STAT 8210 – Applied Regression Analysis

Homework 6

Due April 21, 2020

Connor Armstrong

1. The response variable is medv, median home value in various neighborhoods. With the training data set, fit the model to predict/explain median home values. At this stage, use all of the predictors available. Check this model for adequacy. If anything needs to be fixed, fix it before proceeding. What needed to be fixed, and how did you fix it?

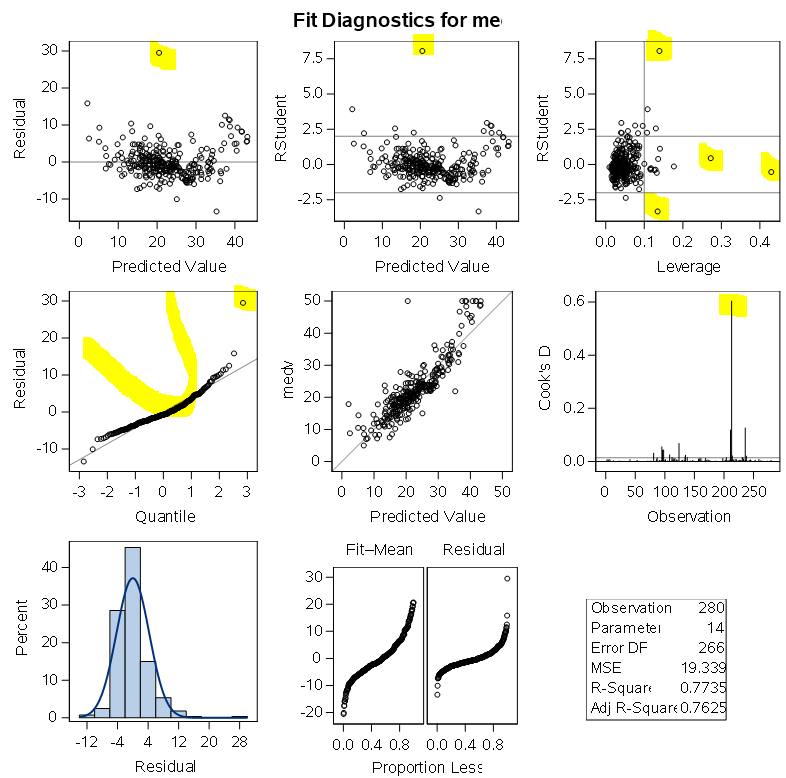
The model was fit using SAS. The following table shows the parameter estimates and other parameters of interest, including the variance inflation factor, which may indicate multicollinearity issues for VIF>10.

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | 1 | 34.14845 | 6.82178 | 5.01 | <.0001 | 0 |
| **crim** | 1 | -0.12409 | 0.05004 | -2.48 | 0.0138 | 2.03334 |
| **zn** | 1 | 0.03693 | 0.01703 | 2.17 | 0.0310 | 2.08727 |
| **indus** | 1 | 0.05196 | 0.07711 | 0.67 | 0.5010 | 3.73579 |
| **chas** | 1 | -0.07974 | 0.97717 | -0.08 | 0.9350 | 1.12416 |
| **nox** | 1 | -18.18536 | 4.77912 | -3.81 | 0.0002 | 4.65428 |
| **rm** | 1 | 4.73988 | 0.54624 | 8.68 | <.0001 | 2.16389 |
| **age** | 1 | -0.00236 | 0.01634 | -0.14 | 0.8851 | 2.97536 |
| **dis** | 1 | -1.45640 | 0.24151 | -6.03 | <.0001 | 3.60805 |
| **rad** | 1 | 0.32534 | 0.08953 | 3.63 | 0.0003 | 8.19887 |
| **tax** | 1 | -0.01747 | 0.00512 | -3.41 | 0.0008 | 9.83299 |
| **ptratio** | 1 | -1.06979 | 0.15907 | -6.73 | <.0001 | 1.80962 |
| **black** | 1 | 0.00767 | 0.00369 | 2.08 | 0.0387 | 1.39983 |
| **lstat** | 1 | -0.40931 | 0.06565 | -6.24 | <.0001 | 2.98380 |

The following table for the full model indicates a root mean square error of 4.398 and adjusted R-squared of 0.76. These values will be compared to other models to determine model adequacy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 4.39761 | **R-Square** | 0.7735 |
| **Dependent Mean** | 22.56071 | **Adj R-Sq** | 0.7625 |
| **Coeff Var** | 19.49233 |  |  |

Residual analysis for the full model shows some regions for concern. Presence of outliers, leverage points, and a “U” shape in the Q-Q probability plot indicate that a transformation might improve the model. The histogram shows a right skew.



