Bode Faleti DSC 423-701 Final Project Report 11/23/2021

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 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 bimodal 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 dependent 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]. In addition, NOX and INDUS have a correlation coefficient of 0.71166 and I believe that is because regions with more non-retail business produce 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 R² 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 ≥ ±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-R², 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 R² methods at this stage. Conveniently, all four selection methods produced the same model with equivalent variables. The R² 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 R² 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-R² (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 R² (0.88974 vs 0.88563) and cross-validated R² (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 R² 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

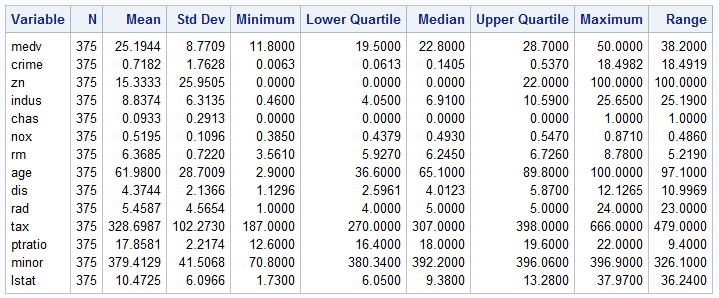


Figure 3 – Median Value by Proximity to Charles River

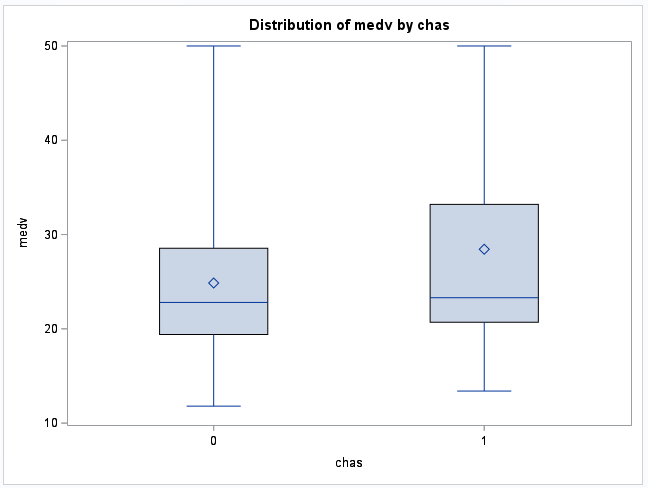


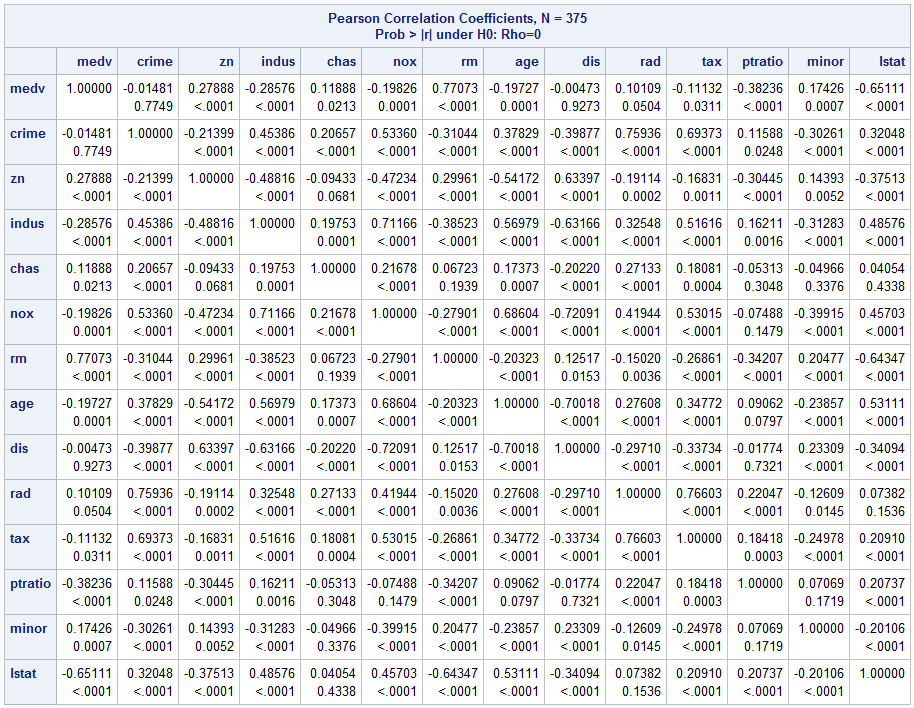
Figure 4 – Pearson Correlation Matrix

Figure 5 – Full Model Regression Output

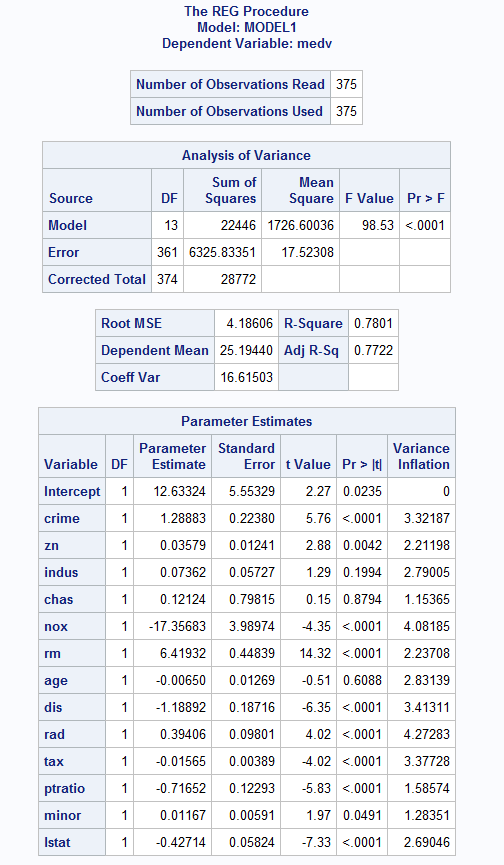


Figure 6 – Studentized Residuals vs MEDV (Full model)

