Modeling Chicago Home Prices with Socioeconomic Variables

Sangwoo Park

February 2017

Abstract

Urban residential housing prices depend on neighborhood characteristics such as crime rate, median household income, transportation accessibility, or quality of education, in addition to the characteristics of the housing unit itself. This study estimates the impact of such neighborhood characteristics on housing prices in the city of Chicago. Data for home prices and socioeconomic factors of neighborhoods from 2012 to 2016 are analyzed using the OLS and the GLS panel regression methods. Results from the random effects model for panel data, which has the best predictability and satisfies all the statistical assumptions, show that crime rate negatively affect median housing prices, and that having a cta station, and having high median yearly income positively effect median housing prices of the community. We also observe a significant interaction effect between household income and quality of public education towards the value of home prices.

1 Introduction

The goal of this project is to predict the median price of residential properties by neighborhoods in Chicago. Any potential buyer of residential property may find out whether home prices in a specific neighborhood are overvalued or undervalued for the neighborhood's socioeconomic characteristics and preferences, e.g. with respect to the crime rate, median household income, transportation accessibility, or quality of education. The working hypothesis is that the median housing price is correlated with the median income, the quality of education, and the existence of a cta station nearby, while anticorrelated with the crime rate.

2 Data

1. Median Home Price (`value`)

   The data for home prices in 77 Chicago community areas are scraped from Trulia, an online residential real estate website. BeautifulSoup4, a Python library for extracting data out of HTML, was used for data scraping. From the gathered data, we calculated the median home price of total homes sold each year in each community area from 2012 to 2016. Not a single home was traded in Riverdale neighborhood from 2012 to 2015 at Trulia, so we imputed the missing values with the median home price sold in Riverdale in 2011. The unit of the variable `value` is US dollars.

2. CTA (`cta`)

   The dummy variable `cta` indicates whether a neighborhood has a CTA station (`cta` = 1) or not (`cta` = 0). I manually created a python dictionary, of which the key is community area code (1-77) and the value is 0 or 1, by looking up all the locations of CTA L stations in Chicago from the Chicago Data Portal.

3. Crime (`crime`)

   The crime rate variable `crime` is the number of total occurrences of violent crime per 100,000 residents in a year. Violent crime consists of murder, assault and burglary. The raw data is obtained from the Chicago Data Portal, and I have cleaned the data by using the Pandas library in Python.

4. Quality of Public Education (`school_score`)

   The quality of education variable `school_score` is the mean score of total progress reports of all public high schools in a neighborhood in specific year. This data is also from the Chicago Data Portal. We used a Python library 'chicago-neighborhood-finder' by Craig M. Booth for converting the (longitude, latitude) of high schools into community area code. The unit of the variable is score that ranges from 0 to 100.

5. Income (`income`)

   The income variable `income` is the mean yearly household income in a community. The raw data is from the Chicago Department of Planning and Development. Since data was only available for years 2010 to 2012, we imputed the data for 2013 to 2016 by using linear interpolation. The variable is in units of US dollars.

All of the python codes that were used can be found in the github repository: `https://github.com/spark01217/project/`. All of the data described above has a ratio-interval level of measurement. Although `cta` variable is an ordinal (dummy) variable, I will treat the variable as a ratio-interval level in the analysis. The minimum unit of analysis is a neighborhood in Chicago, denoted by the variable `community`, ranging from from `community` = 1 (Rogers Park) to `community` = 77 (Edgewater).

3 Analysis

3.1 Model Building: Pooled OLS

	(1) value		(2) value	
school_score	348.8	(0.3756)		
crime	-3.442**	(0.0051)	-3.719**	(0.0018)
income	3.337***	(0.0000)	3.422***	(0.0000)
cta	41672.0***	(0.0000)	41852.1***	(0.0000)
constant	-2838.0	(0.9025)	13971.6	(0.2939)
N	385		385	
R^2	0.5691		0.5682	

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1: Comparison of the two OLS model outputs

