Does distance to Subway affect the price of residential real estate?

Robert V Checco

1. Introduction

The urban real estate market is important for both investors and the general public. Local residents are sensitive to changes in the cost of housing. With an increasing number of opportunities available in cities, rising rents have pushed residents out of their own neighborhoods. As rents rise close to the central business district (CBD), living in peripheral neighborhoods has become more attractive. In New York, Western Brooklyn has grown in popularity, due to the short commute into Manhattan. This paper addresses the following question, is there a relationship between the distance to a subway station and the price of residential real estate sales. It may be that the closer one lives to a station the shorter the commute into Manhattan. Alternatively the presence of public transportation could encourage crime and decrease the property values in the area.

At least three groups are interested in better understanding the relationship between distance to public transit and residential real estate. Trying to understand the way prices of residential properties change relative to distance to the nearest subway is important to three different groups of people. First, this is helpful for urban planners to better understand consumer preferences to living close to public transit. Urban planners could use this information to better understand how subway stations are valued by its residents. For example, understanding the potential price changes of buildings will allow them to create cities that are more cost effective and efficient. One way this could be used would be to increase the supply of housing to bring down the cost of rents in the area.

Second, real estate investors could find it useful to know how distance to subway can be used when placing bids on properties in the area. Having this increased information allows them to make better buying and selling decisions. Third, while individual renters do not purchase residential real estate, the rent that they pay closely maps what the buildings are bought and sold for. This is useful for renters to make decisions whether they will be paying more or less as they live closer to a subway station.

There are studies that create models to better estimate real estate sales prices. The models that are most often used are Hedonic Pricing models. According to Conway and Johnson (1994) it is fair to make an assumption that homes in the same neighborhood are priced in a similar way. Building on this concept is the idea of using descriptive features to approximate the price of the building as a whole.

This papers breaks away from others that try to estimate the effect of proximity to public rail to housing prices in that it is located in New York City where other papers explored countries such as China. This is seen in the research conducted by Yin and Yang (2012) where they explore how subway stations impact land values in Beijing. While authors in the past have tried to quantified this earlier in the 21st century, this paper explores these topics with most recent data.

2. Data

The Data on the sales of residential real estate was collected in the following manner. First, I located a Brooklyn Real Estate data collection on Kaggle that was a subset of the New York City (NYC) Department of Finance ledger of real estate transactions in NYC. From the entire Brooklyn subset of the dataset, I eliminated all transactions that were not for residential rentals. Then the geographic area was narrowed down to all sales conducted in Williamsburg neighborhood of Brooklyn. The data includes areas of Williamsburg to include North, South and East Williamsburg.

I omitted transactions that were for extremely low prices or didn’t have a sale price on file. It could be inferred that these were family to family transactions. These figures would create a strong disturbance in the data and does not provide value to the question that we are attempting to address in this paper. I restricted the years of the data set from 2014-2017. This time period allows us to know a large amount of sales that can be seen to be in a consistent.

To measure the market value of the residential real estate I used the price in USD to measure what the market would value a property at. It is tough to say what a building could be worth without having a record of an actual transaction. For example, a building could be said to be worth one million dollars, but until the building is sold for one million dollars the price of the building is uncertain. Instead of attempting to approximate the value of the buildings in our data set we are using the transactional data from the NYC Dept of Finance. This price cleared the market and can be seen as a clear signal as to how investors value residential real estate.

In order to control for buildings that have much more rentable units, I created a variable by combining two of the variables in the dataset. The new variable that was generated is called price/sqft. This is as it seems, it is the price of the building divided by the square footage of the building. We assume that larger buildings would command more value and by normalizing it with a area parameter we can better compare buildings to other buildings.

To measure the size of the building we used two different variables. The first is square footage of the building. In this paper square footage refers to the building square footage and not the land square footage. This is because many building in the Williamsburg neighborhood have several floors and are built in the vertical direction. In order to compare buildings from one to another it is important to know how big they are.

Second, we use a rentable units variable to describe the rent potential of the building. While a building can be very large, it is important to consider how many rentable units the building has. In order to differentiate between buildings of similar square footage, the rentable units variable allows us to best understand if more units of smaller size affects the price in a meaningful way. This is useful for measuring the value of residential real estate because when investors purchase a building with the intention of renting it out they will try and see how much they can rent out each unit for. Once they approximate this number they then multiply that monthly rent price to understand the monthly rental income for the entire building. Understanding this cash flow allows them to understand at what price the real estate is a good investment.

