Modeling Chicago Home Prices with Socioeconomic Variables

Sangwoo Park

February 2017

1 Abstract

Besides the characteristics of the dwelling unit itself, urban residential housing prices depend on neighborhood characteristics such as crime cate, the median household income of the neighborhood, transportation accessibility, quality of education and etc. This study estimates the impact of neighborhood characteristics on housing prices in the city of Chicago. By using data of home prices and socioeconomic factors of neighborhoods from 2012 to 2016, various regression models using OLS, and panel regression using GLS were executed. The results of the random effects model for panel data, which has the best predictability and satisfies all the statistical assumptions, show that crime rate negatively effect the median housing prices, and that having a cta station, and having high median yearly income positively effect median housing prices of the community. There exists a high interaction effect between household income and quality of public education toward the value of home prices.

2 Discussion

2.1 Introduction

This project involves predicting 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 is overvalued or undervalued for the neighborhood's socioeconomic characteristics and preferences, i.e. crime rate, median household income, transportation accessibility, quality of education, etc. My hypothesis is that as lower the crime rate is, and as higher the median income of the residents in the neighborhood is, as higher the quality of the education is, the higher the median housing prices in the neighborhood will be. Having a cta station will also increase the median price in the neighborhood.

2.2 Data

I have preprocessed the data described below with Edward B. Hayes. All the python codes used are in my github: `https://github.com/spark01217/project/`

All the data described below has a ratio-interval level of measurement. Although `cta` variable is a 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. (Variable `community`, from 1 = Rogers Park to 77 = Edgewater)

1. Median Home Price (`value`)
The data for home prices in 77 Chicago community areas are scraped from 'Trulia,' an online residential real estate site. I have used BeautifulSoup4, a Python library for extracting data out of HTML. I calculated 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 I imputed the missing values with median home price sold in Riverdale in 2011. The unit of the variable `value` is in 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 key is community area code (1-77) and 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`)
Variable `crime` is the number of total occurrences of violent crime per 100000 residents in a year. Violent crime consists of murder, assault and burglery. The raw data is obtained from Chicago Data Portal, and I have cleaned the data by using Pandas library in Python.

4. Quality of Public Education (`school_score`)
Variable `school_score` is the mean score of total progress reports of all public high schools in a neighborhood in specific year. The data is also from Chicago Data Portal. I have used a python library 'chicago-neighborhood-finder' by data scientist 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`)

Variable `income` is the mean yearly household income in a community. The raw data is from Chicago Department of Planning and Development. Since only data for year 2010 to 2012 were available, I imputed the data for 2013 to 2016 by using linear interpolation. The variable unit is in dollars.

3 Analysis

3.1 Correlation between factors

The correlation between cta and the rest of the independent variables are not statistically significant. However, there exists correlation between `crime` and `school_score`, and between `income` and `crime`. The correlation between `income` and `school_score` is 0.552, which is so high (>0.5) that we might consider removing either `income` or `school_score` variable from our model.

```
             | 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
```

Figure 1: Correlation between factors

3.2 Model Building - Pooled OLS

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 inclusion of each variable is statistically significant. According to the regression output in Figure 2, Our model is as following:

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

where i stands for community code number.

As seen in the correlation matrix in Figure 1, we find school score variable unsuitable for our prediction model as 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

