*Introduction*

The objective of this assignment is to perform automated variable selection techniques for identifying the “best” regression model for predicting sale price for homes in the Ames, Iowa area. The first phase includes the assessment of which predictor variables, based on common sense and business justification, made sense to include in a predictive model. After conducting some preliminary exploratory data analysis (EDA), it was concluded that the following predictor variable candidates would be considered in the model:

X1-GrLivArea

X2-LotArea

X3-AgeAtSale

X4-TotalBsmtSF

X5-total\_baths\_calc

X6-TotRmsAbvGrd

X7-highend\_ind Neighborhood grouping

X8-midend\_ind Neighborhood grouping

X9-good\_heating

X10-excl\_kitchen kitchen quality

X11-central\_air

X12-fireplace\_ind

X13-garage\_ind

X14-good\_basement\_ind

X15-concr\_foundation

X16-quality\_index

X17-brick\_exterior

X18-lot\_frontage

X19-new\_bldg

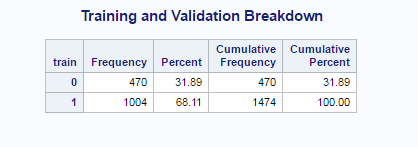
X20-old\_bldg

X21-pos\_cond

X22-recent\_remodel

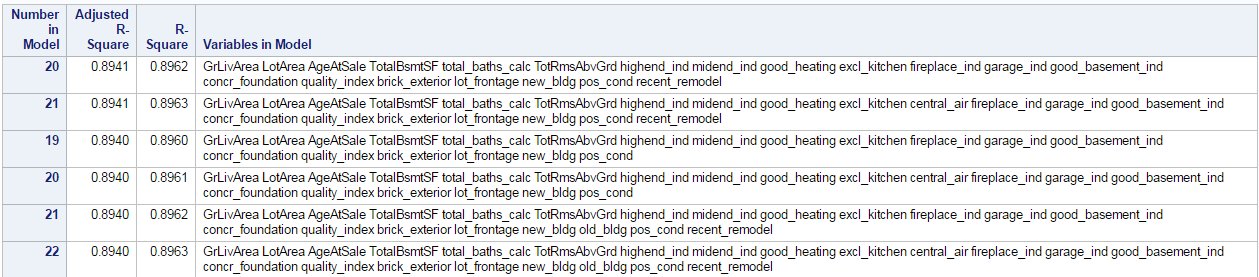
*Training data set*

The data set has been split in test/train data set as shown below. We will test the model accuracy by training the model on the 70% of the data set and validating its accuracy on remaining 30%.



*Adjusted R-Squared Model (Model\_AdjR2)*

The results of the adjusted R-squared variable selection method determined that not all predictor variables candidates (X1-X22) should be included in the final model. The adjusted R-squared of the final model was 0.8941. Table 1 below, you can see the summary output of the variable selection process with the final result in the first row.



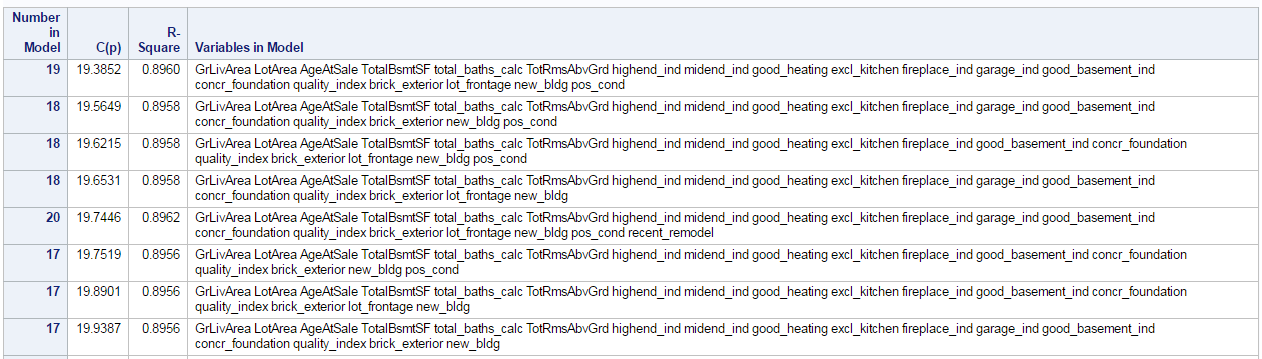
The adjusted R-squared model suggest to exclude variables central\_air(X11) and old\_bldg(X20).

