International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 711-718 www.ijaem.net ISSN: 2395-5252

Application of Principal Component Analysis Method in Evaluating the Factors Affecting Housing Values

Tran Thi Mai, Do Thanh Phuc, Pham Thi Linh
Thai Nguyen University of Economics and Business Administration

Submitted: 15-01-2022 Revised: 23-01-2022 Accepted: 25-01-2022

ABSTRACT: Based on the idea of application of principal component analysis (PCA), the factors affecting the value of houses in a certain geographic area are always concerned by real estate investors, thereby helping them make decisions to invest or not to invest in the field. In this paper, the authors analyze and explore the factors affecting housing value in Boston suburbs in the United States with the help of principal component analysis (PCA).

Keywords: affect; Principal Component Analysis, linear regression model, multivariable,random error.

I. INTRODUCTION

Boston is the metropolis and largest city in the Massachusetts area in the United States. Boston is the largest city in the New England region, with an area of 124 square kilometers and an estimated population of 745,966 in 2019, making it the 24th most populous city in the United States.

The city was the site of several major events during the American Revolution. After the United States gained independence from the British Empire, the city continued to be an important harbor and manufacturing center, as well as a center of education and culture. Boston's rich history makes it attractive to many visitors, so the value of homes here has a strong appeal. Factors affecting housing value in Boston are also of interest to researchers.

Principal Component Analysis (PCA) is a multivariable statistical method widely used by researchers, based on the idea of reducing the number of data dimensions from a multidimensional dataset, but preserving as much information as possible. Assuming the initial study has X_i observed variables in which there is multicollinearity among the X_i variables, through the analysis method PCA will convert into a set of main components PCk with k<<i. On the new variable set PC_k retains the most information of the original data set. The larger the variance, the PC_k principal component, the greater its role in the evaluation of the composite, that is, the greater the amount of information expressed. If the main component PC_i is not enough to represent well the amount of information of the original data set, then continue to choose the main component PC_i, and so on until k principal components (Lin HawiMing, 2013).

After the initial set of observed variables was reduced to a new set of variables PC_1, ...,PC_i, the authors used OLS regression method to analyze the factors affecting the value of houses in Boston suburb.

The principle of the OLS regression method is to minimize the residual variation in the regression. It can be illustrated that when representing observations in the Oxy plane, the OLS regression line is a line that passes through a crowd of data points where the distance from the data points (absolute value of ε) to the line regression is the shortest.

With the total error denoted e, while in the research sample the error is now called residual and denoted ε. The residual variation is calculated as the sum of the squares of all the residuals.

In statistics, the problem we want to evaluate is the information of the population. However, because the population is so large, we cannot get this information. Therefore, we use sample information to estimate or test population information. Likewise with linear regression, the overall regression coefficients such as β_1, β_2 ... or the regression constant β_0 are

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 711-718 www.ijaem.net ISSN: 2395-5252

the parameters we want to know but cannot measure. Therefore, we will use the corresponding parameter from the sample to estimate and then infer the population. The regression equation on the study sample:

$Y= a_0 + a_1X_1 +a_2 X_2 +...+a_kX_k + \varepsilon$
where:
Y: independent variable (continuous variable)
$X_1, X_2, ...,X_k$: dependent variables
a_0: intercept factor
$a_1, a_2, ..., a_k$: regression coefficients of the variables X
ε: noise factor
Regression coefficients $a_0, a_1, a_2, ...,a_k$estimated by the method OLS.
In this article, the author uses data Housing Values in Suburbs of Boston study the influence of factors onHousing Values in Suburbs of Bostonthrough a multivariable linear regression mode.

II. THE METHODOLOGY

2.1. Research steps

The authors use a dataset of 506 observations with 14 variables to study the value of housing in Boston, USA. To analyze this dataset, the author performs the following steps:
Step 1: Use descriptive statistics of the dataset
Step 2: Conduct data cleaning
Step 3: Use principal component analysis to reduce the set of variables
Step 4: Build and test the model
Step 5: Give conclusions.

