PREDICTION OF MEDIAN VALUE OF HOUSE IN BOSTON SUBURBS

A PROJECT REPORT

Presented to the Department of Mathematics and Statistics

California State University, Long Beach

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Mathematics

Option in Statistics

Faculty Reviewer:

Kagba Suaray, Ph.D.

Junyan Deng

NOV 2015

**Table of Contents**

1. **Introduction**
2. **Data**

**3. Methodology**

**4. Variables correlation**

**5. Model Building**

**6. Residual Diagnostics**

**7. Outliers**

**8. Final Model**

**9. Validation of Final Model**

**10. Conclusion**

**Appendix1-** Bibliogphy

**Appendix 2-** SAS Code

1. **Introduction**

Median value is a good measurement to evaluate the house value in a specific area. This median value is related to some elements which are predictors. Understanding what important elements can predict the median value, city officers and the residents can improve these elements to better their house value. The aim of this paper is to determine what important elements can impact the median value, and predict a median value for a specific data set.

1. **Data**

The data were collected by Harrison, D. and Rubinfeld, D.L in 1993, and is taken from the StatLib library which is maintained at Carnegie Mellon University. The data consists of 506 observations and 14 attributes. Figure 1 displays the table of the attributes and a brief description of each variable. The independent variable used in this analysis will be MADV(median value) that is measured in $1000. I split the data into a model building set and a validation set using PROC SURVEYSELECT using 70% of the data into the model building set and 30% into the validation set.

Table 1-the data

|  |  |  |
| --- | --- | --- |
| 1.CRIM | continuous | Per capita crime rate by town |
| 2 ZN | continuous | Proportion of residential land zoned for lots over 25,000 sq.ft |
| 3 INDUS | continuous | proportion of non-retail business acres per town |
| 4 CHAS | indicator | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| 5 NOX | continuous | nitric oxides concentration (parts per 10 million) |
| 6 RM | integer | average number of rooms per dwelling |
| 7 AGE | continuous | proportion of owner-occupied units built prior to 1940 |
| 8 DIS | allocated codes | weighted distances to five Boston employment centers |
| 9 RAD | allocated | index of accessibility to radial highways |
| 10 TAX | continuous | full-value property-tax rate per $10,000 |
| 11 PTRATIO | continuous | pupil-teacher ratio by town |
| 12 B | continuous | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town |
| 13 LSTAT | continuous | Proportion of lower status of the population |
| 14 1MEDV | continuous | Median value of owner-occupied homes in $1000's |

1. **Methodology**

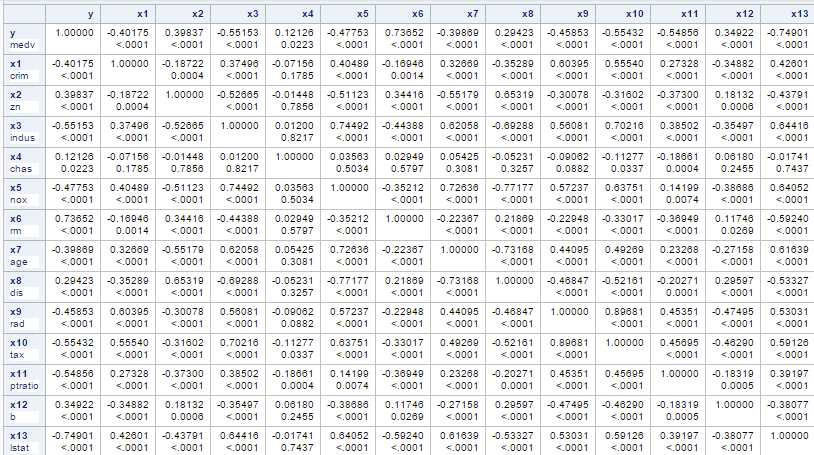
Multiple linear regression is used to determine which subset of predictor variables are significant. I will make a correlation transformation to get an overview importance of the predictor variables. To reduce multicollinearity of higher order variables, I will center the predictor variables by subtracting the mean for each variable. Several different methods, such as, different criterions and stepwise method, will be used to find the optimal model. Weighted least square is used to cancel the influence of outliers. Once reach a “best” model, then ensure the model is free of multicollinearity and conduct diagnostic check, the residual analysis, for any violations of the underlying assumptions. Finally, apply the model to the validation set to check for any bias that may be in the model. If the model has little bias, then interpret the parameters of the final model and their confident intervals, and draw a conclusion from data analysis.

**4 Variables correlation**

Table 2 shows the correlation between variables. The left side column show the relation between dependent variable and dependent variable. Unfortunately all the variable are highly related to the dependent variable. This implies multicollinearity. So centering the dependent variables is necessary to minimize the effect of multicollinearity.

When looking at the data, we can find that x13 and x6 are the most correlated to the dependent. The signs of the association helps us know the effect of independent variable on the dependent variable. These help us to build the model. X1, x3, x5, x7, x9, x10, x11, x13 have negative association while x2, x4, x6, x8, x12 have positive association.

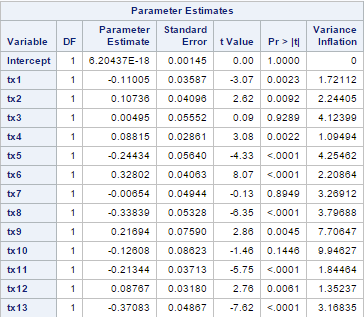
Table 2 Correlation matrix



**5 Model Building**

First we build a standardized regression model to get an over view of the parameters. Make correlation transformation of the predictors as a form txi, i=1, 2….12, 13.From table 3, we can find that the tx3, tx7 and tx10 are not significant. T13, t6 and t8 are most important, and tx5, tx9, and tx11 are less important. T1, t2, t4, and t12 are lest important. This information can be used in added variable method.

