Multiple Linear Regression: Predicting Boston Housing Prices

MIS 749 Spring 2015

Roger(Qiuye) Chen

Samuel Beckom

a) **Why should the data be partitioned into training and validation sets? For what will the training set be used for? For what will the validation set be used?**

The data needs to be partitioned into training and validation sets because in data mining it is more important for the model to be a good predictor of future data as opposed to classical statistics which is more interested in models which explain current data. The training data set is used to estimate the regression coefficients and thus build the model. The validation data set is used to test the model on how well the model performs when new data presents itself.

Fit a multiple linear regression model to the median house price (MEDV) as a function of CRIM, CHAS, and RM.

b) **Write the equation for predicting the median house price from the predictors in the model.**

After running Multiple Linear Regression based on CRIM, CHAS, and RM, we received the following result in table 1.

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	-26.36487	3.657911997	-7.20762769	4.64E-12	-33.5633	-19.1665	146718.3
CRIM	-0.296893	0.04903754	-6.05439434	4.21E-09	-0.39339	-0.20039	3577.683
CHAS	5.213219	1.465334088	3.557699735	0.000435	2.329583	8.096854	640.2803
RM	7.861049	0.577542309	13.61120934	3.48E-33	6.724502	8.997597	8088.495

Table 1 Mutiple Linear Regression based on CRIM, CHAS, and RM

$Y = -26.3649 - 0.2969x_1 + 5.2132x_2 + 7.8610x_3$
Where Y is MEDV, x_1 is CRIM, x_2 is CHAS, and x_3 is RM.

c) **What median price is predicted for a tract in the Boston area that does not bound the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6?**

After plugging in the numbers in the above formula we derived, we get $20,771.41.

d) **Reduce the number of predictors:**

i. **Which predictors are likely to be measuring the same thing among the 14 predictors? Discuss the relationship among INDUS, NOX, and TAX.**

According to Correlation Matrix from Figure 1, RAD and TAX are likely to be measuring the same thing with a 0.910 correlation coefficient. In reality, RAD demonstrates the popularity of the area, which is reflected in TAX.
INDUS, NOX, and TAX are relatively positively related to each other. If the proportion of nonretail business acres per town is high, it's more likely it's an industrialized area, which is a direct factor to Nitric oxide (NOX) emission. That explained the correlation coefficient of 0.764 between INDUS and NOX. It might be the reason that the

government does not encourage people to live in industrialized area since Nitric oxide is bad for health, so that there is also positive correlation between INDUS, NOX and TAX.

	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	CAT. MEDV
MEDV	1														
CRIM	-0.3883	1													
ZN	0.36045	-0.20047	1												
INDUS	-0.48373	0.40658	-0.53383	1											
CHAS	0.17526	-0.05589	-0.0427	0.06294	1										
NOX	-0.42732	0.42097	-0.5166	0.76365	0.0912	1									
RM	0.69536	-0.21925	0.31199	-0.39168	0.09125	-0.30219	1								
AGE	-0.37695	0.35273	-0.56954	0.64478	0.08652	0.73147	-0.24026	1							
DIS	0.24993	-0.37967	0.66441	-0.70803	-0.09918	-0.76923	0.20525	-0.74788	1						
RAD	-0.38163	0.62551	-0.31195	0.59513	-0.00737	0.61144	-0.20985	0.45602	-0.49459	1					
TAX	-0.46854	0.58276	-0.31456	0.72076	-0.03559	0.66802	-0.29205	0.50646	-0.53443	0.91023	1				
PTRATIO	-0.50779	0.28995	-0.39168	0.38325	-0.12152	0.18893	-0.3555	0.26152	-0.23247	0.46474	0.46085	1			
B	0.33346	-0.38506	0.17552	-0.35698	0.04879	-0.38005	0.12807	-0.27353	0.29151	-0.44441	-0.44181	-0.17738	1		
LSTAT	-0.73766	0.45562	-0.41299	0.6038	-0.05393	0.59088	-0.61381	0.60234	-0.497	0.48868	0.54399	0.37404	-0.36609	1	
CAT. MEI	0.78979	-0.15199	0.3653	-0.36628	0.10863	-0.2325	0.64127	-0.1912	0.11889	-0.19792	-0.27369	-0.44342	0.15514	-0.46991	1

Figure 1 Correlation Matrix

ii. **Compute the correlation table for the 13 numerical predictors and search for highly correlated pairs. These have potential redundancy and can cause multicollinearity. Choose which ones remove based on this table.**

