Does being close to the Subway Station affect Housing Prices in Beijing?

A Causal Inference final project for STA256

Yuehe Wen

December 2022

1 Introduction

Housing prices have always been a topic with great interest in China. For the majority of people living and working in the city, one important factor is whether there is convenient access to public transportation so that commuting to the workplace will become easier, especially after 2008 when the Beijing government has issued the restrictions on private drivings. Previous research from Wen(2018) has shown that housing prices increase when the distance of the houses from nearby subway stations decreases in Hangzhou, China. In this project for STA 256, I would like to discover whether there is a causal effect between being close to subway stations and the housing price per square meter in Beijing.

1.1 Dataset

The dataset I will use is from Kaggle(https://www.kaggle.com/datasets/ruiqurm/lianjia). It contains housing prices of Beijing fetched from Lianjia.com. Key variables in the dataset include total price, price, square, number of living rooms, number of drawing rooms, number of kitchens and bathrooms, building type, construction time, renovation

condition, building structure, ladder ratio (which is the proportion between number of residents on the same floor and number of elevator of ladder. It describes how many ladders a resident has on average), whether there is an elevator, whether it is close to the subway, whether it has the five-years-property, district and community average price. Most of the data is traded in 2011-2017, some of them is traded in 2018, and some is even earlier(from 2009 to 2010).

There are several variables that are qualitative and are represented by indicators. For variable *Building Type*, 1 means tower, 2 means bungalow, 3 means combination of plate and tower and 4 means plate. For variable *Renovation Condition,* 4 means hardcover, 3 means simplicity, 2 means rough and 1 means other. For variable *Building Structure,* 6 means steel-concrete composite, 5 means steel, 4 means brick and concrete, 3 means brick and wood, 2 means mixed and 1 means unknown and other.

1.2 Assumptions

I define the treatment *Z* here to be whether close to the subway station or not. *Z* = 1 means that the house is near the subway station thus in the treatment group, and *Z* = 0 indicates that the house is not close to the subway station thus in the control group. In the dataset, 61% of the units are in the treatment group.

In order to make inference from observational study, several strong assumptions are needed for the project.

1.2.1 *Selection bias terms are zero:*

$$E\{Y(0) \mid Z = 1, X\} = E\{Y(0) \mid Z = 0, X\}$$
$$E\{Y(1) \mid Z = 1, X\} = E\{Y(1) \mid Z = 0, X\}$$

This assumption states that the differences in the means of potential outcomes across the treatment and control groups are due to the difference in the observed covariates *X*. In our situation, it means that given the same value of covariates, the housing prices per square meter have the same mean across near-subway groups and not near-subway groups.

1.2.2 *Strong Ignorability proposed by Rubin and Rosenbaum in 1983:*

$$\{Y(1), Y(0)\} \perp\!\!\!\perp Z \mid X$$

This indicates that the potential outcomes (price) are independent of the treatment variables given the observed covariates.

In the project, the covariates *X* are as following:

1) Floor: the height of the house
2) Building Type: the type of house
3) Building Structure: the construction structure of the house
4) Elevator or not: whether the house has access to elevators
5) FiveYearsProperty: whether the house has property rights for five years
6) District: where the house is located in Beijing

These variables are selected as confounders as they might have influences on the potential outcomes. District is an obvious factor that will affect the housing prices. It is common that closer to downtown a district is, higher the housing price in this district. Building type and structure are also determinants as apartments and bungalows usually have different prices, and different types usually have different structures. Elevator can be viewed as an indicator of whether this house is new or not. Five years property

means that if the owner has owned the house for more than five years, there will be a discount on the selling tax.

In the dataset there are some missing values and illegal characters due to web scratching, but these only account for 1% of the whole data, so I just drop these values. Besides, some houses have prices that are very low, which is not realistic in Beijing. According to current market information, I deleted all units which have a price lower than 8000 Yuan as these units only account for 0.2% of the whole dataset.