Table 3 standard regression



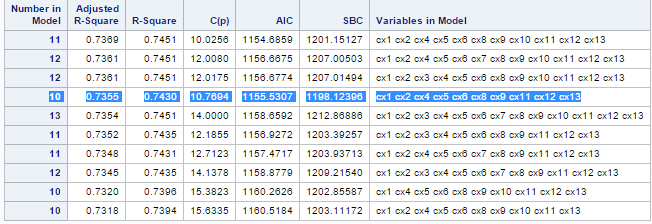
We will use several criteria, such as rsqure, adjrsq, cp, AIC and SBC, to help us select the best model. Stepwise, forward and backward selection also help us validate our choice.

Table 4 shows us the 10 models based on the criteria selection. I select the high light one with 10 variables. Because it has the smallest number of variables in the model with Adjust R-square increase in the corner, and CP closed to 10, and AIC and SBC are small enough.

Stepwise, backward and forward draw the same model with 10 predictors, with cx3, cx7 and cx10 eliminated. These selections method agree with my choice based on criterion.

Although criterion suggest us to choose cx1, cx2, cx4, cx5, cx6, cx8, cx9, cx11, cx12 and cx13, we should use added variable plots to determine if the variables should be added or should be added in a linear way or in a curved way. We begin with cx13 first into the model and add each variable each time into the model.

Table 3 Auto selection based on several criterions.



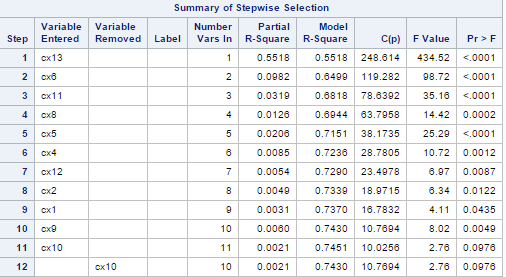
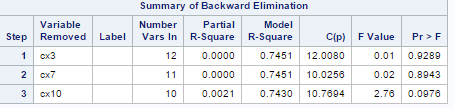
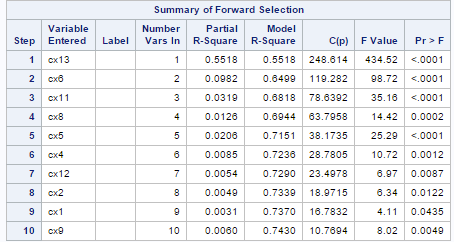
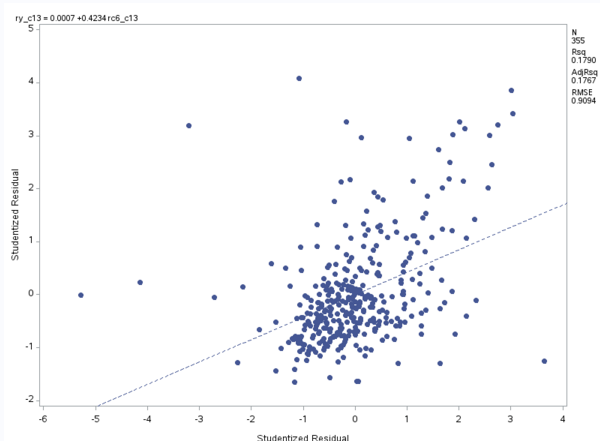
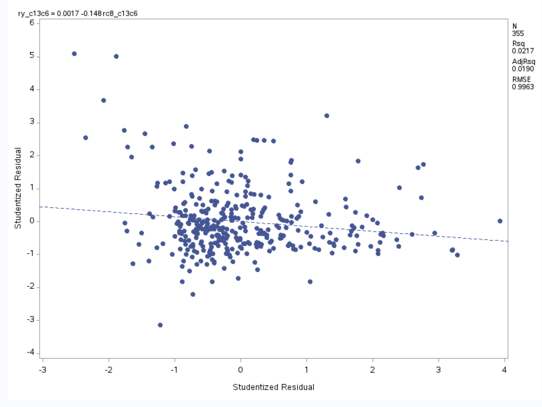
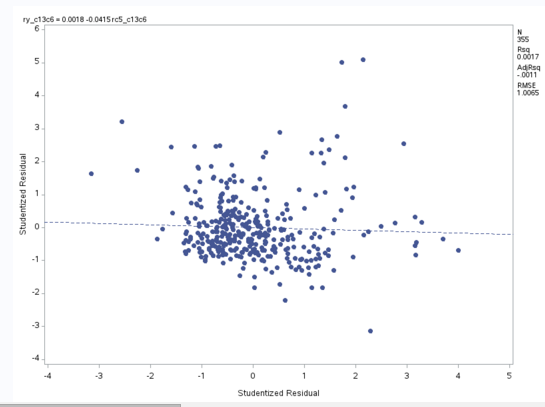
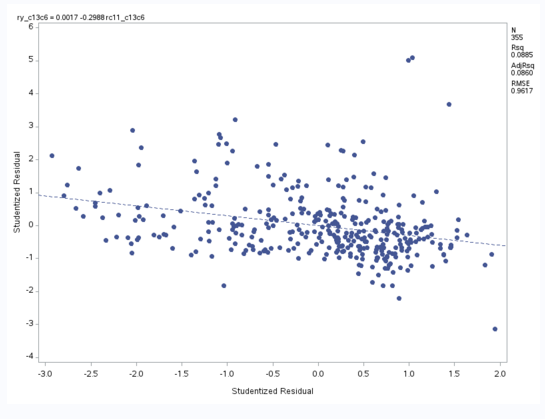
  

