1. Stata is chosen to be our software package

2.Description of data set

The Boston data has 14 variable and 506 observation.

The data frame contains the following variables:

crim: per capita crime rate by town.

zn: proportion of residential land zoned for lots over 25,000 sq.ft.

indus: proportion of non-retail business acres per town.

chas: Charles River value = 1 if tract bounds river, and value = 0 otherwise.

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

age: proportion of owner-occupied units built prior to 1940.

dis: weighted mean of distances to five Boston employment centres.

rad: index of accessibility to radial highways.

tax: full-value property-tax rate per \$10,000.

ptratio: pupil-teacher ratio by town.

black: 1000\*(proportion of blacks by town - 0.63)^2.

lstat: lower status of the population (in percentage).

medv: median value of owner-occupied homes in \$1000s.

3.Research question (Objective)

We are interseted in predicting the median value of owner-occupied home (medv).

We will use Multiple Linear Regression with numerical variable as

our numerical predictors, look at their relationship with medv,

and see how many variation is explained by the model, so we can conclude that

whether our prediction is accurate.

As a statistician, we generally prefer a simpler model.

Our research project will perform feature selecting to prevent overfitting the model.

We will choose the number of predictors using the following criteria.

The variable is that has the highest correlation with medv be selected first.

However, if the variable is highly correlated with the previous selected variable,

then we will choose the variable that has second highest correlation with medv.

We will stop adding predictors when adjusted R square increase only by a small amount.

Next, we will need to check whether our model satisfied Gauss–Markov Theorem.

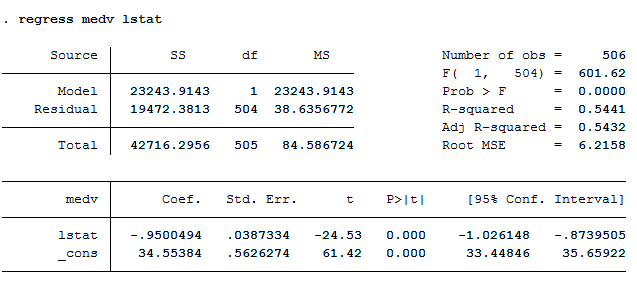
If it is not satisfied, transformation is required, or a new model is needed.

For our research project, we will not be looking at any categorical variable

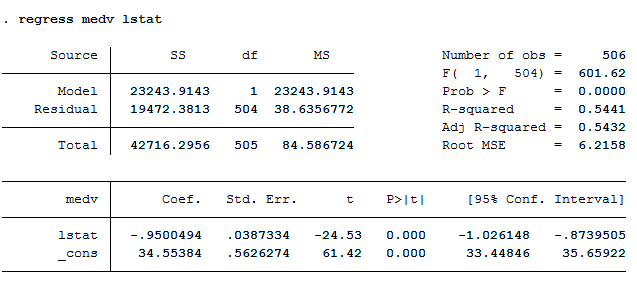
The final model may be harder to interpret with categorical variable involved, so we want to avoid that.

5. Result (Multiple linear regression feature selecting)

We select lstat as our first predictor since it has the highest correlation with our response variable medv

We have adjusted R^2 of 0.5432

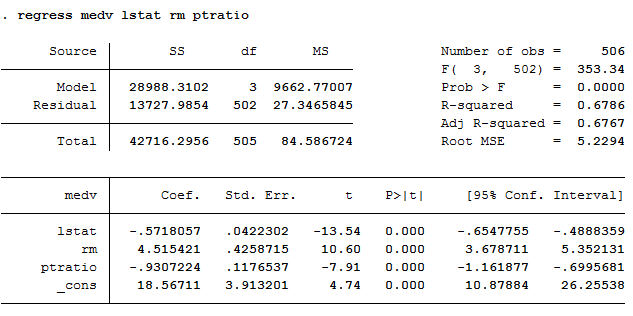
We add another numerical variable rm that is highly correlated with medv



adjusted R^2 increase from 0.5432 to 0.6371.

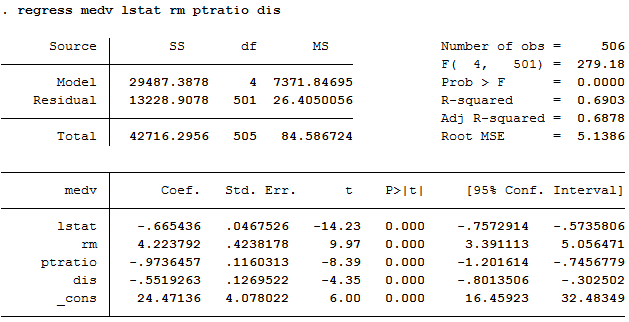
A huge increase in adjusted R^2.

add another numerical variable ptratio that is highly correlated with medv



adjusted R^2 increase from 0.6371 to 0.6767.