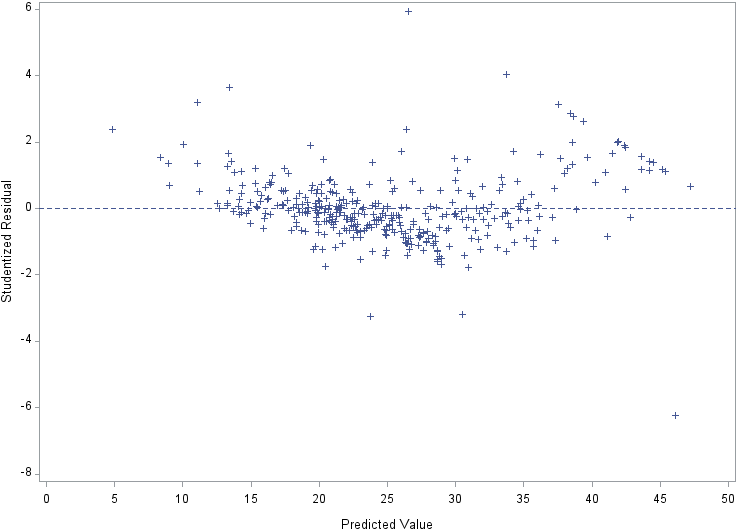


Figure 7 – Studentized Residuals vs DIS (Full model)

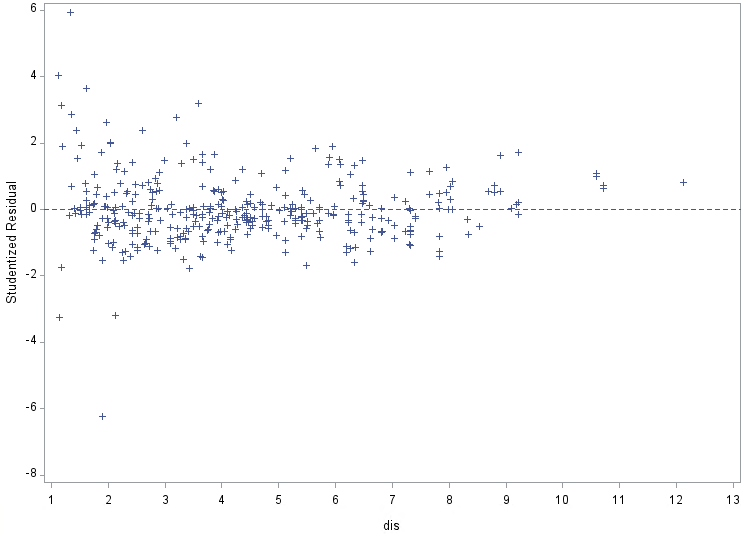


Figure 8 – Studentized Residuals vs LSTAT (Full Model)

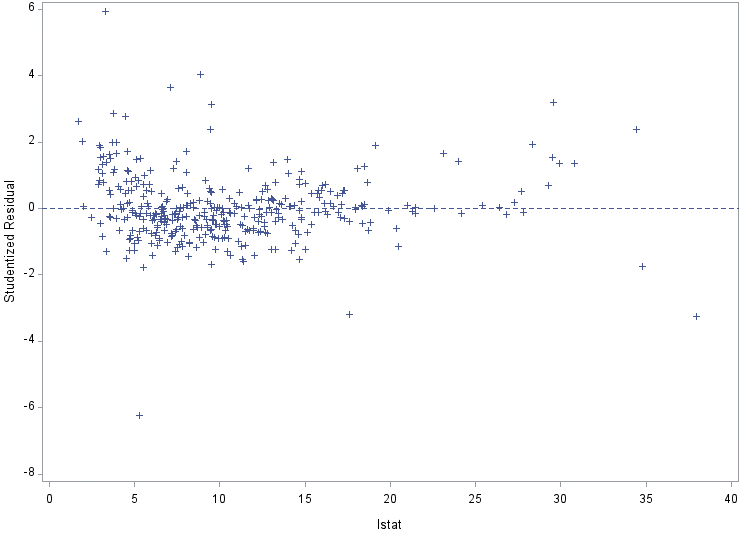


Figure 9 – Normal Probability Plot (Full Model)

