# Hedonic Price Model to Explain Rent Prices in New York City

**Submitted By: Gaurav Bhardwaj**
**Center for Urban Science & Progress**
**Advanced Spatial Analytics**
**Prof. Paul M. Torrens**

## Introduction

When we move to a new city, we often look for places that are close to city centers, public transit, schools and shopping complexes. And as one would assume, the price of the apartment would vary based on the facilities one gets living in that building or the neighborhood. One of the methods that's used in spatial econometrics is Hedonic Price Model, a model that uses internal as well as external factors that may affect the rent of the property. This model has been used for quite some time now to predict the housing prices and has been an important tool for econometric studies of housing markets. The model assumes that people value the characteristics of a good, or the services it provides, rather than the good itself (Ecosystem Valuation, 2014). In the recent years, the real estate prices have spiked even though the underlying characteristics of the properties remain same. The hedonic price model explains these variations well as the model takes into consideration the effect of a variable along with the weight of the variable, i.e. if any given variable contributes more towards the price of the property, it would be given higher weight in order to get proportionate price (Monson, 2009). For this report, I chose 5 variables that I believe affects the rent price of a property, and studied their relationship with the rent price for the property.

## Method & Analysis

Prior to formulating the hedonic price model, it's necessary to study how the variables interact with the price of the property. So, to do that I performed linear regressions and analyzed their adjusted R-square values and the estimated errors. I believe that certain characteristics change considerably for a larger area under consideration, so I limit my analysis to census block group. The variables I considered for my model are (American Community Survey, 2014):

- *Travel Time to Work*: Travel time to work refers to the total number of minutes that it usually took the worker to get from home to work during the reference week. I set a limit of travel time less than 30 minutes for my model.
- *Ratio of Income to Poverty Level in the Past 12 Months:* The Census Bureau uses a set of dollar value thresholds that vary by family size and composition to determine who is in poverty. This variable gives out good information about the overall economic stability of the block group. I wanted to analyze how low ratio of income to poverty would affect the rent price, so I considered the ratio < 1
- *Rooms:* Rooms provide the basis for estimating the amount of living and sleeping spaces within a housing unit. For each unit, rooms include living rooms, dining rooms, kitchens, bedrooms, finished recreation rooms excluding kitchen, bathrooms and balconies.
- *Year Structure Built:* The year the structure was built provides information on the age of housing units. These data help identify new housing construction and measures the disappearance of old housing from the inventory. Specifically, for this reason, I considered the housing units built before the year 2000
- *Gross Rent (Dollars):* Gross rent provides information on the monthly housing cost expenses for renters. This will be dependent variable and the predicted value from the regression model will be predicted gross rent for the block group. The data has been normalized at block group level.

To start with linear regression, I used ordinary least square regression for the above-mentioned variables and got an adjusted R-square values of 0.74 (Appendix Fig.1). This score represents how well the model can fit the variance generated by the explanatory variables. The multicollinearity condition number for the model is 17.95 which means the variables I considered are not correlated with each other (KUCC625 & Jeeshim, 2003).

Although, the regression model hives us good prediction of our dependent variable, there is still one component OLS doesn't take into account. The regression model in no way consider the influence of geography on the variables we define. For the same reason, there have been many hedonic price models that have not been able to predict accurate number for the variables. On similar notes Ayse Can asserts, "The Hedonic Price Model's theoretical foundation is well developed within the traditional urban economics framework, but the empirical work has ignored methodological issues that might result from the spatial nature of data sets" (Can, 1992). Now that we have our regression data, we should check our variables for spatial autocorrelation, as we may get an understanding if some of the variables are essentially spatially clustered in some areas. I ran Global Moran's I for all 5 variables and the scores are attached in Appendix Table 2. The Global Moran's I score for my chosen variables range from 0.25-0.41, so we can infer from these scores that the variables are loosely correlated but have positive correlation. In other words, the variables are not spatially clustered from global perspective of analysis. Once we know we have autocorrelation, we can't rely on traditional regression models. Instead, we use Geographically Weighted Regression model, that can explain the spatially varying relationships of the variables (Lu, Charlton, & Fotheringham, 2011). They also mention, "GWR makes a point-wise calibration around each regression point using a 'bump of influence': around each regression point nearer observations have more influence in estimating the local set of parameters than observations farther away". As my next step, I executed the GWR for my variables and got an adjusted R-square of 0.91. This high value of the model represents that the model has done a good job at spatially adjusting the variance in the model. We still have some improvements to make with our model since each variable may not have similar effect on rent price. We consider this by adjusting our model with spatial weights of the variables that we considered to predict the rent. Spatially-adjusted regression is our way to make better predictions by minimizing the error that could have generated because of unweighted variables. After executing the regression, I got an adjusted R-square value of 0.80 which is again a high value considering the number of variables and spatial autocorrelation (Appendix Fig 2).

## Conclusion

In the analysis section above, we discussed about the R-square values I received for three regression models I ran. The summary is tabulated in Table 1.