First, we disregard the effect of time to the variability of our independent variables, and build a pooled OLS model with all of our independent variables to test whether the inclusion of each variable is statistically significant. According to the regression output in Table 1, Our model is as following:

$$\texttt{value}_i = 3.337 * \texttt{income}_i + 348.790 * \texttt{school\_score}_i - 3.442 * \texttt{crime}_i + 41671.96 * \texttt{cta}_i - 2838.006 + \texttt{e}_i \quad (1)$$

where i stands for the community code number.

	school_score	crime	income	cta
school_score	1.0000			
crime	-0.4334	1.0000		
income	0.5523	-0.4373	1.0000	
cta	-0.0161	0.2024	0.0224	1.0000

Table 2: Correlation between features

However, when correlations between each independent variable are analyzed, we find significant corrleations between `crime` and `school_score`, and between `income` and `crime`, as shown in Table 2. The correlation between `income` and `school_score` is 0.5523, which is so high (>0.5) that we might consider removing either `income` or `school_score` variable from our model. In addition, we find the the p-value for the t-test whether the coefficient β_2 for `school_score` is equal to 0 is 0.376. Since it is logically sound that students from low income families, or area with high crime rate achieve low academic performance, we remove the `school_score` variable from our model. Then, our OLS model becomes

$$\texttt{value}_i = 3.422 * \texttt{income}_i - 3.719 * \texttt{crime}_i + 41852.1 * \texttt{cta}_i + 13971.6 + \texttt{e}_i. \quad (2)$$

The two models exhibit comparable performance in terms of their R^2 values.

3.2 Interaction Effects

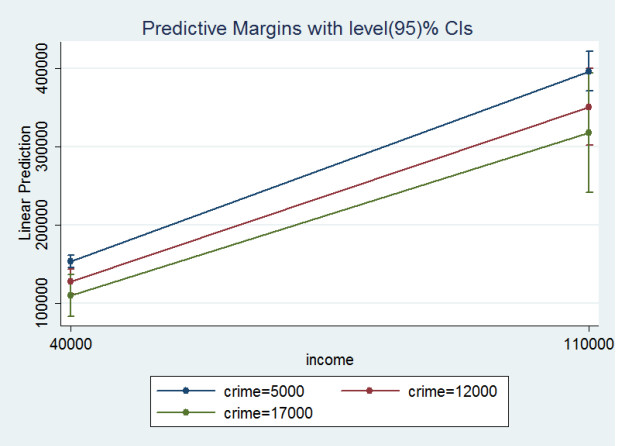
Now, we test whether we should add interaction effects between pairs of independent variables. The presence of a significant interaction indicates that the effect of one independent variable on the dependent variable depends on the values of another independent variable. In Figure 1 we plot multiple regression lines on one graph, holding one independent variable of the pair constant, and find no significant differences in the slope with respect to the value of the independent variable we are holding constant. Moreover, since the p-values of the t-test including interaction terms are not statistically significant, we should not include interaction effect in our pooled regression model.

	(1) value		(2) value		(3) value	
income	3.682***	(0.0000)	2.274***	(0.0000)	3.488***	(0.0000)
crime	-1.893	(0.3825)	-4.684***	(0.0001)	-1.278	(0.5425)
income*crime	-0.0000422	(0.3150)				
cta	42269.5***	(0.0000)	-33632.1*	(0.0500)	58662.8***	(0.0000)
income*cta			1.707***	(0.0000)		
crime*cta					-3.293	(0.1595)
constant	1145.7	(0.9504)	69878.6***	(0.0001)	-634.7	(0.9700)
N	385		385		385	
R^2	0.5694		0.5933		0.5705	

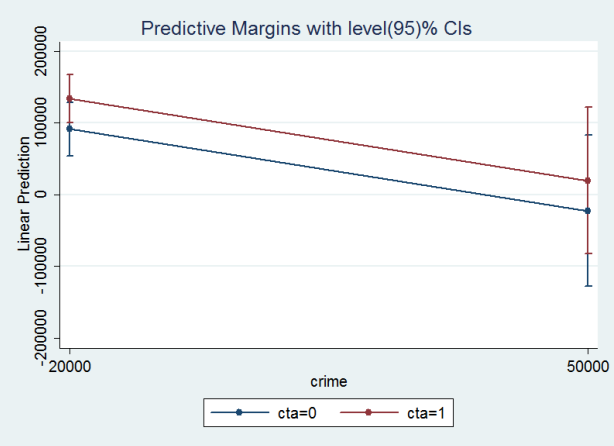
p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