In order to distinguish between building of similar size and rental capacity, I incorporated a categorical variable that accounts for consumer tastes. I call this variable Postwar. This refers to all of the construction of building that occurred after 1945. Buildings that were built after this time were stylistically different than those that were constructed prewar. Building that were built more recently are more likely to have additional amenities such as fitness centers, doorpeople, and lounge areas. In order to incorporate this into the model we created a dummy variable that turns either on or off depending when the building was built. The default is a value of 1 that applies to all construction that was built after 1945, and the 0 would represent construction that was built before 1945. We used the year 1945 as the cutoff because there is a known architectural style that is evident in building built before and after this time period. Also, after the war, customer preferences changed and the way apartment building were constructed in Brooklyn changed.

The last variable was the distance to subway. This was measured by the Google Maps walking feature and it is expressed in meters. It takes the most direct route to the closest Manhattan bound subway station. It is important to note that this is not to the closest subway station but the closest station that goes directly into Manhattan. If there were two stations that were equidistant I chose the station that was closer to Manhattan. For example, If stop ‘a’ was in Manhattan, and stops ‘b’ and ‘c’ were located in Brooklyn, and they were equidistant, I selected stop ‘b’.

sale_price		price sqft		gross_sqft		residential_units		Post War		distance to subway (meters)		year_built	
Mean	6265685.859	Mean	570.6462818	Mean	9988.158824	Mean	10.55882353	Mean	0.141176471	Mean	381.6117647	Mean	1930.441176
Standard Error	1086281.045	Standard Error	28.87347734	Standard Error	1268.22273	Standard Error	1.12045053	Standard Error	0.026784884	Standard Error	16.29443495	Standard Error	2.597234813
Median	2262500	Median	506.8579438	Median	4875	Median	6	Median	0	Median	350	Median	1920
Mode	1900000	Mode	425	Mode	4125	Mode	6	Mode	0	Mode	350	Mode	1910
Standard Deviation	14163372	Standard Deviation	376.4640858	Standard Deviation	16535.60134	Standard Deviation	14.60888758	Standard Deviation	0.349232165	Standard Deviation	212.4534391	Standard Deviation	33.86379888
Sample Variance	2.00601E+14	Sample Variance	141725.2079	Sample Variance	273426111.6	Sample Variance	213.4195962	Sample Variance	0.121963105	Sample Variance	45136.46377	Sample Variance	1146.756874
Kurtosis	35.12561704	Kurtosis	3.626362196	Kurtosis	18.77907661	Kurtosis	17.79005848	Kurtosis	2.351396156	Kurtosis	1.913374333	Kurtosis	2.467589751
Skewness	5.341227372	Skewness	1.524148601	Skewness	3.990830757	Skewness	3.938388586	Skewness	2.079391734	Skewness	1.052148203	Skewness	1.283937717
Range	1248	Range	217	Range	123	Range	100	Range	1	Range	127	Range	215

	00000		3.832626		300						3		
Minimum	200000	Minimum	26.04166667	Minimum	2100	Minimum	4	Minimum	0	Minimum	27	Minimum	1800
Maximum	125000000	Maximum	2199.874293	Maximum	125400	Maximum	104	Maximum	1	Maximum	1300	Maximum	2015
Sum	1065166596	Sum	97009.8679	Sum	1697987	Sum	1795	Sum	24	Sum	64874	Sum	328175
Count	170	Count	170	Count	170	Count	170	Count	170	Count	170	Count	170

3. Empirical Model

The econometric specification used in this paper is a hedonic pricing model. The sale price of residential real estate can be viewed as a linear function of its characteristics. Specifically for this paper it was most appropriate to start out with the size of the apartment. The renting capacity for the building could serve as the strongest indicator on the price. I would make sense that a building with 100 units would be worth more than a building with 10 units, ceteris paribus.

I ran two different linear regression specifications using OLS. They provide a different snapshot of the real estate market in Williamsburg, Brooklyn.