2 Methodology and Results

I will follow the methodology learned from class, especially the part of observational studies in the class notes provided by Professor Peng Ding.

2.1 Regression

The first simple method I will use is to run the OLS of the observed outcome on the treatment indicator, covariates, as well as the interaction between covariates and treatment, which is recommended by Lin (2013). This model is more optimal than the simple linear model as it includes the treatment effect heterogeneity induced by the covariates. The model then is:

$$E(Y|Z,X|) = \beta_0 + \beta_z Z + \beta_x^T X + \beta_{zx}^T XZ$$

The average causal effect then becomes:

$$\tau = \beta_z + \beta_{zx}^T \bar{X}$$

In order for this estimator to be held, the assumption of ignorability and linear model are needed. As mentioned in the notes, if the covariates are centered so that $\bar{X}$ = 0, the

estimator then becomes $\hat{\beta}_z$. Besides, one point estimator is not enough for computing the standard errors, so I will use the method of bootstrap to acquire the standard error of the estimator.

The result of the regression analysis is:

Average treatment effect estimator	Bootstrapped standard error
10113.9	74.4

The result shows that there is a significant positive effect of being close to a subway station on housing prices when conditioning on all the other covariates. We have 95% confidence that houses located near subway stations have 9968.2 ~ 10259.8 higher prices in Yuan than houses not near subway stations. This result can serve as a basic benchmark for this project, as regression is the simplest method.

2.2 Propensity Score Matching

Propensity score is a key concept proposed by Rosenbaum and Rubin (1983b) that is used in causal inference with observational studies. The propensity score is defined:

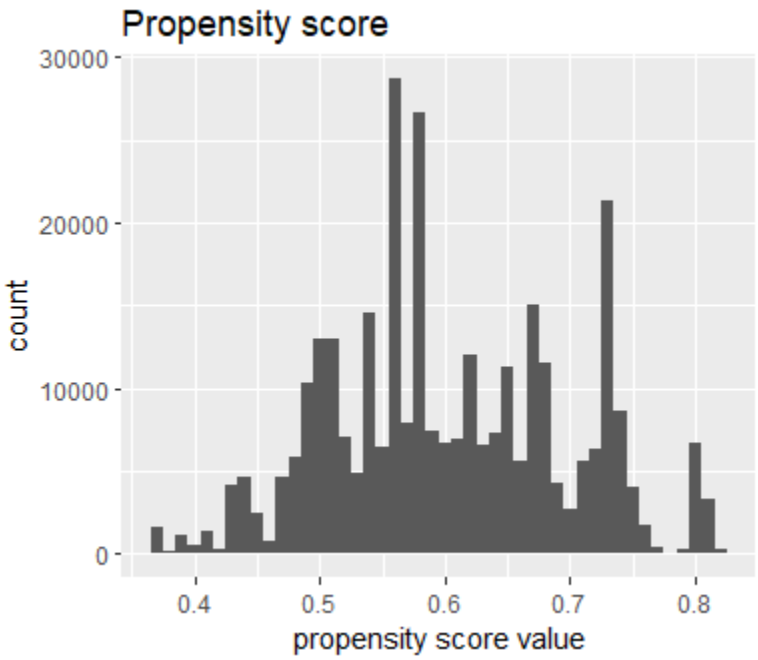
$e(X, Y(1), Y(0)) = pr\{Z = 1 \,|X, Y(1), Y(0)\} = pr(Z = 1|X)$.

Under the assumption of strong ignorability, the propensity score can be reduced to:

$$e(X) = pr(Z = 1|X)$$

, which is the conditional probability of receiving the treatment given the observed covariates. For the propensity score, there is another assumption called overlap or positivity condition, which restricts that $0 < e(X) < 1$.

