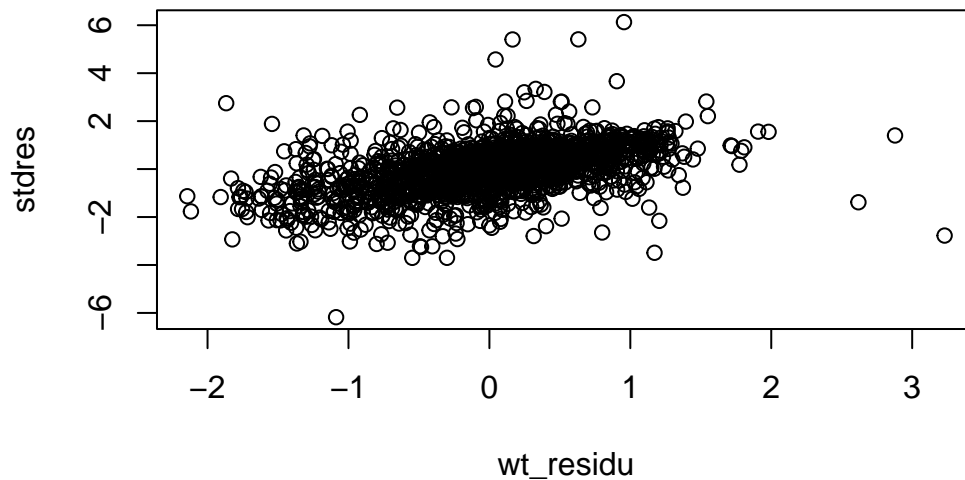
'log Lik.' -711.4933 (df=6)

## p of Standardized Regression Residuals



```
#| echo: false
wt_residu <- sapply(queen, function(x) mean(stdres[x]))

plot(wt_residu, stdres)
```



```
#Note the beta coefficient of the wt_residu. 0.73235. This suggests that there is spatial au
res.lm <- lm(formula=stdres ~ wt_residu)
summary(res.lm)
```

Call:

```
lm(formula = stdres ~ wt_residu)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.3685	-0.4450	0.0585	0.4618	5.4435

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.01281	0.02121	-0.604	0.546
wt_residu	0.73235	0.03244	22.576	<0.0000000000000002 ***

---

Residual standard error: 0.8793 on 1718 degrees of freedom  
Multiple R-squared: 0.2288, Adjusted R-squared: 0.2283  
F-statistic: 509.7 on 1 and 1718 DF, p-value: < 0.000000000000000022

Monte-Carlo simulation of Moran I

```
statistic = 0.3124, observed rank = 1000, p-value < 0.00000000000000022
alternative hypothesis: two.sided
```

[illegible]

```
Call:lagsarlm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
  LNNBELPOV, data = Regression_shpData, listw = queenlist)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.655421	-0.117248	0.018654	0.133126	1.726436

Type: lag

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	3.89845489	0.20111357	19.3843	< 0.00000000000000022
PCTVACANT	-0.00852940	0.00074367	-11.4694	< 0.00000000000000022
PCTSINGLES	0.00203342	0.00051577	3.9425	0.00008063503
PCTBACHMOR	0.00851381	0.00052193	16.3120	< 0.00000000000000022
LNNBELPOV	-0.03405466	0.00629287	-5.4116	0.00000006246

Rho: 0.6511, LR test value: 911.51, p-value: < 0.000000000000000222

Asymptotic standard error: 0.01805

z-value: 36.072, p-value: < 0.000000000000000222

Wald statistic: 1301.2, p-value: < 0.000000000000000222

Log likelihood: -255.74 for lag model

ML residual variance (sigma squared): 0.071948, (sigma: 0.26823)

Number of observations: 1720

Number of parameters estimated: 7

AIC: 525.48, (AIC for lm: 1435)

