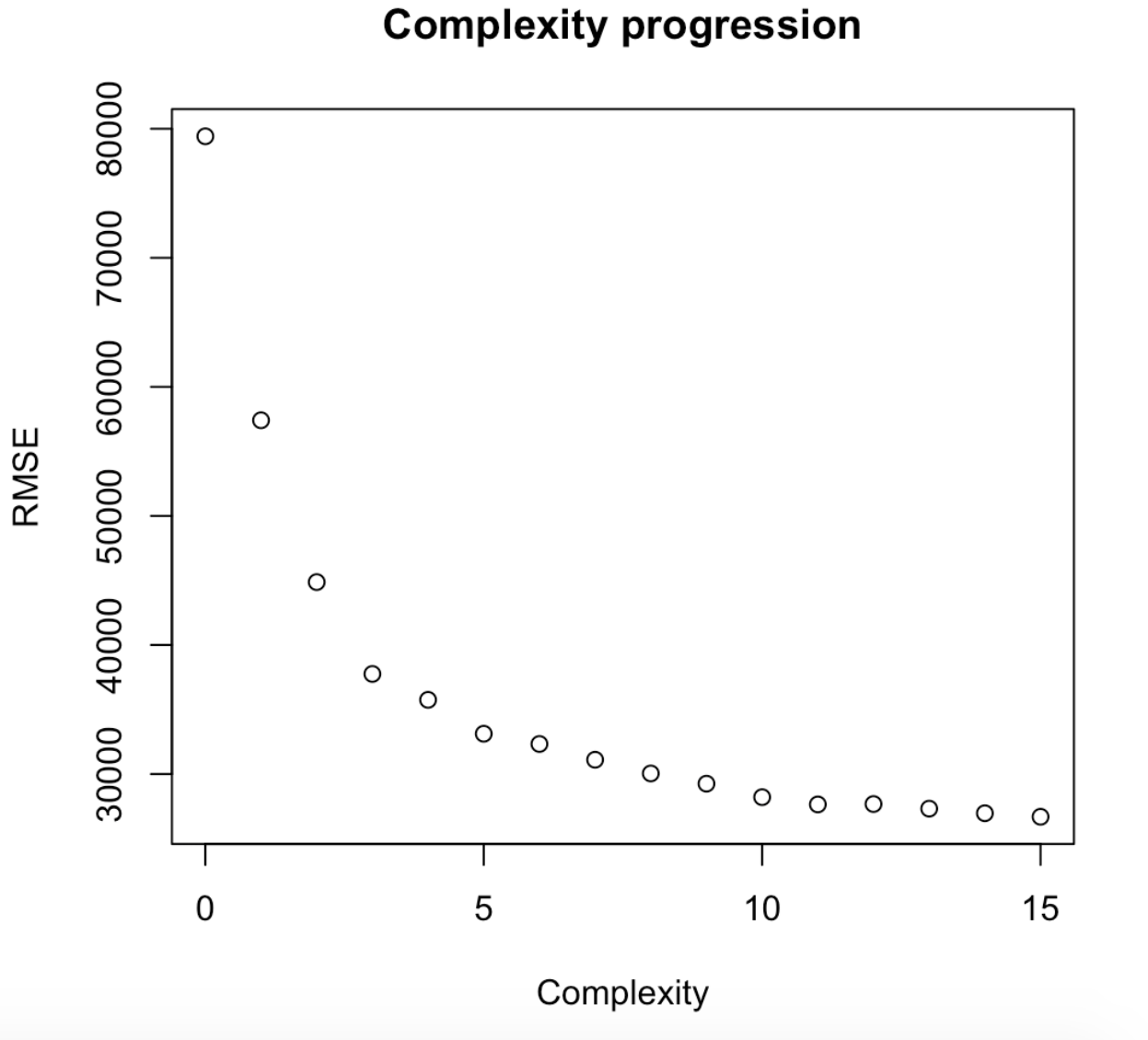
Lab 3

Part 1:

After producing models up to a complexity of 15, the following graph is obtained showing the relationship between complexity and resulting RMSE.



It can be seen that as complexity increases, RMSE decreases, indicating that the full size model is the most accurate model in terms of included variables when determining Sale Price of a house. This is based on the idea that regression gets more accurate the more variables that are included due to the diminishing value of the ommitted variable bias.

Part 2:

We created our model by utilizing forward selection. Essentially we started out with a regression using no explanatory variables and found the RMSE for both the training and testing data. Then we used the “step” function to determine which explanatory variable had the most predicitve power and add that variable into our regression model and calculated the RMSE for the new model. From there, we used the step function to find the variable with the second most explanatory power, added it to the regression, and calculated the RMSE. We repeated this cycle until the RMSE stopped decreasing significantly with each new variable and ultimately ended up incorporating 5 different variables that had a strong predicitve power over the sale price of a home. Those variables were ExterQual, GrLivArea, GarageCars, BsmtFullBath, and LandContour. Ultimately our final RMSE on the testing data was 38,003.59