innovate achieve lead

ECONOMETRIC METHODS

BOSTON HOUSE PRICING MODEL: A LINEAR REGRESSION ANALYSIS

MARCH, 2017

ANIRUDHA K

SHRAVAN T

RISHABH MUNDADA

DIVYAM PATRO ASHOK

UTKARSH AGRAWAL

BACKGROUND OF STUDY

From many years due to the increasing real estate price bubble many individuals have lost track of the intrinsic housing prices in the market.They rely on the prices quoted by the real estate brokers which are often overvalued .It has left many of us wondering how to properly estimate the housing price values since many of the consumers are not aware of the methods to calculate these prices.In order to determine housing prices we have used multiple regression in which we have included many determinants that influence the market price of the house.The intent of this paper is to use regression analysis to determine the housing prices for a particular set of values of attributes,also to see the correlation between the attributes and the price of house.Values of the correlation between each attribute and the dependent variable(price of house) will show the major attributes(factors) which influence the housing price.

This model has been presented below using a case study of boston.Housing Prices has been one of the most prevailing problems in Great Boston, and as per the report card 2013 published by The Kitty and Michael Dukakis Center for Urban and Regional Policy Northeastern University. They have observed that despite of rising house prices and declining household income housing market recovery has took place in Boston in 2013 ever since 2008.The data for houses in Boston is used to present the results of the regression analysis.

INTRODUCTION

Predictive modeling is a statistical approach that is used to build a prediction function from the observed data. Then this function is used to estimate a value of a dependent variable for new data set or the new observed data.

A commonly used method in predictive modeling is regression analysis that has been applied to a wide range of application domains. So it is to build a model (i.e. function f) from an observed data set D such that the model will predict the outcome of a new input x as f(x) with the best probability.The domain of X is a set of independent variables, which is used to predict the dependent variable. Various methods have been developed for predictive modeling and among them,multivariate linear regression is perhaps one of the most commonly used and relatively easy to build.

Now say if there are n number of independent variables then the dependent variable y can be expressed in terms of independent variables as follows:

$y = c0 + c1x1 + \ldots + cnxn + error.$

Also if the relationship between the dependent and independent variable is non-linear then new variables for non-linear terms can be created, by replacing those with linear terms.

Say, for example

$y = c0 + c1x1 + c2z2 + c3z3 + eror$ where $z2 = x2^2$, and $z3 = lnx3$

LITERATURE REVIEW

Housing is both a consumption good and an asset for individuals. If someone owns a house then it is an asset for him and if someone pays rent for living in the house then the house is a service for them. Housing goods are differentiated (i.e house is a differentiated product) because the choice of house depends on various attributes. These attributes on which the individual bases his choice contribute in some ways to the price of the house. The dependency of the price on the various factors is determined by regression analysis after which we know the estimated dependence on the various attributes is determined, this ultimately helps in determining the future rates of the houses.
The most famous regression model used in this analysis is the ***hedonic pricing model*** which was developed by *Lancaster and Sherwin Rosen*[1].
Rosen stated that "goods are valued for their utility-bearing attributes or characteristics". Rosen's definition of hedonic prices led to the use of Hedonic pricing which is defined as a model identifying price factors according to which price is determined both by internal characteristics of the good being sold and external factors affecting it. He continued his research to show that the price of the house depends on both the internal house attributes and the external factors. In economic terms it can be seen that cost of construction is directly proportional to the internal housing attributes and the external factors determine the land price and the market price at which the house is priced in the market. Based on the above research we have compiled a list of both internal and external attributes, among them the attributes are chosen such that we measure the effect of the characteristics only on the sales price of the house. The attributes used in our regression model are mentioned in the further sections.

OBJECTIVES OF STUDY

1. *Use the hedonic multiple regression model to examine the influence of various neighborhood attributes on housing prices in Boston in an attempt to discover the most suitable explanatory variables.*

2. *Fit a linear regression model that best explains the variation in the response variable i.e. median value of owner-occupied homes in Boston (in thousands of dollars).*[2]

[1] http://www.stern.nyu.edu/networks/phdcourse/Rosen_Hedonic_prices.pdf

[2] The data as mentioned in next page is of the time period 1970-80. therefore our model would be applicable to that time that period.