LM test for residual autocorrelation

test value: 67.737, p-value: 0.00000000000000022204

Monte-Carlo simulation of Moran I

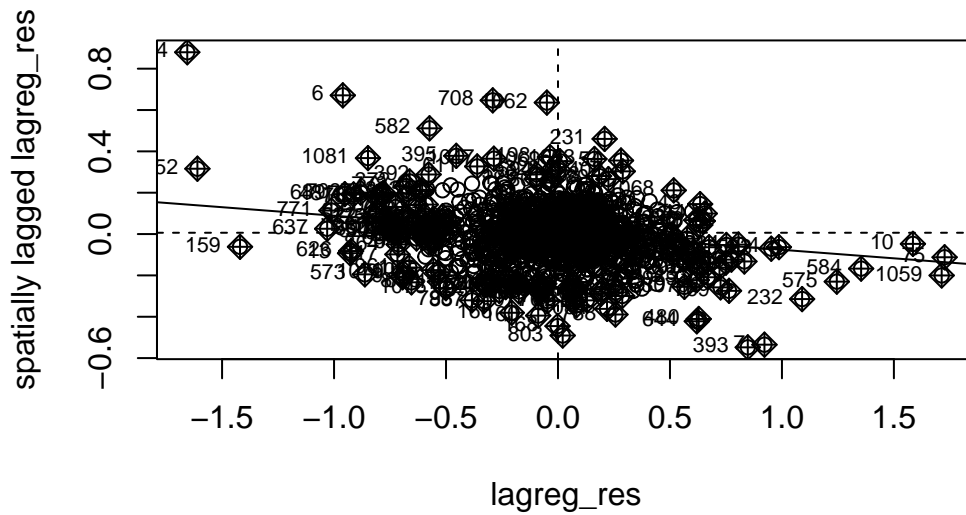
data: lagreg\_res

weights: queenlist

number of simulations + 1: 1000

statistic = -0.082412, observed rank = 1, p-value = 0.002

alternative hypothesis: two.sided



```
Call:errorsarlm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
  LNNBELPOV, data = Regression_shpData, listw = queenlist)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.926477	-0.115408	0.014889	0.133852	1.948664

Type: error

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	10.90643422	0.05346780	203.9814	< 0.00000000000000022
PCTVACANT	-0.00578308	0.00088670	-6.5220	0.00000000006937
PCTSINGLES	0.00267792	0.00062083	4.3134	0.00001607388269
PCTBACHMOR	0.00981293	0.00072896	13.4615	< 0.00000000000000022
LNNBELPOV	-0.03453408	0.00708933	-4.8713	0.00000110882040

Lambda: 0.81492, LR test value: 677.61, p-value: < 0.000000000000000222

Asymptotic standard error: 0.016373

z-value: 49.772, p-value: < 0.000000000000000222

Wald statistic: 2477.2, p-value: < 0.000000000000000222

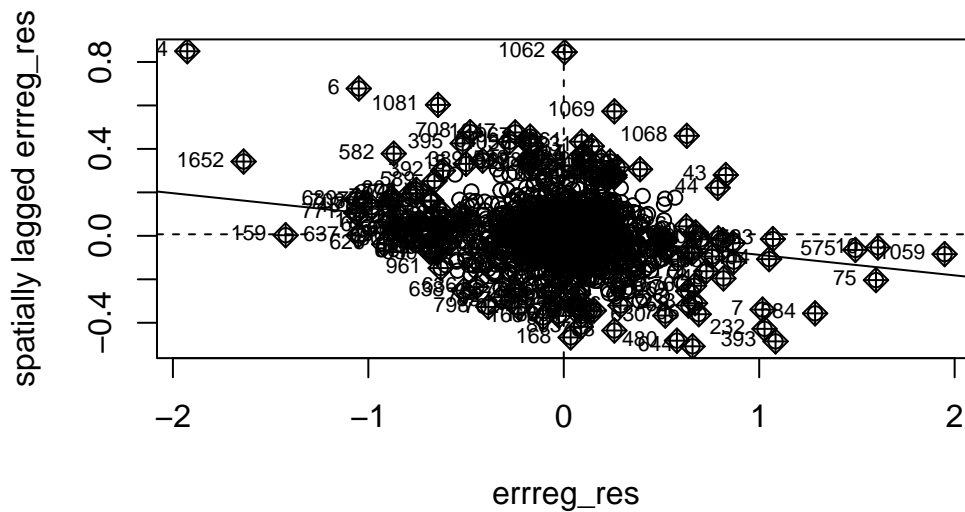
Log likelihood: -372.6904 for error model

ML residual variance (sigma squared): 0.076551, (sigma: 0.27668)  
Number of observations: 1720  
Number of parameters estimated: 7  
AIC: 759.38, (AIC for lm: 1435)

Monte-Carlo simulation of Moran I

data: errreg\_res  
weights: queenlist  
number of simulations + 1: 1000

statistic = -0.094532, observed rank = 1, p-value = 0.002  
alternative hypothesis: two.sided





# 1 Introduction (Google Doc at end)

## 2 Methods

### 2.1 The Concept of Spatial Autocorrelation (Angel)

Spatial autocorrelation builds upon the 1st Law of Geography, Tobler's Law, which states that all things are related but near things are more related than further observations.

Spatial autocorrelation examines the relationship of values of the same variable at nearby locations. A positive relationship is indicated by related values at nearby locations and a negative relationship is indicated by significantly different values at nearby locations. Moran's I is the statistical measure used to detect spatial autocorrelation, giving a value closer to 1 for positive spatially dependent values, -1 for negative spatially dependent values, and 0 for values that exhibit no spatial autocorrelation. The formula for calculating Moran's I is as follows:

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^N \sum_{j=1}^N W_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

$\bar{X}$  is the mean of the variable X,  $X_i$  is the variable value at a particular location  $i$ ,  $X_j$  is the variable value at another location  $j$ ,  $W_{ij}$  is the value given by the weight matrix of location  $i$  relative to  $j$ , and  $n$  is the number of observations. This formula captures the covariance between neighboring values, standardized by the overall variance. For each pair of locations  $i$  and  $j$ , Moran's I measures how similarly their values deviate from the mean, multiplies those deviations, weights the result by their spatial proximity, and normalizes by the overall variance.

This analysis uses a queen contiguity weight matrix, which defines neighbors as polygons sharing either a border or a vertex. The matrix is square,  $n \times n$ , where  $n$  is the number of observations, that assigns a value of 1 for neighbors and 0 otherwise. Unless there is a strong theoretical reason for using a particular weight matrix, statisticians will sometimes use multiple weight matrices (other contiguity-based measurements like rook or distance-based measurements) to ensure results are robust rather than just a byproduct of a single spatial definition.