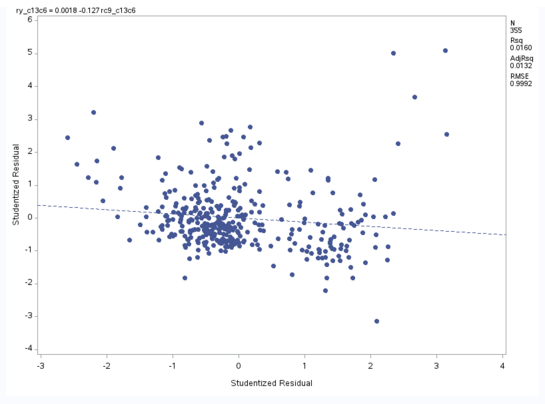
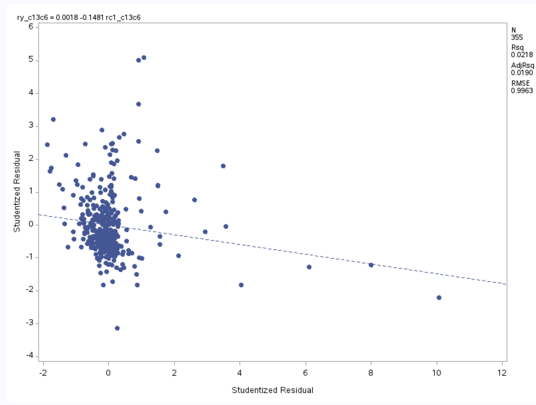
Figure 1 added variable into the model.

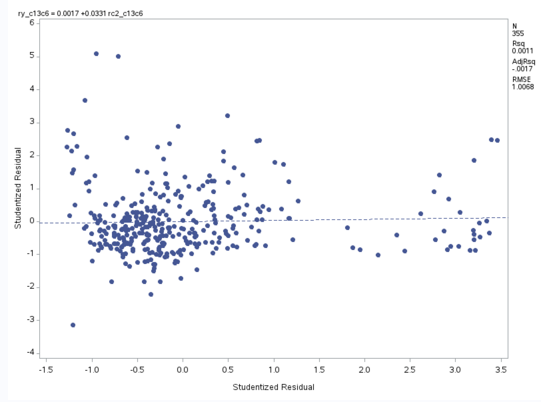
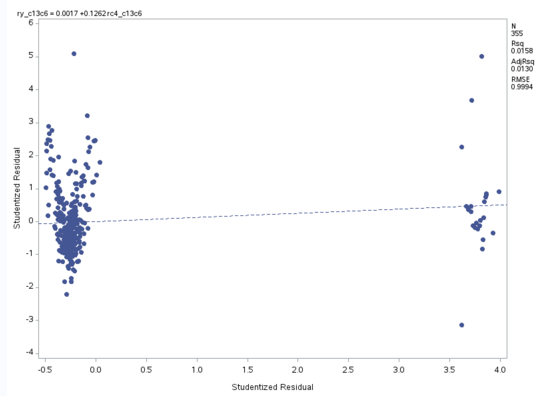
Cx6 cx8

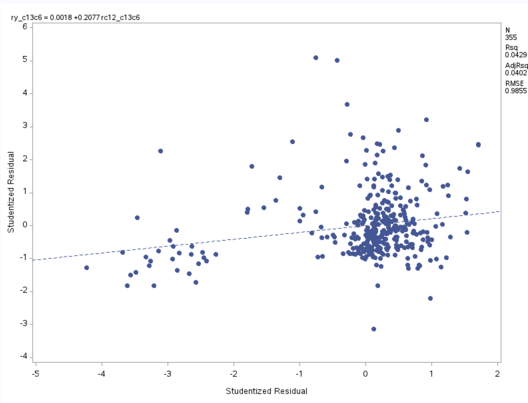
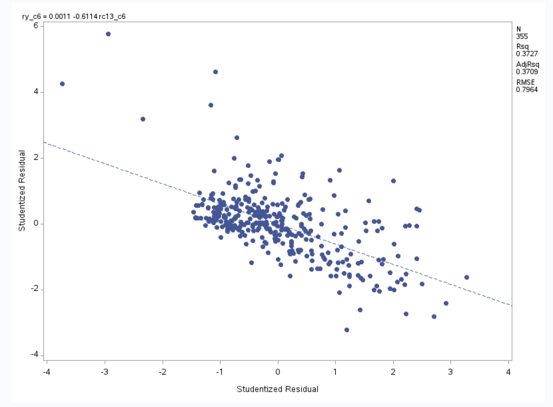
Cx5 cx11

Cx9 cx1

Cx2 cx4

Cx12 cx13 given cx6 in the model

From Figure 1 we can find that cx6, cx12 and cx13 should be stay in the model. Cx8, cx5, cx11, cx9, cx2, and cx4 are negligible small effect on the model when cx6, cx12, and cx13 have already in the model, because the slop is so flat. The slop of cx1 looks deep. But when look into the plot, we find out that that slop is due to the points far away which maybe outlier. Neglect those point, the slop of cx1 is still flat. We also find that all the independent variables should be added in the model with approximate linear function. Further function form needs residual analysis.

Table 5 show the regression output from SAS. VIF ‘s show multicollinearity is not a problem in the model. The F value of the model is 205.61, and the p value is less than 0.001. This shows that the model is significant. Cx13 is negative associate with the response variable but cx6 and cx12 are positive associate with the response variable. All the relations with the response variable agree with the correlation matrix in table 2.