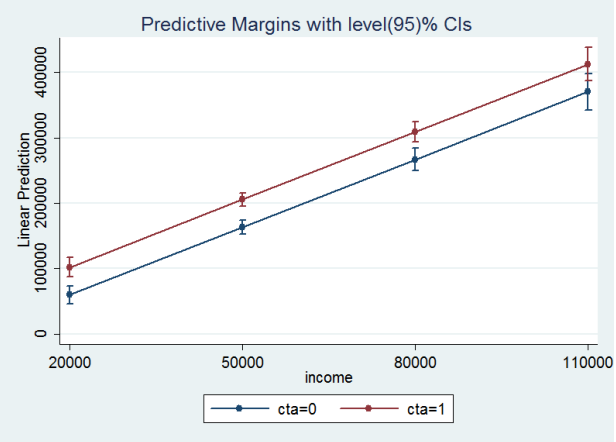
Table 3: Regression output of OLS models with single interaction term between two factors



(a) Interaction between `income` and `crime`



(b) Interaction between `crime` and `cta`



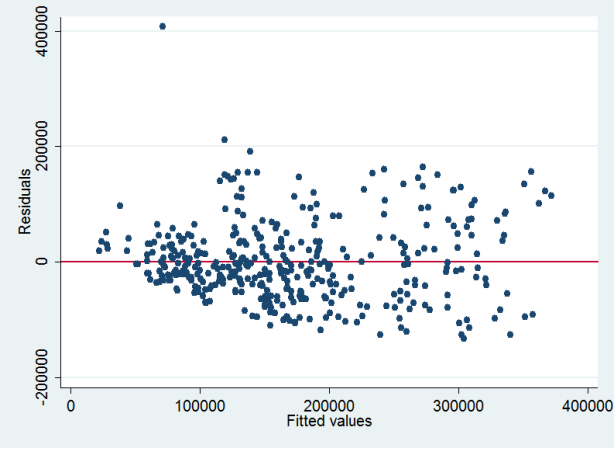
(c) Interaction between `income` and `cta`

Figure 1: Plots of interaction effects between independent variables

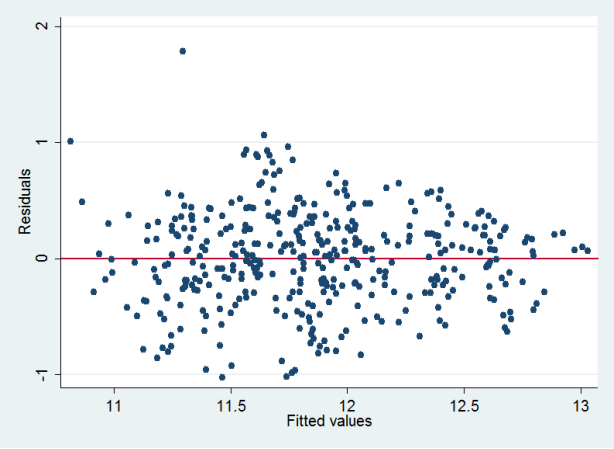
3.3 Model Robustness: Pooled OLS

3.3.1 Homoscedasticity of Errors

In order to test homoscedasticity of the errors, we plot residuals against the fitted values to see whether the error terms are the same across all values of independent variables that predict our dependent variable. In the left plot of Figure 2, we detect heteroscedasticity as we see a pattern in the residual-fitted plot where the residuals grow as the fitted values increase. To treat this observed heteroscedasticity, we apply a log-transformation to our dependent variable `value`. The residual-fitted plot of our new model has randomly distributed residuals across the entire range of fitted values, as shown in the right panel of Figure 2.



