Lauren’s Models

The PCR and PLS models were helpful in determining how many components to include in the model. PCR validation plot shows that at about 19 components, the smallest cross-validation error occurs and then it levels out. The Kaggle score for this model was 0.4188. Partial Least Squares MSE decreased steadily from 1 to 10 components used, and then continued decrease slightly until about 20 where it plateaued. The Kaggle score for this model was 0.3301. Both models showed that approximately 20 variable models would be best. Forward and Backward selection gave the same two 4 variable models that included Lot Area, Overall Quality, Year Built, and Ground Floor Living Area. Overall Quality was the most significant variable. Since there was too many variables to generate usable graphs, I took the best 4 term model and added interaction and polynomial terms to boost the R squared value. I ended up with this model: LotArea+poly(OverallQual,5)+YearBuilt+GrLivArea\*LotArea+OverallCond. This model increased the R squared from 0.8219 to 0.8534. This is a relatively small increase in R squared but I am only dealing with 4 variables. There are residual vs. leverage plots that I think I may use for the presentation, but I need to do more work on them to make sure I am explaining them correctly. They’re on github. :) My models were all pretty consistent. Forward and backward gave the same 4 variable model and PCR and PLS gave similar results for number of components.