*Maximum R-Squared Model (Model\_MaxR)*

Like the adjusted R-squared variable selection method, the results of the maximum R-squared variable selection method determined that not all predictor variable candidates (X1-X22) should be included in the final model. The R-squared of the final model was 0.8963. In Table below, we can see the ANOVA and parameter estimates for the final suggested model from the maximum R-squared variable selection method. The model suggests excluding variables central\_air(X11) and old\_bldg(X20) as none of them is statistically significant.



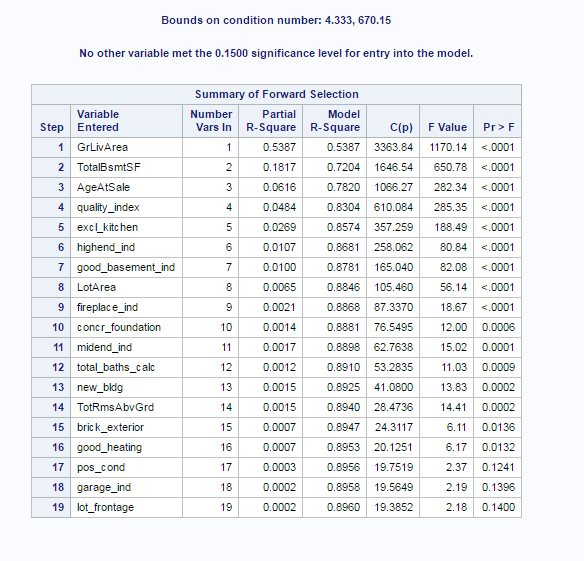
*Mallow’s Cp Model (Model\_MCp)*

As with the previous two models, the results of the Mallow’s Cp variable selection method also determined that not all predictor variable candidates (X1-X22) should be included in the final model. The Mallow’s Cp of the final model was 19.3852. In Table 3 below, we can see some of the output from the Mallow’s Cp variable selection process. The best model is shown in the first row of the table. 

Apart from old\_bldg(X20) and central\_air(X11) the model suggest to exclude recent\_remodel(X22) variable.

*Forward Selection Model (Model\_F)*

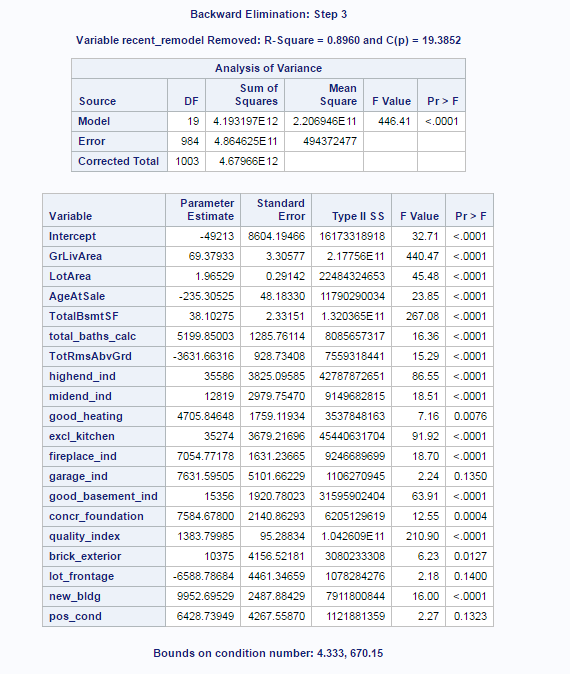
Like proceeding variable selection methods, the results of the forward variable selection method also determined that not all predictor variable candidates (X1-X22) should be included in the final model. In table below, we can see the summary of the forward selection method and you’ll notice that the p-value from the nested F-tests did not increase until the 10th variable was entered into the model. For this method, we chose a *slentry* value of 0.15 as our threshold for variables to be allowed to enter into the model.