Take interaction term into the model. From table 6 we find out that R square increases to 0.724 from 0.637. We should take interaction into the model.

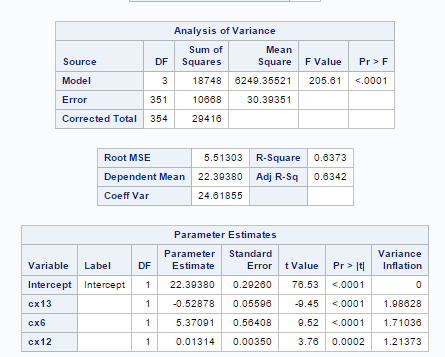
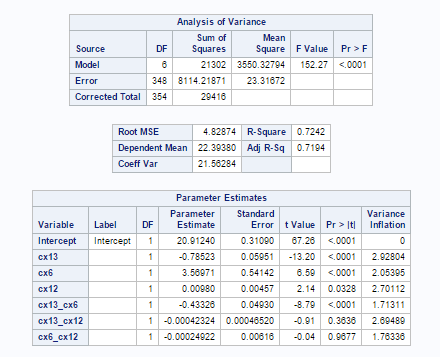
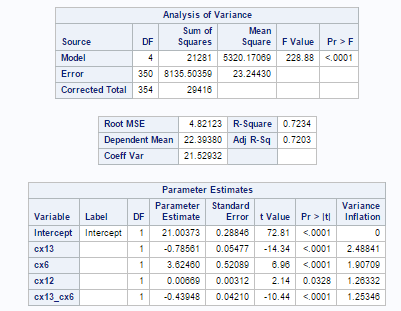
Table 5 preliminary regression output without interaction terms 

Table 6 Preliminary SAS output with interaction terms

1. **Residual Diagnostics**

We have already build a final model with 3 predictor and one interaction term. Before analysis, we have to check the assumption of the model. The assumptions are

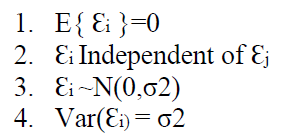
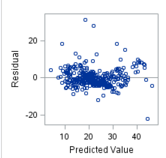
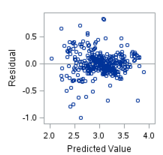


Figure 2

Before transformation log transformation

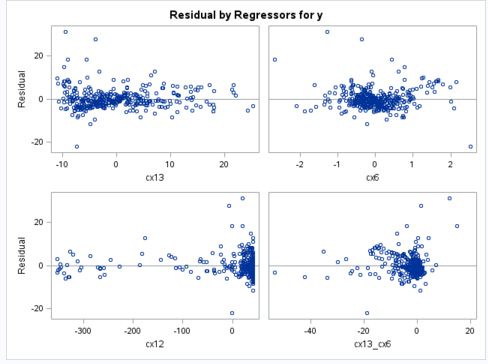
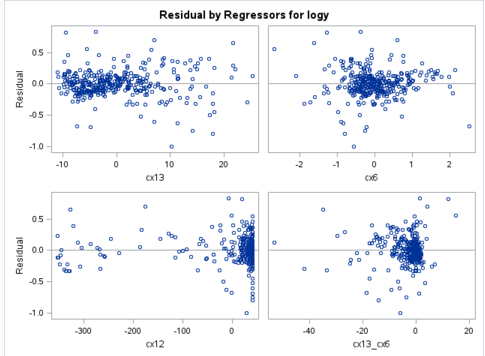
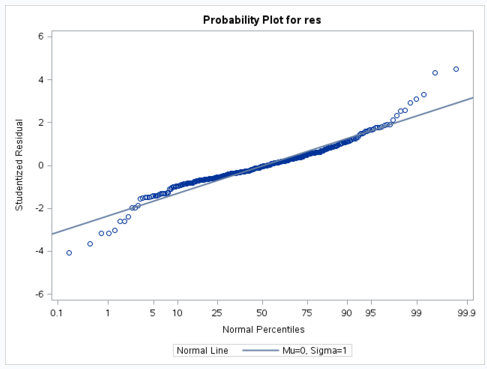
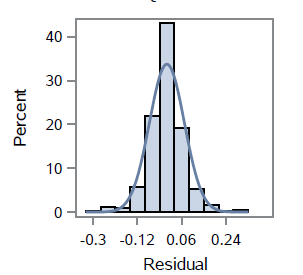
 

Figure3 normality plot and histogram

We can check residuals vs predicted value and predictors to check for independent and constant variance. From figure 2, on the left side we find out that the residual favor on the positive side. On the right side, using box-cox transformation, lambda=0.19. After transformation and get the residual plot, right side from figure 2. The residuals disperse more averagely and unstructured. And also the plots of residuals vs predictors improve, unstructured and disperse averagely. This satisfies constant variance and independent on predictors and dependent variable, which means error terms are independent (It is not a time series problem). Look at the normality plot in figure 3. We can find that 95% points fall in the strait line, which proves to be approximate normal distribution. The histogram shows normal. Heavy two tails cannot be canceled, this may be because the data is exactly t distribution. Spline transformation can good fit the data with R square up to 85%. But it is out of the scope of linear regression.

Box-cox transformation on dependent variable also improves the R square by 1 percent and lowers MSE very much from 23.24 to 0.00509.

**7 Outliers**

7.1 Checking outliers

Using the criteria of hii> 2p/n to detect x outliers, where hii is the diagonal elements of the hat matrix, and where 2p/n=2\*14/355=0.07887.

Using studentized deleted residuals to detect y outliers. 