Model 1: $Price = B_0 + B_1 sqft + B_2 rentalunits + B_3 postwar + B_4\ distsubway + u$

The first model has a dependent variable of sale price, and independent variables of square footage, residential units, post-war and distance to subway. This model attempted to create a model that would approximate the sale price of a building.

Model 2: Price/Sqft = $B_0 + B_1$rentalunits + B_2postwar + B_3 distsubway + u

The second model differs from the first in that the dependent variable is a fraction. It is the price per square foot. This is a variable that I computed by dividing the listed sale price by the listed square feet of the building. This helps control for the new construction that has been being built in Williamsburg. We would expect the price of new construction in the area to follow a similar price per square foot as other sales in the area. Using this control technique to better explain these high prices, will allow us to make better approximation for the Beta term for the distance to subway variable.

The second one has a dependent variable of sale price divided by square footage. It contains independent variables of residential units, post-war and distance to subway. I removed the square footage from the right side of the equation by dividing it out and dividing both side of the equation. It would not make sense mathematically to have it remain on both sides of the equation.

4. Empirical Results

After running the first regression specification the following regression output was generated.

H_0: $Beta_{\text{distance to subway}} = 0$

H_1: $Beta_{\text{distance to subway}} \neq 0$

Regression Statistics	
Multiple R	0.904875161
R Square	0.818799057
Adjusted R Square	0.814406307
Standard Error	6101663.719
Observations	170

ANOVA

	df	SS	MS	F	Significance F
Regression	4	2.77586E+16	6.93965E+15	186.3978221	4.30775E-60
Residual	165	6.143E+15	3.72303E+13		
Total	169	3.39016E+16			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-3124380.599	1049342.897	-2.977463903	0.003344477	-5196251.097	-1052510.1	-5196251.097	-1052510.1
gross_sqft	-243.0256524	77.71457642	-3.127156624	0.002086688	-396.4688582	-89.58244659	-396.4688582	-89.58244659
residential_units	1060396.067	88504.16284	11.98131289	3.37762E-24	885649.4092	1235142.724	885649.4092	1235142.724

Post War	5580152.859	1498914.047	3.722797095	0.000269998	2620628.569	8539677.148	2620628.569	8539677.148
distance to subway	-437.2924771	2238.686099	-0.195334432	0.845371475	-4857.456462	3982.871508	-4857.456462	3982.871508

The first thought of this output was complete confusion because a few of the terms did not make economic sense. First, the negative intercept seemed to be completely incorrect. This is because if we assume there is a plot of land that is vacant and has no building we would assume that there still exists a value for the plot of land. Even if the intercept was a smaller term is would make sense, but for it to be negative makes no economic sense. If the land has negative value then how is it less than zero? What is it about the land that can possibly give it a value less than zero?

My guess is that in this regression specification there were problems estimating the price with a linear OLS model. The reason I believe this to be true is because in recent years there has been increase real estate development in the geographic area this paper addresses. The buildings that are being built are significantly bigger than the ones that previously exists in the past. This very new construction seems to be captured in the PostWar variable and any new construction seems to have an extremely high premium in the area.

The area studied in this paper has seen an uptick in gentrification as evident in the increasing number of new construction in the area. One way to control for this would be to increase a dummy variable for new construction. This may be a wise approach in a future version of this paper. The sale price of these units may be tied to investor confidence in the area which may value the buildings much higher than in the past. This could make these recent sales to be outliers to the dataset and skew the regression towards the extreme values. In the future I could limit the data points to cap the total rentable units as to not receive this error.

The coefficient of the square footage does not make sense in economic terms either. Being negative means that as the square footage goes up the price should go down. This clearly makes no logical sense and I again think this has something to do with the regression model I used and the data points that were chosen for this project.

The coefficient on the distance to subway was negative, meaning that for every meter increase in distance from the subway, the price of the residential real estate fell roughly 437 dollars. As mentioned in the beginning of the paper, I was not sure whether this beta term would be positive or negative. Being negative means that there is a premium in being located nearby a Manhattan bound subway station.

In order to check if the coefficient is meaningful, I conducted a two-tailed test and received an output of -0.2. This is below the critical value of significance of 1.96 and therefore I fail to reject the null hypothesis at the 0.05 level of significance. This regression was inconclusive and we are unable to say whether there is a relationship between distance to subway station, and the price of residential real estate.