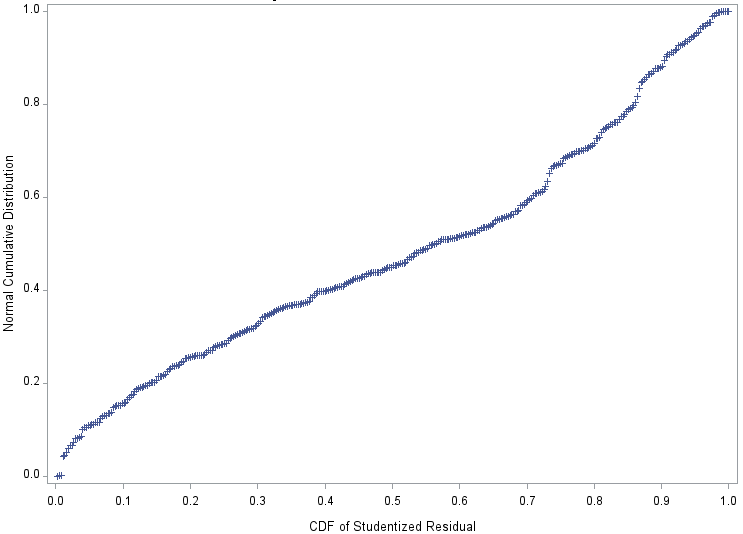


Figure 10 – Histogram of MEDV

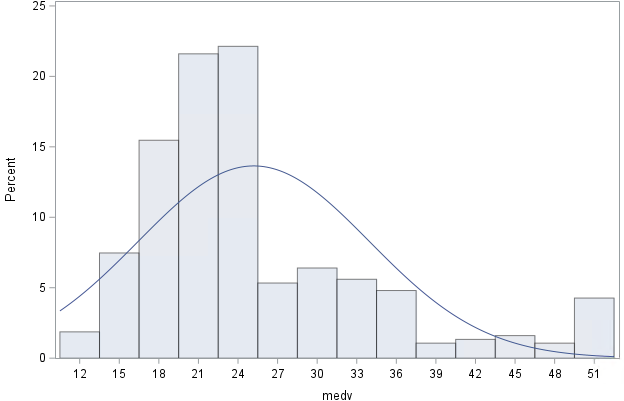


Figure 11 – Histogram of CRIME

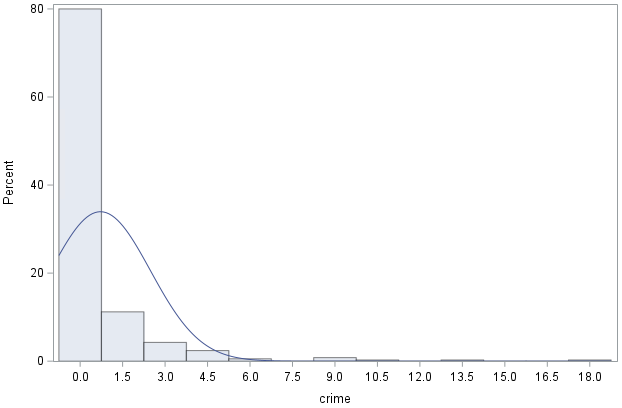


Figure 12 – Histogram of ZN

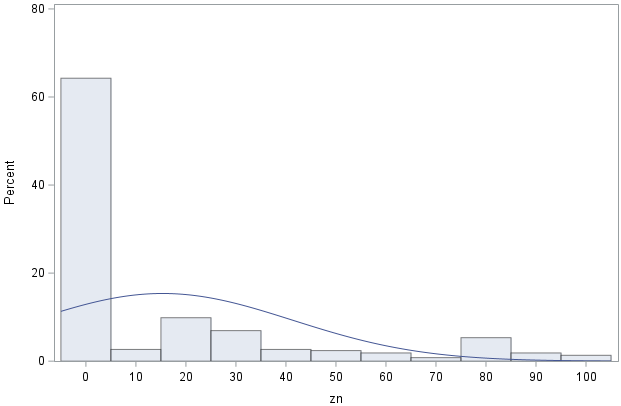


Figure 13 – RAD Frequency Table

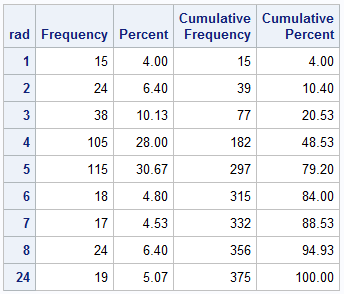


Figure 14 – MEDV vs DIS

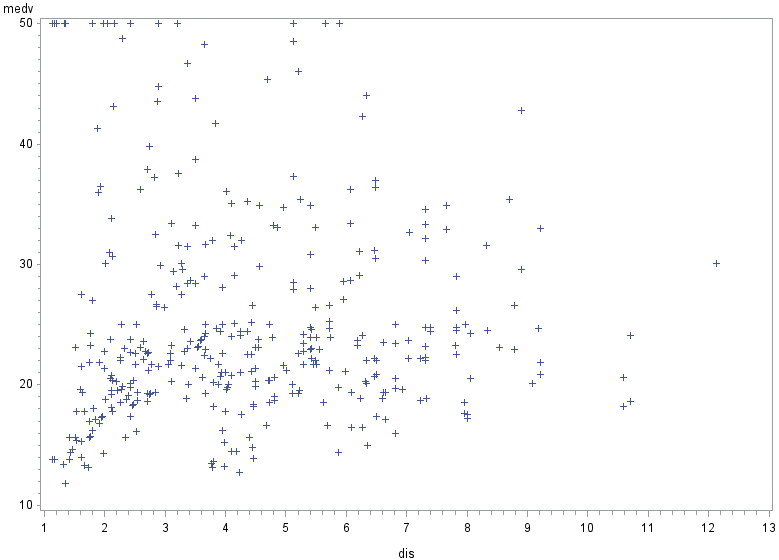
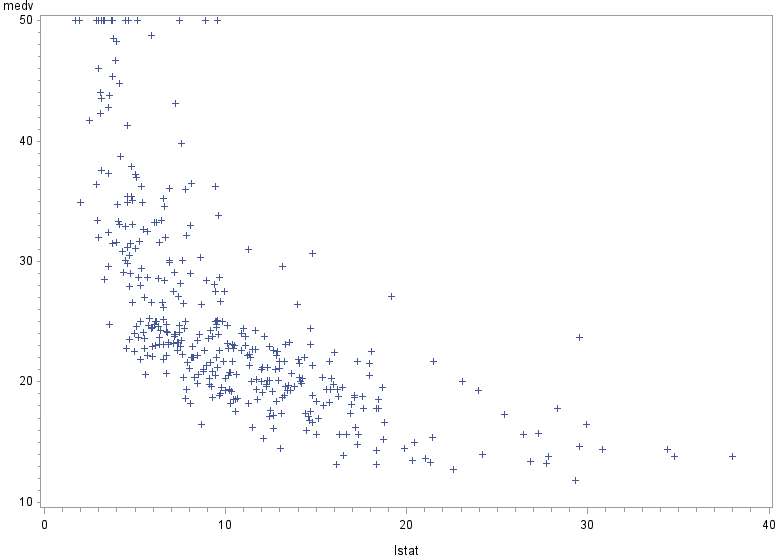


Figure 15 – MEDV vs LSTAT



|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
| R² | 0.7801 | 0.8209 | 0.8492 |
| Adusted-R² | 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)

