INTRODUCTION

The goal of this report is to investigate the relevance and degree of certain factors that can be utilized to predict the median value of owner-occupied homes in Boston. The dataset was compiled by David Harrison Jr and Daniel L Rubinfeld for their paper 'Hedonic Housing Prices and the Demand for Clean Air' which focused on Boston's housing market in the early 1970s. The goal of their original paper was to investigate issues associated with using housing market data to assess one's willingness to pay for clean air. This report, however, merely intends to develop a regression model that can be used as a template for similar analyses in the future.

The dataset features information for census tracts in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970 along with other relevant sources. Figure 1 in the appendix organizes the variables by definition and general sources as described by the original paper [1]. The dataset, regression model, and prediction results were respectively investigated and synthesized using SAS 9.4. The provided dataset was clean except for one particular observation (#280). The LSTAT and MEDV fields had a typo that was reasonably resolved in Microsoft Excel based on the structure of the rest of the dataset. [1]

I have preconceived notions about the relationship and measured importance of these factors in determining housing value. Without investigating the variables and units deeply, I expect location, size, and age to be the most important factors in the valuation of a home. Ultimately, the variables within the set contribute in various ways to those three prevailing factors. Moreover, I expect the following variables to have a positive relationship: ZN, CHAS, ZN and RAD. I expect the following variables to have a negative relationship: CRIME, INDUS, NOX, AGE, DIS, TAX PTRATIO, MINOR, and LSTAT.

ANALYSIS

The analysis portion of the report is divided into a few sections. Section I described initial exploration of the dataset. Section II discusses the first model featuring all variables and any modifications that reasonably need to be made. Section III details the model selection process. Section IV examines the resulting model, the corresponding knowledge, and its possible shortcomings. Section V spotlights two housing value predictions based on the chosen regression model.

Section I – Exploration

The variables in this dataset are mostly continuous (numeric) in nature except for CHAS which differentiates census tracts that are bounded by the river from those that do not. The analysis began with a survey of initial descriptives to identify if there are any outstanding values for any of the variables (Figure 2). The first thing I recognized was that CRIME had a maximum value of 18.492 which was significantly different from its mean (0.7182), median (0.1405) and upper quartile values (0.5370). An inspection of the crime rate distribution confirms an obvious right skew. (Figure 11) While I was relieved to see that there were other observations with crime rights greater than 1 and a few in the double

[1] I submitted an updated CSV file of the dataset to D2L.

digits, roughly 90% of census tracts in this dataset have a crime rate of less than 3%. Similar to CRIME, the descriptive table also made me aware of an unusual maximum value for ZN. The corresponding histogram confirmed my suspicions of a right skew. (Figure 12) Approximately, 64.2% (241 out of 375) census tracts in this dataset had a ZN value of 0. Interestingly enough, the distribution appears to be binomial if you ignore 64% of the data (which is not practical).

The distribution of RAD also requires discussion. As expressed in Figure 13, about 95% of the observations in the dataset feature RAD values between 1 and 8 while the remaining 5% of the dataset has RAD values of 24. I initially considered discretizing this variable into 8 dummy variables but I opted not to because the radial highway index is apparently continuous. The highway access index was calculated on a town basis. [1] It is intriguing that the dataset features a subset which such lopsided values relative to the rest of set. However, local knowledge of Boston metropolitan geography may reveal a region that is very dense with highway access relative to the rest of the SMSA.

Given that CHAS is a binary dummy variable, a boxplot was immediately produced to distribute home values by proximity to the Charles River (Figure 3). While the ranges were equal regardless of CHAS status, census tracts that bounded the river had median home values on average compared to otherwise. It can be argued that proximity to greenspace and a riverfront view likely inflates prices of the homes along the river.

Normality, linearity, constant variance and independence are assumptions that must be validated during the construction of a linear regression. While constant variance and independence are harder to confirm before the analysis of residual values, the outcome of normality and linearity validation can be hinted during the exploration stage. The distribution of the independent variable is the primary way to assess normality prior to producing a regression output. According to its corresponding histogram, median home values are not normally distributed. (Figure 10) The figure is right-skewed although not as right-skewed as some of the other variables discussed above. A normal probability plot will confirm this but I anticipate that a transformation will be required to make the final model more viable. In terms of linearity, the scatterplots for DIS (Figure 14) and LSTAT (Figure 15) appeared to be nonlinear. A residual analysis would confirm that relationship but I believe a higher order or logarithmic transformation may be appropriate for these two variables.

For the purposes of this report, interaction terms were not incorporated into this analysis for multiple reasons. Primarily, it was believed that the variables were isolated enough that I could not theorize the effect of one variable could be different depending on the prevalence of another variable. That being said, a survey of the Pearson correlation matrix provided early survey into possible multicollinearity (Figure 4). While none of the independent variables feature correlation coefficients that are ±0.9, two mild correlations warrant explanations. RAD and CRIM have a correlation coefficient of 0.75936 and I believe that's due to the fact that city centers (which usually have greater highway access) have higher rates of crime. [6]. City centers (that are often urban) have NOX and INDUS have a correlation coefficient of 0.71166 and I believe that is because regions with a lot of non-retail business produces more air pollutants. Non-store retailing refers to retailing that takes place outside traditional brick-and-mortar (physical) locations. [5] While it's mostly defined by online retailing in today's

economy, non-retail business in the late 70s was dominated by factories so the pathway to pollution is highly conceivable.

Section II – First Model

Figure 5 is the regression output for the full model before anything has been done to either the dataset or the variables. The model has an F-value of 98.53 and a p-value that is less than .001 meaning we can reject the null hypothesis and conclude that at least one of the dependent variables featured in this study has a statistically significant effect on median home values. Furthermore, the adjusted R^2 value of 0.7722 indicates that 77.22% of the variation in median home values can be explained by this preliminary model. Most of the variables are statistically significant as well except for INDUS, CHAS, and AGE. However, there are some issues with this model. Firstly, the assumption of constant variance and independence are violated as indicated by the studentized residual plots associated with MEDV (Figure 6). Secondly, the early indication of a normality violation is confirmed by the normal probability plot. (Figure 9) While the shape of the plot is not as egregious as the histogram, as a result, I believe that a logarithmic transformation is necessary to stabilize the variance a bit. In addition, my earlier suspicions about the distributions of DIS and LSTAT were confirmed after studying their residual plots. Similar to MEDV, both plots violate constant variance and independence because they each have a funnel-shaped distribution of their points. (Figure 7 and Figure 8) A transformation of these dependent variables will likely improve the model as well. Figure 16 illustrates the improvement of the model after transformations of the aforementioned variables.

Analysis of diagnostics reveal further issues with the model. Figure 5 does not signal multicollinearity because all the variance inflation factors (VIF) for each parameter are not close to 10 (and that remained true over the course of the transformations). However, there are several outliers and influential points present in the dataset after the transformations were applied. The studentized residual threshold for an outlier in this report was $\geq \pm 3$ and the Cook's D threshold for an influential point was ≥ 0.01067 (also known as $4/n$). I only removed points that were simultaneously identify as outlier and influential. Figure 17 presents the studentized residual and Cook's D values for those points. Figure 18 is the regression output of the model after transformation and removal of the aforementioned outlier/influential points. The residual plots for the transformed variables and the normal probability plot all stabilized post-transformation and post-outlier removal. Figure 19 refers to LNMEDV (natural log of MEDV), Figure 20 refers to LNDIS (natural log of DIS), Figure 21 refers to LNLSTAT (natural log of LSTAT) and Figure 22 refers to the normal probability plot.

Section III – Final Model Selection and Validation

The variables have been adequately transformed and outlier points have been appropriately removed but all the variables that were presented in the dataset still exist within the model. At this stage, CRIME, ZN, INDUS, CHAS, AGE, and MINOR appear to be statistically insignificant (although MINOR is arguably at a p-value of 0.0507). At this stage, I employed a hybrid approach to finding an appropriate model. I first applied the GLMSELECT procedure to the dataset (after trimming out outliers and influential points) to produce two different datasets and then compared their training and testing

performances to choose the best model. Next, I employed the SURVEYSELECT procedure to the trimmed dataset via a holdout method to produce a superior model that can be compared to the model chosen from the GLMSELECT procedure. Finally, I compared the test and training performances of the final two models on the same holdout sets to determine an overall model for use. Each of the following paragraphs will go into more detail about each step and then this section will finish with an explanation of why the final model was chosen.

GLMSELECT is a procedure that gives SAS the ability to perform a k-fold cross-validation on a dataset with the goal of producing an appropriate regression model based on a selection method of choice. Using this procedure, I produced two 10-fold cross-validation models and held out 25% of the dataset for testing within each fold. The first model was produced using the stepwise selection method and the second model was produced using the backward selection method. Figure 23 presents the stepwise selection output with variable selection summary, CVPRESS values, corresponding error charts and values, and cross-validation estimates for each fold. Figure 24 presents the exact same figure except for the backward selection method. Figure 25 is a table that summarizes the key that metrics that help determine which of these two models is and would perform better on unseen data. Firstly, the main difference between the stepwise and backward models is that the stepwise output has an additional variable (MINOR). Both models feature the following variables: CRIME, NOX, RM, AGE, LNDIS, RAD, TAX, PTRATIO, and LNLSTAT. Their performances are also very similar in performance as well (as Figure 25 shows). While the backward model is slightly superior in terms of F-Value and RMSE, the stepwise model has marginal advantages in terms of a higher adjusted-R^2, lower ASE values for both training and tests sets, and a lower CVPRESS. In addition, the stepwise model also has a smaller difference between its two ASE values as well (0.00028 vs 0.00204). While the stepwise model appears to be slightly superior, the backward model will be kept in consideration as SAS's other model selection procedure is employed.

SURVEYSELECT is a SAS procedure that can split a dataset into training set and tests and provide users direct access to the datasets themselves. Firstly, I split the trimmed dataset and then applied 4 selection methods to the training set in order to yield two good models for consideration. I applied stepwise, backward, Mallows' CP, and Adjusted R^2 methods at this stage. Conveniently, all four selection methods produced the same model with equivalent variables. The R^2 value of the model was 0.8961 and the Cp value was 8.8429. Furthermore, the model actually matched the stepwise output from the GLMSELECT procedure. It appears that the transformations and dataset trimming in combination with an already high 0.8438 R^2 left little room for significant improvement via removal of certain variables.

Finally, validation methods were applied to the recently synthesized test set to ultimately compare the performance metrics of the two GLMSELECT models on the same SURVEYSELECT test set produced in the previous step (since the stepwise GLMSELECT model is equivalent to the best SURVEYSELECT output). From this point on in the report, the two models being compared will GLMSELECT outputs; one produced with the stepwise selection method and the other produced with the backward selection method. Figure 26 features the regression output of the stepwise model on the training set and Figure 27 features the test metrics of the stepwise model on the test set. Figure 28 features the regression output of the backward model on the test set and Figure 29 features the test

metrics of the backward model on the test set. Figure 30 is a table that conveniently summarizes all the relevant metrics of each model's training and test performances.

In terms of the training set, the stepwise model marginally outperforms the backward model. The former has smaller values for RMSE (0.10375 vs 0.10503) and higher for adjusted-R^2 (0.8922 vs 0.8895) while the latter has a greater F-value (246.98 vs 227.75). In terms of the test set, the stepwise marginally outperforms the backward model in all 5 test metrics; RMSE (0.0959 vs 0.0975), MAE (0.0745 vs 0.0765), adjusted R^2 (0.88974 vs 0.88563) and cross-validated R^2 (0.00636 vs 0.00747). Overall, the model metrics have a slight preference for the GLMSELECT Stepwise model that has 10 variables over the GLMSELECT backward model that has just 9 variables. Given that the defining variable in this comparison is MINOR, I'm even more inclined to choose the model that includes it. Beyond the data, there is evidence that demographics play a role in housing prices nationwide.[7] Specifically, Boston is also renown for being racially-segregated and it was worse at the time that this dataset was compiled. [8] The slightly more complicated model I believe will be more applicable in the real world.

Section IV – Findings and Results

Figure 31 is the regression output for the final model. The model has an F-value of 301.56 and a p-value that is less than .001 meaning we can reject the null hypothesis and conclude that at least one of the dependent variables featured in this study has a statistically significant effect on median home values. Furthermore, the adjusted R^2 value of 0.8917 indicates that 89.17% of the variation in median home values can be explained by this preliminary model. All of the variables are statistically significant as well as indicated by their small p-values. The model assumptions are adequately satisfied as well as the earlier transformations heavily stabilized the patterns that originally occurred with MEDV as well as DIS and LSTAT. Figure 32 exhibits a highly-stabilized normal probability plot as well. Once again, multicollinearity remains a non-factor. Of course, two outliers came about as a result of analysis but that pales in comparison to the 9 found at the beginning of this report.