(a) Residuals before the log-transformation of `value`



(b) Residuals after the log-transformation of `value`

Figure 2: Comparison of residuals before and after the log-transformation of `value`

The log-transformed model is given by

$$\log(\texttt{value})_i = .0000188 * \texttt{income}_i + -.0000424 * \texttt{crime}_i + .2270081 * \texttt{cta}_i + 11.1215 + \texttt{e}_i. \tag{3}$$

3.3.2 Normality of Residuals

To test the normality of residual assumption of our OLS model, we employ a quantile-quantile plot of residuals. As the Q-Q plot in Figure 3 exhibits a linear distribution of the points, we can confirm that our residuals are normally distributed.

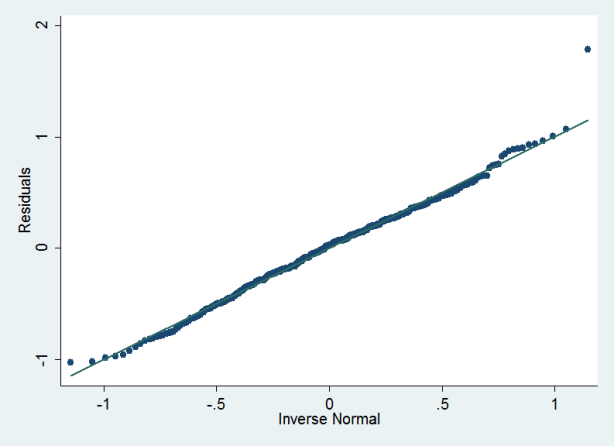
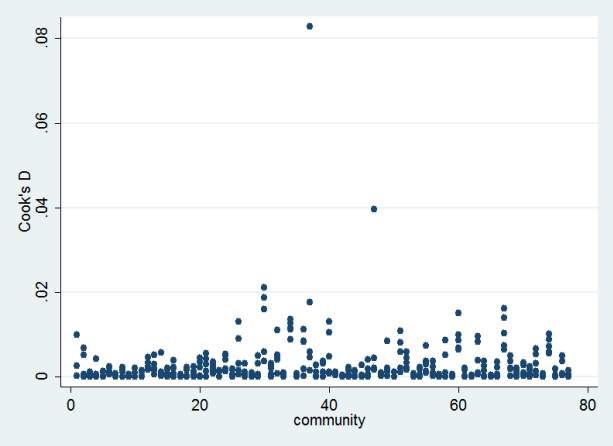


Figure 3: Q-Q plot of residuals

3.3.3 Linearity and Outliers

From the scatter plot of the dependent and the independent variables in Figure 4, we fail to see any nonlinear patterns. Next, in order to detect any outlier observations that negatively affect our regression model, we turn our attention to the cook's distance plot in Figure 4, which is a combination of each observation's leverage and residual. As a rule of thumb, I dropped all the observations that had cook's distance greater than $4/N$ where N is the number of total observations, resulting in 19 dropped observations. From the outlier-cleaned data, we obtain a new model as following:

$$\log(\texttt{value})_i = .0000193 * \texttt{income}_i + -.0000404 * \texttt{crime}_i + .2231793 * \texttt{cta}_i + 11.08814 + \texttt{e}_i. \quad (4)$$



(a) Cook's distance of all observations

(b) Scatter plots of all independent variables

Figure 4: Linearity and outliers

	(1) log_value	
income	0.0000206***	(0.0000)
crime	-0.0000377***	(0.0000)
cta	0.210***	(0.0000)
constant	11.00***	(0.0000)
N	359	
R^2	0.6002	

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Regression output after removing outliers

3.3.4 Autocorrelation

So far, we have found no evidence of failures in assumption that may lead to lack of robustness in our models. However, here we find that our current model violates the assumption of independence of errors. If we plot the residuals of our OLS model for each year, we see that our residuals fluctuate over time, which may suggest that there is an omitted variable that is time-variant and affecting our dependent variable directly. In practice, there are many time-variant factors that affect median home prices of neighborhood. Especially, macroeconomic fluctuations such as yearly changes in the interest rate significantly affect home prices. If the interest rate increases, the resulting increase in amortization of mortgage decreases potential home buyer's purchasing power, lowering the demand of the homes being sold. Therefore, instead of using OLS on the pooled data, I should treat the data as panel data in order to control for the differences in the variance of our independent variables over time, and use a model different from OLS to account for serial or panel correlation.

