|  |
| --- |
| Math 6388 |
| Homework #2 – Chapter 3 |
| Linear Regression Model using Forward, Backward, Step-Wise and All Subsets Selection Methods |
|  |
| **Akbar A Esfahani** |
| **2/14/2011** |

|  |
| --- |
|  |

# The data

The data is from a city assessor that was interested in predicting home sale prices as a function of various characteristics of the home and the surrounding property.

The data consists of sale price, finished square feet, # of bedrooms, # of bathrooms, AC, garage size (# of cars fitted), pool, year built, quality (1=high q, 2=med q, 3=low q), style, lot size, and adjacency to highway (1=yes, 0=no).

# Backward selection method

The final model:

The backward selection method produced the following model:

Price = Area + # of Bedrooms + Age + Quality + Style + Lot Size, at an alpha value of 0.01.

The model has a mean prediction interval of (152237, 405635), which makes the test set mean of price fall right in, at 280188.80.

The model has a mean confidence interval of (264871, 293001), which makes the mean of the test data set fall right in, at 276549.50.

The test model has a mean error of 0.008587438, that is; the test model performs at about 99.90% with the selected variables.

# Forward Selection Method

The Final Model:

Again, the final model is: Price = Area + # of Bedrooms + Age + Quality + Style + Lot Size.

So, again, the prediction and confidence intervals will be the same as above, as well as the mean error of the model.

# Best Subset Selection Method

The Final Model:

With the best subset method, since we have to rely on the R^2 values, the model becomes a bit tricky. According to the R^2 values, choosing Price = Area + Age + Quality + Style + Lot Size, will have the same effect as choosing Price = Area + # of Bedrooms + Age + Quality + Style + Lot Size. To verify this assumption, I tested the former to compare with two models above. The error rate for the first model does indeed do better than the model used above, when comparing the test set error rates. While the model with six variables has an error rate of 0.0086, the model with 5 variables has an error rate of 0.007. So one can see why the best subset method would give equal R^2 values for both models, but when we consider the adjustment needed with degrees of freedom, we come to the conclusion that the model with five variables is the better model.

# Step-wise Subset Selection Method

The Final Model:

With the Step-wise method, I used the build in R function “Step ()” to make my selections. Using the AIC as selection criterion, the final model does differ from the previous two models substantially. The model suggested by Step is: Price = Area + # of Bedrooms + AC + Garage Size + Age + Quality + Style + Lot Size + Highway Adjacency. While the model did not consider an alpha level, it is worthwhile to note that the three new variables introduced here are above the .05 level.

# Conclusion

As far as how the models stack up against each other, the following table is the result from the error analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Model 1 | Model 2 | Model 3 |
|  | Model error rates | 0.008587438 | 0.006952181 | 0.008344884 |
| Model 1 | 0.008587438 | 0 | - 0.001635257 | - 0.00242598 |
| Model 2 | 0.006952181 | 0.001635257 | 0 | 0.001392703 |
| Model 3 | 0.008344884 | 0.000242598 | - 0.001392703 | 0 |

Model 1 is forward/backward selection, Model 2 is Best Subset selection, and Model 3 is Step-wise selection.

As we can see from the above error analysis, the best subset method performed better than the other two models, while the Step-wise model performed better than the forward/backward model.

The testing procedures showed that all three models performed very well and the errors were well within a margin that we can reasonably choose either of the models. Never the less, I would recommend the Best subset selection model, as it is the most Parsimonious model from the three. 9 All three models used the same training and test set and thus all three output parameters are comparable. The final model is:

# Things for Further Considerations

To come to a conclusion about the model that I selected, I believe the model should be tested further. To that affect, I ran my model in a Monte Carlo simulation with 10000 iterations. The summary output of the simulation confirmed the model selected. All the variables were within less than one standard deviation from their mean, thus the model is stable and ready for use within the framework of its requirement.