The appropriate Bonferroni critical value is t(1-α/2n;n-p-1). In our model,

t(1-0.05/(2\*n),n-p-1)=t(1-0.05/(2\*355),355-14-1)=t(0.9999296,340)=-3.851. Take its absolute value is 3.851. the absolute value of data statistic greater than 3.851 should be considered to be y outliers.

Table 7

X outliers detected by hii.

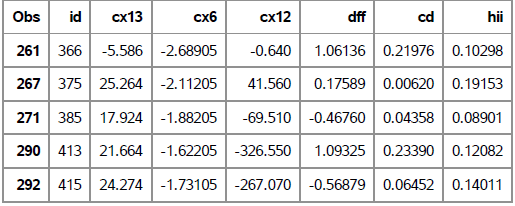


Table 8

Y outliers detected by studentized deleted residual

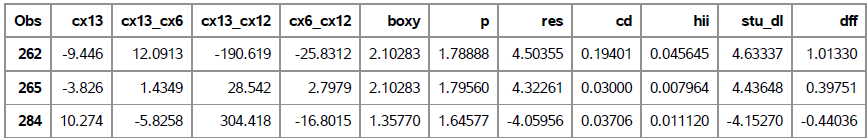
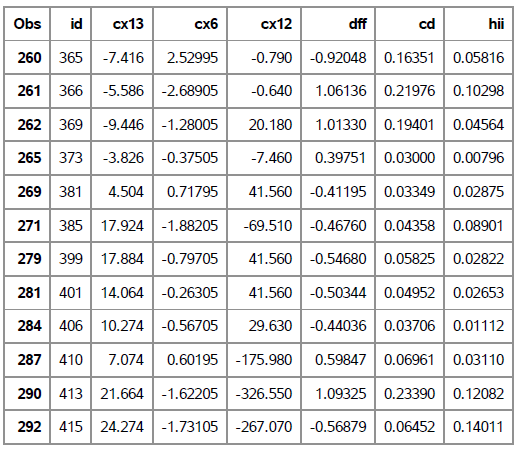


Table 9

X and Y outliers detected by DFFITS



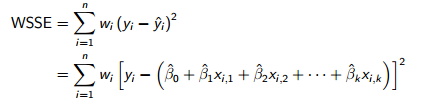
  DFFITS is a diagnostic meant to show how influential a point is. The formula is



The letters DF stand for the difference between the fitted value  including ith case and the fitted value  excluding ith case. S(i) is the stander error of ith case, hii is the diagonal elements of hat matrix. Values larger than 2\*sqrt(p/n) in absolute value are considered highly influential.

7.2 weighted Least squares

Instead of minimize the regular sum of square of residual. We can minimize the weighted least squares.

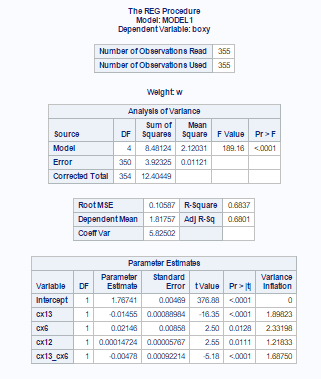
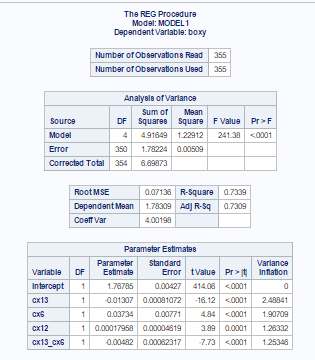
 

When  is unknown, we use  to estimate. But the disadvantage of weighted least squares is obvious. If the estimate is biased or by small number of replicates, it would lead to bad model.

In our data set, see table 10, we find out that using weighted method makes the model worse. We should not admit weighted least squares method.

Table 10

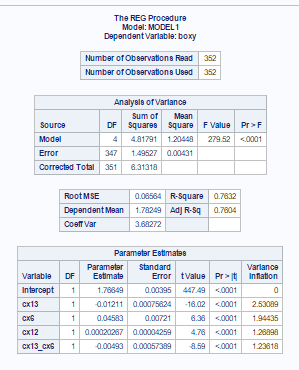
Unweighted and weighted method.

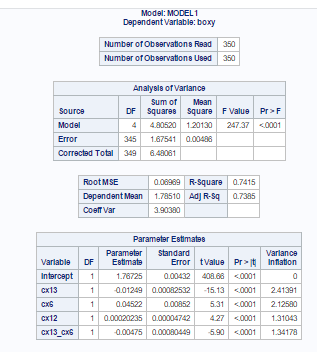


7.3 comparing delete outliers and keep outliers

Since we cannot remedy the effect or outlier using weighted least squares method. We cannot simply delete these outliers. Compare the difference of the model parameters between deleting and keeping the outliers. In table 11, we find out that deleting only y outlier improve R square, but there are no other difference between deleting only Y outliers and deleting both Y and X outliers. Comparing table 11 and table 10, we find out that the parameter of cx6 increase slightly from 0.037 to 0.045 and R square increases only 3%. So we conclude that the outliers do not impact the model too much.

