The analysis of factors that contributes to the median housing price in Boston

MATH 564 PROJECT

PINGCHUAN MA; XIAOJIAN LIAN; CHUNLIANG ZHAO; YUENA CHEN

The analysis of factors that contributes to the median housing price in Boston

1. Problem Description

1.1 Background information:

Houses, as one of the most important investments a household might have invested, are always of pretty high prices. Thus, when we making the decisions to buy a house, we care for the its price and value. Here, we collect the median values for certain blocks in Boston and other contributes that might potentially affect the housing prices. By using these data, we try to find the factors that most important contributes to the median housing prices in the block. We hope to reveal the factors behind the housing prices so that we make housing investments, we can make better decisions.

1.2 Objective:

The main objective is to analyze the collected data in order to figure out the possible important contributes to the median housing prices in a block. Throughout the collected data, we hope to find eliminate those data that are not of great importance and focus only on the factors that will affect the median housing prices significantly. We try to answer the following question:

1. Are all the factors from the collected data important?
2. What are the factors that have great importance?
3. Will our models be comprehensive and statistically correct?

2. Data Information and Methodology

2.1 Data Description and Information

The data collected here total has 13 factors that are used in this project; they are all potential reason for the differences in median housing prices. The details of these factors will be stated below:

MEDV— Median value of the housing prices in a certain block. Note that these are all owner occupied houses instead of empty houses which might not reflect the true value. For example, empty houses without owner residing in it will not be affected by the community the house is at too much.

CRIM—The crime rate of certain block. We believe that crime rate is considered as a good indication of the housing prices. No one would buy a house that is in a block with high crime rate.

ZN—The proportion of the residential land zone that are 25000 square feet or higher. These might become one of the factor is because that larger residential land usually have bigger community size, it might also reflect the community that a specific house might be in.

INDUS—Proportion of non-retail business in town. This is a questionable source, with more non-retail business in the block, the housing prices might fall down as it might become a more business area like rather than a cozy place to set a house.

CHAS—This is a dummy variable, representing if the block is by the Charles river. As common sense, a house that is close to the river usually has a higher price as residing beside rivers usually gives a better view.

NOX—nitric oxides concentration (parts per 10 million), as a gas in the air, overdosing nitric oxide will not be good for human body. Thus this might also affect the housing prices if the block has a very high concentration of nitric oxides.

RM—Room per dwelling. This is a very important reason when we are considering houses, more rooms always are welcomed when buying a house.

AGE—Proportion of these investigated houses that are built prior to 1940. People might want newer houses for numerous reasons. Thus this might become a factor that affect the median housing prices in a designated area.

DIS— Weighted distance from the five employment center in Boston. This is an important factor as we believe closer distance to the workplace usually brings more convenience and thus results in a higher value of house.

RAD—Index of accessibility to radial highway. Being able to access the highway means easier to access the outside world and thus brings up the median values in certain block.

TAX—Full-value property-tax rate per $10,000. This might become an important reason as tax is one of the very important reason for home owner to consider when purchasing a house.

PTRATIO—Pupil to teacher ratio in the designated area. A higher PTRATIO means the education resources will be more abundant in the block. This will result in higher value of houses as parents will always consider places that have better education resources.

B—The blacks in the certain area. Meaning no offense but for many cases, blacks might not be employed or might be homeless. They might become potential risks. Thus they might be a factor that affect the housing prices in a certain area.

LSTAT—The percentage of people that have lower social status in the block. People care about their community they live in. This is a reflection of people's concern about the community when they make house purchasing decision.

2.2 Methodology

The basic idea to do this is to use multiple linear regression, a classical model that is widely used in many fields, to analysis the factors stated above in order to come out a linear relationship between the median housing values and factors we collected. Then by using t-statistic and p-value, we determine the factors that are less important. By using backward stepwise regression, we determine the simplified model. Then we consider the potential quadratic terms by looking at the regression plot; after determining which factors to become quadratic terms, we build new models and use partial F test to test the significance of the new models.