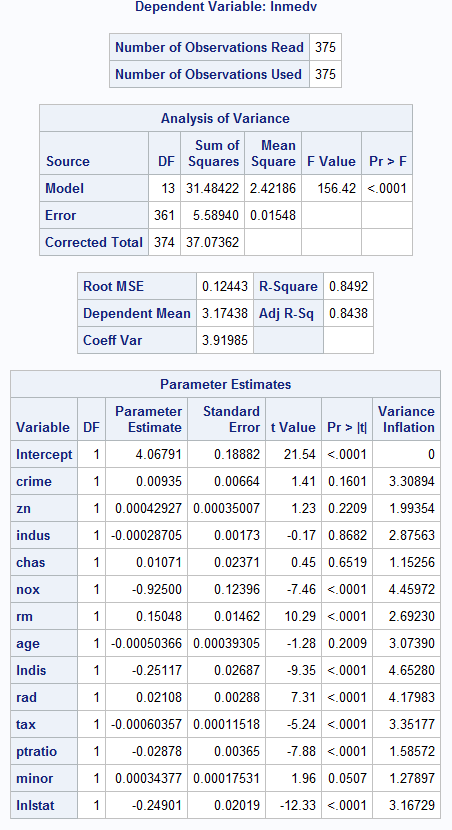


Figure 19 – Studentized Residuals vs LNMEDV (Full Model after Transformations and Outlier Removal)

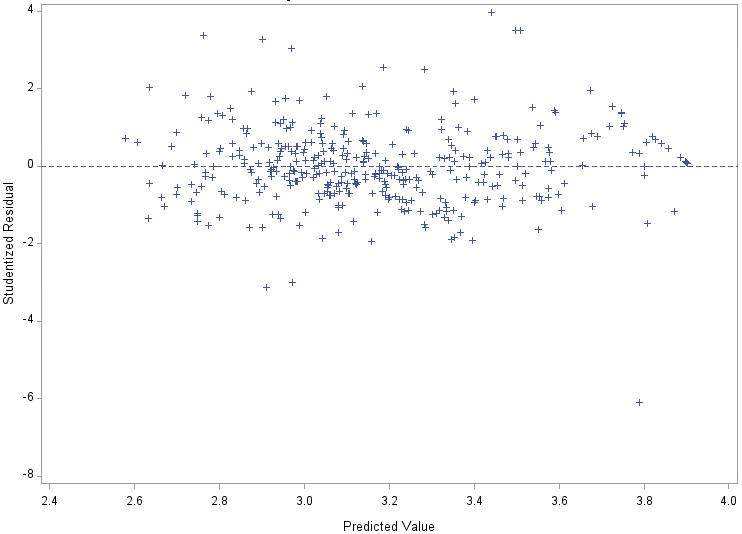


Figure 20 – Studentized Residuals vs LNDIS (Full Model after Transformations and Outlier Removal)

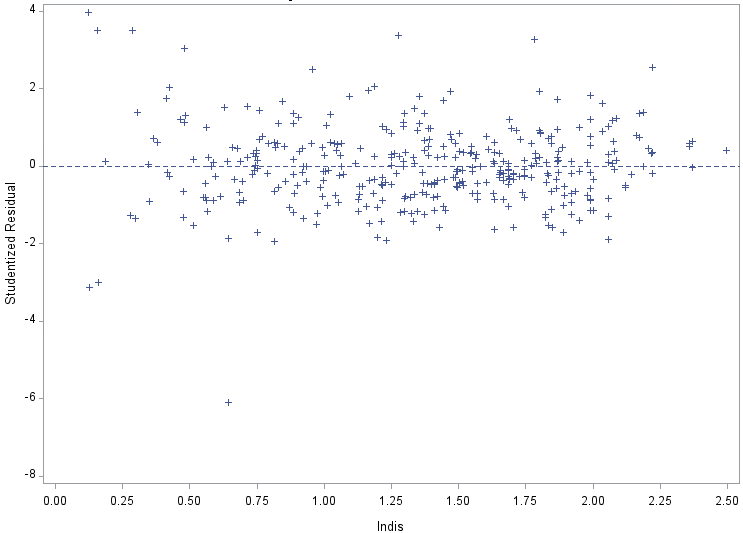


Figure 21 – Studentized Residuals vs LNLSTAT (Full Model after Transformations and Outlier Removal)

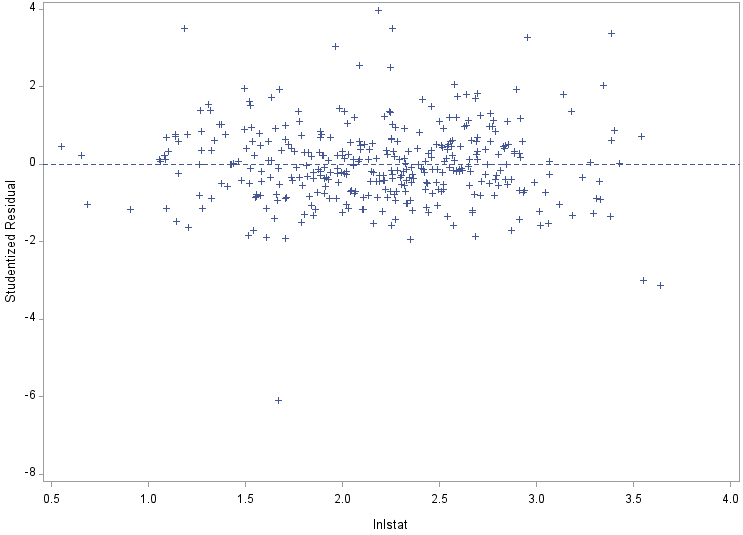


Figure 22 – Normal Probability Plot (Full Model after Transformations and Outlier Removal)

