# Analyzing the effects of Subway proximity on real estate prices in London [*]

Aman Rana
1007034692
aman.rana@mail.utoronto.ca

Pierre Sarrailh
1006770111
pierre.sarrailh@mail.utoronto.ca

ECO375: Applied Econometrics
University of Toronto
Department of Economics
5th December, 2024

## Abstract

Following Zhou et al. (2019) we saw that the completion of subway lines increased the prices of real estate in their proximity. Our contribution will be to look at these effects in London. We found precise null effects by a multivariate linear regression of distance from a station on the log of house prices, with controls for time and district. In a regression of log price on the distance from the nearest station with time and district controls, the slope coefficient is -0.0000268 with a standard error of 0.000000687, a statistically significant relationship of such small effect it led us to conclude precise null effects. The effect is economically insignificant, even a 10km distance from a station (which is above the 95th percentile of our distance data) would lead to only a 0.2% decrease in house price. On more relative terms, a standard deviation difference in distance is only associated with a 14% of a standard deviation change in log house price. A possible further investigation would be to look at the treatment effect of having a nearer subway station using a difference-in-difference (DiD) framework. Using the DiD framework we would be aiming to investigate the effect of specifically new subway stations on house prices while accounting standardizing over previously built ones.

[*] Code and data supporting this analysis is available at: `https://github.com/Aman-Rana-02/`

# 1 Introduction

The relationship between transport infrastructure and real estate prices has been a topic of interest in econometrics for many years now. Accessibility to public transit has been long established as a key factor influencing the price of residential properties, particularly in dense urban environments where vehicle mobility is greatly constrained. In the context of the Greater London Area subway stations are the most efficient method of transportation for most commuters and thereby play a critical role in shaping the greater housing market. To this end in this paper we investigate how the euclidean distance from a residence to its nearest subway station affects market price

By focusing on the Greater London Area our study contributes to already established relationships between subway proximity and house prices. Early studies such as those done by Muth (1969) are cornerstones in establishing a relationship between densification and land use in cities. More recently a greater number of studies have found a significant positive relationship between transit proximity and real estate prices in cities such as those in Shanghai (Zhou et al., 2019) where subway stations proximity was specifically correlated to increased house prices at statistically significant results. However London presents unique opportunities for research. As the historical pioneer in subway technology the city has had the opportunity to develop around an already established subway system and culture. This study extends the findings of existing literature by using highly specific spatial data and leveraging advanced analysis techniques to create clear and insightful findings on the pricing dynamics of the London underground.

The rest of the paper is structured as follows: Section 2 describes the data used in this analysis. Section 3 outlines the methodology and model specifications used to measure the subway effect. Section 4 presents the results of our analysis. Section 5 discusses the limitations of our methodology and concludes. Appendix A holds figures and tables.

# 2 Data

We began our analysis by sourcing data from ONS, the national statistics arm of the UK government, which provided panel data containing each lease transfer (home sale) that has occurred in the U.K. from 1995 to today. We pruned this dataset to only contain observations within the Greater London Area. This dataset amounts to approximately 350 thousand observations. Each unit of observation represents a single occurrence of a lease transfer in London containing the date, coordinates, neighborhood, and price of each observation's lease transfer. The second source of our data was a table containing the coordinates of each subway station in the London underground and the date at which it was opened. To begin our analysis we combined these two datasets by using the date and coordinates of our houses to search for the closest subway station to the current house in the second dataset. This allowed us to get the minimum distance to a subway station at the time of its sale which we added to our ONS dataset as the variable calculating the euclidean distance to the closest station and a variable for the coordinate of the closest subway. For all houses which were further than 10 km away from a subway station we set their distance to be equal to 10 km to account for outliers in our data. Delving into the summary statistics of our data, which can be found in tables 1 to 4, we can look at how the data is spread across our different variables. For our main variables of interest, the distance to subway stations, we see that we have an average distance of 2821 meters with a variance of 2915 meters. This means that on average houses are located quite far from a subway station but also contain vast amounts of variation. From our earlier work on the matter the reason for this high variation and mean is due to the fact that a lot of houses are located very close to subway stations but there is a very long tail in our data with a significant number of houses located over 5 kilometers away. On the categorical side of things we see that our houses are very evenly spread between our neighborhoods with the largest neighborhood only being 2 points off of the mean. Similarly, our lease transfers were also very evenly spread between years with each year accounting for approximately 2% of our dataset. This is good for the robustness of our future models since the even spread of data between our neighbourhoods would provide us enough data to account for outliers within neighbourhoods. On the other hand the spread between newly built and old homes is quite dismal with 97% of lease transfers being accounted for by the sale of pre-owned homes and as a result it will be more sensitive to outliers.

## 3 Regression Analysis

For our regression analysis we created three different formulas of increasing specificity to best capture the strength of our variable of interest as a predictor for price change. We began by looking at the basic relationship between the distance to a subway station and the log of price. We chose to use a log linear formula as our findings look at the effect of distance on price, or in other words how much change from our average home price does 1 meter of distance have. After looking at Figure 1 we realized that the effects of distance may be closely related to price only for houses very near to a subway station. Therefore, we created a dummy variable which calculated the effect on prices for houses within X meters of a subway station and outside that range. For our second formula we found that our errors were dependent on our variable of interest, meaning we had omitted variable bias so for our third formula we accounted for omitted variables. Finally for our fourth formula we also looked at non-linear relationships between our variable of interest and our other explainer variables.