The equation for this regression models is as follows:

Lnmedv = 3.3875 + 0.03193*CRIME – 0.8359*NOX + 0.21365*RM – 0.00108*AGE – 0.021603*LNDIS + 0.01578*RAD – 0.00057811*TAX – 0.02632*PTRATIO + 0.00048985*MINOR – 0.1909 LNLSTAT

The following interpretations of the equation assume that everything else remains constant in terms of describing any particular variable's effect. If the crime rate increased by one percent, the median home value of a census tract would increase by 3.24%. An increase in nitrogen oxide concentration by 1 part per hundred million would decrease the median home value of a census tract by 56.65%. If the average number of rooms per dwelling increased by 1, there would be a 23.82% increase in median home value within an average census tract. A one percent increase in proportion of homes built prior to 1940 would decrease median home value by 0.108%. A single unit increase in the natural log of the weighted distance from five employment centers would decrease the median home value of a census tract by -2.13%. A one unit increased in the radial highway access index would increase the median home value of a census tract by 1.59%. A 1% increase in the tax rate per $10,000 would decrease the median home value of a census tract by just 0.058%. A one unit increase in the pupil-

teacher ratio would decrease the median home value of a census tract by 2.59%. A one unit increase in MINOR would increase the median home value of a census tract by 0,049%. Finally, a unit increase in the natural log of the proportion of the population that is lower status would decrease the median home value of a census tract by 17.38%.

NOX, LNSTAT, LNDIS, and RM are the most influential predictors in the model because they have the highest absolute value beta coefficient values as well as the greatest standardized estimate values in the regression model.

Section V – Predictions

Earlier sections in this report have validated the chosen model as adequate for unseen data. In this section, I explain the production of two median home value predictions based on certain variable values. The values for these observations were produced randomly with Python taking the minimum and maximum values of each variable into consideration. (Figure 33) For the sake of efficiency, Figure 34 is a table that displays all the values for each relevant variable in detail while Figure 35 is the prediction output.

Given the values for each variable, the first observation has a predicted median home value of $17,167 with a confidence interval of $15,572 to $18,927 and a prediction interval of $13,745 to $21,443. The second observation has a predicted median home value of $39,769 with a confidence interval of $34,230 to $46,210 and a prediction interval of $30,976 to $51,060. Given that the values for each variable were well within the range of the dataset, I have high confidence in these estimated values.

Future Work

There were some internal and external pathways in terms of future research with this such. From an internal perspective, this perspective led to some interesting outputs. Primarily, I was surprised that CRIME would end up with a possible parameter. Logic would tell you that property is more valuable in census tracts that are safer. However, there's an argument to be made that the positive effect of convenient location (downtown areas with accessible amenities) outweighs the negative effect of greater likelihood of witnessing or being the victim of a crime. Secondly, within the dataset, there was a class imbalance in terms of observations where CHAS = 1. 91% of the dataset was not bounded by the river. I wonder if that proportion is representative of Boston in general since the dataset used here is a subset of the dataset used in the original study by Harrison and Rubinfeld. In addition, there's a pragmatic argument to be made for a parabolic relationship between median home values and MINOR. On the one hand, an increase in minority proportion usually leads to a decrease in home value in areas where minorities are deemed undesirable white people. On the other hand, I've witnessed neighborhoods and suburban areas in Chicagoland that charge a premium for housing that have very high minority rates due to market discrimination. The external research about Boston's history of segregation definitely informed my beliefs about what other factors can contribute to median home values. Finally, I believe that income or some measure of wealth per capita should have been included in the dataset because socioeconomic status absolute plays a role in property value in real life.

Appendix

Figure 1 – Variables used in the Dataset		
Variables	Definition	Source
MEDV	Median value of owner-occupied homes	1970 U.S. Census
CRIME	Per capita crime rate by town	Federal Bureau of Investigation (1970)
ZN	Proportion of a town's residential land zoned for lots greater than 25,000 square feet	Metropolitan Area Planning Commission (1972)
INDUS	Proportion nonretail business acres per town	Vot, Ivers, and Associates [2]
CHAS	Charles River dummy variable (=1 if tract bounds the Charles River, =0 if otherwise)	1970 U.S. Census Tract maps
NOX	Nitrogen oxide concentrations in pphm (annual average concentration in parts per 100 million)	TASSIM [3]
RM	Average number of rooms per dwelling	1970 U.S. Census
AGE	Proportion of owner-occupied units built prior to 1940 built prior to 1940	1970 U.S. Census
DIS	Weighted distances to five employment centers in the Boston region	Schnare [4]
RAD	Index of accessibility to radial highways	MIT Boston Project
TAX	Full value property tax rate per $10,000	Massachusetts Taxpayers Foundation (1970)
PTRATIO	Pupil-teacher ratio by town school district	Massachusetts Department of Education (1971-1972)
MINOR	Calculated as $1000*(MINK - 0.63)^2$ where MINK is the proportion of minorities by town	1970 U.S. Census
LSTAT	Proportion of population that is lower status	

Figure 2 – Initial Descriptives

Variable	N	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range
medv	375	25.1944	8.7709	11.8000	19.5000	22.8000	28.7000	50.0000	38.2000
crime	375	0.7182	1.7628	0.0063	0.0613	0.1405	0.5370	18.4982	18.4919
zn	375	15.3333	25.9505	0.0000	0.0000	0.0000	22.0000	100.0000	100.0000
indus	375	8.8374	6.3135	0.4600	4.0500	6.9100	10.5900	25.6500	25.1900
chas	375	0.0933	0.2913	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
nox	375	0.5195	0.1096	0.3850	0.4379	0.4930	0.5470	0.8710	0.4860
rm	375	6.3685	0.7220	3.5610	5.9270	6.2450	6.7260	8.7800	5.2190
age	375	61.9800	28.7009	2.9000	36.6000	65.1000	89.8000	100.0000	97.1000
dis	375	4.3744	2.1366	1.1296	2.5961	4.0123	5.8700	12.1265	10.9969
rad	375	5.4587	4.5654	1.0000	4.0000	5.0000	5.0000	24.0000	23.0000
tax	375	328.6987	102.2730	187.0000	270.0000	307.0000	398.0000	666.0000	479.0000
ptratio	375	17.8581	2.2174	12.6000	16.4000	18.0000	19.6000	22.0000	9.4000
minor	375	379.4129	41.5068	70.8000	380.3400	392.2000	396.0600	396.9000	326.1000
lstat	375	10.4725	6.0966	1.7300	6.0500	9.3800	13.2800	37.9700	36.2400

Figure 3 – Median Value by Proximity to Charles River

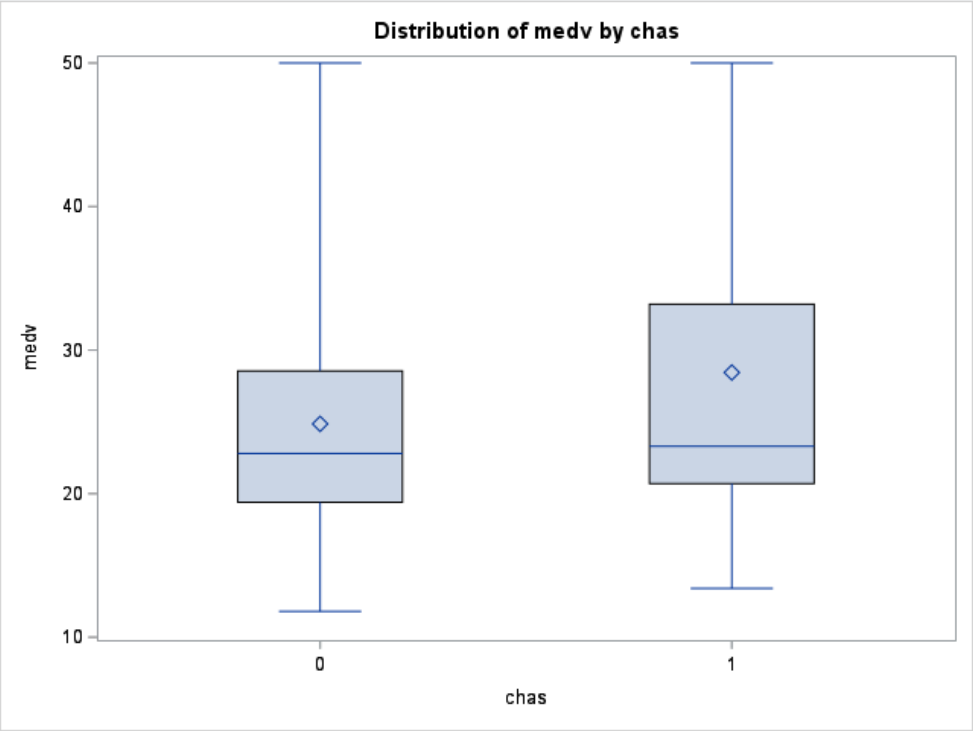


Figure 4 – Pearson Correlation Matrix

Pearson Correlation Coefficients, N = 375
Prob > |r| under H0: Rho=0

	medv	crime	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	minor	lstat
medv	1.00000	-0.01481 0.7749	0.27888 <.0001	-0.28576 <.0001	0.11888 0.0213	-0.19826 0.0001	0.77073 <.0001	-0.19727 0.0001	-0.00473 0.9273	0.10109 0.0504	-0.11132 0.0311	-0.38236 <.0001	0.17426 0.0007	-0.65111 <.0001
crime	-0.01481 0.7749	1.00000	-0.21399 <.0001	0.45386 <.0001	0.20657 <.0001	0.53360 <.0001	-0.31044 <.0001	0.37829 <.0001	-0.39877 <.0001	0.75936 <.0001	0.69373 <.0001	0.11588 0.0248	-0.30261 <.0001	0.32048 <.0001
zn	0.27888 <.0001	-0.21399 <.0001	1.00000	-0.48816 <.0001	-0.09433 0.0681	-0.47234 <.0001	0.29961 <.0001	-0.54172 <.0001	0.63397 <.0001	-0.19114 0.0002	-0.16831 0.0011	-0.30445 <.0001	0.14393 0.0052	-0.37513 <.0001
indus	-0.28576 <.0001	0.45386 <.0001	-0.48816 <.0001	1.00000	0.19753 0.0001	0.71166 <.0001	-0.38523 <.0001	0.56979 <.0001	-0.63166 <.0001	0.32548 <.0001	0.51616 <.0001	0.16211 0.0016	-0.31283 <.0001	0.48576 <.0001
chas	0.11888 0.0213	0.20657 <.0001	-0.09433 0.0681	0.19753 0.0001	1.00000	0.21678 <.0001	0.06723 0.1939	0.17373 0.0007	-0.20220 <.0001	0.27133 <.0001	0.18081 0.0004	-0.05313 0.3048	-0.04966 0.3376	0.04054 0.4338
nox	-0.19826 0.0001	0.53360 <.0001	-0.47234 <.0001	0.71166 <.0001	0.21678 <.0001	1.00000	-0.27901 <.0001	0.68604 <.0001	-0.72091 <.0001	0.41944 <.0001	0.53015 <.0001	-0.07488 0.1479	-0.39915 <.0001	0.45703 <.0001
rm	0.77073 <.0001	-0.31044 <.0001	0.29961 <.0001	-0.38523 <.0001	0.06723 0.1939	-0.27901 <.0001	1.00000	-0.20323 <.0001	0.12517 0.0153	-0.15020 0.0036	-0.26861 <.0001	-0.34207 <.0001	0.20477 <.0001	-0.64347 <.0001
age	-0.19727 0.0001	0.37829 <.0001	-0.54172 <.0001	0.56979 <.0001	0.17373 0.0007	0.68604 <.0001	-0.20323 <.0001	1.00000	-0.70018 <.0001	0.27608 <.0001	0.34772 <.0001	0.09062 0.0797	-0.23857 <.0001	0.53111 <.0001
dis	-0.00473 0.9273	-0.39877 <.0001	0.63397 <.0001	-0.63166 <.0001	-0.20220 <.0001	-0.72091 <.0001	0.12517 0.0153	-0.70018 <.0001	1.00000	-0.29710 <.0001	-0.33734 <.0001	-0.01774 0.7321	0.23309 <.0001	-0.34094 <.0001
rad	0.10109 0.0504	0.75936 <.0001	-0.19114 0.0002	0.32548 <.0001	0.27133 <.0001	0.41944 <.0001	-0.15020 0.0036	0.27608 <.0001	-0.29710 <.0001	1.00000	0.76603 <.0001	0.22047 <.0001	-0.12609 0.0145	0.07382 0.1536
tax	-0.11132 0.0311	0.69373 <.0001	-0.16831 0.0011	0.51616 <.0001	0.18081 0.0004	0.53015 <.0001	-0.26861 <.0001	0.34772 <.0001	-0.33734 <.0001	0.76603 <.0001	1.00000	0.18418 0.0003	-0.24978 <.0001	0.20910 <.0001
ptratio	-0.38236 <.0001	0.11588 0.0248	-0.30445 <.0001	0.16211 0.0016	-0.05313 0.3048	-0.07488 0.1479	-0.34207 <.0001	0.09062 0.0797	-0.01774 0.7321	0.22047 <.0001	0.18418 0.0003	1.00000	0.07069 0.1719	0.20737 <.0001
minor	0.17426 0.0007	-0.30261 <.0001	0.14393 0.0052	-0.31283 <.0001	-0.04966 0.3376	-0.39915 <.0001	0.20477 <.0001	-0.23857 <.0001	0.23309 <.0001	-0.12609 0.0145	-0.24978 <.0001	0.07069 0.1719	1.00000	-0.20106 <.0001
lstat	-0.65111 <.0001	0.32048 <.0001	-0.37513 <.0001	0.48576 <.0001	0.04054 0.4338	0.45703 <.0001	-0.64347 <.0001	0.53111 <.0001	-0.34094 <.0001	0.07382 0.1536	0.20910 <.0001	0.20737 <.0001	-0.20106 <.0001	1.00000