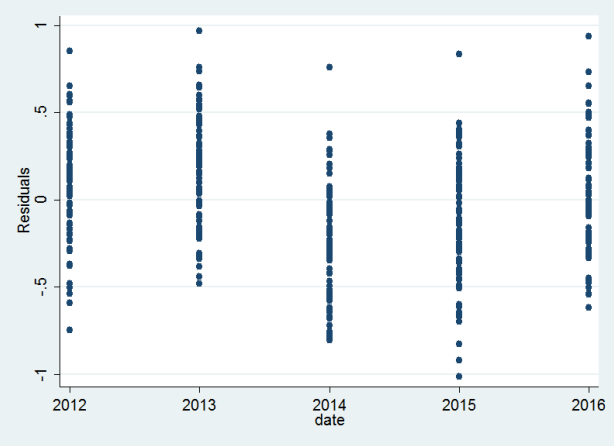


Figure 5: Distribution of residuals for each year

3.4 Model Building: Panel Data Analysis

We start by decarling the our dataset is a panel data, where the panel is `community` variable that represents 77 different neighborhoods in Chicago, and where the time variable is `date` that represents years from 2012 to 2016. To keep a balanced panel, we do not perform the outlier cleaning shown in Figure 4. We then consider two unobserved effects models for time series analysis; the fixed effects and the random effects models. While the fixed effects model controls for the the effects of time-invariant variables with time-invariant effects, the random effects model allows us to estimate the effects of time-invariant variables by assuming all unobserved factors are uncorrelated with the observed independent variables. Since potential unobserved factors that affect median home prices in Chicago neighborhoods include both time-invariant variables such as quality of local amenities (i.e. stores, gyms, theaters, etc.) or accessibility to transportation, and time-variant variables such as the interest rate as mentioned before, the random effects model is expected to have higher merit.

We also perform the Durbin-Wu-Hausman test, which can compare the consistency of coefficients for each independent variable under fixed and random effects models. First, we 'manually' execute a stepwise regression to build models to compare, with the regular (non-logarithmic) housing price as our dependent variable. Then, we employ the forward selection method, where we start with an empty model and repeat the process of adding one independent variable whose coefficient will be the most statistically significant. The threshold for including a variable will be a p-value of 0.1. Interaction effects are added by the same criterion. The resulting random effects model is

$$\begin{aligned}\texttt{value}_{i,t} = 1.913 * \texttt{income}_{i,t} - 2012.523 * \texttt{school\_score}_{i,t} + 0.0287 * \texttt{school\_score}_{i,t} * \texttt{income}_{i,t} \\ +35901.52 * \texttt{cta}_i + 100658.3 + \texttt{u}_{i,t} + \texttt{e}_{i,t},\end{aligned} \tag{5}$$

where i stands for community code from 1 to 77 and t stands for time from 2012 to 2016. Here, $\texttt{u}_{i,t}$ is the between-entity error, i.e. error that is caused by omitted variables that have different value across panels, and $\texttt{e}_{i,t}$ is the within-entity error caused by omitted variables that have same value across panels. Similarly, the resulting fixed effects model is given by

$$\texttt{value}_{i,t} = -716.271 * \texttt{school\_score}_{i,t} + 12.923 * \texttt{crime}_{i,t} + 138798 + \alpha_{i,t} + \texttt{u}_{i,t} \tag{6}$$

where $\alpha_{i,t}$ is the time-invariant error and $\texttt{u}_{i,t}$ is the between-entity error described above. The p-value of 0.0372 in Hausman test states that we should choose the random effects model. Therefore, our final model becomes the random effects model given in Eq. 5. For example, if median household income \$50000, crime rate is 4200, and mean of public high school transcripts is 50 in Hyde Park in year 2020, the median housing price is calculated as

$$\texttt{value}_{41,2020} = 1.913 * 50000 - 2012.523 * 50 + 0.0287 * (50000 * 50) + 35901.52 * 1 + 100658.3 = \$203333.67. \tag{7}$$

	(1) value	
income	1.913*	(0.0385)
school_score	-2012.5**	(0.0036)
income*school_score	0.0287*	(0.0462)
cta	35901.5*	(0.0122)
constant	100658.3*	(0.0186)
N	385	
R^2	0.5528	

