Boston Housing Data Report

Wenxuan Gu

1.Introduction

Boston housing data is collected by Harrison David, JR and Rubinfeld, D.L and it was used in Economics & Management paper 'Hedonic price and demand for clean air' by J. Environ in 1978. The major purpose of this paper is to present the relationship between air pollution and housing market in Boston metropolitan area.

The quantitative housing market approach is applied to investigate households willingness-to-pay on housing market and air quality demand. This research raises the point that households are much more willing to reside in units with less air pollution than more air pollution. Parameter "NOX" is used as the index of air quality. This indicator has a negative effect on housing price which describes nitric oxides concentration in Boston metropolitan area with unit parts per 10 million.

The Boston housing data analysis report in this paper conducts a similar strategy which uses parameters to explore the relationship between housing median price and all potential variables in the provided data set. The goal of this paper is to explore holistic and unseen parameters within the observations in order to fit a good model for the median price of owner-occupied properties in Boston metropolitan area. In other words, this analysis focuses on all given parameters and their effects instead of mainly concentrating on air quality and households willingness-to-pay.

2.Data

This data set is a subset of an original data set which only contains 375 observations in total. Variable nox is one of the parameters that has been applied for Harrison David, JR and Rubinfeld, D.L's research. There is a total of 13 exploratory variables in this data set described as the following order: crime, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, minor, lstat (**Table 1**). The independent variable in this project is medv which presents the median value of owner-occupied homes with unit in $1000's.

Table 1 shows all 14 original variables and their brief descreptives where "crime" is crime rate by town; "zn" is the proportion of residential land over 25.000 sq.ft; "indus" is proportion of non-retail business by town; "chas" is the Charles River dummy variable; "rm" Is the average room per resident; "age" is the ratio of owner-occupied units before 1940; "dis" is measured distances to employment centers; "rad" is the calculated index to highways; "tax" is the full property tax per $10,000; "ptratio" is the pupil-

teacher ratio by town; “minor” is the ratio of minorities by town and lastat is the percentage lower status of the population.

Table 1
Summary statistics for original variables

Variable	Mean	Std. Dev.	Min	Max
medv	25.114	8.817	5.100	50.000
crime	0.7182	1.762	0.006	18.498
zn	15.333	25.950	0.000	100.000
indus	2.609	6.313	0.460	25.650
chas	1.824	0.291	0.000	1.000
nox	1.198	0.109	0.385	0.871
rm	6.369	0.722	3.561	8.780
age	61.980	28.700	2.900	100.000
dis	4.374	2.137	1.129	12.127
rad	5.459	4.565	1.000	24.000
tax	328.699	102.273	187.000	666.000
ptratio	17.858	2.217	12.600	22.000
minor	379.412	41.506	70.800	396.900
lstat	10.487	2.257	6.070	37.970

Notes: All variables contain 375 non-missing observations. All variables are numerical variables expect 1 dummy variable which is chas(= 1 if tract bounds river; = 0 otherwise).

By examining all variables in the given data set, it is clear to see 9 variables that are given as proportion variables which are “crime, zn, indus, nox, age, dis, ptratio, minor and lsatat ,” 3 continuous variables which are “rm tax and medv”, 1 dummy variable which is “chas,” and 1 categorical or index variables which is “rad.” Before stepping to further data analysis phase, it is always valuable to query weather all these variables can be applied and used directly to build future models.

For instance, proportion variables or ratio variables cannot be used directly since their distributions are usually skewed. Dummy variables need to be return to zeros and ones since pure texts are not helpful for analysis. Index variables might also need transformations, but it depends on situations and their actual effects to model generation.

Specifically, we need to explore out the independent variables that have the potential to fit, the independent variables that violet certain statistical rules during the process and the ones that need to be transformed before using them.

3.Analysis & Findings

As stated in the previous paragraph, the goal of this project is to apply the given variables to successfully generate a model which fits sufficient variables for independent variable medv. The following section will be focusing on the specific methods and process that used in this project. A total of 8 general steps are presented as the following:

3.1 Import

Importing the original data set is the first step that enables us to examine the data we have. By running "proc print" in SAS, we are able to load the given "boston.csv" file to SAS system. By examining all the observations in this step, 375 observations are included in this dataset and there are 14 variables in total including independent variable medv. By verifying it again with the given project description, none missing observations are found in this phase and we rename it to bos_house before proceeding to the exploratory step.

3.2 Exploration

The purpose of proceeding exploration is to check and verify the essential data statistics from given data set in order to acknowledge each predictor in a quantitative way. Thus, we apply "means" to show all essential descreptives such as mean, standard error, quantiles, minimums, and maximums. This allows us to see how these variables present in bos_house dataset. Frequency checking might not be as essential as visualizing the descreptives, but it provides a way of seeing each variable separately. Scatter plot and correlation plot are generated during this phase to present a brief view of how each independent variable interacting with response variable medv.

By looking at the correlation matrix, we see that variable rm has the strongest positive association with medv and dis has the weakest positive association with medv. Variable lstat has the strongest negative association with medv and crime has the weakest negative association with medv. However, these findings are explorations at the very beginning exploration phase, they should only be considered as necessary findings instead of any conclusion for model selection and analysis. By graphing out sgscatter plot out, we can also see that even though some variables present relatively strong or weak associations

to response variable, majority of the independent variables are non-linear corresponding to response. By checking the pattern more specifically, gplots are presented on each variable to visualize the linearity pattern with respect to response variable medv. The same result from gplots provides a hint that proportion variables will need to be transformed in order to behave normally in a linear regression. This means that if regressors do not present a linear pattern, then they must be skewed in their histograms. As we predicted, all of these proportion variables are either left skewed or right skewed. Taking necessary actions to fix their skewness is important and should be proceeded in the next step.

3.3 Transformation

In the transformation phase, a new dataset named "bos_analysis" is created that includes all original predictors plus transformation variables and three interaction terms. As we discussed in section 3.2, variables with proportions present skewness and all need to be fixed in an optimal way. That is by transforming them of applying the most famous "log" transformation. Therefore, new log variables ln_crime, ln_zn, ln_indus, ln_nox, ln_age, ln_dis, ln_minor are created by applying log on them. By running histograms again to check, the result of "log" transformation fixes most proportion variables with skewness issues. Additionally, the unusual variable zn applies a slightly different log transformation method. The reason is because zn is a variable that has more than 200 zero observations. By only taking a "log" of variable zn would not solve any problem because log(0) has "undefined" solution. In order to avoid missing observation issues, we can add a small value "c" to itself and make zero observations none-zero values. By doing this, it not only solves the issue of missing observation but also does not modify the original variable much since 0.001 is a relatively small value. So that we have ln_zn written as log(zn + 0.001). Eliminating this variable is a way but that means we will lose a variable in the future and taking we don't expect this offset to happen in our dataset.

The reason that lstat needs to be transformed using "quadratic" method is because its gplot present a quadratic pattern and the best way to solve it is to make it a square variable. As we observed in the correlation plot in section 3.2, we might be able to create few interaction terms for future use. The reason that rm is selected to combine with variable chas, lstat and ptratio is because rm has a linear relationship with response variable and rm has the strongest positive effect as shown. In addition, using rm to combine with other variables such as dis or age can also be interesting but picking chas, lstat and ptratio does not have a specific reason. A summary of transformed new variables and the original variables are presented in **Table 2** which has shown below.

Table 2
Descreptives for transformed & non-transformed variables

Variable	Mean	Std. Dev.	Min	Max
medv	25.114	8.817	5.100	50.000
ln_crime	-1.663	1.537	-5.064	2.918
ln_zn	-3.164	5.041	-6.908	4.605
ln_indus	1.916	0.756	-0.777	3.245
chas	0.093	0.291	0.000	1.000
ln_nox	-0.674	0.189	-0.955	-0.138
rm	6.369	0.722	3.561	8.780
ln_age	3.965	0.653	1.065	4.605
ln_dis	1.350	0.516	0.121	2.495
rad	5.459	4.565	1.000	24.000
tax	328.699	102.273	187.000	666.000
ptratio	17.858	2.217	12.600	22.000
ln_minor	5.928	0.174	4.260	5.984
lstat2	147.071	191.325	2.999	1441.720

Notes: All variables contain 375 non-missing observations. All variables are numerical variables expect 1 dummy variable which is chas(= 1 if tract bounds river; = 0 otherwise).

3.4 Reassemble

Reassembly is a phase that is designed in this project for general review before validation and model generation phase. This is a step that keeps us on track and provides a clear view of what variables we have and what variables that we need to use for future steps. Therefore, a new dataset “bos_transform” is created here that contains all important variables we created in section 3.3 and repetitive variables are dropped out here which are crime, zn, indus, nox, age, dis, minor, lstat. The transformation of these variables are kept in this phase. By re-check all the variables we have now for future use, we have medv as our response, in dependent variable ln_crime, ln_zn, ln_indus, ln_nox, rm, ln_age, ln_dis, rad, tax, ptratio, ln_minor and lstat2. So that we have 14 variables in total including response medv and 8 of them are transformed in section 3.2.

3.5 Validation (train & test sets)

The validation process is the most important step before any model generation. Without model validation process, data and variables might not be used appropriately and correctly. Thus, two methods are implemented during the process in order to make sure that we will be using the appropriate variables and data for future use. The first method is by applying the fundamental "surveyselect" method. It is a way that splits data to train and test set with a ratio of 75% and 25% in this project based on prior generated dataset bos_analysis. After splitting two datasets, we create a new dataset named "var_all" for future use. In this process, the main reason of creating new dataset is to generate a new response variable "new_y" for our training set. The value of variable new_y is equal to original response variable medv. This means that any observations that are chosen by "surveyselect" are marked ad ones in the first column which is "selected." Each selected observation(training set) has their own response which is variable new_y. A full model that includes all 13 verified and transformed variables plus 3 interaction terms generated in section 3.2 are used for the following full model test. By choosing stepwise selection and fitting all full variables in. It is clear to see in the result that 281 observations are used in this process and 94 observations are the missing value ones. This is correct because we always need to check if we are using training set to generate models during the validation process. The R-Sq value is 0.8144 and the Adj R-Sq is 0.8075 which means that this full model explains almost 80% of the variation based on the training set that we use. A total of 10 independent variables are picked by stepwise selection and all of them have significant values less than 0.05 besides variable ln_crime.

After setting up and using our training set, the next step is to create new response variable for testing set and compute their performance. For our testing set, response variable yhat is created which has the equal value to medv and new_y as well. By processing the yhat, we are able to see that a total of 93 "NA" testing set observations are assigned with values. By doing two of these two important setups, we are finally able to compare them and compute the performance of our testing set. The rmse value is 3.32987 and the mae value is 2.39367 for our test set which are lower than 3.84371. By using the correlation, we can calculate that our testing set has R-Sq (0.92552^2) equal to 0.8565 which is 5% higher than what we have for our training set. By doing survey select method, the result meets our expectation and shows that training set performs better than training set.

*medv =ß0 + ß1rm +ß2rm*lstat + ß3ln_dis + ß4lstat^2 + ß5rad + ß6ln_nox +*

*ß7rm*ptratio + ß8ptratio +ß9tax + ß10ln_crime+ u (M1)*

As stated in the beginning of 3.5, the other method is also applied here which is the “glmselect” method or “cross-validation” method. This method is a more automatic procedure because it combines dataset splitting and model generating steps together and directly shows the combined output. Thus, two different selection methods (stepwise and backward) are both applied here to strengthen our model validation process. By using stepwise cross-validation, 9 variables are presented in total and ln_cirme is dropped here because it is not statistically significant. By checking the difference between ASE(train) and ASE(test), it shows the difference is less than 1 which is ideal because we always expect the difference between them to be as close as possible. The F values is 138.71 which is not bad to a dataset that has a total of 288 training observations. The R-Sq value is 0.8179 and Adj R-Sq is 0.8120 which are similar to the result we get by using “surveyselect” method.

medv =ß0 + ß1ln_nox +ß2rm + ß3ln_dis + ß4rad + ß5tax + ß6ptratio +

*ß7lstat^2 + ß8rm*lstat +ß9rm*ptratio + u (M2)*

By using backward cross-validation, a total 12 variables are presented in the result except the intercept. Variable, ln_zn, rm_chas, ln_minor and ln_crime are dropped at this procedure based on 280 training observations. ln_indus is calculated to be a potential candidate for removal. By checking the difference between ASE(train) and ASE(test), the difference is also small which is only 1.18. The F values in this case is 107.87 which is 30.9 lower than using the “stepwise” cross-validation. The R-Sq value is 0.8290 and Adj R-Sq is 0.8213. They are about only 1% to 2% higher than the result of “stepwise” validation but with additional 3 regressors which indicates. By checking its goodness of fit value again, this indicates that even though these additional variables raises the R-Sq by a very small amount, but they might not become good regressors to fit in the model. By adding them in, we lose 30.9 goodness-of-fit. Thus, we might need to consider dropping these variables for our final model generation although they explain a little more variation in general.

medv =ß0 + ß1ln_indus +ß2chas + ß3ln_nox + ß4rm + ß5ln_age + ß6ln_dis +

*ß7rad + ß8tax +ß9ptratio + ß10lstat^2 + ß11rm*lstat + ß12rm_ptratio + u (M3)*

3.6 Model Generation

By doing the previous validation procedure in section 3.5, we see that additional variables cause significant reduction of F-value although they produce a small amount of increment in M3. Variable ln_cirme is presented in all three validation processes as an insignificant variable and needs to be

dropped. In general, by summarizing the result of M1 and M2 with two different validation methods, we are able to pick out a propriety fitted mode as the following:

*medv =ß0 + ß1rm +ß2rm*lstat + ß3ln_dis + ß4lstat^2 + ß5rad + ß6ln_nox +*

*ß7rm*ptratio + ß8ptratio +ß9tax + u (M0)*

By generating the regression of this model, we see that a total of 374 out of 375 observations are used and the goodness-of-fit has a value of 192.02. The R-Sq and Adj R-Sq are 0.8260 and 0.8217 which means this model explains over 82% variation of this dataset in general. The RMSE value is 3.70198 which is similar to the training sets where we calculated in section 3.5. By checking the p-value throughout all generated variables, we see that all variables satisfy the significance level check, and all have p-values less than 0.0001. By checking the standardized estimate, we see that rm is the variable with the strongest influence.

The concern occurs here by checking variance inflation that variable rm, rm_ptratio and ptratio are over the threshold 10. However, the cause of this problem is reasonable. If we recall the transformation process in section 3.3, we see that rm_lstat and rm_ptratio are two interaction terms. This means that two interaction terms are strongly correlated to rm, and it is acceptable that all three interacted variables have variance inflation above the threshold. To verify this fact, we eliminate two interaction terms rm_lstat and rm_ptratio variables one by one. The collinearity effect goes away as we proceed the verification. Thus, we are able to neglect this collinearity issue caused by interaction terms. At the same time, we also see that the R-Sq drops quickly from over 82% to 75% by checking both R-Sq and Adj R-Sq after two steps of interaction terms elimination. The F-value also drops dramatically from 192.02 to 162.05 which tells us that two interaction terms has a relatively strong influence towards the generated model.

3.7 Cleaning

The questions comes that is there any way that we can make the model we have a little better. With R-Sq and Adj R-Sq over 82% and goodness-of-fit over 190, this model can be considered as a descent regression to predict our response. However, there are still flaws that can be fixed which means we need to check for model diagnostics and assumptions to clean up our dataset and improve.

Therefore, we check student residual plots, predicted residual plots and normal cumulative distribution in this step to visualize how each variable behaves in this model. We see in the appendix result that at

least over 5 influential points and outliers can be found on the residual plot on each variable. This causes variables assumption: constant variance, independence and linearity might not satisfy although the model might still fit. For instance, variable ln_dis and ptratio have suspicious patterns and have some evenly distributed residuals which lead to unsatisfied model assumptions. The most important finding that we need to be aware is the graph of CDF student residual plot. It shows a "S shaped" curve in terms of normality and should be fixed by eliminating all outliers and influential points. Therefore, we need fix this issue by doing the following two steps: automatically visualizing the influential points and outliers and automatically pick out these observations.

The metric used here is by checking CookD to measure the observations that are greater than 4 divides by overall 374 used observations and pick out influential points. By checking whether Rstudent value over or equal to 3, we are able to pick out all potential outliers. By performing these two steps together of investigation, we successfully pick out 20 observations. Some of these 20 observations are both influential points and outliers in bos_analysis dataset. These variables are deleted and cleaned dataset named clean_bos_analysis is created for further analytics use. The result of cleaning this dataset is ideal based on the new statistics: F-value, RMSE, R-Sq and Adj R-Sq. The cleaning procedure keeps 355 observations in total, and the goodness-of-fit shows a dramatic improve from 192.02 to 344.50 which is 1.8 times better than the uncleaned bos_analysis dataset. The significance level on all variables remain great condition. The RMSE decreases from 3.70198 to 2.56348. The most surprising result is that the R-Sq improves another 8% from 0.8260 to 0.9001. The Adj R-Sq is 0.8975 which is relatively high comparing to any of the Adj R-Sq values in previous regression models. By summarizing these regression statistics, we might conclude that this is a reliable model that has a high goodness-of-fit value and explains 90% variation based on clean_bos_analysis dataset.

To make our finding more concrete and convincible, student residual plots on all regressors are generated again for model assumption check and Normal Cumulative Distribution plot is generated to visualize the improvement of normality. Based on appendix results, the improvement can be visualized clearly. By comparing with uncleaned data, we see that student residual and predicted plots on each variable are scattered much more randomly around the 0 residual line. By checking the CDF of Studentized Residual, we see a significant improve based on the shape of it. The result shows an ideal 45 degree "straight line" that has nearly no curves.

*medv = -87.165 + 20.731rm – 0.1575rm*lstat – 4.3902ln_dis + 0.0179lstat^2 + 0.4023rad – 8.3411ln_nox – 0.7655rm*ptratio + 4.2861ptratio – 0.0132tax + u (M0)*

3.8 Prediction

The last step for testing whether this model fits well to predict our response is by applying it with values. This model prediction procedure enables us to apply reasonable values on regressors in our model to generate out a response value. Thus, three different prediction scenarios and results are generated in the appendix.

The first prediction has the scenario as the following:

rm = 5 rm_lstat = 20 ln_dis = 0 lstat2 = 0 rad = 5 ln_nox = -0.5 rm_ptratio = 100 ptratio = 21 tax = 0

The way to interpret it correctly is that a house in Boston with 5 rooms, index of accessibility to highway is equal to 5, the nitric oxides concentration is -0.005 /parts per 10 million, the interaction term between rm and ptratio is 100, and pupil-teacher ratio is 21. By using this scenario, our predicted value is $32.9817K for a property with such features.

The third prediction has the scenario as the following:

rm = 6.3 rm_lstat = 75.7 ln_dis = 0 lstat2 = 155.9 rad = 8 ln_nox = -0.67 rm_ptratio = 96.31 ptratio = 15.4 tax = 0

Different numbers are substituted in this situation, and we set lstat2 equal to 155.9 in this case to see if there would be any difference. The result shows that a house in Boston with such features has a predicted value of $35.4037K back in 1970s or before.

As we can see that as tax and weight distance remain unknown, we are able to include that property in Boston metropolitan area can be generally reflected through these predictors and some of them play very important roles of predicting market house value. As we can see from the result in a quantitative way that a house with more rooms, higher nitric oxides concentration (nitric oxides are chemicals that are beneficial to human body and the lack of it might cause diseases and health problems), lower percentage of population, higher accessibility to highways tend to have a higher price in market back in 1970s. If we combine these prediction 1 and 2 together, we can also read the result as: a house in Boston with 1 more room, 12.4% lower minority population than the other, medium level of accessing to highway, 0.17% nitrogen oxides higher and pupil-teacher rate is a 6.6% lower has a prediction value of $2.422 higher value in 1970s.

However, everything seems correct when using such a model to make median house value predication. The second prediction is set as a scenario that takes out two major regressors which are variable rm and ln_nox. The reason for doing it is to test whether M0 is still able to predict a relevant result. The result in appendix of our second prediction is a suspicious and can be investigated further.

The second prediction has the scenario as the following:

rm = 0 rm_lstat = 30 ln_dis = 2 lstat2 = 100 rad = 24 ln_nox = 0 rm_ptratio = 120 ptratio = 18 tax = 100

As we see from this scenario, all other variables are assigned with values besides rm and ln_nox. The prediction result for medv in this case is $-107.9056 and this is something we might not expect to happen. By reviewing our **Table 2**, we see that the maximum of Boston housing price in this dataset is $50,000 and this value is twice as large as our maximum for response variable with a negative sign. It is clearly not correct but what does it tell us? By viewing our model once again, we might find from this prediction result that this becomes the downside of our model. M0 has its intercept with value equal to -87.165 which means by the sum of all regressors will need to offset it with an approximate value of 130 to 140 in order to calculate out our predicted value.

Without adding the major regressors, our prediction result will be off by a lot since other variables in this model might not be able to contribute as much effects as rm and ln_nox. By relating to Harrison David, JR and Rubinfeld, D.L's research, we recall that their research mainly concentrates on air quality and housing price so that nitrogen oxides percentage would definitely play an important role of housing price prediction. The other regressor rm presents the average room of a property. By relating to the ground truth of real estate price, we also recognize that a property with more rooms has high square feet which leads to higher price. Thus, we do need to include rm and ln_nox when predicting median price of houses in Boston metropolitan area. Without them, we will not be able to generate a correct and accurate result that we expect.

4.Future Work

Additional avenues that worth exploring based on what we have done would be model application, in other words, it can be understood as how good is our model that predicts not only the median price of house in Boston metropolitan area but also other cities housing price. The transformation that we have done in this project is based on a standard metric where logarithm transformation are used on proportion-based variables. Even though our model shows a relatively satisfying result, we still lose some interesting predictors in the end. For instance, variables such as ln_indus, ln_zn, ln_crime are all

eliminated during the process due to their insignificance. However, we know that factors such as crime rate, ratio of non-retail business acres or residential land acres might also be related to housing price, but they are not included in our final model(M0). The assumption might be that Boston has a low crime rate in 1970s, most industrial centers are in suburban areas and not every household has much land in metropolitan area. These assumptions might cause that variables as such are neglected in model generation of predicting house values. However, does that mean such variables are useless and unusable? The answer is not "probably not" because the model we build only applies to a very small amount of variation of data. If we step back, we reconsider whether it is applicable on Great Chicago area or Downtown Manhattan back in 1970s, it would be a whole different story. For example, crime ratio or minority group ratio will probably make a huge impact just on Chicago housing price. Thus, a more universal model might be generated based on this study in order to apply on different datasets as a further study. This indicates that some variables might need to be recalculated before use.

The current result shown is relatively good. It has 9 predictors without intercept and has an R-Sq of 90%. If we keep working on what we have, we might need to focus on dummy variables. That means we will conduct future exploration on dummy variable chas (Charles River dummy) and set it as our logistic response variable. By viewing the location of Charles River, it is only 1.2 miles from city central and the most famous college town Cambridge. A future analysis can be conducted by fitting variables to a classification model to investigate houses with certain features and median values are how much close to Charles River. That is, by providing such features and proportion values based on given data set, the logistic model is able to predict their locations in a quantitative way.

There are two research materials helped me solving obstacles in this project. The first article is named "How should I transform non-negative data including zeros?" posted on Stack Exchange. Is specifically discussed the methods of the transformation of non-negative data. The reason that this method helps a lot because variable zn has a lot of zeros in the dataset as stated in section 3.3, so that by only taking it with logarithm is not helpful and this will get us more than half missing observations. With highly skewed non-negative data, an alternative method is stated in this material that log(x + c) can be applied for transformation and c can be set with a small positive value. By applying such method, we prevent variable zn causing missing value problem in section 3.3.

Another article is named "Interpreting Log Transformation in a Linear Model" from University of Virginia Library. It specifically explains the rules for interpretation log-transformed variables which I included in section 3.8. A total of 8 variables are generated in section 3.3. In order to understand each log variable

better, we need to interpret them correctly. As this article stated, any log transformation independent variables can be interpreted that as 1% increase of independent variable, there is a (coefficient/100) increment to independent variable. This is different from interpreting log-transformed dependent variable.

5.Research Discussion

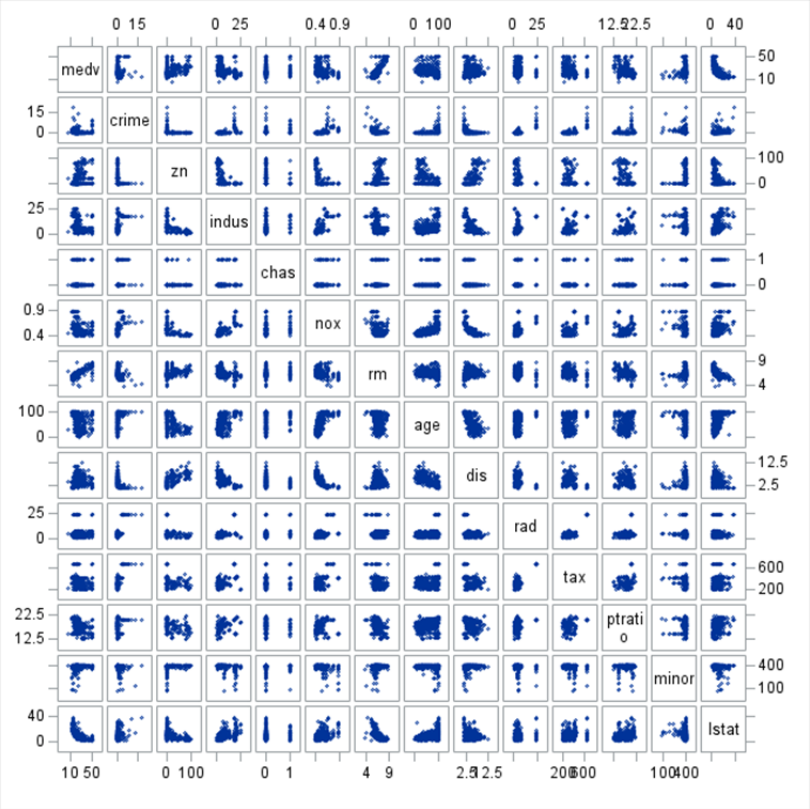
By reading the article "How to use regularization to prevent model overfitting," it brings a great concern that what data scientists need to be aware while building models. It points out that "overfitting" can become an issue of fitting complex parameters and large data sets. In general, the mechanism "shrinkage" is applied to solve such problems which zeros out some parameters, but this becomes an offset which loses sufficient amount of useful information. Therefore, using "shrinkage" to solve such problems might not always be a good idea in practice. Instead, by applying "regularization" correctly will strongly help use to avoid "overfitting" problems. Even though we do not have more than 3 high power or complex parameters in this project, that does not mean overfitting will not be concerned in relevant models or future studies. The type of "regularization" can be introduced as the articles stated including "using a simpler model," "applying subsample of features" or even combine both of them. These methods can be applied as soon as we enter to model-validation phase.

Taking a macro-view on the project that we are working on, we also see that there are many approaches that can be improved. The most important aspect is "data transformation and predictive analysis." As the article "Predictive Interaction for Data Transformation" specifies, we learn that "human work involved data transformation represents a major bottleneck in nowadays." This means that the so - called transformation phase and the methods being used that we present in this project is not efficient and optimal for data transformation. The method we apply to the parameters based on the given dataset is by taking logarithm, making quadratic and interaction terms. However, these steps that we perform are all based on our domain knowledge, humans understanding of this dataset and fundamental rules of statistics. In other words, these elements are the guides that helped us but none of them are guaranteed accurate or efficient. If we recall to the steps, we did in section 3.3, obstacles occurred on transforming variable "zn" since it contains more than half zeros. After 3 times adjusting and modifying the logarithm method, we eventually applied log(x + c) to remove the problem of missing observation issues. Imaging that a data with more than millions or thousand millions of observation and hundreds of features, human work will never be possible to apply any reliable techniques on data

transformation. Thus, the method named “DSL(compilation, disambiguation, prediction, interaction, and visualization” is introduced in this paper which automatically and intelligently list and predict the potential transformation result. The most successful example is Tableau since it becomes one of the systems that has the ability to not only manipulate data transformation but provide visualization of transformed data.

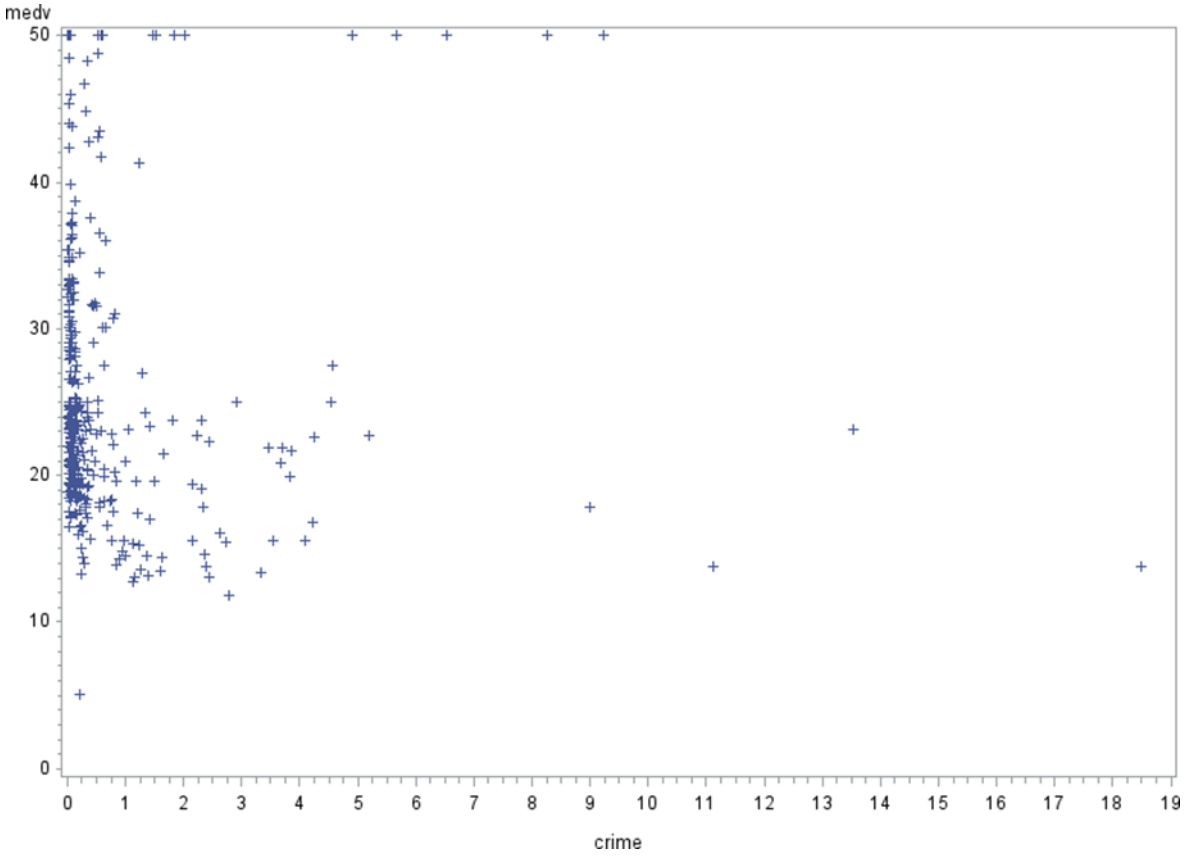
The more advanced topic that is related to this project is a hot topic for decades called “Zero-shot learning (ZSL)” This methodology is more related to the prediction phase that we did in the last part of this project. After the generation of our final model, predictions on response variable was made but the second prediction result was a lot off of what we expected. This issue might be solved and understood by the application of ZSL. The article “Cluster-based zero-shot learning for multivariate data” provides a comprehensive view of the logic of ZSL and its applications. When solving real-world problems, the circumstances are not as ideal as we dela with the given dataset which means we will not be able to have training data or test data in practice. ZSL as introduced as a cluster-based method that uses limited features to generate clustering for solving unseen classes in predictions. This method is largely applied in bioinformatics and medical fields. If recall to the second prediction in section 3.8, we might realize that without two major parameters in our model, it is still possible to generate a reasonable result by applying ZSL. More specifically, by classify these observations to different clusters based on 374 observations are used in this regression, we are able to generate different clusters based on their features and automatically reconstruct the model. By doing so, we can not only predict data within the range but also making reliable predictions on unseen class observations.