Analysis of the regressor residual plots demonstrate a lack of homogeneity of variance for variables crim, zn, and dis. Variance stabilizing transformations such as logarithm, square root, or reciprocal transformations can correct this assumption violation. There is a clear “U” shape in the residual plot for the variables rm and lstat. A normality test for the residuals will determine whether a transformation is necessary to proceed.

The Anderson-Darling normality test returns a p-value of less than 0.005, which indicates that the residuals are not normally distributed, and that a transformation is likely necessary to proceed.

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.898324 | **Pr < W** | <0.0001 |
| **Kolmogorov-Smirnov** | **D** | 0.107792 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 0.966513 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 5.358209 | **Pr > A-Sq** | <0.0050 |

Studentized residuals are effective for determining influential observations (undesirable). The observations with the 10 largest (absolute value) studentized residual are displayed in the table below. These observations will be removed from the dataset to determine whether/how this will impact the regression model.

| **Obs** | **medv** | **IDnum** | **predict** | **resid** | **rstudent** | **dffits** | **studabs** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 50 | 369 | 20.4965 | 29.5035 | 8.05194 | 3.23731 | 8.05194 |
| **2** | 17.9 | 413 | 2.0526 | 15.8474 | 3.91834 | 1.36728 | 3.91834 |
| **3** | 21.9 | 365 | 35.2857 | -13.3857 | -3.33340 | -1.31396 | 3.33340 |
| **4** | 50 | 187 | 37.4785 | 12.5215 | 2.94844 | 0.60521 | 2.94844 |
| **5** | 50 | 167 | 38.4684 | 11.5316 | 2.76181 | 0.79239 | 2.76181 |
| **6** | 50 | 163 | 38.6957 | 11.3043 | 2.73606 | 0.89061 | 2.73606 |
| **7** | 23.7 | 215 | 13.3196 | 10.3804 | 2.55674 | 0.98787 | 2.55674 |
| **8** | 15 | 376 | 25.0749 | -10.0749 | -2.37319 | -0.55492 | 2.37319 |
| **9** | 46.7 | 229 | 36.6593 | 10.0407 | 2.37222 | 0.58595 | 2.37222 |
| **10** | 27.9 | 408 | 18.2415 | 9.6585 | 2.25828 | 0.45825 | 2.25828 |

After the observations were removed, the model improved substantially. The normality condition is still not satisfied.

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 13 | 16191 | 1245.44460 | 115.55 | <.0001 |
| **Error** | 256 | 2759.37462 | 10.77881 |  |  |
| **Corrected Total** | 269 | 18950 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 3.28311 | **R-Square** | 0.8544 |
| **Dependent Mean** | 22.08852 | **Adj R-Sq** | 0.8470 |
| **Coeff Var** | 14.86342 |  |  |

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | 1 | 9.69843 | 5.58948 | 1.74 | 0.0839 | 0 |
| **crim** | 1 | -0.08616 | 0.03790 | -2.27 | 0.0238 | 2.01048 |
| **zn** | 1 | 0.03208 | 0.01290 | 2.49 | 0.0136 | 2.13195 |
| **indus** | 1 | 0.01100 | 0.05877 | 0.19 | 0.8516 | 3.73469 |
| **chas** | 1 | 0.72680 | 0.75673 | 0.96 | 0.3377 | 1.11783 |
| **nox** | 1 | -7.90087 | 3.69377 | -2.14 | 0.0334 | 4.89561 |
| **rm** | 1 | 6.40059 | 0.46586 | 13.74 | <.0001 | 2.37199 |
| **age** | 1 | -0.02246 | 0.01344 | -1.67 | 0.0959 | 3.41166 |
| **dis** | 1 | -1.02373 | 0.18394 | -5.57 | <.0001 | 3.64807 |
| **rad** | 1 | 0.23327 | 0.06793 | 3.43 | 0.0007 | 7.95375 |
| **tax** | 1 | -0.01592 | 0.00386 | -4.13 | <.0001 | 9.41614 |
| **ptratio** | 1 | -0.80590 | 0.12142 | -6.64 | <.0001 | 1.83015 |
| **black** | 1 | 0.01203 | 0.00284 | 4.23 | <.0001 | 1.40380 |
| **lstat** | 1 | -0.26521 | 0.05538 | -4.79 | <.0001 | 3.45491 |

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.974232 | **Pr < W** | <0.0001 |
| **Kolmogorov-Smirnov** | **D** | 0.083476 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 0.373252 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 2.096136 | **Pr > A-Sq** | <0.0050 |

**Still failed AD test**



1. Use forward, backward, and stepwise selection to select a model.

**Forward Selection:** Sle you used: \_\_\_0.2\_\_\_\_

In the table below, put an X in the “Forward selection” row for each predictor in this final model. (Ignore the last column of the table until question 5.)