Figure 5 – Full Model Regression Output

The REG Procedure
Model: MODEL1
Dependent Variable: medv

Number of Observations Read	375
Number of Observations Used	375

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	22446	1726.60036	98.53	<.0001
Error	361	6325.83351	17.52308		
Corrected Total	374	28772			

Root MSE	4.18606	R-Square	0.7801
Dependent Mean	25.19440	Adj R-Sq	0.7722
Coeff Var	16.61503		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	12.63324	5.55329	2.27	0.0235	0
crime	1	1.28883	0.22380	5.76	<.0001	3.32187
zn	1	0.03579	0.01241	2.88	0.0042	2.21198
indus	1	0.07362	0.05727	1.29	0.1994	2.79005
chas	1	0.12124	0.79815	0.15	0.8794	1.15365
nox	1	-17.35683	3.98974	-4.35	<.0001	4.08185
rm	1	6.41932	0.44839	14.32	<.0001	2.23708
age	1	-0.00650	0.01269	-0.51	0.6088	2.83139
dis	1	-1.18892	0.18716	-6.35	<.0001	3.41311
rad	1	0.39406	0.09801	4.02	<.0001	4.27283
tax	1	-0.01565	0.00389	-4.02	<.0001	3.37728
ptratio	1	-0.71652	0.12293	-5.83	<.0001	1.58574
minor	1	0.01167	0.00591	1.97	0.0491	1.28351
lstat	1	-0.42714	0.05824	-7.33	<.0001	2.69046

Figure 6 – Studentized Residuals vs MEDV (Full model)

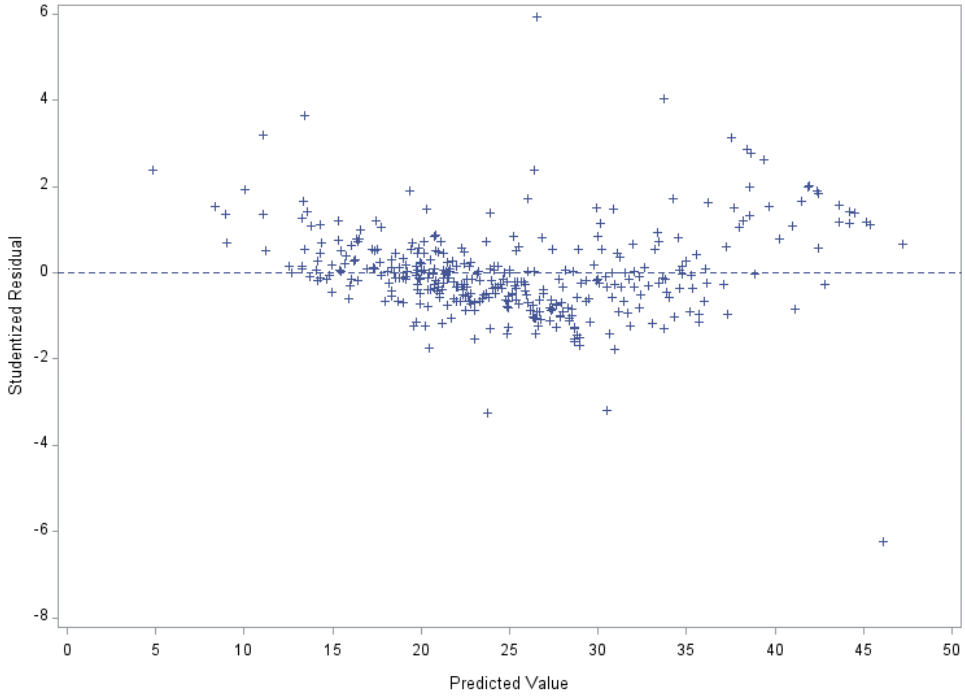


Figure 7 – Studentized Residuals vs DIS (Full model)

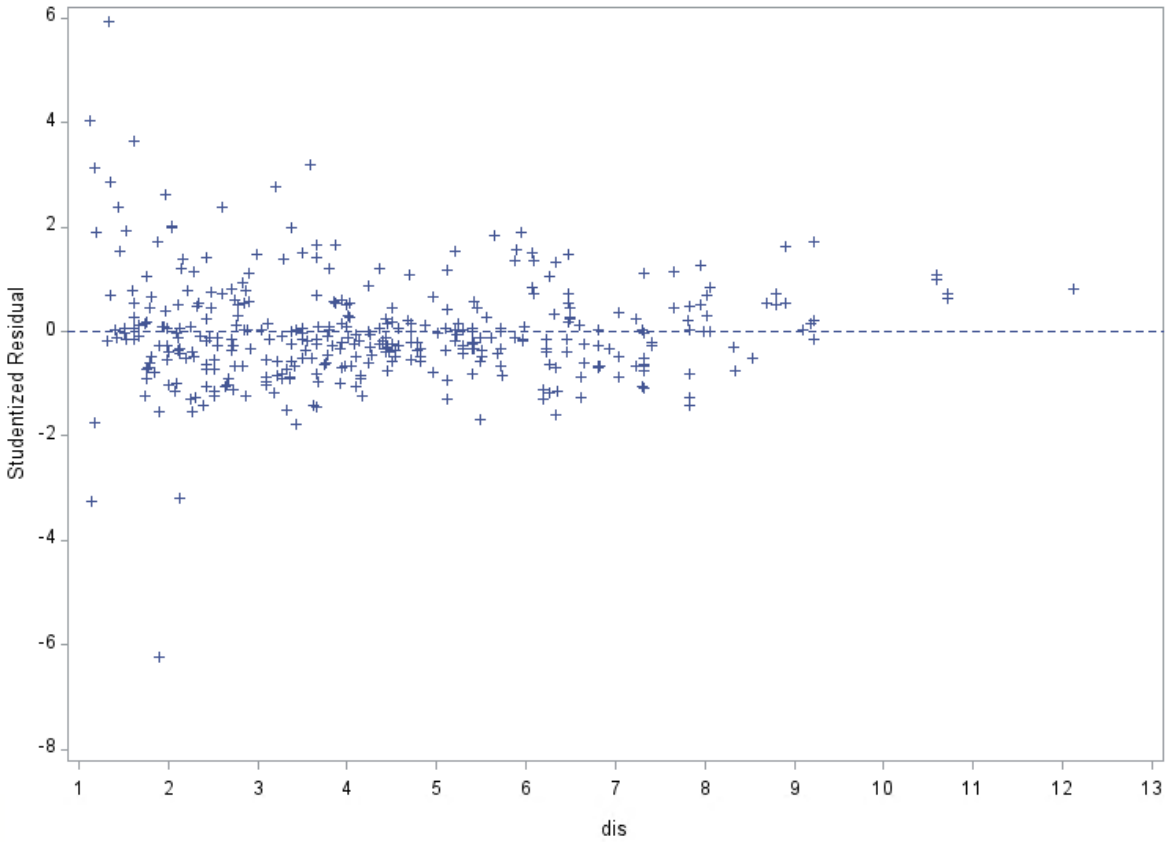


Figure 8 – Studentized Residuals vs LSTAT (Full Model)

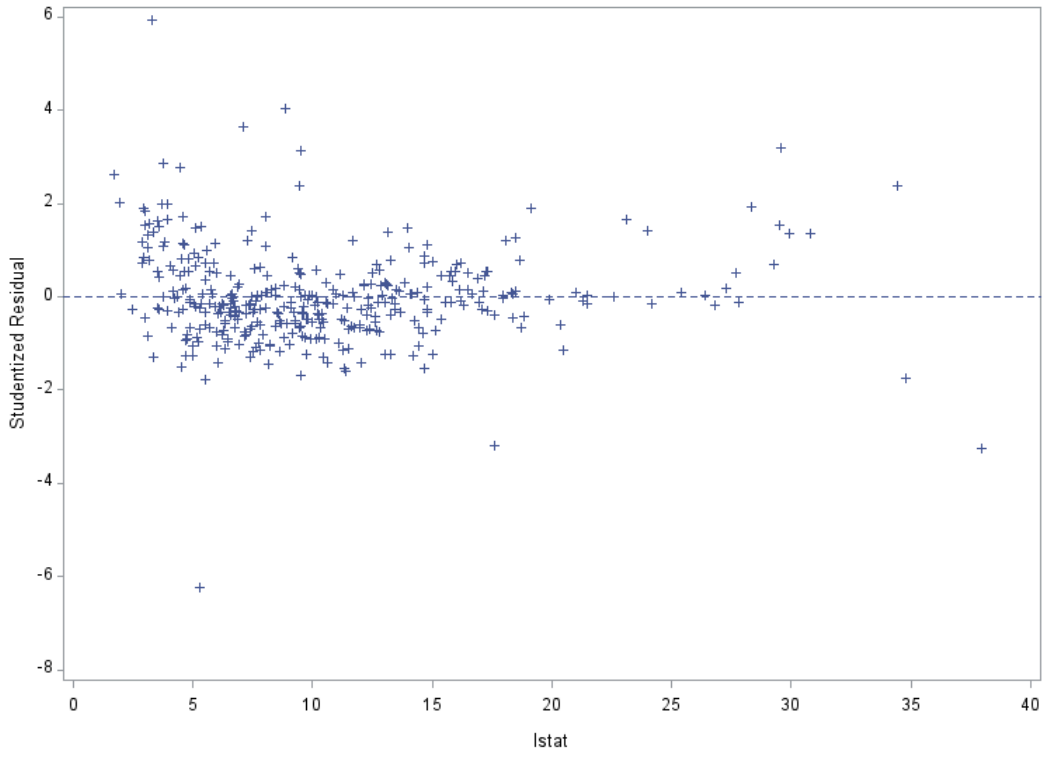


Figure 9 – Normal Probability Plot (Full Model)