3. Analysis Process and Results

3.1 Multiple Linear Regression Result

In this section, we regret the median housing values onto all the potential factors, building a full model. And we test the significance using the t-test. The model below did not shows only the logic of this regression, i.e., only the factors considered being showed and not the coefficients.

The result as follows:

Model: MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT + Intercept

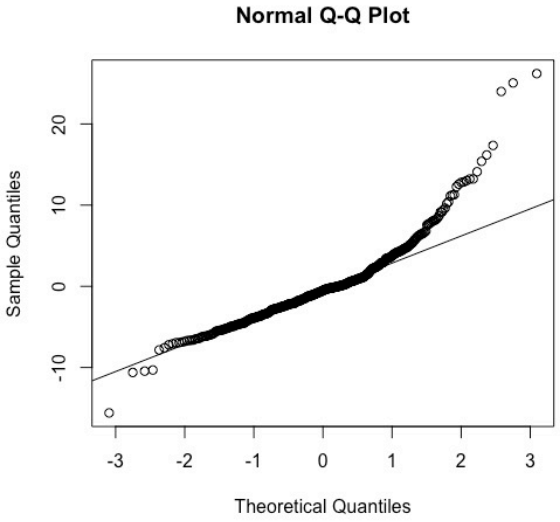
Coeff Name:	*Estimates*	*Standard Error*	*T-Statistic*	*P-Value*
Intercept	3.646e+01	5.103e+00	7.144	3.28e-12
CRIM	-1.080e-01	3.286e-02	-3.287	0.001087
ZN	4.642e-02	1.373e-02	3.382	0.000778
INDUS	2.056e-02	6.150e-02	0.334	0.738288
CHAS	2.687e+00	8.616e-01	3.118	0.001925

NOX	-1.777e+01	3.820e+00	-4.651	4.25e-06
RM	3.810e+00	4.179e-01	9.116	< 2e-16
AGE	6.922e-04	1.321e-02	0.052	0.958229
DIS	-1.476e+00	1.995e-01	-7.398	6.01e-13
RAD	3.060e-01	6.635e-02	4.613	5.07e-06
TAX	1.233e-02	3.760e-03	-3.280	0.001112
PTRATIO	-9.527e-01	1.308e-01	-7.283	1.31e-12
B	9.312e-03	2.686e-03	3.467	0.000573
LSTAT	-5.248e-01	5.072e-02	-10.347	< 2e-16

This model has an adjusted R-Square ~0.73 and a F-Stat = 108.1.

First, we perform t-test on the variables, and then use the p-value to exam its significance. We keep the variables that are significant on 0.01 level and remove the variables that is not significant, in other word, these variables have p-values greater than 0.01. Base on this fact, we found INDUS and AGE have relatively large p-values, which are 0.738288 and 0.958229 respectively. This fact suggests us to remove the IDUS and AGE to better fit the model.

Besides, within the variables that are significant on 0.01 level, we can also use P-value to show which variables are more significant than the others. We can use significant level 0.001 to separate between highly significant and slightly less significant. It turns out CRIM, CHAS and TAX are less significant than others. We will further use backward analysis to test on the variables.



3.2 QQ-Plot

From the Normal quantile-quantile plot, the model we built has an extremely fatter tail in the higher part, indicating that the model will not be very normal. Thus further discussion of this model is needed. We would always want a model that are more close to the normal distribution so that are assumptions when setting up the model will not be violated.

3.3 Backward Elimination

In statistics, stepwise regression includes regression models in which the choice of predictive variables is carried out by an automatic procedure. Stepwise model comparison is an iterative model evaluation.

As in section 3.1, We have tried to get the linear model with testing the significance of different parameters. To verify the correctness of our model, we are doing backward elimination. The backward elimination starts with all variables, iteratively removing those of low importance.

The Akaike information criterion(AIC) is considered as the most important criterion for the importance of the model. We do the backward elimination as following.