In R, we can assess the statistical significance of Moran's I using a Monte-Carlo permutation test via the `moran.mc()` function. This method computes Moran's I for the original variable, then randomly shuffles the variable values 999 times, recalculating Moran's I for each permutation. We evaluate where our original Moran's I falls by ranking it relative to the randomly permuted Moran's I, either in descending order for positive autocorrelation or ascending for negative. A pseudo p-value is then calculated by dividing the rank of the original Moran's I by the total number of permutations, estimating the likelihood of observing such a value under

spatial randomness. We test for the null hypothesis,  $H_0$ , no spatial autocorrelation against the two-sided alternative  $H_a$ , positive spatial autocorrelation or negative spatial autocorrelation. Visual diagnostics include a histogram of the permuted data's Moran's I values with the original Moran's I highlighted, and a Moran scatterplot comparing the block groups' original variable values to the average of its neighbors, also known as spatially lagged variable values. A clear pattern in the scatterplot would suggest spatial autocorrelation while randomness implies no spatial autocorrelation.

Local spatial autocorrelation examines how similar or dissimilar values at one location are to nearby locations. Rather than describing the overall global spatial patterns, local spatial autocorrelation pinpoints areas of spatial clustering or spatial outliers. We test for local spatial autocorrelation using LISA, Local Indices of Spatial Association. In this case, we use the local Moran's I as a statistical measure of local spatial autocorrelation. In R, we compute Local Moran's I using the `localmoran()` function. Conceptually, the statistic is calculated by taking the deviation of a value at location  $i$  from the global mean, multiplying it by the weighted average of its neighbors' deviations, and normalizing by the total variance across all locations. The null hypothesis for local spatial autocorrelation is  $H_0$ , no local spatial autocorrelation at location  $i$  while the two-sided alternative hypothesis is  $H_a$ , a positive or negative spatial autocorrelation at location  $i$ . The `localmoran()` function implements a permutation-based test for statistical significance. For each location, the value at  $i$  is held constant while the values of its neighbors are randomly shuffled. A two-sided pseudo p-value is then computed based on the rank of the original  $I_i$  relative to the permuted values' Moran's I. This p-value is returned by indexing the results of the `localmoran` function with  $\Pr(z \neq E(I_i))$  and can be conceptualized as the probability that the observed  $I_i$  is significantly different from what we'd expect under spatial randomness in either direction (positive or negative). Visually, we can create a map that shows the spatial distribution of Local Moran's I p-values and clusters to further assess for local spatial autocorrelation.

## 2.2 A Review of OLS Regression and Assumptions (Angel)

Ordinary Least Squares (OLS) regression estimates the relationship between a dependent variable and one or more independent variables by minimizing the sum of squared differences between observed and predicted values. In the context of multiple regression, OLS quantifies the unique contribution of each predictor to the outcome while controlling for the influence of the others. Unlike simple regression, which models the dependent variable using a single predictor, multiple regression incorporates several predictors, each with a coefficient representing its effect on the dependent variable. Multiple regression relies on several key assumptions, most of which mirror the assumptions of simple regression: linearity between the dependent variable and each predictor, normally distributed residuals, randomness of residuals— indicating that observations are not systematically related, homoscedastic of residuals or constant variance across all values, a continuous dependent variable, and, a unique assumption for multiple regression, no perfect multicollinearity between predictors. A more comprehensive overview of

Ordinary Least Squares (OLS) regression can be found in Homework 1: *Using OLS Regression to Predict Median House Values in Philadelphia*.