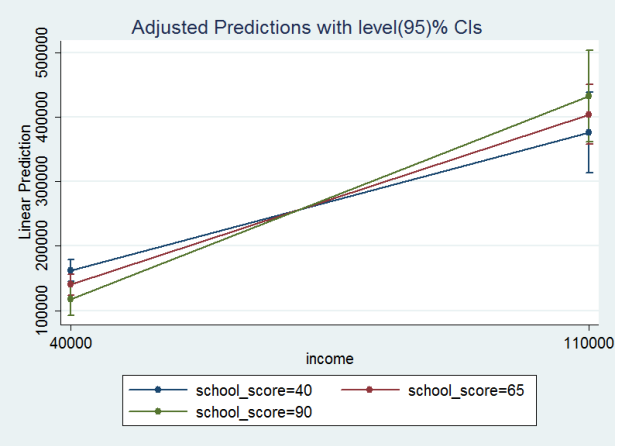
p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

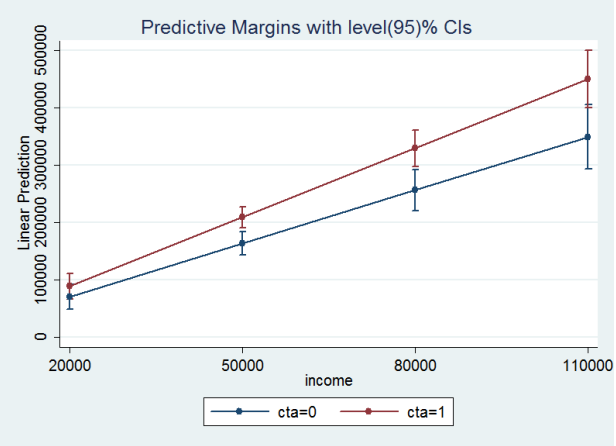
Table 5: Regression output of the random effects model

3.5 Interaction Effects

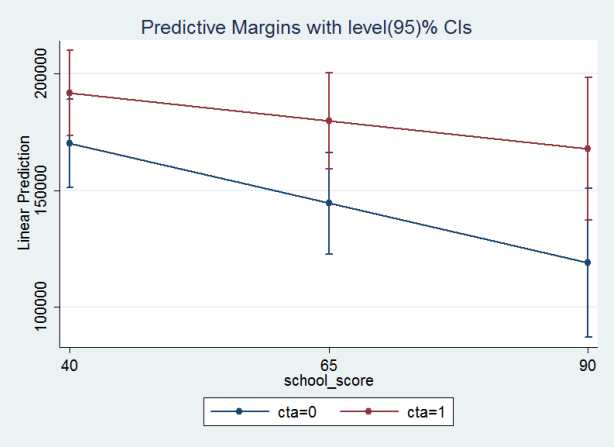
According to the Figure 6, we see that the interaction effect between `income` and `school_score` is statistically significant. Although we see slope differences in the margins graph by different `cta` value in Figure 6 (b) and (c), the interaction effects are negligible according to the Table 6.



(a) Interaction between `income` and `school_score`



(b) Interaction between `income` and `cta`



(c) Interaction between `school_score` and `cta`

Figure 6: Plots of interaction effects between independent variables

	(1) value		(2) value		(3) value	
income	1.913*	(0.0385)	2.814***	(0.0000)	3.639***	(0.0000)
school_score	-2012.5**	(0.0036)	-752.7**	(0.0057)	-750.7	(0.0552)
cta	35901.5*	(0.0122)	-26727.1	(0.4129)	34888.0	(0.2921)
income*school_score	0.0287*	(0.0462)				
income*cta			1.397*	(0.0330)		
school_score*cta					24.92	(0.9634)
constant	100658.3*	(0.0186)	65021.5*	(0.0224)	27993.9	(0.3123)
N	385		385		385	
R^2	0.5528		0.5615		0.5443	