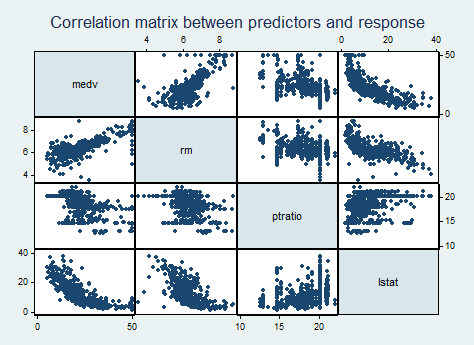
adjusted R^2 increase by a large amount, which is good.

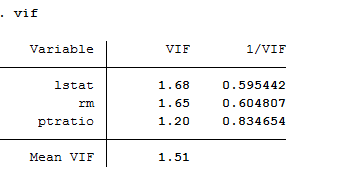
To prevent multiple collinearity, we select numerical variable dis that is moderately correlated with medv and not highly with the predictors we already selected 

adjusted R^2 square increase only by around 1% this time.

Simpler model is always preferred.

Therefore, model with subset size of 3 is our best subset.



4.Descriptive Statistics 

1.rm has strong positive correlation with medv

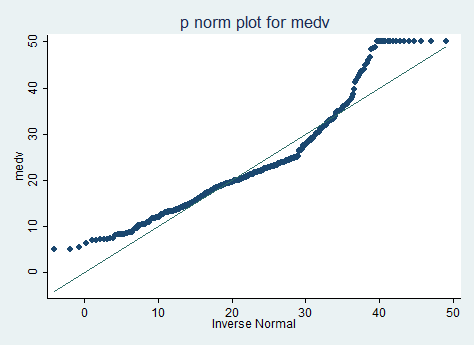
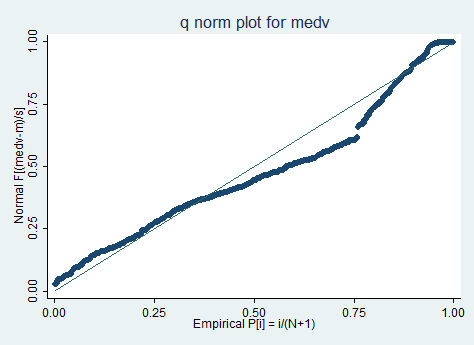
2.ptration has negative correlation with medv

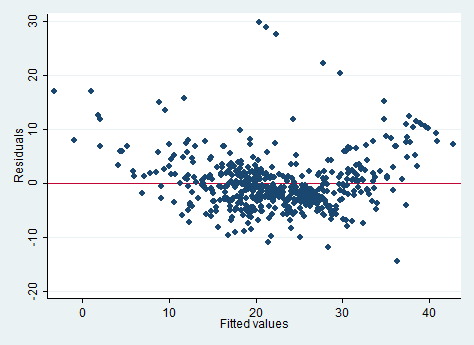
3.lstat has strong negative correlation with medv

4.rm and lstat has moderately negative correlating with each other.

5.Other predictor doesn't seem to have correlating with each other.

Since the VIF is below 5, we don't have multiple collinearity issue.



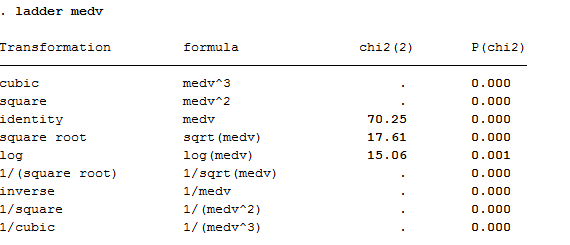


The graph suggests that we have issues with normality, and

the residual vs fitted plot shows that the residuals are not competently independent.