The highlighted cells in Figure 1 are the correlation coefficients bigger than 0.7. Now let's compare the correlation matrix with the predictor coefficients from the regression model. If the sign of the correlation of 13 variables to MEDV is different to the coefficient from the regression model (Table 3), it is an indication that multicollinearity might occur. In this case, we want to consider the following suspicious predictors in Table 2:

Potential Predictors to be removed	Correlation	Coefficient
INDUS	-0.48373	0.0443746
AGE	-0.37695	0.0033188
DIS	0.24993	-1.641532
RAD	-0.38163	0.3436012

Table 2 Potential Predictors which have different signs for correlation to MEDV and coefficient to the model

Input Variables	Coefficient
Intercept	37.499522
CRIM	-0.111592
ZN	0.0704388
INDUS	0.0443746
CHAS	3.8213311
NOX	-18.67856
RM	3.2403711
AGE	0.0033188
DIS	-1.641532
RAD	0.3436012
TAX	-0.0141
PTRATIO	-0.78402
B	0.0109766
LSTAT	-0.564079

Table 3 Predictors' coefficients to the regression model

Go back to the correlation matrix, we can observe that the potential predictors to be removed are also the ones correlate to each other the most, which further confirms that these four predictors are producing redundant prediction. But we do not want to remove all of them. Instead, predictor(s) that represent the other variables the most will be left in the model. In determine that, we should look closely at each variable's coefficient of the regression formula. Soon we can find out that INDUS and AGE are not contributing the model as we expected, which indicates that they are the sub-contributing factors to the model. In addition, TAX is another factor we should be aware of because its high correlation to MEDV speaks differently than its low coefficient to the model. Theoretically, TAX is determined based on other factors in real life. In the meanwhile, the other two variables (DIS and RAD) are the true direct-contributing factors to the model.

If we cross INDUS, AGE, and TAX in the correlation matrix (Figure 2), we find out there is only one correlation coefficient is bigger than 0.7 between NOX and DIS, which further confirms that we removed the correct redundant variables. The reason neither DIS nor NOX is removed is they both contributing significantly to the model.

	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	CAT. MEDV
MEDV	1														
CRIM	-0.3883	1													
ZN	0.36045	-0.20047	1												
INDUS	-0.48373	0.40658	-0.53383	1											
CHAS	0.17526	-0.05589	-0.0427	0.06294	1										
NOX	-0.42732	0.42097	-0.5166	0.76365	0.0912	1									
RM	0.69536	-0.21925	0.31199	-0.39168	0.09125	-0.30219	1								
~~AGE~~	~~-0.37695~~	~~0.35273~~	~~-0.56954~~	~~0.64476~~	~~0.08652~~	~~0.73147~~	~~-0.24026~~	~~1~~							
DIS	0.24993	-0.37967	0.66441	-0.70803	-0.09918	-0.76923	0.20525	-0.74788	1						
RAD	-0.38163	0.62551	-0.31195	0.59513	-0.00737	0.61144	-0.20985	0.45602	-0.49459	1					
~~TAX~~	~~-0.46854~~	~~0.58276~~	~~-0.31456~~	~~0.72076~~	~~-0.03559~~	~~0.66802~~	~~-0.29205~~	~~0.50646~~	~~-0.53443~~	~~0.91023~~	~~1~~				
PTRATIO	-0.50779	0.28995	-0.39168	0.38325	-0.12152	0.18893	-0.3555	0.26152	-0.23247	0.46474	0.46085	1			
B	0.33346	-0.38506	0.17552	-0.35698	0.04879	-0.38005	0.12807	-0.27353	0.29151	-0.44441	-0.44181	-0.17738	1		
LSTAT	-0.73766	0.45562	-0.41299	0.6038	-0.05393	0.59088	-0.61381	0.60234	-0.497	0.48868	0.54399	0.37404	-0.36609	1	
CAT. MEI	0.78979	-0.15199	0.3653	-0.36628	0.10863	-0.2325	0.64127	-0.[illegible]912	0.11889	-0.19792	-0.27369	-0.44342	0.15514	-0.46991	1

Figure 2 Correlation Matrix with INDUS, AGE, and TAX Crossed Out

iii. Use an exhaustive search to reduce the remaining predictors as follows: First, choose the top three models. Then run each of these models separately on the training set, and compare their predictive accuracy for the validation set. Compare the RMSE and average error, as well as lift charts. Finally, describe the best model.

After running multiple linear regressions with the remaining 10 variables (Figure 4), we are at a much better position to choose the top three models.