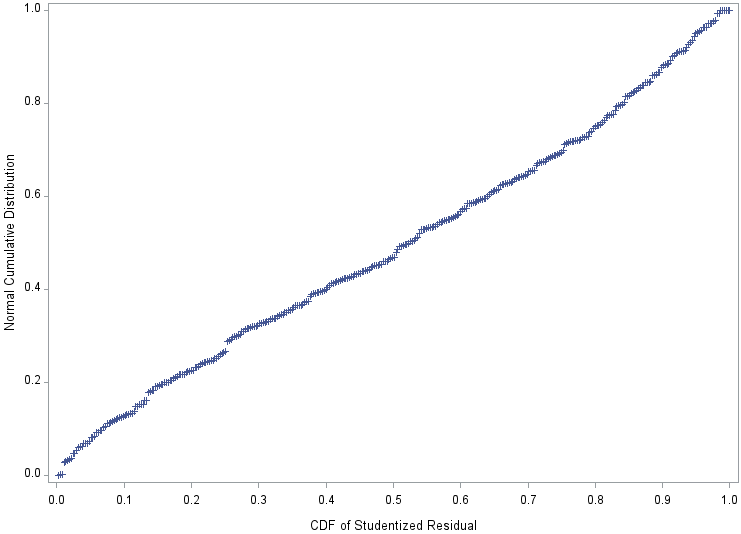
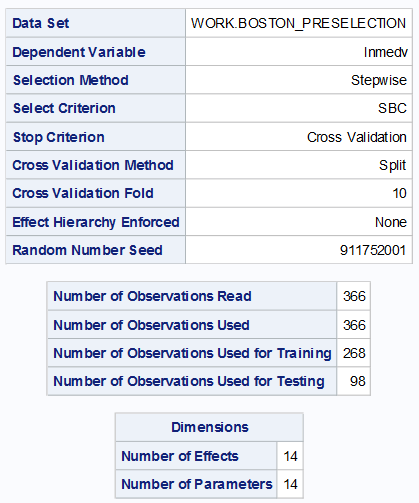
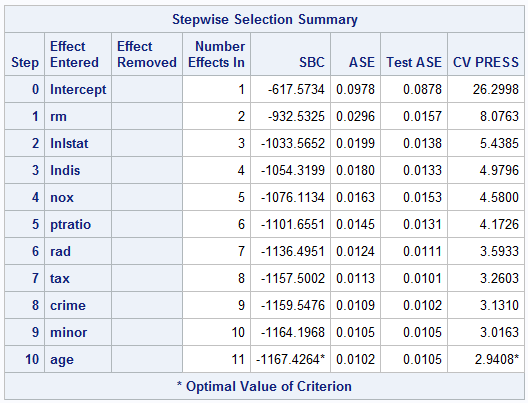
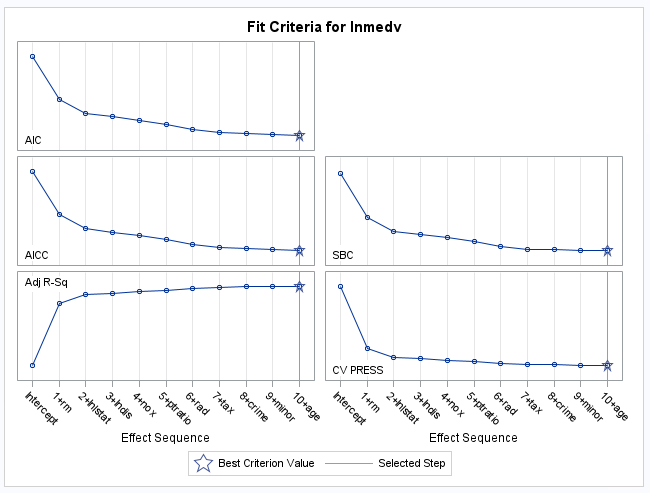
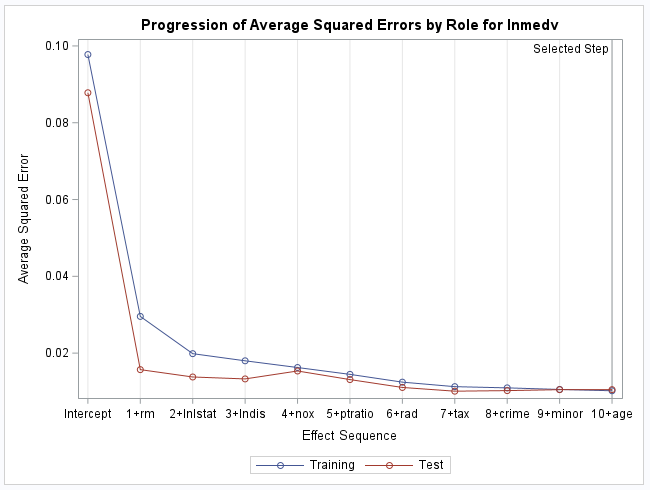


Figure 23 – PROC GLMSELECT RESULTS (Stepwise Method)