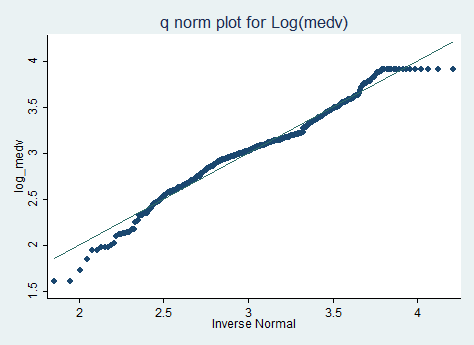
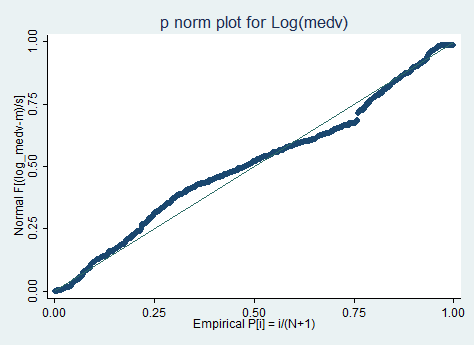
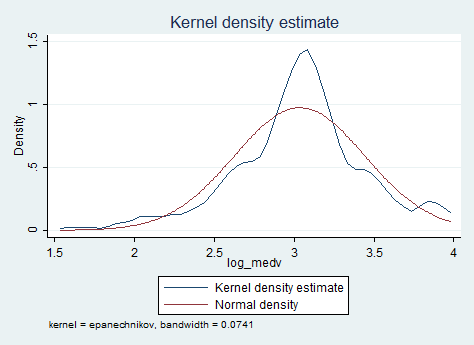
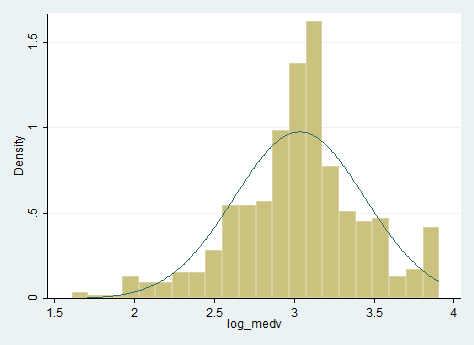
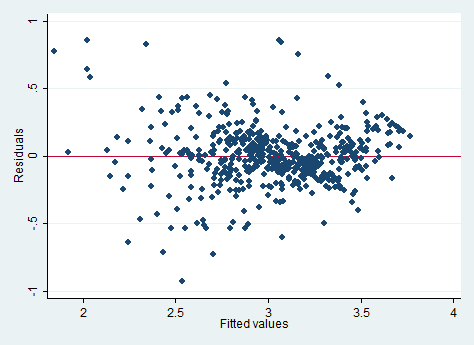
The assumption for Gauss–Markov Theorem is not met.

Transformation is required.



Here, we use ladder function to find the best transformation.

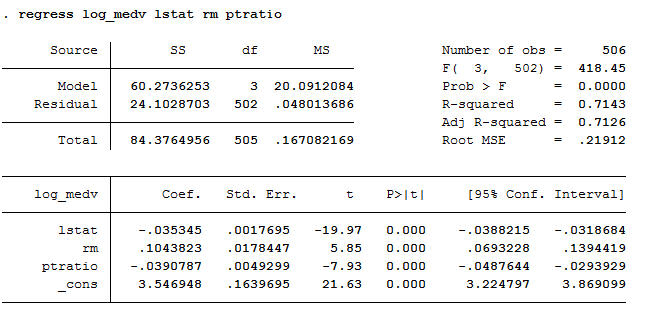
The output suggests log transformation since it has the lowest chi square value.



We check the conditions on normality, residuals.

The graph suggests log transformation is fairly normal and residual is independent.

Therefore, it is a pretty good transformation.



(Final model)

Log(medv) = 3.546948 + -0.035345\*lstat + 0.1043823\*rm - 0.0390787\*ptratio

6.Interpretation

our final model Has log(medv) as the response variable.

lstat, rm, ptration, is our predictors.

All the predictor has p value of 0.

There is 0 % of chance mistakenly rejected the null hypothesis.

Therefore, we reject the null, and conclude that all

predictors have significant relationship with log(medv).

Keep everything constant,

For each unit increase in lstate, there is -0.035345 decrease in log(medv).

For each unit increase in rm, there is 0.1043823 incrase in log(medv).

For each unit increase in ptration, there is -0.390787 decrease in log(medv).

The model has adjusted R square of 0.7126.

Therefore, we know that 71.26% of variance in log(medv) is explained by the model.

7. Conclusion

Our project did not use any advanced statistically technic.

However, the criteria we follow to select the predictor appears to work

well. With only 3 predictors, we could explain the variation of

medv by 71.26%. This suggests that the medina value of owner-occupied homes is

highly depended on average number of rooms per dwelling, lower status of the population,

and pupil-teacher ratio by town.

To explain in detail, the median house price increases when the average number of rooms per dwelling increases.

This is a pretty common phenomenon. The more rooms usually mean a house has higher value.

On the other hand, the median value of owner-occupied home decreases when lower status of the population and pupil-teacher ratio decreases.

For family who is concerned about their children's education and their living environment,

it makes perfect sense that the value of the medina value of owner-occupied home drops.

Overall, this model should make a pretty reasonable prediction on medv.

Since multiple linear regression is limited on its flexibility, this is a good result.

If we want to make a more accurate prediction, more advanced statistical technic is required.