For our analysis, for all our formulas we assumed that our data was i.i.d since logically the sale of one house does not have an effect on others selling their own homes. This assumption may be questioned but it is outside the scope of this paper and discussed further in our limitations section. In our data cleaning section we also clipped all our large outliers to be equal to 10 km therefore we can assume our data contains no large outliers. For all our calculations we used robust standard error and therefore aren't concerned about the homoscedasticity of errors. We discuss omitted variable bias further in each respective subsection

3.1 Simple Linear Regression $Log(Price) = 12.52 - .0000441 * Min_dist$ Using this formula we were able to get statistically significant null effects for distance on price. From our table 2, see that the coefficient of our variable of interest is -.0000431 at above the 5We calculate the omitted variable bias for our variable of interest on our formula. To account for this we then calculate our next formula using a distance dummy This simple regression model is misspecified. Intuitively the variation of house prices can be explained by more factors. We expect that this leads to a coefficient with a downward bias, the subway distance explaining variation caused by other factors: size, year of observation and age of house. 3.2 Simple Linear Regression With a Dummy Variable $Log(Price) = B_0 + B_1 * dist_dummy$ The mean of our errors over each year was not equal to 0, as seen in table X. To account for this 3.3 Multiple Linear Regression With a Dummy Variable $Log(Price) = B_0 + B_1 * dist_dummy + B_n * Ommited_n$ 3.4 Multiple Regression With Linear and Non-Linear effects $Log(Price) = B_0 + B_1 * dist_dummy + B_2 * dist^2 + B_n * Ommited_n$

# 4 Results

Our features are uncorrelated, the predicted vs actual fit is on the 45 degree line, and our errors are normal.

In table 3 we increase the robustness of our analysis by adding controls for additional characteristics which likely biased our estimation from our simple linear regression. After holding characteristics regarding neighborhood, year, and whether the house was newly built or not, the significance of distance to a subway station on price decreases to -0.0000268. Due to the decrease in the magnitude in the coefficient for subway distance, we concluded that these characteristics were causing a negative bias on our variable. We include the square of the distance as well, to account for non-linearity noticed, that houses further away from subway stations have a more negative relationship than a linear fit suggests. We also see that holding these variables constant there were other independent variables which accounted for a much greater percentage of the difference in price. As expected the year the house was sold had a high significance with houses sold in 2024 being associated with a 2% price increase compared to the mean house. This trend was similar with other years with houses being sold closer to the modern day being associated with greater price increases. This kind of relationship is expected as the price of real estate has tended to increase in the past 3 decades. We also noted a higher significance relating to the neighborhood in which the house was sold than our original independent variable of closest subway station. Affluent neighborhoods such as Kensington were associated with a 1.5

# 5 Discussion

## 5.1 Limitations of Results

The analysis of our limitations can be broken into internal and external validity. Our regression is misspecified and has an omitted variable bias. One potential omitted variable could be house size. We would expect house size to explain some of the variation in house price since larger houses command higher prices. However without further analysis it would be difficult to estimate the effect of house size since larger houses are in less dense areas of the city where property is also generally cheaper. We also run into the problem of simultaneous causality, areas that become trendy and develop would command higher housing prices, which could justify a new subway station nearby. We attempt to control for this using time-invariant district fixed effects but that does not solve the time-variant relationship where growing neighborhoods get subway allocations. In a future work we would frame a subway being built as a treatment, and attempt a difference-in-difference analysis, so we can gauge economic significance and price effects of subways without having to consider some of the other omitted variables. In our analysis we also assumed that our data was i.i.d. However we can assure that this may not be true as we saw in the 2008 financial crisis when many people sell houses at the same time it leads to a depreciation in house price which leads to more people selling their homes. This means that one person selling their house may affect the decision of another person to sell their own home. The external validity of our regression is limited. The population studied are homes in the Greater London Area, with observations from 1995 to 2024. We can use our model to make inferences on homes that fit within this sample space, however, expect pricing dynamics to be different between cities and regimes. For example, in smaller cities we might expect price variation to be independent of subway locations if there is sufficient coverage.

## 5.2 Conclusion

Houses closer to subway stations in the Greater London Area see higher prices than those further away. This analysis includes controls for time, district, and a polynomial fit for the distance and price relationship. We would warn against assuming external validity, and exogenous factors like governance and housing density could affect inferences. Our results are statistically significant but economically insignificant.

# References

Muth, R. F. (1969). *Cities and Housing: The Spatial Pattern of Urban Residential Land Use.* Chicago and London: The University of Chicago Press.

Zhou, Z., H. Chen, L. Han, and A. Zhang (2019, Feb). The effect of a subway on house prices: Evidence from shanghai. *Real Estate Economics 49*(S1), 199–234.

# Appendix

## A Appendix