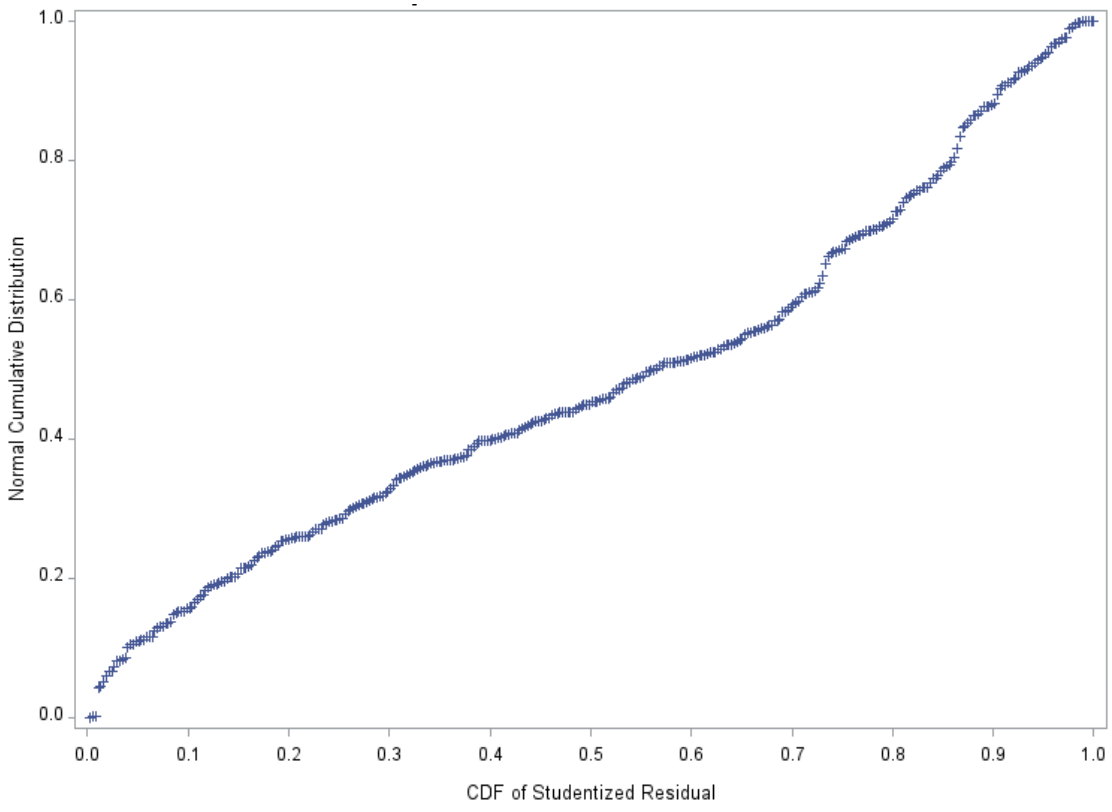


Figure 10 – Histogram of MEDV

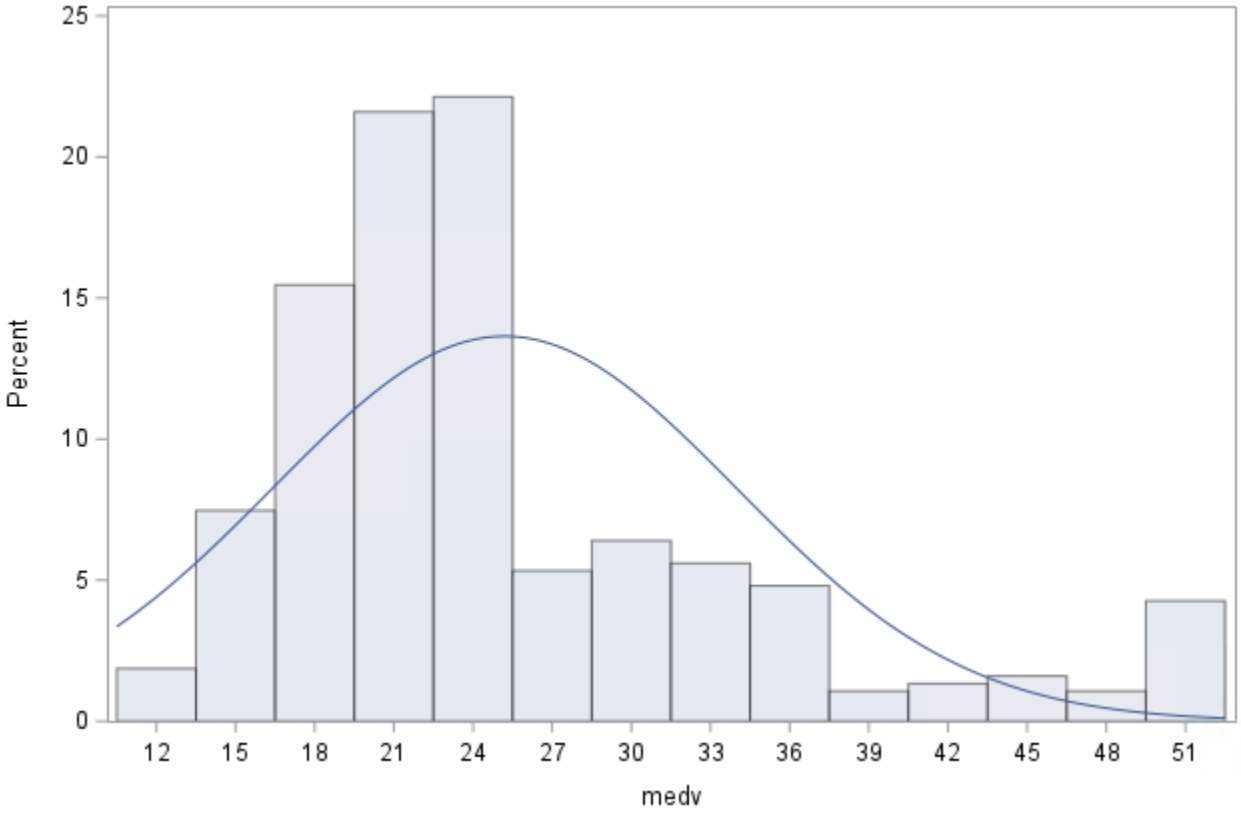


Figure 11 – Histogram of CRIME

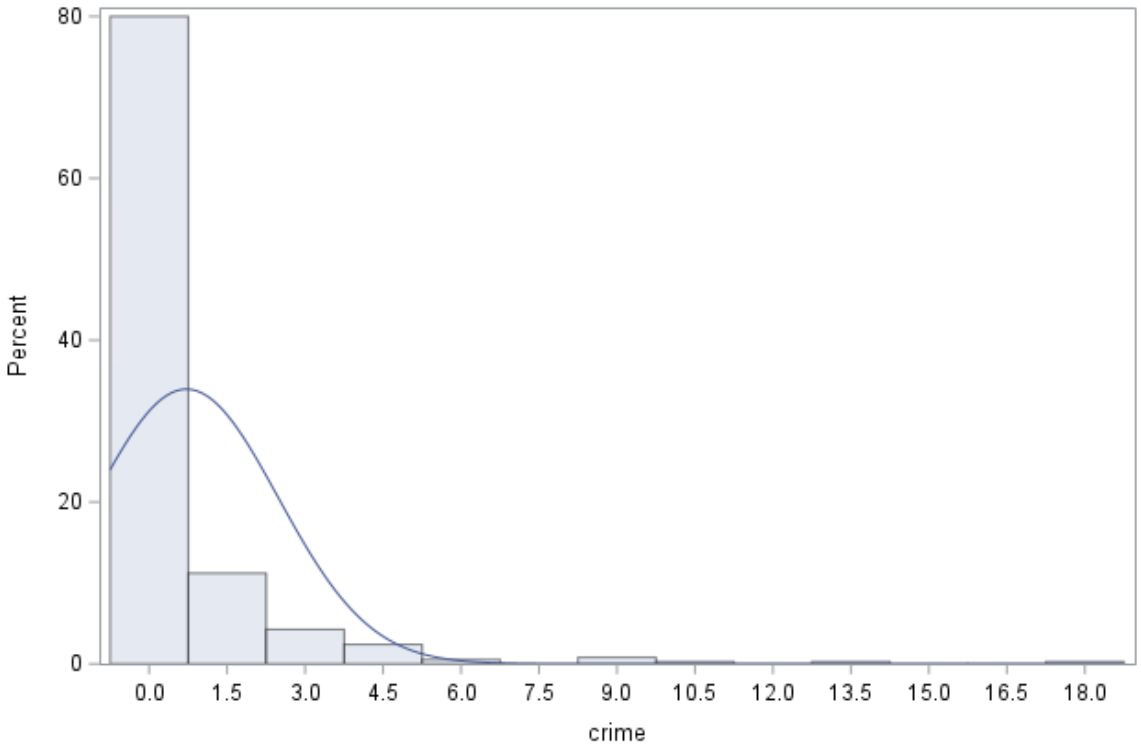


Figure 12 – Histogram of ZN

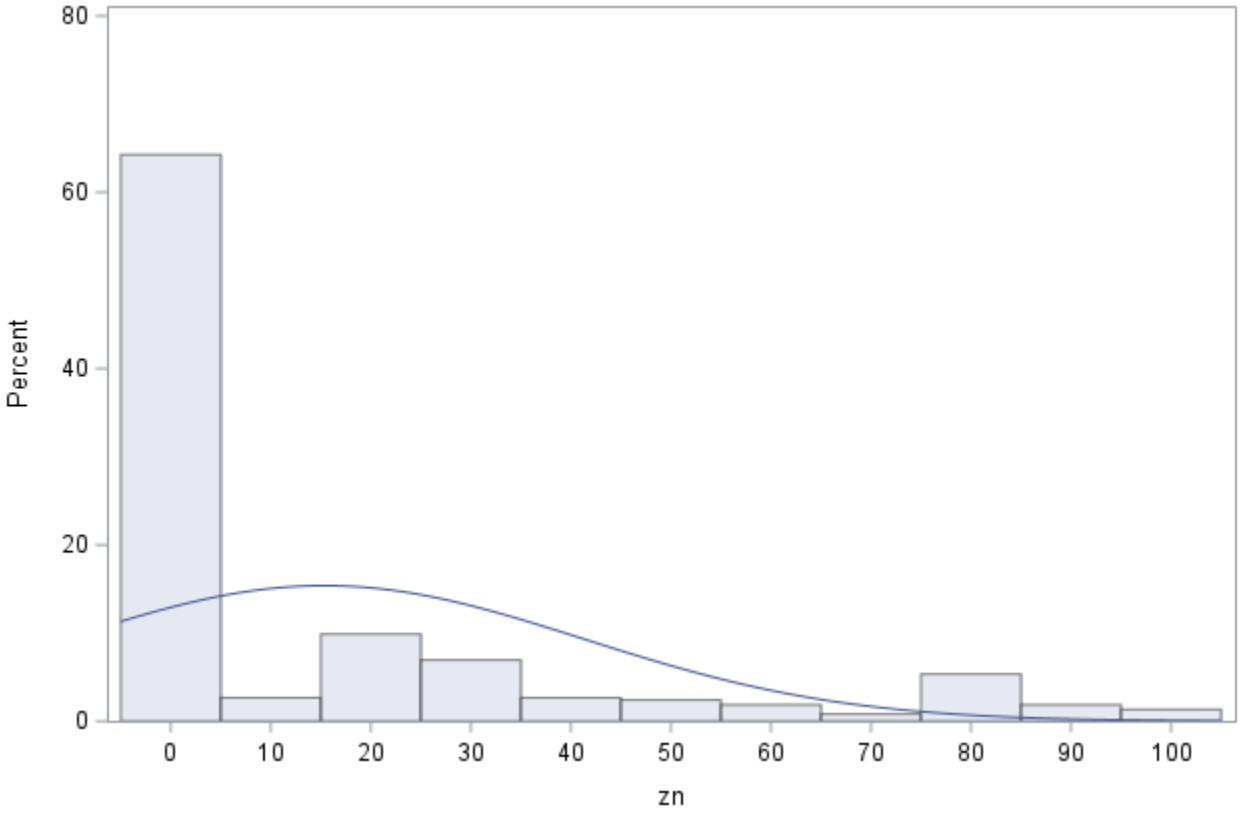


Figure 13 – RAD Frequency Table

rad	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	15	4.00	15	4.00
2	24	6.40	39	10.40
3	38	10.13	77	20.53
4	105	28.00	182	48.53
5	115	30.67	297	79.20
6	18	4.80	315	84.00
7	17	4.53	332	88.53
8	24	6.40	356	94.93
24	19	5.07	375	100.00