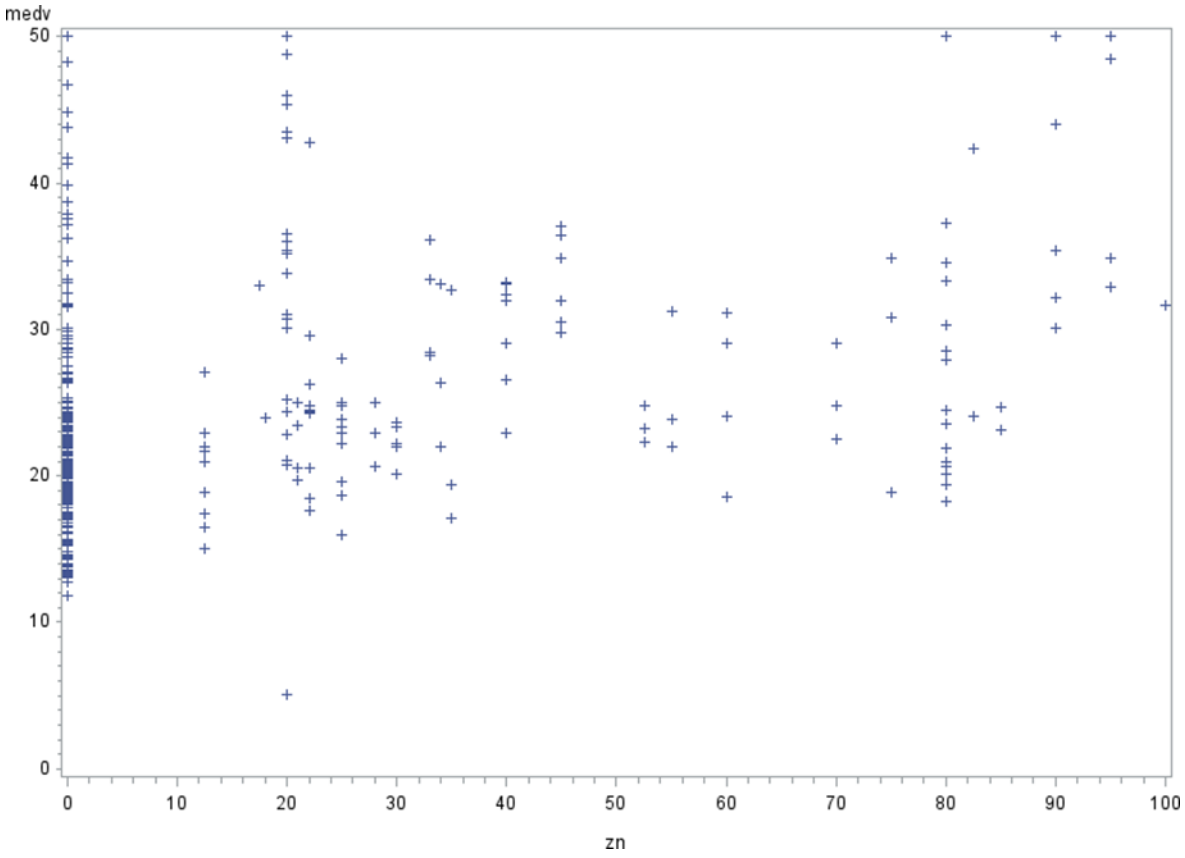
Appendix

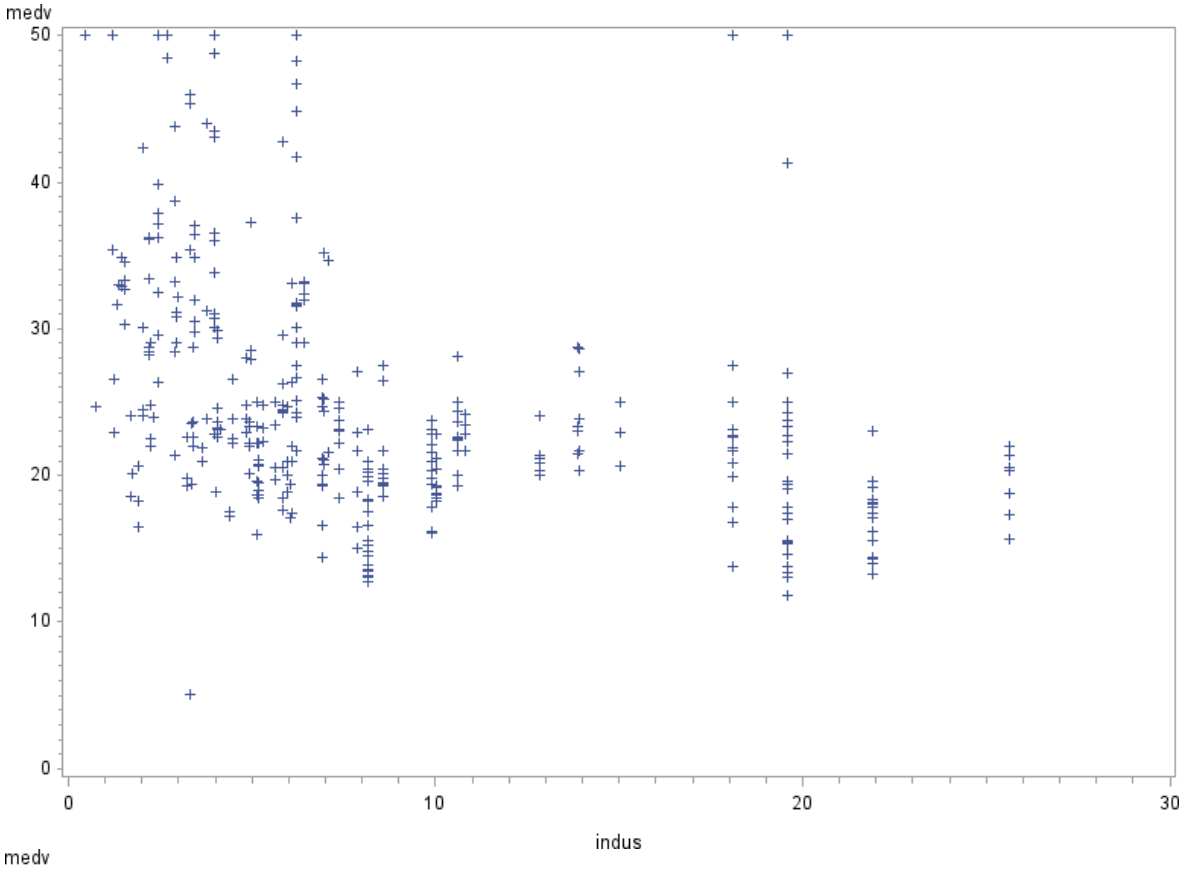


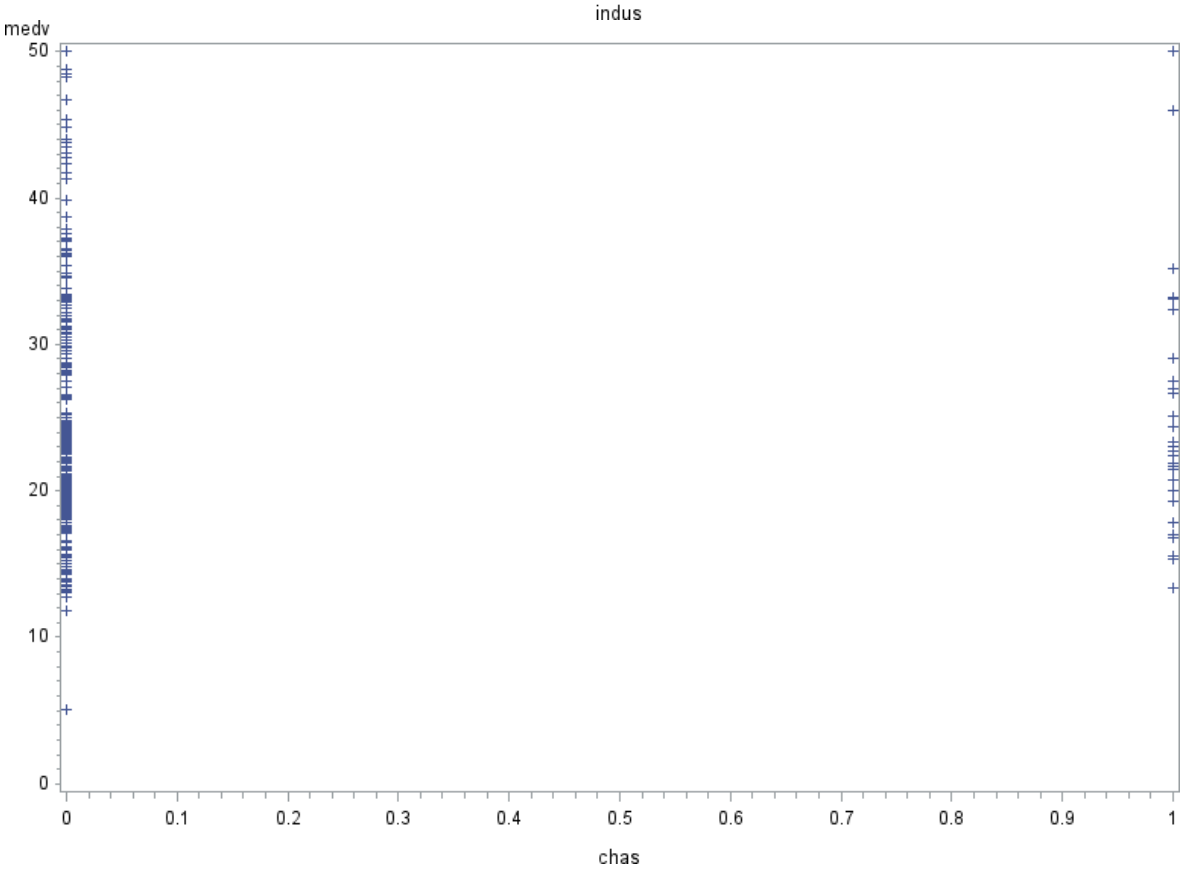
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

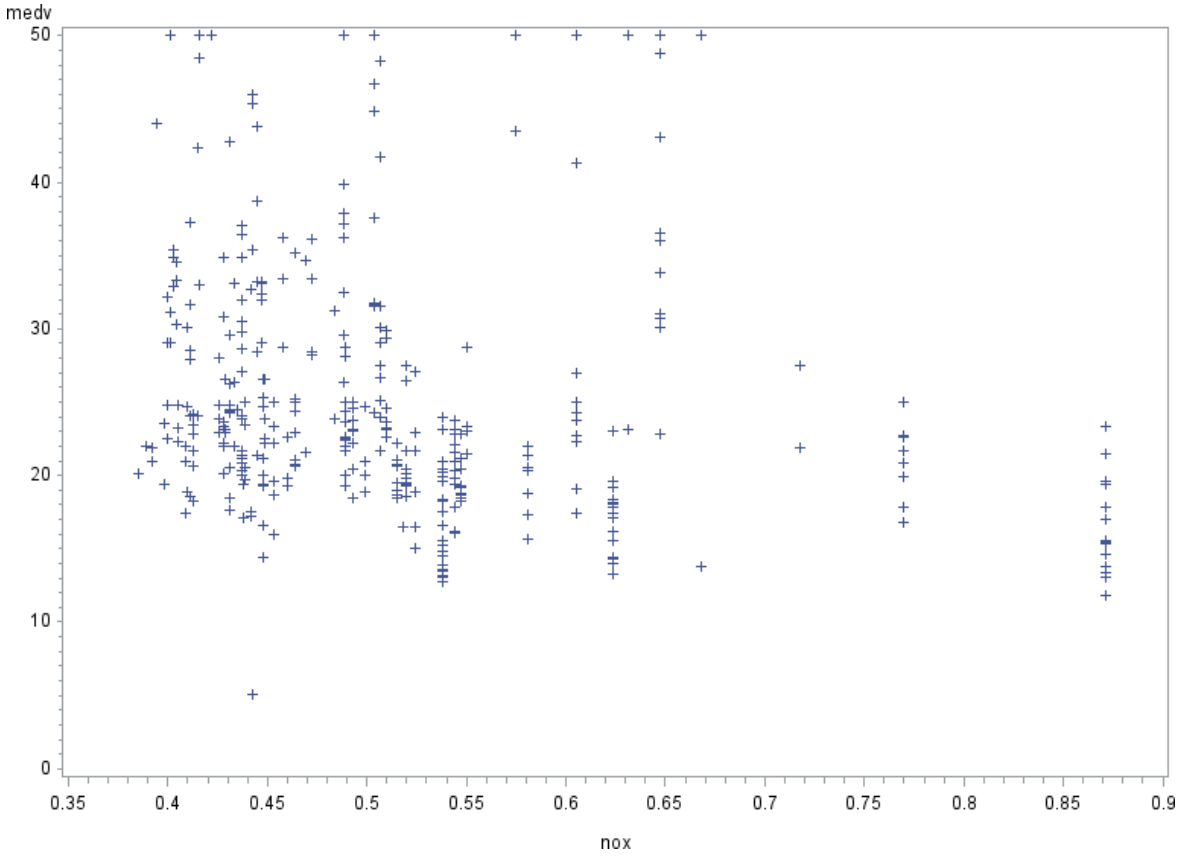
	medv	crime	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	minor	lstat
medv	1.00000 375	-0.01211 0.8151 375	0.27579 <.0001 375	-0.27633 <.0001 375	0.12118 0.0189 375	-0.19087 0.0002 375	0.76112 <.0001 375	-0.18680 0.0003 375	-0.00354 0.9455 375	0.10148 0.0496 375	-0.10071 0.0513 375	-0.36822 <.0001 375	0.16952 0.0010 375	-0.65017 <.0001 374
crime	-0.01211 0.8151 375	1.00000 375	-0.21399 <.0001 375	0.45386 <.0001 375	0.20657 <.0001 375	0.53360 <.0001 375	-0.31044 <.0001 375	0.37829 <.0001 375	-0.39877 <.0001 375	0.75936 <.0001 375	0.69373 <.0001 375	0.11588 0.0248 375	-0.30261 <.0001 375	0.32017 <.0001 374
zn	0.27579 <.0001 375	-0.21399 <.0001 375	1.00000 375	-0.48816 <.0001 375	-0.09433 0.0681 375	-0.47234 <.0001 375	0.29961 <.0001 375	-0.54172 <.0001 375	0.63397 <.0001 375	-0.19114 0.0002 375	-0.16831 0.0011 375	-0.30445 <.0001 375	0.14393 0.0052 375	-0.37513 <.0001 374
indus	-0.27633 <.0001 375	0.45386 <.0001 375	-0.48816 <.0001 375	1.00000 375	0.19753 0.0001 375	0.71166 <.0001 375	-0.38523 <.0001 375	0.56979 <.0001 375	-0.63166 <.0001 375	0.32548 <.0001 375	0.51616 <.0001 375	0.16211 0.0016 375	-0.31283 <.0001 375	0.48465 <.0001 374
chas	0.12118 0.0189 375	0.20657 <.0001 375	-0.09433 0.0681 375	0.19753 0.0001 375	1.00000 375	0.21678 <.0001 375	0.06723 0.1939 375	0.17373 0.0007 375	-0.20220 <.0001 375	0.27133 <.0001 375	0.18081 0.0004 375	-0.05313 0.3048 375	-0.04966 0.3376 375	0.03980 0.4429 374
nox	-0.19087 0.0002 375	0.53360 <.0001 375	-0.47234 <.0001 375	0.71166 <.0001 375	0.21678 <.0001 375	1.00000 375	-0.27901 <.0001 375	0.68604 <.0001 375	-0.72091 <.0001 375	0.41944 <.0001 375	0.53015 <.0001 375	-0.07488 0.1479 375	-0.39915 <.0001 375	0.45612 <.0001 374
rm	0.76112 <.0001 375	-0.31044 <.0001 375	0.29961 <.0001 375	-0.38523 <.0001 375	0.06723 0.1939 375	-0.27901 <.0001 375	1.00000 375	-0.20323 <.0001 375	0.12517 0.0153 375	-0.15020 0.0036 375	-0.26861 <.0001 375	-0.34207 <.0001 375	0.20477 <.0001 375	-0.64301 <.0001 374
age	-0.18680 0.0003 375	0.37829 <.0001 375	-0.54172 <.0001 375	0.56979 <.0001 375	0.17373 0.0007 375	0.68604 <.0001 375	-0.20323 <.0001 375	1.00000 375	-0.70018 <.0001 375	0.27608 <.0001 375	0.34772 <.0001 375	0.09062 0.0797 375	-0.23857 <.0001 375	0.52991 <.0001 374
dis	-0.00354 0.9455 375	-0.39877 <.0001 375	0.63397 <.0001 375	-0.63166 <.0001 375	-0.20220 <.0001 375	-0.72091 <.0001 375	0.12517 0.0153 375	-0.70018 <.0001 375	1.00000 375	-0.29710 <.0001 375	-0.33734 <.0001 375	-0.01774 0.7321 375	0.23309 <.0001 375	-0.34165 <.0001 374
rad	0.10148 0.0496 375	0.75936 <.0001 375	-0.19114 0.0002 375	0.32548 <.0001 375	0.27133 <.0001 375	0.41944 <.0001 375	-0.15020 0.0036 375	0.27608 <.0001 375	-0.29710 <.0001 375	1.00000 375	0.76603 <.0001 375	0.22047 <.0001 375	-0.12609 0.0145 375	0.07366 0.1551 374
tax	-0.10071 0.0513 375	0.69373 <.0001 375	-0.16831 0.0011 375	0.51616 <.0001 375	0.18081 0.0004 375	0.53015 <.0001 375	-0.26861 <.0001 375	0.34772 <.0001 375	-0.33734 <.0001 375	0.76603 <.0001 375	1.00000 375	0.18418 0.0003 375	-0.24978 <.0001 375	0.20695 <.0001 374
ptratio	-0.36822 <.0001 375	0.11588 0.0248 375	-0.30445 <.0001 375	0.16211 0.0016 375	-0.05313 0.3048 375	-0.07488 0.1479 375	-0.34207 <.0001 375	0.09062 0.0797 375	-0.01774 0.7321 375	0.22047 <.0001 375	0.18418 0.0003 375	1.00000 375	0.07069 0.1719 375	0.20479 <.0001 374
minor	0.16952 0.0010 375	-0.30261 <.0001 375	0.14393 0.0052 375	-0.31283 <.0001 375	-0.04966 0.3376 375	-0.39915 <.0001 375	0.20477 <.0001 375	-0.23857 <.0001 375	0.23309 <.0001 375	-0.12609 0.0145 375	-0.24978 <.0001 375	0.07069 0.1719 375	1.00000 375	-0.20030 <.0001 374
lstat	-0.65017 <.0001 374	0.32017 <.0001 374	-0.37513 <.0001 374	0.48465 <.0001 374	0.03980 0.4429 374	0.45612 <.0001 374	-0.64301 <.0001 374	0.52991 <.0001 374	-0.34165 <.0001 374	0.07366 0.1551 374	0.20695 <.0001 374	0.20479 <.0001 374	-0.20030 <.0001 374	1.00000 374