DATA DESCRIPTION :

1. Title: Boston Housing Data

2. Sources:

(a) Origin: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon university.

(b) Creator: Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air',

J. Environ. Economics & Management, vol.5, 81-102, 1978.

(c) Date: July 7, 1993

(d) web source : https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

3. Number of Instances: 506

4. Number of Attributes: 13 continuous attributes (including "class" attribute "MEDV"), 1 binary-valued attribute.

5. Attribute Information:

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 : nitric 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 distances to five Boston employment centers
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. B : 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT : % lower status of the population
14. MEDV : Median value of owner-occupied homes in $1000's

6. Missing Attribute Values: None.

PRIMARY ANALYSIS OF ALL THE VARIABLES:

CRIM		*ZN*		*INDUS*		*CHAS*	
Mean	3.613523557	Mean	11.3636364	Mean	11.1367787	Mean	0.06916996
Median	0.25651	Median	0	Median	9.69	Median	0
Mode	0.01501	Mode	0	Mode	18.1	Mode	0
Standard Dev	8.601545105	Standard Dev	23.322453	Standard Dev	6.86035294	Standard Dev	0.253994041
Skewness	5.223148798	Skewness	2.22566632	Skewness	0.29502157	Skewness	3.405904172
Minimum	0.00632	Minimum	0	Minimum	0.46	Minimum	0
Maximum	88.9762	Maximum	100	Maximum	27.74	Maximum	1
NOX		*RM*		*AGE*		*DIS*	
Mean	0.554695059	Mean	6.28463439	Mean	68.5749012	Mean	3.795042688
Median	0.538	Median	6.2085	Median	77.5	Median	3.20745
Mode	0.538	Mode	5.713	Mode	100	Mode	3.4952
Standard Dev	0.115877676	Standard Dev	0.70261714	Standard Dev	28.1488614	Standard Dev	2.105710127
Skewness	0.729307923	Skewness	0.40361213	Skewness	-0.5989626	Skewness	1.011780579
Minimum	0.385	Minimum	3.561	Minimum	2.9	Minimum	1.1296
Maximum	0.871	Maximum	8.78	Maximum	100	Maximum	12.1265
RAD		*TAX*		*PTRATIO*		*B*	
Mean	9.549407115	Mean	408.237154	Mean	18.4555336	Mean	356.6740316
Median	5	Median	330	Median	19.05	Median	391.44
Mode	24	Mode	666	Mode	20.2	Mode	396.9
Standard Dev	8.707259384	Standard Dev	168.537116	Standard Dev	2.16494552	Standard Dev	91.29486438
Skewness	1.004814648	Skewness	0.66995594	Skewness	-0.8023249	Skewness	-2.890373712
Minimum	1	Minimum	187	Minimum	12.6	Minimum	0.32
Maximum	24	Maximum	711	Maximum	22	Maximum	396.9
		LSTAT		*MEDV*			
		Mean	12.6530632	Mean	22.5328063		
		Median	11.36	Median	21.2		
		Mode	8.05	Mode	50		
		Standard Dev	7.14106151	Standard Dev	9.19710409		
		Skewness	0.90646009	Skewness	1.10809841		
		Minimum	1.73	Minimum	5		
		Maximum	37.97	Maximum	50		