Figure 14 – MEDV vs DIS

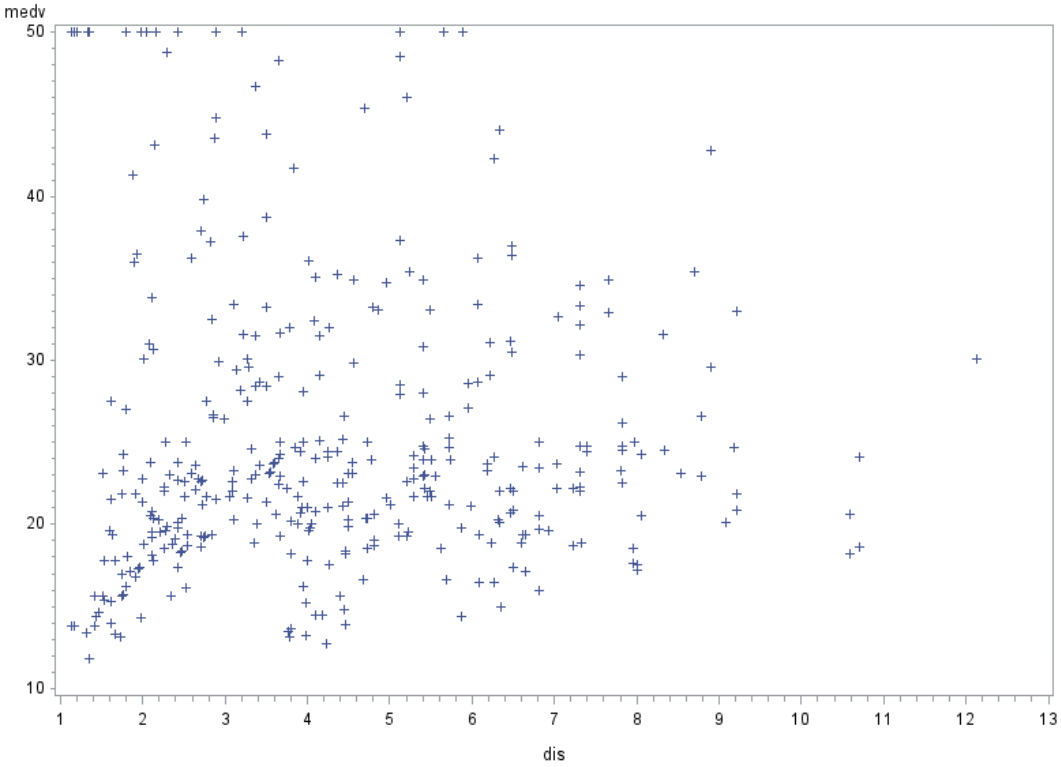


Figure 15 – MEDV vs LSTAT

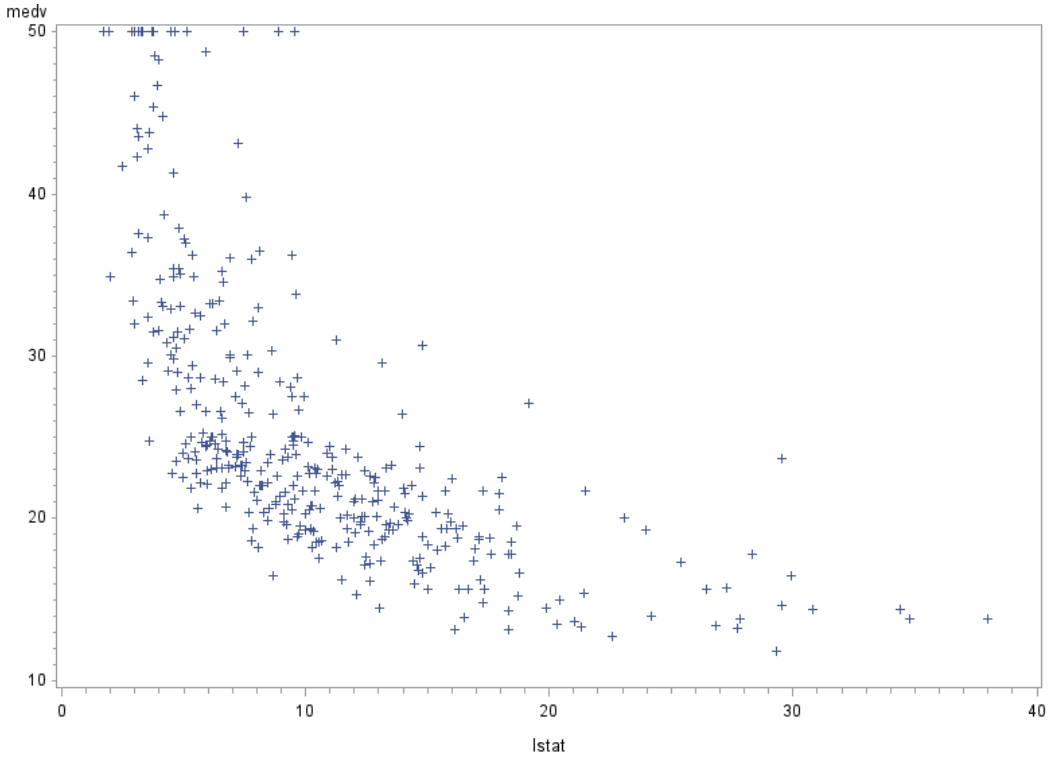


Figure 16 – Progression of Model after Variable Transformation			
	Full Model	After log-transformation of MEDV	After log-transformation of DIS, LSTAT
F-Value	98.53	127.32	156.42
RMSE	4.18606	0.13561	0.12443
R^2	0.7801	0.8209	0.8492
Adusted-R^2	0.7722	0.8145	0.8438

Figure 17 – Outliers/Influential Points (Removed after Variable Transformations)		
Observation #	Studentized Residual Value	Cook's D value
8	3.262	0.031
215	3.366	0.065
365	-6.091	0.434
366	3.035	0.127
369	3.511	0.147
372	3.492	0.089
373	3.964	0.1
374	-3.013	0.109
375	-3.134	0.597

Figure 18 – Regression Output (Full Model after Transformations and Outlier Removal)

Dependent Variable: lnmedv

Number of Observations Read	375
Number of Observations Used	375

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	31.48422	2.42186	156.42	<.0001
Error	361	5.58940	0.01548		
Corrected Total	374	37.07362			

Root MSE	0.12443	R-Square	0.8492
Dependent Mean	3.17438	Adj R-Sq	0.8438
Coeff Var	3.91985		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.06791	0.18882	21.54	<.0001	0
crime	1	0.00935	0.00664	1.41	0.1601	3.30894
zn	1	0.00042927	0.00035007	1.23	0.2209	1.99354
indus	1	-0.00028705	0.00173	-0.17	0.8682	2.87563
chas	1	0.01071	0.02371	0.45	0.6519	1.15256
nox	1	-0.92500	0.12396	-7.46	<.0001	4.45972
rm	1	0.15048	0.01462	10.29	<.0001	2.69230
age	1	-0.00050366	0.00039305	-1.28	0.2009	3.07390
lndis	1	-0.25117	0.02687	-9.35	<.0001	4.65280
rad	1	0.02108	0.00288	7.31	<.0001	4.17983
tax	1	-0.00060357	0.00011518	-5.24	<.0001	3.35177
ptratio	1	-0.02878	0.00365	-7.88	<.0001	1.58572
minor	1	0.00034377	0.00017531	1.96	0.0507	1.27897
lnlstat	1	-0.24901	0.02019	-12.33	<.0001	3.16729

Figure 19 – Studentized Residuals vs LNMEDV (Full Model after Transformations and Outlier Removal)

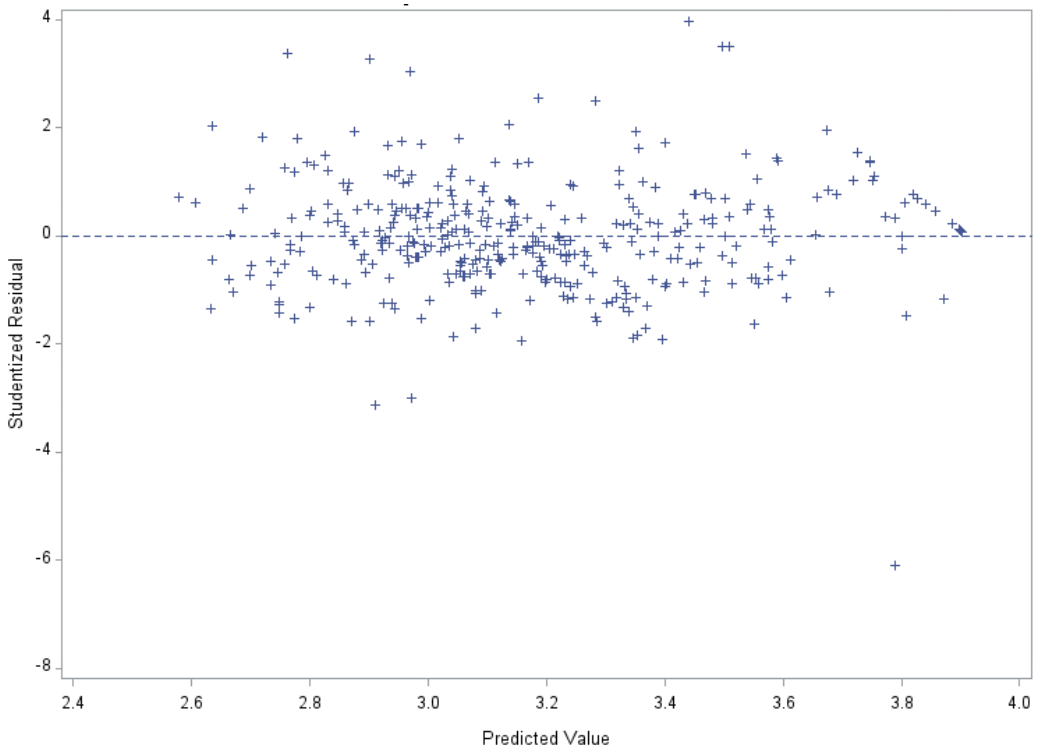


Figure 20 – Studentized Residuals vs LNDIS (Full Model after Transformations and Outlier Removal)

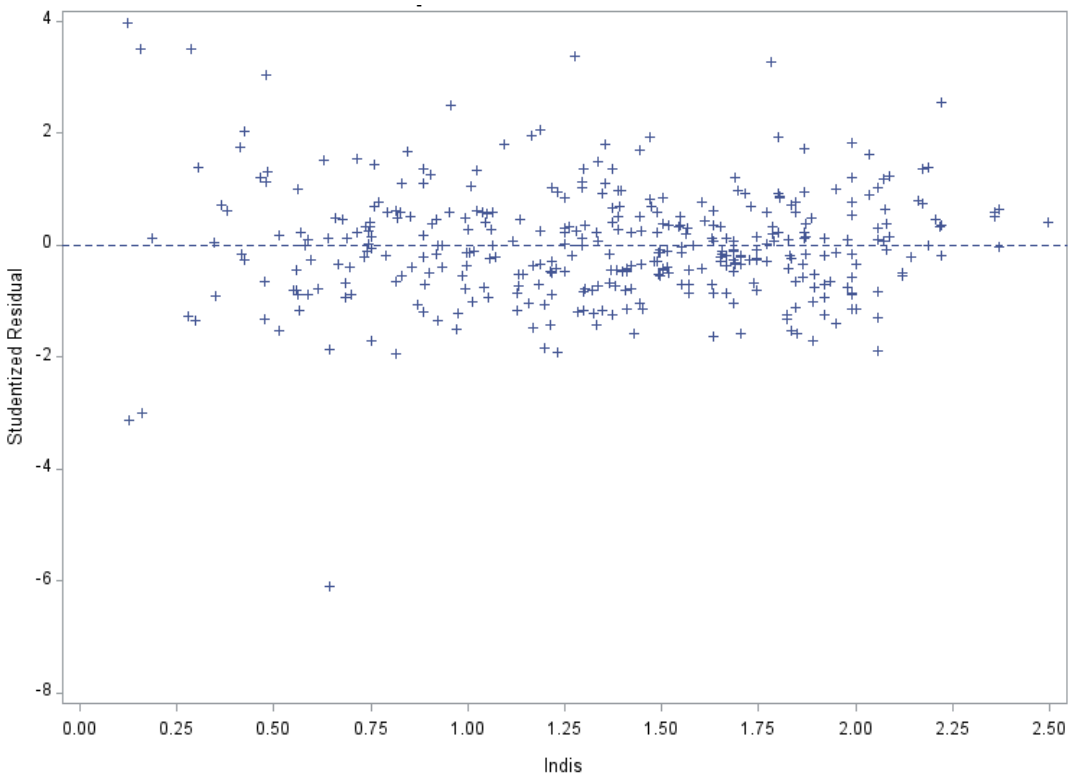


Figure 21 – Studentized Residuals vs LNLSTAT (Full Model after Transformations and Outlier Removal)

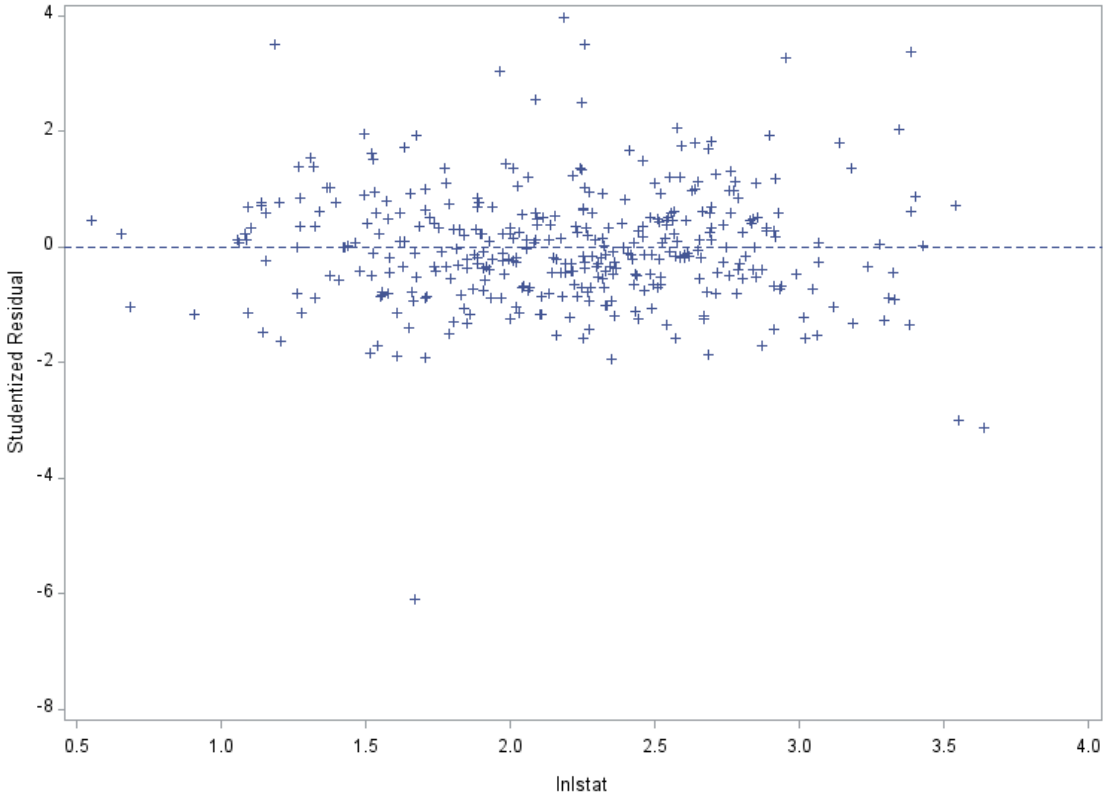


Figure 22 – Normal Probability Plot (Full Model after Transformations and Outlier Removal)