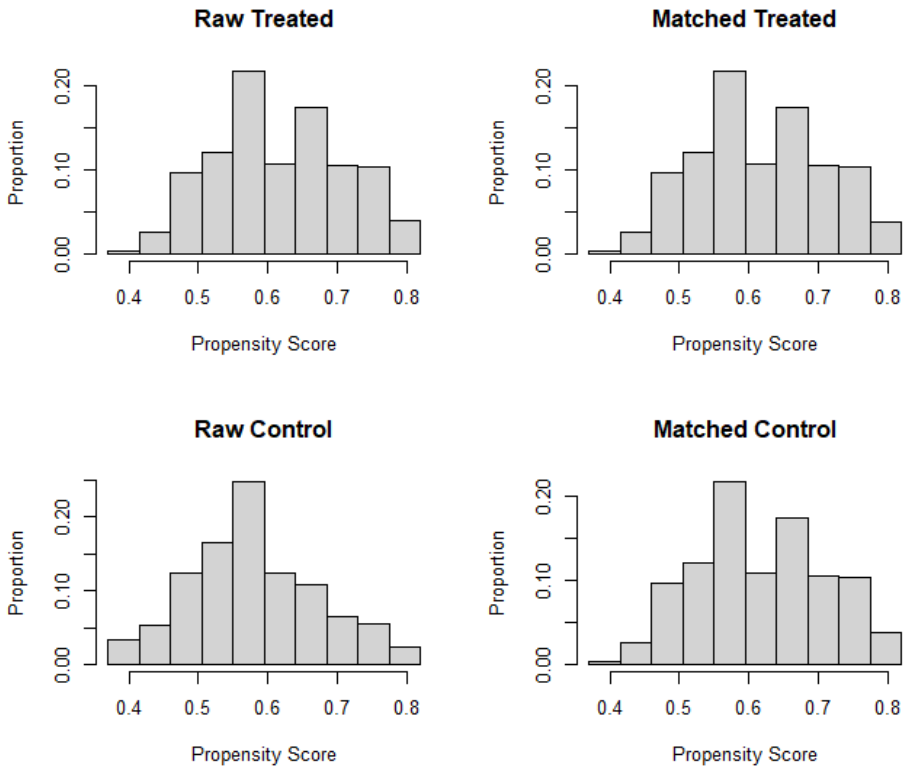
In general studies, the propensity score is unknown and not discrete. So I will first obtain the estimated propensity score $\hat{e}(X)$ via a logistic model. The estimated propensity scores are:



This histogram shows that most of the propensity scores fall into the range of 0.4 ~ 0.8, which satisfies the overlap assumption. There are two peaks in the graph, one is around 0.55 and the other is around 0.75. Since the treatment is binary, it is reasonable to have such scenarios, but the values are a little skewed right due to some possible underlying covariates bias, resulting in that the model is more confident to classify the unit as near to the subway stations. Ideally, the distribution of the propensity scores should be around 0.5, showing that given the covariates, each unit has the same probability of receiving treatment or not. Next step, I will use the propensity score matching in order to see if it is possible to eliminate the selection bias to some extent.

One flaw for this method is that in the dataset, the number of treated units is larger than the number of controlled units, so common distance matching methods such as nearest

neighbor cannot be used in this situation. I choose the method of generalized full matching, which uses a clustering algorithm to assign every unit in the treated group to a subclass in the controlled group.



The result shows the raw propensity scores of control and treated groups, as well as the matched scores for two groups. The distribution of the treated group does not change much after matching, while the control group has a flatter and more uniform distribution after matching. Besides, compare the standard mean difference and variance ratio before and after:

	Std. Mean Difference Before	Std. Mean Difference After matching	Variance Ratio Before	Variance Ratio After matching
Distance	0.4034	-0.0001	1.0087	0.9997

Floor	0.2367	0.0321	1.1052	1.0450
Building Type	-0.2904	0.0004	1.2888	0.9945
Building Structure	0.1943	-0.0039	0.9154	1.0052
Elevator	0.2146	0.0398	0.9423	0.9820
Five Years Property	0.1244	-0.0184	0.9301	1.0138
District	-0.2285	-0.0025	1.0035	1.0559

Standard mean difference approaching 0 or variance ratio approximating to 1 indicate good covariates balance. It can be seen that before matching, there are imbalances between the treated and non-treated groups, which is a sign of selection bias before treatment; while this imbalance fades away after the usage of propensity score matching.