As previously mentioned, a core assumption of OLS regression is the randomness of residuals or, in other words, the independence of residuals from one another. When spatial autocorrelation is present, this indicates that values of a variable at nearby locations are related to one another, violating the assumption of independence. As a result, the OLS error term may contain insightful spatial patterns rather than random noise and beta coefficients may be inefficient estimates. In practice, this can manifest as systematic over- or under-prediction. We can statistically quantify the spatial autocorrelation of residuals by calculating the Moran's I of the OLS residuals, Moran's I which we previously introduced as a measure of bidirectional spatial dependence.

Another way to assess spatial dependence in OLS residuals is to regress each residual on its spatially lagged counterpart. A spatial lag refers to the value of a variable at neighboring locations, which, in this case, are defined by the queen weights matrix. By creating spatially lagged residuals and regressing each residual on its lagged value, we can test whether residuals are systematically related across space. In the statistical summary of this regression, the slope  $b$  represents the coefficient of the lagged residuals when predicting the original residuals. This slope quantifies the strength and direction of spatial dependence. If it is significantly different from zero, it suggests that residuals are spatially autocorrelated, meaning the assumption of independence is violated and the OLS estimates may be compromised. Visually, this relationship can be assessed using a scatterplot of OLS residuals against the weight or spatially lagged residuals.

In addition to estimating regression coefficients, we can use various libraries in R to perform statistical tests that assess other assumptions of OLS. One assumption we can test for is homoscedasticity, which refers to the constant variance of residuals across all predicted values. If residuals vary systematically with predicted values, this indicates heteroscedasticity, or non-constant variance, which violates the assumption. This assumption is closely tied to the independence of errors as residuals that show heteroscedasticity could imply non-random or dependent variance which can compromise estimation. A simple visual diagnostic involves plotting OLS residuals by predicted residuals. In R, we can use the `whitestrip` and `lmtest` library to perform three commonly used statistical tests: the Breusch-Pagan Test, the Koenker-Bassett Test or studentized Breusch-Pagan Test, and the White Test. Each test evaluates the null hypothesis of homoscedasticity or no heteroscedasticity and the alternative hypothesis of heteroscedasticity. If the result p-value for these tests is less than 0.05, then we can reject the null hypothesis for the alternate hypothesis of heteroscedasticity.

Another key assumption of OLS regression is the normality of errors. Residuals should behave like random noise, containing no systematic structure, and should follow a normal distribution. A simple visual diagnostic of residual normality is to plot residuals using a histogram. In R, we can perform a statistical diagnostic of normality by using the Jarque-Bera Test, available through the `tseries` package. The null hypothesis for the Jarque-Bera Test is that the residuals follow a normal distribution while the alternative hypothesis is non-normality or

non-normal distribution. If the resulting p-value is less than 0.05, we reject the null hypothesis in favor of the alternative, indicating a violation of the normality assumption.

## 2.3 Spatial Lag and Spatial Error Regression (Sujan)

## 2.4 Geographically Weighted Regression (Ming)

# 3 Results

## 3.1 Spatial Autocorrelation (Angel)

### Random Permutation Test/ Monte-Carlo Simulation Table of LNMEDHVAL

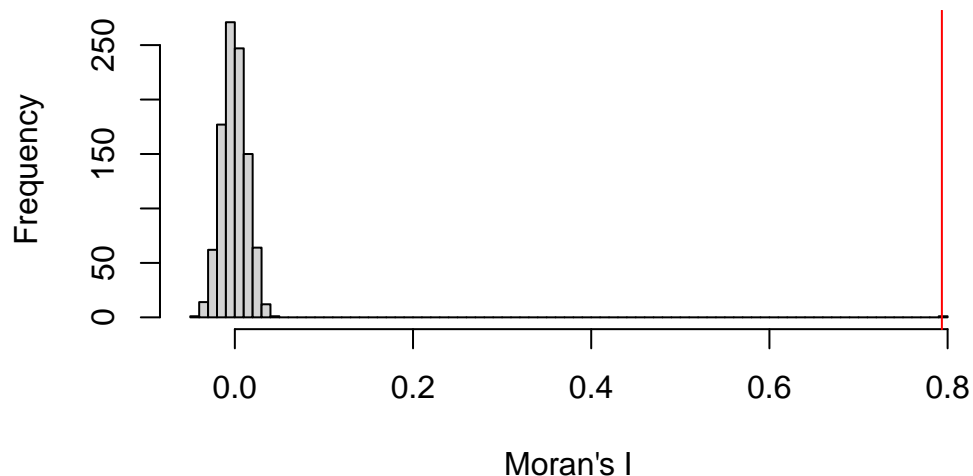
Monte-Carlo simulation of Moran I