Table 11

Delete y outliers delete x and y outliers



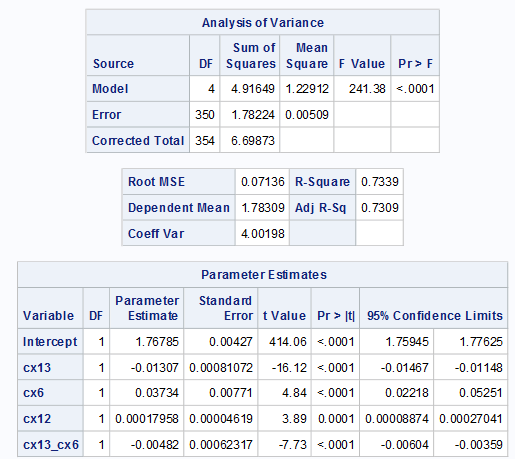
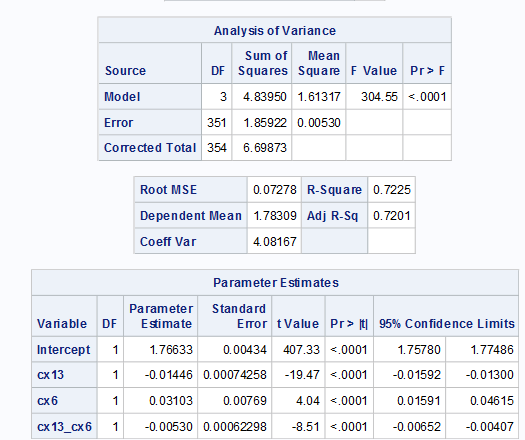
**8 Final model**

We reach the final model that include cx13, cx6, cx12 and the interaction between cx13 and cx6. The parameter of c12 is negligible small. Delete c12 in our model R square does not decrease much. And other parameters do not change much. So for interpretation better we can delete c12 and reach our final model only include cx13, cx6 and the interaction between them.

The ANOVA table show that the model is significant under the P-value of F test less than 0.0001. The P-value of each parameter is less than or equal to 0.0001. R square is 0.7225, which is good fit of the model. 72.25 of the variation of response variable is explained by the model. We also get the 95% confidence limits of the parameters. Notice that here we have center the predictor, so the parameters themselves explain the average effect of increasing one unit in the center variable has the effect on the box-cox transformation of the response variable.

Table 12 final model

Include cx12 delete cx12

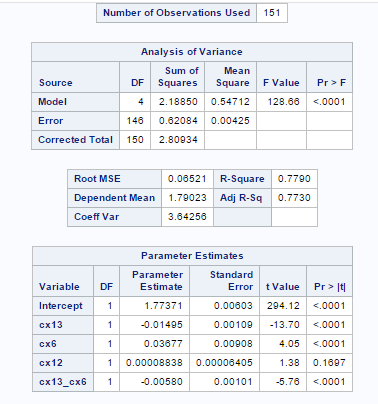
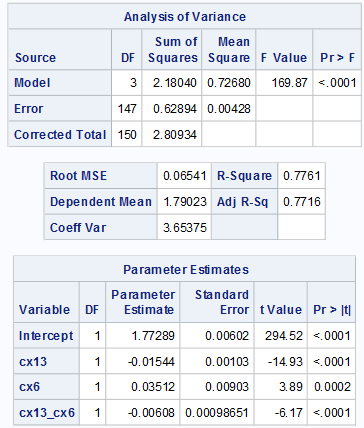
**9 Validation of the model**

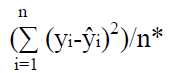
We use the test data split from the raw data to get the predictive power of our model. We use the exact method of our model building the get the ANOVA table from the test data and compare the difference. We find out that the parameter is similar except cx12. Cx12 becomes insignificant and only a half of the parameter of c12 in our model. This reinforces our consideration of deleting c12 in our final model. These evidences prove that our model has high predictive power.

We also can obtain MSPR( mean square prediction error) to check our model. If MSPR is very closed to the MSE from our model, then our model has good predictive power.

Table 13 validation model

Include cx12 final model

MSPR=

Yi is the value of the response variable in the ith validation case.

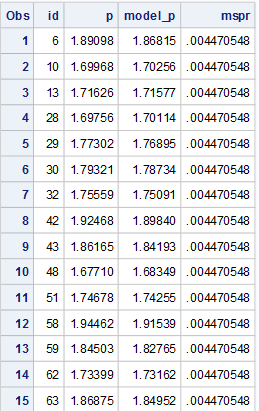
Y(hat)i is the predicted value of the ith validation case based on the model building data set.

n\* is the number of cases in the validation set.

In our test data set, we get MSPR=0.00447, which is very closed the MSE=0.0053 from our final model. This indicates that our model has high predictive power.

Table 14

MSPR(only shows 15 observation)



**10 Conclusion**

The purpose of this analysis is to determine what important factors can predict the median value of housing in Boston. In our final model, we find out that LSTAT (Proportion of lower status of the population), RM (average number of rooms per dwelling) and their interaction contain much information of the median value of house. Decrease the proportion of lower status of the population and increase the average number of rooms can increase the median value of the house in Boston. Although other predictors can predict the median value of house, but they contribute little when comparing to LSTAT and RM. Weighted least squares has its own disadvantage which assume the variance is known or easy to assess. Biased estimate would lead to bad model. Our paper only concerns the multiple linear regression method, although other methods can get a better prediction, such as spline transformation of predictors and adding higher order of predictors.

Appendix 1

Kutner, Michael H. Applied linear statistical models. Boston: McGraw-Hill Irwin, 2005.