The R-Squared of this regression was convincing in that it was 0.81. This means that the model does a fine job of describing the price of residential real estate in Williamsburg, Brooklyn. Although the intercept and coefficients frighten me, I included this in my findings because of the strong R-Square. It seems to somehow do a good job of fitting the data. There is not as much left in the unobserved term as the specification seen below.

The second regression specification yielded more logical results.

H_0: $\text{Beta}_{\text{distance to subway}} = 0$

H_1: $\text{Beta}_{\text{distance to subway}} \neq 0$

Regression Statistics	
Multiple R	0.210227743
R Square	0.044195704
Adjusted R Square	0.026922132
Standard Error	371.3619031
Observations	170

ANOVA

	df	SS	MS	F	Significance F
Regression	3	1058556.06	352852.02	2.558573577	0.056893957
Residual	166	22893004.07	137909.6631		
Total	169	23951560.13			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	526.2005349	62.36093048	8.437984021	1.51787E-14	403.0777477	649.3233221	403.0777477	649.3233221
residential_units	1.95783864	2.178001425	0.898915224	0.369999873	-2.342315384	6.257992663	-2.342315384	6.257992663
Post War	178.3059453	91.11262616	1.956983931	0.052026699	-1.582975506	358.1948662	-1.582975506	358.1948662
distance to subway	-0.003666894	0.134479672	-0.027267273	0.978279312	-0.269177878	0.26184409	-0.269177878	0.26184409

Unlike the previous regression specification this regression made much more economic sense. It is important to first note that the reason that the coefficients appear to be at different magnitudes is because in this regression, as mentioned above, uses price/square foot as the dependent variable.

The intercept term means that the price per square foot in the Williamsburg neighborhood, excluding all other variables begins at 526 dollars a square foot. This is much more meaningful than the previous regression specification because even without rentable units there is still a price for the land.

The coefficient of the distance to subway was negative in this regression as well. In this specific calculation the price of residential real estate per square foot decreases 1 cent every 3 meters further from the subway. This means that people are interested in living near a public transit line and that developers and renters in the area place a premium on public transit. In order to see if this coefficient has statistical significance I conducted a two tailed test. The p-value in this regression was -0.03 well below the critical value of 1.96. As a result I fail to reject the null hypothesis at the 0.05 level of significance and it is found to be inconclusive whether distance to subway significantly affects the price of residential real estate in terms of price per square foot.

It is important to note that in this regression specification the model does not seem to accurately describe the price per square foot as the r-squared is 0.04. This means that the model does a poor job in describing the price of apartments in terms of square footage. There must be a lot of factors outside of my independent variables that are contributing to this variation. Future research will be conducted in order to sort out this confusion.

5. Conclusion

With the available data and econometric specification used, there is not sufficient evidence to say there is a positive or negative relationship between the distance to a subway and the price of residential real estate. While this paper was hoping to discover a meaningful output, It is a start in trying to answer this question. Going forward, by incorporating different regression techniques, as well as increasing the size of the dataset could lead to a more conclusive finding. I am unable to reject the null hypothesis that there is no relationship between distance to subway and the price of residential real estate in Williamsburg, Brooklyn.

In future research it could be useful to incorporate other subway lines that would allow riders to connect to Manhattan bound subway lines. In this research I mainly focused on subway stops that allowed direct access into Manhattan without having to transfer in Brooklyn. Maybe instead of measuring the distance in meters, I could create an index that would explain how quick on average it would take for them to access Manhattan including walking, riding, and transferring subway lines. A future paper on this subject could expand the years of sales data which would increase the number observations present in the dataset.

Conway, Delores A., and David Dale-Johnson. "Multivariate Analysis of Real Estate Prices." *Lecture Notes-Monograph Series*24 (1994): 439-44. http://www.jstor.org/stable/4355822.

Yan, Bin, and Jiawan Yang. "Land Values Impacts of Subway Stations A Case Study of Beijing City." https://smartech.gatech.edu/bitstream/handle/1853/43479/BinYan_Land Values Impacts of SubwayStations.pdf;jsessionid=E89433B5B72FD4A4C5C4428CB0CD3C2B.smartech?sequence=1.