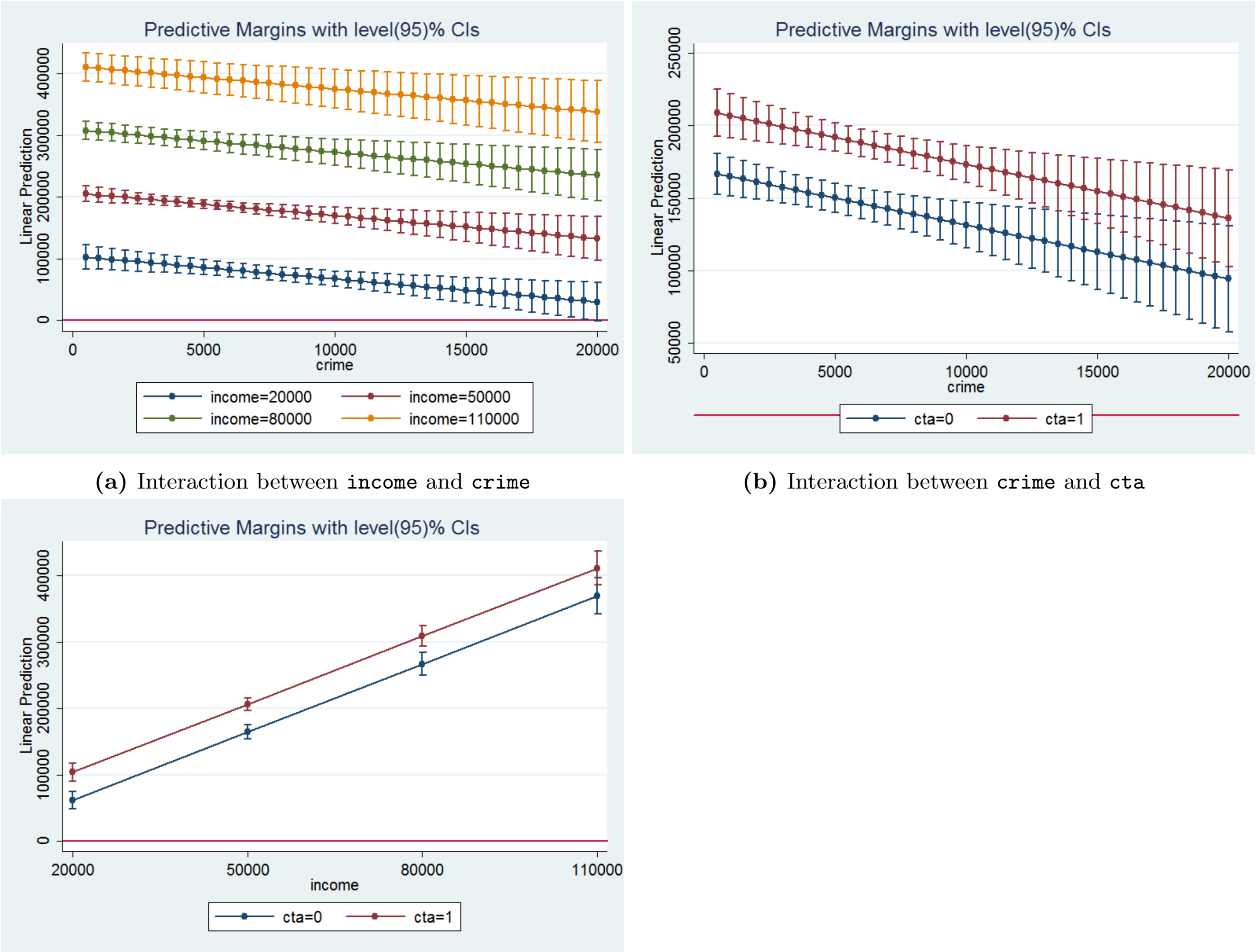
$$\texttt{value}_i = 3.337 * \texttt{income}_i + 348.790 * \texttt{school\_score}_i - 3.442 * \texttt{crime}_i + 41671.96 * \texttt{cta}_i - 2838.006 + \texttt{e}_i \tag{2}$$

```
-----------------------------------------------------------------
                      (1)                       (2)
                    value                      value
-----------------------------------------------------------------
income              3.337*** (15.49)       3.422*** (17.71)

school_score        348.8    (0.89)

crime              -3.442**  (-2.81)      -3.719**  (-3.15)

cta               41672.0*** (5.75)      41852.1*** (5.77)

constant          -2838.0    (-0.12)     13971.6    (1.05)
-----------------------------------------------------------------
N                     385                       385
-----------------------------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Figure 2: Comparison of two OLS Model Outputs

3.3 Interaction Effect

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 is different at different values of the other independent variable. We plot multiple regression lines on one graph, holding one independent variable of the pair constant. We do not see a significant difference in the slope by different value of 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.



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

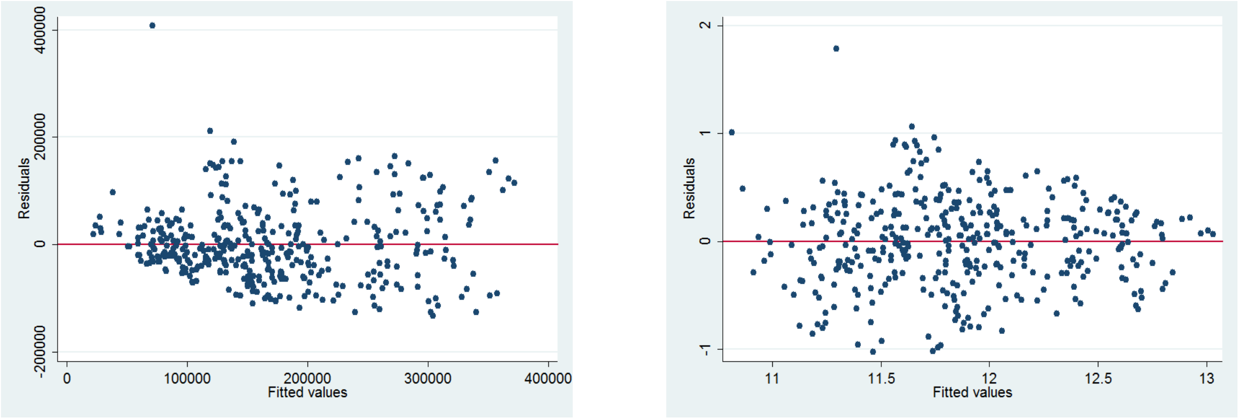
(b) Interaction between `crime` and `cta`

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

Figure 3: Plots of interaction effects between independent variables

3.4 Model Assumptions

1. Homoscedasticity of errors
In order to test homoscedasticity of the errors, we plot residuals against the fitted value to see whether the error terms are the same across all values of independent variables that predicts our dependent variable. In the left plot of figure 4, we detect heteroscedasticity as we see that there exists a pattern in the residual-fitted plot such that the residuals grow as the fitted value increases. We fix heteroscedasticity by the log-transformation of our dependent variable `value`. The residual-fitted plot of our new model has randomly distributed residuals across the entire range of fitted values in right plot of figure 4.



(a) Residual Plot before Log Transformation of `value` **(b)** Residual Plot after Log Transformation of `value`

Figure 4: Residual Plots

Then, our new model becomes

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