						Model										
#Coef	RSS	Cp	R²	Adjusted R²	Probabil	1	2	3	4	5	6	7	8	9	10	11
1	25404.1731	631.5259	0	0	0	Intercept										
2	12165.8175	147.0567	0.5211	0.5195	0	Intercept										LSTAT
3	10581.5458	90.8396	0.5835	0.5807	0	Intercept					RM					LSTAT
4	9785.9956	63.6055	0.6148	0.6109	0	Intercept					RM			PTRATIO		LSTAT
5	9313.003	48.2245	0.6334	0.6285	0	Intercept			CHAS		RM			PTRATIO		LSTAT
6	8984.6411	38.1582	0.6463	0.6404	0	Intercept				NOX	RM	DIS		PTRATIO		LSTAT
7	8615.2392	26.5838	0.6609	0.654	0.0001	Intercept			CHAS	NOX	RM	DIS		PTRATIO		LSTAT
8	8379.7982	19.932	0.6701	0.6623	0.0022	Intercept		ZN	CHAS	NOX	RM	DIS		PTRATIO		LSTAT
9	8170.3405	14.2351	0.6784	0.6697	0.028	Intercept		ZN	CHAS	NOX	RM	DIS		PTRATIO	B	LSTAT
10	8088.8502	13.2406	0.6816	0.6718	0.0403	Intercept		ZN	CHAS	NOX	RM	DIS	RAD	PTRATIO	B	LSTAT
11	7973.4503	11	0.6861	0.6754	1	Intercept	CRIM	ZN	CHAS	NOX	RM	DIS	RAD	PTRATIO	B	LSTAT

Figure 3 The Best Two Models

According to Mallows Cp and Adjusted R^2, the best three models ranked in order should be the ones with coefficient 11, 10 and 9 (Notice one coefficient is for the intercept). The next step is to compare the predictive errors from these three models.

11 Coefficents (10 Variables)			10 Coefficents (9 Variables)			9 Coefficents (8 Variables)		
Training Data Scoring - Summary Report			Training Data Scoring - Summary Report			Training Data Scoring - Summary Report		
Total sum of squared errors	RMS Error	Average Error	Total sum of squared errors	RMS Error	Average Error	Total sum of squared errors	RMS Error	Average Error
7973.45	5.121372	-2.5E-15	8088.85	5.1583	-1.2E-14	8170.341	5.184218	-5.3E-15
Validation Data Scoring - Summary Report			Validation Data Scoring - Summary Report			Validation Data Scoring - Summary Report		
Total sum of squared errors	RMS Error	Average Error	Total sum of squared errors	RMS Error	Average Error	Total sum of squared errors	RMS Error	Average Error
3568.787	4.203244	0.149827	3688.44	4.273126	0.072098	3698.111	4.278724	0.050852

Figure 4 Error Comparison

Based on SSE, RMSE, and Average Error on the validation data (Figure 5), it supports our earlier rankings for the top three models:

1. The model with 11 coefficients (10 Variables: CRIM, ZN, CHAS, NOX, RM, DIS, RAD, PTRATIO, B, LSTAT)
2. The model with 10 coefficients (9 Variables: ZN, CHAS, NOX, RM, DIS, RAD, PTRATIO, B, LSTAT)
3. The model with 9 coefficients (8 Variables: ZN, CHAS, NOX, RM, DIS, PTRATIO, B, LSTAT)

The lift charts of the three models look almost identical (Figure 5, 6, 7):