Methods used by the author in the study: descriptive statistics; factor analysis, regression analysis and model testing to give weights the impact of the independent variables on the dependent variable. In this paper, the author uses R 4.0.5 software to explore and process datasets.

2.2.Datasets

The study uses data frame has 506 observations and 14 variables about Housing Values in Suburbs of Boston.
The code table for the observed variables is described in Table 1 with 13 independent variables and 1 dependent variable (medv).

Table 1: Code sheet for predictors in the data

	Variable	Description
1	crim	per capita crime rate by town.
2	zn	proportion of residential land zoned for lots over 25,000 sq.ft.
3	indus	proportion of non-retail business acres per town.
4	chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
5	nox	nitrogen oxides concentration (parts per 10 million).
6	rm	average number of rooms per dwelling.
7	age	proportion of owner-occupied units built prior to 1940.
8	dis	weighted mean of distances to five Boston employment centres.
9	rad	index of accessibility to radial highways.
10	tax	full-value property-tax rate per \$10,000.
11	ptratio	pupil-teacher ratio by town.
12	black	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
13	lstat	lower status of the population (percent).
14	medv	median value of owner-occupied homes in \$1000s.

III. RESEARCH RESULTS

3.1.Catergorical variable anaysis

Table 2 gives readers an overview of the data set that the authors use for analysis. The data set includes 14 continuous variables in which: medv is the dependent variable and 13 independent variables.

Table 2: Descriptive statistics for data set variables

crim	zn	indus	black	lstat
Min. : 0.00632	Min. : 0.00	Min. : 0.46	Min. : 0.32	Min. : 1.73
1st Qu.: 0.08205	1st Qu.: 0.00	1st Qu.: 5.19	1st Qu.:375.	1st Qu.: 6.95

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 711-718 www.ijaem.net ISSN: 2395-5252

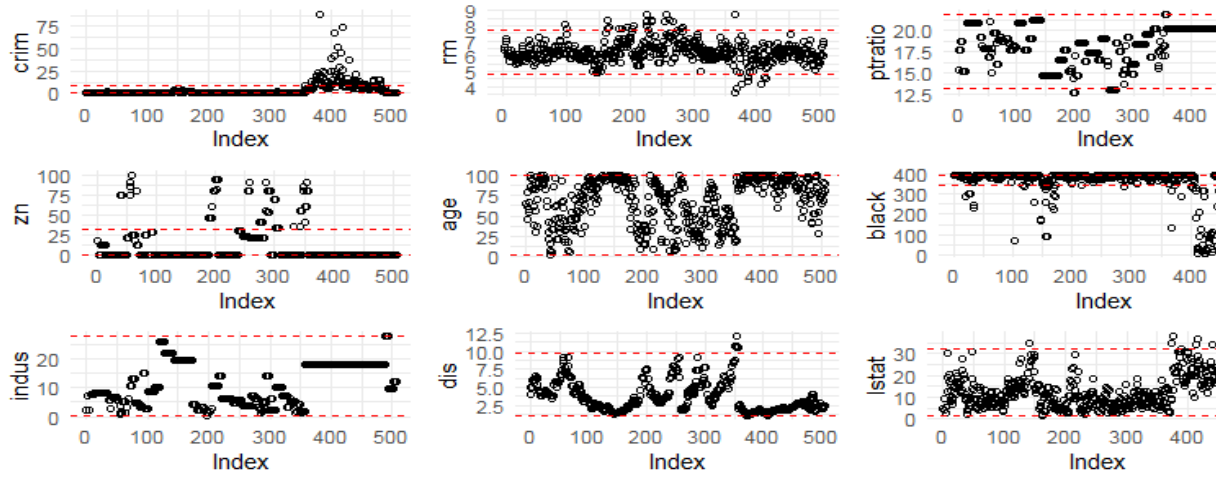
```
Median : 0.25651   Median :  0.00   Median : 9.69  Median :391.44    Median :11.
Mean   : 3.61352   Mean   : 11.36   Mean   :11.14  Mean   :356.67    Mean   :12.65
3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10  3rd Qu.:396.23    3rd Qu.:16.95
Max.   :88.97620   Max.   :100.00   Max.   :27.74  Max.   :396.90    Max.   :37.97
chas        nox              rm            age           medv
 Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 5.00
 1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02    1st Qu.:17.02
 Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50    Median :21.20
 Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57    Mean   :22.53
 3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08    3rd Qu.:25.00
 Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00    Max.   :50.00
 dis              rad              tax           ptratio
Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60
1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40
Median : 3.207   Median : 5.000   Median :330.0   Median :19.05
Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46
3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20
Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00
```