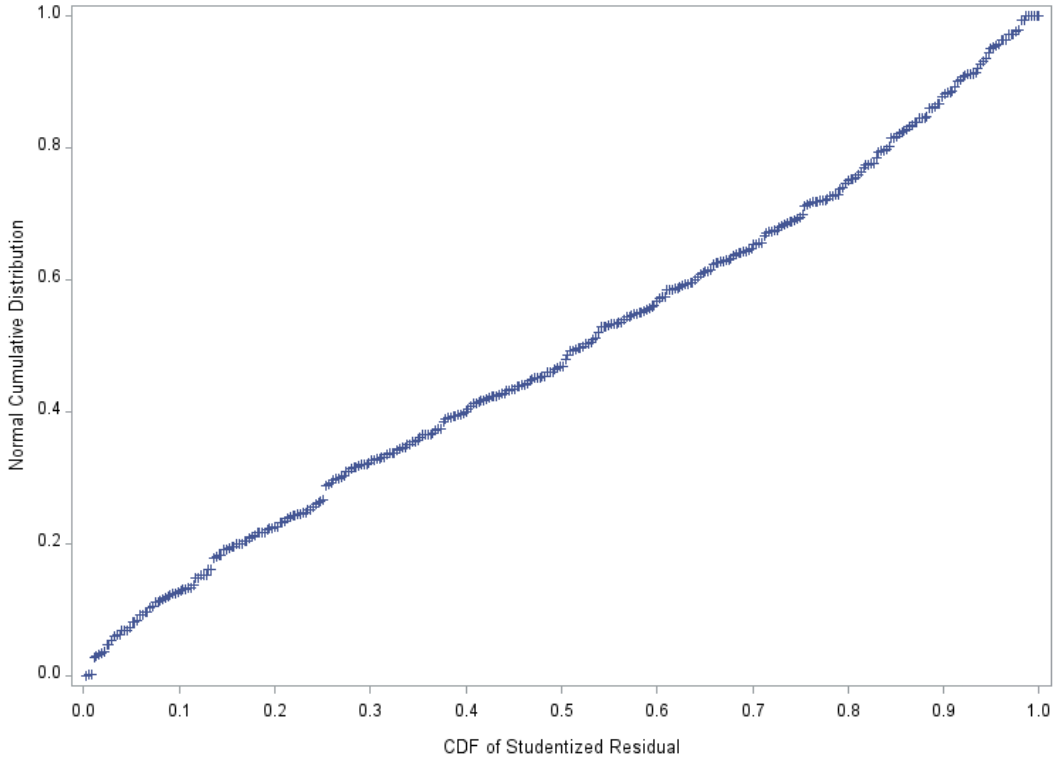


Figure 23 – PROC GLMSELECT RESULTS (Stepwise Method)

Data Set	WORK.BOSTON_PRESELECTION
Dependent Variable	lnmedv
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	10
Effect Hierarchy Enforced	None
Random Number Seed	911752001

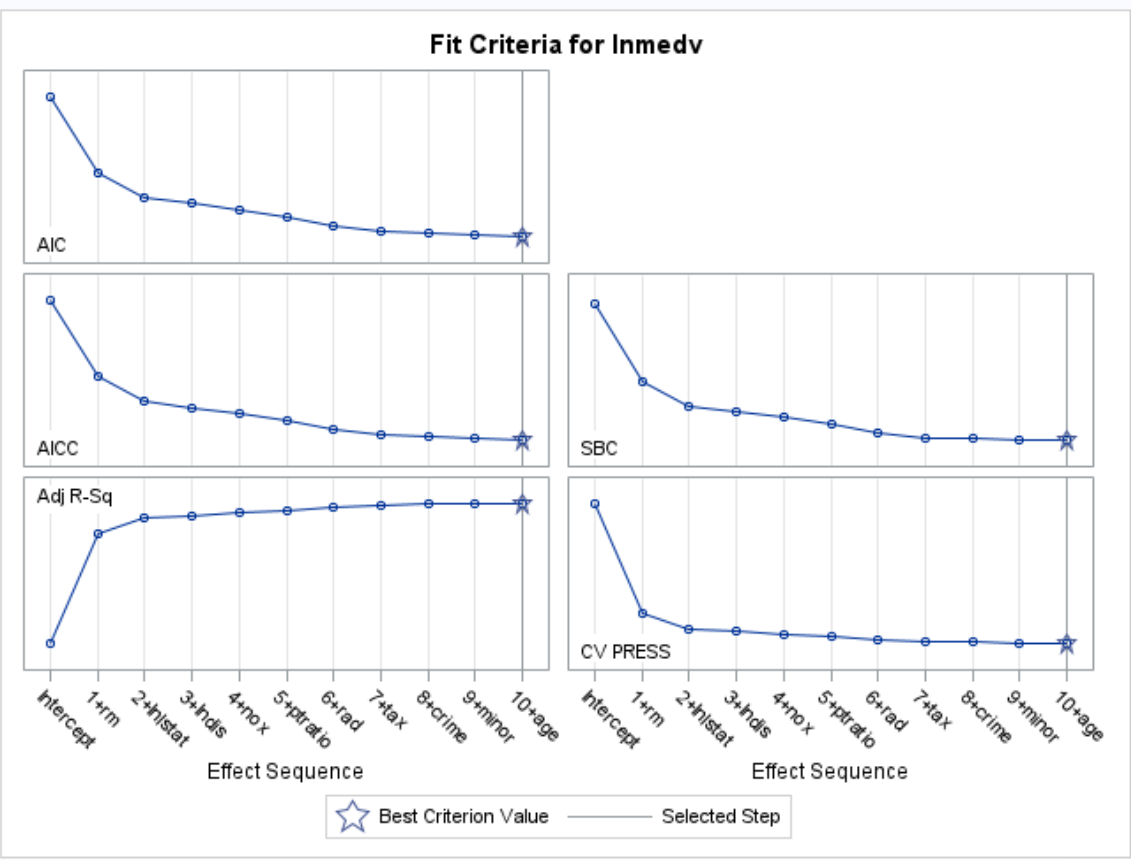
Number of Observations Read	366
Number of Observations Used	366
Number of Observations Used for Training	268
Number of Observations Used for Testing	98

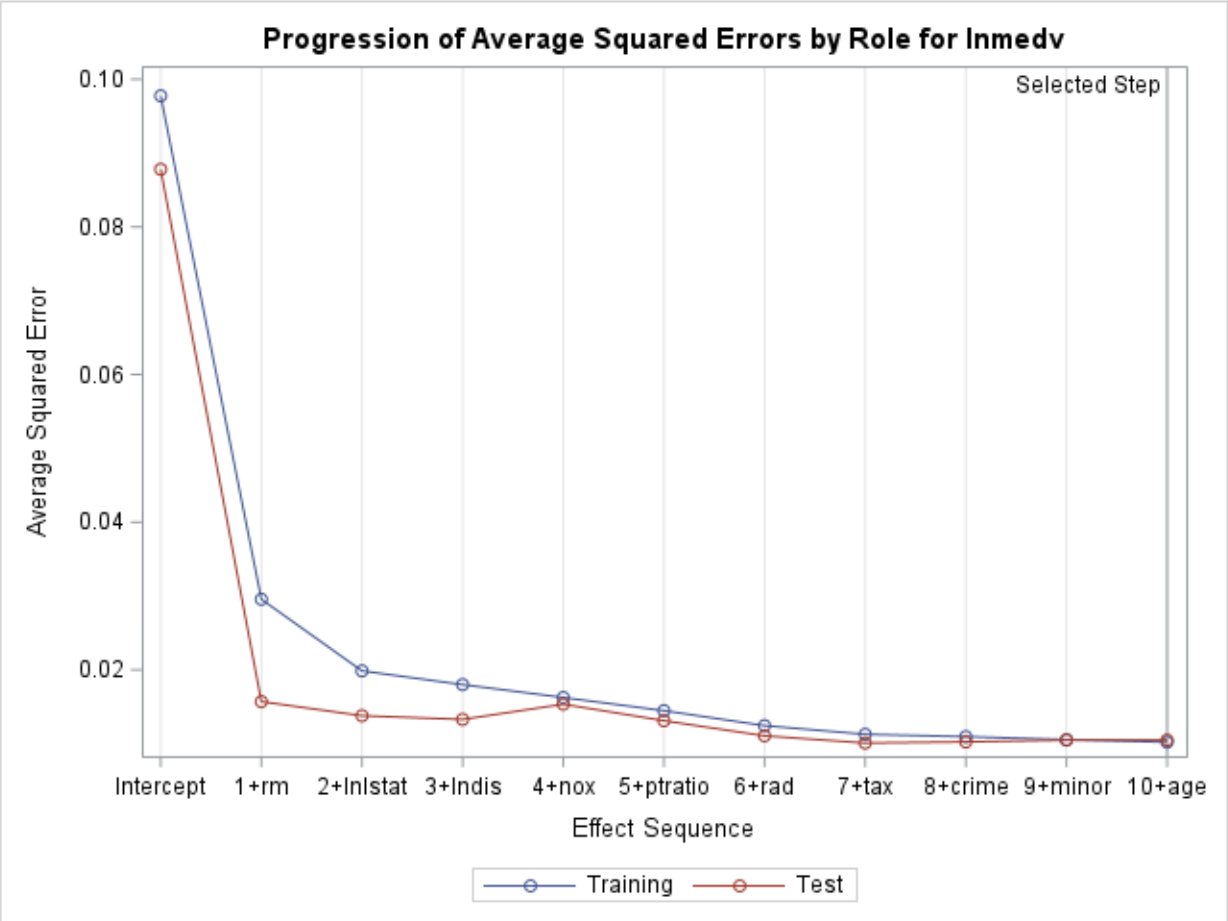
Dimensions	
Number of Effects	14
Number of Parameters	14

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0	Intercept		1	-617.5734	0.0978	0.0878	26.2998
1	rm		2	-932.5325	0.0296	0.0157	8.0763
2	lnlstat		3	-1033.5652	0.0199	0.0138	5.4385
3	lndis		4	-1054.3199	0.0180	0.0133	4.9796
4	nox		5	-1076.1134	0.0163	0.0153	4.5800
5	ptratio		6	-1101.6551	0.0145	0.0131	4.1726
6	rad		7	-1136.4951	0.0124	0.0111	3.5933
7	tax		8	-1157.5002	0.0113	0.0101	3.2603
8	crime		9	-1159.5476	0.0109	0.0102	3.1310
9	minor		10	-1164.1968	0.0105	0.0105	3.0163
10	age		11	-1167.4264*	0.0102	0.0105	2.9408*

* Optimal Value of Criterion





Effects:	Intercept crime nox rm age lndis rad tax ptratio minor lnlstat

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	10	23.46657	2.34666	220.66
Error	257	2.73310	0.01063	
Corrected Total	267	26.19967		

Root MSE	0.10312
Dependent Mean	3.16441
R-Square	0.8957
Adj R-Sq	0.8916
AIC	-936.92722
AICC	-935.70369
SBC	-1167.42636
ASE (Train)	0.01020
ASE (Test)	0.01048
CV PRESS	2.94083

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	3.519184	0.201928	17.43
crime	1	0.044399	0.010277	4.32
nox	1	-0.924837	0.117972	-7.84
rm	1	0.202045	0.016087	12.56
age	1	-0.001158	0.000395	-2.93
lndis	1	-0.238389	0.024854	-9.59
rad	1	0.015704	0.002879	5.45
tax	1	-0.000627	0.000107	-5.88
ptratio	1	-0.024583	0.003345	-7.35
minor	1	0.000628	0.000188	3.33
lnlstat	1	-0.212515	0.022115	-9.61

Figure 24 – PROC GLMSELECT RESULTS (Stepwise Method)

Data Set	WORK.BOSTON_PRESELECTION
Dependent Variable	lnmedv
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	10
Effect Hierarchy Enforced	None
Random Number Seed	304619001

Number of Observations Read	366
Number of Observations Used	366
Number of Observations Used for Training	286
Number of Observations Used for Testing	80