```
data: Regression_shpData$LNMEDHVAL
weights: queenlist
number of simulations + 1: 1000
```

```
statistic = 0.79356, observed rank = 1000, p-value < 2.2e-16
alternative hypothesis: two.sided
```

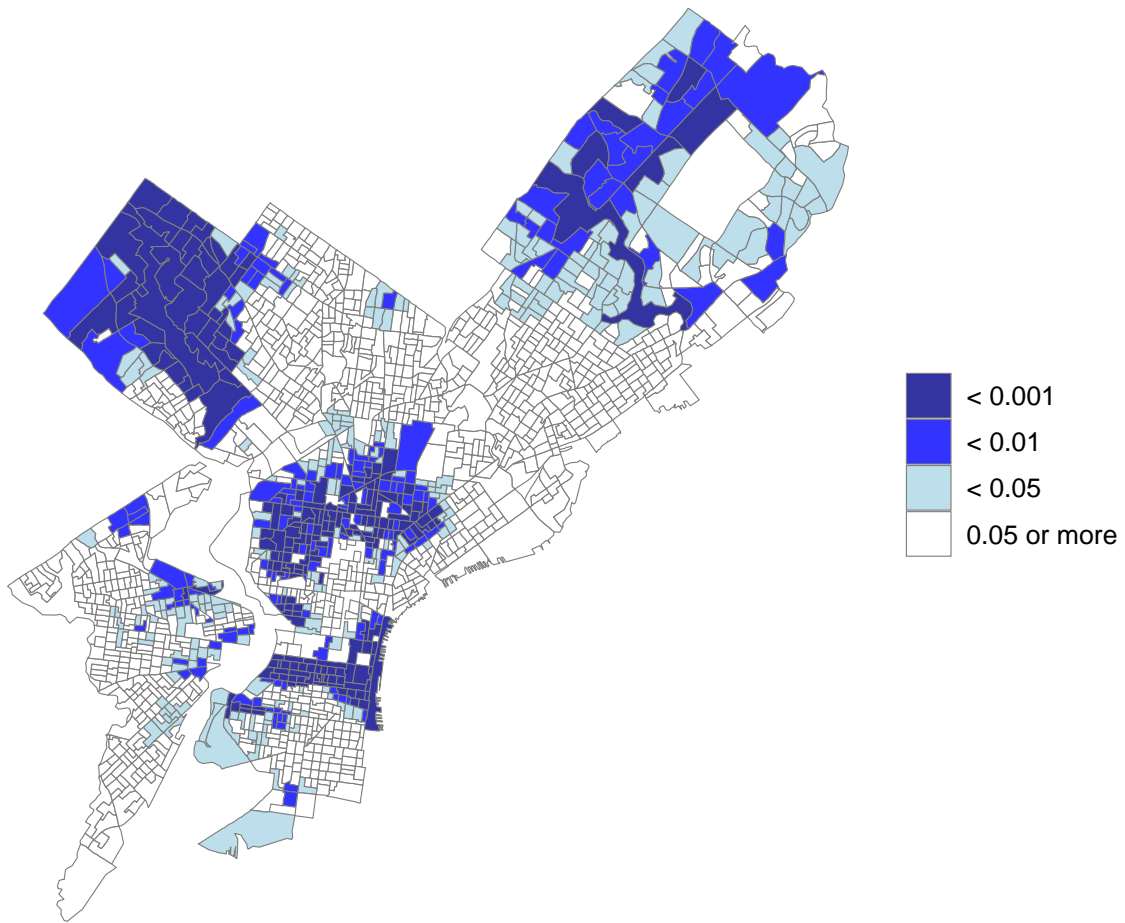
Our global Moran's I value, 0.793, of our dependent variable was considerably different from 0, indicating high positive spatial autocorrelation. Our random permutation test suggests that our Moran's I was statistically significant as it returned a p-value of less than 0.00000000000000022 which falls into the statistically significant threshold ( $p < 0.05$ ). The rank of our observed Moran's I further enforces our findings, suggesting that none of the 999 randomly permuted simulations produced a value as extreme as the one observed. Thus, we can reject the Moran's I null hypothesis of no spatial autocorrelation as well as consider LNMEDHVAL to be significantly spatially autocorrelated.

**Distribution of Moran's I Values From Logged Median House  
Permutation Tests  
Red Line = Observed Moran's I**

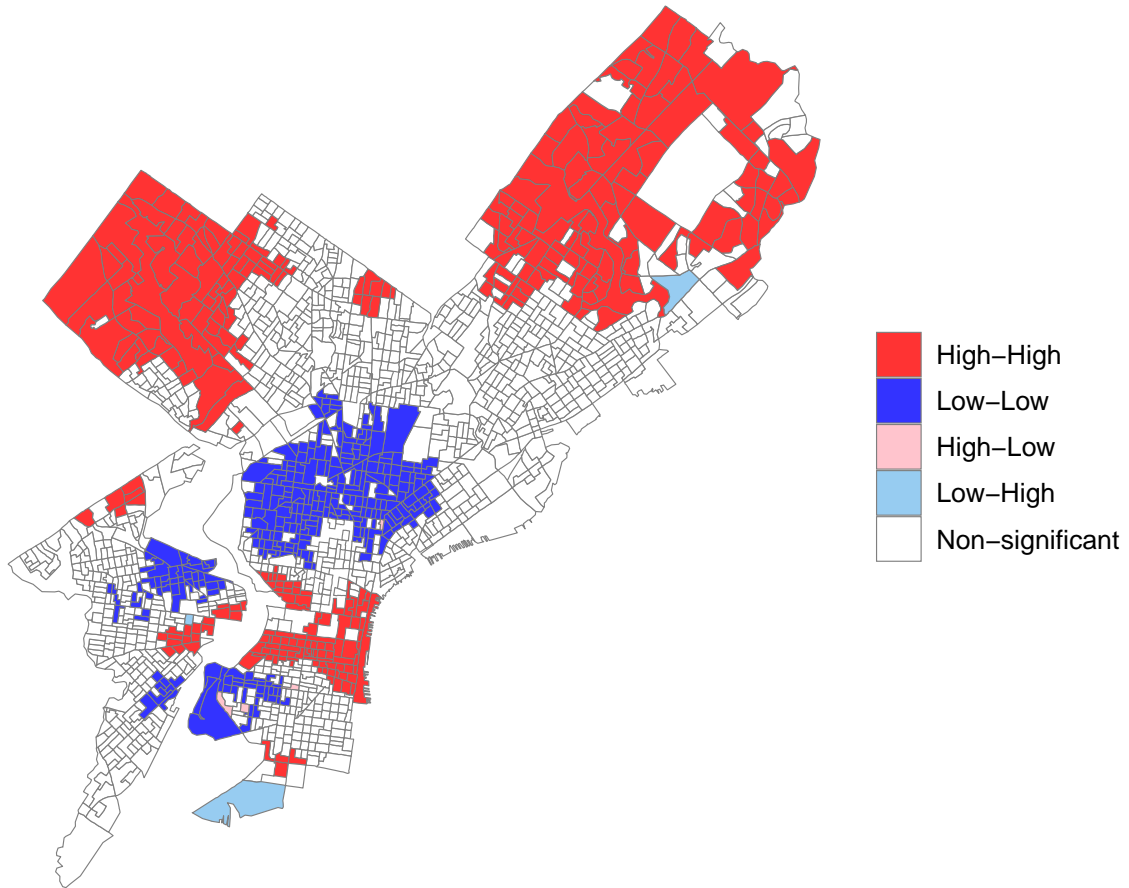


Visually, the histogram of the distribution of global Moran's I values from all the permutation tests shows that our original Moran's I (highlighted by the red vertical line) towers far above the other Moran's I. This histogram further indicates that the possibility of retaining our original Moran's I under true spatial randomness is low.

## LISA P-Value Map



## LISA Cluster Map



The Local Moran's I analysis was conducted to identify spatial clusters and outliers in the distribution of LNMEDHVAL across Philadelphia. Two maps were generated: a significance (p-value) map that highlights areas where spatial autocorrelation is statistically significant and a cluster map that classifies areas into High-High, Low-Low, High-Low, Low-High, and Not Significant clusters based on local spatial relationships. A majority of North East Philadelphia, Upper North Philadelphia and Center City as well as parts of West Philadelphia such as Wynnefield and University City exhibited high values surrounded by other high-value neighbors. A majority of North Philadelphia, Parkside in West Philadelphia, Kingsessing in Southwest Philadelphia, and parts of the South Philadelphia neighborhood exhibited Low-Low relationships or low values surrounded by other low-value neighbors. Only a few areas exhibited Low-High spatial relationships or low values surrounded by other low value neighbors: parts of the South Philadelphia neighborhood and one block group in Torresdale of North East Philadelphia. The Significance Map confirms that most of the identified clusters fall within statistically significant zones ( $p < 0.05$ ) while the rest of Philadelphia was rendered Not Significant.

### 3.2 A Review of OLS Regression and Assumptions: Results (Angel)

#### OLS Regression Summary Table

Call:

```
lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
    LNNBELPOV, data = Regression_shpData)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.25817	-0.20391	0.03822	0.21743	2.24345

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.1137781	0.0465318	238.843	< 2e-16 ***
PCTVACANT	-0.0191563	0.0009779	-19.590	< 2e-16 ***
PCTSINGLES	0.0029770	0.0007032	4.234	2.42e-05 ***
PCTBACHMOR	0.0209095	0.0005432	38.494	< 2e-16 ***
LNNBELPOV	-0.0789035	0.0084567	-9.330	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

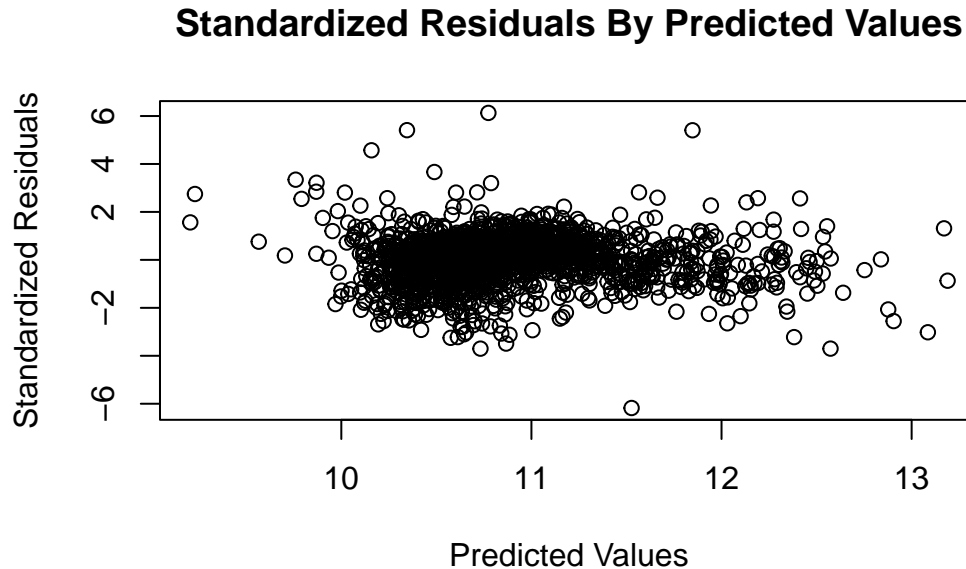
Residual standard error: 0.3665 on 1715 degrees of freedom

Multiple R-squared: 0.6623, Adjusted R-squared: 0.6615

F-statistic: 840.9 on 4 and 1715 DF, p-value: < 2.2e-16



Our OLS results considered all of our model's predictors, PCTBACHMOR, LNNBELOPOV, PCTSINGLES, and PCTVACANT to be significant and returned an  $R^2$  of 0.6623 meaning that approximately 66% of the variance in logged median house values could be explained by the model.



In our initial visual test for heterodasticity in HW 1, we concluded that our scatter plot of our standardized residuals showed general homoscedasticity or consistent variance of residuals. We decided there was general uniformity of the standardized residuals as most were between -2 and positive 2. There were some outliers that extend past -4 and 4 but determined they did not dominate the overall pattern. We also observed no funneling affect or any other pattern of non-constant variance.