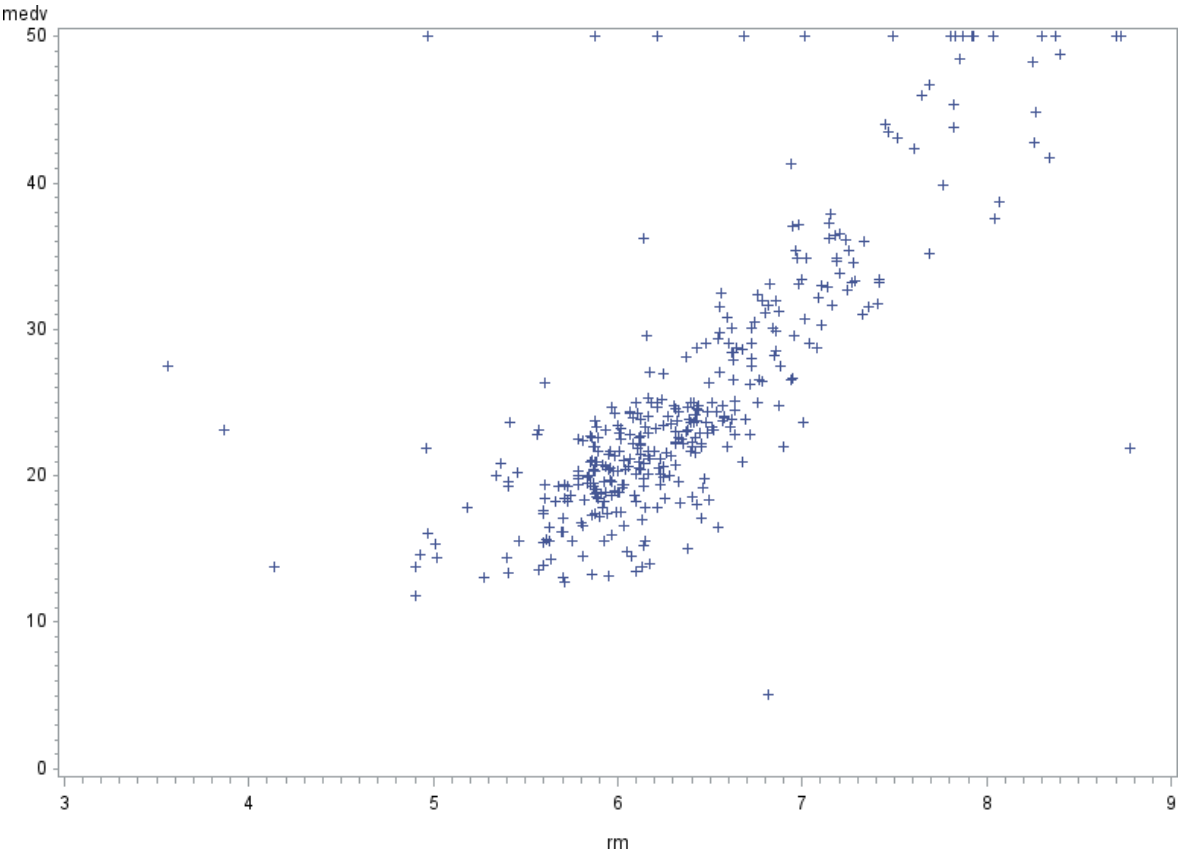

medv
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
crime

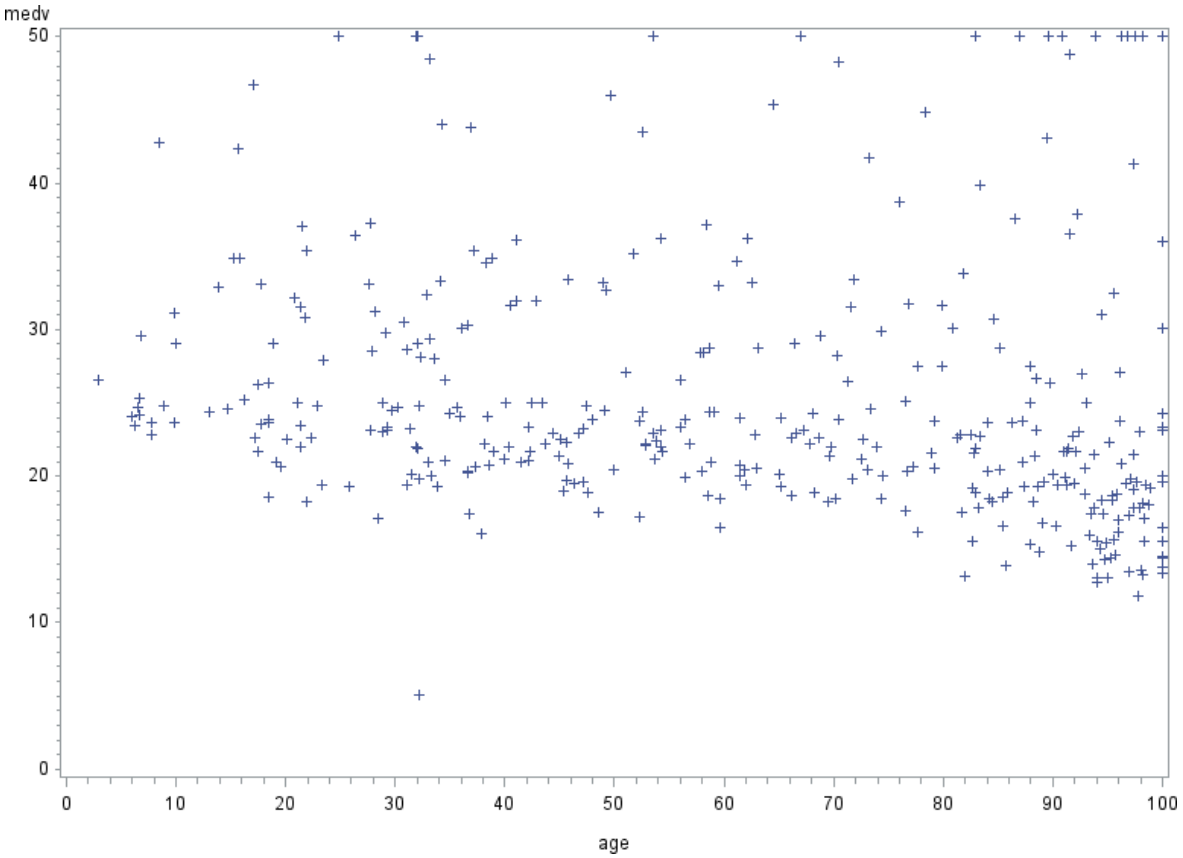

medv
50
40
30
20
10
0
0 10 20 30 40 50 60 70 80 90 100
zn

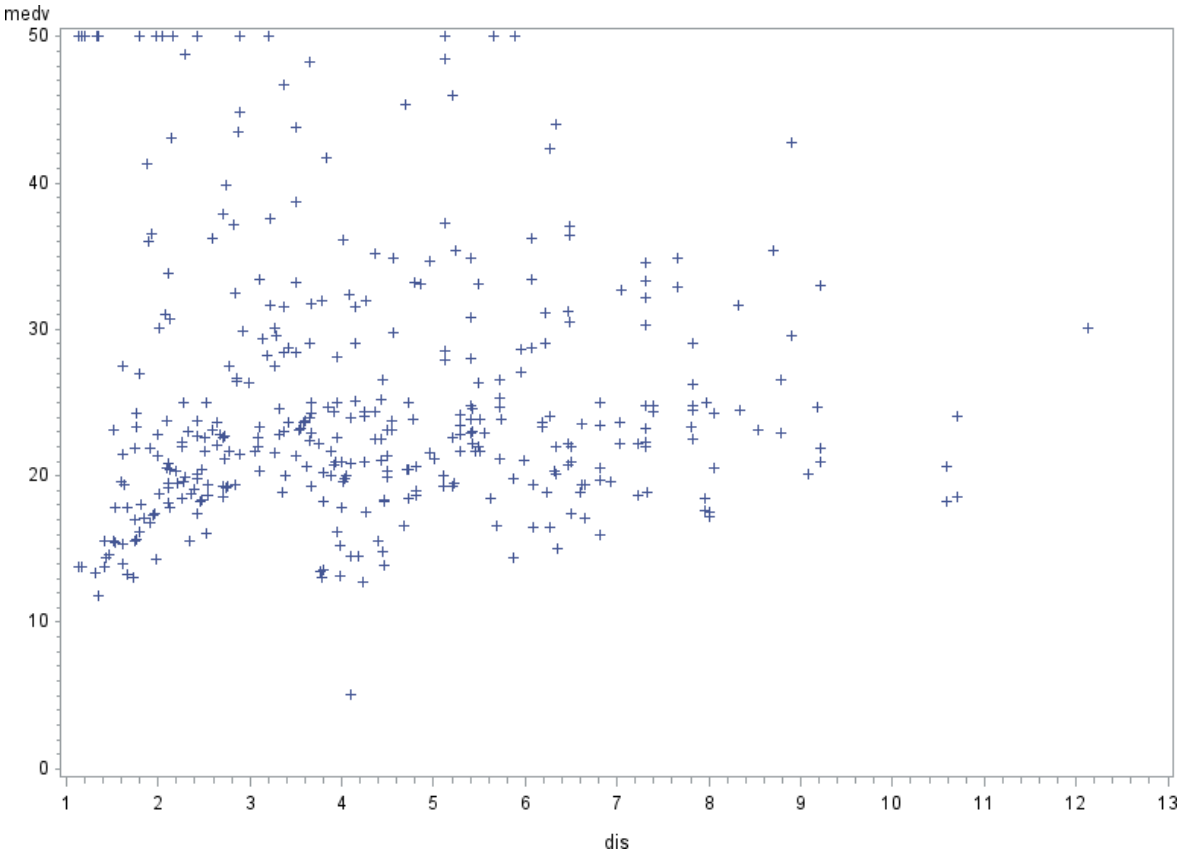

medv
50
40
30
20
10
0
0
10
20
30
indus

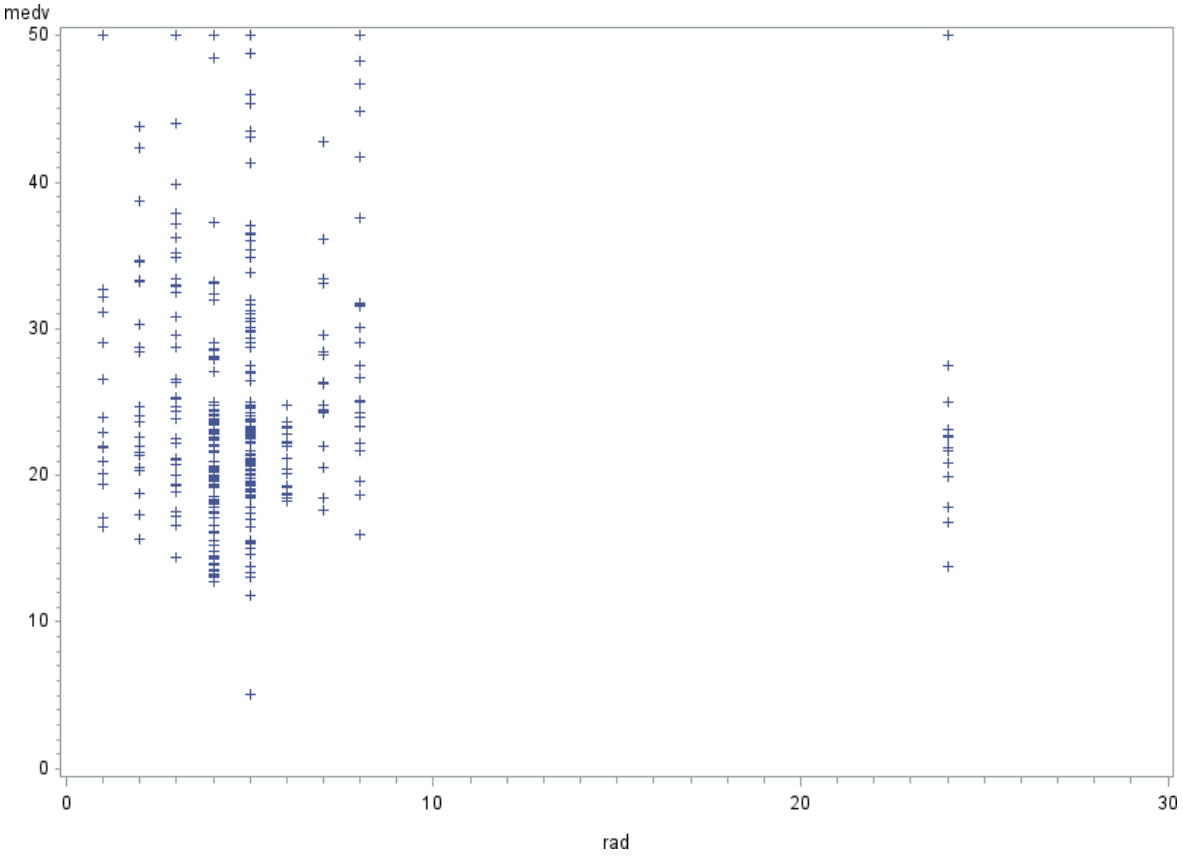

medv
50
40
30
20
10
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
chas

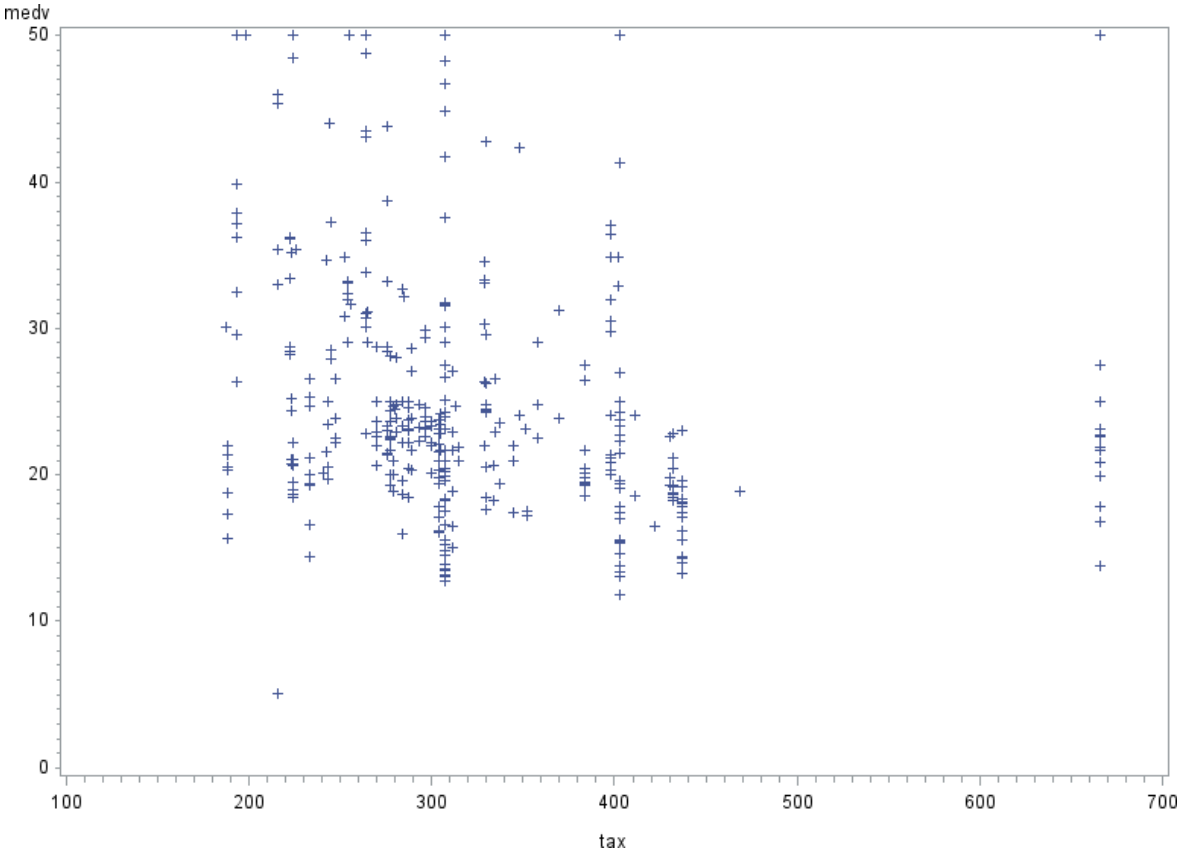

medv
50
40
30
20
10
0
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
nox

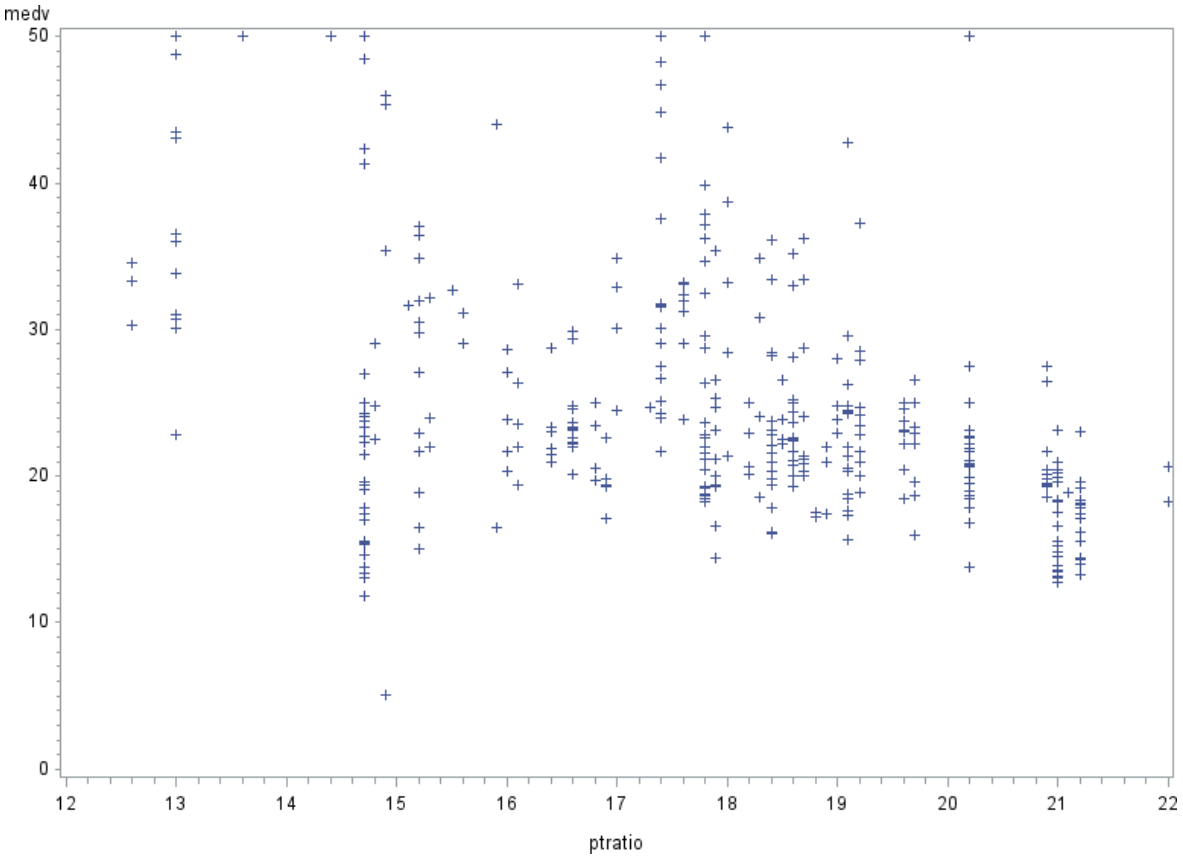

medv
50
40
30
20
10
0
3
4
5
6
7
8
9
rm

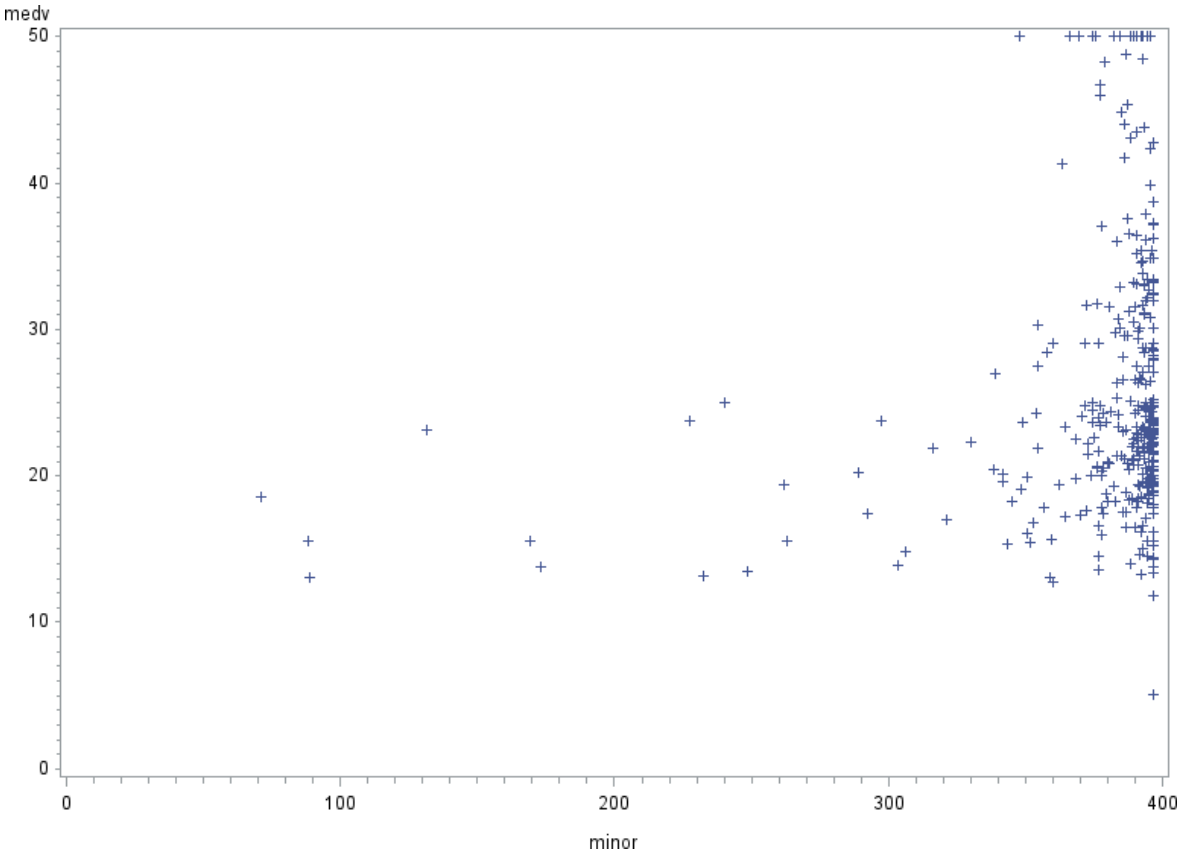

medv
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
age

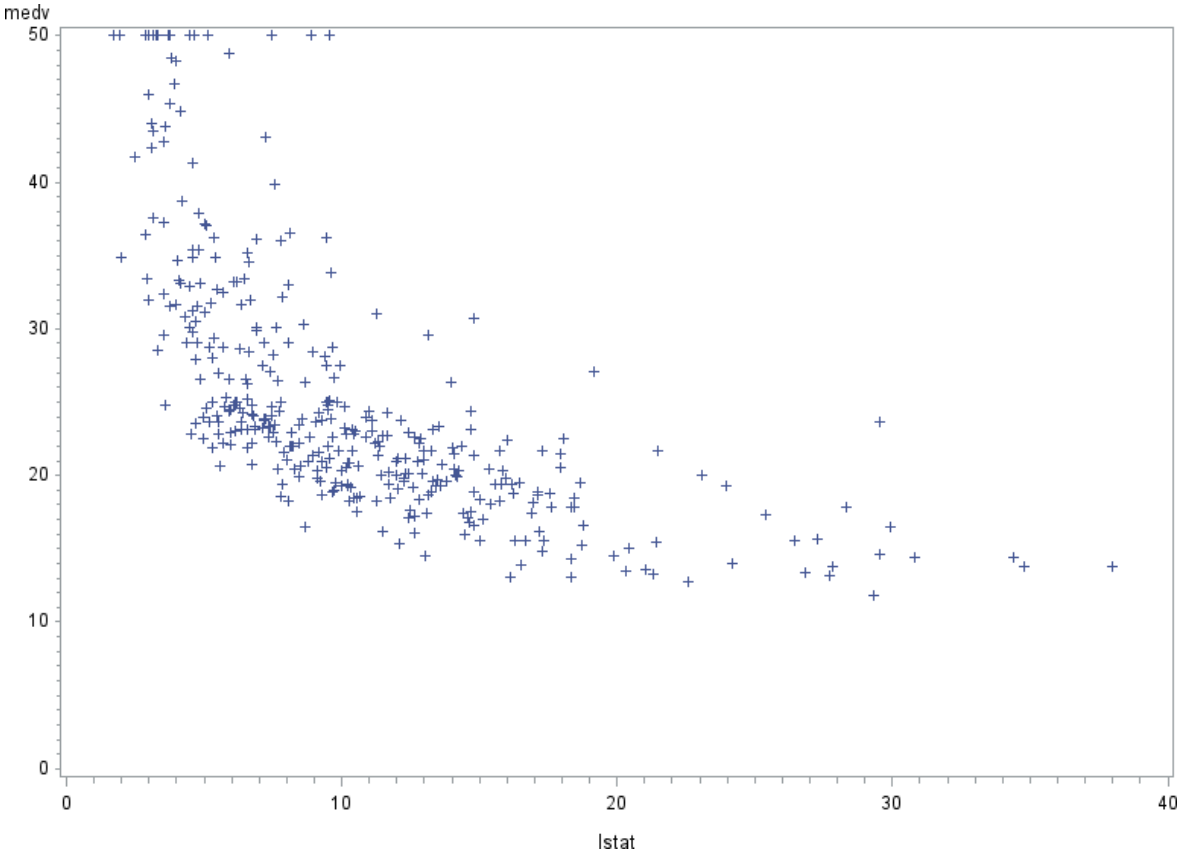

medv
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
dis

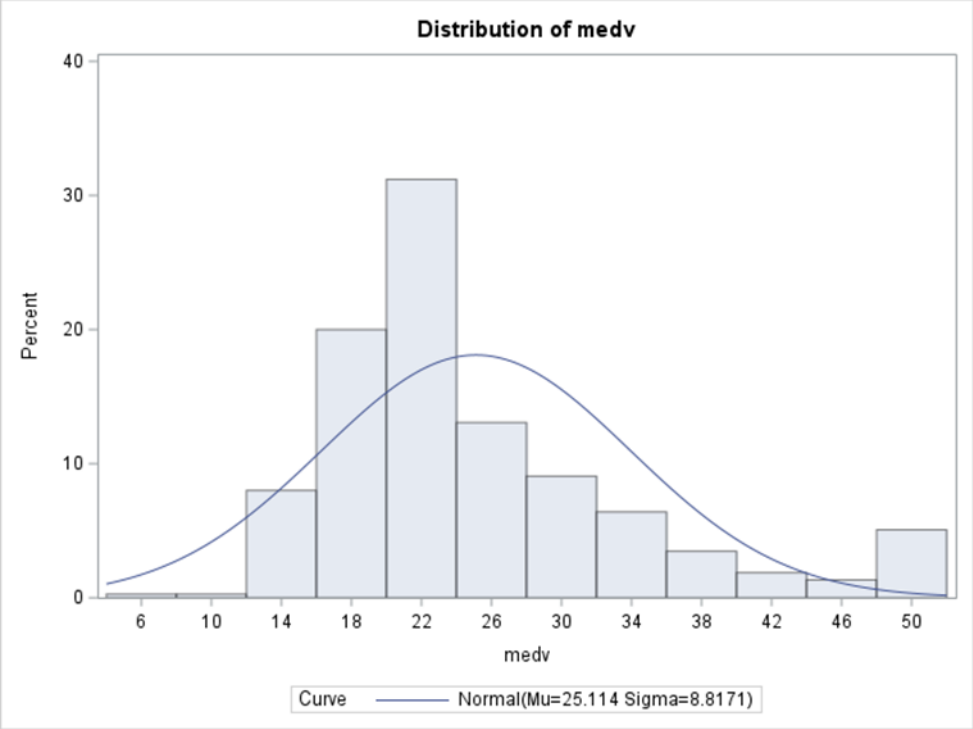

medv
50
40
30
20
10
0
0
10
20
30
rad

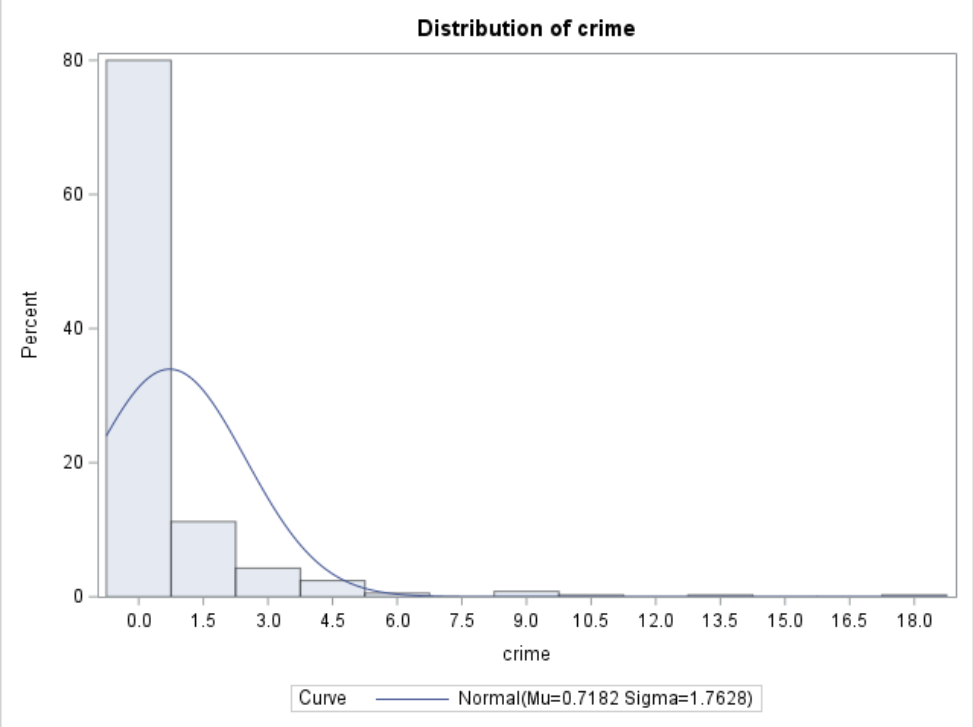

medv
50
40
30
20
10
0
100
200
300
400
500
600
700
tax

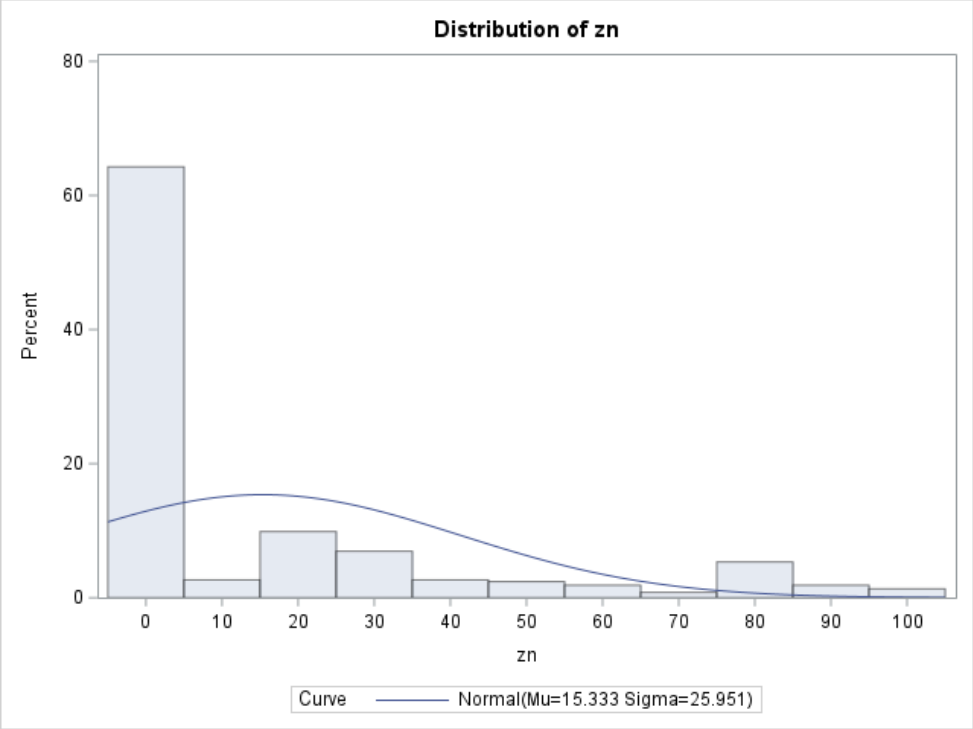

medv
50
40
30
20
10
0
12
13
14
15
16
17
18
19
20
21
22
ptratio

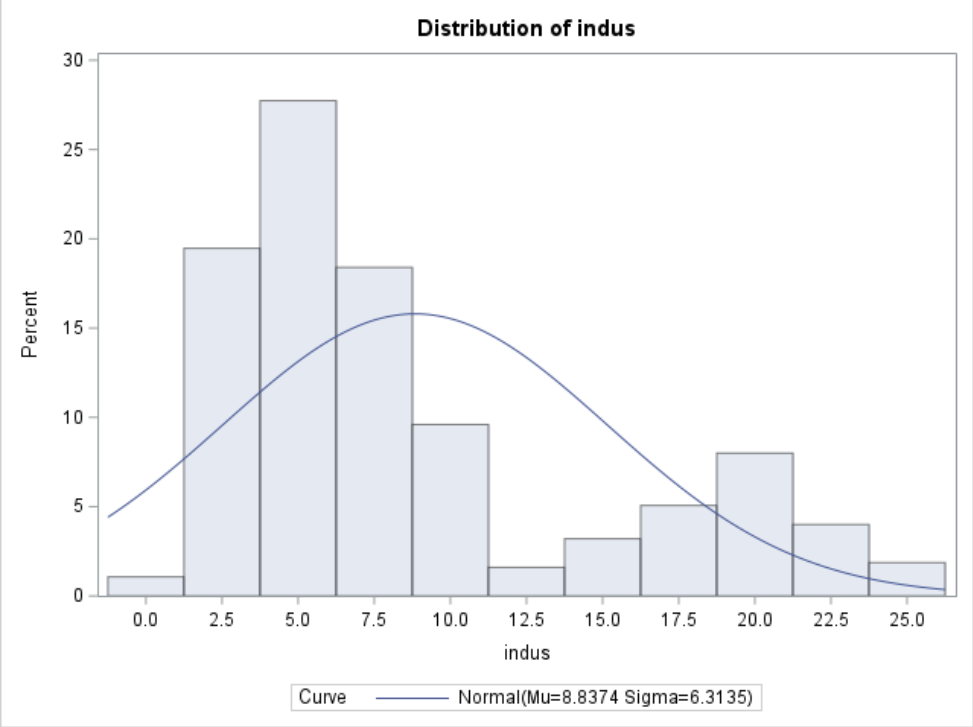

medv
50
40
30
20
10
0
0
100
200
300
400
minor

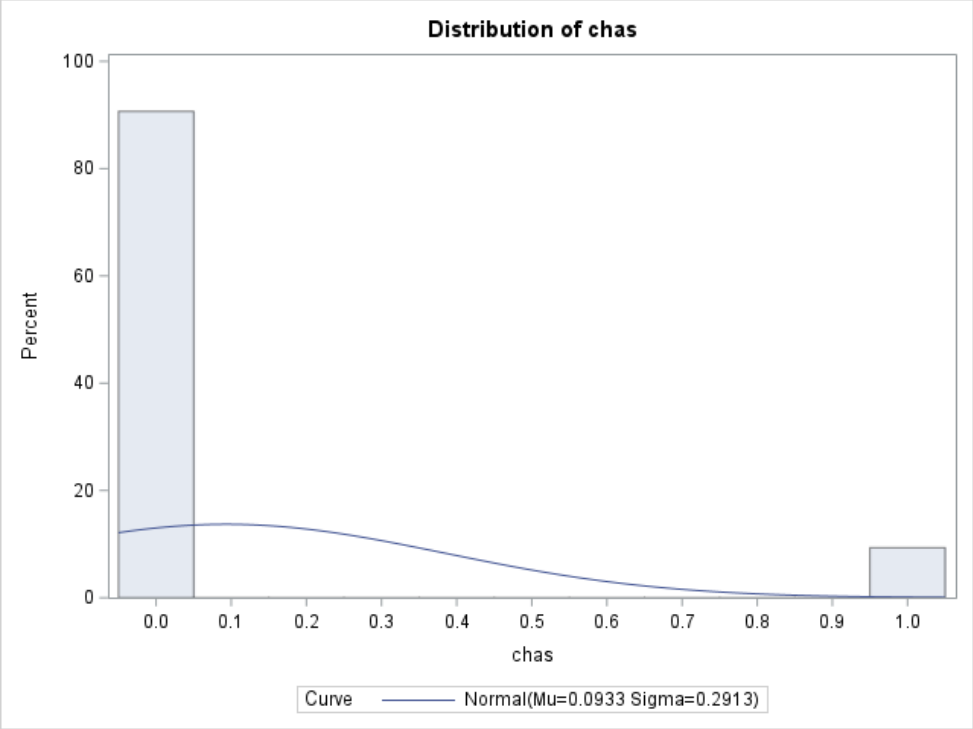

medv
50
40
30
20
10
0
0
10
20
30
40
lstat

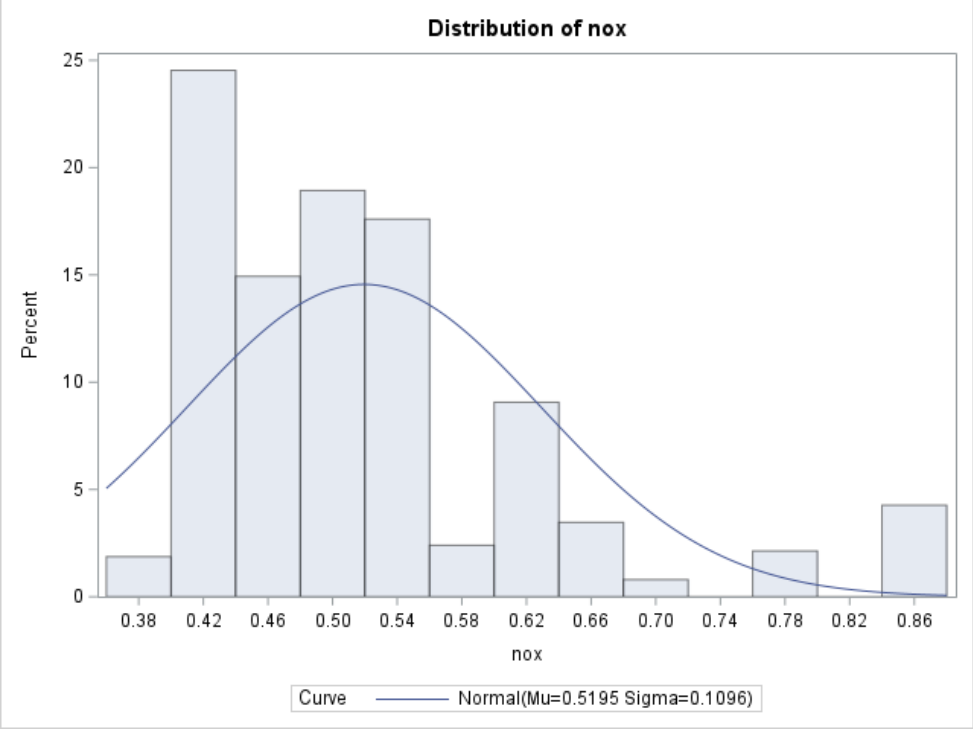

Distribution of medv
Percent
40
30
20
10
0
6
10
14
18
22
26
30
34
38
42
46
50
medv
Curve
Normal(Mu=25.114 Sigma=8.8171)

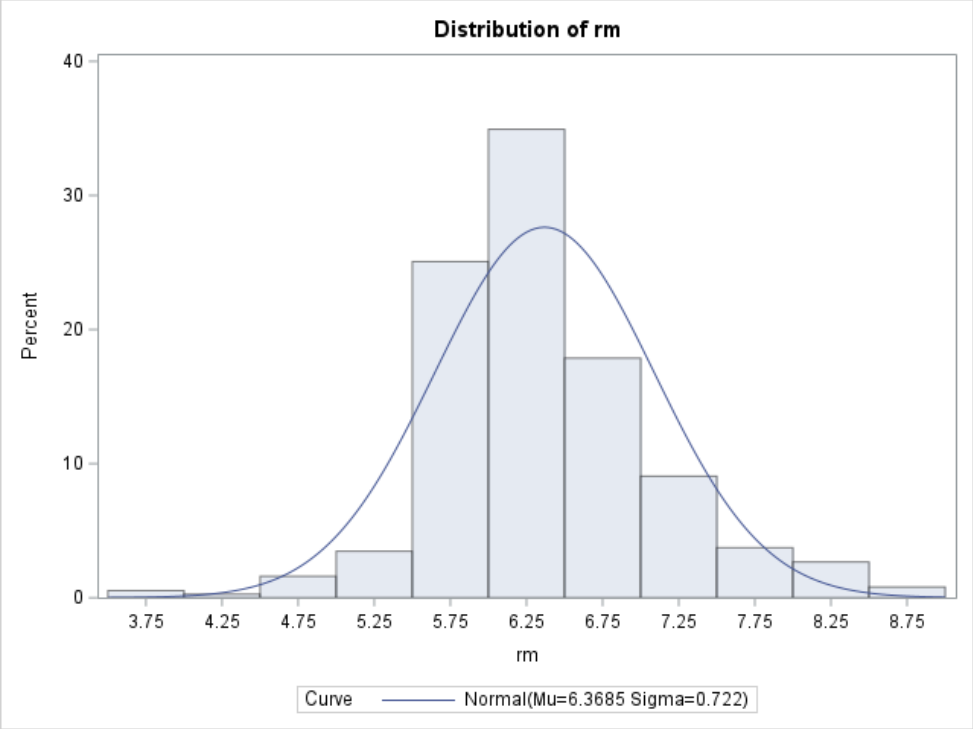

Distribution of crime
Percent
80
60
40
20
0
0.0
1.5
3.0
4.5
6.0
7.5
9.0
10.5
12.0
13.5
15.0
16.5
18.0
crime
Curve
Normal(Mu=0.7182 Sigma=1.7628)

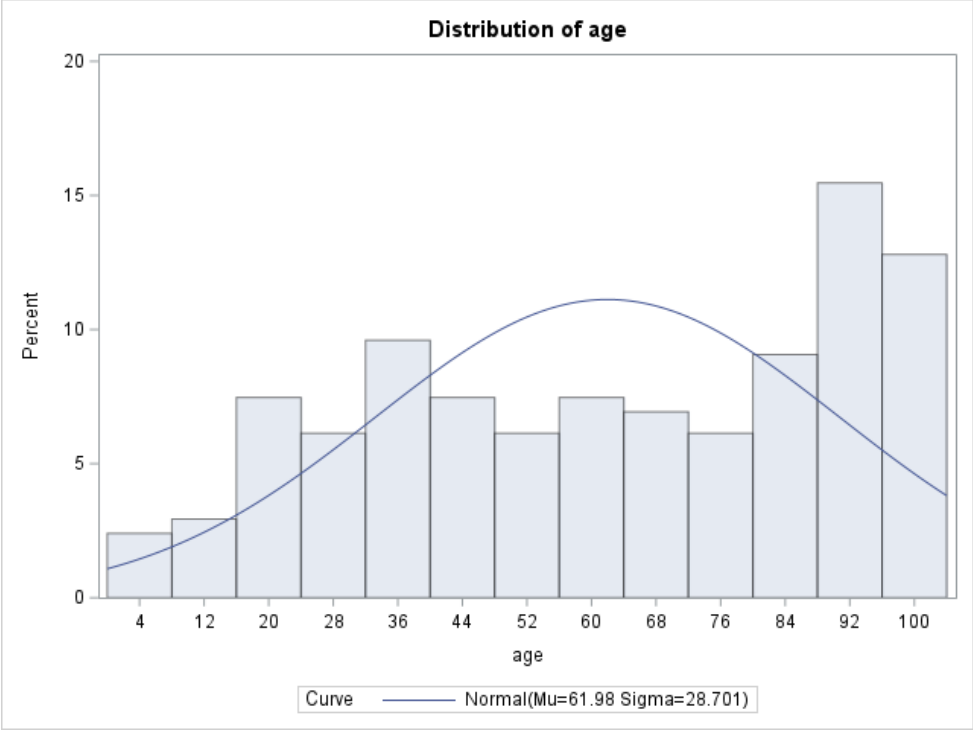

Distribution of zn
Percent
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
zn
Curve
Normal(Mu=15.333 Sigma=25.951)

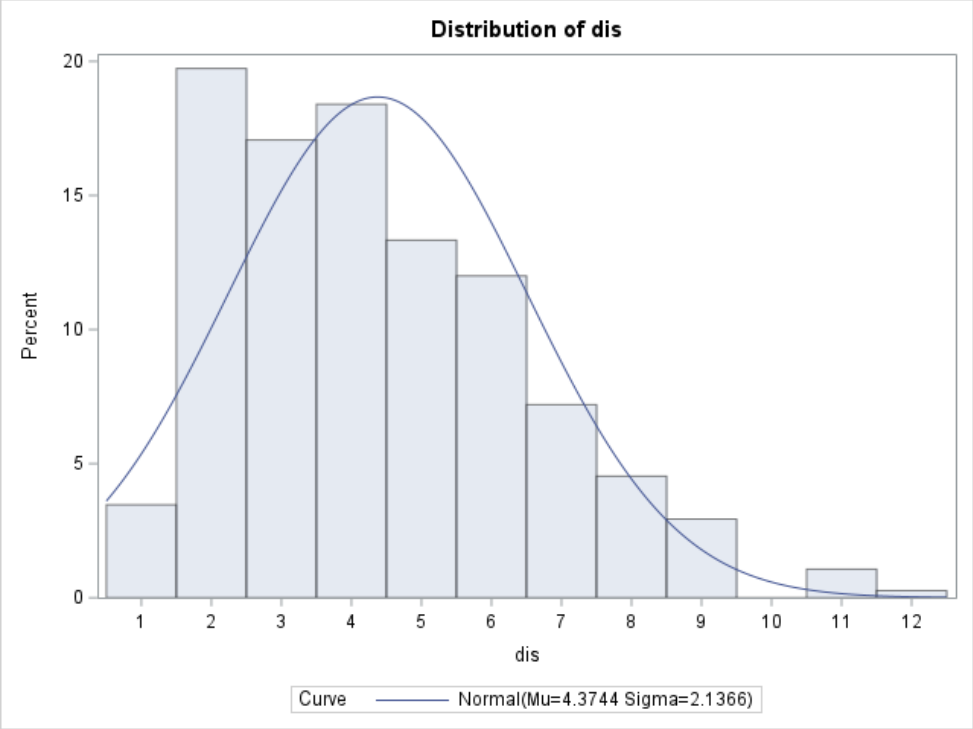

Distribution of indus
Percent
30
25
20
15
10
5
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
indus
Curve
Normal(Mu=8.8374 Sigma=6.3135)

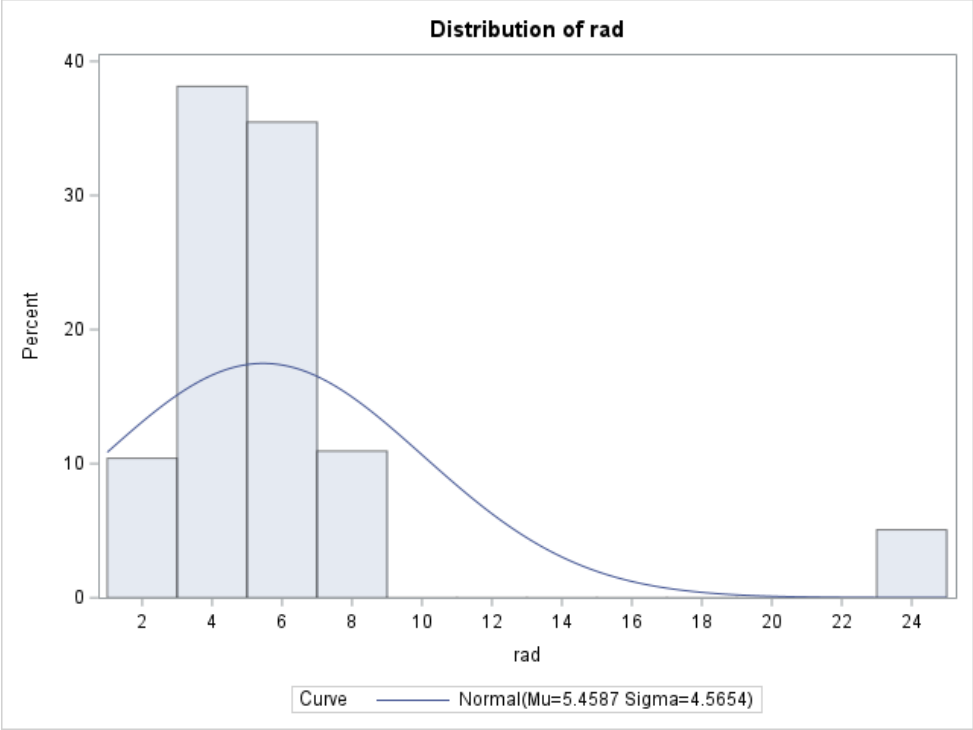

Distribution of chas
Percent
100
80
60
40
20
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
chas
Curve
Normal(Mu=0.0933 Sigma=0.2913)

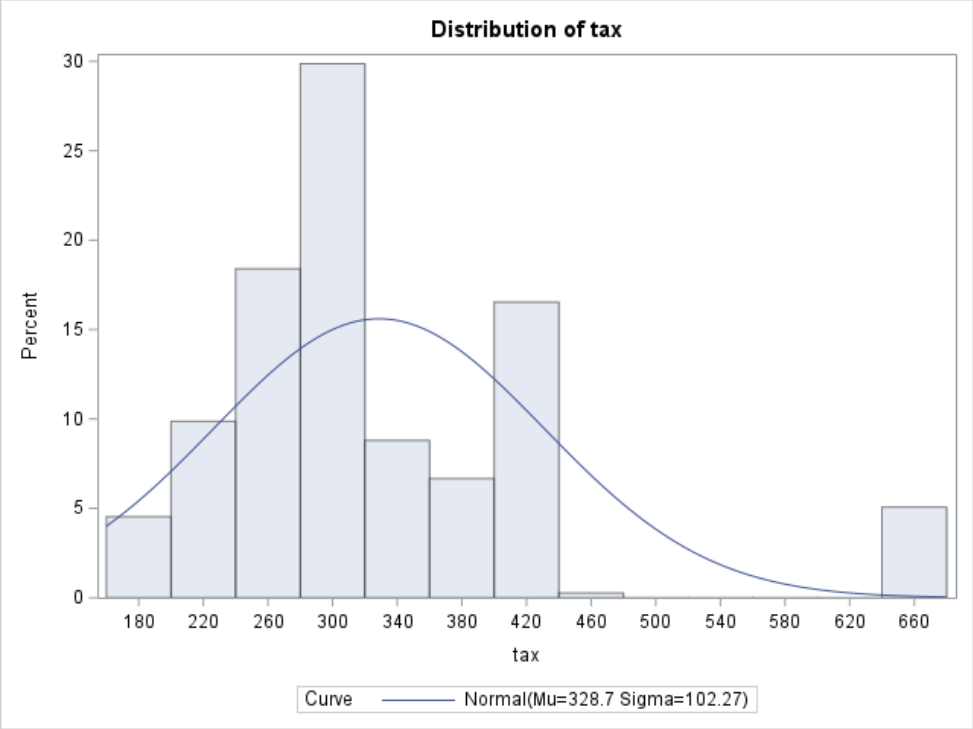

Distribution of nox
Percent
25
20
15
10
5
0
0.38
0.42
0.46
0.50
0.54
0.58
0.62
0.66
0.70
0.74
0.78
0.82
0.86
nox
Curve
Normal(Mu=0.5195 Sigma=0.1096)

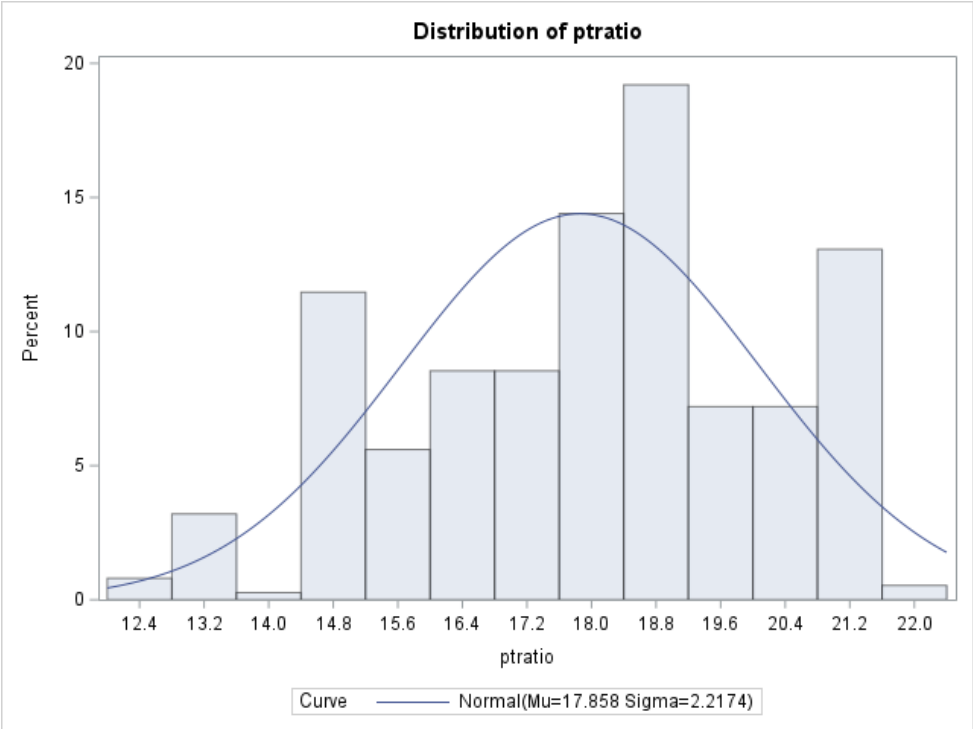

Distribution of rm
Percent
40
30
20
10
0
3.75
4.25
4.75
5.25
5.75
6.25
6.75
7.25
7.75
8.25
8.75
rm
Curve
Normal(Mu=6.3685 Sigma=0.722)

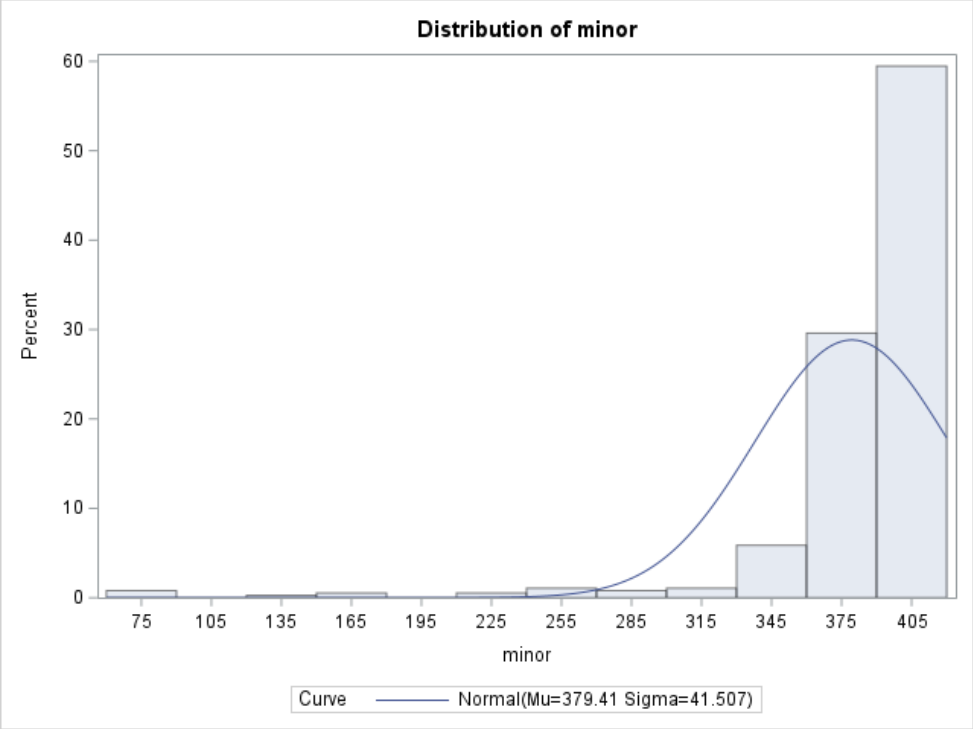

Distribution of age
Percent
20
15
10
5
0
4
12
20
28
36
44
52
60
68
76
84
92
100
age
Curve
Normal(Mu=61.98 Sigma=28.701)