p-values in parentheses

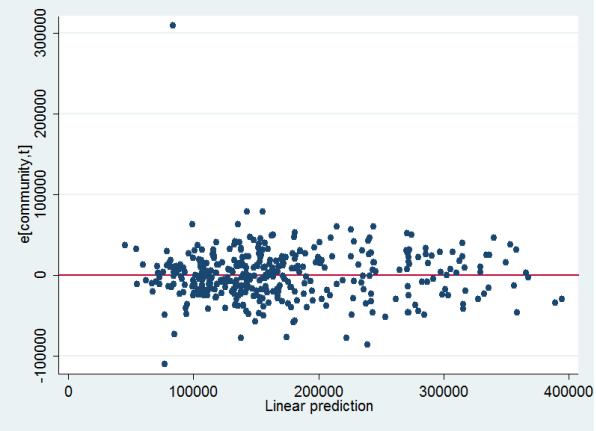
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Regression output of random effect models with single interaction term between two factors

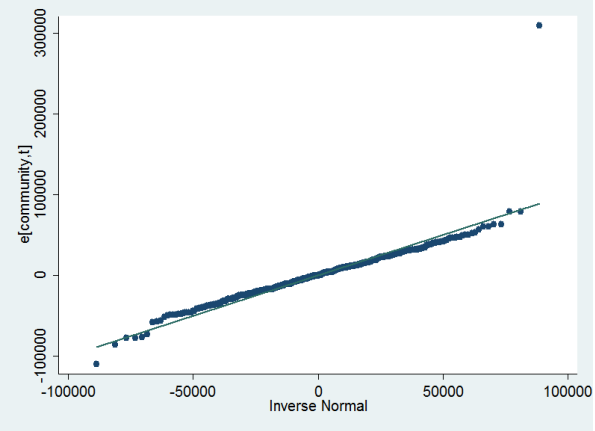
3.6 Model Robustness: Panel Data Analysis

1. Homoscedasticity of errors
In Figure 7, the residual-fitted plot does not display any heteroscedasticity.

2. Normality of residuals
In Figure 7, we see that the errors are normally distributed as the residuals are distributed linearly in the Q-Q plot.



(a) Residual-Fitted Plot



(b) Q-Q Plot of residuals

Figure 7: Homoscedasticity and Normality

3. Linearity
Linearity between independent variables and dependent variable is independent of the regression model we select. We have proven linear relationship between pairs of independent variable and dependent variable in Figure 4.

4 Discussion

We discuss the interpretation for the most robust random effects model. For an arbitrary neighborhood, a 1 dollar increase in median yearly household income will lead to a 1.913 dollar increase in the median housing price. If the neighborhood has a cta station, the median housing price in the neighborhood will increase by 35,901 dollars compared to a neighborhood with the same level of other socioeconomic factors. For a default neighborhood with 0 crime rate, 0 median household income and 0 median school score will have median housing price of 100,658 dollars. It is interesting to see a negative coefficient for the variable `school_score`. For an additional 1 point increase in the average transcript score for public high school students, the median home price will decrease by 2,012 dollars. Such a negative coefficient is balanced by the positive interaction term between income and school score. The greater the yearly median income of residents, the greater the effect of school score will have on increasing the median home prices. In other words, as a household has a higher yearly income, there would be more demand for a family to relocate to a community that has better quality of public education. This model explains 55.28% of the total variability in the median home prices in different neighborhoods per each year.

5 Extensions

5.1 Literature Review and Linkages

Islam (2012) suggests that urban residential house prices depend on intangible factors, especially neighborhood characteristics. He performed regression analyses using average housing prices in Edmonton, Alberta, Canada as the dependent variable, against independent variables including the crime rate, adjacency to ravines, household income, and population density. His pooled OLS model has shown that crime rate is a negligible factor for determining housing prices, and adjacency to ravines and household income are statistically significant. This study confirms his claim that median household income in the neighborhood affects housing prices significantly, but our own pooled OLS model (with autocorrelation issues) shows that crime rate affect median home prices significantly in Chicago.

5.2 Limitations

First, since the income data is generated by linear interpolation, the actual median income may be different from the actual data. Therefore, there exists a possibility that interaction effect between income and school score will be statistically insignificant if we use the actual data. Moreover, if we use the model to predict the median home prices of neighborhoods outside Chicago, our statistical model may suffer from overfitting issues as the median income, crime rate, and home prices of the data are extremely polarized between the rich downtown area and the far south and southwest areas that suffer from extreme poverty.