### Breusch-Pagan Test Results

Breusch-Pagan test

```
data:  reg1
BP = 113.19, df = 4, p-value < 2.2e-16
```

The p-value from the Breusch-Pagan test suggests, however, that the residuals in the OLS regression model likely exhibit heteroscedasticity. The Breusch-Pagan test evaluates whether the residuals from a regression model exhibit constant variance. Since the resulting p-value falls below the conventional significance threshold of 0.05, we reject the null hypothesis of

homoscedasticity in favor of the alternative hypothesis that the residuals have non-constant variance.

### **Studentized Breusch-Pagan Test Results**

studentized Breusch-Pagan test

```
data: reg1
BP = 42.868, df = 4, p-value = 1.102e-08
```

The p-value from the studentized Breusch-Pagan test, a more robust version of our initial Breusch-Pagan test, also suggests heteroscedasticity since the p-value is less than 0.00000001102, well below the conventional threshold of 0.05, allowing us to reject the null hypothesis of homoscedasticity.

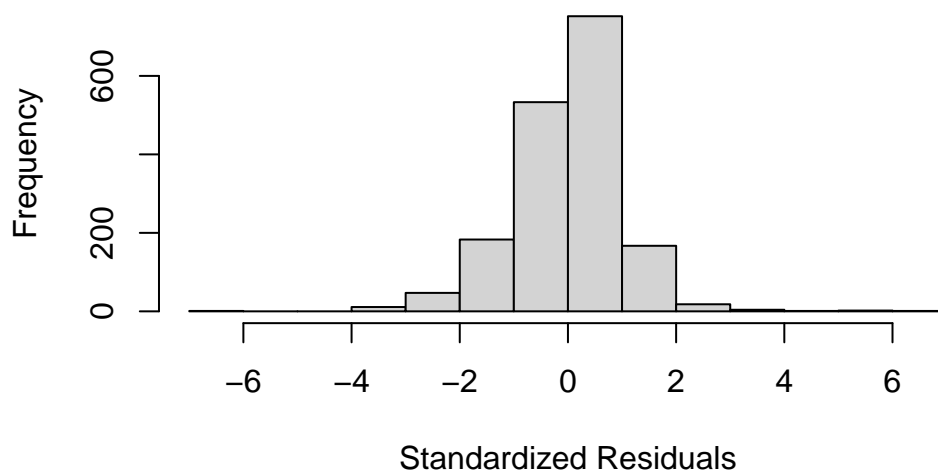
### **White Test Results**

White's test results

```
Null hypothesis: Homoskedasticity of the residuals
Alternative hypothesis: Heteroskedasticity of the residuals
Test Statistic: 43.94
P-value: 0
```

The White's test is another statistical assessment we used to test whether the residuals from an OLS regression model exhibit constant variance. The results show that p-value from this final statistical test is effectively zero, providing strong evidence against the null hypothesis of homoscedasticity. This result reinforces the presence of heteroscedasticity in the model. All three statistical measures heteroscedasticity contradict our initial visual assessment, which did not clearly indicate a violation. The discrepancy highlights that visual diagnostics alone may be insufficient for detecting non-constant variance and emphasizes the importance of formal statistical testing in validating model assumptions.

## Histogram of Standardized Regression Residuals



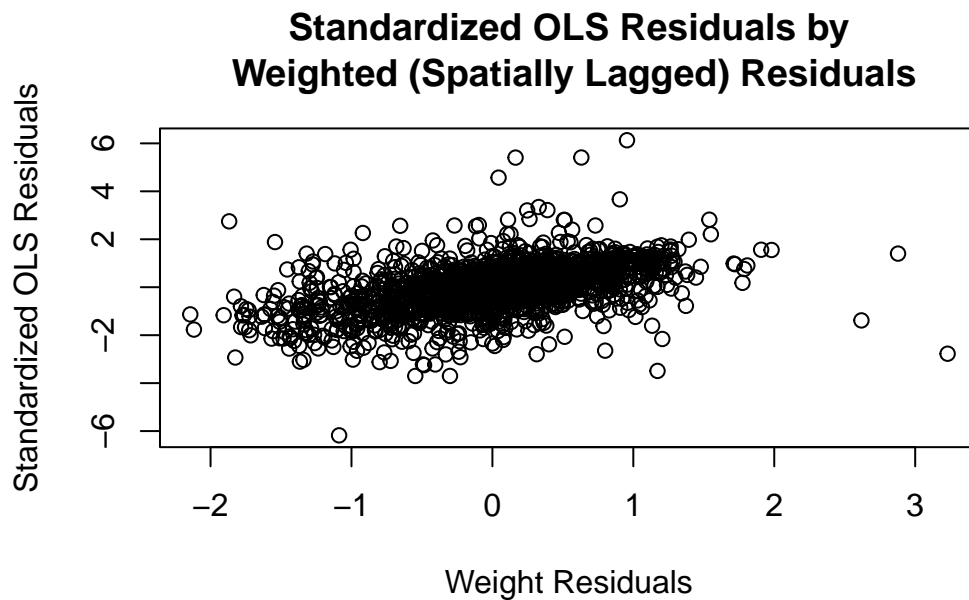
In our initial visual assessment for the assumption of normally distributed residuals, we concluded that the histogram of the standardized residuals showed the normality in residuals needed per our assumption and supported the need for the logarithmic transformations we performed to achieve normality.