Start: AIC=1589.64

MEDV = CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT

	DF	*Sum of Sq*	*RSS*	*AIC*
-AGE	1	0.06	11079	1587.7
-INDUS	1	2.52	11081	1587.8
NONE			11079	1589.6
-CHAS	1	218.97	11298	1597.5
-TAX	1	242.26	11321	1598.6
-CRIM	1	243.22	11322	1598.6
-ZN	1	257.49	11336	1599.3
-B	1	270.63	11349	1599.8
-RAD	1	479.15	11558	1609.1
-NOX	1	487.16	11566	1609.4
-PTRATIO	1	1194.23	12273	1639.4
-DIS	1	1232.41	12311	1641.0
-RM	1	1871.32	12950	1666.6
-LSTAT	1	2410.84	13490	1687.3

Step: AIC=1587.65

MEDV = CRIM + ZN + INDUS + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LSTAT

	DF	*Sum of Sq*	*RSS*	*AIC*
-INDUS	1	2.52	11081	1585.8
NONE			11079	1587.7
-CHAS	1	219.91	11299	1595.6
-TAX	1	242.24	11321	1596.6
-CRIM	1	243.20	11322	1596.6
-ZN	1	260.32	11339	1597.4
-B	1	272.26	11351	1597.9
-RAD	1	481.09	11560	1607.2
-NOX	1	520.87	11600	1608.9
-PRTATIO	1	1200.23	12279	1637.7
-DIS	1	1352.26	12431	1643.9
-RM	1	1959.55	13038	1668.0
-LSTAT	1	2718.88	13798	1696.7

Step: AIC=1585.76

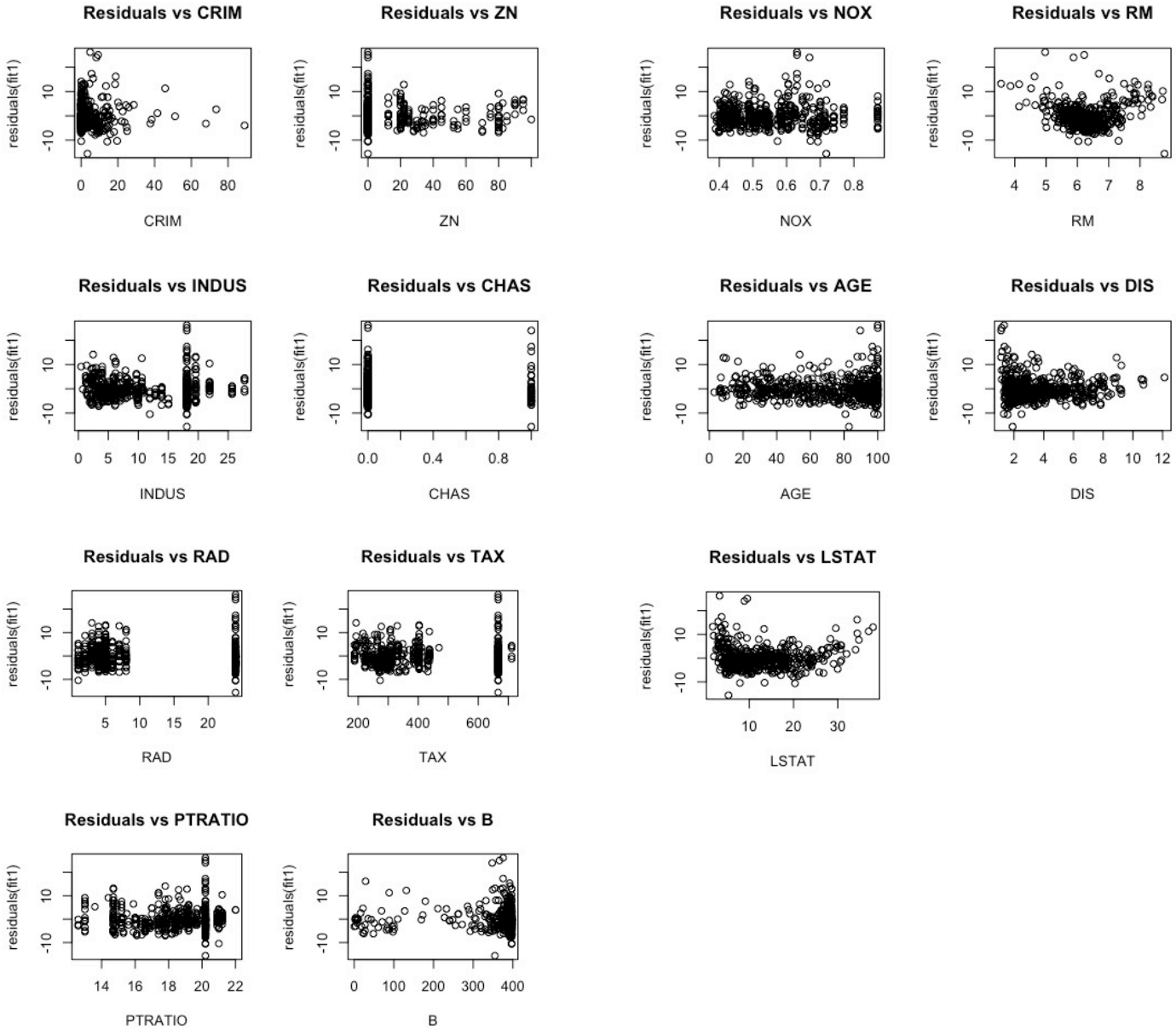
MEDV = CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LSTAT

