Assignment # 7

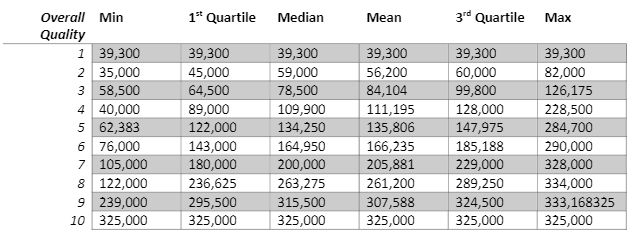
May 2021

1)

After reviewing the variables in the Ames Dataset, I will explore the additional variables of overall quality, overall condition, and land slope.

Overall Quality:

This variable was viewed in previous assignments and tested high in correlation to sales price, it is also a seeming natural connection to sales price and will be tested further.



Avg. mean = 31.744.44

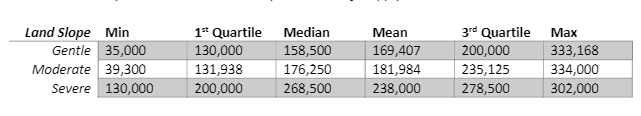
Overall Condition:

Overall condition sems very similar to overall quality and would be a great variable to test against sales price. Multicollinearity of overall condition and overall quality is something to keep in mind though when viewing results.

Average mean = 28,508

Land Slope:

Over the past assignments, a surprising variable that showed some affect on sales price was land scope. Here, we will explore it further and see if scale of the slope affects the sales price.



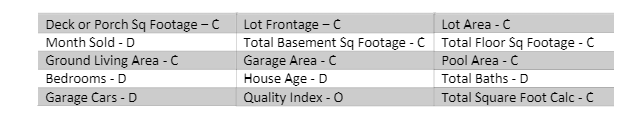
Avg. mean = 34,297

The three average means are all very close together in range with the minimum being overall condition at 28k and the max being land slope at 34k. Land Slope proved to be the highest mean and thus the variable to further study when comparing against sales price. The difference in means between each level supports the logic of dummy coding for each categorical variable as the changes are not equal and a constant value will be left out of the model.

2)

Train: 1213; test: 515; Total after sampling population: 1,728

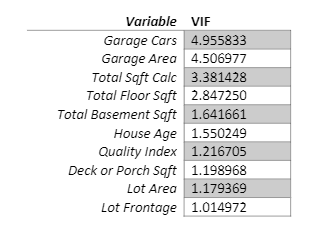
3)



I chose these 15 predictor variables based on our past assignments and correlation. The three model procedures I selected are forward, backward, and stepwise.

Forward:

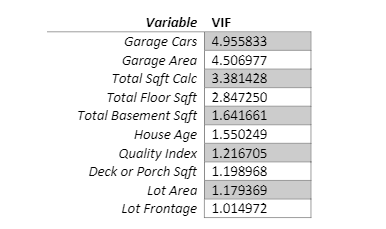
Predicted Sales Price = Total SqFt + Garage Cars + Quality index - House Age + Total Floor Sqft + Total basement sqft + lot Area + Deck or Porch Sqft + Garage Area + Lot Frontage



The vif values are very low and thus none require to be removed because of these low values.

Backward:

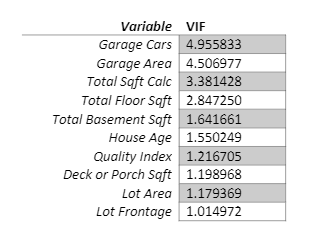
Predicted Sales Price = Total SqFt + Garage Cars + Quality index - House Age + Total Floor Sqft + Total basement sqft + lot Area + Deck or Porch Sqft + Garage Area + Lot Frontage



The vif values are very low and thus none require to be removed because of these low values.

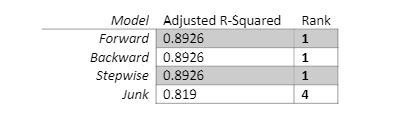
Stepwise:

Predicted Sales Price = Total SqFt + Garage Cars + Quality index - House Age + Total Floor Sqft + Total basement sqft + lot Area + Deck or Porch Sqft + Garage Area + Lot Frontage

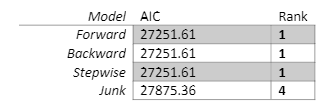


The vif values are very low and thus none require to be removed because of these low values.

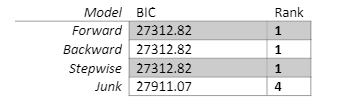
Adjusted r-squared:



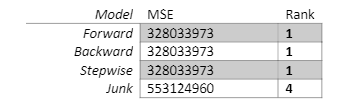
AIC:



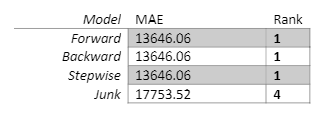
BIC:



MSE:



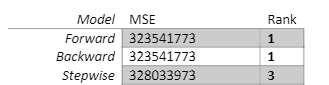
MAE:



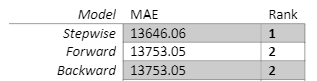
All three models produced the same or similar results, and all had the same model fit ranking of #1.

4)

MSE:



MAE:



The stepwise model has different MSE and MAE scores from the test sample but I hade worse mse scores than the forward and backward models, but did have better mae values. To decide between the two for a better-out-of-sample fit, it is important to note that better predictive accuracy in-sample would likely mean our model is more accurate with the current data, and if it has a higher out-of-sample accuracy. The forecast values will have a higher accuracy/ This is the reason we are creating the different models and suing comparative metrics to find the best fitting model.