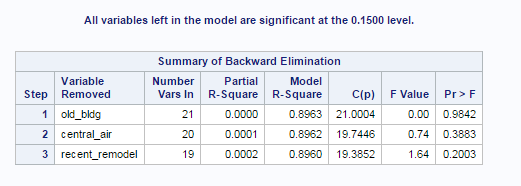


The model suggest exclusion of central\_air(X11), old\_bldg(X20), recent\_remodel(X22) variables

*Backward Model (Model\_B)*

Also this time the results of the backward variable selection method determined that not all predictor variable candidates (X1-X22) should be included in the final model. In table below, we can see the summary of the beckward elimination method. For this method, we chose a *slentry* value of 0.15 as our threshold for variables to be allowed to enter into the model.

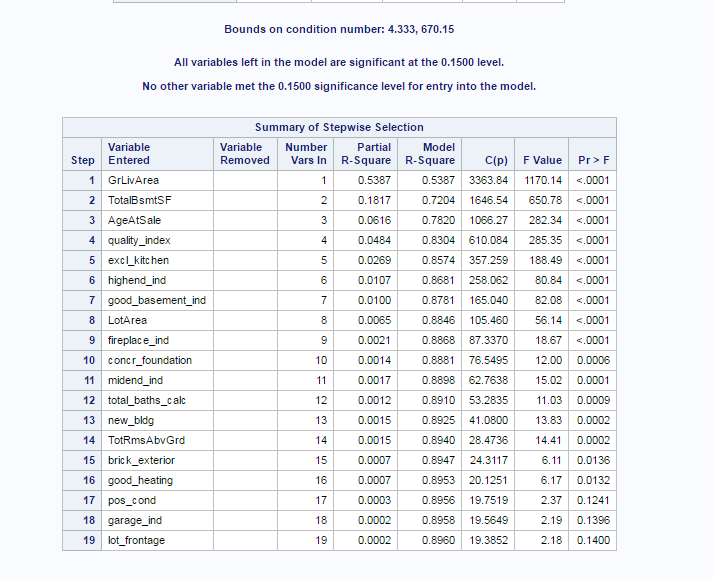




The model suggest exclusion of central\_air(X11), old\_bldg(X20), recent\_remodel(X22) variables

*Stepwise Selection Model (Model\_S)*

The stepwise selection method was the final option used for model selection. As with previous methods, the stepwise selection method indicated that not all predictor variable candidates should remain in the model. The stepwise variable selection summary is shown below.



At the 0.1500 level variable X20-X22 were excluded from the model.

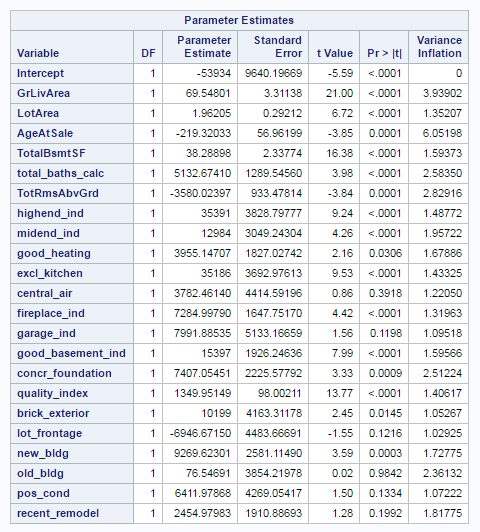
*Model Comparison*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Model\_AdjR2 | Model\_MaxR | Model\_MCp | Model\_F | Model\_B | Model\_S |
|  | Predictors  excluded | X11,X20 | - | X11,X20,X22 | X11,X20,X22 | X11,X20,X22 | X11,X20,X22 |
|  | Adjusted R2 | 0.8941 | 0.8940 | 0.8940 | 0.8940 | 0.8940 | 0.8940 |
|  | AIC | 20118.9967 | 20122.23 | 20118.67 | 20118.67 | 20118.67 | 20118.67 |
|  | BIC | 20121.9478 | 20125.31 | 20121.51 | 20121.51 | 20121.51 | 20121.51 |
|  | Mallow’s Cp | 19.7446 | 23 | 19.3852 | 19.3852 | 19.3852 | 19.3852 |
|  | MAE | 15942671.02 | 15929924.83 | 15971168.33 | 15971168.33 | 15971168.33 | 15971168.33 |
|  | MSE | 485650958256 | 485282607503 | 486462517503 | 486462517503 | 486462517503 | 486462517503 |
|  | MAE | 7074841.93 | 7057884.68 | 7100654.29 | 7100654.29 | 7100654.29 | 7100654.29 |
|  | MSE | 200370436990 | 199633617931 | 202198438096 | 202198438096 | 202198438096 | 202198438096 |