	DF	*Sum of Sq*	*RSS*	*AIC*
None	1		11081	1585.8
-CHAS	1	227.21	11309	1594.0
-CRIM	1	245.37	11327	1594.8
-ZN	1	257.82	11339	1595.4
-B	1	270.82	11352	1596.0
-TAX	1	273.62	11355	1596.1
-RAD	1	500.92	11582	1606.1
-NOX	1	541.91	11623	1607.9
-PTRATIO	1	1206.45	12288	1636.0
-DIS	1	1448.94	12530	1645.9
-RM	1	1963.66	13045	1666.3
-LSTAT	1	2723.48	13805	1695.0

From the backward elimination process, we got the reduced model without two independent variables, AGE and INDUS. This is the model that gave the lowest AIC value step by step. The previous model that we regress has the same independent variables with this one. So we can see that for the linear part, we get an appropriate model with independent variables CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B and LSTAT.

3.4 Residual Analysis

Here we use the residual plots to determine which factors should be regarded as quadratic terms. Any residual plots that display a curvature indicate that we should take quadratic into account.



After we plot the residuals plots for each variable, we found couple interesting patterns that suggests possible quadratic relationship between the variable and medium house value. Residual plot for RM has downward trend on its left tail and upward trend on its right tail, which form an opening upward Parabola. This fact suggests RM might have quadratic term to fits this pattern. Similarly, residual plot for LSTAT also has an opening upward Parabola, with decreasing trend on its left tail and slightly increasing trend on its right tail. This pattern also suggests the existence of quadratic term for the variable LSTAT. In conclusion, RM and LSTAT need to have quadratic term.

3.5 Quadratic terms

Here we build 3 models:

1. Quadratic RM factor: MEDV ~ CRIM + ZN + CHAS + NOX + RM + I(RM^2) + DIS + RAD + TAX + PTRATIO + B + LSTAT
2. Quadratic LSTAT factor: MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LSTAT + I(LSTAT^2)
3. Adding both the quadratic terms: MEDV ~ CRIM + ZN + CHAS + NOX + RM + I(RM^2) + DIS + RAD + TAX + PTRATIO + B + LSTAT + I(LSTAT^2)

We then test the significance of the quadratic terms using partial F-test comparing this to the new model that INDUS and AGE being removed.

Test 1: Testing the RM quadratic term

RES.DF	*RSS*	*DF*	*SUM OF SQ*	*F*	*PR(>F)*
494	11081.4				
493	8429.9	1	2651.5	155.06	<2.2e^-16

The F statistic displays as 155.06 with a P-value much smaller than 0.001, indicating that this is significant even at a 1% level.