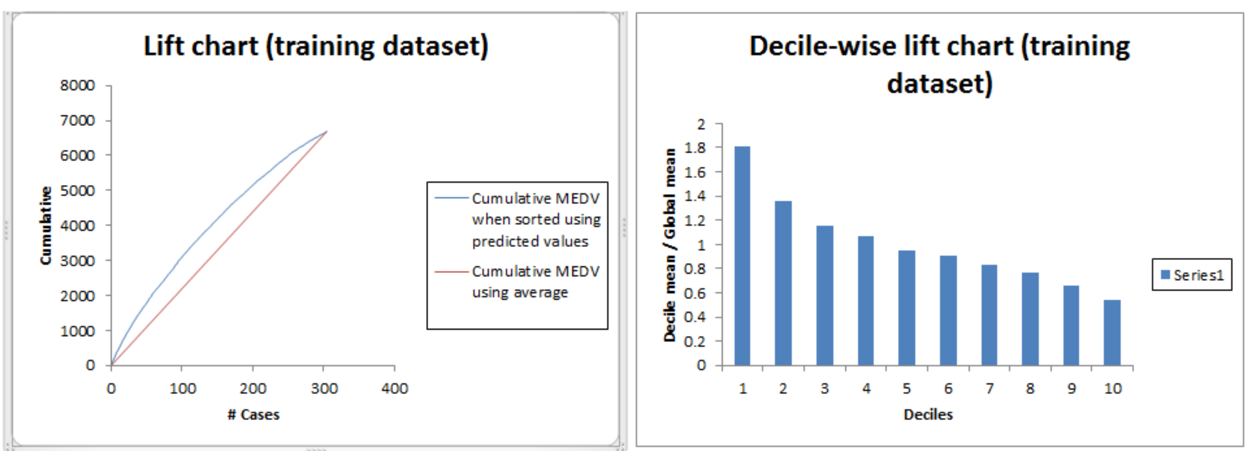

Figure 5. 11 Coefficients lift charts

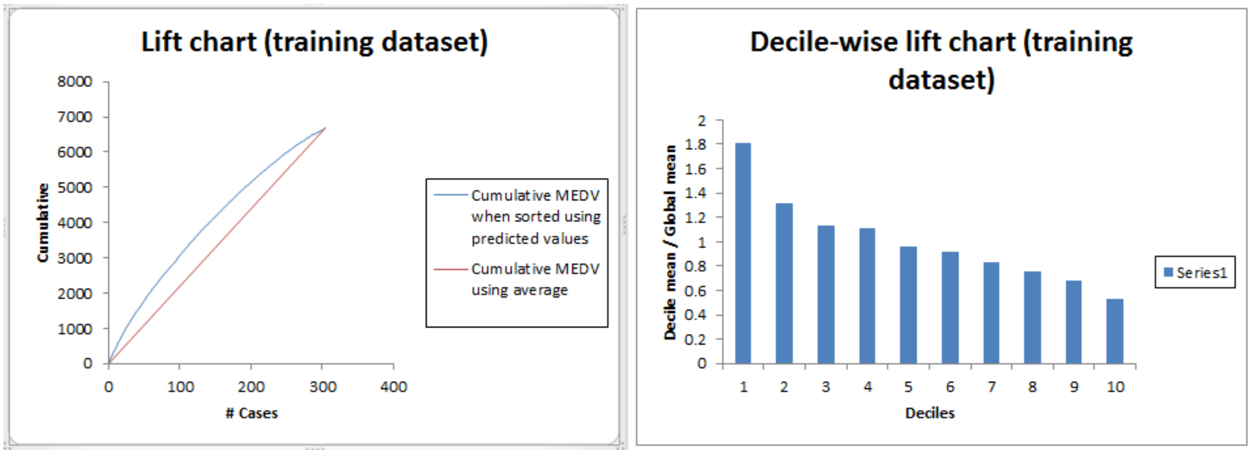

Figure 6. 10 Coefficients lift charts

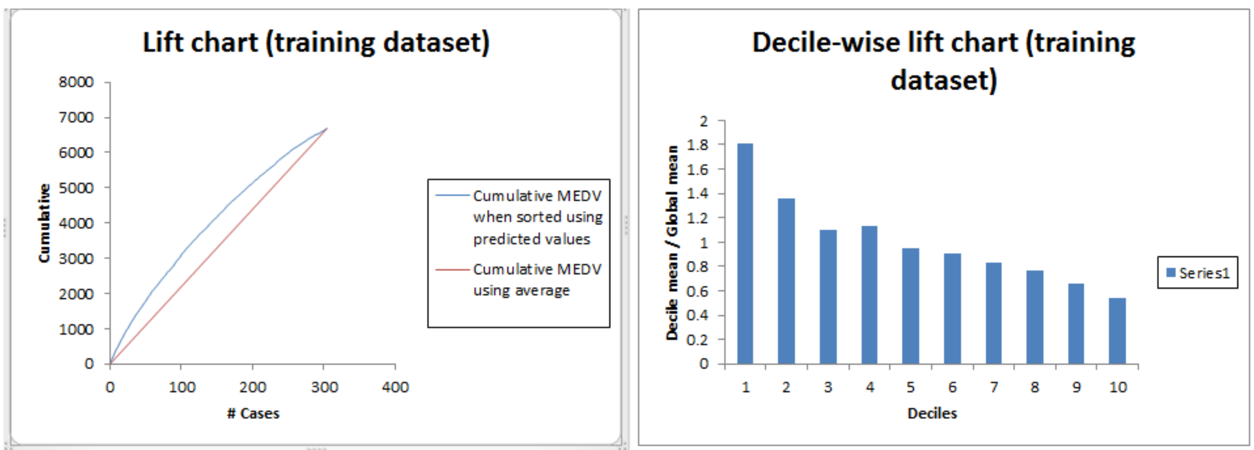

Figure 7. 9 Coefficients lift charts