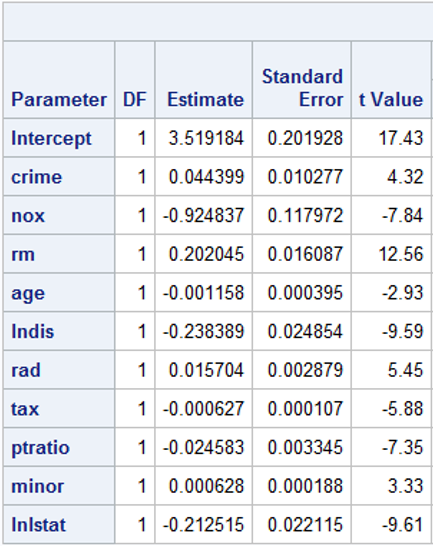
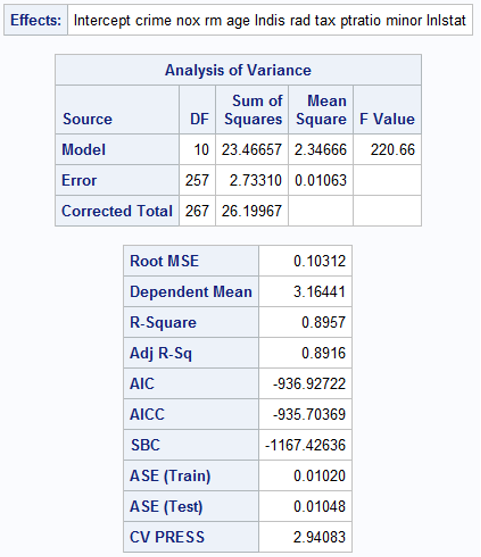
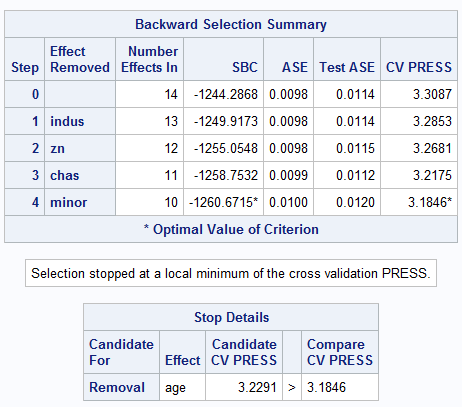
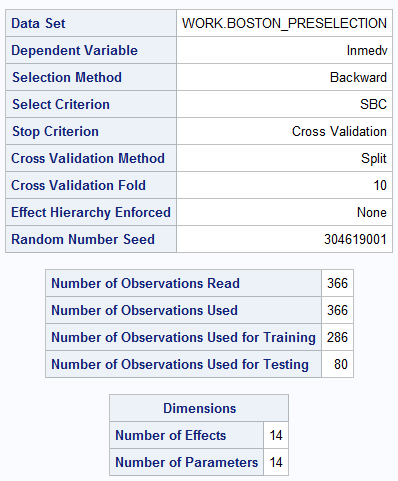
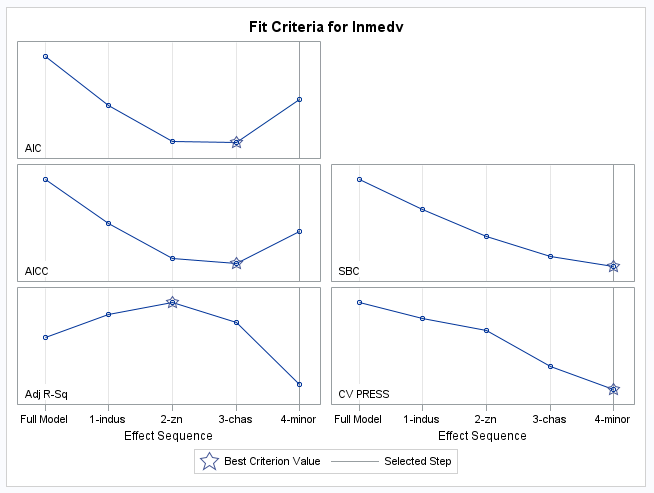
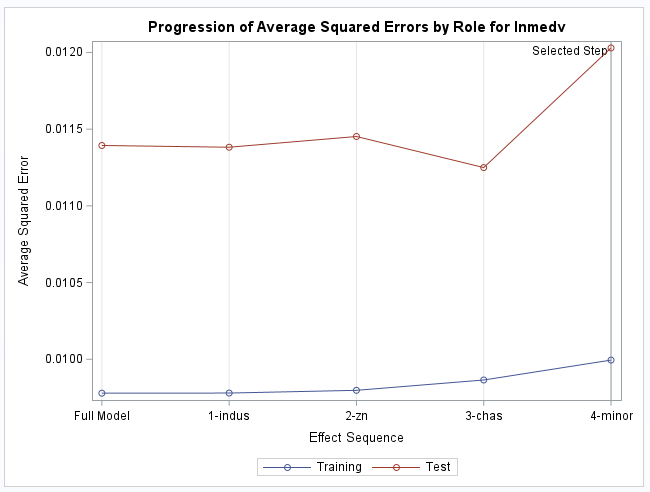
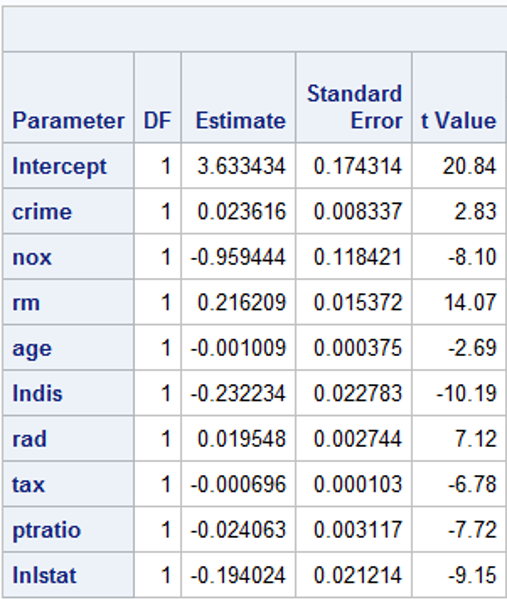
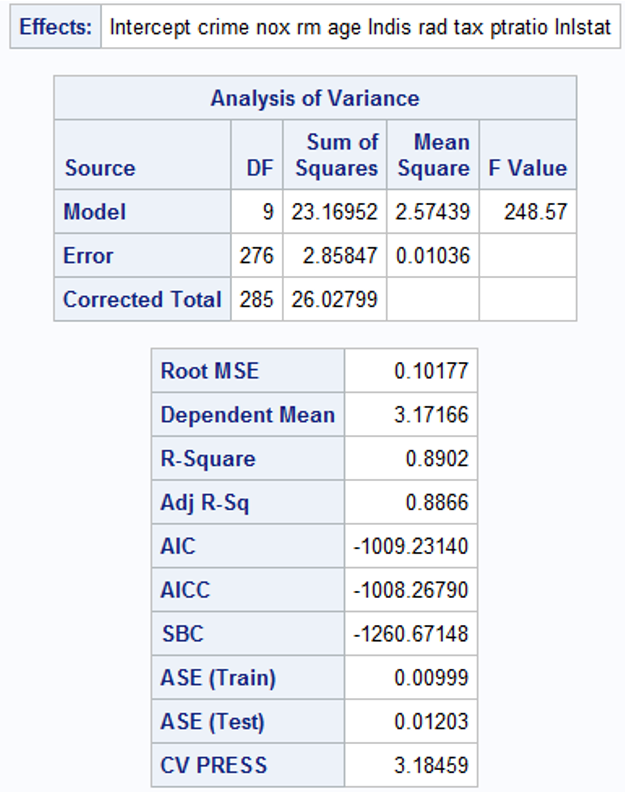