Sample project, author unknown

Data set is taken from UCI website

<https://archive.ics.uci.edu/ml/datasets/Housing>

Appendix 2

SAS code

**data** rawdat;

infile “F:\course\510\hw\project\housing.data”;

input id x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 y;

label x1="crim"

x2="zn"

x3="indus"

x4="chas"

x5="nox"

x6="rm"

x7="age"

x8="dis"

x9="rad"

x10="tax"

x11="ptratio"

x12="b"

x13="lstat"

y="medv";

datalines;

**run**;

proc surveyselect data=rawdat out=split rate=0.7 outall;

run;

proc sql;

create table build as select \* from split where selected=1;

quit;

proc sql;

create table test as select \* from split where selected=0;

quit;

proc print data=build;

var x1-x13 y;

run;

proc print data=test;

var x1-x13 y;

run;

proc corr data=build;

var y x1-x13;

run;

proc sql;

create table a as

select \*, (y-mean(y))/(std(y)\*sqrt(count(y)-1)) as ty,

(x1-mean(x1))/(std(x1)\*sqrt(count(x1)-1)) as tx1,

(x2-mean(x2))/(std(x2)\*sqrt(count(x2)-1)) as tx2,

(x3-mean(x3))/(std(x3)\*sqrt(count(x3)-1)) as tx3,

(x4-mean(x4))/(std(x4)\*sqrt(count(x4)-1)) as tx4,

(x5-mean(x5))/(std(x5)\*sqrt(count(x5)-1)) as tx5,

(x6-mean(x6))/(std(x6)\*sqrt(count(x6)-1)) as tx6,

(x7-mean(x7))/(std(x7)\*sqrt(count(x7)-1)) as tx7,

(x8-mean(x8))/(std(x8)\*sqrt(count(x8)-1)) as tx8,

(x9-mean(x9))/(std(x9)\*sqrt(count(x9)-1)) as tx9,

(x10-mean(x10))/(std(x10)\*sqrt(count(x10)-1)) as tx10,

(x11-mean(x11))/(std(x11)\*sqrt(count(x11)-1)) as tx11,

(x12-mean(x12))/(std(x12)\*sqrt(count(x12)-1)) as tx12,

(x13-mean(x13))/(std(x13)\*sqrt(count(x13)-1)) as tx13,

(x1-mean(x1)) as cx1,

(x2-mean(x2)) as cx2,

(x3-mean(x3)) as cx3,

(x4-mean(x4)) as cx4,

(x5-mean(x5)) as cx5,

(x6-mean(x6)) as cx6,

(x7-mean(x7)) as cx7,

(x8-mean(x8)) as cx8,

(x9-mean(x9)) as cx9,

(x10-mean(x10)) as cx10,

(x11-mean(x11)) as cx11,

(x12-mean(x12)) as cx12,

(x13-mean(x13)) as cx13

from build;

proc reg data=a;

model ty=tx1-tx13/vif;

run;

proc reg data=a;

model y=cx1-cx13/vif selection= adjrsq aic sbc cp rsquare best=10;

run;

proc reg data=a;

model y=cx1-cx13/vif selection=stepwise sle=0.25 sls=0.05;

run;

proc reg data=a;

model y=cx1-cx13/vif selection=backward sls=0.05;

run;

proc reg data=a;

model y=cx1-cx13/vif selection=forward sle=0.05 sls=0.05;

run;

/\*using added method to check. start from c13\*/

proc reg data=a noprint;

model y cx6=cx13 ;

output out=res student=ry\_c13 rc6\_c13;

run;

proc reg data=res noprint;

model ry\_c13=rc6\_c13;

symbol1 i=sm70s v=dot;

plot ry\_c13\*rc6\_c13;

run;

/\*cx6 is important givin cx13 in the model\*/

/\*check cx13 given cx6 in the model\*/

proc reg data=a noprint;

model y cx13= cx6;

output out=res student=ry\_c6 rc13\_c6;

run;

proc reg data=res;

model ry\_c6=rc13\_c6;

symbol1 i=sm70s v=dot;

plot ry\_c6\*rc13\_c6;

run;

/\*cx6 is important givin cx13 is in the model, add cx8\*/

proc reg data=a noprint;

model y cx8=cx13 cx6;

output out=res student=ry\_c13c6 rc8\_c13c6;

run;

proc reg data=res noprint;

model ry\_c13c6=rc8\_c13c6;

symbol1 i=sm70s v=dot;

plot ry\_c13c6\*rc8\_c13c6;

run;

/\*cx8 is not imporatn givin cx13 and cx6 are in the model\*/

/\*add cx5\*/

proc reg data=a noprint;

model y cx5=cx13 cx6;

output out=res student=ry\_c13c6 rc5\_c13c6;

run;

proc reg data=res noprint;

model ry\_c13c6=rc5\_c13c6;

symbol1 i=sm70s v=dot;

plot ry\_c13c6\*rc5\_c13c6;

run;

/\*cx5 is not important\*/

/\*add cx11\*/

proc reg data=a noprint;

model y cx11=cx13 cx6;

output out=res student=ry\_c13c6 rc11\_c13c6;

run;

proc reg data=res noprint;

model ry\_c13c6=rc11\_c13c6;

symbol1 i=sm70s v=dot;

plot ry\_c13c6\*rc11\_c13c6;

run;

/\*cx11 is not important\*/

/\*add cx9\*/

proc reg data=a noprint;

model y cx9=cx13 cx6;

output out=res student=ry\_c13c6 rc9\_c13c6;

run;

proc reg data=res noprint;

model ry\_c13c6=rc9\_c13c6;

symbol1 i=sm70s v=dot;

plot ry\_c13c6\*rc9\_c13c6;

run;

/\*cx9 is not important\*/

/\*add cx1\*/

proc reg data=a noprint;

model y cx1=cx13 cx6;

output out=res student=ry\_c13c6 rc1\_c13c6;

run;

proc reg data=res noprint;

model ry\_c13c6=rc1\_c13c6;

symbol1 i=sm70s v=dot;

plot ry\_c13c6\*rc1\_c13c6;

run;

/\*cx1 is not important\*/

/\*add cx2\*/

proc reg data=a noprint;

model y cx2=cx13 cx6;

output out=res student=ry\_c13c6 rc2\_c13c6;

run;

proc reg data=res noprint;

model ry\_c13c6=rc2\_c13c6;

symbol1 i=sm70s v=dot;

plot ry\_c13c6\*rc2\_c13c6;

run;