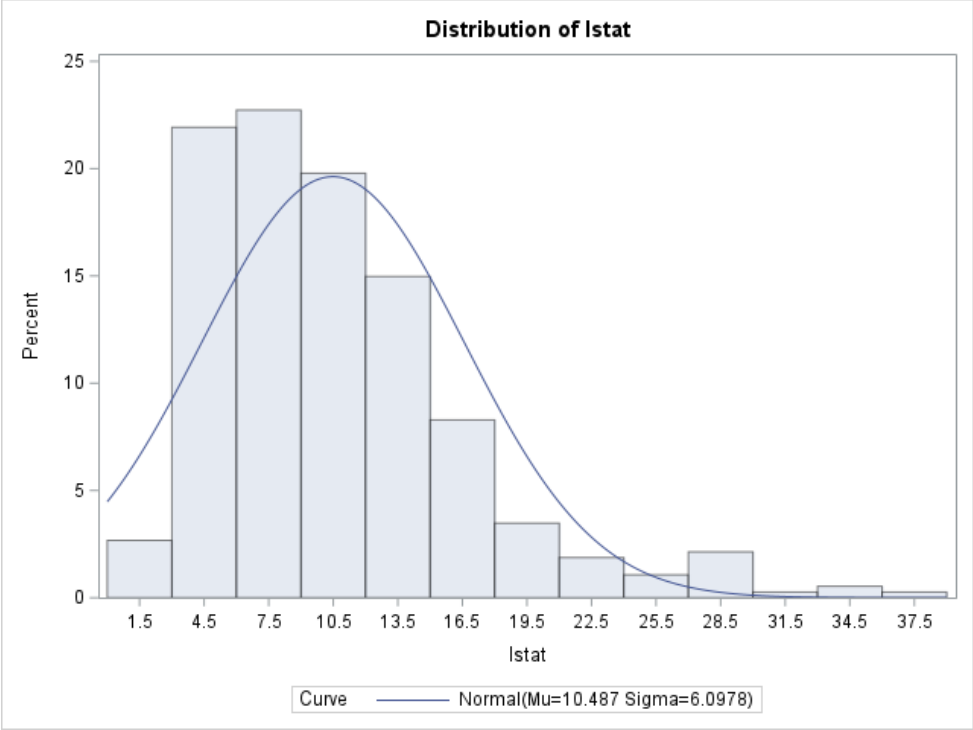

Distribution of dis
Percent
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
dis
Curve
Normal(Mu=4.3744 Sigma=2.1366)

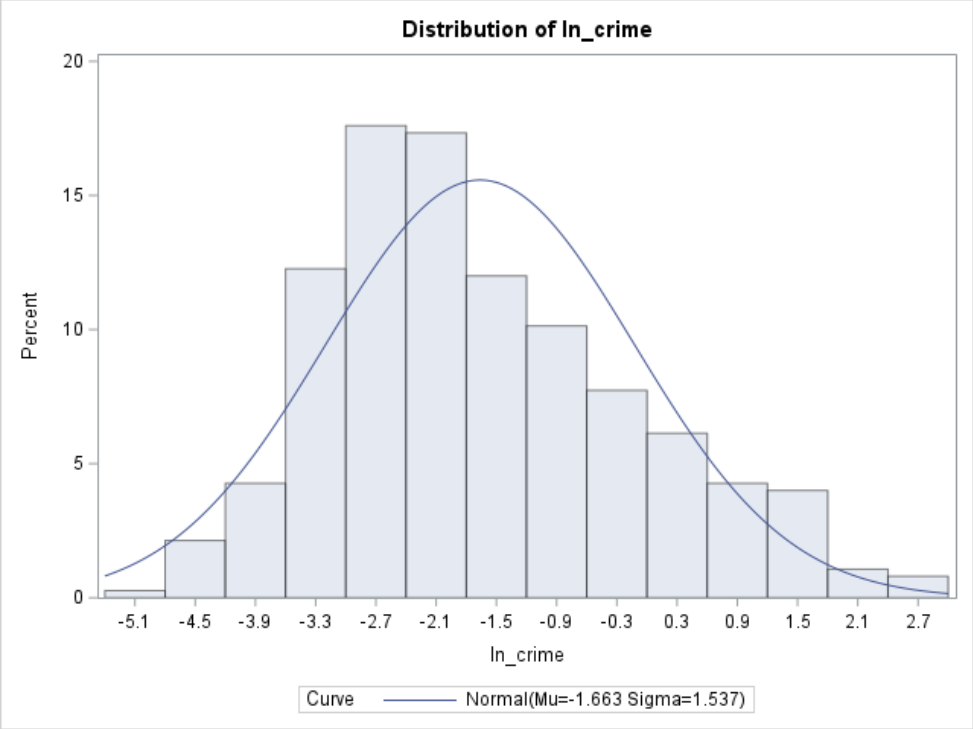

Distribution of rad
Percent
40
30
20
10
0
2
4
6
8
10
12
14
16
18
20
22
24
rad
Curve
Normal(Mu=5.4587 Sigma=4.5654)

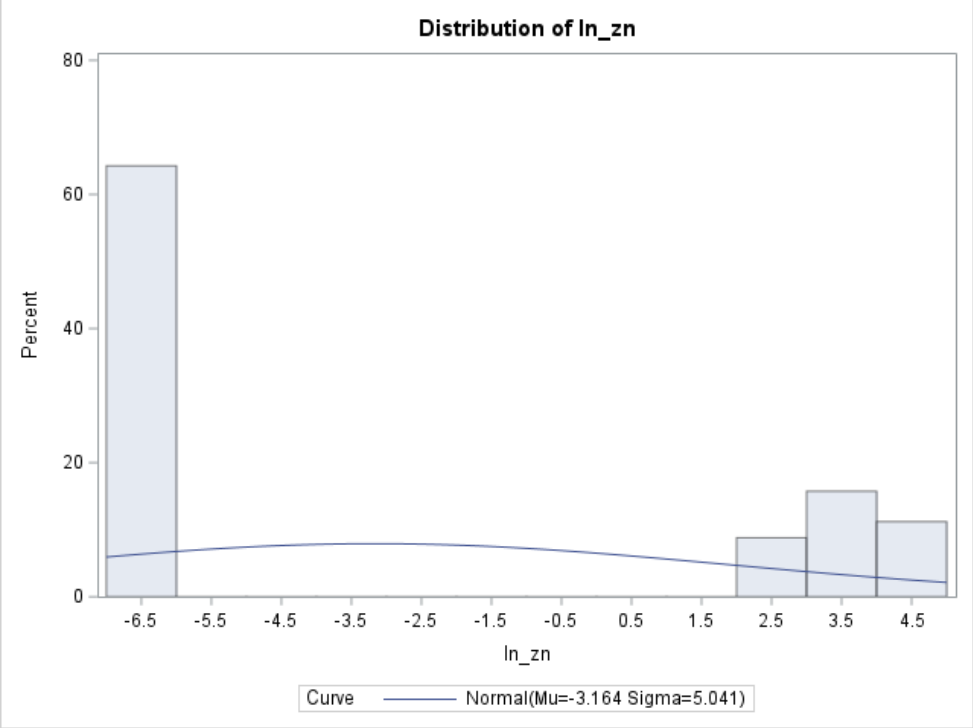

Distribution of tax
Percent
30
25
20
15
10
5
0
180
220
260
300
340
380
420
460
500
540
580
620
660
tax
Curve
Normal(Mu=328.7 Sigma=102.27)

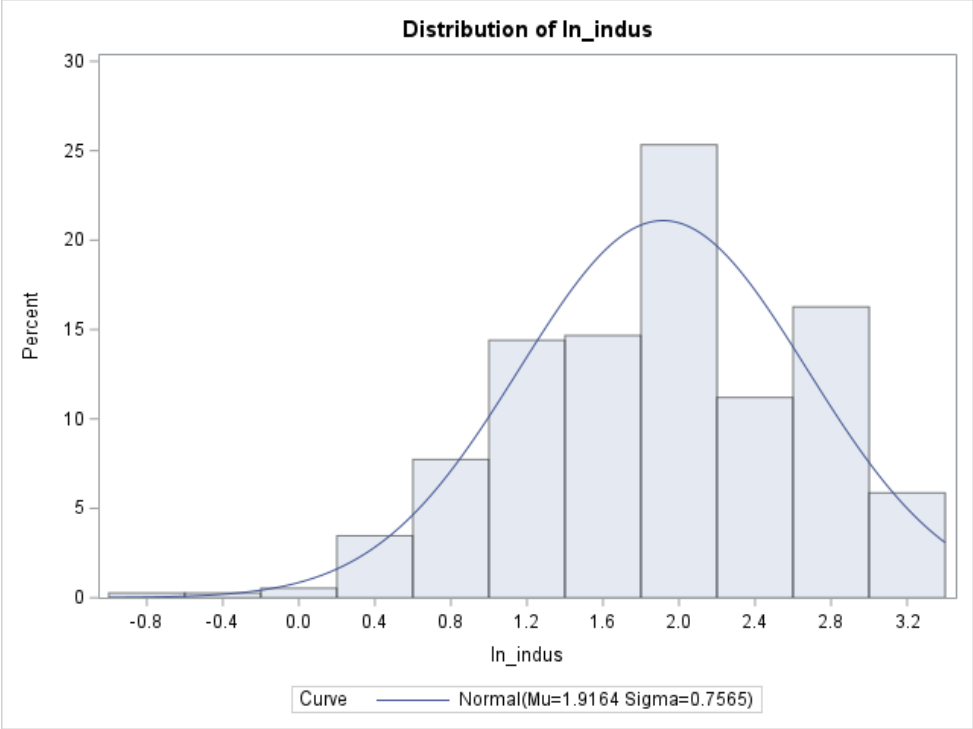

Distribution of ptratio
Percent
20
15
10
5
0
12.4
13.2
14.0
14.8
15.6
16.4
17.2
18.0
18.8
19.6
20.4
21.2
22.0
ptratio
Curve
Normal(Mu=17.858 Sigma=2.2174)

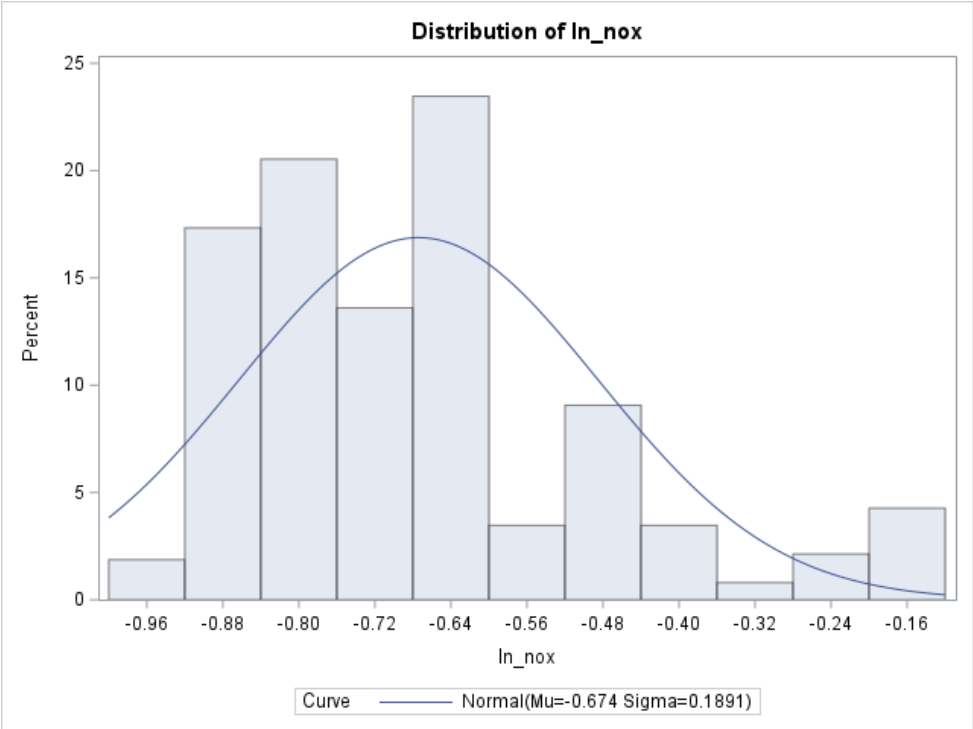

Distribution of minor
Percent
60
50
40
30
20
10
0
75
105
135
165
195
225
255
285
315
345
375
405
minor
Curve
Normal(Mu=379.41 Sigma=41.507)

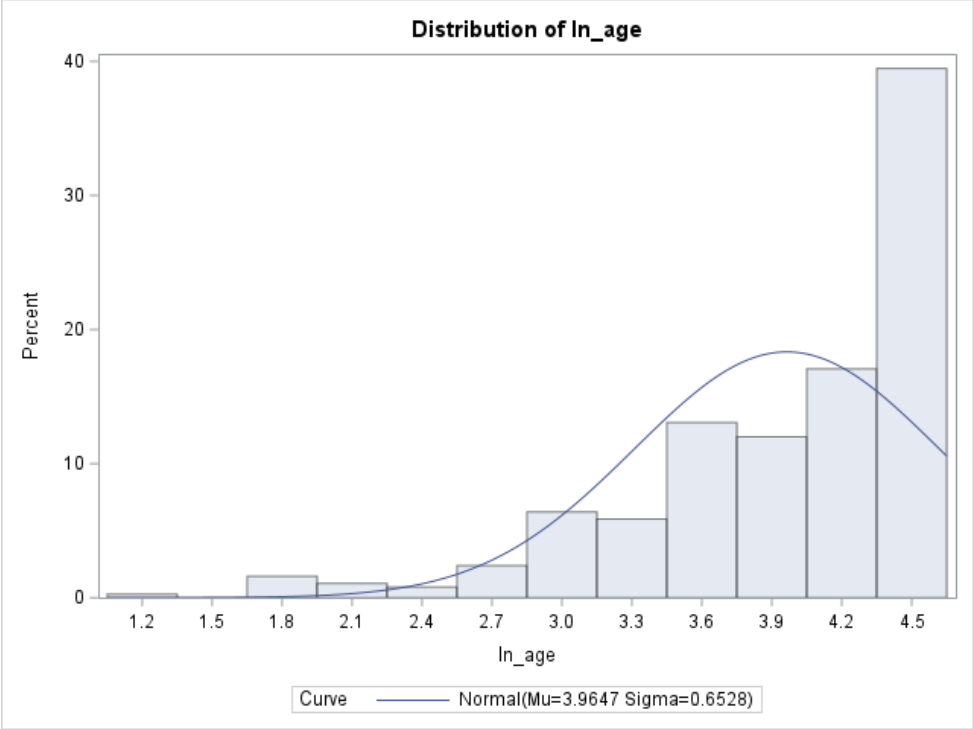

Distribution of lstat
Percent
25
20
15
10
5
0
1.5
4.5
7.5
10.5
13.5
16.5
19.5
22.5
25.5
28.5
31.5
34.5
37.5
lstat
Curve
Normal(Mu=10.487 Sigma=6.0978)

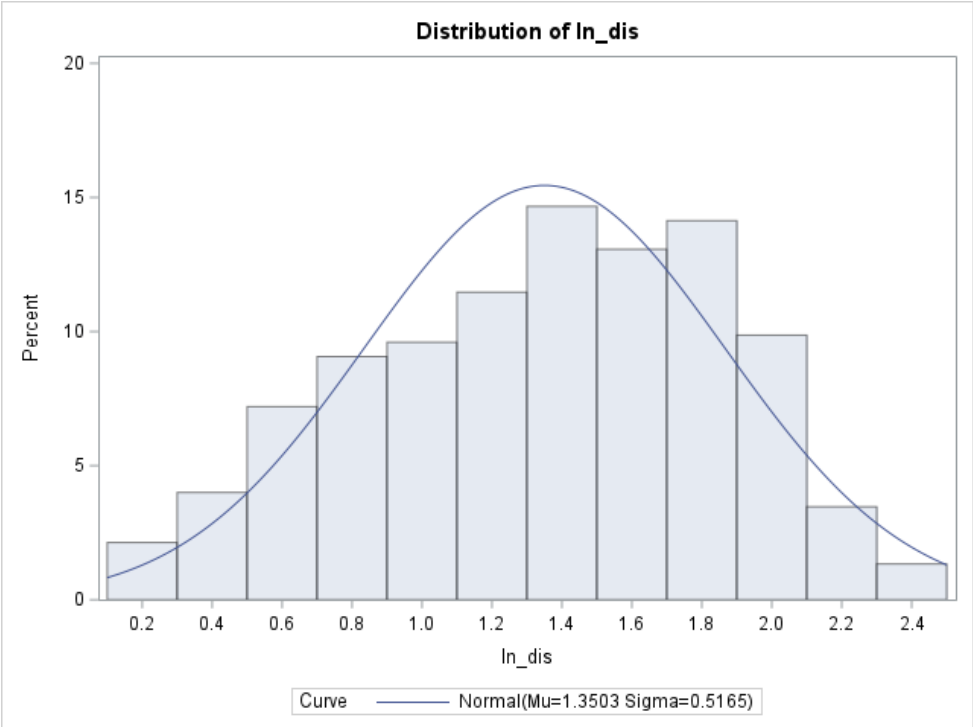

Distribution of ln_crime
Percent
20
15
10
5
0
-5.1
-4.5
-3.9
-3.3
-2.7
-2.1
-1.5
-0.9
-0.3
0.3
0.9
1.5
2.1
2.7
ln_crime
Curve
Normal(Mu=-1.663 Sigma=1.537)

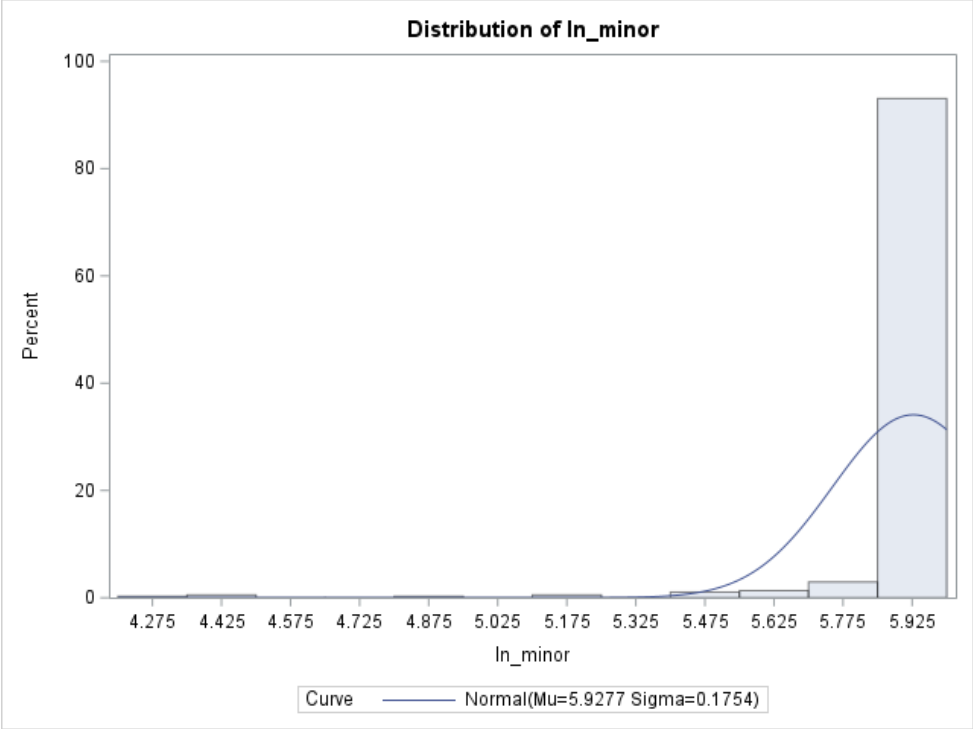

Distribution of ln_zn
Percent
80
60
40
20
0
-6.5
-5.5
-4.5
-3.5
-2.5
-1.5
-0.5
0.5
1.5
2.5
3.5
4.5
ln_zn
Curve
Normal(Mu=-3.164 Sigma=5.041)

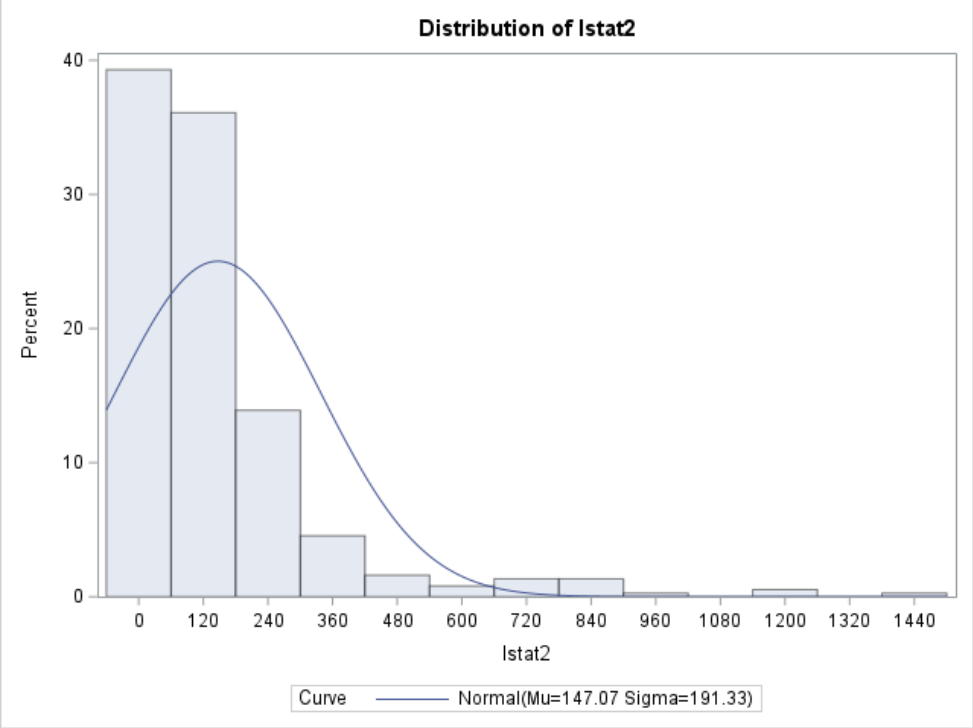

Distribution of ln_indus
Percent
30
25
20
15
10
5
0
-0.8
-0.4
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
ln_indus
Curve
Normal(Mu=1.9164 Sigma=0.7565)


Distribution of ln_nox
Percent
25
20
15
10
5
0
-0.96
-0.88
-0.80
-0.72
-0.64
-0.56
-0.48
-0.40
-0.32
-0.24
-0.16
ln_nox
Curve
Normal(Mu=-0.674 Sigma=0.1891)


Distribution of ln_age
Percent
40
30
20
10
0
1.2
1.5
1.8
2.1
2.4
2.7
3.0
3.3
3.6
3.9
4.2
4.5
ln_age
Curve
Normal(Mu=3.9647 Sigma=0.6528)


Distribution of ln_dis
Percent
20
15
10
5
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
ln_dis
Curve
Normal(Mu=1.3503 Sigma=0.5165)


Distribution of ln_minor
Percent
100
80
60
40
20
0
4.275
4.425
4.575
4.725
4.875
5.025
5.175
5.325
5.475
5.625
5.775
5.925
ln_minor
Curve
Normal(Mu=5.9277 Sigma=0.1754)

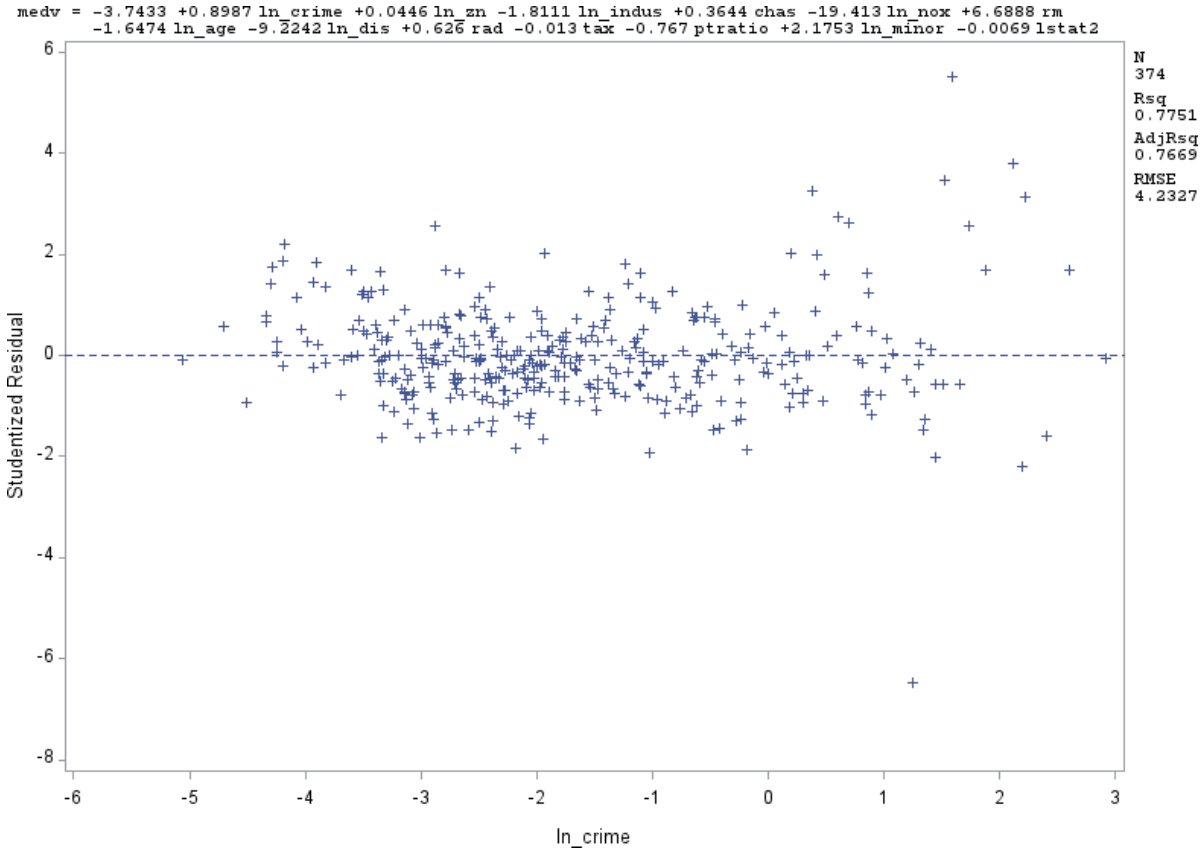

Distribution of lstat2
Percent
40
30
20
10
0
0
120
240
360
480
600
720
840
960
1080
1200
1320
1440
lstat2
Curve
Normal(Mu=147.07 Sigma=191.33)

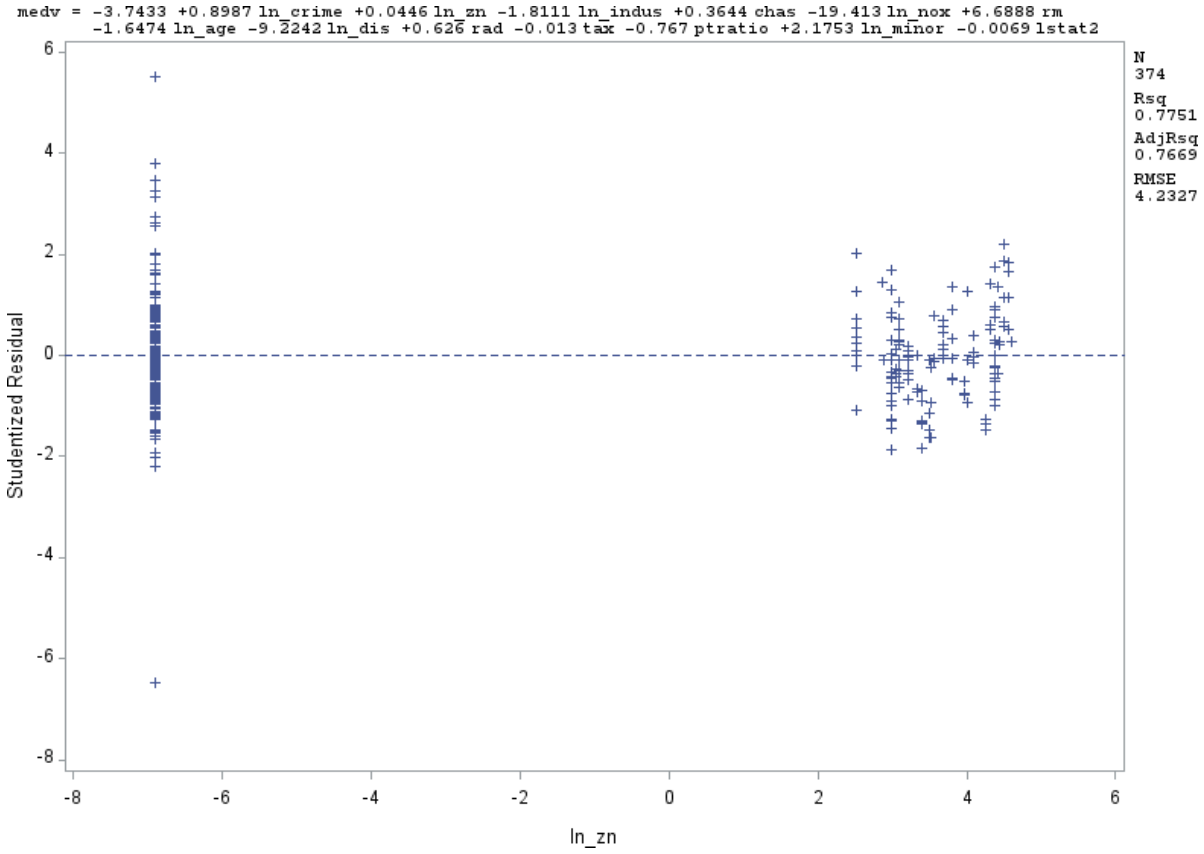
Number of Observations Read	375
Number of Observations Used	374
Number of Observations with Missing Values	1

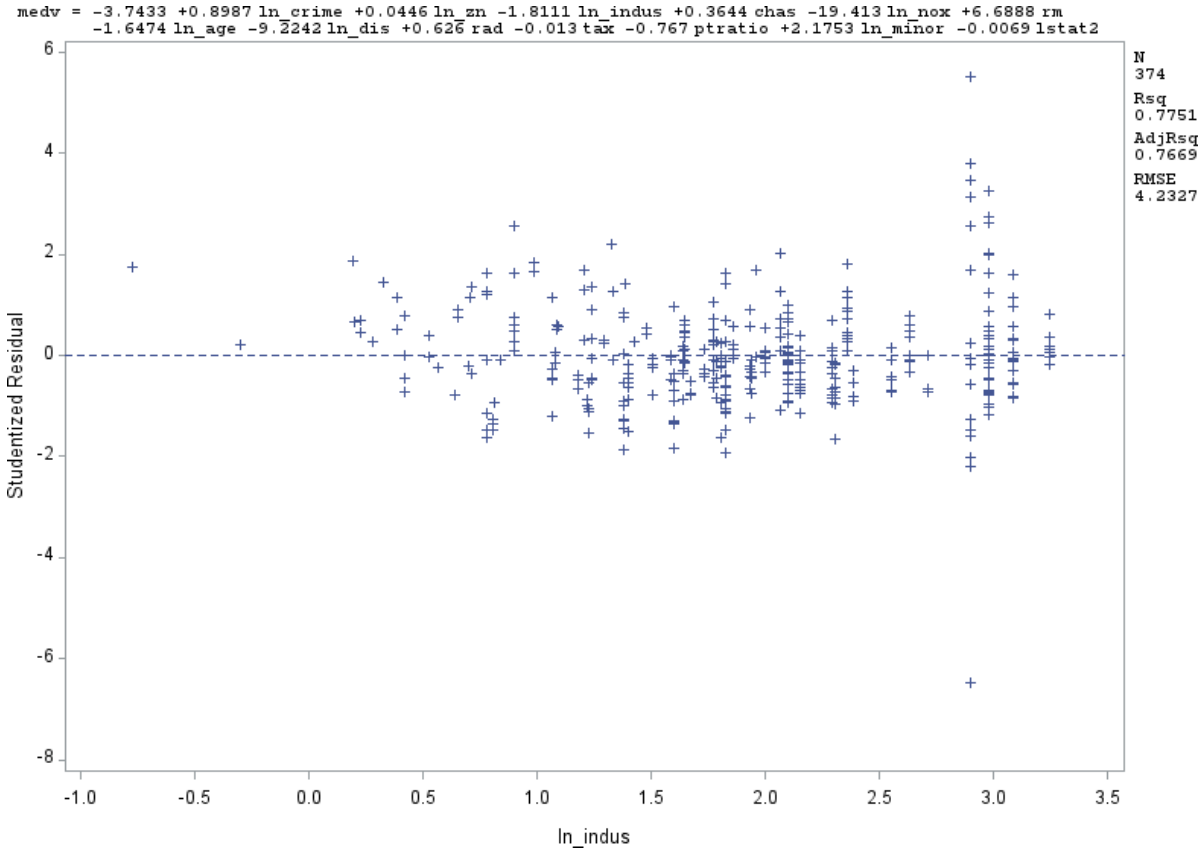
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	22224	1709.50628	95.42	<.0001
Error	360	6449.67328	17.91576		
Corrected Total	373	28673			

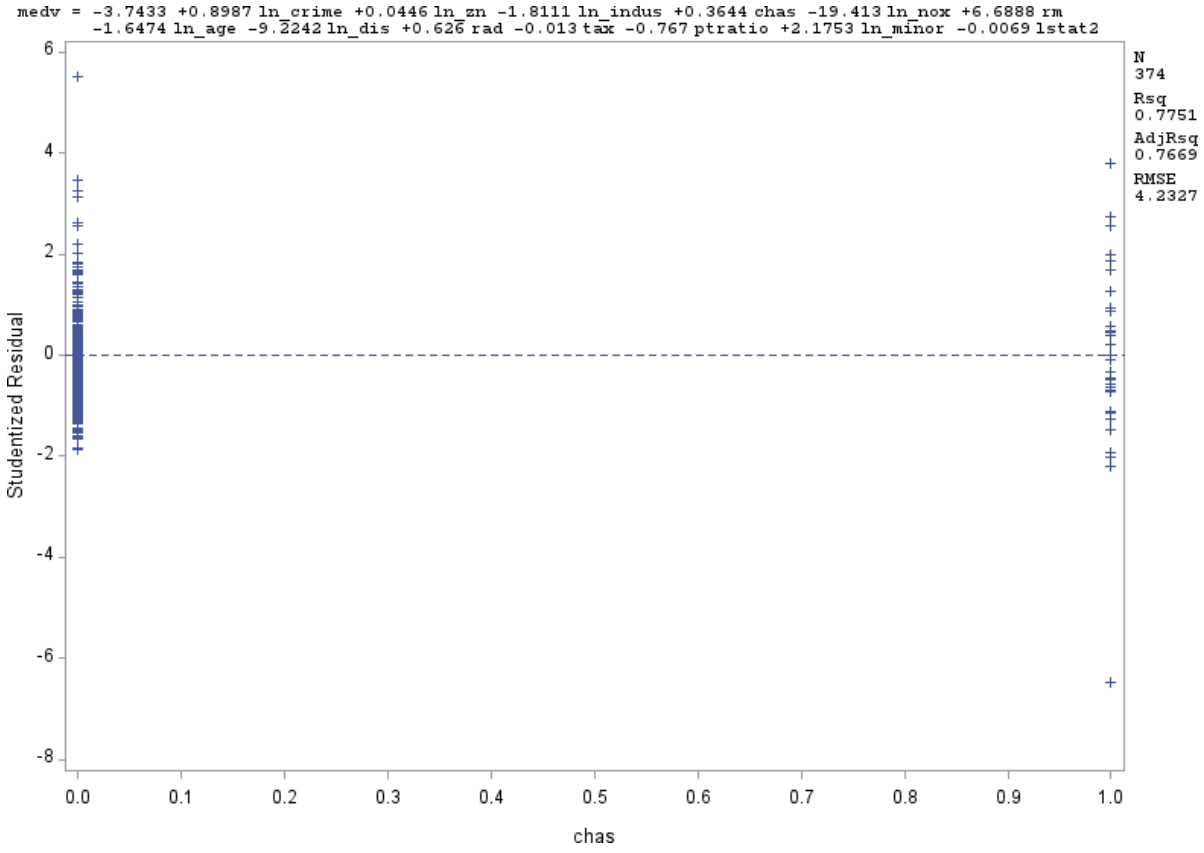
Root MSE	4.23270	R-Square	0.7751
Dependent Mean	25.16791	Adj R-Sq	0.7669
Coeff Var	16.81785		