The average treatment effect and standard error after propensity score matching are calculated via direct comparison and fitted Lin's regression model $E(Y|Z,X|) = \beta_0 + \beta_z Z + \beta_x^T X + \beta_{zx}^T XZ$ separately. The results of the two methods are as follows:

	Average Treatment Effect	Standard Error	95% CI
Direct Comparison	4548.7	77.9	(4396.0, 4701.4)
Regression Model	4694.46	61.9	(4538.4, 4850.5)

The two results do not differ from each other much, both indicating that there exists an approximate 4300 Yuan higher difference in price between houses near subway stations and houses not near subway stations. This value differs a lot from the previous

simple regression outcome, which is 10121, as the matching method balances the covariates so that the potential selection bias in the regression model is eliminated. However, there might exist some problems in the propensity score matching. As noted in the work of King and Nielsen (2019), the process of propensity score matching is to approximate a complete randomized experiment rather than a blocked randomized experiment within the observational datasets, so it might fail to eliminate the covariate imbalance that can be eliminated by blocking. Further, in our dataset it is impossible to pair each unit. Therefore, a propensity score stratification method might be more appropriate for the dataset.

2.3 Propensity Score Stratification

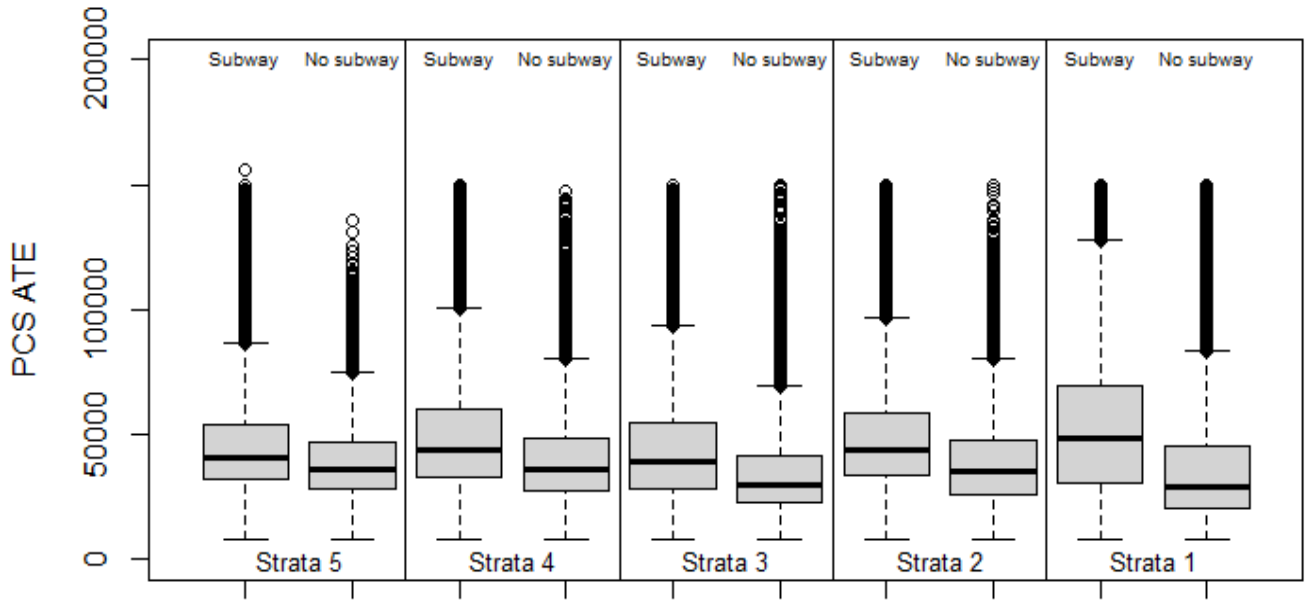
Stratification is designed to reduce the bias due to undesirable treatment allocation in complete randomized experiments. The principle is to generate subgroups first, known as strata, based on covariates and then conduct randomization experiments within each strata. For propensity scores, it is assumed that they are known and only take *K* possible values $\{e_1, \ldots, e_K\}$ where *K* is much smaller than n. Then the formula becomes:

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid e(X) = e_k \quad (k = 1, \ldots, K).$$

Then, we have *K* independent complete randomized experiments within strata of the propensity scores. In reality, the propensity scores are not known and not discrete, so the common method is to discretize the estimated propensity score by *K* quantiles to obtain $\hat{e}'(X) : \hat{e}'(X_i) = e_k$. I will estimate the causal effect within each stratum and construct the final estimator by a weighted average at a different value of *K*.

First I will use $K = 5$, which is suggested by Rosenbaum and Rubin (1984) as it is able to remove approximately 90% of the bias in the unadjusted estimate.

The boxplot of treatment and control group within each strata is:



It can be directly observed that within each strata of propensity scores, there are obvious positive treatment effects in housing prices for treatment and control groups. This result is roughly consistent with what we got from propensity score matching, while here we can examine the treatment effect more in detail. The following table compares the average treatment effect and standard error for K taking 5, 10 and 20:

	Average Treatment Effect	Standard Error
K = 5	10088.6	74.2
K = 10	9631.0	75.6
K = 20	9390.6	75.7

It can be seen that when K grows larger, the standard error becomes larger while the treatment effect decreases. As mentioned in previous research, the suggested value of K is between 5 and 10 as larger K will have less additional effects and also means there are lower samples in one strata.

2.4 Inverse Propensity Score Weighting

The next estimator I will use is the method proposed by Horvitz and Thompson (1952): the inverse propensity score weighting (IPW) estimator, also called HT estimator for the average causal effect:

$$\hat{\tau}^{\text{ht}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)},$$

where $\hat{e}(X_i)$ is the estimated propensity score. But this estimator has a problem that it is not invariant to location transformation of the outcome. To fix this, Hajek(1971) modified this estimator to become more stable:

$$\hat{\tau}^{\text{hajek}} = \frac{\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^{n}\frac{Z_i}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}}{\sum_{i=1}^{n}\frac{1-Z_i}{1-\hat{e}(X_i)}}.$$

Similarly for the regression outcome, I calculated the IPW average treatment effect estimator and obtained the standard error by bootstrap.

Truncation Level	HT ATE	HT Std.	Hajek ATE	Hajek Std.
0	10000.78	78.371	10171.94	78.8
0.01	10000.78	88.3	10171.94	88.0
0.05	10000.78	71.01	10171.94	70.7

0.1	10000.78	69.4	10171.94	68.7

The HT estimator is very unstable and sensitive to extreme propensity scores values (values that are very close to 0 or 1). The result table is consistent with my findings before. The histogram of propensity scores has shown that most of the scores fall into the range of 0.4 to 0.9, and there are few values that are very close to 0 or 1. It indicates that the propensity scores are under the assumption of overlap, so there should not be great changes in the average treatment effects in different truncation levels. And the HT estimator under this non-extreme situation achieves a similar performance as the Hajek estimator.

Both of the estimators indicate that for houses near subway stations, their prices are about 10000 Yuan higher than those that are not near subway stations. The result value is similar to the result obtained in propensity score stratification, which is 10088 Yuan when $K = 5$. It can be concluded that both inverse propensity score weighting with truncation and stratification can make the propensity scores more robust and reduce the impact of potential bias.