The data set is described by basic parameters such as: Min; Max ; 1st Qu ; Median; Mean; 3rd Qu. Next, the authors clean up the dataset for analysis.

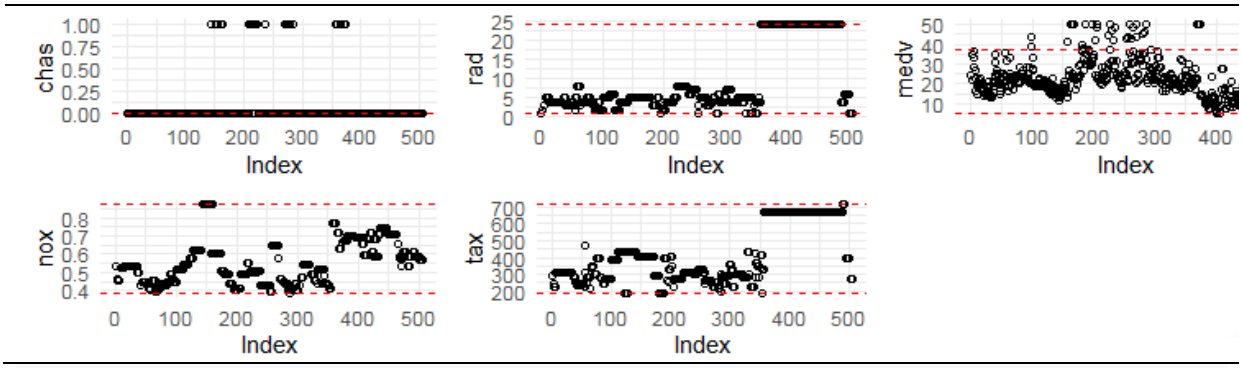
2.3 Data cleaning

The authors use scatter plot to illustrate the data set, the results are shown in Figure 1 for 14 observed variables of the raw data set.

Figure 1: Scatter plot



International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 711-718 www.ijaem.net ISSN: 2395-5252



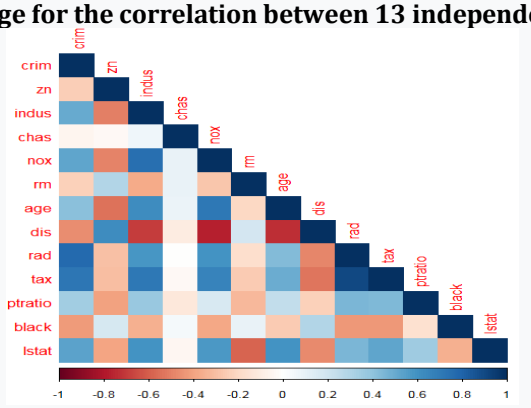
Source: Calculation on software R

The variables included in the analysis are all variables and continuous, so the author uses a scatter plot to observe outliers in the data set. An observation x is said to be outlier in the data set if $x>Q_3+ 1.5. IQR$ or $x>Q_1 - 1.5. IQR$, where Q_1, Q_3 are the 1st and 3rd quartiles of the So the cleaned data set of 484 observations was included for principal component analysis and regression analysis.

Next, the author analyzes the correlation between 13 independent variables (except the dependent variable medv). The results are illustrated in Figure 2. Dark orange shows sample, IQR =$Q_3 - Q_1$. Figure 1 is a scatter plot of 14 continuous variables, however, based on the characteristic properties of some variables, the author has re-selected the threshold for outliers such as zn, black, and medv, resulting in 22 rejected observations. from the data set.

variables with strong negative correlation while dark green represents variables with strong positive correlation. Figure 2 shows that there are many strong correlations between the independent variables, which are the basis for the author to use principal component analysis (PCA).