TABLE. 1

CORRELATION ANALYSIS BETWEEN VARIABLES:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	1													
ZN	-0.20	1												
INDUS	0.406	-0.53	1											
CHAS	-0.05	-0.04	0.062	1										
NOX	0.420	-0.51	0.763	0.091	1									
RM	-0.21	0.311	-0.39	0.091	-0.30	1								
AGE	0.352	-0.56	0.644	0.086	0.731	-0.24	1							
DIS	-0.37	0.664	-0.70	-0.09	-0.76	0.205	-0.74	1						
RAD	0.625	-0.31	0.595	-0.00	0.611	-0.20	0.456	-0.49	1					
TAX	0.582	-0.31	0.720	-0.03	0.668	-0.29	0.506	-0.53	**0.910**	1				
PTRATIO	0.289	-0.39	0.383	-0.12	0.188	-0.35	0.261	-0.23	0.464	0.460	1			
B	-0.38	0.175	-0.35	0.048	-0.38	0.128	-0.27	0.291	-0.44	-0.44	-0.17	1		
LSTAT	0.455	-0.41	0.603	-0.05	0.590	-0.61	0.602	-0.49	0.488	0.543	0.374	-0.36	1	
MEDV	-0.38	0.360	-0.48	0.175	**-0.42**	**0.695**	-0.37	0.249	-0.38	**-0.46**	**-0.50**	0.333	**-0.73**	1

Table.2

Here we see that MEDV is significantly correlated with variables NOX, RM, TAX, PTRATIO, LSTAT.

We see that it is positively correlated with variables like RM , DIS .This is explained by the fact that as average number of rooms per dwelling increases the median value of the house is expected to go up.

Also MEDV is negatively correlated with variables like NOX,CRIM,LSTAT,TAX as in places where pollution, crimes, property tax is more ,people prefer not to stay in these places hence prices are expected to decrease.

Also we see that the variables RAD and TAX are highly correlated. This may affect our regression analysis due to the problem of multicollinearity . This problem and how we have solved is discussed in further sections.

REGRESSION ANALYSIS:

STAGE 1

Here we fit the linear regression model without any modification to the data set.

Below table is the output after fitting the model.

Regression Statistics	
Multiple R	0.860605987
R Square	0.740642664
Adjusted R Square	**0.733789726**
Standard Error	4.745298182
Observations	506

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	13	31637.5108	2433.65	108.077	6.722E-135
Residual	492	11078.7846	22.5179		
Total	505	42716.2954			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	36.45948839	5.10345881	7.14407	3.3E-12	26.432226	46.48675	26.432226	46.486751
CRIM	-0.10801136	0.03286499	-3.2865	0.00109	-0.1725844	-0.04344	-0.1725844	-0.0434383
ZN	0.046420458	0.01372746	3.38158	0.00078	0.01944878	0.073392	0.0194488	0.0733921
INDUS	0.020558626	0.06149569	0.33431	0.73829	-0.1002679	0.141385	-0.1002679	0.1413852
CHAS	2.686733819	0.86157976	3.11838	0.00193	0.99390419	4.379563	0.9939042	4.3795634
NOX	-17.7666112	3.81974371	-4.6513	4.2E-06	-25.271634	-10.2616	-25.271634	-10.261589
RM	3.809865207	0.41792525	9.11614	2E-18	2.98872677	4.631004	2.9887268	4.6310036
AGE	0.000692225	0.01320978	0.0524	0.95823	-0.0252623	0.026647	-0.0252623	0.0266468
DIS	-1.47556685	0.19945473	-7.398	6E-13	-1.867455	-1.08368	-1.867455	-1.0836787
RAD	0.306049479	0.06634644	4.6129	5.1E-06	0.17569217	0.436407	0.1756922	0.4364068
TAX	-0.01233459	0.00376054	-3.28	0.00111	-0.0197233	-0.00495	-0.0197233	-0.0049459
PTRATIO	-0.95274723	0.13082676	-7.2825	1.3E-12	-1.2097953	-0.6957	-1.2097953	-0.6956992
B	0.009311683	0.00268596	3.46679	0.00057	0.00403431	0.014589	0.0040343	0.0145891
LSTAT	-0.52475838	0.05071528	-10.347	7.8E-23	-0.6244036	-0.42511	-0.6244036	-0.4251131