Dimensions	
Number of Effects	14
Number of Parameters	14

Backward Selection Summary

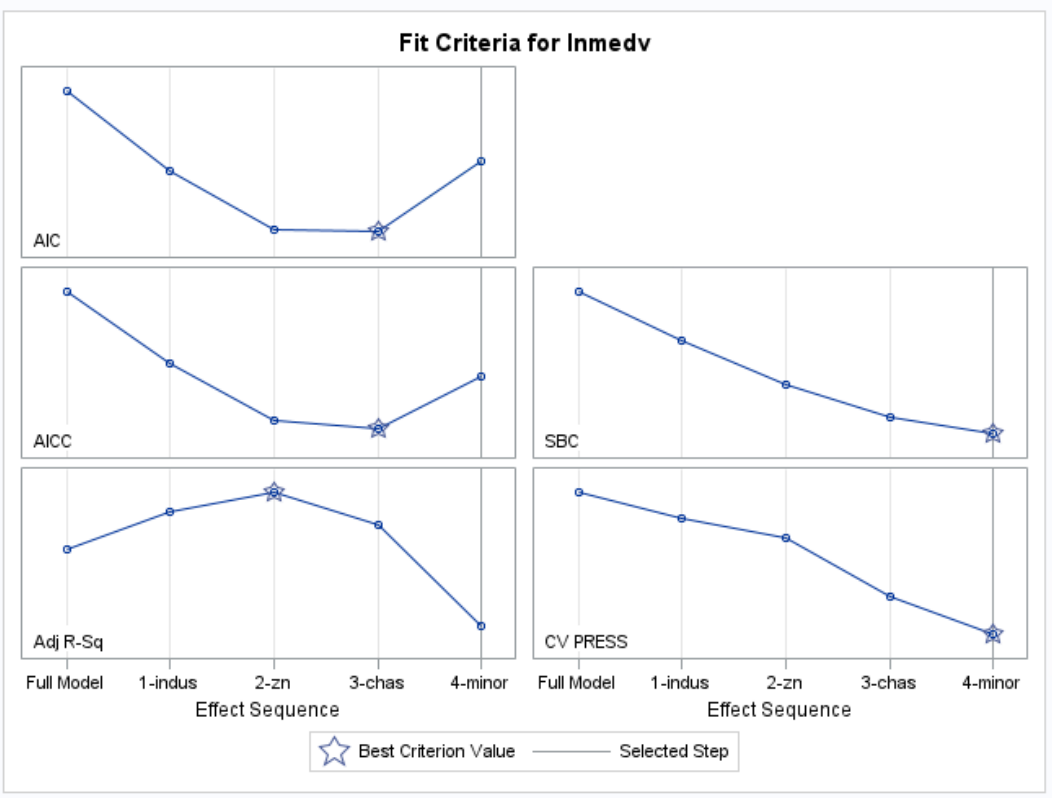
Step	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0		14	-1244.2868	0.0098	0.0114	3.3087
1	indus	13	-1249.9173	0.0098	0.0114	3.2853
2	zn	12	-1255.0548	0.0098	0.0115	3.2681
3	chas	11	-1258.7532	0.0099	0.0112	3.2175
4	minor	10	-1260.6715*	0.0100	0.0120	3.1846*

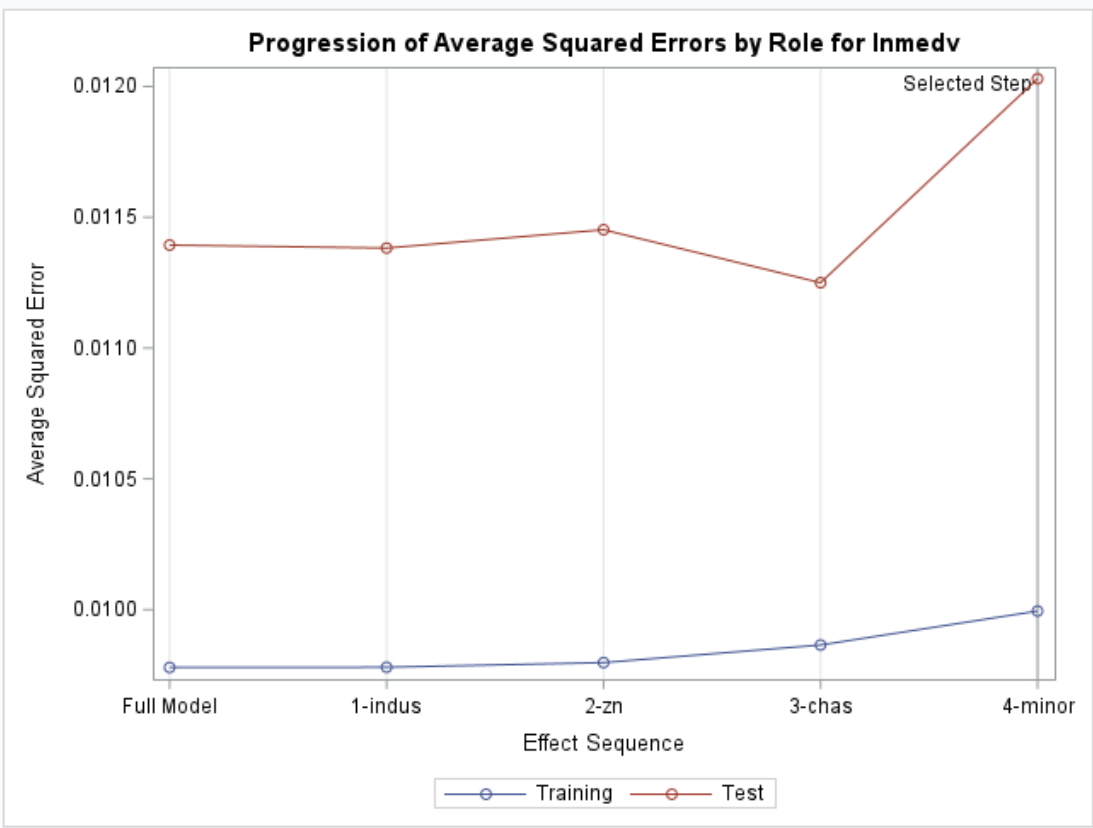
* Optimal Value of Criterion

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

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	age	3.2291	>	3.1846





Effects:	Intercept crime nox rm age lndis rad tax ptratio lnlstat

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	9	23.16952	2.57439	248.57
Error	276	2.85847	0.01036	
Corrected Total	285	26.02799		

Root MSE	0.10177
Dependent Mean	3.17166
R-Square	0.8902
Adj R-Sq	0.8866
AIC	-1009.23140
AICC	-1008.26790
SBC	-1260.67148
ASE (Train)	0.00999
ASE (Test)	0.01203
CV PRESS	3.18459

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	3.633434	0.174314	20.84
crime	1	0.023616	0.008337	2.83
nox	1	-0.959444	0.118421	-8.10
rm	1	0.216209	0.015372	14.07
age	1	-0.001009	0.000375	-2.69
lndis	1	-0.232234	0.022783	-10.19
rad	1	0.019548	0.002744	7.12
tax	1	-0.000696	0.000103	-6.78
ptratio	1	-0.024063	0.003117	-7.72
lnlstat	1	-0.194024	0.021214	-9.15

Figure 25 – Summary of PRC GLMSELECT Outputs		
Parameters	Stepwise	Backward
# of Variables in Model	10	9
# of Training Observations	268	286
# of Test Observations	98	80
F-Value	220.66	248.57
RMSE	0.10312	0.10177
R^2	0.8957	0.8902
Adjusted R^2	0.8916	0.8866
ASE (Train)	0.0102	0.00999
ASE (Test)	0.01048	0.01203
CVPRESS	2.94083	3.18459

Figure 26 – Regression Output of GLMSELECT Stepwise Model

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: new_Invalue

Number of Observations Read	366
Number of Observations Used	275
Number of Observations with Missing Values	91

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	24.51617	2.45162	227.75	<.0001
Error	264	2.84180	0.01076		
Corrected Total	274	27.35797			

Root MSE	0.10375	R-Square	0.8961
Dependent Mean	3.16182	Adj R-Sq	0.8922
Coeff Var	3.28138		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.45156	0.19868	17.37	<.0001
crime	1	0.03209	0.00892	3.60	0.0004
nox	1	-0.87981	0.12355	-7.12	<.0001
rm	1	0.21085	0.01617	13.04	<.0001
age	1	-0.00068185	0.00039037	-1.75	0.0819
lndis	1	-0.20889	0.02445	-8.54	<.0001
rad	1	0.01452	0.00286	5.08	<.0001
tax	1	-0.00057366	0.00010291	-5.57	<.0001
ptratio	1	-0.02718	0.00342	-7.95	<.0001
minor	1	0.00045427	0.00016485	2.76	0.0063
lnlstat	1	-0.20394	0.02217	-9.20	<.0001

Figure 27 – Test Output of GLMSELECT Stepwise Model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	91	0.095888	0.074553

Pearson Correlation Coefficients, N = 91 Prob > |r| under H0: Rho=0		
	lnmedv	yhat
lnmedv	1.00000	0.94326 <.0001
yhat Predicted Value of new_lnvalue	0.94326 <.0001	1.00000

Figure 28 – Regression Output of GLMSELECT Backward Model

Number of Observations Read	366
Number of Observations Used	275
Number of Observations with Missing Values	91

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	24.43443	2.71494	246.09	<.0001
Error	265	2.92353	0.01103		
Corrected Total	274	27.35797			

Root MSE	0.10503	R-Square	0.8931
Dependent Mean	3.16182	Adj R-Sq	0.8895
Coeff Var	3.32195		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.63420	0.18961	19.17	<.0001
crime	1	0.02450	0.00858	2.85	0.0047
nox	1	-0.92841	0.12380	-7.50	<.0001
rm	1	0.21266	0.01635	13.00	<.0001
age	1	-0.00067009	0.00039517	-1.70	0.0911
Indis	1	-0.21228	0.02472	-8.59	<.0001
rad	1	0.01589	0.00285	5.58	<.0001
tax	1	-0.00058801	0.00010405	-5.65	<.0001
ptratio	1	-0.02675	0.00346	-7.74	<.0001
lnlstat	1	-0.20338	0.02244	-9.06	<.0001

Figure 29 – Test Output of GLMSELECT Backward Model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	91	0.097514	0.076566

Pearson Correlation Coefficients, N = 91 Prob > |r| under H0: Rho=0		
	lnmedv	yhat
lnmedv	1.00000	0.94108 <.0001
yhat Predicted Value of new_lnvalue	0.94108 <.0001	1.00000

Figure 30 – Table of Relevant Training and Test Metrics

TRAIN		
Metrics	GLMSELECT Stepwise (10 vars)	GLMSELECT Backward (9 vars)
RMSE	0.10375	0.10503
R^2	0.8961	0.8931
Adjusted R^2	0.8922	0.8895
F-Value	227.75	246.98
Residuals	Good	Good
TEST		
Metrics	GLMSELECT Stepwise (10 vars)	GLMSELECT Backward (9 vars)
RMSE	0.095888	0.097514
MAE	0.074553	0.076566
R^2	0.88974	0.88563
Adjusted R^2 (calculated)	0.87596	0.87282
Cross-Validated R^2	0.00636	0.00747

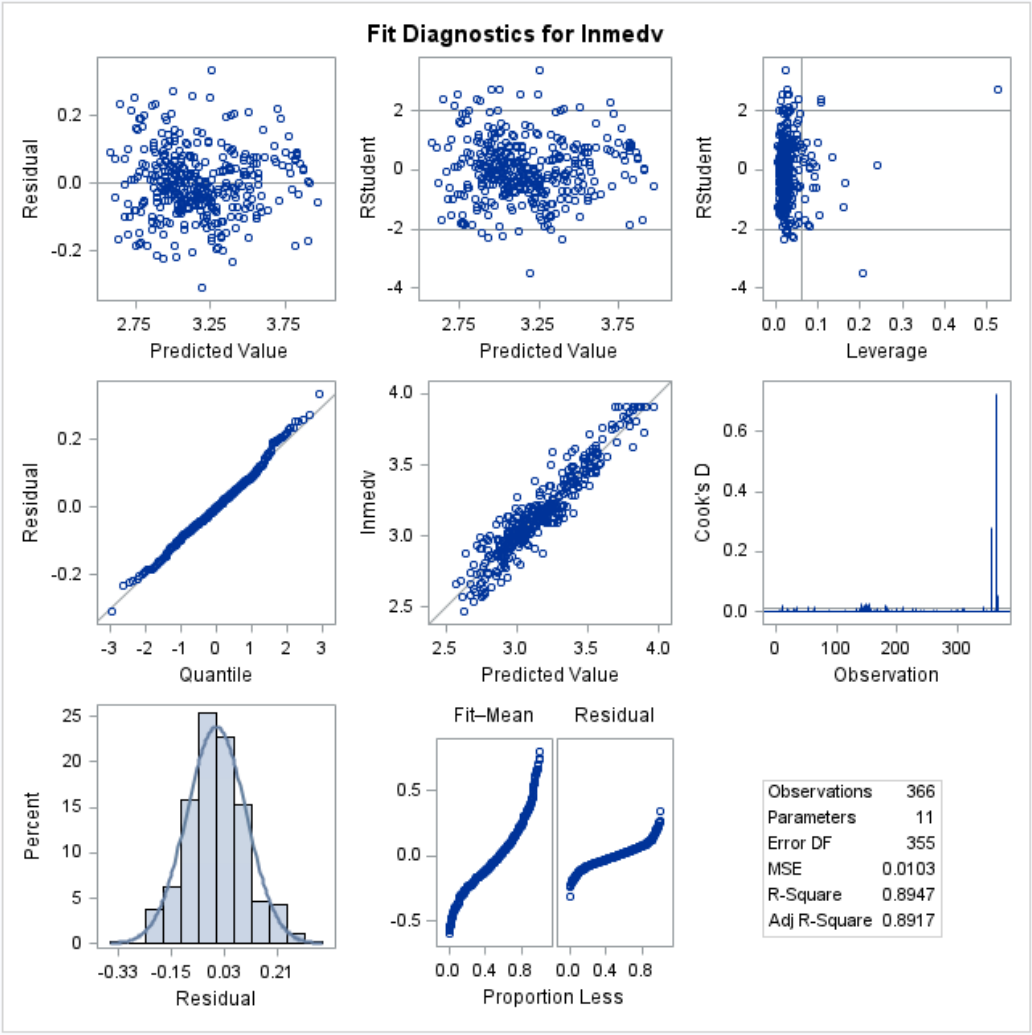
Figure 31 – Final Model Regression Output