Figure 2: Image for the correlation between 13 independent variables



Source: Calculation on software R

3.2.The principal component method

Here, the author uses the principal component method for 13 independent variables (except for the dependent variable medv.

From the cleaned data set of 484 observations and 13 independent variables, it is replaced by 5 main components Comp.1; Comp.2; Comp.3; Comp.4; Comp.5 has no correlation. Due to the characteristics of the study, the authors used 5 main components and these 5 main components contributed to explain 81.84% of the information of the original data set, the results are shown in Table 3.

Table 3: Main components of PCA

Importance of components:					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	2.4422	1.2674	1.1633	0.93405	0.91819

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 711-718 www.ijaem.net ISSN: 2395-5252

Proportion of Variance	0.4588	0.1236	0.1041	0.06711	0.06485
Cumulative Proportion	0.4588	0.5823	0.6864	0.75355	0.81840
Eigenvalue	5.9640	1.6062	1.3533	0.8724	0.8430
Cu.vari.percent	45.877	58.233	68.644	75.355	81.840

Source: Calculation on software R

The main component upload coefficients of the observed variables are shown in Table 4

Table 4: Table of load factors of main components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
crim	0.298723158	-0.31512674	0.197224917	-0.014777073	0.09349482
zn	-0.246671098	-0.32925919	0.286588526	-0.293621432	0.26625260
indus	0.342537415	0.11264982	-0.015021432	-0.004255634	0.03007065
chas	0.006144749	0.43258070	0.324954814	0.258592462	0.76601652
nox	0.340578918	0.22295172	0.132192539	-0.191216757	-0.06008678
rm	-0.172660408	0.08369249	0.613237902	0.268227570	-0.43245394
age	0.306257442	0.33471806	-0.001767919	-0.071755415	-0.16749770
dis	-0.318910411	-0.34708979	-0.076337159	-0.026297555	0.23612073
rad	0.320684501	-0.31156365	0.267687362	0.216720066	0.06100869
tax	0.340664224	-0.26113248	0.202224414	0.124660216	0.06088517
ptratio	0.201267951	-0.28750782	-0.340481536	0.630245893	0.01084701
black	-0.200581352	0.23310235	-0.255816985	0.399184996	0.06351062
lstat	0.302676047	-0.03144103	-0.282596447	-0.336546191	0.21602566

Source: Calculation on software R

Looking at Table 4, we have

Comp.1= 0,3.crim – 0,25. zn +0,34. indus + 0,006. chas + 0,34. nox – 0,172. rm + 0,3. age – 0,32. dis +0,32. rad + 0,34. tax + 0,2. ptratio– 0,2. black + 0,3.lstat.

Main component Comp.1 has a positive relationship with variables crim; indus; chas; nox; age; rad; tax; ptratio; lstat with the respective weights: 0.3; 0.34; 0.006; 0.34; 0.3; 0.32; 0.34; 0.2 and 0.3 and at the same time has a negative relationship with the variables: zn; rm; dis; black with the respective weights: -0.25; -1.72; -0.32; -0.2.

Comp.2= -0,31.crim + 0,33. zn +0,11. indus + 0,43. chas + 0,22. nox – 0,08. rm + 0,33. age – 0,34. dis – 0,31. rad - 0,26. tax – 0,28. ptration – 0,23. black – 0,03.lstat .

Main componentComp.2 has a positive relationship with the following variables: zn;indus; chas; nox; age with corresponding weights: 0.33; 0.11; 0.43; 0.22; 0.33 and has a negative relationship with the variables: crim; rm; dis; rad; tax; ptration; black; lstat with the respective weights: -0.31; -0.88; -0.34; -0.31;-0.26; -0.28; -0.23; -0.03.

Comp.3= 0,2.crim +0,29. zn - 0,015. indus + 0,32. chas + 0,13. nox + 0,61. rm - 0,001. age – 0,07. dis +0,27. rad + 0,2. tax – 0,34. ptration – 0,25. black – 0,03.lstat.