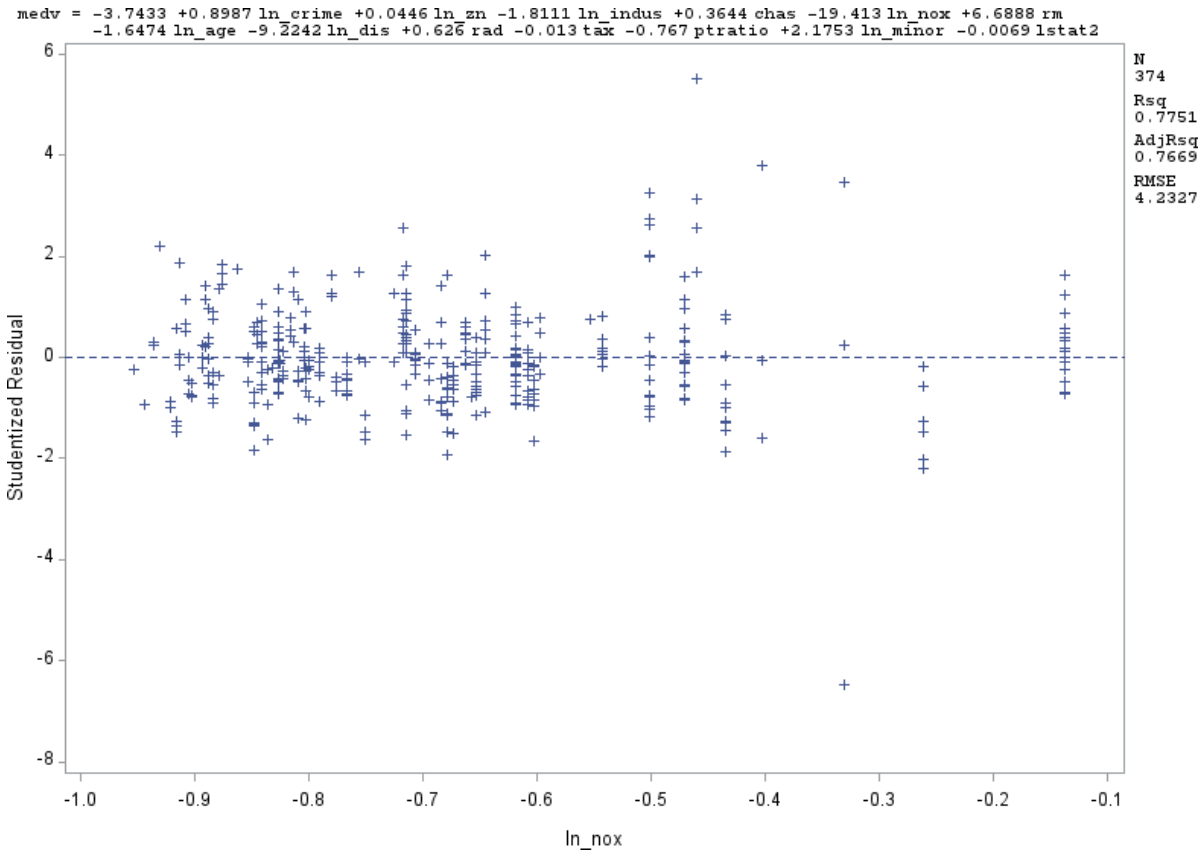
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.74326	8.75842	-0.43	0.6694
ln_crime	1	0.89875	0.28754	3.13	0.0019
ln_zn	1	0.04456	0.06380	0.70	0.4854
ln_indus	1	-1.81114	0.48820	-3.71	0.0002
chas	1	0.36439	0.80911	0.45	0.6527
ln_nox	1	-19.41258	2.74739	-7.07	<.0001
rm	1	6.68877	0.41402	16.16	<.0001
ln_age	1	-1.64744	0.46313	-3.56	0.0004
ln_dis	1	-9.22418	0.89756	-10.28	<.0001
rad	1	0.62600	0.08523	7.35	<.0001
tax	1	-0.01304	0.00369	-3.54	0.0005
ptratio	1	-0.76702	0.12623	-6.08	<.0001
ln_minor	1	2.17529	1.36808	1.59	0.1127
lstat2	1	-0.00686	0.00152	-4.52	<.0001

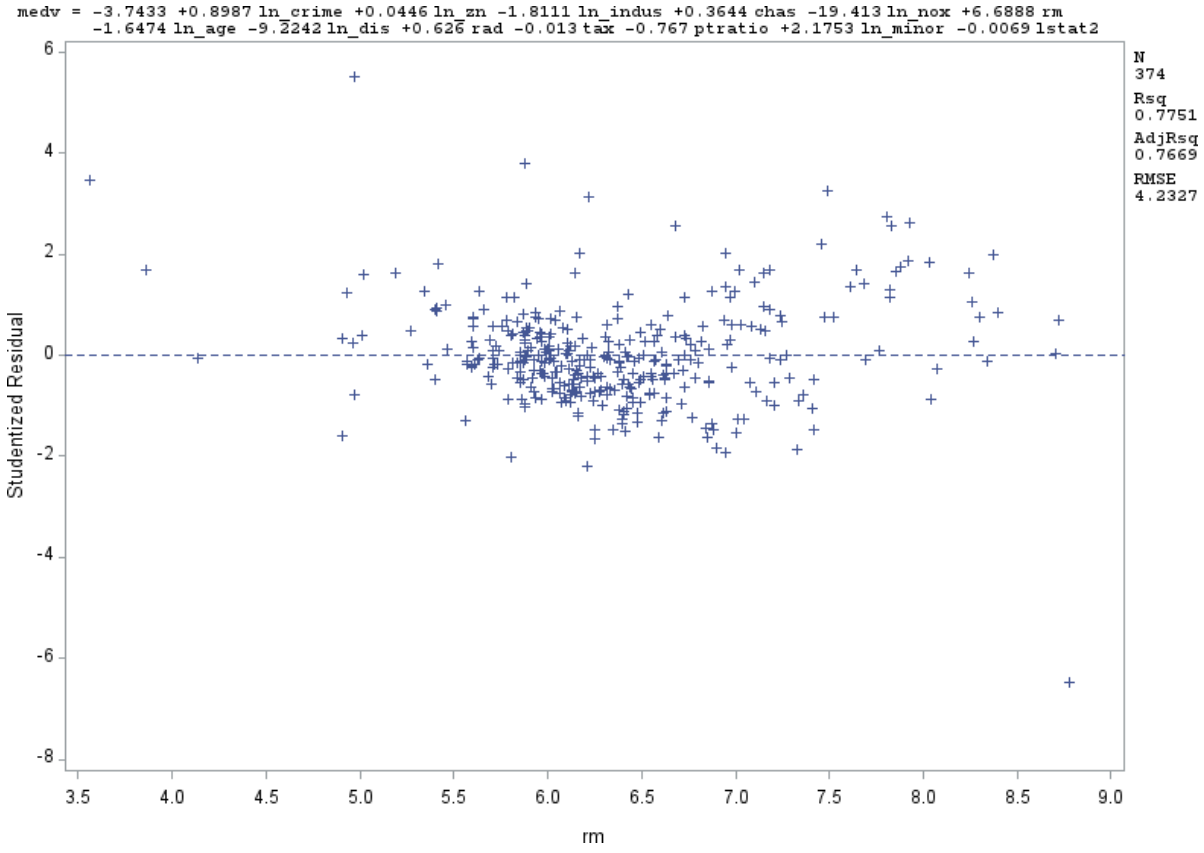

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
ln_crime

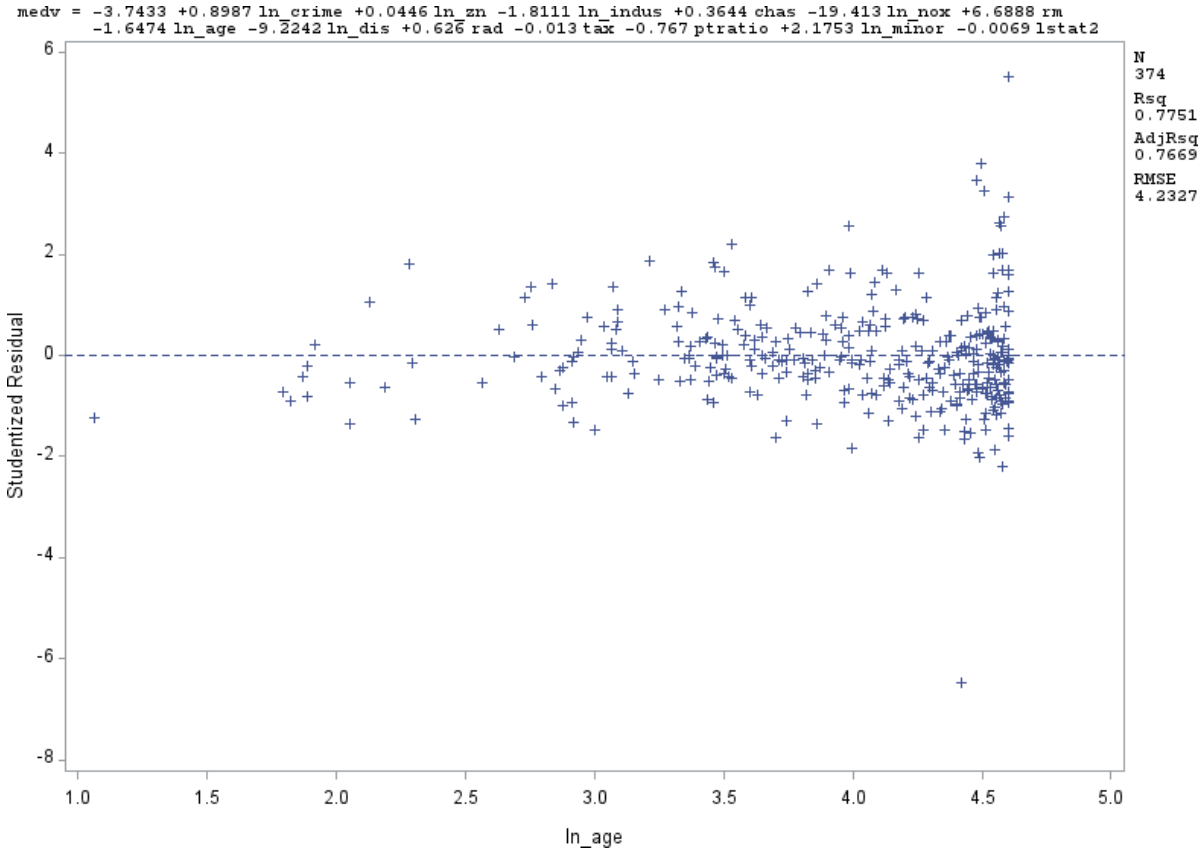

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
ln_zn

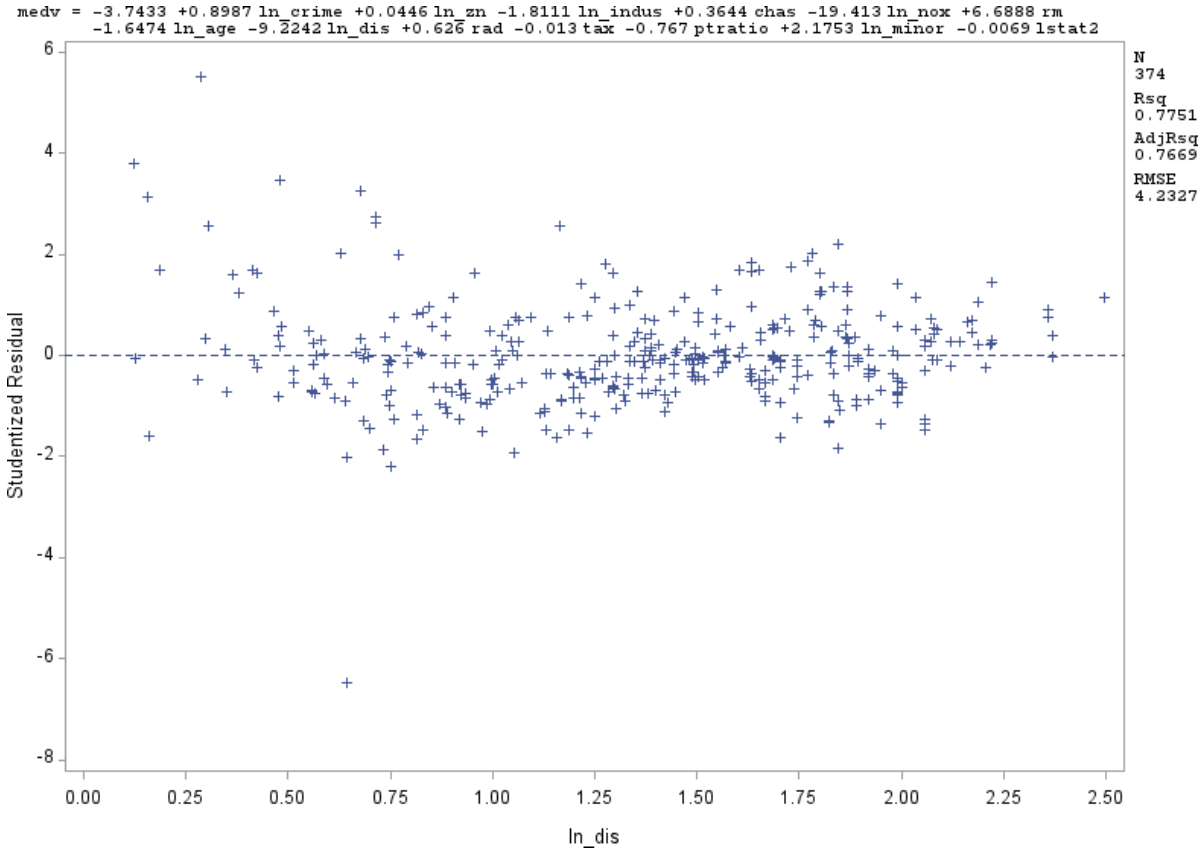

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
ln_indus

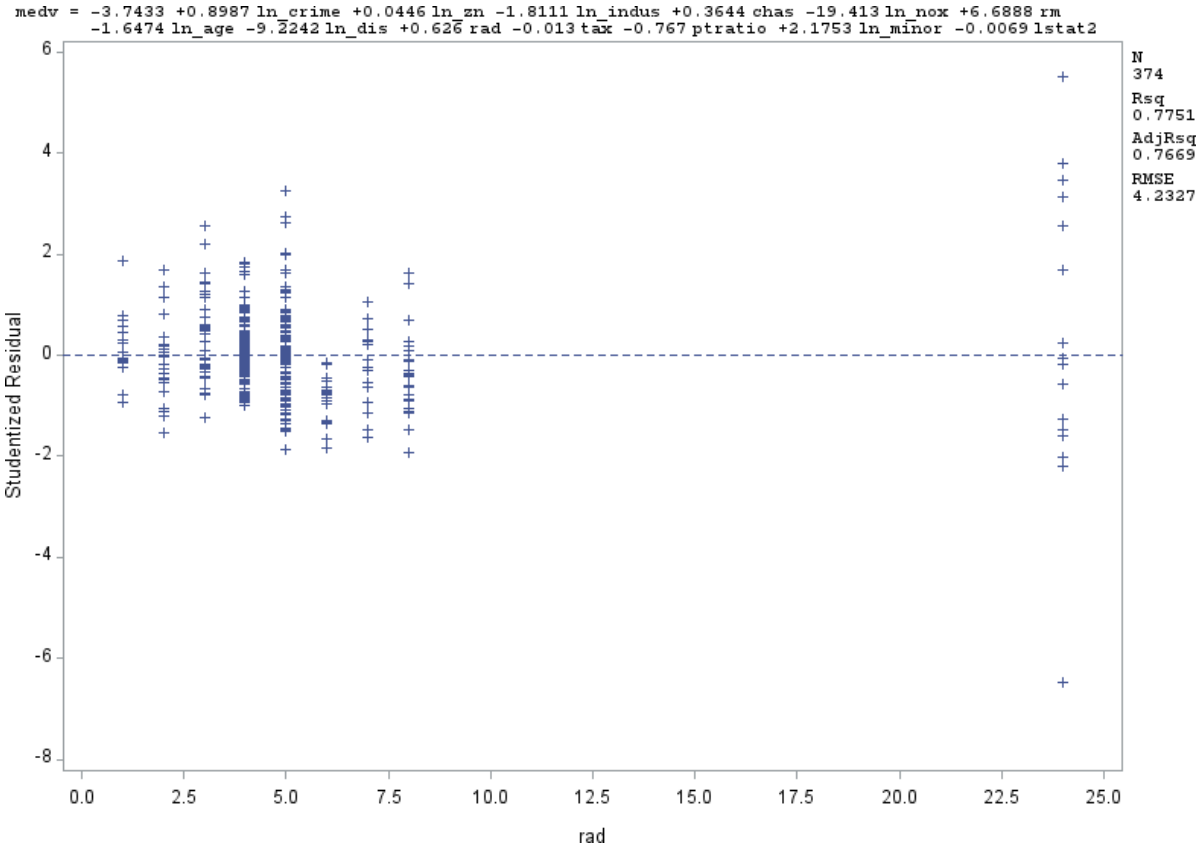

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
chas

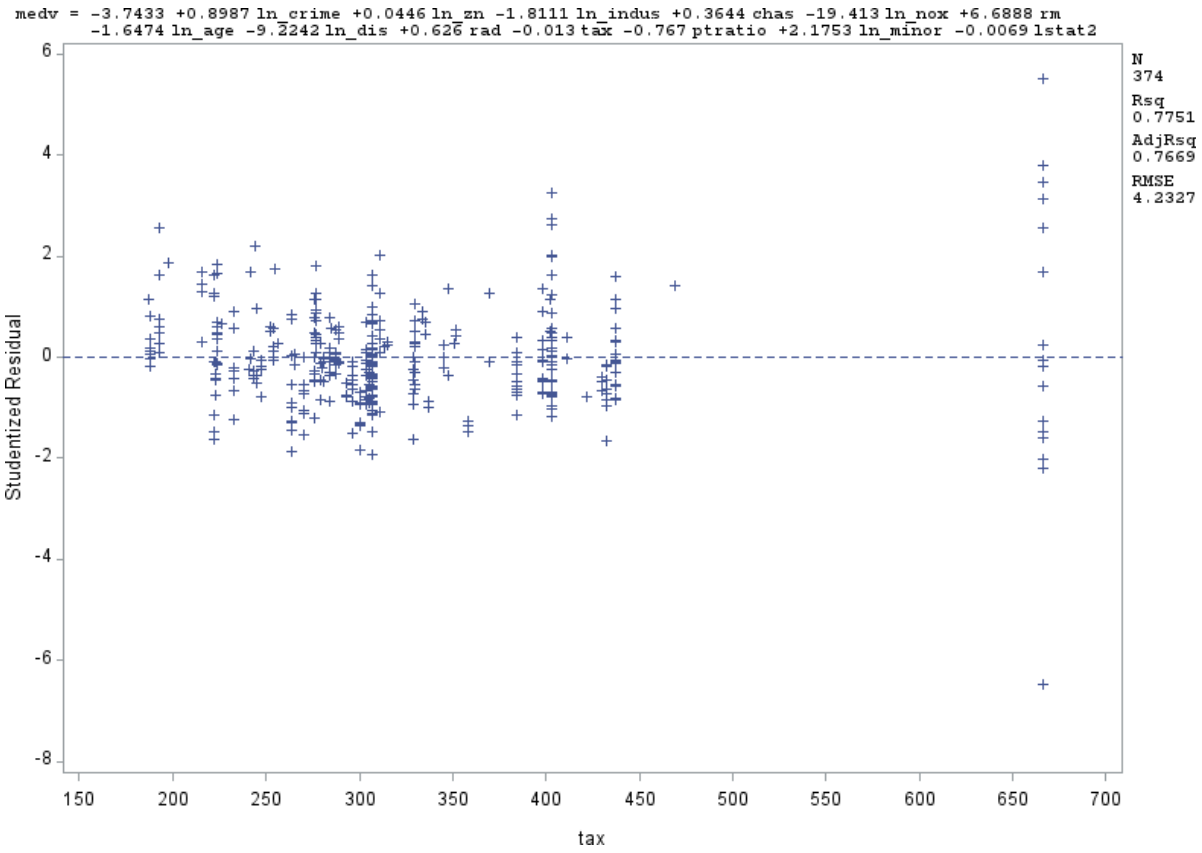

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
ln_nox

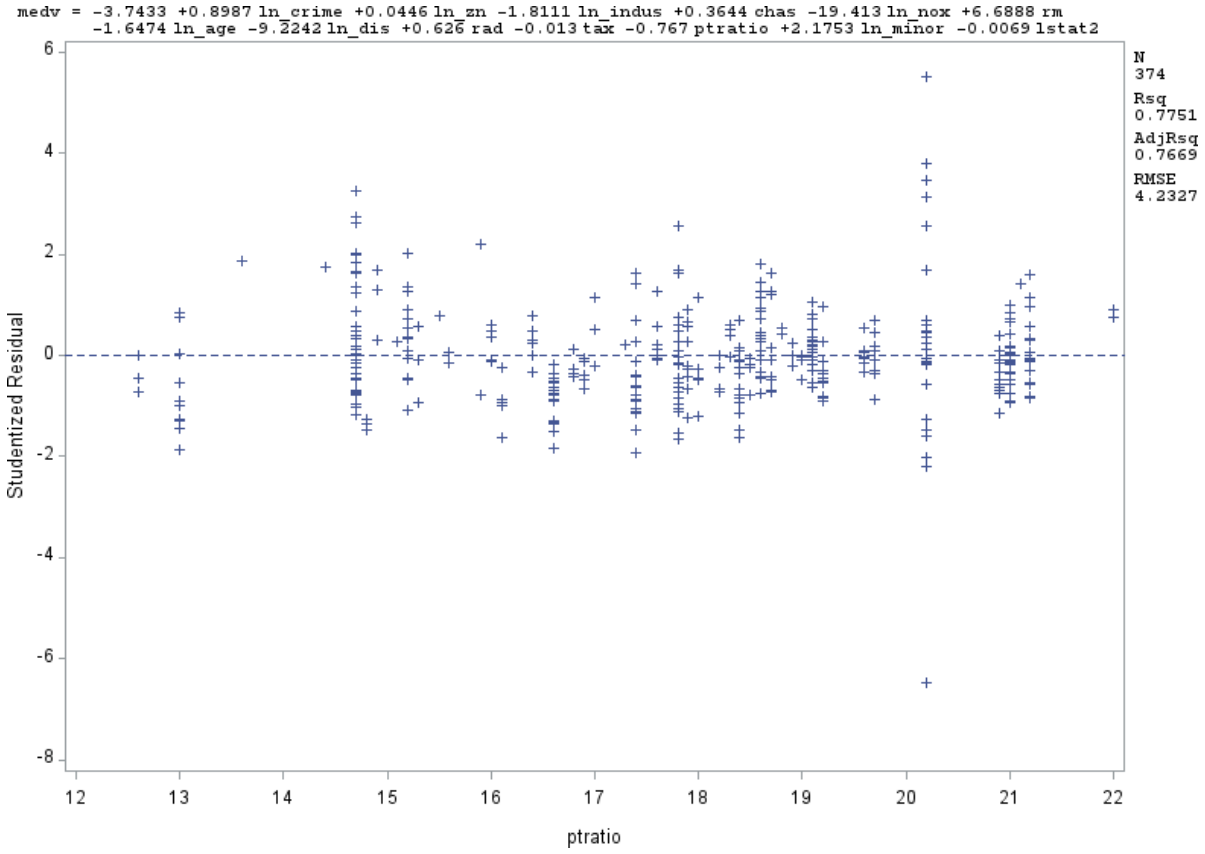

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
rm

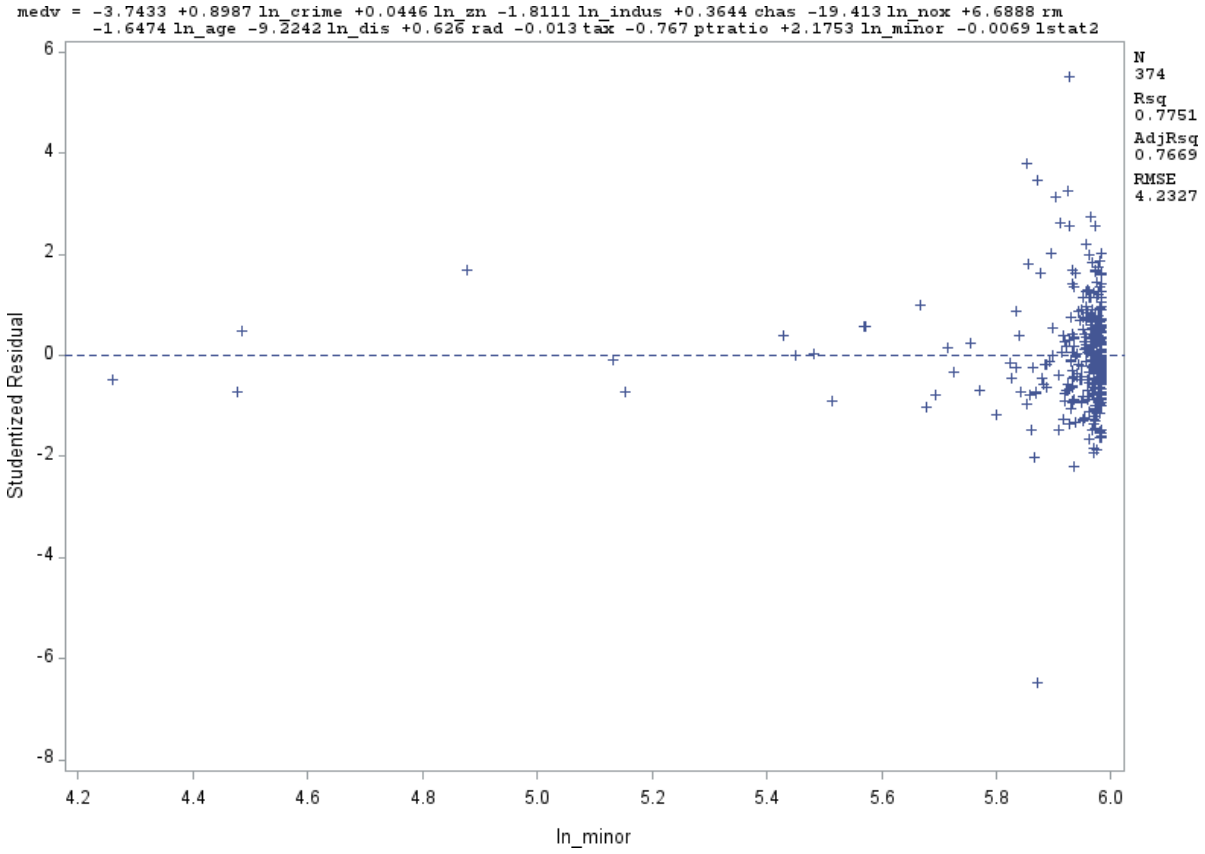

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
ln_age

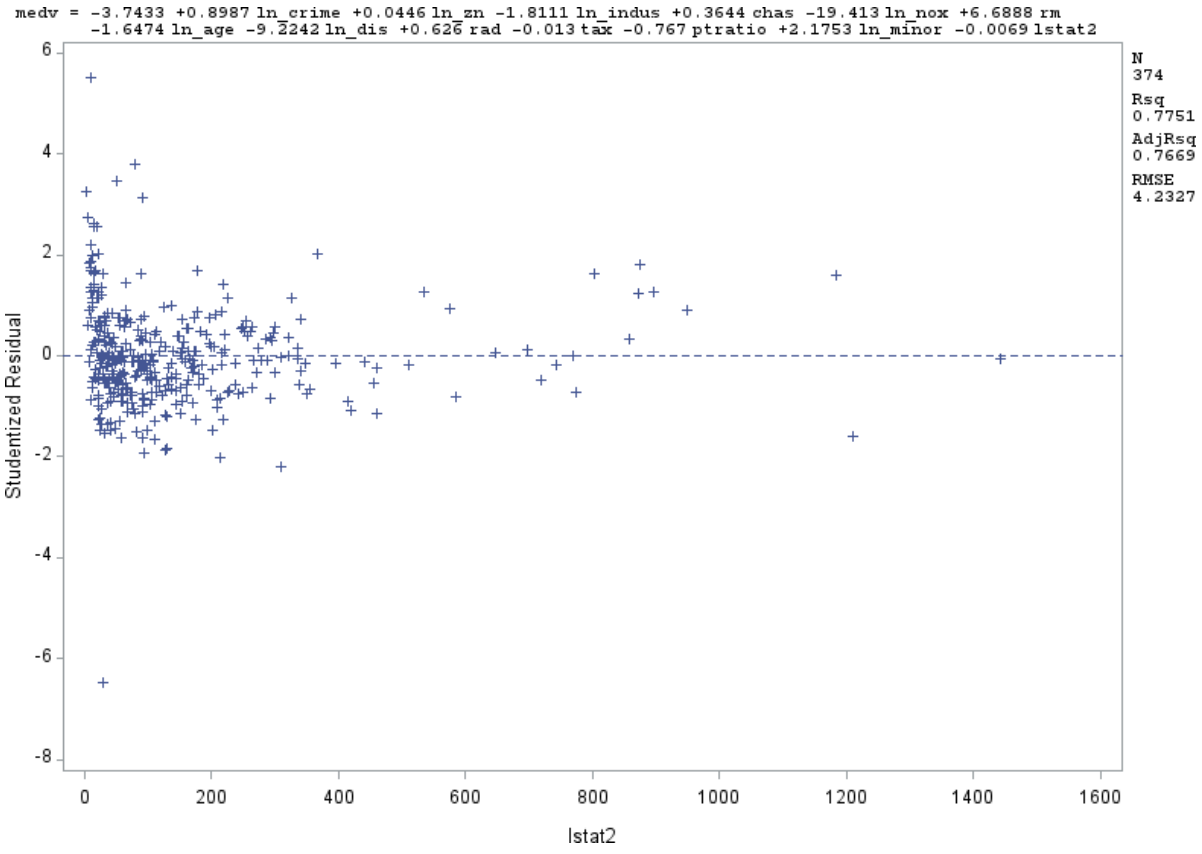

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
ln_dis

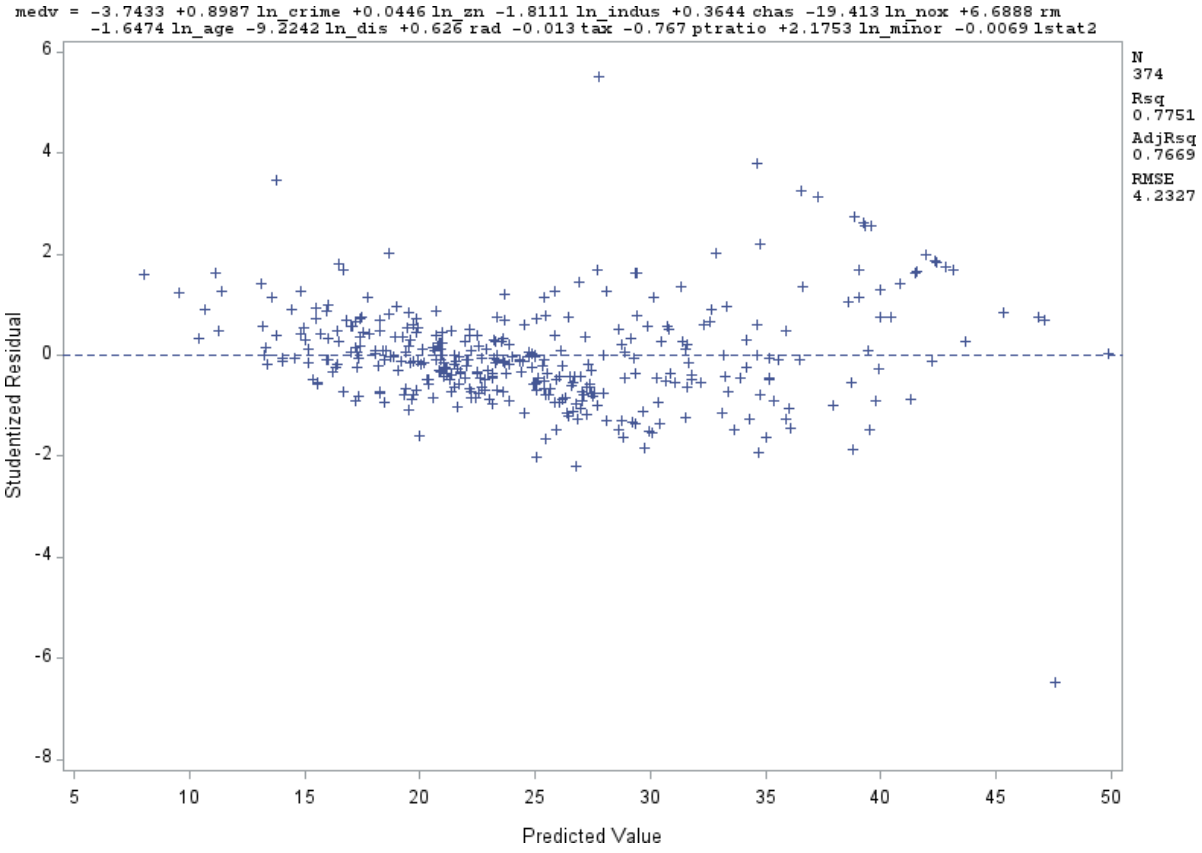

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
rad

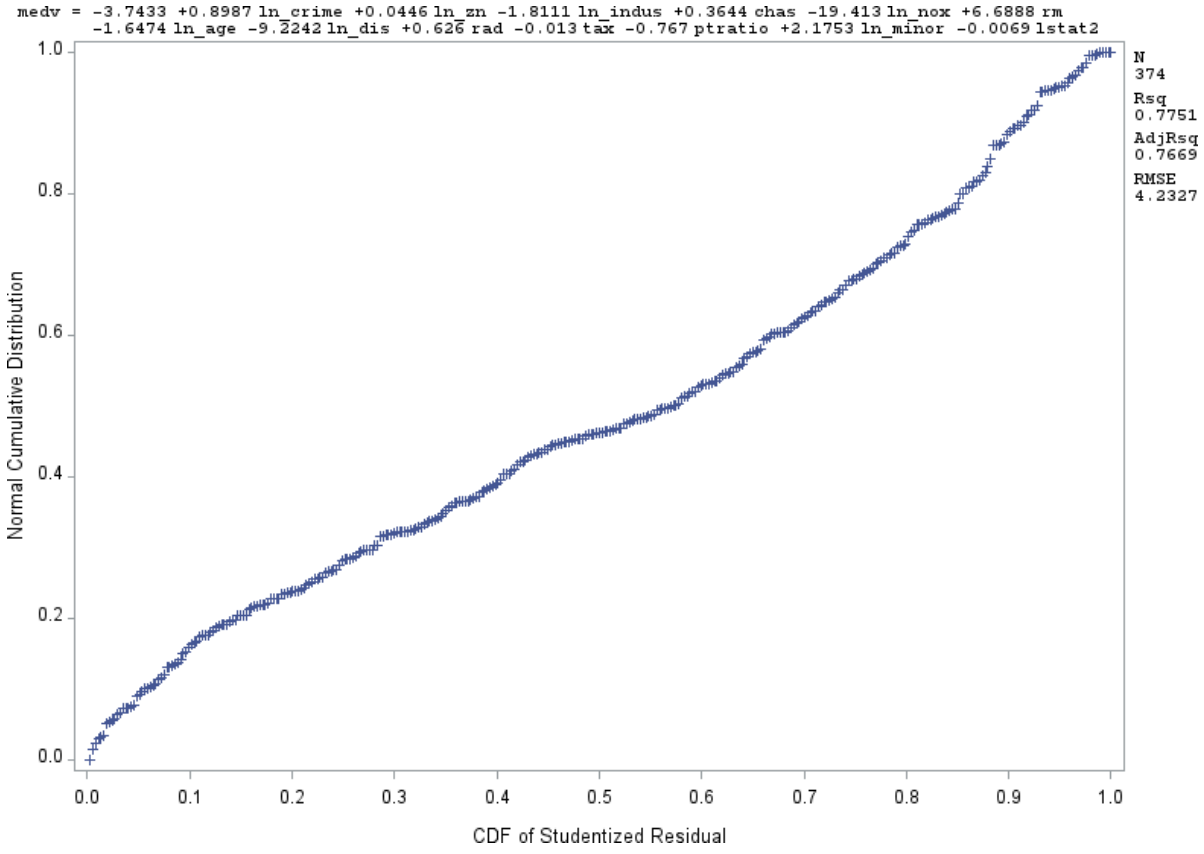

medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
tax


medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
ptratio


medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
ln_minor


medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
lstat2


medv = -3.7433 +0.8987 ln_crime +0.0446 ln_zn -1.8111 ln_indus +0.3644 chas -19.413 ln_nox +6.6888 rm
-1.6474 ln_age -9.2242 ln_dis +0.626 rad -0.013 tax -0.767 ptratio +2.1753 ln_minor -0.0069 lstat2
N
374
Rsq
0.7751
AdjRsq
0.7669
RMSE
4.2327
Studentized Residual
Predicted Value



Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	rm		1	0.5294	0.5294	424.950	313.84	<.0001
2	rm_lstat		2	0.0781	0.6075	310.453	55.32	<.0001
3	ln_dis		3	0.0660	0.6735	214.003	56.00	<.0001
4	lstat2		4	0.0468	0.7203	146.193	46.18	<.0001
5	rad		5	0.0157	0.7360	124.709	16.40	<.0001
6	ln_nox		6	0.0200	0.7560	96.9098	22.44	<.0001
7	rm_ptratio		7	0.0310	0.7870	52.6734	39.73	<.0001
8	ptratio		8	0.0194	0.8064	25.7005	27.30	<.0001
9	tax		9	0.0064	0.8129	18.1073	9.31	0.0025
10	ln_crime		10	0.0015	0.8144	17.8000	2.25	0.1347

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	375
Number of Observations Used	281
Number of Observations with Missing Values	94

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	17506	1750.58760	118.49	<.0001
Error	270	3989.01240	14.77412		
Corrected Total	280	21495			

Root MSE	3.84371	R-Square	0.8144
Dependent Mean	24.91993	Adj R-Sq	0.8075
Coeff Var	15.42425		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-80.43546	15.06157	-5.34	<.0001
rm	1	19.61163	2.26821	8.65	<.0001
rm_lstat	1	-0.15897	0.01878	-8.47	<.0001
ln_dis	1	-6.41842	0.83116	-7.72	<.0001
lstat2	1	0.01482	0.00301	4.92	<.0001
rad	1	0.55195	0.08546	6.46	<.0001
ln_nox	1	-16.00436	2.90212	-5.51	<.0001
rm_ptratio	1	-0.75334	0.12925	-5.83	<.0001
ptratio	1	4.10862	0.84113	4.88	<.0001
tax	1	-0.01195	0.00384	-3.12	0.0020
ln_crime	1	0.42418	0.28275	1.50	0.1347

The SAS System

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	93	3.32987	2.39367

The SAS System

The CORR Procedure

2 Variables:	medv yhat

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
medv	93	25.91720	8.79036	2410	13.10000	50.00000	
yhat	93	25.48415	8.21385	2370	12.58725	46.31393	Predicted Value of new_y

Pearson Correlation Coefficients, N = 93 Prob > |r| under H0: Rho=0		
	medv	yhat
medv	1.00000	0.92552 <.0001
yhat Predicted Value of new_y	0.92552 <.0001	1.00000

The GLMSELECT Procedure

Data Set	WORK.BOS_ANALYSIS
Dependent Variable	medv
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	480937001

Number of Observations Read	375
Number of Observations Used	374
Number of Observations Used for Training	288
Number of Observations Used for Testing	86

Dimensions	
Number of Effects	17
Number of Parameters	17

The SAS System

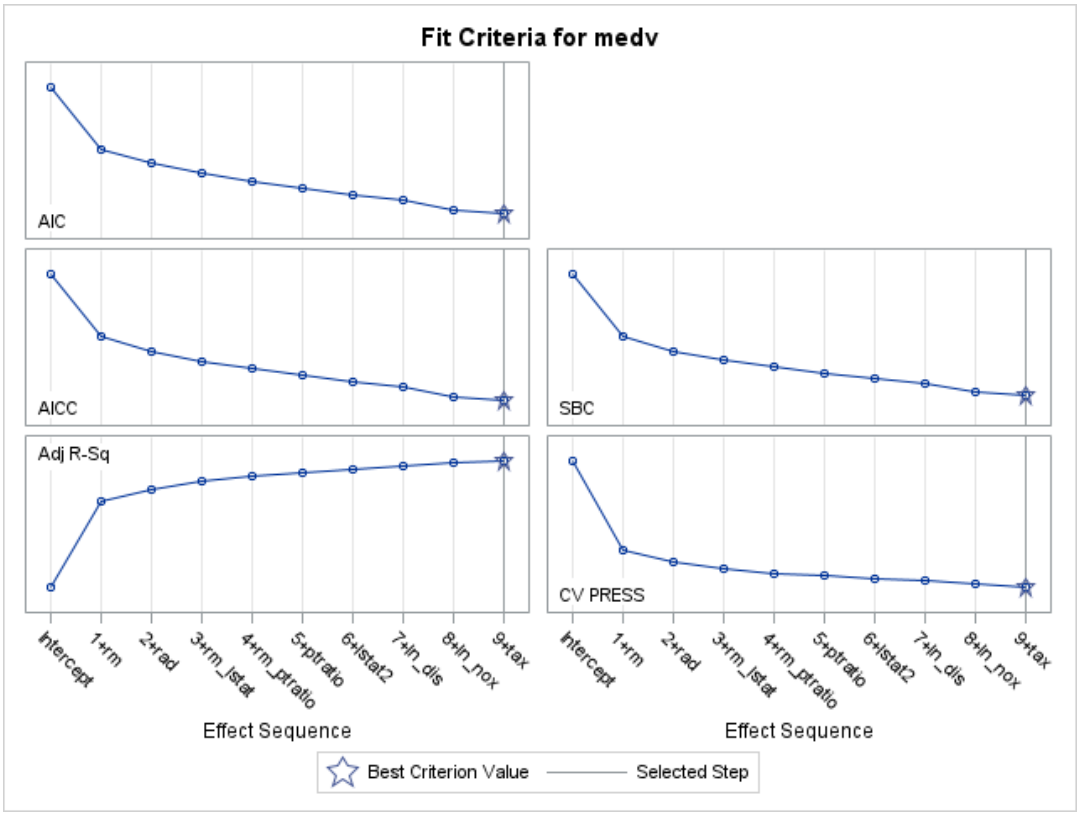
The GLMSELECT Procedure

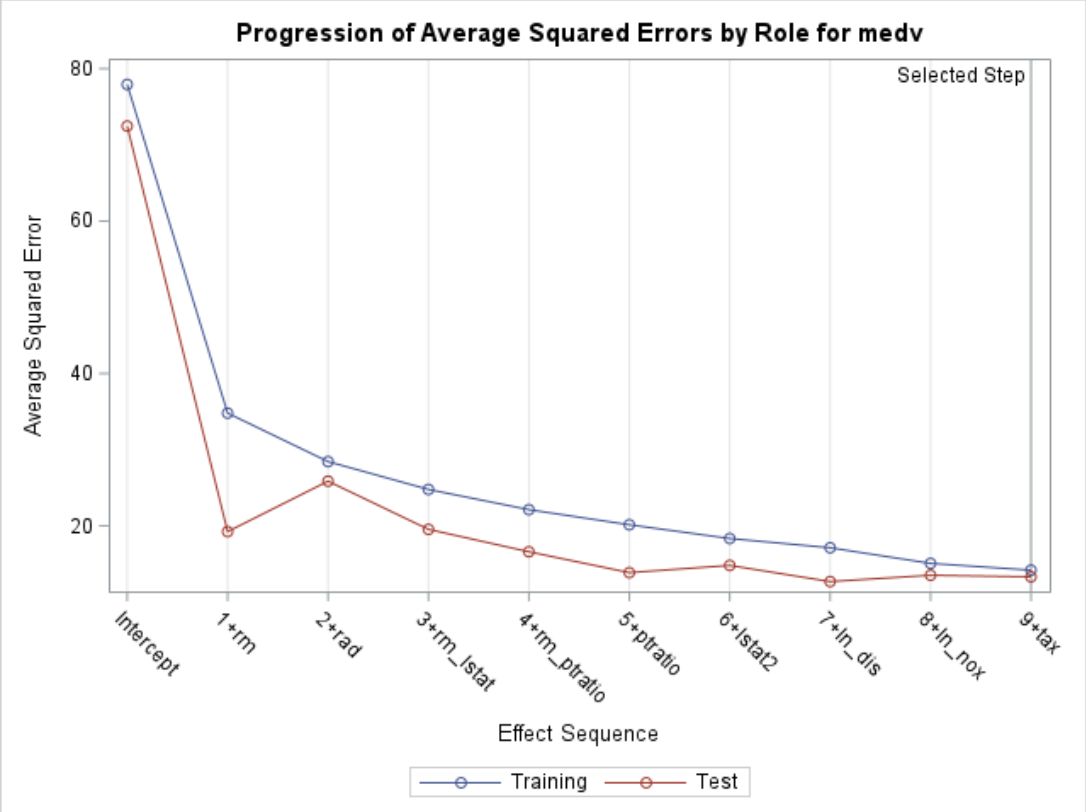
Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0	Intercept		1	1260.1238	77.9265	72.4644	22559.6127
1	rm		2	1033.5884	34.7967	19.2463	10147.1347
2	rad		3	981.1542	28.4400	25.8682	8647.7204
3	rm_lstat		4	947.1606	24.7816	19.5451	7653.0625
4	rm_ptratio		5	920.2195	22.1290	16.6042	6939.5281
5	ptratio		6	898.8746	20.1482	13.8657	6647.8671
6	lstat2		7	877.5261	18.3444	14.8134	6246.5416
7	ln_dis		8	863.4295	17.1280	12.6926	5863.3934
8	ln_nox		9	832.4114	15.0797	13.5220	5425.9615
9	tax		10	820.6166*	14.1928	13.3119	5075.2989*

* Optimal Value of Criterion

Selection stopped as adding or dropping any effect does not improve the selection criterion.


Fit Criteria for medv
AIC
AICC
SBC
Adj R-Sq
CV PRESS
Intercept
1+rm
2+rad
3+rm_lstat
4+rm_ptratio
5+ptratio
6+lstat2
7+ln_dis
8+ln_nox
9+tax
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors by Role for medv
Selected Step
Average Squared Error
80
60
40
20
Intercept
1+rm
2+rad
3+rm_lstat
4+rm_ptratio
5+ptratio
6+lstat2
7+ln_dis
8+ln_nox
9+tax
Effect Sequence
Training
Test

The selected model is the model at the last step (Step 9).

Effects:	Intercept ln_nox rm ln_dis rad tax ptratio lstat2 rm_lstat rm_ptratio

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	9	18355	2039.48097	138.71
Error	278	4087.51735	14.70330	
Corrected Total	287	22443		

Root MSE	3.83449
Dependent Mean	25.23194
R-Square	0.8179
Adj R-Sq	0.8120
AIC	1073.98698
AICC	1074.94350
SBC	820.61659
ASE (Train)	14.19277
ASE (Test)	13.31187
CV PRESS	5075.29891

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	230	58	709.4078
2	230	58	1128.1569
3	230	58	1088.6299
4	231	57	1478.1177
5	231	57	670.9865
Total			5075.2989

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	-84.399189	15.279181	-5.52	-88.8830	-98.4750	-92.3295	-44.5330	-95.1553
ln_nox	1	-14.929033	2.600798	-5.74	-13.7941	-15.2988	-16.6260	-13.5389	-16.4529
rm	1	20.237789	2.312550	8.75	21.1466	22.6412	21.4783	13.4881	21.8406
ln_dis	1	-6.853363	0.834850	-8.21	-6.7704	-7.1372	-7.2115	-6.2572	-7.0440
rad	1	0.683327	0.083577	8.18	0.6508	0.5972	0.7373	0.7934	0.6847
tax	1	-0.015749	0.003779	-4.17	-0.0172	-0.0144	-0.0161	-0.0145	-0.0151
ptratio	1	4.388020	0.850387	5.16	4.7462	5.4762	4.7942	1.7500	4.9069
lstat2	1	0.019474	0.003620	5.38	0.0200	0.0201	0.0291	0.0158	0.0159
rm_lstat	1	-0.174790	0.021170	-8.26	-0.1863	-0.1864	-0.2112	-0.1458	-0.1537
rm_ptratio	1	-0.782954	0.130775	-5.99	-0.8335	-0.9601	-0.8452	-0.3578	-0.8756

The GLMSELECT Procedure

Data Set	WORK.BOS_ANALYSIS
Dependent Variable	medv
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	481359001

Number of Observations Read	375
Number of Observations Used	374
Number of Observations Used for Training	280
Number of Observations Used for Testing	94

Dimensions	
Number of Effects	17
Number of Parameters	17

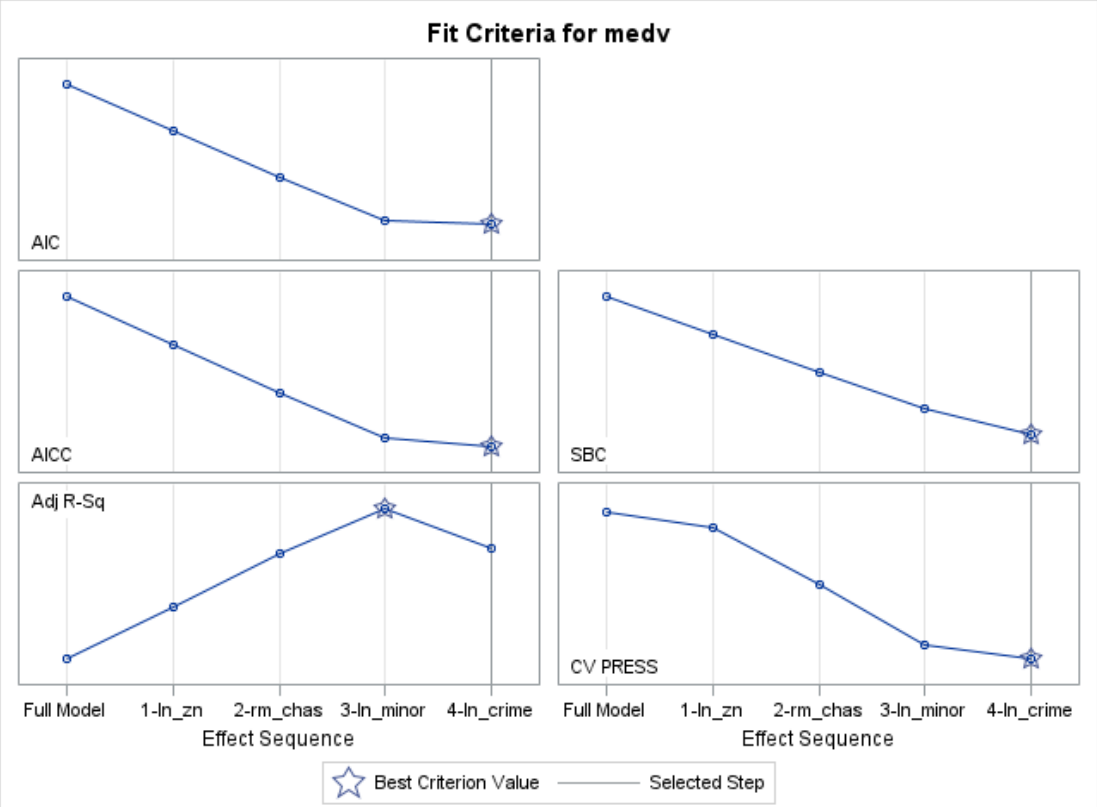
The SAS System

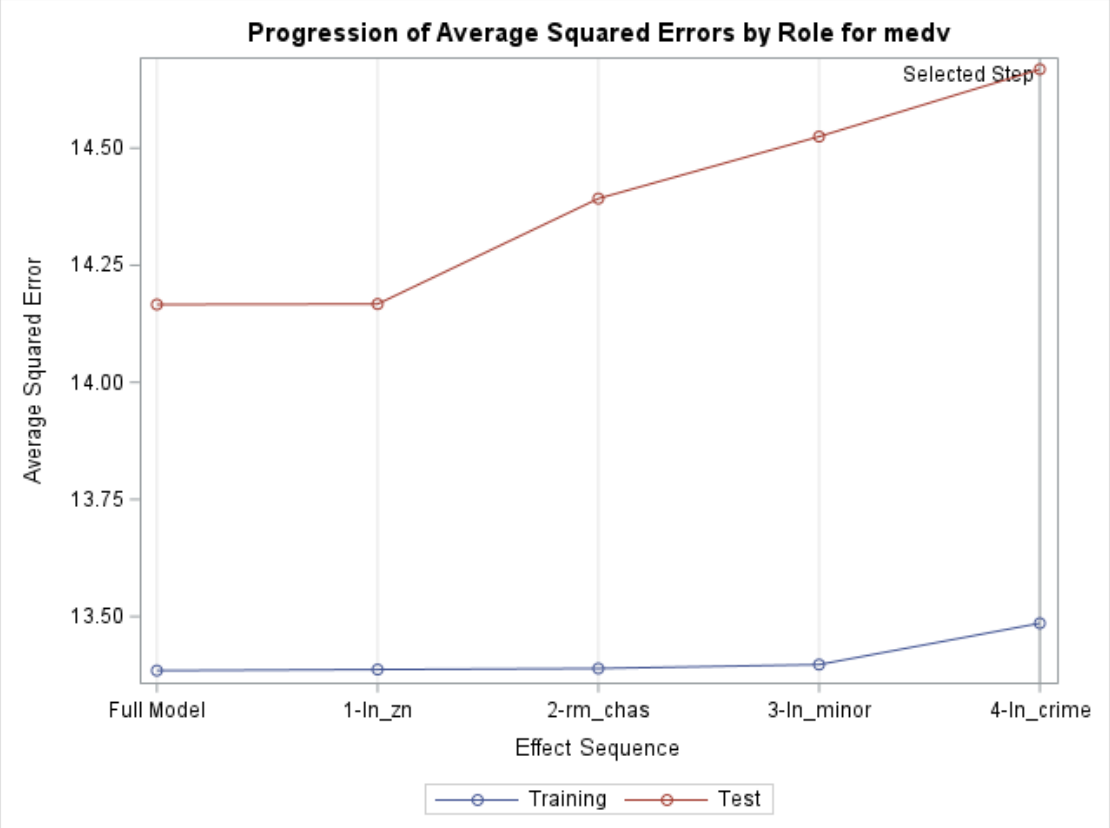
The GLMSELECT Procedure

Backward Selection Summary						
Step	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0		17	822.1393	13.3845	14.1661	4317.0121
1	ln_zn	16	816.5524	13.3868	14.1671	4306.8011
2	rm_chas	15	810.9643	13.3891	14.3923	4269.4588
3	ln_minor	14	805.5032	13.3974	14.5247	4229.3126
4	ln_crime	13	801.7055*	13.4856	14.6684	4220.4630*
* Optimal Value of Criterion						

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details				
Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	ln_indus	4231.7615	>	4220.4630


Fit Criteria for medv
AIC
AICC
SBC
Adj R-Sq
CV PRESS
Full Model
1-ln_zn
2-rm_chas
3-ln_minor
4-ln_crime
Effect Sequence
Best Criterion Value
Selected Step

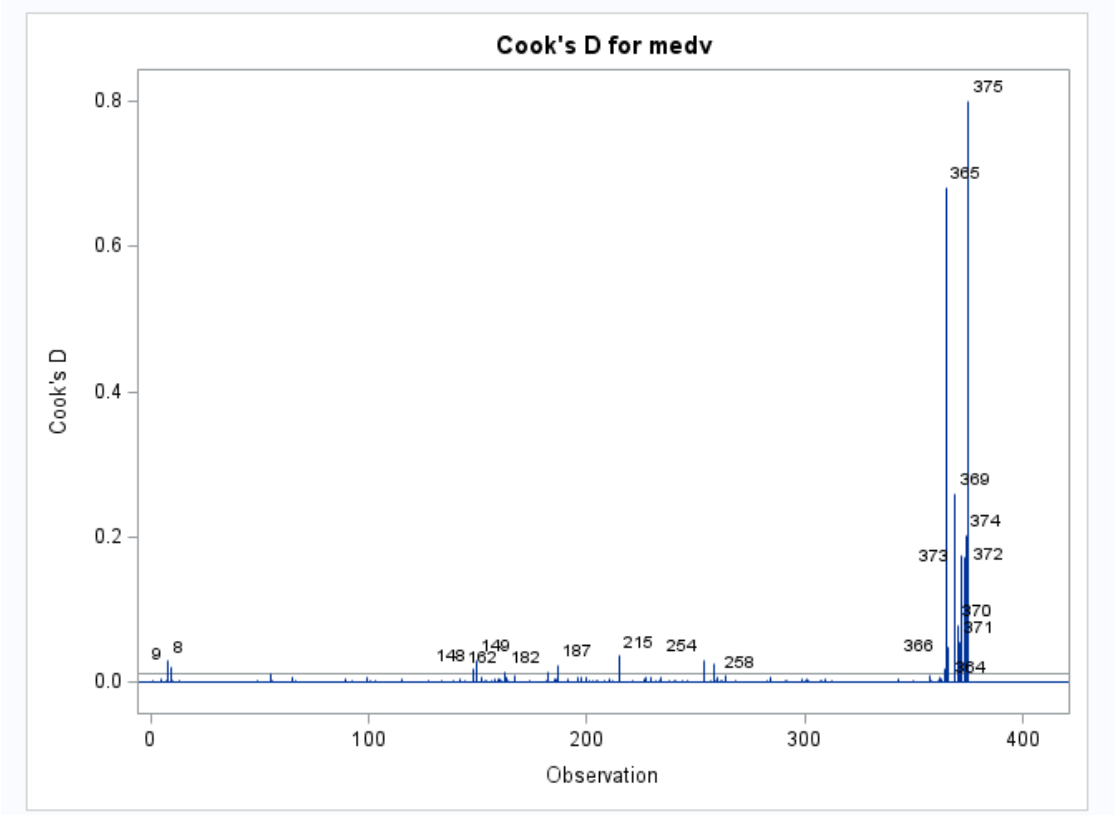

Progression of Average Squared Errors by Role for medv
Selected Step
Average Squared Error
14.50
14.25
14.00
13.75
13.50
Full Model
1-ln_zn
2-rm_chas
3-ln_minor
4-ln_crime
Effect Sequence
Training
Test

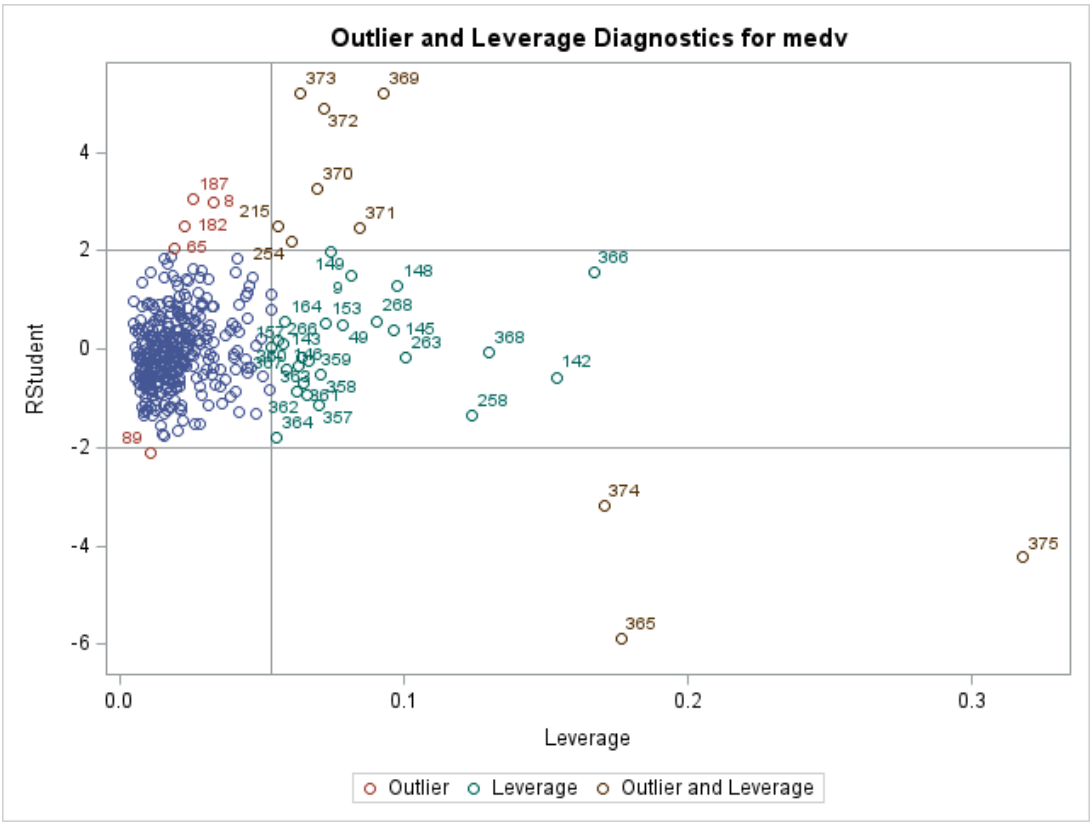
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	12	18307	1525.56663	107.87
Error	267	3775.95444	14.14215	
Corrected Total	279	22083		

Root MSE	3.76061
Dependent Mean	25.07964
R-Square	0.8290
Adj R-Sq	0.8213
AIC	1036.45328
AICC	1038.03819
SBC	801.70555
ASE (Train)	13.48555
ASE (Test)	14.66835
CV PRESS	4220.46305

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	224	56	886.4086
2	224	56	739.9622
3	224	56	706.7242
4	224	56	940.7288
5	224	56	946.6392
Total			4220.4630

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	-63.690417	16.399332	-3.88	-71.28500	-58.0642	-68.0683	-56.4054	-61.75544
ln_indus	1	-0.503709	0.472673	-1.07	-0.52923	-0.5305	-0.2651	-0.4394	-0.70805
chas	1	2.493535	0.850376	2.93	2.94378	2.7963	1.6348	2.8442	2.37026
ln_nox	1	-12.078572	2.684352	-4.50	-11.08348	-10.2363	-12.3348	-13.6218	-13.40352
rm	1	17.667292	2.529018	6.99	19.21631	17.1330	18.3679	16.1138	16.98523
ln_age	1	-0.734250	0.539822	-1.36	-1.18658	-0.9307	-0.6532	-0.4978	-0.30037
ln_dis	1	-7.043043	0.909086	-7.75	-7.15495	-6.8300	-6.8845	-7.0259	-7.23506
rad	1	0.578631	0.081076	7.14	0.59399	0.5166	0.6311	0.5917	0.57023
tax	1	-0.011160	0.003741	-2.98	-0.01007	-0.0113	-0.0154	-0.0108	-0.00836
ptratio	1	3.392387	0.921507	3.68	3.97227	3.1172	3.6604	2.6658	3.33141
lstat2	1	0.010744	0.002945	3.65	0.00826	0.0108	0.0104	0.0115	0.01458
rm_lstat	1	-0.137159	0.020239	-6.78	-0.12704	-0.1401	-0.1354	-0.1306	-0.16140


Cook's D for medv
Cook's D
Observation
0.0
0.2
0.4
0.6
0.8
0
100
200
300
400
375
365
369
374
373
372
370
371
366
364
9
8
148
162
149
182
187
215
254
258


Outlier and Leverage Diagnostics for medv
RStudent
Leverage
0.0
0.1
0.2
0.3
4
2
0
-2
-4
-6
Outlier
Leverage
Outlier and Leverage
373
369
372
370
187
8
215
371
182
65
254
149
148
366
9
164
153
268
145
263
368
142
49
359
358
362
364
357
258
89
374
375
365

The REG Procedure
Model: MODEL1
Dependent Variable: medv

Number of Observations Read	375
Number of Observations Used	374
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	23685	2631.63881	192.02	<.0001
Error	364	4988.50568	13.70469		
Corrected Total	373	28673			

Root MSE	3.70198	R-Square	0.8260
Dependent Mean	25.16791	Adj R-Sq	0.8217
Coeff Var	14.70914		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-76.56739	12.82434	-5.97	<.0001	0	0
rm	1	19.27412	1.94854	9.89	<.0001	1.58858	53.96282
rm_lstat	1	-0.15592	0.01606	-9.71	<.0001	-0.56904	7.18646
ln_dis	1	-6.45478	0.71206	-9.06	<.0001	-0.38071	3.69046
lstat2	1	0.01369	0.00253	5.40	<.0001	0.29865	6.40044
rad	1	0.54359	0.06941	7.83	<.0001	0.28343	2.74054
ln_nox	1	-12.66416	2.17894	-5.81	<.0001	-0.27332	4.62702
rm_ptratio	1	-0.73338	0.11160	-6.57	<.0001	-1.28963	80.57726
ptratio	1	4.01055	0.72279	5.55	<.0001	1.01322	69.76508
tax	1	-0.01349	0.00315	-4.28	<.0001	-0.15732	2.83184

The REG Procedure
Model: MODEL1
Dependent Variable: medv

Number of Observations Read	375
Number of Observations Used	374
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	22393	2799.10107	162.68	<.0001
Error	365	6280.44641	17.20670		
Corrected Total	373	28673			

Root MSE	4.14810	R-Square	0.7810
Dependent Mean	25.16791	Adj R-Sq	0.7762
Coeff Var	16.48168		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-94.01703	14.22795	-6.61	<.0001	0	0
rm	1	20.89400	2.17534	9.60	<.0001	1.72209	53.56721
ln_dis	1	-6.98226	0.79554	-8.78	<.0001	-0.41183	3.66898
lstat2	1	-0.00751	0.00144	-5.21	<.0001	-0.16390	1.65185
rad	1	0.69695	0.07574	9.20	<.0001	0.36338	2.59865
ln_nox	1	-18.28950	2.35362	-7.77	<.0001	-0.39473	4.29988
rm_ptratio	1	-0.80385	0.12478	-6.44	<.0001	-1.41355	80.23645
ptratio	1	4.24478	0.80944	5.24	<.0001	1.07240	69.68736
tax	1	-0.01300	0.00353	-3.68	0.0003	-0.15165	2.83112

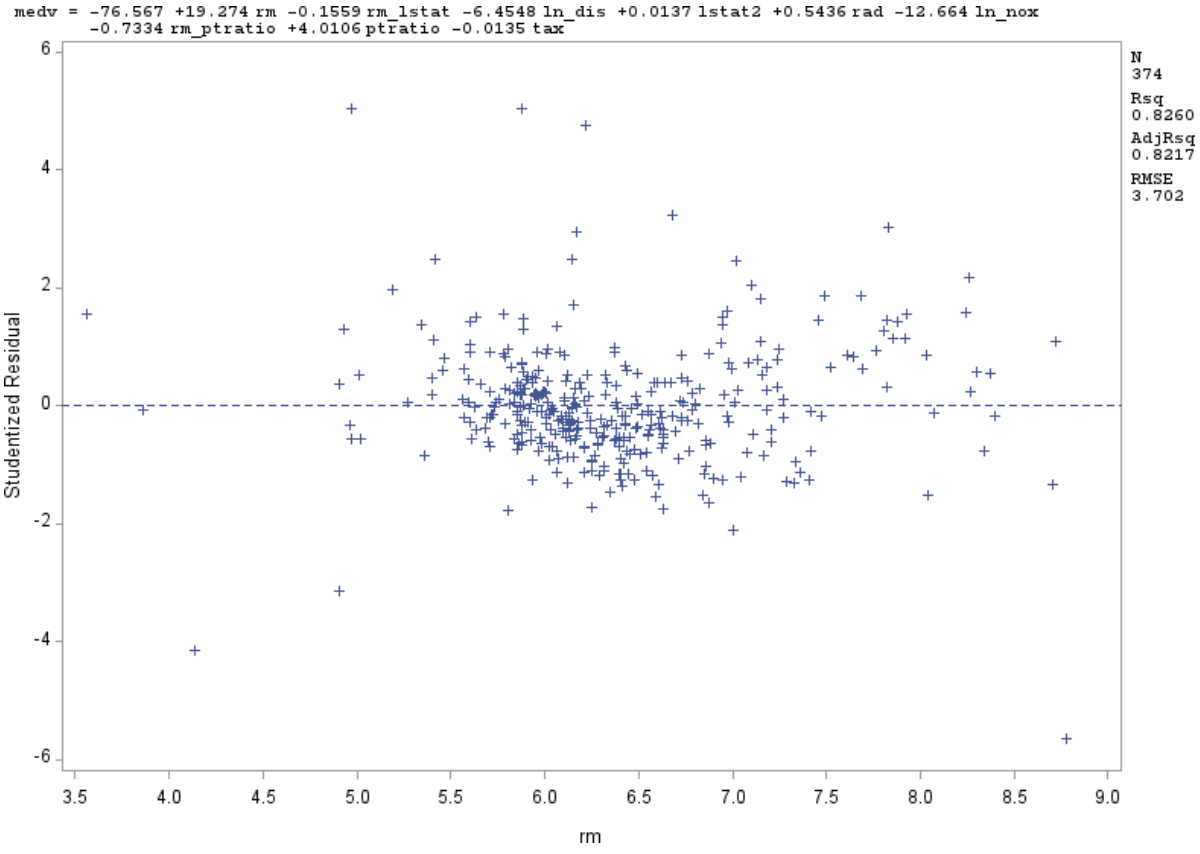
The REG Procedure
Model: MODEL1
Dependent Variable: medv