2.5 Doubly Robust Estimator

When addressing real-life observational study, sometimes we cannot be 100% sure about whether the propensity scores or the outcome model are both correctly specified. The doubly robust estimator presents a good way to combine the propensity scores and regression model in a way that we don’t have to rely on either of them. It is also possible to augment the IPW estimator by the imputed outcomes, which strengthens the

theoretical properties of strong ignorability and overlap. The outcomes can be expressed with parameters β_i:

$$\begin{aligned}\mu_1(X) &= E\{Y(1) \mid X\} = E\{Y \mid Z = 1, X\}, \\ \mu_0(X) &= E\{Y(0) \mid X\} = E\{Y \mid Z = 0, X\}\end{aligned}$$

If the outcome model is correctly specified, then:

$$\mu_1(X, \beta_1) = \mu_1(X) \text{ and } \mu_0(X, \beta_0) = \mu_0(X).$$

A working model is also posited for the propensity score $e(X, \alpha)$, indexed by the parameter α. If the propensity score is correctly specified, then $e(X, \alpha) = e(X)$.

Then the doubly robust estimator is defined as:

$$\begin{aligned}\tilde{\mu}_1^{\text{dr}} &= E\left[\frac{ZY}{e(X,\alpha)} - \frac{Z - e(X,\alpha)}{e(X,\alpha)}\mu_1(X,\beta_1)\right], \\ \tilde{\mu}_0^{\text{dr}} &= E\left[\frac{(1-Z)Y}{1-e(X,\alpha)} - \frac{e(X,\alpha) - Z}{1-e(X,\alpha)}\mu_0(X,\beta_0)\right]\end{aligned}$$

And the treatment effect is defined as:

$$\hat{\tau}^{dr} = \hat{\mu}_1^{dr} - \hat{\mu}_0^{dr}.$$

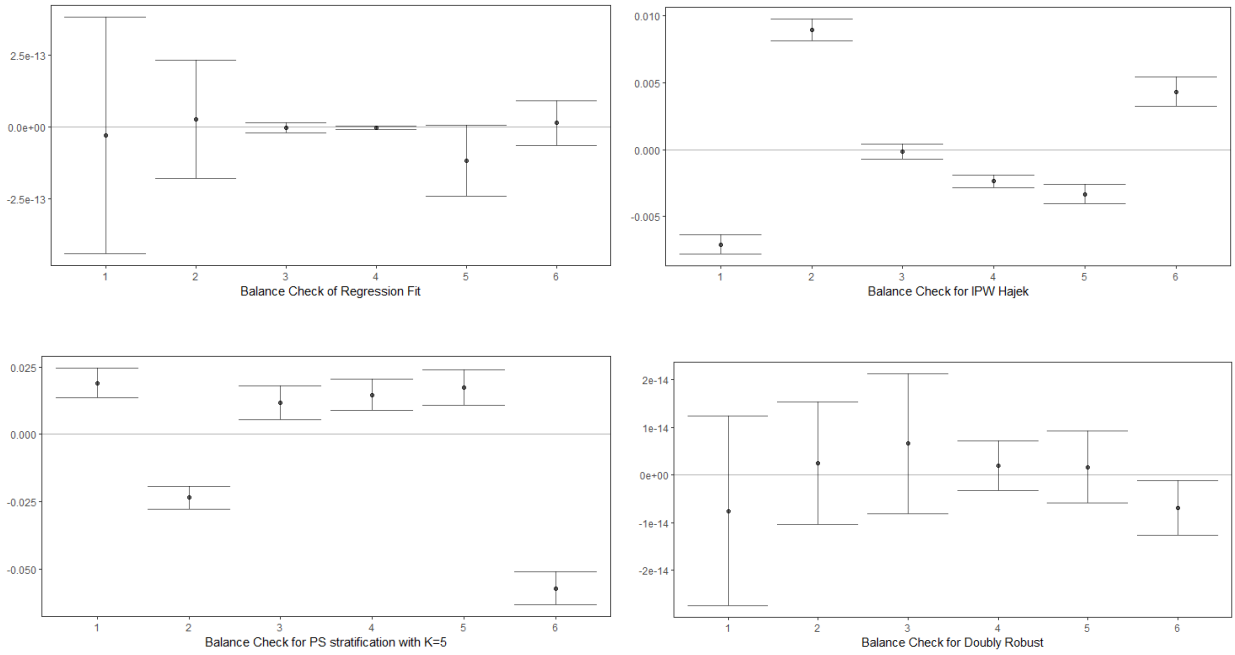
The result of the doubly robust estimator is:

Truncation Level	Doubly Robust ATE	Doubly Robust Std.
0	10197.67	66.92
0.01	10197.68	66.11
0.05	10197.67	71.85
0.1	10197.68	75.54

It can be seen that the average treatment effect generated by the doubly robust method is consistent with most of the previous estimators, which shows that there is a positive effect of being close to subway stations

3 Covariates Balance Check

For all the previous methods used, the results are under the assumption of strong ignorability and unconfoundedness for measured or unmeasured variables. So it is important to measure the validity of these assumptions by checking the balance of covariates.

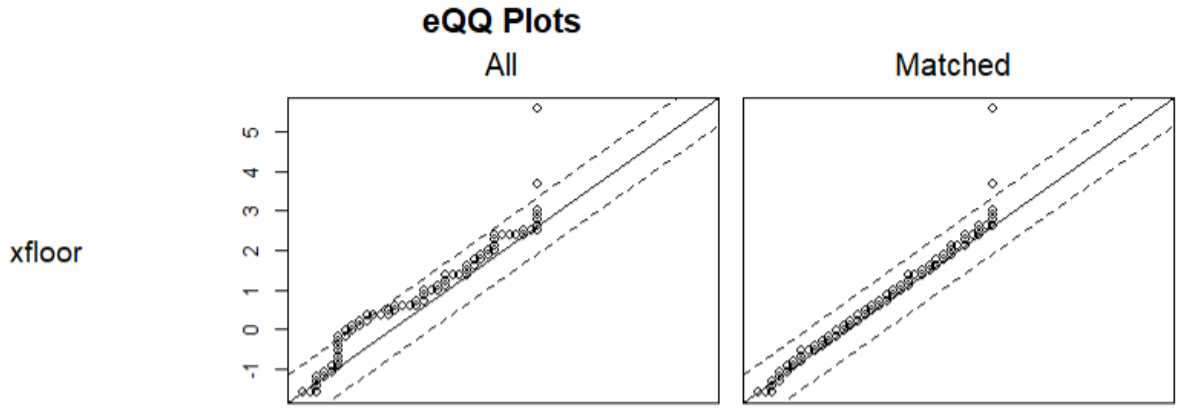


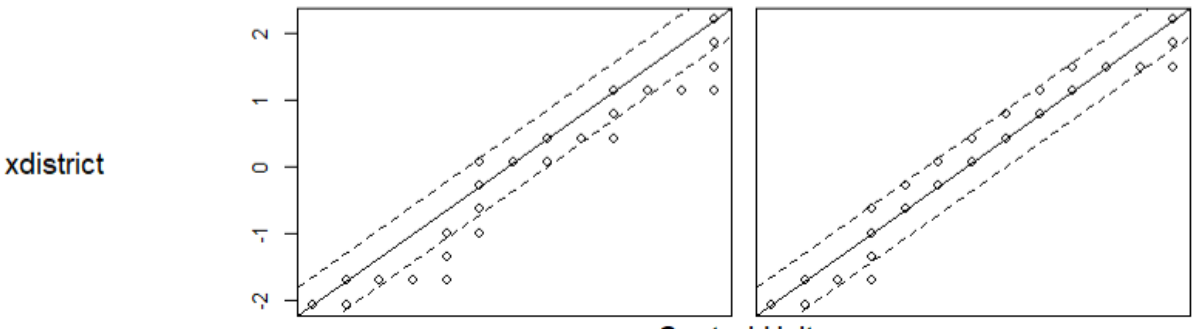
The above figures show the balance check for regression fit, Hajek estimator, propensity score stratification with K = 5, and the doubly robust estimator. Although these four methods all generate a consistent result that there is a positive effect (about 10,000 Yuan higher) of being close to the subway stations on the housing price, it can be spotted from the balance plots that there are significant bias treatment effects