### Jarque Bera Test Results

#### Jarque Bera Test

```
data: reg1$residuals
X-squared = 778.96, df = 2, p-value < 2.2e-16
```

The Jarque-Bera test statistically evaluates whether the residuals from a regression model follow a normal distribution. In this case, the p-value is less than 0.00000000000000022, well below the conventional threshold of 0.05, allowing us to reject the null hypothesis that the residuals are normally distributed. This contrast between the visual and statistical diagnostic for normality further reinforces the need for formal assessments of regression assumptions.



The scatter plot between the OLS residuals and their spatially lagged counterparts has a discernible linear trend, suggesting that residuals are not randomly distributed in space and, instead, exhibit spatial dependence.

#### Summary of Weight Residual Regressed on OLS Residuals

Call:

```
lm(formula = stdres ~ wt_residu)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.3685	-0.4450	0.0585	0.4618	5.4435

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.01281	0.02121	-0.604	0.546
wt_residu	0.73235	0.03244	22.576	<2e-16 ***

---

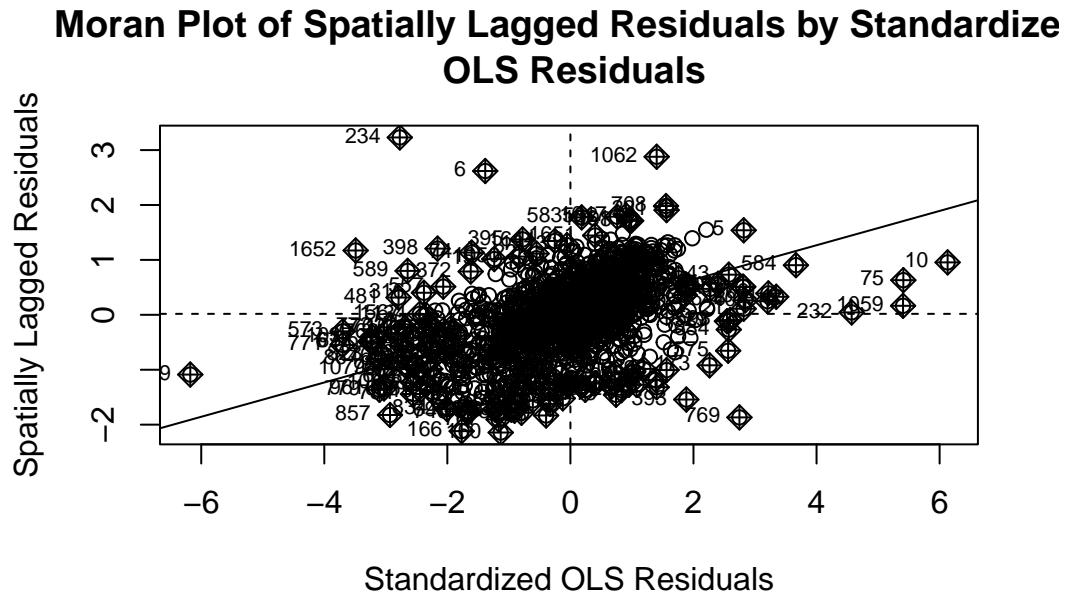
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8793 on 1718 degrees of freedom

Multiple R-squared: 0.2288, Adjusted R-squared: 0.2283

F-statistic: 509.7 on 1 and 1718 DF, p-value: < 2.2e-16

The  $b$  coefficient of 0.73235 from the summary table of the weighted residuals regressed on the OLS residuals quantifies this relationship and indicates a strong positive association between each residual and the average residuals of its spatial neighbors.



In the Moran scatter plot of the spatially lagged residuals by the OLS residual, the clustering of values along the diagonal line indicates a clear pattern of spatial dependence, suggesting that residuals at one location tend to resemble those of neighboring locations.

#### Random Permutation Test/ Monte-Carlo Simulation Table for OLS Regression

Monte-Carlo simulation of Moran I

```
data: stdres
weights: queenlist
number of simulations + 1: 1000

statistic = 0.3124, observed rank = 1000, p-value < 2.2e-16
alternative hypothesis: two.sided
```

The observed Moran's I value, 0.3124, indicates moderate positive spatial autocorrelation. Out of 1000 permutations, our observed statistic was the most extreme, meaning none of the randomized simulations produced a Moran's I as large. Our p-value of less than

0.000000000000000022 was much less than the statistical threshold, indicating our observed Moran's I was statistically significant.

Thus, the Moran's I and the  $b$  coefficient from the regression of spatially lagged residuals on OLS residuals both reinforce the presence of spatial autocorrelation. Together, these diagnostics consistently point to the violation of the OLS assumptions of independent errors the need for spatial modeling. Specifically, these findings motivate us to use the spatial error and spatial lag regression models, which explicitly account for spatial dependence.

### **3.3 Spatial Lag and Spatial Error Regression Results (Sujan)**

### **3.4 Geographically Weighted Regression Results (Ming)**

## **4 Discussion (Google Doc at end)**