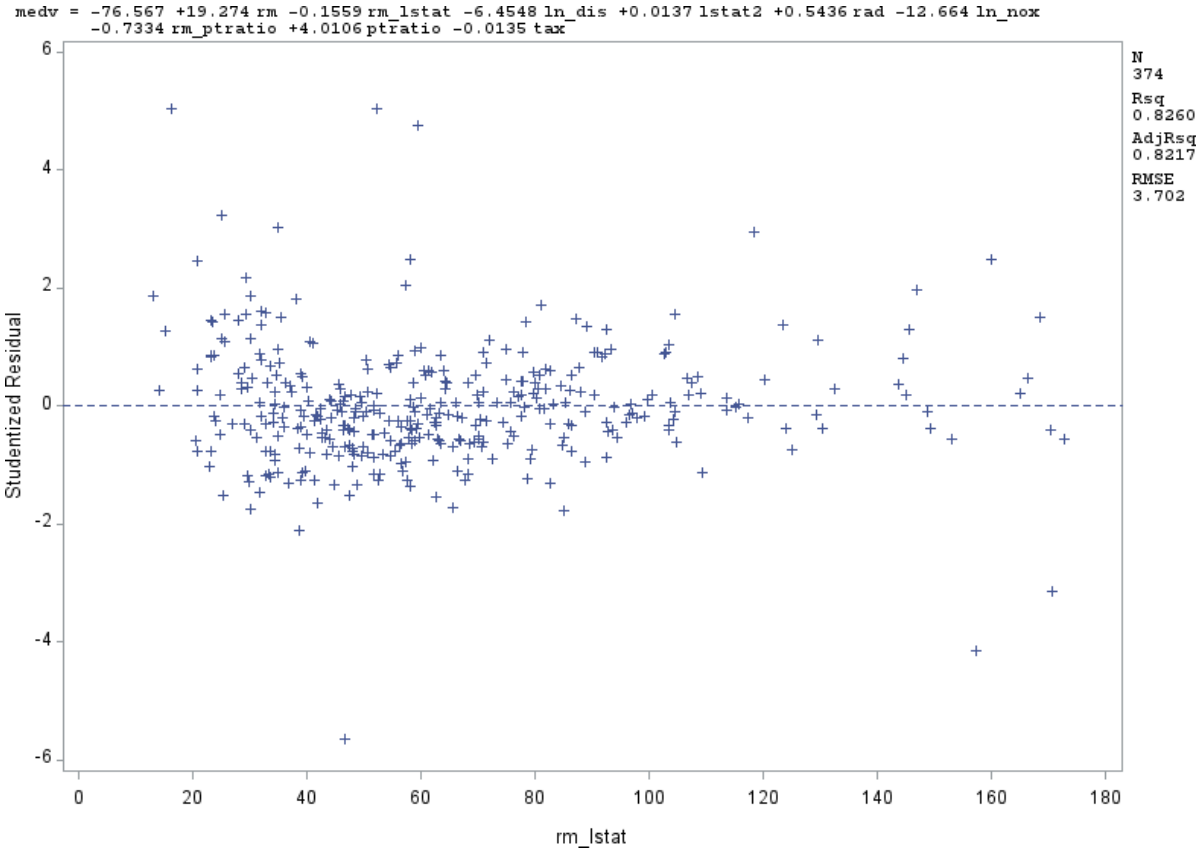
Number of Observations Read	375
Number of Observations Used	374
Number of Observations with Missing Values	1

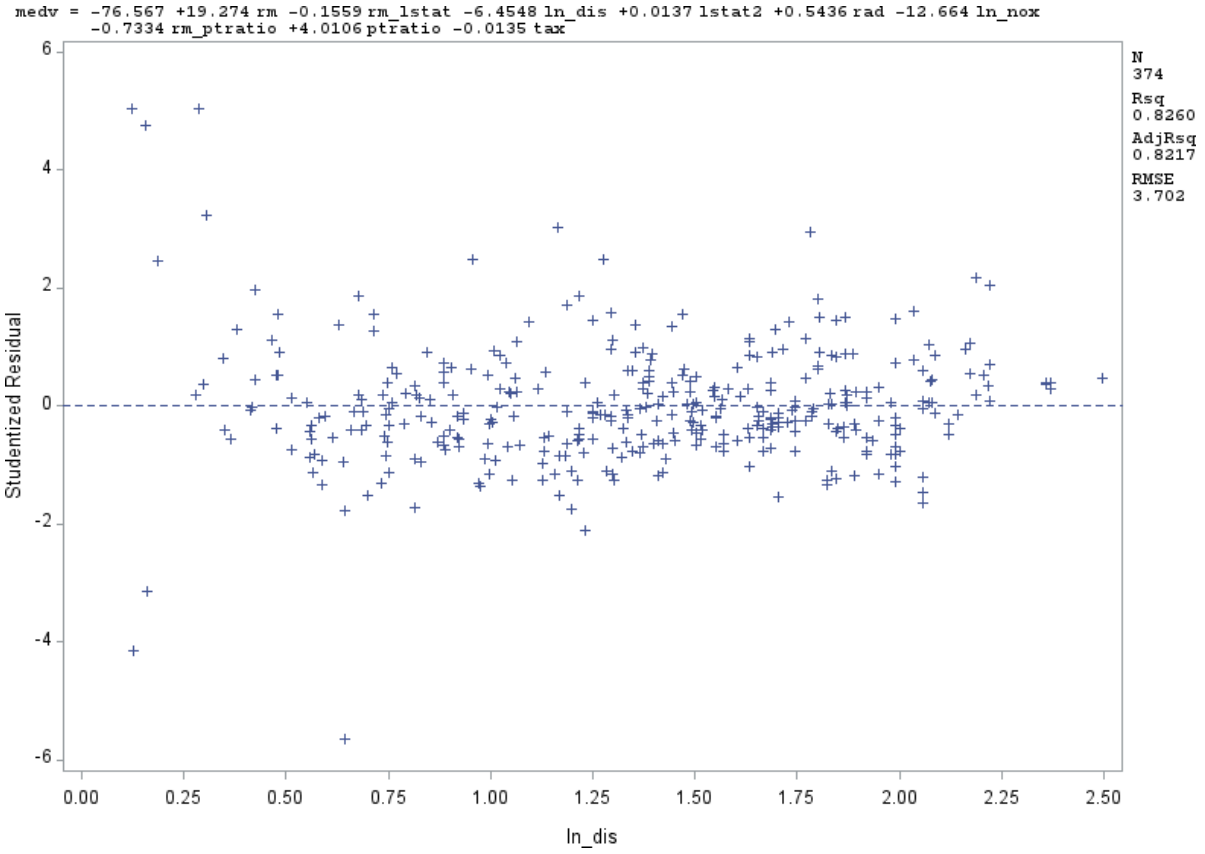
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	21679	3096.96591	162.05	<.0001
Error	366	6994.49360	19.11064		
Corrected Total	373	28673			

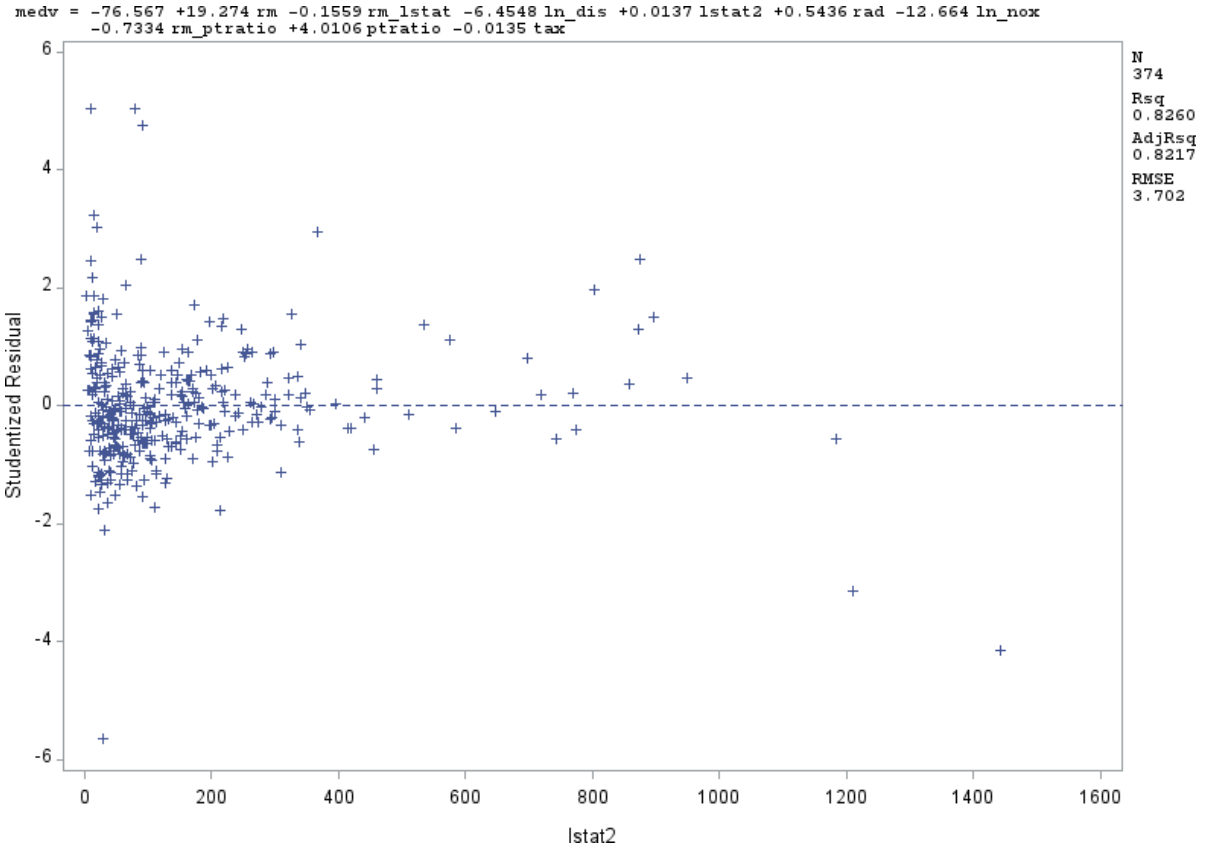
Root MSE	4.37157	R-Square	0.7561
Dependent Mean	25.16791	Adj R-Sq	0.7514
Coeff Var	17.36962		

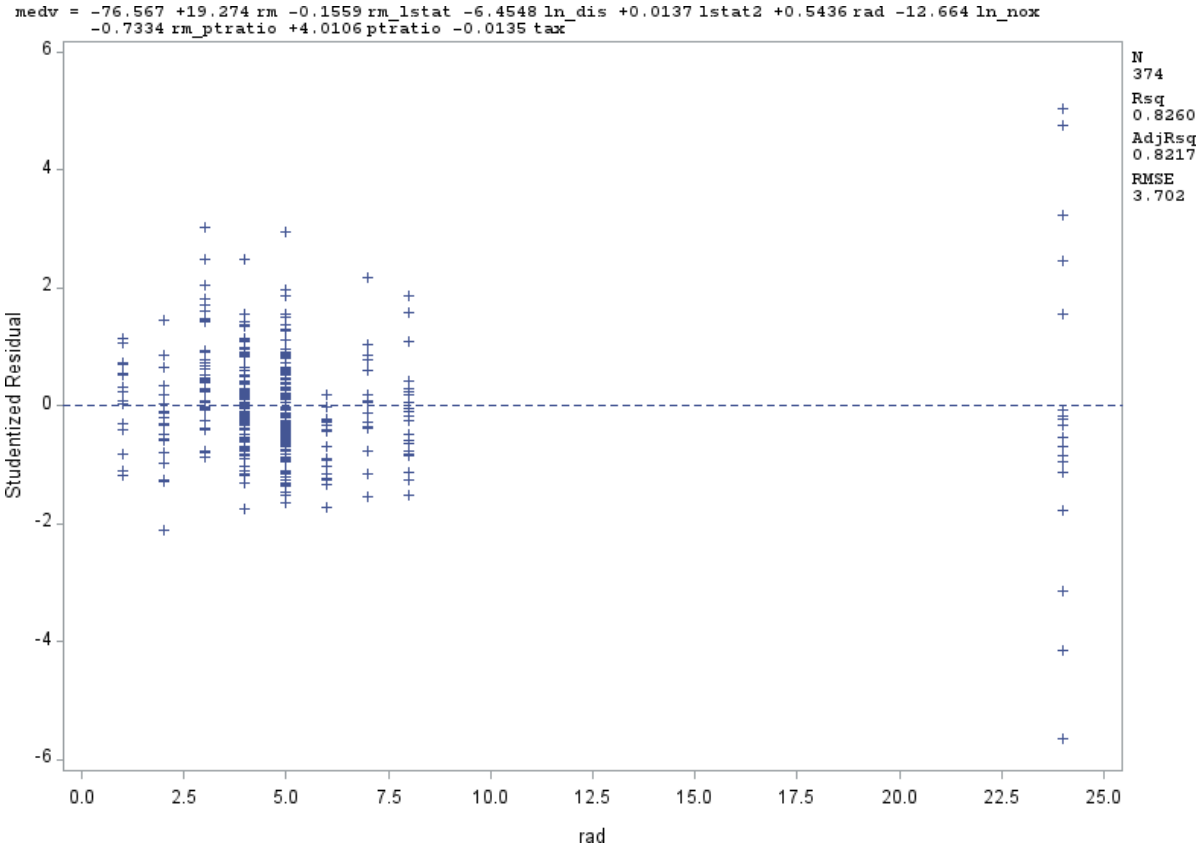
Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-5.87409	4.11108	-1.43	0.1539	0	0
rm	1	7.10149	0.40536	17.52	<.0001	0.58531	1.67479
ln_dis	1	-7.71714	0.82974	-9.30	<.0001	-0.45517	3.59354
lstat2	1	-0.00669	0.00151	-4.42	<.0001	-0.14608	1.63909
rad	1	0.75608	0.07923	9.54	<.0001	0.39421	2.56048
ln_nox	1	-20.73307	2.44799	-8.47	<.0001	-0.44747	4.18819
ptratio	1	-0.92248	0.11440	-8.06	<.0001	-0.23305	1.25323
tax	1	-0.01333	0.00372	-3.58	0.0004	-0.15547	2.83054

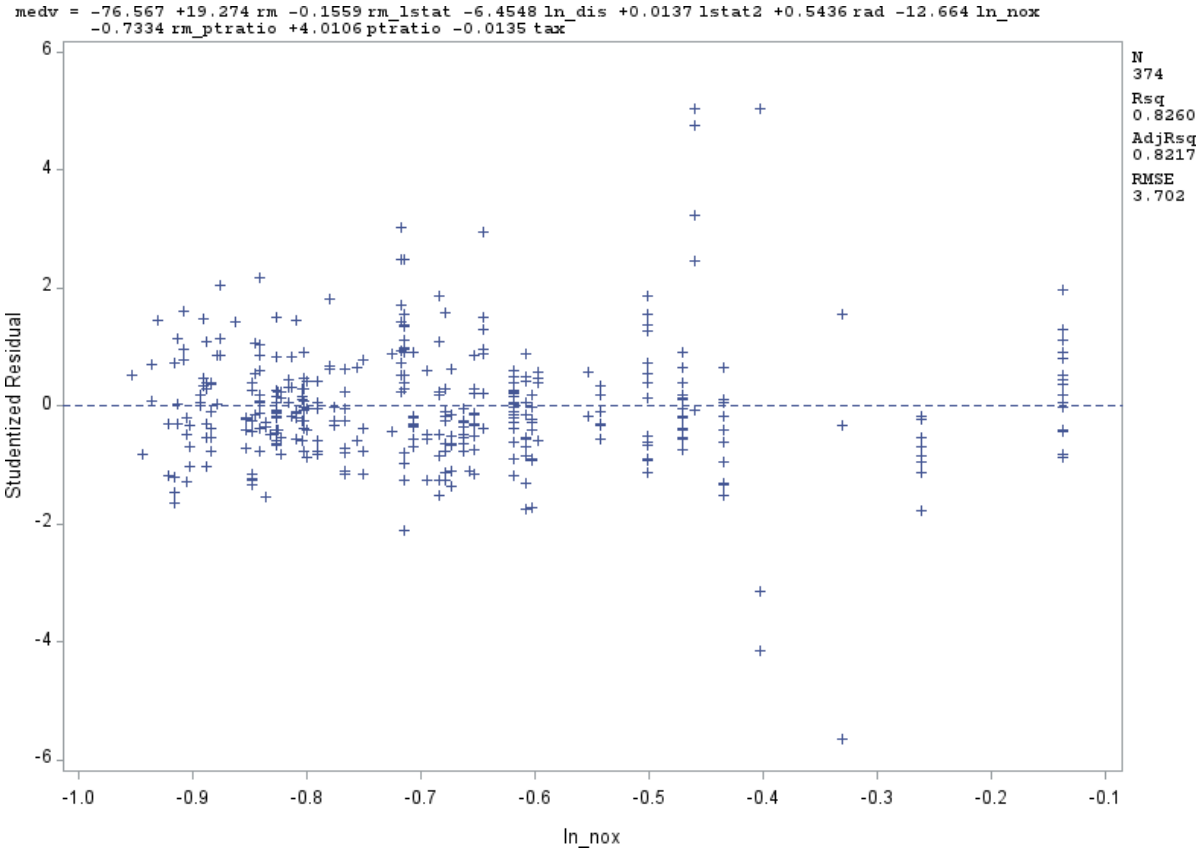

medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
rm

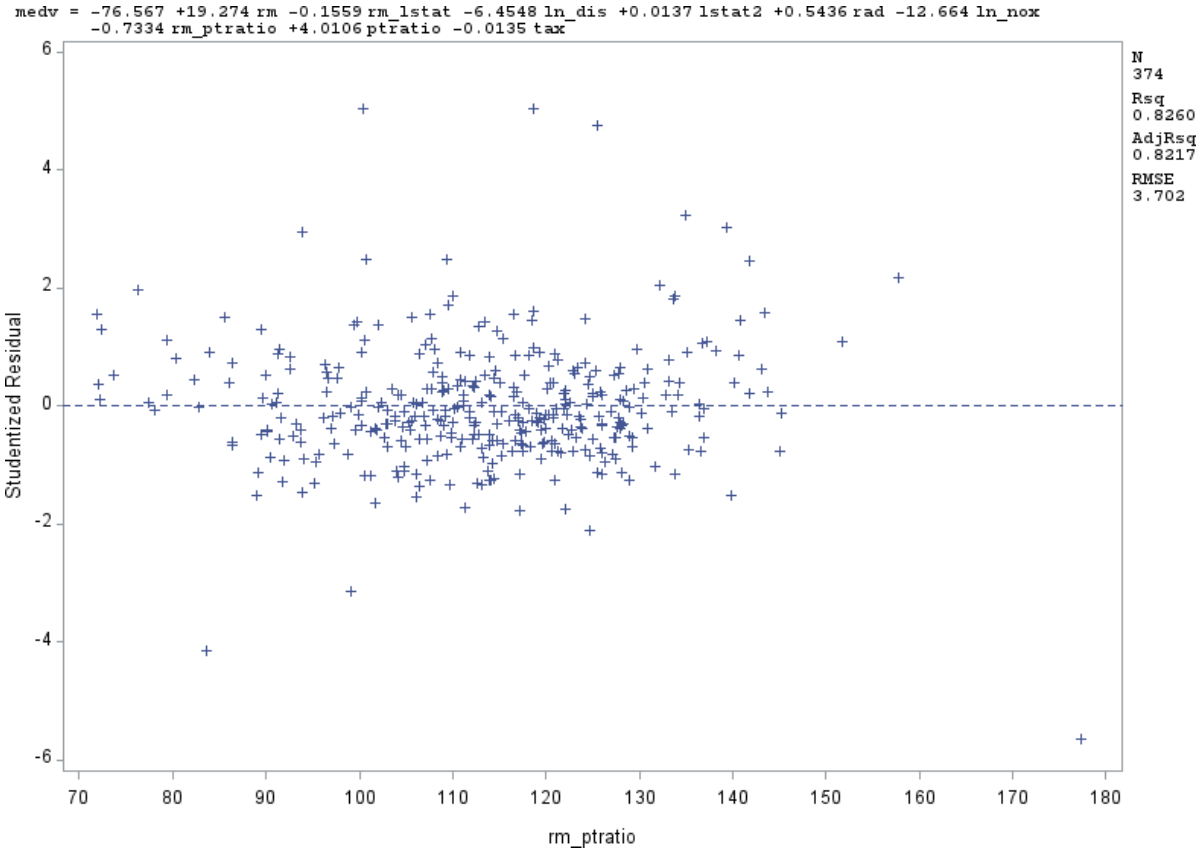

medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
rm_lstat

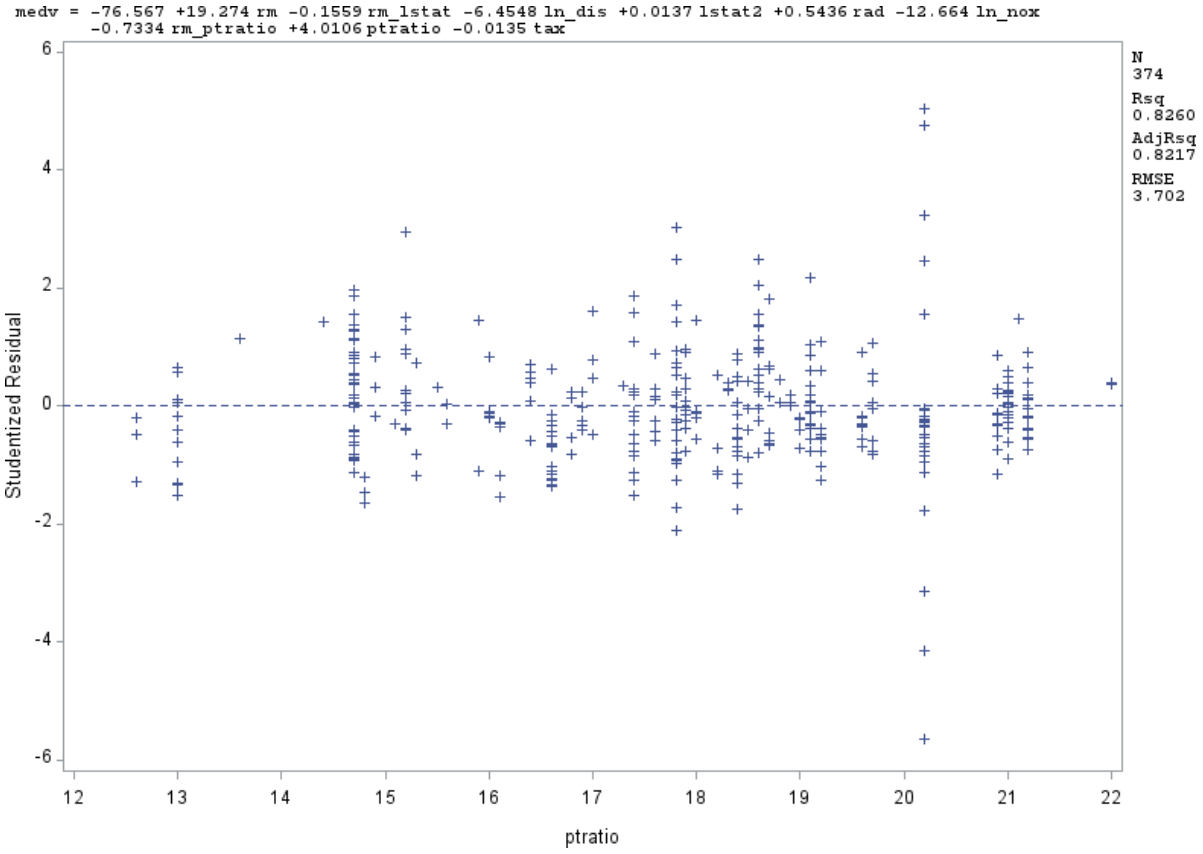

medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
6
4
2
0
-2
-4
-6
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
ln_dis

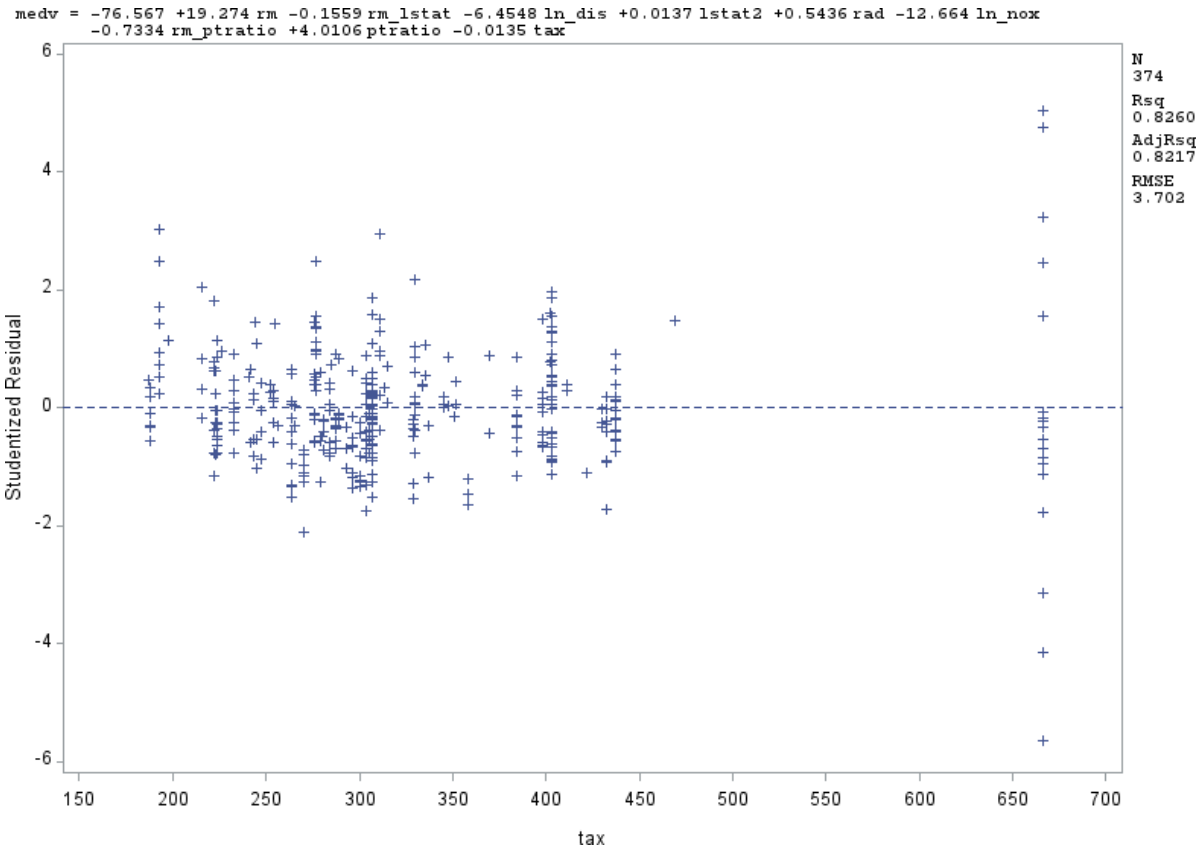

medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
6
4
2
0
-2
-4
-6
0
200
400
600
800
1000
1200
1400
1600
lstat2

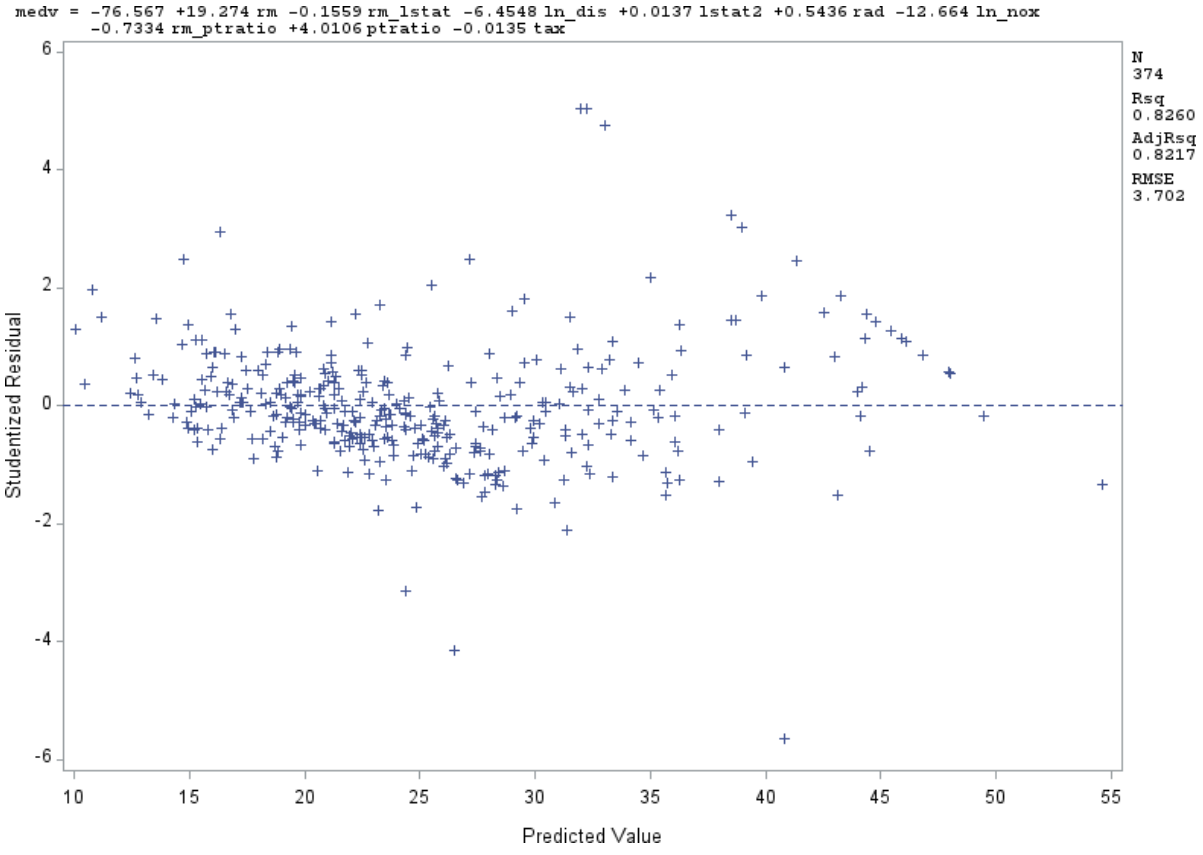

medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
rad

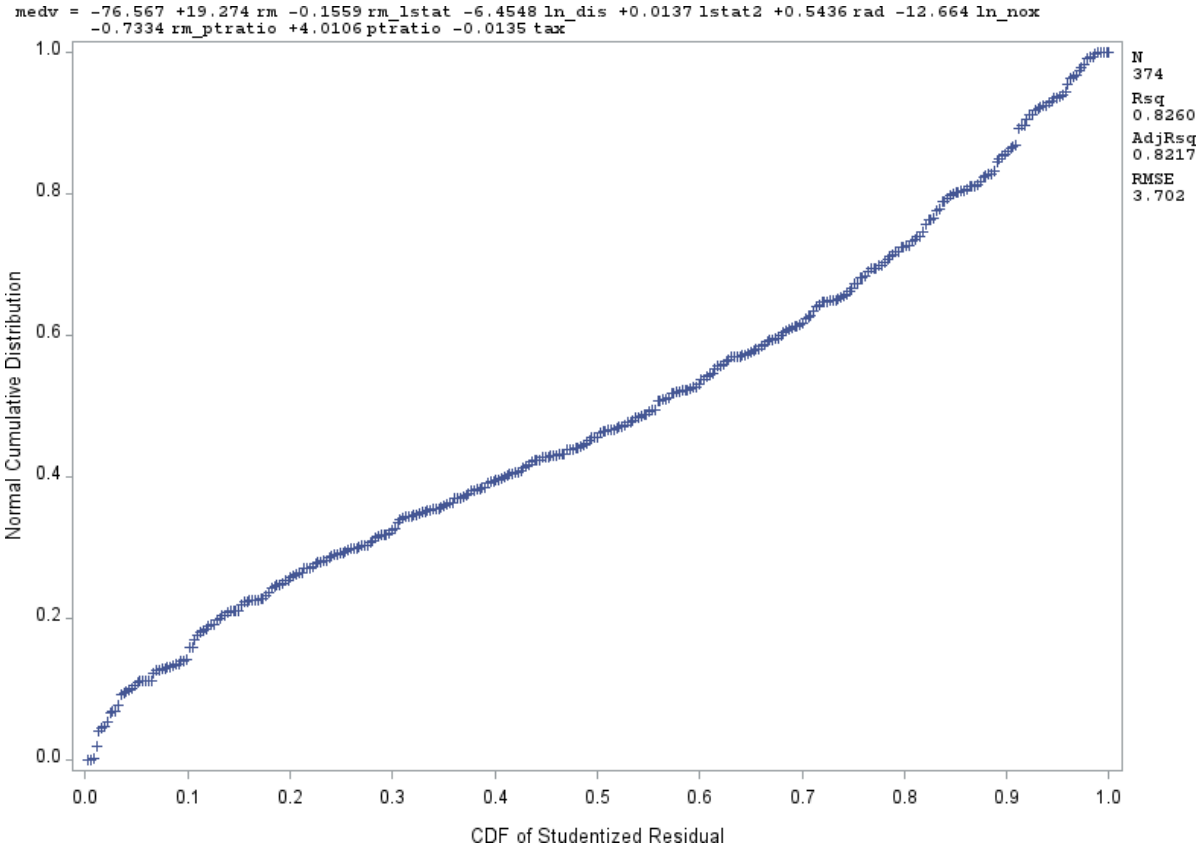

medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
ln_nox


medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
rm_ptratio


medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
ptratio


medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
tax


medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Studentized Residual
Predicted Value


medv = -76.567 +19.274 rm -0.1559 rm_lstat -6.4548 ln_dis +0.0137 lstat2 +0.5436 rad -12.664 ln_nox
-0.7334 rm_ptratio +4.0106 ptratio -0.0135 tax
N
374
Rsq
0.8260
AdjRsq
0.8217
RMSE
3.702
Normal Cumulative Distribution
CDF of Studentized Residual
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7
0.9

The REG Procedure
Model: MODEL1
Dependent Variable: medv

Number of Observations Read	355
Number of Observations Used	354
Number of Observations with Missing Values	1