```
      Source |       SS       df       MS              Number of obs =     385
-------------+------------------------------           F(  3,   381) =  169.76
       Model |  86.3944428     3  28.7981476           Prob > F      =  0.0000
    Residual |  64.6332523   381  .169641082           R-squared     =  0.5720
-------------+------------------------------           Adj R-squared =  0.5687
       Total |  151.027695   384  .393301289           Root MSE      =  .41188

------------------------------------------------------------------------------
   log_value |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0000188   1.15e-06    16.30   0.000     .0000165    .0000211
       crime |  -.0000424   7.06e-06    -6.01   0.000    -.0000563   -.0000286
         cta |   .2270081   .0433045     5.24   0.000     .1418624    .3121539
       _cons |    11.1215    .079422   140.03   0.000     10.96534    11.27766
------------------------------------------------------------------------------
```

Figure 5: Regression Output for Logarithmic Transformations

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 displays linear distribution of the points, we can confirm that our residuals are normally distributed.

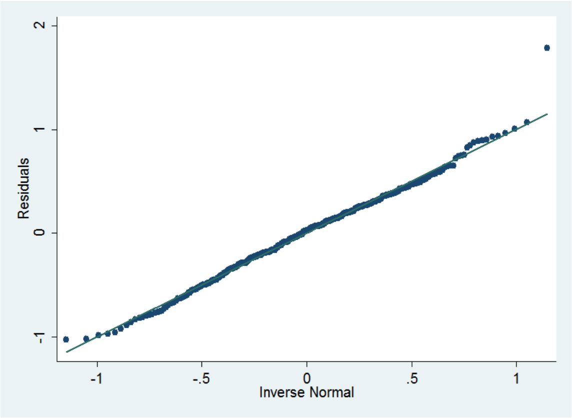
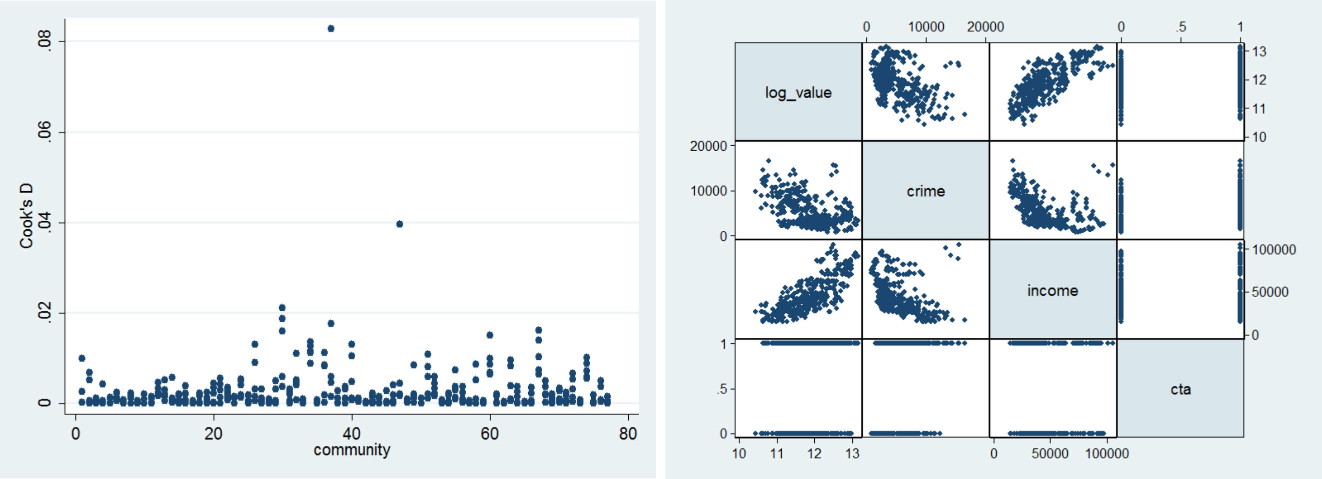


Figure 6: Q-Q Plot of Residuals

3. Linearity and Outliers
In order to detect any observations that negatively affect our regression model, I diagnose the cook's distance plot, which is a combination of each observation's leverage and residual. As the rule of thumb, I dropped all the observations that had cook's distance greater than 4/N (number of total observations). Therefore I dropped 19 observations. Then, our new model becomes

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

As we see the scattersplot between our dependent variable and our independent variables, we fail to see any nonlinear patterns.



(a) Cook's Distance of all observations **(b)** Matrix Scattersplot of All Factors in Model

Figure 7: Linearity and Outlier

4. Autocorrelation
However, the model we have used insofar violates the independence of errors. If we plot the residuals of our OLS model versus date, we see that our residuals fluctuates over time period, which may suggest that there is an omitted variable that is time-variant and affects 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 change in interest rate significantly affects home prices. If interest rate increases, then increase in amortization of mortgage decreases potential home buyer's purchasing power, lowering demand of the homes being sold.