Figure 24 – PROC GLMSELECT RESULTS (Stepwise Method)









|  |  |  |
| --- | --- | --- |
| 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 |
| R² | 0.8957 | 0.8902 |
| Adjusted R² | 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

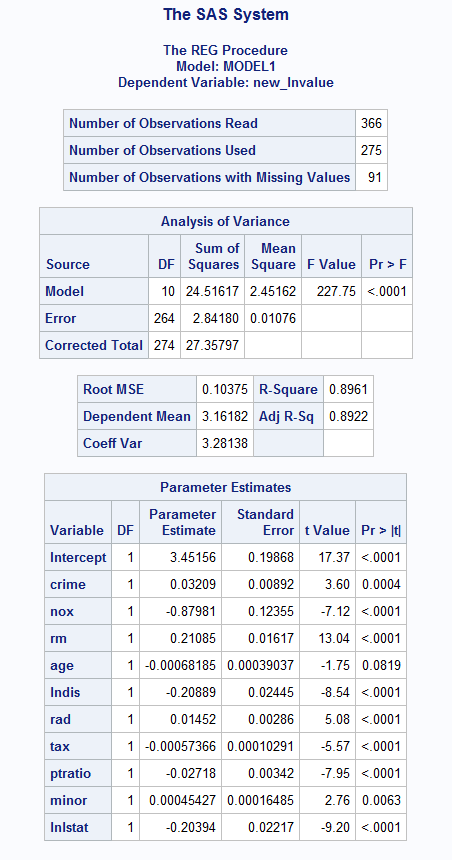


Figure 27 – Test Output of GLMSELECT Stepwise Model

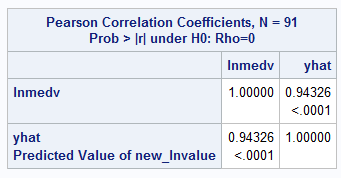
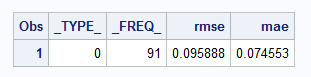


Figure 28 – Regression Output of GLMSELECT Backward Model

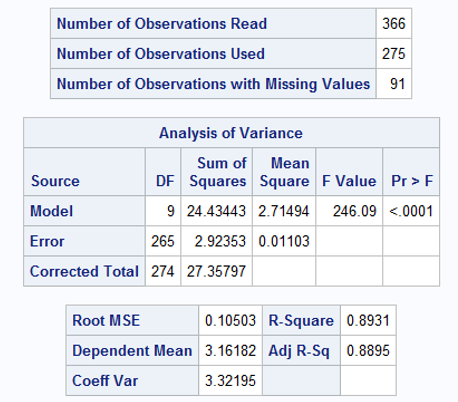
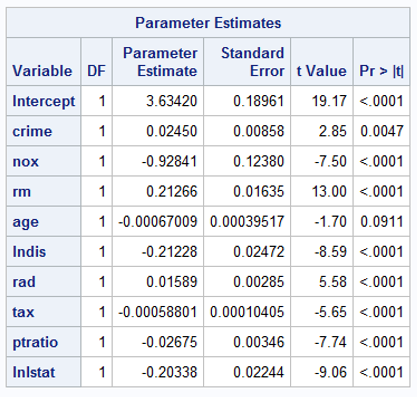
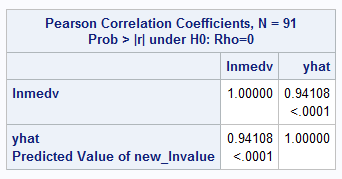
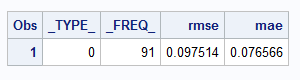
 

Figure 29 – Test Output of GLMSELECT Backward Model



|  |  |  |  |
| --- | --- | --- | --- |
| Figure 30 – Table of Relevant Training and Test Metrics | | | |
|  | | | |
| TRAIN | | |
| Metrics | GLMSELECT Stepwise (10 vars) | GLMSELECT Backward (9 vars) |
| RMSE | 0.10375 | 0.10503 |
| R² | 0.8961 | 0.8931 |
| Adjusted R² | 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 |
| R² | 0.88974 | 0.88563 |
| Adjusted R² (calculated) | 0.87596 | 0.87282 |
| Cross-Validated R² | 0.00636 | 0.00747 |