Main componentComp.3 has a positive relationship with variables:crim; zn; chas; nox;rm; rad; tax with the respective weights: 0.2; 0.29; 0.32; 0.13; 0.61; 0.27; 0.2 and has a negative relationship with the variables: indus; age; dis; ptration; black; lstat with the respective weights: -0.015; -0.001; -0.07; -0.34;-0.25; -0.03.

Comp.4= -0,014.crim –0,29. zn - 0,04. indus + 0,26. chas -0,19. nox - 0,26. rm - 0,07. age – 0,027. dis +0,22. rad + 0,12. tax+0,63. ptration+ 0,4. black – 0,33.lstat .

Main componentComp.4 has a positive relationship with the following variables: chas; rad; tax; ptration; black with the respective weights: 0.26; 0.22; 0.12; 0.63; 0.4 also has a negative relationship with the variables: crim; zn; indus; nox; rm; age; dis; lstat with the respective weights: -0.014; -0.29; -0.04; -0.19;-0.26; -0.07; -0.027; -0.33.

Comp.5= 0,09.crim + 0,27. zn + 0,03. indus + 0,76. chas -0,06. nox - 0,43. rm - 0,16. age + 0,23. dis +0,061. rad + 0,06. tax+0,01. ptration+ 0,06. black + 0,21.lstat .

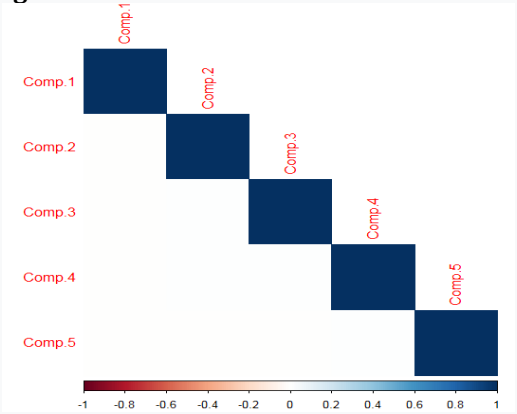
Main component Comp.5 has a positive relationship with the following variables: crim; zn; indus; chas; dis; rad; tax; ptration; lstat; black with the respective weights: 0.09; 0.27; 0.03; 0.76; 0.23; 0.061; 0.06; 0.01; 0.06; 0.21 at the same time has a negative relationship with

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 711-718 www.ijaem.net ISSN: 2395-5252

the variables: nox; rm; age with the respective weights: -0.06; -0.43; -0.16.

To analyze the impact of 13 observed variables: crim; zn; indus; chas; nox; rm; age; dis; rad; tax; ptratio; black; lstat on the dependent variable medv. Instead, the team now studies the impact of five key components: Comp.1; Comp.2; Comp.3; Comp.4; Comp.5 on the dependent variable medv. We see between these 5 main components the correlation is almost zero with the results shown in Figure 3. This is a good condition for performing multivariable linear regression.

Figure 3: Image of the correlation between the 5 main components



Source: Calculation on software R

3.3Regression analysis results

The author conducts multivariable linear regression analysis with the main components: Comp.1; Comp.2; Comp.3; Comp.4; Comp.5 and the dependent variable is medv from the open source R calculation, the results are shown in Table 5.

Table 5:Regression analysis results

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)		2.5 %	97.5 %
(Intercept)	22.64814	0.15635	144.853	<2e-16	***	22.340918	22.955363
Comp.1	-2.26241	0.06402	-35.338	<2e-16	***	-2.388209	-2.136609
Comp.2	-4.21668	0.12337	-34.180	<2e-16	***	-4.459088	-3.974276
Comp.3	1.71639	0.13440	12.771	<2e-16	***	1.452301	1.980475
Comp.4	-1.43673	0.16739	-8.583	<2e-16	***	-1.765644	-1.107815
Comp.5	2.20198	0.17028	12.931	<2e-16	***	1.867388	2.536582