Therefore, instead of using OLS on the pooled data, I should treat the data as a panel data in order to control for the differences in variance in our independent variables over time, and use different model than OLS in order to account for serial or panel correlation.

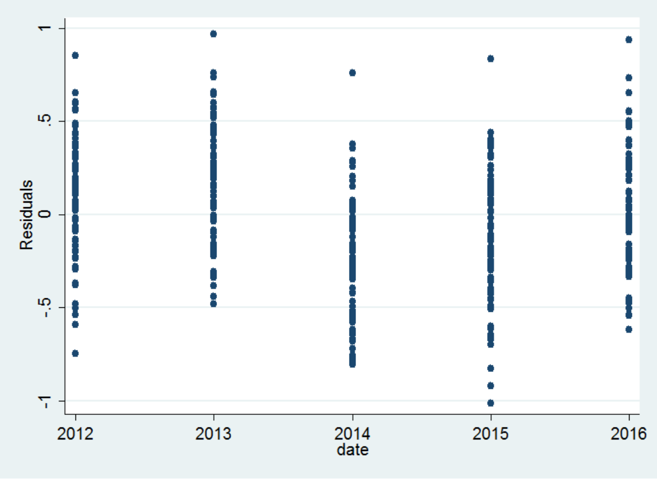


Figure 8: Scattersplot of residuals versus year

3.5 Panel Data Analysis

I declare that my data is a panel data of which panel is `community`, which represents 77 different neighborhood s in Chicago, and time is `date`, which represents year from 2012 to 2016. Among many regression models that are used in time series analysis, there are two unobserved effects model; fixed effects and random effects model. While the fixed effects model controls for the the effects of time invariant variables with time-invariant effects, random effects model lets us estimate effects of time-invariant variables by assuming all unobserved factors are uncorrelated with the observed independent variables. Since potential unobserved factors that affects 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 interest rate as mentioned before, using random effects model is logically correct.

I also execute Durbin-Wu-Hausman Test which can compare consistency of coefficients of each variables under fixed and random effects model. First, I 'manually' execute stepwise regression to build models to compare. I use regular (non-logarithmic) housing price as my dependent variable. I will employ forward selection method by which I start with an empty model, repeating the process of adding one variable of which coefficient will be most statistically significant. The threshold for including a variable will be a p-value of 0.1. Then, I add interaction effects by same algorithm. I am left with `income`, `school_score`, `cta`, and the interaction effect of `income` and `school_score` for my final random effects model. I repeat the same process using fixed effects model. I am left with `school_score` and `crime`. The p-value of 0.0372 in Hausman test states that we should choose the random effects model.