5.3 Further Studies

The only data that were publicly available were the census data and the data from the Chicago Data Portal. The model could have better predictive power if we could specify the number of commuters from each CTA L stations, and include the number of Metra & bus passengers as a holistic variable "accessibility to transportation means." More variables such as unemployment rate, final level of education, or the fraction of residents of working age could be introduced. If a user of this model wishes to predict a specific housing value, he/she can modify my web-scraping code for extracting median price value so that one can extract median price per square feet and multiply the value by the size of the residential property.

6 Stata Code

```
insheet using "\matrix.csv"
corr school_score crime income cta *correlation between factors

eststo: quietly regress value school_score crime income cta *pooled OLS
eststo: quietly stepwise, pr(.2): regress value school_score crime income cta *stepwise regression
esttab, p(4) r2(4) wide, using example.tex


*** view interaction effect between income and crime
quietly regress value income crime c.income#c.crime cta
quietly margins, at(income=(40000 110000) crime=(5000 12000 17000)) vsquish
marginsplot

***view interaction effect between cta and income
quietly margins, at(income=(20000(30000)110000) cta=(0 1)) vsquish
marginsplot

***view interaction effect between cta and crime
quietly margins, at(crime=(50000 20000) cta=(0 1)) vsquish
marginsplot

***Regression output of random effect models with single interaction term between two factors
est clear
eststo: regress value income crime c.income#c.crime cta
eststo: regress value income crime c.income#c.cta cta
eststo: regress value income crime c.crime#c.cta cta

***check for heteroscedascity
predict resid, residuals
predict yhat
graph twoway scatter resid yhat, yline(0)

*** Diagnose heteroscedasticity after log transformation of the dependent variable
gen log_value = log(value)
quietly regress log_value income crime cta
drop resid
drop yhat
predict resid, residuals
predict yhat
graph twoway scatter resid yhat, yline(0)

***check for outlier and linearity
predict di, cook
graph twoway scatter di community
summarize(di)
drop if di >= 4/385
est clear
eststo: quietly regress log_value income crime cta
esttab, p(4) r2(4) wide, using example1.tex
graph matrix log_value crime income cta

***check for normality of residuals
qnorm resid
```

```
***check for autocorrelation
graph twoway scatter resid date

clear all
insheet using "\matrix.csv"

***declare data to be panel data (panel = community, time = date)
tsset community date

***compare final models using fixed/random effects by hausman test
* I skip the manual stepwise regression(forward selection) in the do file.
gen income_school = income*school_score * interaction effect between school and income
quietly xtreg value crime school_score, fe
estimates store fixed
quietly xtreg value income school_score cta income_school, re
est clear
est store intonly
est restore intonly
estimates store random
hausman random fixed

***view interaction effect between cta and income
quietly xtreg value income school_score c.income#c.cta, re
quietly margins, at(income=(20000(30000)110000) cta=(0 1)) vsquish
marginsplot

***view interaction effect between cta and crime
quietly xtreg value income school_score c.school_score#c.cta, re
quietly margins, at(school_score=(40 65 90) cta=(0 1)) vsquish
marginsplot

*** view interaction effect between income and school_score
quietly xtreg value income school_score c.income#c.school_score, re
quietly margins, at(income=(40000 110000) school_score=(40 65 90)) vsquish
marginsplot

***Regression output of random effect models with single interaction term between two factors
est clear
eststo: quietly xtreg value income school_score cta c.income#c.school_score, re
eststo: quietly xtreg value income school_score cta c.income#c.cta, re
eststo: quietly xtreg value income school_score cta c.school_score#c.cta, re
esttab, p(4) r2(4) wide, using example4.tex

***check for heteroscedascity
drop yhat
drop resid
predict yhat
predict resid, e
graph twoway scatter resid yhat, yline(0)

***check for normality of residuals
qnorm resid

***check for autocorrelation
graph twoway scatter resid date
```

7 References

Shahidul Islam, 2012. Impact of neighborhood characteristics on Housing Prices. ASBBS 19 (1), 443-451