\*perform forward selection with sle=0.2;

title 'Forward selection (sle=0.2)';

**proc** **reg** data = delset;

model medv = crim zn indus chas nox rm age dis rad tax ptratio black lstat / selection=forward sle=**0.2** vif;

**run**;

**Backward Selection:** Sls you used:\_\_\_0.15\_\_\_\_

Put an X in the “Backward” row for each predictor in this final model

\*perform backward selection with sls=0.15;

title 'Backward selection (sls=0.15)';

**proc** **reg** data = delset;

model medv = crim zn indus chas nox rm age dis rad tax ptratio black lstat / selection=backward sls=**0.15** vif;

**run**;

**Stepwise Selection:** SLE you used:\_\_\_0.15\_\_\_ SLS you used: \_\_\_0.15\_\_\_

Put an X in the “Stepwise” row for each predictor in this final model

\*perform stepwise selection with sle=0.15 and sls=0.15;

ods rtf;

ods graphics on;

title 'Stepwise selection (sls=0.15, sle=0.15)';

**proc** **reg** data = delset;

model medv = crim zn indus chas nox rm age dis rad tax ptratio black lstat / selection=stepwise sls=**0.15** sle=**0.15** vif;

**run**;

Do all three methods yield the same model? \_Yes\_. If not, do at least two of them agree? \_\_N/A\_\_.

1. Use all possible regressions to find the three models with the best adjusted R2, the three models with the best AIC values, and the three models with the best BIC values. Put an X in the appropriate rows for each predictor left in the models.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Crim | zn | indus | chas | nox | rm | age | dis | rad | tax | ptratio | black | lstat | Model # | Criterion you’ll use to pick best model:  Adjusted R2  Value of this criterion when model applied to test data |
| Forward Selection | X | X |  |  | X | X | X | X | X | X | X | X | X | 1 | 0.6983 |
| Backward | X | X |  |  | X | X | X | X | X | X | X | X | X | 1 | 0.6983 |
| Stepwise | X | X |  |  | X | X | X | X | X | X | X | X | X | 1 | 0.6983 |
| Best Adj R2 | X | X |  |  | X | X | X | X | X | X | X | X | X | 1 | 0.6983 |
| 2nd Best Adj R2 | X | X |  | X | X | X | X | X | X | X | X | X | X | 2 | 0.7269 |
| 3rd Best Adj R2 | X | X | X |  | X | X | X | X | X | X | X | X | X | 3 | 0.6969 |
| Best AIC | X | X |  |  | X | X | X | X | X | X | X | X | X | 1 | 0.6983 |
| 2nd Best AIC | X | X |  |  | X | X |  | X | X | X | X | X | X | 4 | 0.6996 |
| 3rd Best AIC | X | X |  | X | X | X | X | X | X | X | X | X | X | 2 | 0.7269 |
| Best BIC | X | X |  |  | X | X | X | X | X | X | X | X | X | 1 | 0.6983 |
| 2nd Best BIC | X | X |  |  | X | X |  | X | X | X | X | X | X | 4 | 0.6996 |
| 3rd Best BIC | X | X |  | X | X | X | X | X | X | X | X | X | X | 2 | 0.7269 |

1. Pick which criterion you will use to select the best model when you fit these models to the test data set (for example, adjusted R2.) There are many criteria you might consider for the best model, but for time’s sake pick **one criterion** in advance that you will be using. Write that into the last column of the table.

Fit all models in the table to the test data set, and for each record the value of the selection criterion you are using in the last column. (Note: many of your models may be the same, so you should not need 12 Proc Regs here!) Which model seems to be the best based on this criterion?

The model with the highest R2 adjusted for the above variable selection combinations fitted to the test dataset is Model 2, which contains all of the variables of the full model except for “indus”

1. Briefly discuss the following:
   1. What are the advantages of these automatic variable selection methods?

Proper usage of statistical software to determine which combination of variables yields the best desired characteristic for a regression model is very useful, especially whenever the number of variables is large and making those comparisons by hand would take a prohibitively long time.

* 1. What are the disadvantages?

Blindly allowing a computer to select the “best” model might cause an analyst to miss an important piece of information which might have otherwise been found had a more thorough residual analysis been performed throughout the model selection process.

* 1. How do you feel about using automatic selection methods as opposed to a more hands-on approach?

It is a powerful tool but not a replacement for thorough analysis and assumption validation. Whenever using a regression model for a real-life purpose, there are real implications and consequences to removing or adding variables to a predictive model. Variable selection decisions must be made with more reasoning than “the software said this model has the best adjusted R2,” for example.

* 1. Which approach (automatic selection or hands-on) do you think you are likely to use more in your statistical career? Why?

It would depend on the amount of time I was allowed to spend on performing thorough residual analysis for the models. I think that automatic variable selection will be a good place to start to find out which models deserve a more “hands-on” analysis.