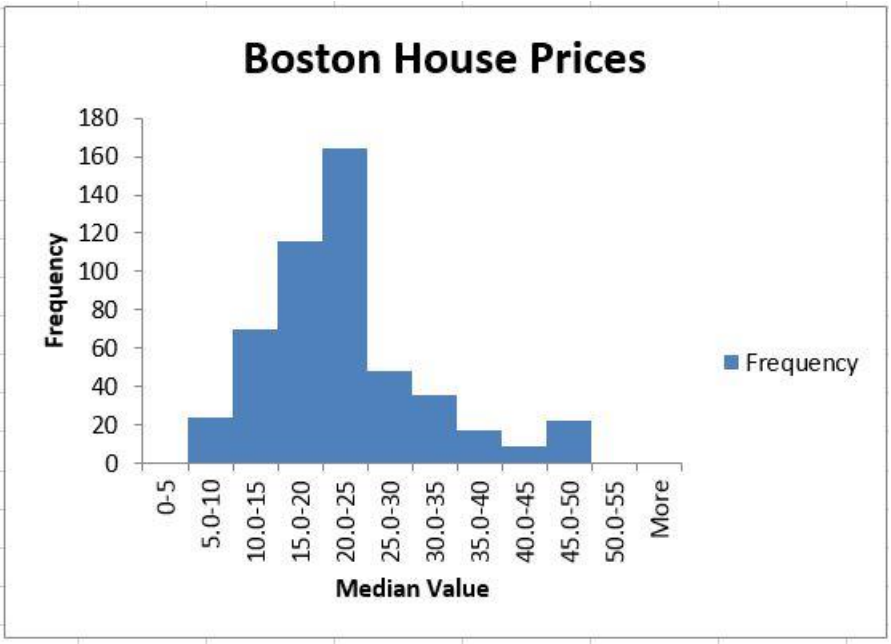
Here we see that: adjusted R-squared = 0.7338

F value = 108.077

In further stages will try to improve the above value by studing the skewness of dependent variable MDEV, significance of various dependent variables , removing multicollinearity.

STAGE 2:

Here we study the skewness of dependent variable so as to make appropriate transformation.



From the above histogram we see that MDEV is positively skewed.

MEDV	
Mean	22.53281
Median	21.2
Mode	50
Standard Dev	9.197104
Skewness	**1.108098**
Minimum	5
Maximum	50

Also its skewness value is 1.1 as obtained from the above table.

Hence a natural logarithmic transformation would be appropriate.

Now we fit the regression model by taking the ln(MDEV) as the dependent variable.

The summary of the new regression fit is presented in the following page.

Regression Statistics	
Multiple R	0.8886173
R Square	0.7896407
Adjusted R Square	**0.7840825**
Standard Error	0.1899368
Observations	506

ANOVA					
	df	*SS*	*MS*	***F***	*Significance F*
Regression	13	66.62711513	5.12516	**142.066**	4.067E-157
Residual	492	17.74937707	0.03608		
Total	505	84.37649219			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.1020423	0.204272598	20.0812	5.9E-66	3.70068803	4.5034	3.700688	4.50339655
CRIM	-0.0102715	0.001315464	-7.8083	3.5E-14	-0.01285616	-0.00769	-0.012856	-0.0076869
ZN	0.0011725	0.00054946	2.13385	0.03335	9.2886E-05	0.00225	9.289E-05	0.00225204
INDUS	0.0024668	0.002461445	1.00217	0.31675	-0.00236946	0.0073	-0.002369	0.00730302
CHAS	0.1008876	0.034485854	2.92548	0.0036	0.03312989	0.16865	0.0331299	0.16864532
NOX	-0.7783993	0.152890226	-5.0912	5.1E-07	-1.07879766	-0.478	-1.078798	-0.478001
RM	0.0908331	0.016728004	5.43	8.9E-08	0.05796593	0.1237	0.0579659	0.12370021
AGE	0.0002106	0.000528739	0.39832	0.69057	-0.00082826	0.00125	-0.000828	0.00124947
DIS	-0.0490873	0.007983436	-6.1486	1.6E-09	-0.06477317	-0.0334	-0.064773	-0.0334015
RAD	0.0142673	0.002655603	5.37252	1.2E-07	0.00904956	0.01948	0.0090496	0.019485
TAX	-0.0006258	0.00015052	-4.1574	3.8E-05	-0.00092151	-0.00033	-0.000922	-0.00033
PTRATIO	-0.0382715	0.005236512	-7.3086	1.1E-12	-0.04856018	-0.02798	-0.04856	-0.0279828
B	0.0004136	0.000107509	3.84681	0.00014	0.00020233	0.00062	0.0002023	0.0006248
LSTAT	-0.0290355	0.002029945	-14.304	4.7E-39	-0.03302395	-0.02505	-0.033024	-0.0250471

Here we see that: adjusted R-squared = 0.784

F value = 142.066

Hence we see that there is an improvement in adjusted R-squared and the F value of the test.