Therefore, the Hedonic Price Model equation according to my analysis would be

$$y = 0.25a - 0.05b + \ 0.61c - 0.06d - 136.77$$

where,
a - *Ratio of Income to Poverty Level in the Past 12 Months*
*b - Year Structure Built*
*c – Rooms*
*d - Travel Time to Work*

From my results, the model above can predict the gross rent with high accuracy and adjusting for all the errors and spatial lags. The model has been tested for autocorrelation and multicollinearity and the results have been tabulated in the Appendix. Through this project, I got a good sense how urban data cannot be predicted using only the statistical models. These models fail to account for other factors, like in this project was geographical influence. The variables have to be adjusted because different variables affect the geography in a different way and therefore giving same weight to everyone would downgrade the predicted values. Although these models are good, we should be at the lookout for errors such as heteroscedasticity, normal distribution of errors about regression line and residual independence. After considering all possible options, the predicted model is the best fit for the variables considered.

## Bibliography

American Community Survey. (2014, 07 21). *American Community Survey and Puerto Rico Community Survey 2014.* Retrieved from Census.gov: https://www2.census.gov/programs-surveys/acs/tech_docs/subject_definitions/2014_ACSSubjectDefinitions.pdf

Can, A. (1992, 02 01). *Specification and estimation of hedonic housing price models.* Retrieved from ScienceDirect: https://ac.els-cdn.com/0166046292900394/1-s2.0-0166046292900394-main.pdf?_tid=spdf-88cfac4e-58f0-40fa-8ae5-eba36e23a212&acdnat=1519577187_b25f006d07d529aa6d792ffbb30f6f50

Ecosystem Valuation. (2014, 05 01). *Hedonic Pricing Method*. Retrieved 02 22, 2018, from ecosystemvaluation.org: http://www.ecosystemvaluation.org/hedonic_pricing.htm

KUCC625, & Jeeshim. (2003, 05 09). *Multicollinearity in Regression Models.* Retrieved from http://sites.stat.psu.edu/: http://sites.stat.psu.edu/~ajw13/SpecialTopics/multicollinearity.pdf

Lu, B., Charlton, M., & Fotheringham, A. S. (2011, 10 01). *Geographically Weighted Regression Using a Non-Euclidean Distance Metric with a Study on London House Price Data.* Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S1878029611001447

Monson, M. (2009, 07 01). *Valuation Using Hedonic Pricing Models.* Retrieved from scholarship.sha.cornell.edu: https://scholarship.sha.cornell.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1058&context=crer

# Appendix

| | OLS | GWR | Spatial Lag |
|---|---|---|---|
| **R-Square** | **0.745382** | **0.9517** | **0.8002** |
| **Adjusted R-Square** | **0.745225** | **0.9083** | **-** |
| **Intercept** | **-63.3158** | **-** | **-136.77** |

**Table 1: Regression Results for Models**

| | Global Moran's I Score |
|---|---|
| **Gross Rent** | *0.32* |
| **Travel Time** | *0.42* |
| **Ratio of Income to Poverty** | *0.37* |
| **Rooms** | *0.29* |
| **Year Built** | *0.28* |

**Table 2: Global Moran's I Scores for Variables**

```
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  NYC_Data
Dependent Variable  :     B25063e1  Number of Observations: 6492
Mean dependent var  :      324.573  Number of Variables   :    5
S.D. dependent var  :      253.825  Degrees of Freedom    : 6487

R-squared           :     0.745382  F-statistic           :     4747.59
Adjusted R-squared  :     0.745225  Prob(F-statistic)     :           0
Sum squared residual:1.06497e+008  Log likelihood        :    -40715.2
Sigma-square        :        16417  Akaike info criterion :     81440.3
S.E. of regression  :      128.129  Schwarz criterion     :     81474.2
Sigma-square ML     :      16404.3
S.E of regression ML:      128.079

-----------------------------------------------------------------------------
       Variable      Coefficient       Std.Error     t-Statistic    Probability
-----------------------------------------------------------------------------
          CONSTANT      -63.3158         3.33119       -19.0069      0.00000
        Travel_Tim    -0.0368909       0.0148525       -2.48382      0.01302
          B25017e1       0.65644       0.0174504        37.6174      0.00000
        Year_Built    -0.0506065       0.0171497       -2.95086      0.00318
        Inc_Pov_Ra      0.285838      0.00641696        44.5442      0.00000
-----------------------------------------------------------------------------
```

**Fig 1: Results for OLS Model**

```
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : NYC_Data
Spatial Weight      : GeoDa_SW_GRent
Dependent Variable  :     B25063e1  Number of Observations: 6492
Mean dependent var  :      324.573  Number of Variables   :    6
S.D. dependent var  :      253.825  Degrees of Freedom    : 6486
Lag coeff.   (Rho)  :     0.346336

R-squared           :     0.800265  Log likelihood        :    -39991.2
Sq. Correlation     : -             Akaike info criterion :     79994.5
Sigma-square        :      12868.4  Schwarz criterion     :     80035.2
S.E of regression   :      113.439

-----------------------------------------------------------------------------
       Variable      Coefficient       Std.Error        z-value     Probability
-----------------------------------------------------------------------------
        W_B25063e1      0.346336      0.00877811        39.4545      0.00000
          CONSTANT       -136.77          3.4007       -40.2184      0.00000
        Travel_Tim    -0.0630397       0.0132029       -4.77469      0.00000
          B25017e1      0.613986       0.0154989        39.6149      0.00000
        Year_Built    -0.0553727       0.0151847       -3.64662      0.00027
        Inc_Pov_Ra      0.247514      0.00574441        43.0878      0.00000
-----------------------------------------------------------------------------
```

**Fig 2. Regression Results for Spatial Lag Model**