Test 2: Testing the LSTAT quadratic term

RES.DF	*RSS*	*DF*	*SUM OF SQ*	*F*	*PR(>F)*
494	11081.4				
493	9107.3	1	1974.1	106.86	<2.2e^-16

The F statistic displays as 106.86 with a P-value much smaller than 0.001, indicating that this is significant even at a 1% level.

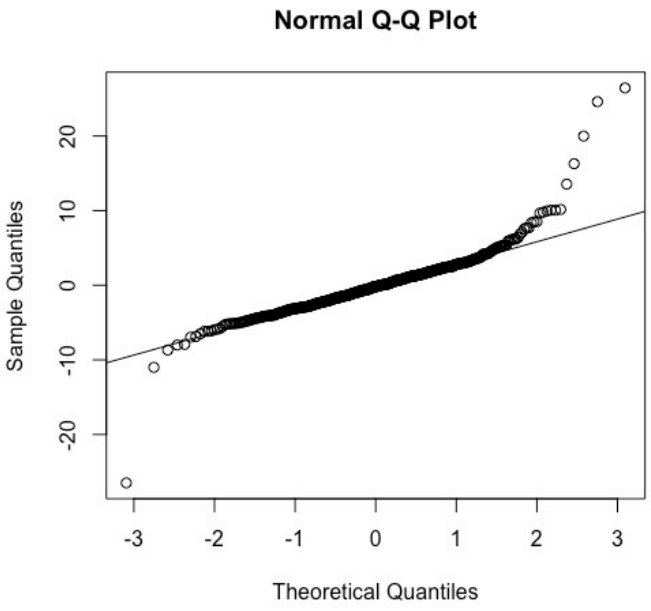
Test 3: Testing both RM and LSTAT quadratic terms being added

RES.DF	*RSS*	*DF*	*SUM OF SQ*	*F*	*PR(>F)*
494	11081.4				
493	7809.9	2	3271.4	103.05	<2.2e^-16

The F statistic displays as 103.05 with a P-value much smaller than 0.001, indicating that this is significant even at a 1% level.

By using partial F test we come to the conclusion that both the RM term and LSTAT term should be added to the model. Thus our new complete model should be:

MEDV ~ CRIM + ZN + CHAS + NOX + RM + I(RM^2) + DIS + RAD + TAX + PTRATIO + B + LSTAT + I(LSTAT^2)



3.6 QQ-Plot for the newly set model

The new model has a QQ plot that looks absolutely more normal than the original one. Aside from the outliers at the very far end, the whole plot looks much smaller tails.

4. *Conclusion*

The factors considered in this model include the property factors, like RM- how many room the house has, which can be used to evaluate the value of asset itself; environment factors, say, CRIM- criminal rates, the social environment, and CHAS- the natural environment factor, which can evaluate whether it is a suitable place to live; additional value factors, e.g. PTRATIO- the ratio of pupil and teachers, which is an important factor considered by specified people like parents looking for places that can offer better education. Thus this model can be considered as a profound model that contains most of the important factor that may influence the value of the real estate.

By a first analyzing of the model we constructed, we can see that, in general, most of the factors are statistically significant except INDUS and AGE. The age of the real estate is not a statistically significant factor is an interesting result since in China, it is indeed a very important factor that have influence on the value of a house. This suggest us that when we are looking for a house in a foreign country, we must keep in mind that there may be huge value divergence.

By further check we can find that the model with the quadratic terms fit the data better since the normal plots of the residual shows this model satisfies the assumption of the errors more strongly and that the standard error has drop from 4.7 to 3.9. Also the F-stat and the coefficient of determination has both increased. These just suggest that the rooms per doweling and the quality of community (percentage of lower status) have more significant impact on house's price. This is reasonable since more rooms means larger place to live and people are willing to pay more on the house with higher utility and that a community of higher quality usually indicates better safety, better education environment and possibly higher level of social relationship.

Although the final model has fitted the data very well, we can still have further analysis to improve this model. A main insufficient of this model is that we haven't check whether there should be any interaction variable included in this model. However, due to the length of the report and the large amount of variable and possible huge amount of interaction terms, we

think it is acceptable to not include interaction variable in this model, which already fits the data very well.