Now in order to further improve our model we see that from the t- stat values, the variables ZN, INDUS, AGE are not significant.

Also from correlation matrix (Table.2) we see that there is high collinearity between RAD and TAX. Hence in further analysis to improve the fit we ignore the TAX variable, due to which multicollinearity can be reduced.

STAGE 3:

Results of the new fit obtained by ignoring the variables ZN, INDUS, AGE, TAX below.

Regression Statistics	
Multiple R	0.8838704
R Square	0.7812269
Adjusted R Square	**0.7772572**
Standard Error	0.1929154
Observations	506

ANOVA					
	df	*SS*	*MS*	***F***	*Significance F*
Regression	9	65.91718544	7.324132	**196.7988**	1.877E-157
Residual	496	18.45930675	0.037216		
Total	505	84.37649219			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.021302	0.205166499	19.60019	9.12E-64	3.61819947	4.42440463	3.61819947	4.424404627
CRIM	-0.0100087	0.001329386	-7.52882	2.44E-13	-0.0126206	-0.0073968	-0.0126206	-0.00739678
CHAS	0.1161515	0.034666865	3.350505	0.000868	0.04803951	0.18426353	0.04803951	0.184263532
NOX	-0.8712566	0.139596664	-6.24124	9.32E-10	-1.1455303	-0.5969829	-1.1455303	-0.59698289
RM	0.1014707	0.016293064	6.227847	1.01E-09	0.06945877	0.13348263	0.06945877	0.133482632
DIS	-0.0442353	0.006551952	-6.75146	4.1E-11	-0.0571083	-0.0313623	-0.0571083	-0.03136226
RAD	0.0056058	0.001629488	3.440196	0.00063	0.00240421	0.00880731	0.00240421	0.008807308
PTRATIO	-0.0426888	0.004917489	-8.68102	5.68E-17	-0.0523505	-0.0330272	-0.0523505	-0.03302717
B	0.0004319	0.000108808	3.9698	8.26E-05	0.00021817	0.00064573	0.00021817	0.000645729
LSTAT	-0.0288434	0.001930463	-14.9412	6.04E-42	-0.0326363	-0.0250505	-0.0326363	-0.02505049

Here we see that: adjusted R-squared = 0.777

F value = 196.7

Hence we see that there is a significant improvement in the F value after the removing the variables ZN, INDUS, AGE, TAX from the model.

Therefore the final model is:

ln(MDEV) = 4.021302049 - 0.010008699(CRIM) + 0.116151521(CHAS) - 0.871256594(NOX) +0.1014707(RM) - 0.044235257(DIS) + 0.005605758(RAD) - 0.042688843(PTRATIO) + 0.000431947(B) - 0.028843381(LSTAT) + <error>

CONCLUSION

Various statistical techniques were used to eliminate predictors and irrelevant observations. After examining the final model we can say that house prices are higher in areas with lower crime, lower pupil-teacher ratio. House prices also tend to be higher closer to the Charles River, and houses with more rooms are pricier. The most interesting factors to consider are nitrogen oxide levels and distance to the main employment centers. On one hand people would want to live closer to their place of employment yet it is reasonable to say that pollution levels are higher as one moves closer major employment centers. The model suggests that people would prefer to live further away from their place of employment if it meant lower levels of pollution, which is an interesting point to consider. Now considering the internal factor (average number of rooms) the value of the residence is expected to increase with increase in this factor. On a concluding note, it is important to note that the data for this report was collected several decades ago. In the years since, there is no doubt that pollution levels have risen and it would be interesting to examine the ways in which that affects house pricing in Boston today.