/\*cx2 is not important\*/

/\*add cx4\*/

proc reg data=a noprint;

model y cx4=cx13 cx6;

output out=res student=ry\_c13c6 rc4\_c13c6;

run;

proc reg data=res noprint;

model ry\_c13c6=rc4\_c13c6;

symbol1 i=sm70s v=dot;

plot ry\_c13c6\*rc4\_c13c6;

run;

/\*cx4 is not important\*/

/\*add cx12\*/

proc reg data=a noprint;

model y cx12=cx13 cx6;

output out=res student=ry\_c13c6 rc12\_c13c6;

run;

proc reg data=res noprint;

model ry\_c13c6=rc12\_c13c6;

symbol1 i=sm70s v=dot;

plot ry\_c13c6\*rc12\_c13c6;

run;

proc reg data=a;

model y=cx13 cx6 cx12/vif;

run;

data ac;

set a;

cx13\_cx6=cx13\*cx6;

cx13\_cx12=cx13\*cx12;

cx6\_cx12=cx6\*cx12;

run;

proc reg data=ac;

model y=cx13 cx6 cx12 cx13\_cx6 cx13\_cx12 cx6\_cx12/vif;

run;

proc reg data=ac;

model y=cx13 cx6 cx12 cx13\_cx6/vif;

output out=res r=res p=p;

run;

proc gplot data=res;

plot res\*p;

plot res\*(cx13 cx6 cx12 cx13\_cx6);

run;

proc transreg data=ac;

model boxcox(y/lambda=-3 to 3 by 0.01)=identity(cx13 cx6 cx12);

run;

data ac;

set ac;

boxy=y\*\*0.19;

run;

proc reg data=ac;

model boxy=cx13 cx6 cx12 cx13\_cx6/vif influence;

output out=res1 student=res p=p rstudent=stu\_dl h=hii cookd=cd dffits=dff ;

run;

proc univariate data=res1;

var res;

probplot/normal(mu=0 sigma=1 color=red);

run;

/\*exam outliner\*/

/\*2p/n=2\*14/355=0.07887\*/

proc print data=res1;

var id cx13 cx6 cx12 dff cd hii;

where hii>0.07887;

run;

/\*t(1-0.05/(2\*n),n-p-1)=t(1-0.05/(2\*355),355-14-1)=t(0.9999296,340)=-3.851\*/

proc print data=res1;

where stu\_dl>3.851 or stu\_dl<-3.851;

run;

/\*2\*sqrt(p/n)=2\*sqrt(14/355)=0.3972\*/

proc print data=res1;

var id cx13 cx6 cx12 dff cd hii;

where dff>0.3972 or dff<-0.3972;

run;

/\*F(0.05,14,355)=0.47\*/

proc print data=res2;

var id cx13 cx6 cx12 dff cd hii;

where cd>0.47;

run;

/\*weighted method\*/

data wls;

set res1;

absr=abs(res);

run;

proc reg data=wls;

model absr=cx13 cx6 cx12 cx13\_cx6;

output out=weigh p=s;

run;

data weighted;

set weigh;

w=1/(s\*\*2);

run;

proc reg data=weighted;

weight w;

model boxy=cx13 cx6 cx12 cx13\_cx6/vif influence;

run;

proc reg data=weighted;

weight w;

model boxy=cx13 cx6 cx12 cx13\_cx6;

run;

/\*delete y outlier\*/

data res2;

set res1;

where -3.851<stu\_dl<3.851;

run;

proc reg data=res2;

model boxy=cx13 cx6 cx12 cx13\_cx6/vif;

output out=res3 student=res1 p=p1;

run;

proc gplot data=res3;

plot res1\*p1;

run;

/\*delet x y outlier\*/

data res4;

set res1;

where -3.851<stu\_dl<3.851;

where -0.3972<dff<0.3972;

where hii<0.07887;

run;

proc reg data=res4;

model boxy=cx13 cx6 cx12 cx13\_cx6/vif;

output out=res5 student=res1 p=p1;

run;

proc gplot data=res5;

plot res1\*p1;

run;

/\*final model\*/

proc reg data=ac;

model boxy=cx13 cx6 cx12 cx13\_cx6/p cli clm clb;

run;

/\*model validation\*/

proc sql;

create table tes as

select \*,

(y\*\*0.19) as boxy,

(x1-mean(x1)) as cx1,

(x2-mean(x2)) as cx2,

(x3-mean(x3)) as cx3,

(x4-mean(x4)) as cx4,

(x5-mean(x5)) as cx5,

(x6-mean(x6)) as cx6,

(x7-mean(x7)) as cx7,

(x8-mean(x8)) as cx8,

(x9-mean(x9)) as cx9,

(x10-mean(x10)) as cx10,

(x11-mean(x11)) as cx11,

(x12-mean(x12)) as cx12,

(x13-mean(x13)) as cx13

from test;

quit;

data b;

set tes;

cx13\_cx6=cx13\*cx6;

run;

proc reg data=b;

model boxy=cx13 cx6 cx12 cx13\_cx6;

run;

proc reg data=b;

model boxy=cx13 cx6 cx13\_cx6;

output out=valid p=p;

run;

PROC SQL;

CREATE TABLE MSPR AS

SELECT \*,

(**1.76785**-**0.01307**\*cx13+**0.03734**\*cx6 -**0.00482**\*cx13\_cx6) as model\_p,

(mean((boxy-model\_p)\*\*2)) as mspr

from valid;

quit;

proc print data=mspr;

var id p model\_p mspr;

run;