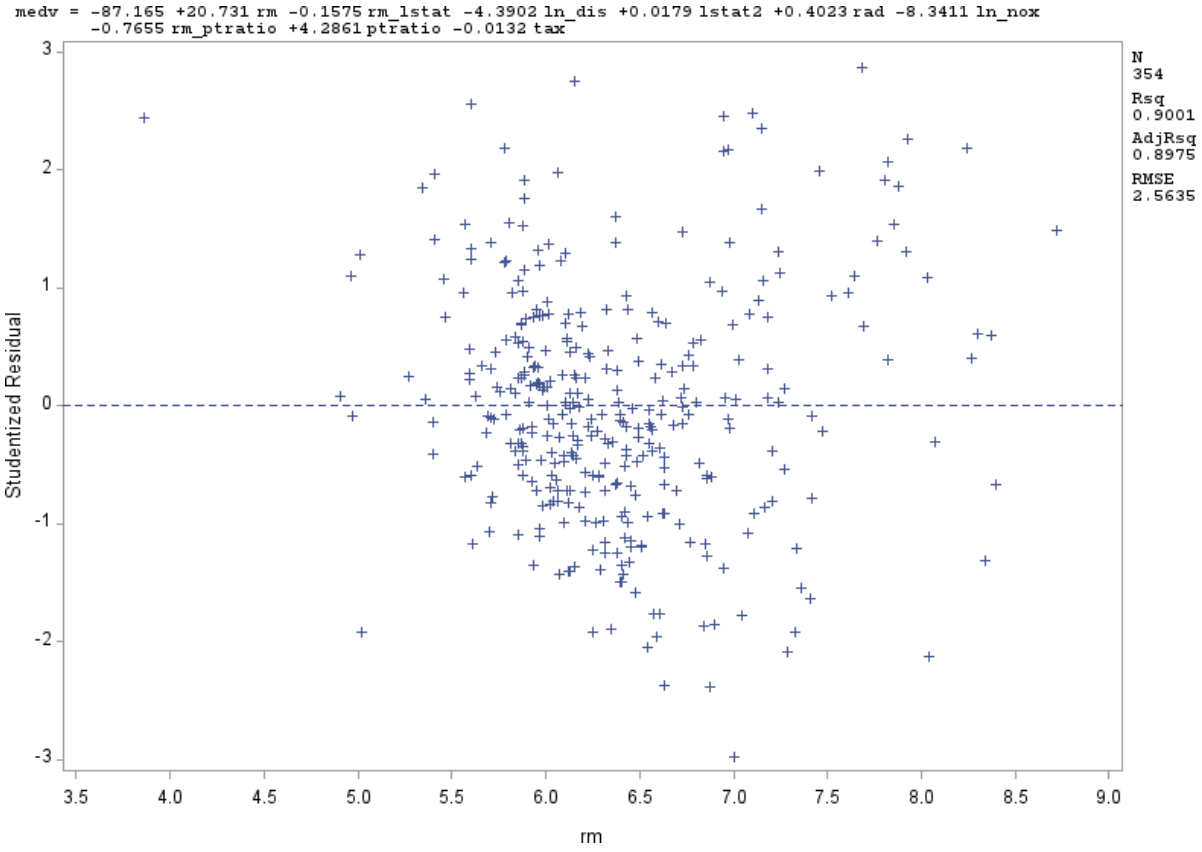
Analysis of Variance

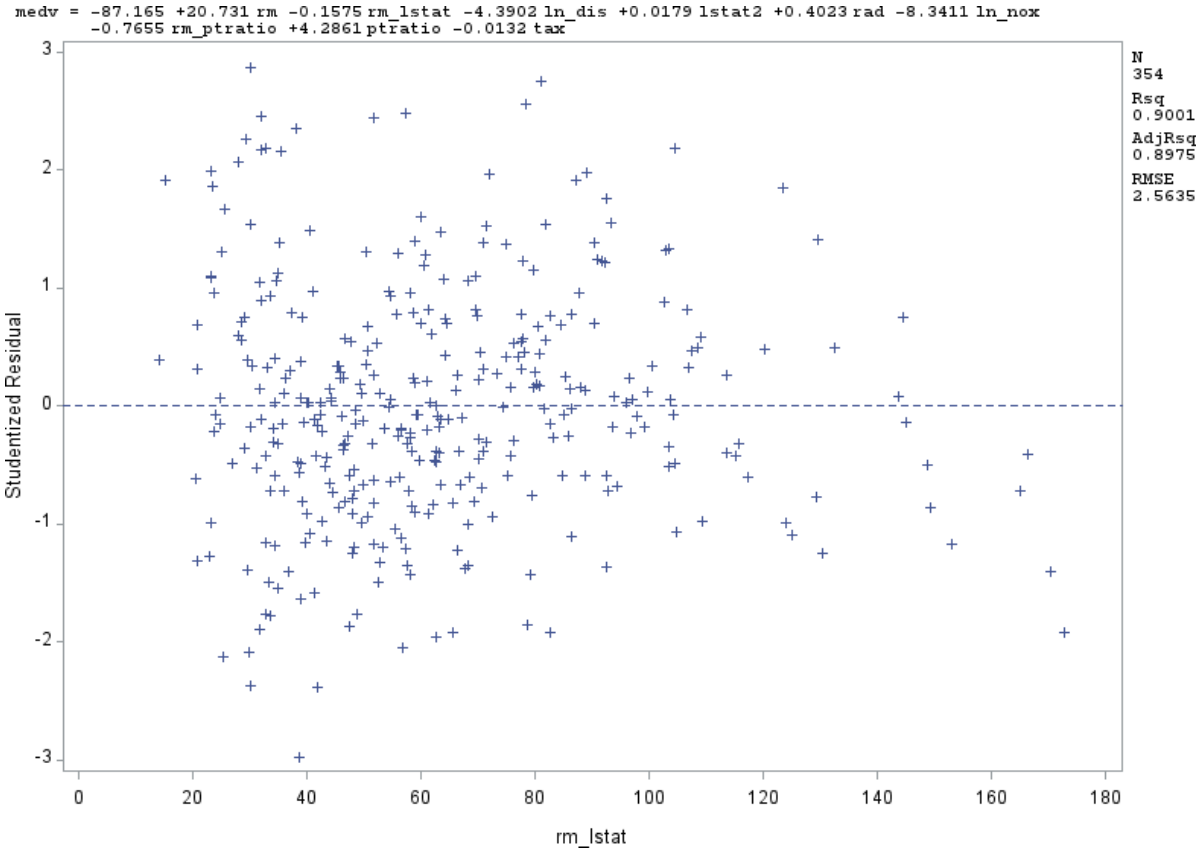
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	20375	2263.84722	344.50	<.0001
Error	344	2260.57038	6.57143		
Corrected Total	353	22635			

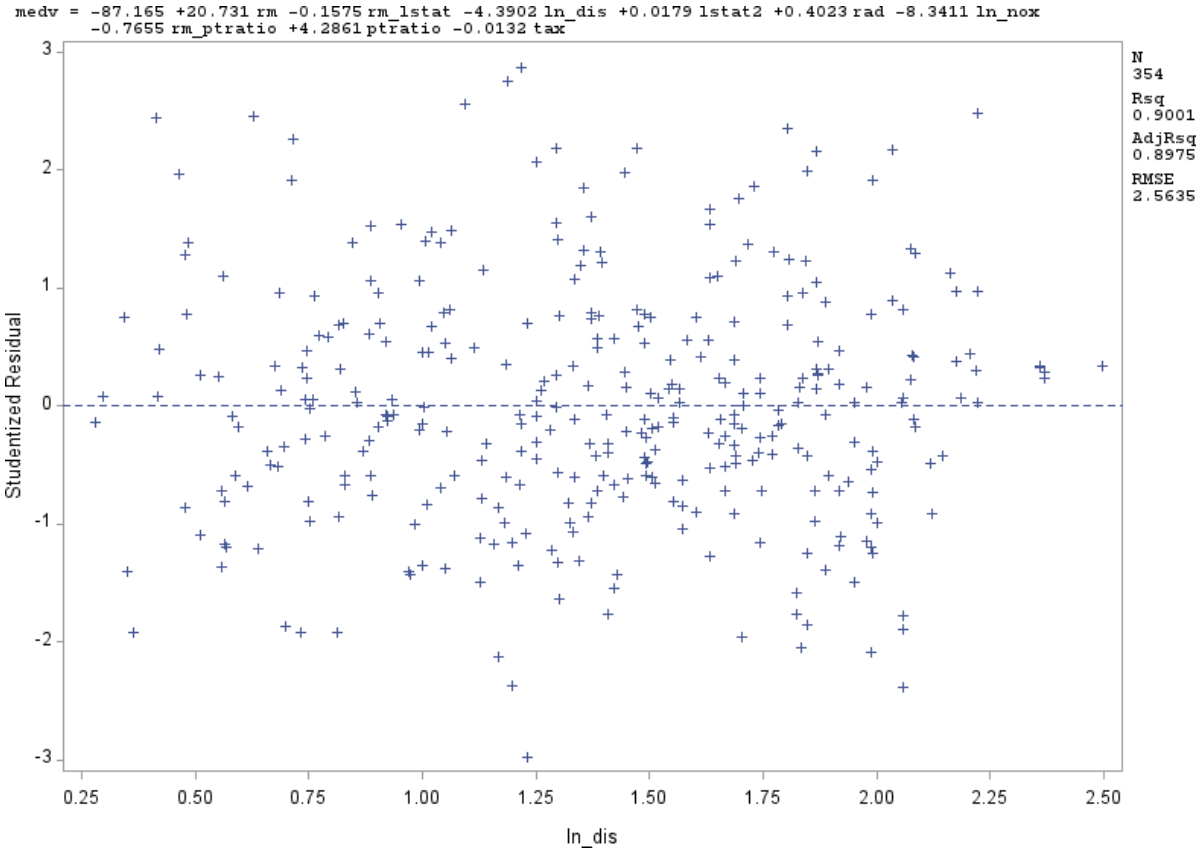
Root MSE	2.56348	R-Square	0.9001
Dependent Mean	24.69011	Adj R-Sq	0.8975
Coeff Var	10.38261		

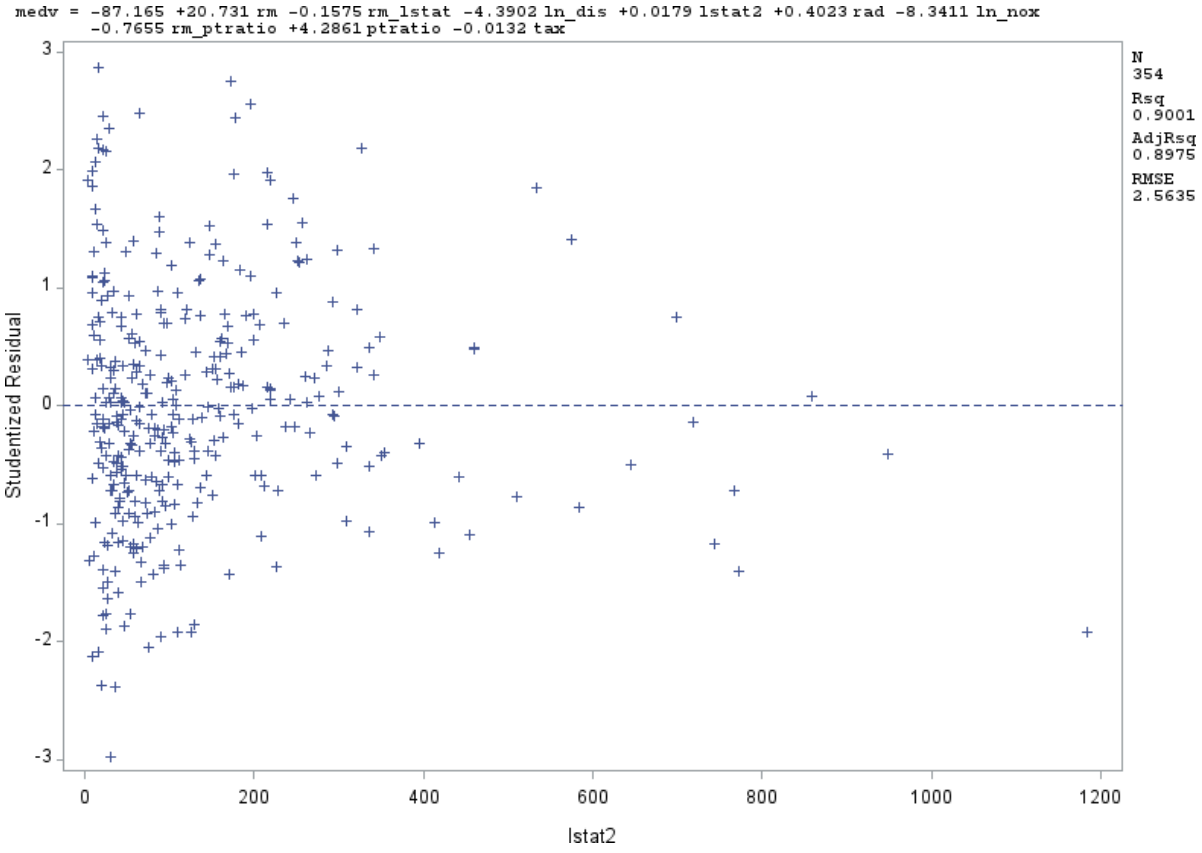
Parameter Estimates

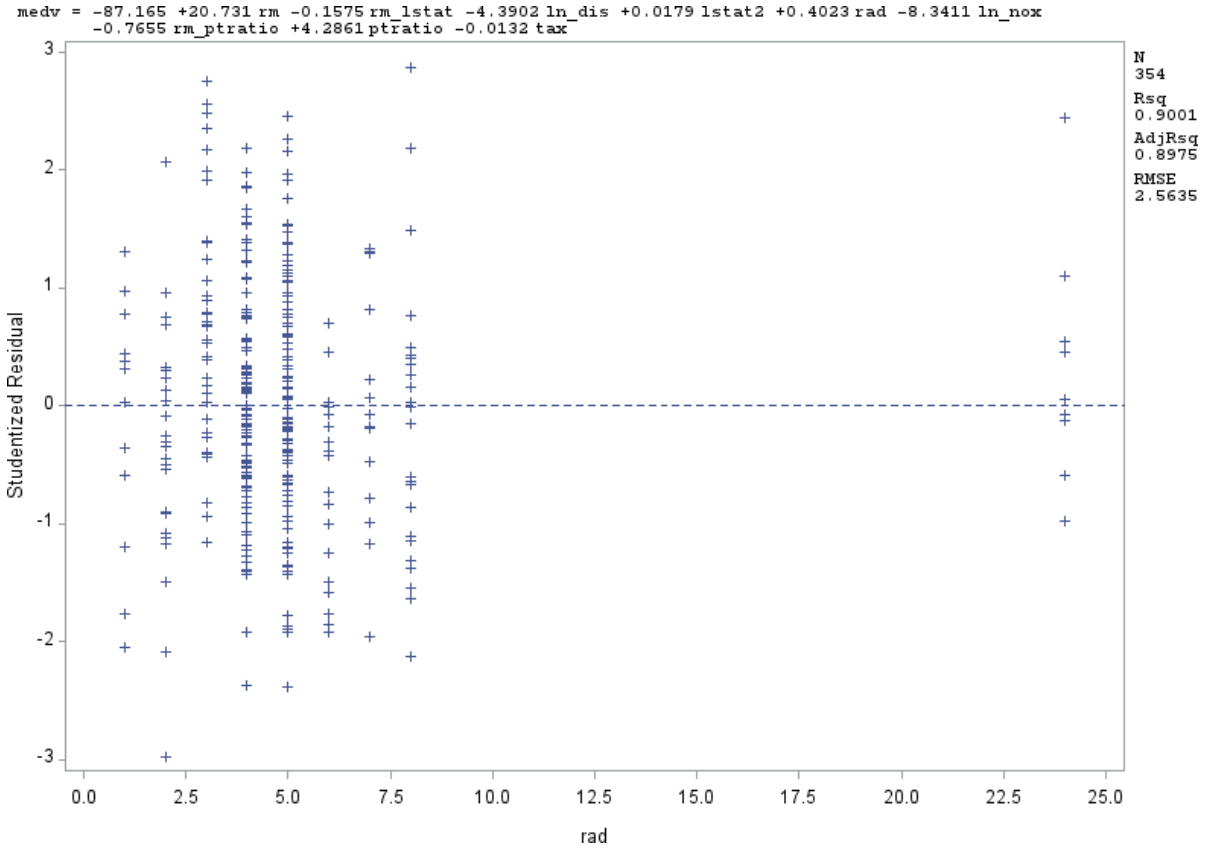
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-87.16519	10.13683	-8.60	<.0001
rm	1	20.73129	1.56496	13.25	<.0001
rm_lstat	1	-0.15748	0.01354	-11.63	<.0001
ln_dis	1	-4.39016	0.53940	-8.14	<.0001
lstat2	1	0.01793	0.00245	7.31	<.0001
rad	1	0.40229	0.05468	7.36	<.0001
ln_nox	1	-8.34114	1.62395	-5.14	<.0001
rm_ptratio	1	-0.76550	0.09210	-8.31	<.0001
ptratio	1	4.28607	0.59152	7.25	<.0001
tax	1	-0.01325	0.00222	-5.97	<.0001

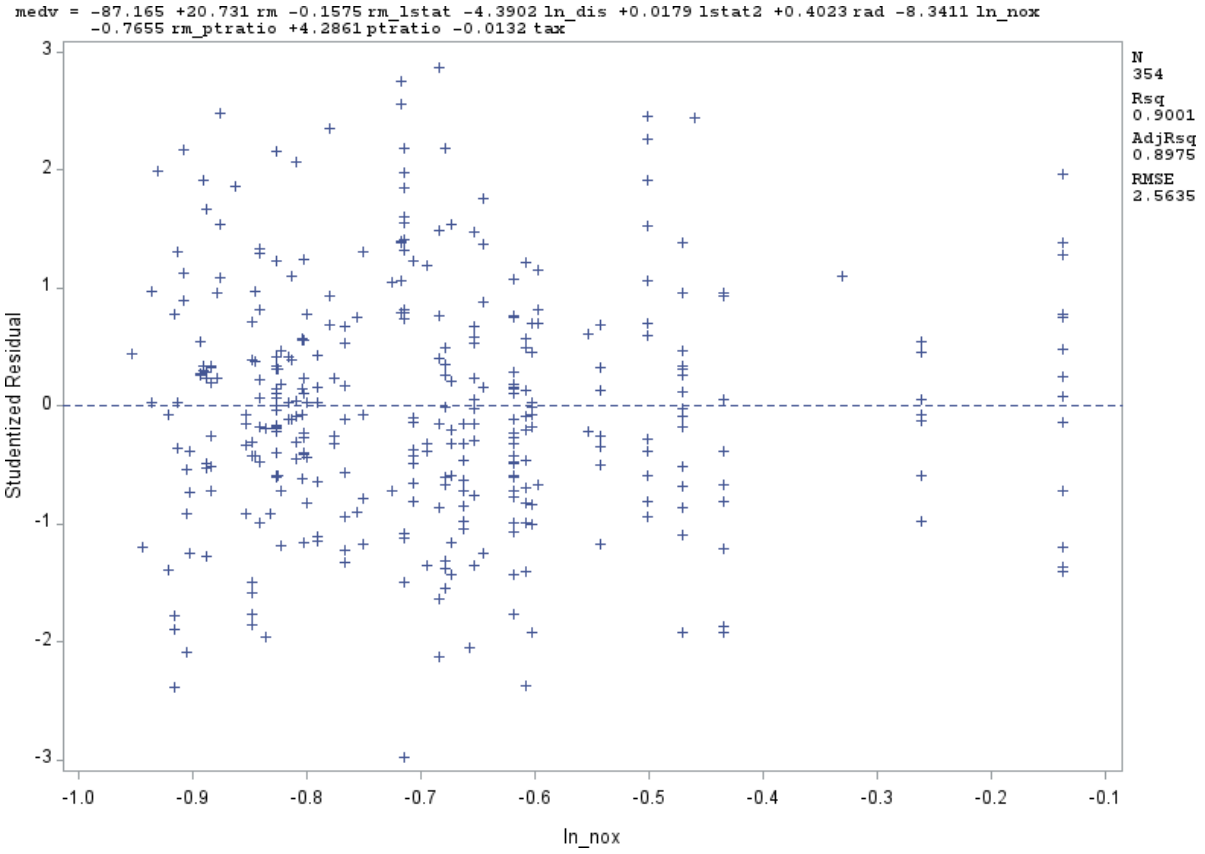

medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
rm

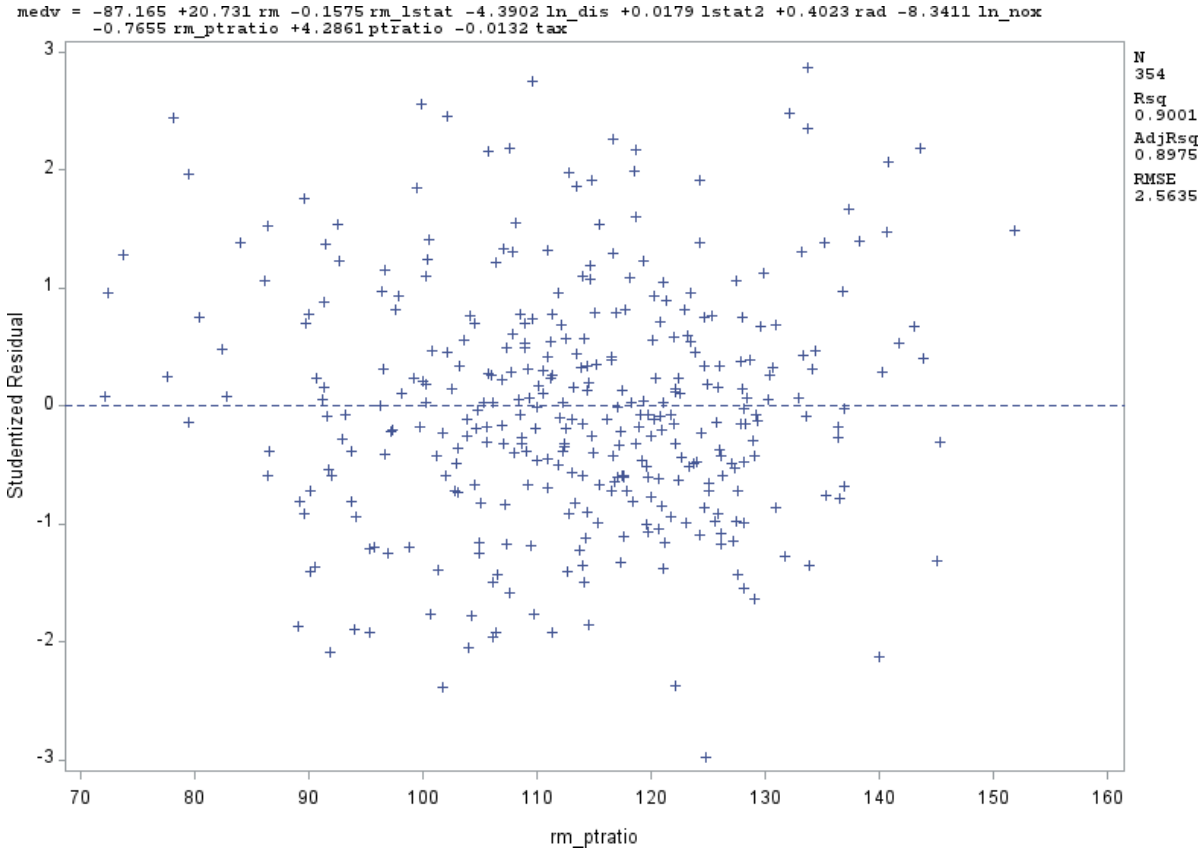

medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
rm_lstat

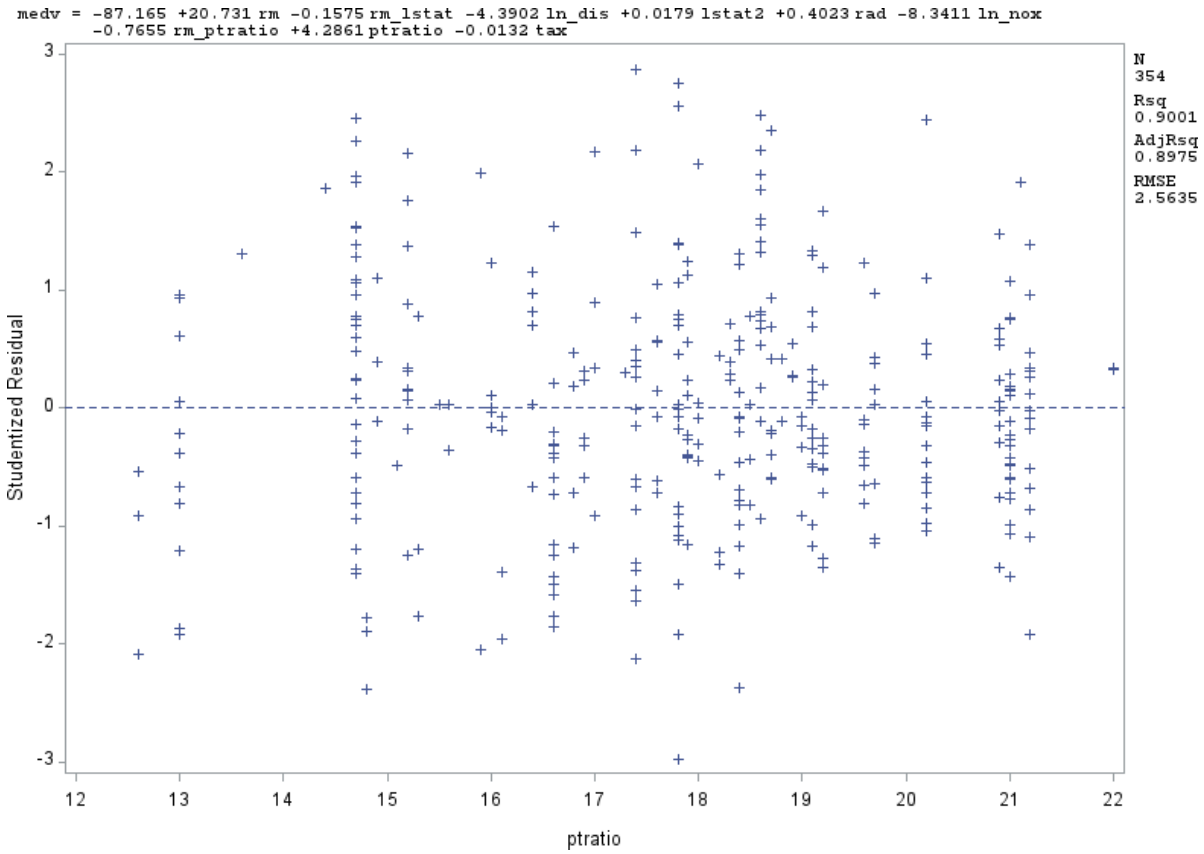

medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
ln_dis

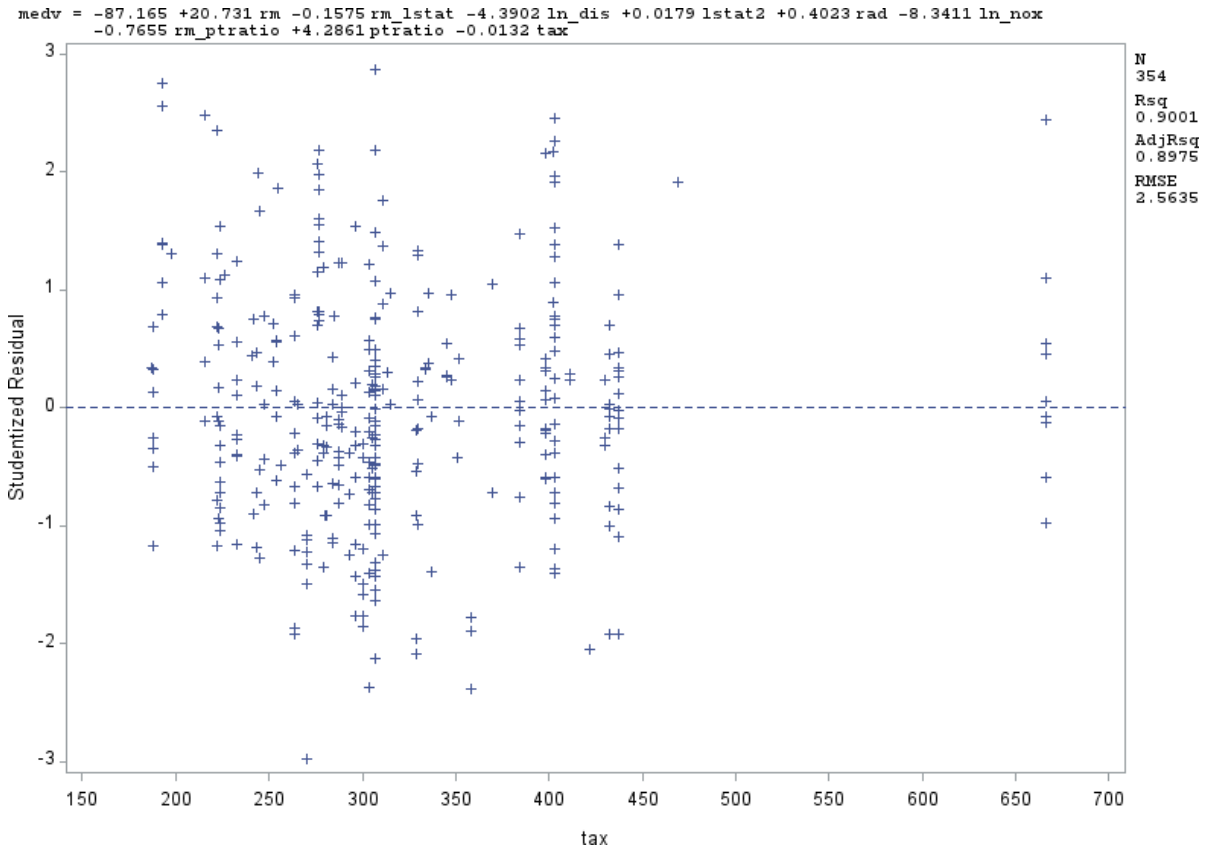

medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
lstat2

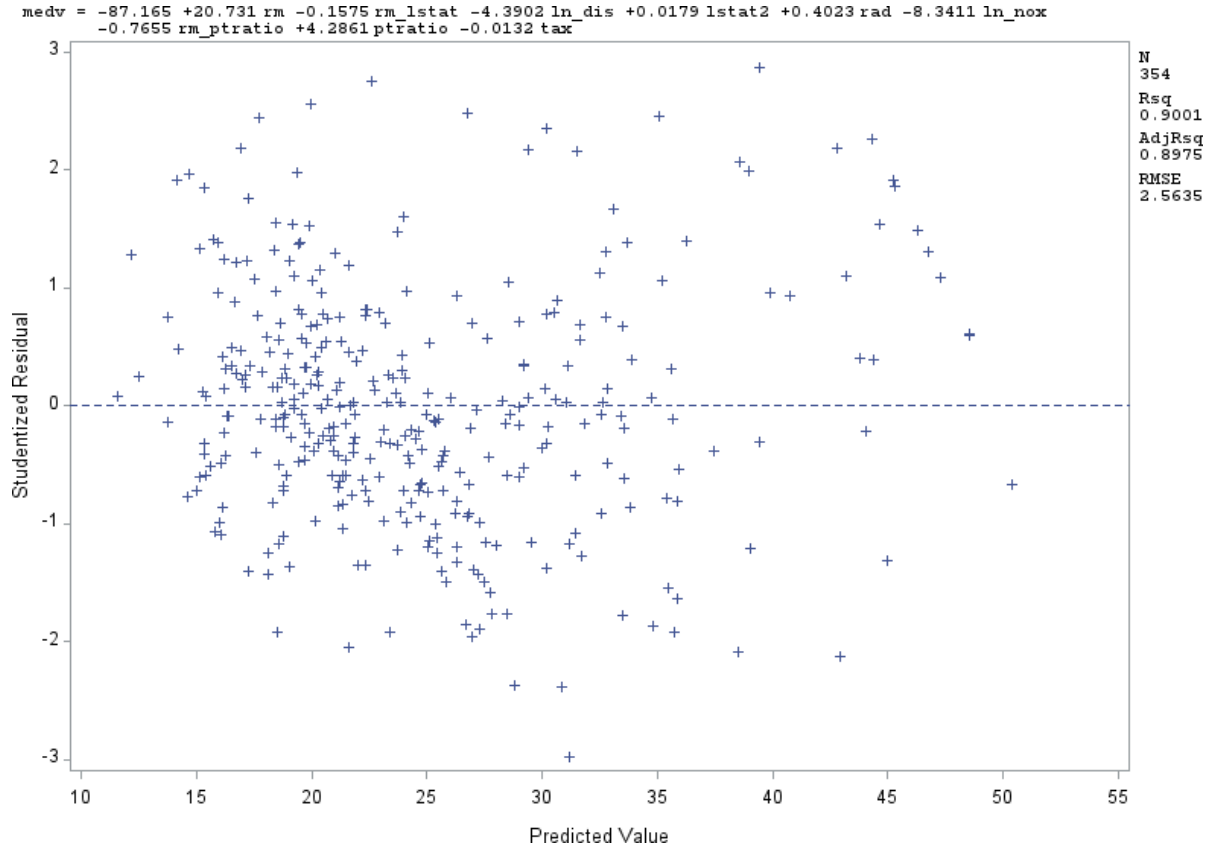

medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
rad

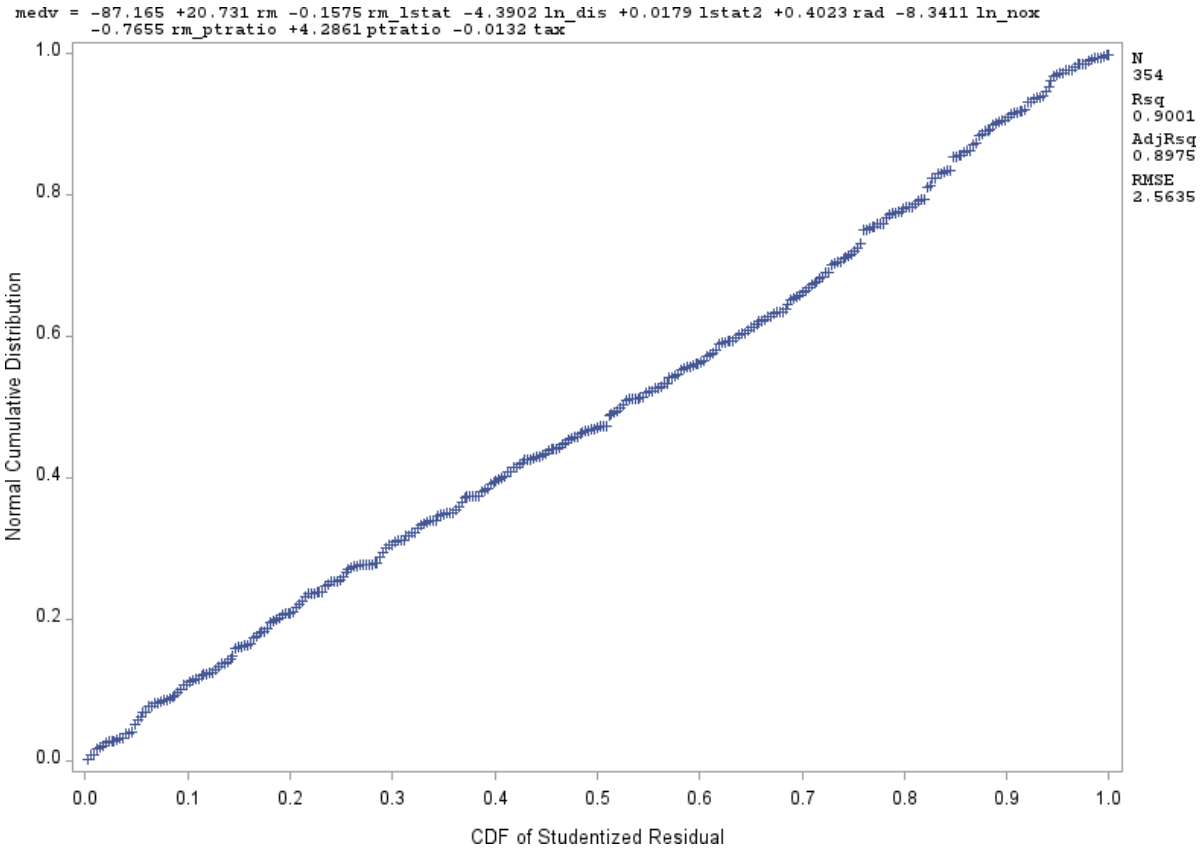

medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
ln_nox


medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
rm_ptratio


medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
ptratio


medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
tax


medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Studentized Residual
Predicted Value


medv = -87.165 +20.731 rm -0.1575 rm_lstat -4.3902 ln_dis +0.0179 lstat2 +0.4023 rad -8.3411 ln_nox
-0.7655 rm_ptratio +4.2861 ptratio -0.0132 tax
N
354
Rsq
0.9001
AdjRsq
0.8975
RMSE
2.5635
Normal Cumulative Distribution
CDF of Studentized Residual
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	32.9817	1.5646	29.9043	36.0591	27.0747	38.8887	.

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	-107.9056	10.5734	-128.7023	-87.1090	-129.3048	-86.5065	.

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	35.4037	1.1686	33.1052	37.7022	29.8624	40.9449	.

References

Berkeley Law. https://www.law.berkeley.edu/files/Rubinfeld_06.09.04_pdfs/Review%20of%20Economics%20and%20Statistics/RES_Air%20pollution_1978.pdf.

jcf2d, Written by. "University of Virginia Library Research Data Services + Sciences." *Research Data Services + Sciences*, https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/.

Rob HyndmanRob Hyndman 50.8k2222 gold badges126126 silver badges177177 bronze badges, et al. "How Should I Transform Non-Negative Data Including Zeros?" *Cross Validated*, 1 Oct. 1958, https://stats.stackexchange.com/questions/1444/how-should-i-transform-non-negative-data-including-zeros.

Hui Li on The SAS Data Science Blog. "How to Use Regularization to Prevent Model Overfitting." *The SAS Data Science Blog*, 6 July 2017, https://blogs.sas.com/content/subconsciousmusings/2017/07/06/how-to-use-regularization-to-prevent-model-overfitting/.

Predictive Interaction for Data Transformation. https://idl.cs.washington.edu/files/2015-PredictiveInteraction-CIDR.pdf.

Hayashi, Toshitaka; Fujita. "Cluster-Based Zero-Shot Learning for Multivariate Data, Journal of Ambient Intelligence and Humanized Computing." *DeepDyve*, Springer Berlin Heidelberg, 28 June 2020, https://www.deepdyve.com/lp/springer-journal/cluster-based-zero-shot-learning-for-multivariate-data-eqc3WWToHu?key=springer.