Figure 31 – Final Model Regression Output

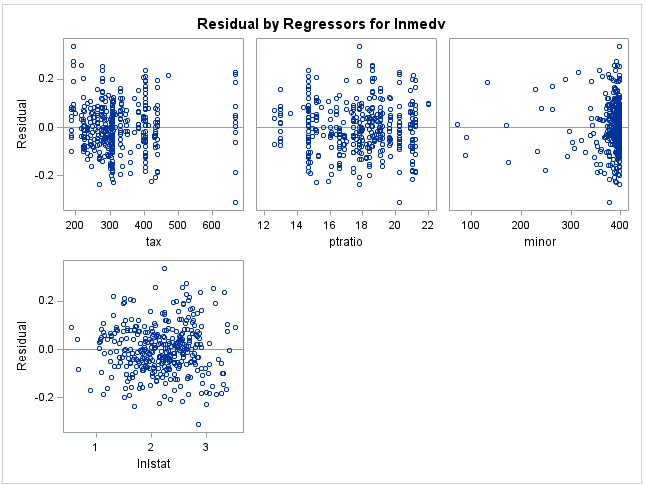
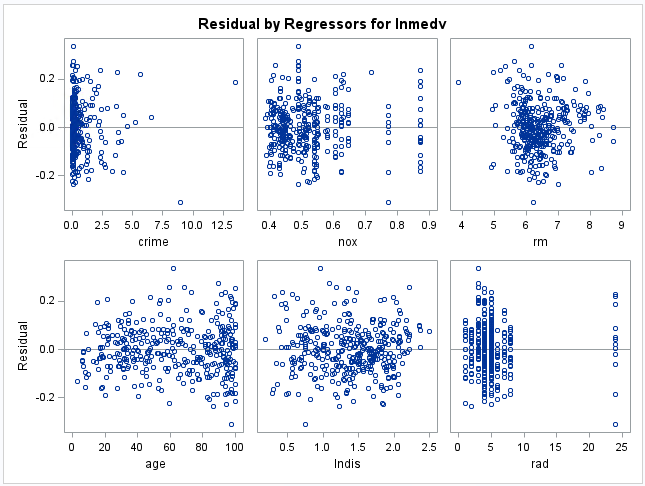
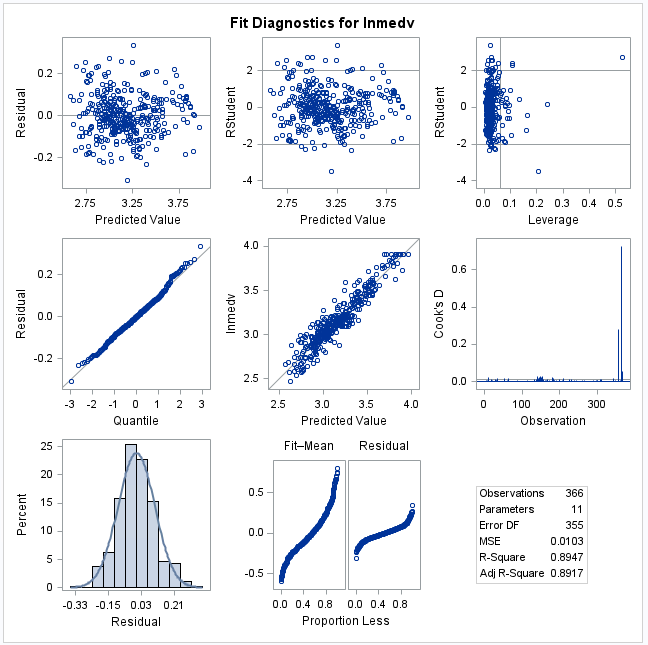
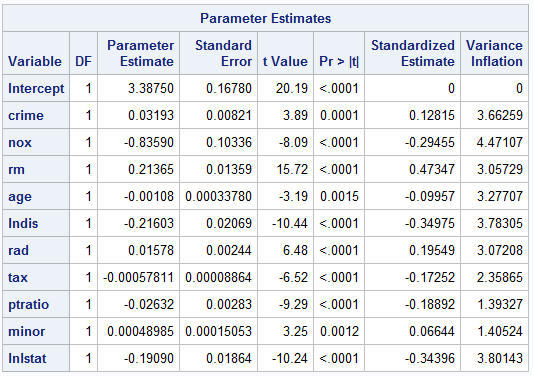
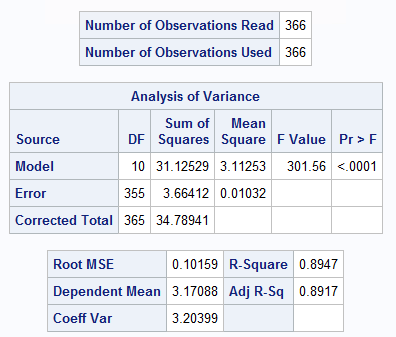


Figure 32 – Normal Probability Plot for Final Model

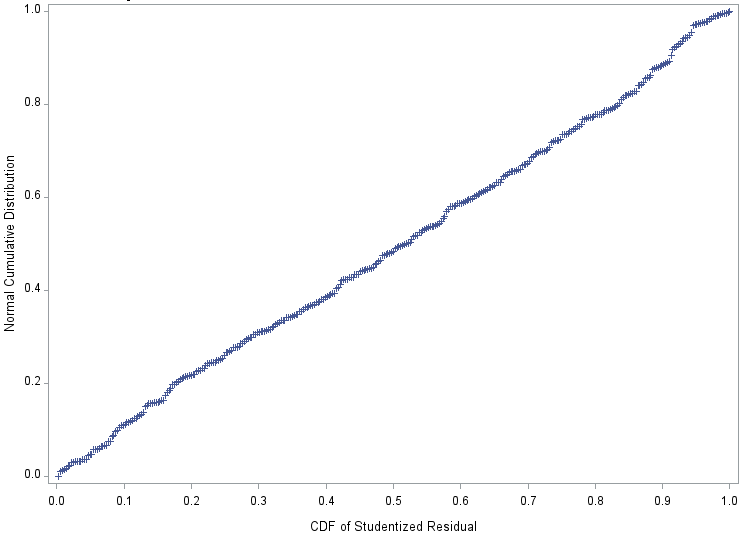
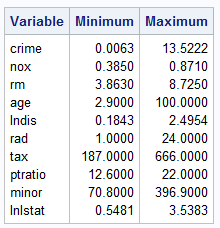
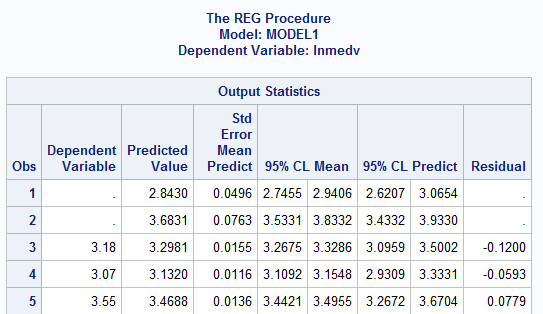


Figure 33 – Table of Minimums and Maximums for each Variable in Trimmed Dataset



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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



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