5)

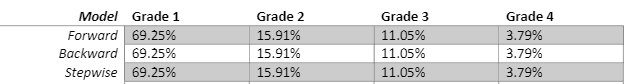
G1 = within 10% of actual value

G2 = within 15% of actual value and outside scope of G1

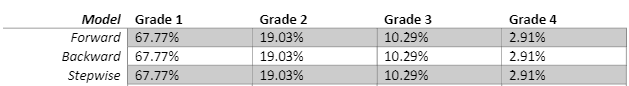
G3 = within 25% actual value and outside scope of G2

G4 = anything outside the scope of G1-3

In Sample:



Out-of-Sample:

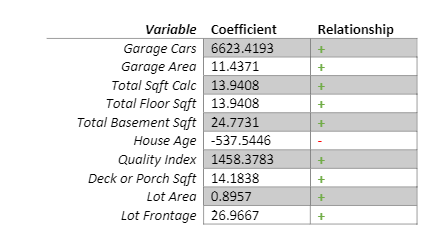


The in sample and out of sample results produced the same values for all three of the forward, backward, and stepwise models with the out of sample results closely resembling the training data results. Lastly, I noted that if the model is accurate within Grade 1 more than 50% of the time, each of our models are of good quality.

6)

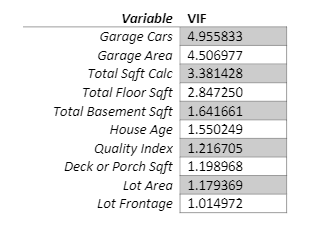
Step 4 showed that the best models are the forward and backward models with the highest rankings for fit and their mse values, so I will use the forward model moving forward to re-visit, clean up, and conduct residual diagnostics. We must further understand our models outputs, standardized residuals, and distributions.

Multicollinearity:



House age appears to be the only coefficient with a negative relationship to our target variable of sales price which is in line with how the housing market works.

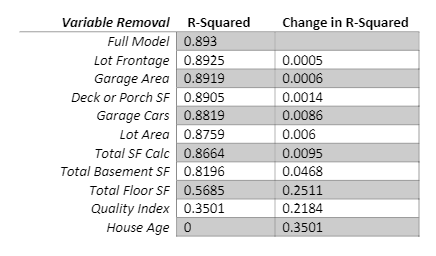
Vif:



The vif values don’t indicate multicollinearity high enough to leave any of the variables out of the final model.

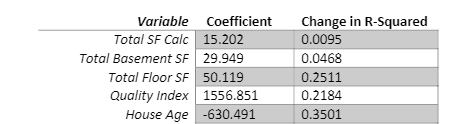
R-Squared values:

Due to large sample size we are going to remove variables one by one to analyze the changes in r-squared – keeping those that impact the predictive ability.



Lot frontage, garage area, deck or porch sf, garage cars, lot area all have values less than 1 percentage, we will remove these variables as the don’t contribute to the impact on predictive ability. Keeping these variables has a total r-squared value of .8759 with only .0171 less than the full model after removing the five variables.

Coefficients with change in r-squared:



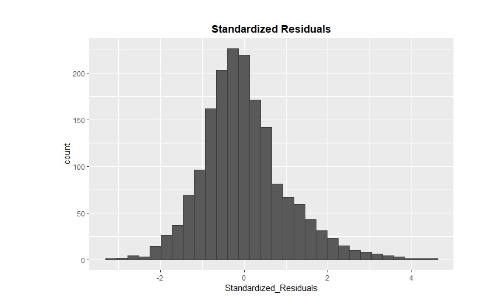
All of the variables had high r-squared change values and thus should be kept as they impact the predicted value.

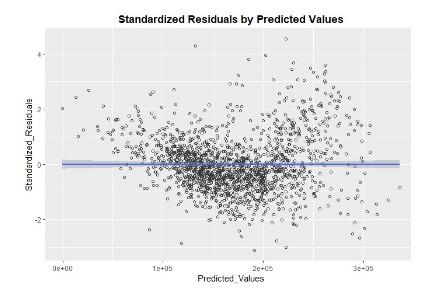
Dummy coded variables:

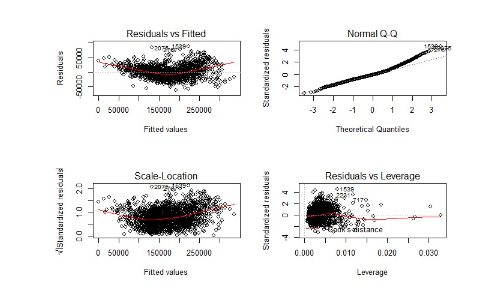
I did not include any dummy coded variables as they were not needed in the final model, however, if there had been they would have been included for the categorical variables.

Unequal Slopes:

Quality index is the only categorical model we have chosen for the final model and it is a combination of overall quality and condition of home. The interaction between this variable and the four others do not appear to be a logical possibility, so there is no need to test.







There are quite a few outliers in the theoretical quantiles and standardized residuals, but none of our plots show any indication that the assumptions of our model are false. The standardized residuals are normally distributed and there appears to be an equal distribution above and below the trend line in predicted values compared to actual values.

7)

I liked that we went back to the Ames data set for this assignment and have been using it to build on most of our assignments. It has been good to be able to delve deeper into the dataset and build upon our previous models and underlying assumptions about the data. In this assignment, I was concerned by the amount of outliers that are present and wonder if more fitting would be required to get a more accurate model. This would likely need more training data, and since we only have a snapshot int time for Ames, would probably require more years worth of data to be able to view time lapses or changes over time. Independent factors like economy and gdp would also have to be taken into consideration for the outcomes. The max fitting model and the simple model seem both to be good models but would need to be applied in the right situations. The max fitting model seems to be more powerful but the simpler model more intuitive, it would also depend on who the target audience is.