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
Residual standard error: 3.44 on 478 degrees of freedom
Multiple R-squared: 0.8551, Adjusted R-squared: 0.8536
F-statistic: 564.2 on 5 and 478 DF, p-value: < 2.2e-16

Source: Calculation on software R

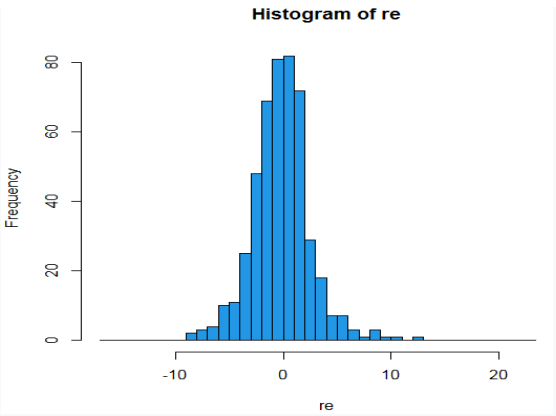
From Table 5 we have the following representation
medv = 22,64 – 2,26.Comp.1 – 4,21.Comp.2 +1,71. Comp.3 – 1,42. Comp.4 + 2,2. Comp.5
Variables :Comp.1; Comp.2; Comp.3; Comp.4; Comp.5 in the multivariable linear regression model is very statistically significant with the P-value = 2.2e -6 < 5% and explains 85% of the change of the dependent variable medv.

2.5 Model testing

Hypothesis 1: Test for the normal distribution of residuals

The author uses Historgram to check the normal distribution of residuals in the model, looking at the graph we see that the residuals have a fairly normal distribution, so hypothesis 1 is satisfied.
Figure 3:Historgram image for residuals

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 711-718 www.ijaem.net ISSN: 2395-5252



Assumption 2: The variance of the random error is constant

The author tests the following pair of hypotheses with the ncvTest function in the car .command package

H0: variance of error constant

H1: Variance of error changes

The results of the test with Chisquare = 0.01575827 and p = 0.9001>0.05. Therefore, rejecting H1 and accepting H0 means that the model satisfies hypothesis 2: The variance of random error is constant.

Assumption 3: Check for multicollinearity

Variance Inflation Factorof the variables given in Table 5

Table 6: Variance Inflation Factor (VIF)

Biến	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
VIF	1	1	1	1	1

Source: Calculation on software R

Looking at Table 6, the VIF index = 1, concluding that the model does not have multicollinearity, hypothesis 3 is satisfied.

Hypothesis 4: The expectation of random error is zero

The author tests the following pair of hypotheses, with the t-test function in the software R.

H0: Expected random error is 0

H1: Expectation of random error other than 0

Table 7: Expectation test results of random error)

```
One Sample t-test
data:  re
t = -1.659e-16, df = 483, p-value = 1
alternative hypothesis: true mean is not
equal to 0.95 percent confidence interval:
 -0.4396062  0.4396062
sample estimates:mean of x
-3.711735e-17
```

With the results in Table 7, it shows that p-value=1>5% accepts hypothesis H0 and rejects hypothesis H1. Conclusion the model satisfies hypothesis 4.

IV. CONCLUDE

The article used principal component analysis combined with OLS least squares method to show the factors affecting housing value in Boston.The variable Comp2 has the strongest negative impact on the variable medv. Specifically, when the variable Comp.2 increases

1 times, the variable medv decreases 4.21 times;Next is the influence of variables comp.1 and Comp.5 on the dependent variable medv. Specifically, when the variable Comp.1 increases 1 times, the variable medv decreases 2.26 times and when the variable Comp.5 increases 1 times, the variable medv increases 2.2 times. The variable that has the least influence on the variable medv is Comp.3, ie, when the variable Comp.3 increases 1 times, the variable medv increases 1.71 times

REFERENCE

[1]. Harrell, F. E. (2015), Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis, 2nd edition, Springer – Verlag, Cham.
[2]. Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5, 81–102.
[3]. Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.
[4]. Joseph F. Hair Jr., William C. Black, Barry J. Babin, Rolph E. Anderson (2014) Multuivariate Data Analysis.