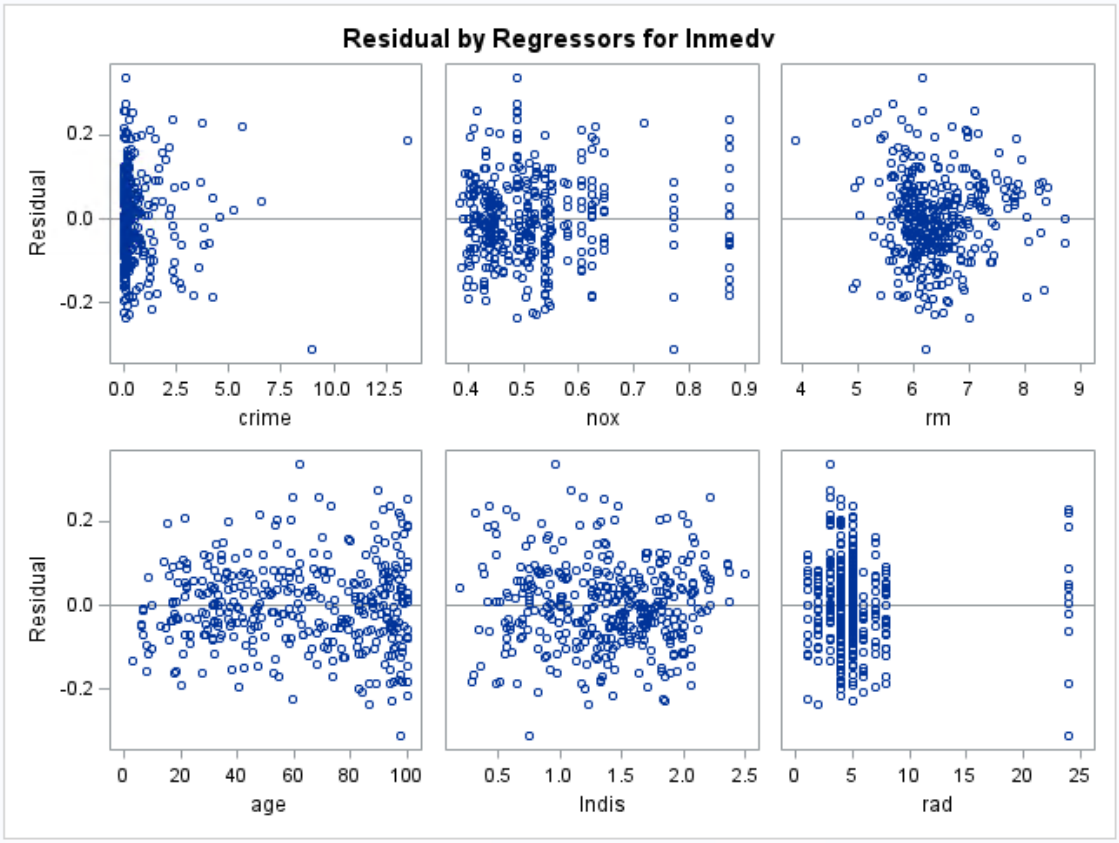
Number of Observations Read	366
Number of Observations Used	366

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	31.12529	3.11253	301.56	<.0001
Error	355	3.66412	0.01032		
Corrected Total	365	34.78941			

Root MSE	0.10159	R-Square	0.8947
Dependent Mean	3.17088	Adj R-Sq	0.8917
Coeff Var	3.20399		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	3.38750	0.16780	20.19	<.0001	0	0
crime	1	0.03193	0.00821	3.89	0.0001	0.12815	3.66259
nox	1	-0.83590	0.10336	-8.09	<.0001	-0.29455	4.47107
rm	1	0.21365	0.01359	15.72	<.0001	0.47347	3.05729
age	1	-0.00108	0.00033780	-3.19	0.0015	-0.09957	3.27707
Indis	1	-0.21603	0.02069	-10.44	<.0001	-0.34975	3.78305
rad	1	0.01578	0.00244	6.48	<.0001	0.19549	3.07208
tax	1	-0.00057811	0.00008864	-6.52	<.0001	-0.17252	2.35865
ptratio	1	-0.02632	0.00283	-9.29	<.0001	-0.18892	1.39327
minor	1	0.00048985	0.00015053	3.25	0.0012	0.06644	1.40524
lnlstat	1	-0.19090	0.01864	-10.24	<.0001	-0.34396	3.80143




Residual by Regressors for lnmedv
Residual
0.2
0.0
-0.2
0.0 2.5 5.0 7.5 10.0 12.5
crime
0.4 0.5 0.6 0.7 0.8 0.9
nox
4 5 6 7 8 9
rm
0 20 40 60 80 100
age
0.5 1.0 1.5 2.0 2.5
lndis
0 5 10 15 20 25
rad

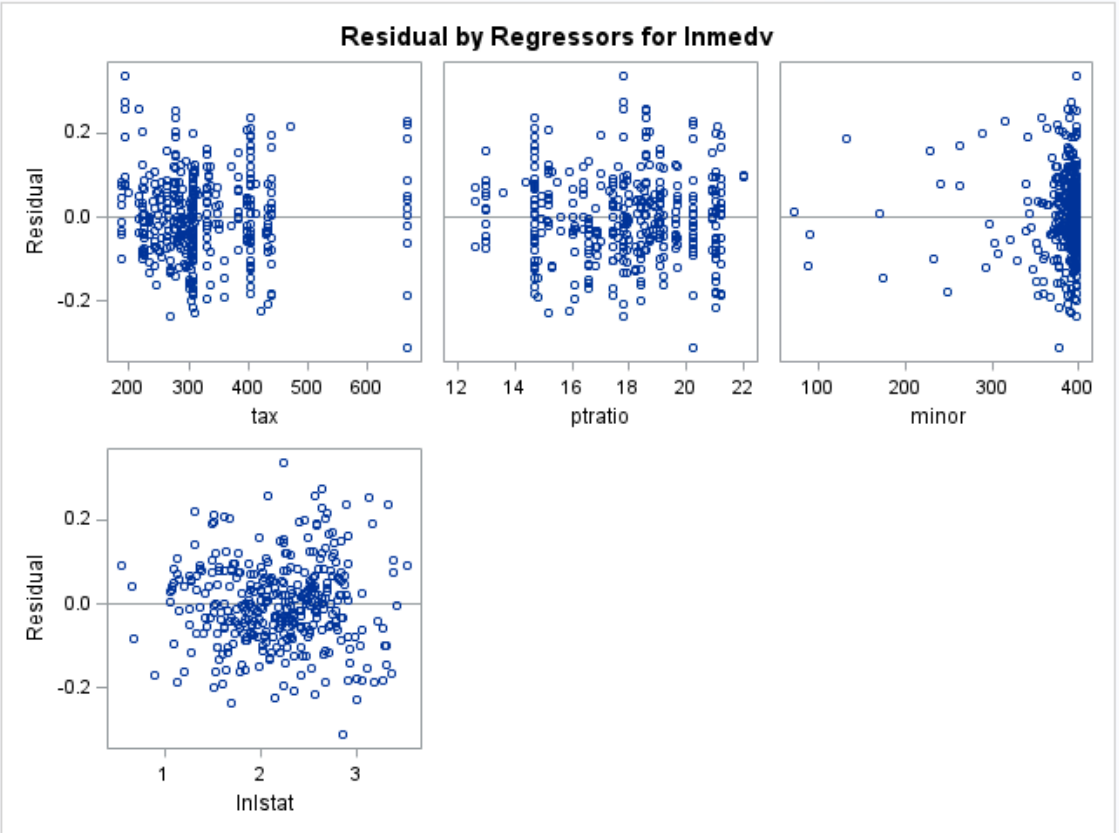

Residual by Regressors for lnmedv
Residual
0.2
0.0
-0.2
200 300 400 500 600
tax
12 14 16 18 20 22
ptratio
100 200 300 400
minor
1 2 3
lnlstat

Figure 32 – Normal Probability Plot for Final Model

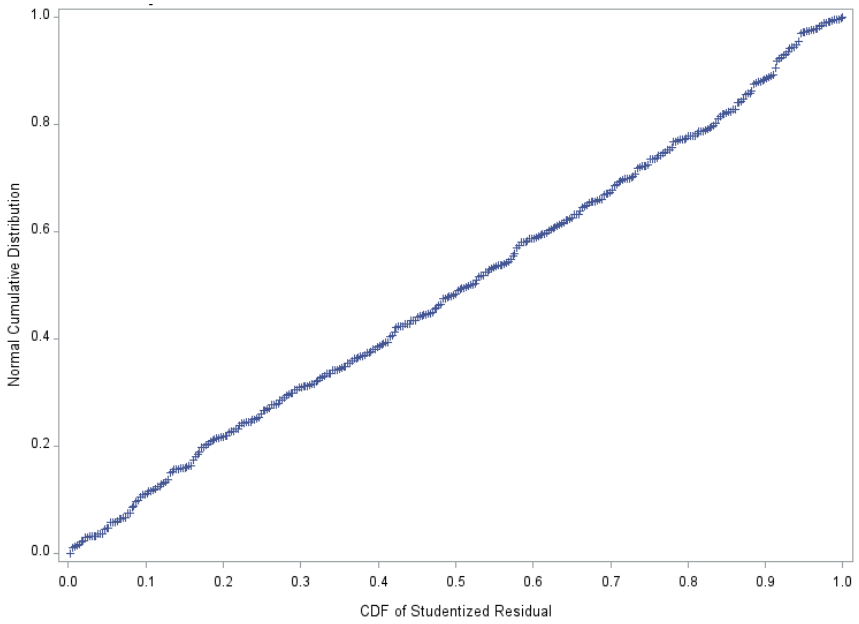


Figure 33 – Table of Minimums and Maximums for each Variable in Trimmed Dataset

Variable	Minimum	Maximum
crime	0.0063	13.5222
nox	0.3850	0.8710
rm	3.8630	8.7250
age	2.9000	100.0000
lndis	0.1843	2.4954
rad	1.0000	24.0000
tax	187.0000	666.0000
ptratio	12.6000	22.0000
minor	70.8000	396.9000
lnlstat	0.5481	3.5383

Figure 34 – Prediction Values

Observation	crime	nox	rm	age	dis	rad	tax	ptratio	minor	lstat
1	0.277	0.718	6.418	35.5	11.8234	15	283	14.6	330.53	23.56
2	9.254	0.452	7.941	18.3	7.4003	5	520	17.2	169.87	4.23

Figure 35 – Prediction Table Output

The REG Procedure
Model: MODEL1
Dependent Variable: lnmedv

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	2.8430	0.0496	2.7455	2.9406	2.6207	3.0654	.
2	.	3.6831	0.0763	3.5331	3.8332	3.4332	3.9330	.
3	3.18	3.2981	0.0155	3.2675	3.3286	3.0959	3.5002	-0.1200
4	3.07	3.1320	0.0116	3.1092	3.1548	2.9309	3.3331	-0.0593
5	3.55	3.4688	0.0136	3.4421	3.4955	3.2672	3.6704	0.0779

Research References

1. Harrison, D and Rubinfield, D.L., Hedonic prices and the demand for clean air', J. Environ. Economics & Management, Vol. 5, 81-102, (1978)
2. Vogt, Ivers, and Associates, Comprehensive land use inventory report, Department of Commerce and Development, Commonwealth of Massachusetts, Boston, Mass. (1965).
3. G. K. Ingram and G. R. Fauth, "TASSIM: A Transportation and Air Shed Simulation Model," Final Report to the U. S. Department of Transportation, National Technical Information Service, Springfield, Va. (May 1974).
4. A. Schnare, An empirical analysis of the dimensions of neighborhood quality, Unpublished Ph.D. Dissertation, Harvard University (1973).
5. Market Business News. (2019, May 6). *What is non-store retailing? Definition and examples*. https://marketbusinessnews.com/financial-glossary/non-store-retailing/
6. Glaeser, E. L., & Sacerdote, B. (1999). Why is There More Crime in Cities? *Journal of Political Economy*, 107(S6), S225–S258. https://doi.org/10.1086/250109
7. Harris, D. R. (1999). "Property Values Drop When Blacks Move in, Because...": Racial and Socioeconomic Determinants of Neighborhood Desirability. *American Sociological Review*, 64(3), 461–479. https://doi.org/10.2307/2657496
8. Nathan Kantrowitz. (1979). Racial and Ethnic Residential Segregation in Boston 1830-1970. *The Annals of the American Academy of Political and Social Science, 441,* 41–54. http://www.jstor.org/stable/1043293