There has been slight variation in variable selection between models . Table below shows model fit criteria from the models created using the training sample. Based on the results in Table 6, we can see that the predictive ability of the final model built from the training sample. Based on the results, we can see that the predictive ability of the final model built from the training sample performed very well with the test sample data.

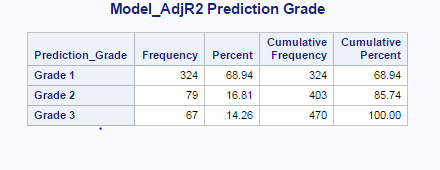
*Multicollinearity*

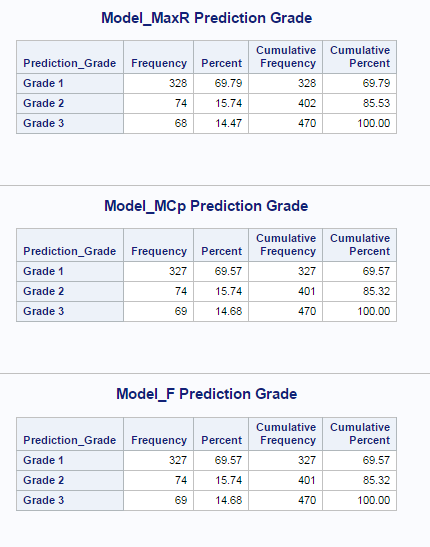
Values 10 or above for variance inflation analysis may cause serious problem for data analysis. It seems that there are not at immediate risk for multicollinearity even though the value for AgeAtSale is a little bit high.

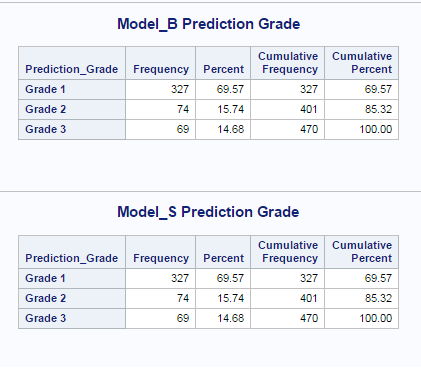


*Operational validation*

To assess the operational accuracy of the final model, we placed the predictive scores, absolute value for each observation’s actual vs predicted value for the response variable SalePrice, into three categories: Grade 1 (within 10% of the observed value), Grade 2 (between 10-15% of the observed value), Grade 3 (everything else). Below the results for each one of the models.







Based on the results of the prediction scores above, we can see that 85% of the predicted values were within 15% of the observed value.

*Conclusion*

In order to see effects of the automated variable selection I had to choose quite a few variables. The model might be hard to interpret for the business users, although it performs quite well. It has been proven that simpler models repeatedly run on larger data sets perform better in the long run.

It is quite clear how automated variable selection can be valuable when designing a predictive model. That should be combined with domain knowledge of stakeholders in order to achieve the best results.