especially for the Hajek estimator and propensity score stratification. This indicates that our propensity score model might be poorly specified. It is also noted that the doubly robust estimator, though having a mis-specified propensity score model, still forces a good covariate balance, which is a key benefit of this method.

The covariates floor and district have been the strongest confounder that will affect the housing price, which is in some way consistent with the reality. The research from Wonseok (2019) has already mentioned that the effect of subway stations on housing prices is greatly related to the average income and convenience of certain neighborhoods. In Beijing, districts near the downtown area usually mean closer distances to higher-paid workplaces and more convenient life, so even a small and old house might have a very high price. The ideal way to eliminate this confounder's effect is to use the method of matching within each district, but unfortunately there is not enough to conduct a full analysis on it.

As noted before, the propensity score matching method generates a result of 4,000 Yuan effect which is different from all the other methods. The major reason is that there are more treated units in the datasets than the controlled units, so it is impossible to match every treated unit to one or more controlled units, thus giving a biased result.





The above plot shows the QQ plot for those subgroups before and after matching. It can be seen that after matching, these two variables have better balances, so if there are more datasets, matching might become a good method.

4 Conclusion

In the project, I have applied several widely used methods in causal inference to deal with observational study and discovered that houses being close to the subway stations have higher prices of about 10,000 Yuan than houses not being close to the subway stations in Beijing. This result is intuitive and there also exists previous research stating the positive effect of access to public transportation on housing prices in other cities. However, the numerical value should still be under consideration as most of the methods do not seem to have an ideal covariate balance, which indicates that there might be some confounder effect on the treatment outcome from either measured covariates or unmeasured covariates. This has always been an important but hard problem to be solved for observational study as the assumption is too strong and the dataset I used does not have enough information.

Nevertheless, it can be observed that though facing imbalance, many properties such as the robustness of the doubly robust estimator and the consistency of all the estimators did seem to be followed as in the theories.

5 Reference

Wen H, Gui Z, Tian C, Xiao Y, Fang L. Subway Opening, Traffic Accessibility, and Housing Prices: A Quantile Hedonic Analysis in Hangzhou, China. Sustainability. 2018; 10(7):2254. https://doi.org/10.3390/su10072254

Winston Lin. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." The Annals of Applied Statistics, 7(1) 295-318 March 2013. https://doi.org/10.1214/12-AOAS583

Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, vol. 70, no. 1, 1983, pp. 41–55. *JSTOR*, https://doi.org/10.2307/2335942 .

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis, 27*(4), 435-454. https://doi.org/10.1017/pan.2019.11

Paul R. Rosenbaum & Donald B. Rubin (1984) Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, Journal of the American Statistical Association, 79:387, 516-524, DOI: 10.1080/01621459.1984.10478078

Wonseok Seo, Hyung Kwon Nam,Trade-off relationship between public transportation accessibility and household economy: Analysis of subway access values by housing size,Cities,Volume 87,2019,247-258,264-2751, https://doi.org/10.1016/j.cities.2018.11.004 .

Kaggle Dataset, “Housing price in Beijing”, https://www.kaggle.com/datasets/ruiqurm/lianjia

Professor Peng Ding’ s lecture notes, “CausalInference2021”, from bCourses