```
Random-effects GLS regression                   Number of obs      =       385
Group variable: community                       Number of groups   =        77

R-sq:  within  = 0.0601                         Obs per group: min =         5
       between = 0.6058                                        avg =       5.0
       overall = 0.5528                                        max =         5

                                                Wald chi2(4)       =    133.75
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       value |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |    1.91317   .9243271     2.07   0.038     .1015224    3.724818
school_score |  -2012.523   691.6266    -2.91   0.004    -3368.086   -656.9597
income_school|   .0287026   .0143979     1.99   0.046     .0004833    .0569219
         cta |   35901.52   14332.65     2.50   0.012     7810.044       63993
       _cons |   100658.3   42788.55     2.35   0.019     16794.31    184522.3
```

```
--------------+----------------------------------------------------------------
      sigma_u |  59719.276
      sigma_e |  34757.905
          rho |  .74696597   (fraction of variance due to u_i)
-------------------------------------------------------------------------------
```

Therefore, our final model is a random-effects model:

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

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

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

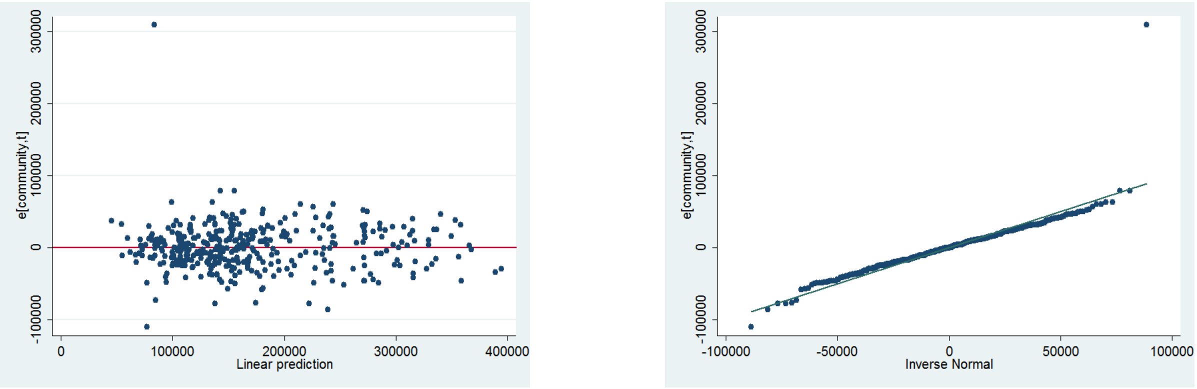
3.6 Model Interpretation

For an arbitrary neighborhood, 1 dollar increase in median yearly household income will lead to 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 35901 dollar compared to a neighborhood with the same level of other socio-economic factors. For a default neighborhood with 0 crime rate, 0 median household income and 0 median school score will have median housing price of 100658.3 dollars. It is interesting to see a negative coefficient for the variable `school_score`. For additional 1 point increase in the average transcript score for public high school students, the median home price will decrease by 2012 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.

3.7 New Model Assumptions

1. Homoscedasticity of errors
In figure 9, the residual-fitted plot does not display any heteroscedasticity.

2. Normality of residuals
According to the linear residuals in Q-Q plot, the errors are normally distributed.



(a) Residual-Fitted Plot

(b) Q-Q Plot of residuals

Figure 9: 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 previous matrix scattersplot.

4 Extensions

4.1 Limitations

First, since the income data is generated by linear interpolation, the actual median income data 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.

4.2 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 prediction 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, percent of residents with working age, etc. could be introduced. If the user of the model wants 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.

5 Stata Code

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

regress value school_score crime income cta *pooled OLS
stepwise, pr(.2): regress value school_score income crime cta *stepwise regression

*** view interaction effect between crime and income
regress value crime income cta
est store intonly
est restore intonly
***import data
margins, at(crime=(500(500) 20000) income=(20000(30000)110000)) vsquish
marginsplot, yline(0)

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

***view interaction effect between cta and crime
margins, at(cta=(0 1) crime=(0(5000)20000)) vsquish
marginsplot, yline(0)

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

```
*** Since we observe a pattern in the scatterplot of residuals versus fitted values,
* We execute a log transformation to the dependent variable and check for heteroscedascity again.
gen log_value = log(value)
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
regress log_value income crime cta
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
xtreg value crime school_score, fe
estimates store fixed
xtreg value income school_score cta income_school, re
estimates store random
hausman random fixed

***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
```