Prediction of Housing Prices in King County, Washington

Ronny Malkowski1, Nicholas Nikolov1, Alexandria A. Pokorny1

1 Department of Mathematics, Illinois State University, Normal, IL 61761

rmalkow@ilstu.edu, ndnikol@ilstu.edu, aapokor@ilstu.edu

**Abstract**

Determining the price of a home was once an abstract process based on intuition and rough comparison to neighboring homes. However, in the modern day it would not seem to be sufficient to determine profits on mere intuition. Proper housing price modeling within relatively fixed time is affected by various house parameters such as square footage, number of bedrooms, number of bathrooms, etc. This project utilizes several selection and assessment techniques to develop the most superior model for predicting the expected price of a home. Data was collected as housing sales from May 2014 - May 2015 in King County, Washington. Evaluation

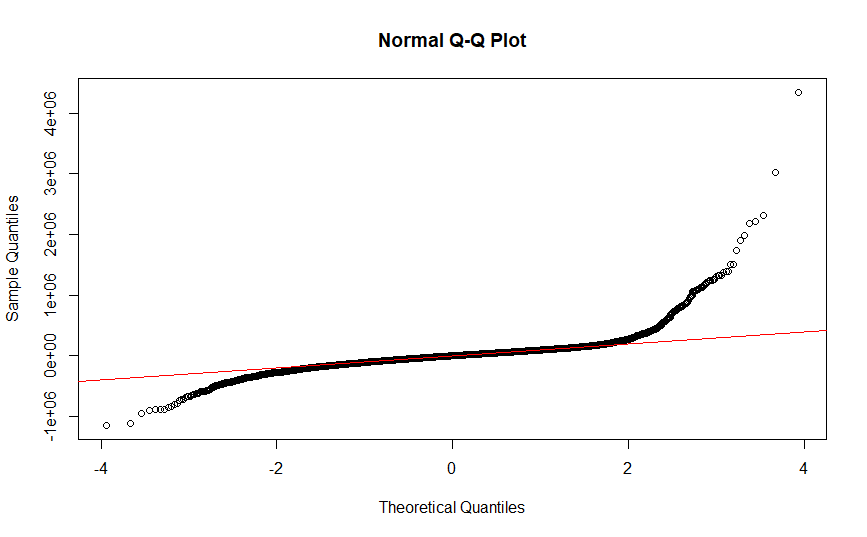
**Introduction**

[insert intro here]

**Results**

Our goal is to predict housing prices accurately with the highest attainable accuracy which we shall define to be the model with the lowest root mean squared error (RMSE).

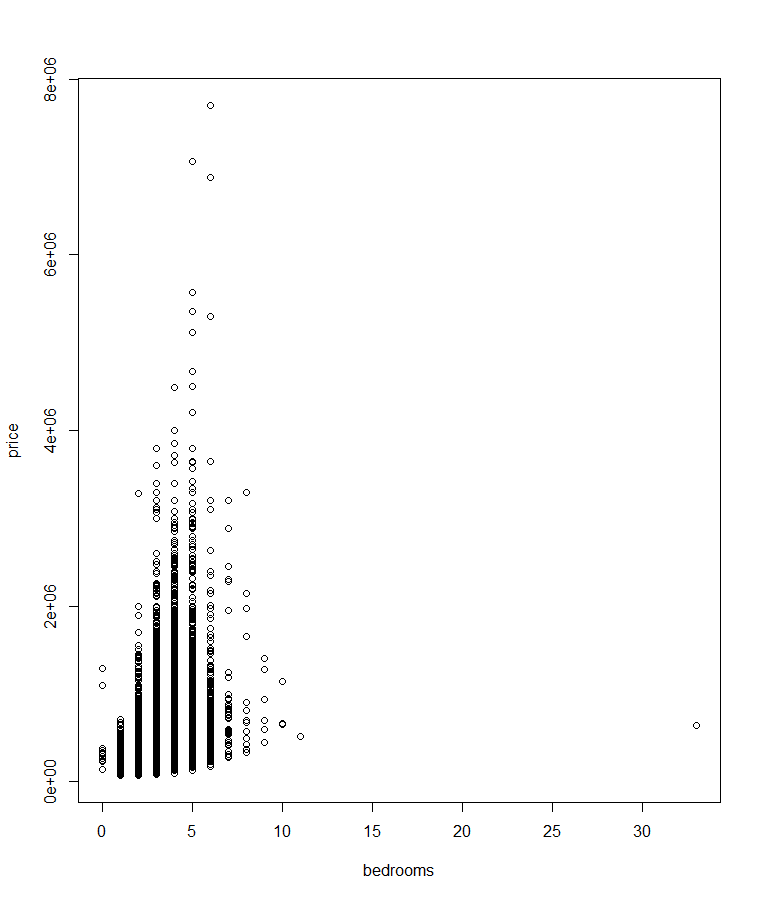
Ordinary Least Squares Process

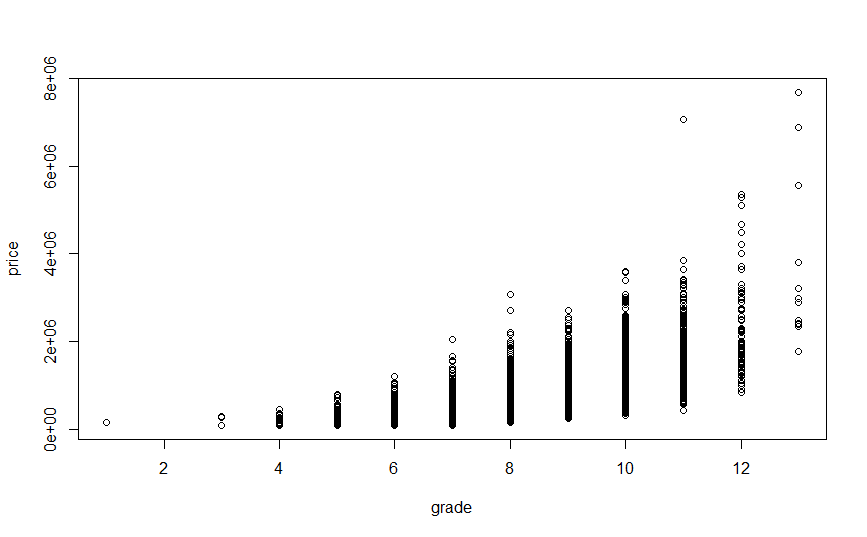
Before proceeding with development of a proper model, various adequacy measures had to be checked. Initially we observed residuals behavior through a residual plot as shown.

This plot shows the residuals plotted in black and two horizontal red lines representing limits at three standard deviations from the mean. This allows us to observe how violations to the normality assumption are operating. Further analysis demonstrated that the data was relatively controlled with 98.7% of data falling within the limits. 116 observations were greater than the upper limit and 47 fell below the lower limit. This was the first piece of evidence suggesting outliers having higher than expected prices. This had an apparent impact on the normality plot which suggests a positive skew. We were suspicious of the effect these outliers were having on our model. Before considering logarithmic transformations, considerations were made for omitting large outliers due to the difficulty for simple regression models to predict outliers. The difficulty with ordinary least squares regression is the not-so-robust nature that suffers from severe outliers.

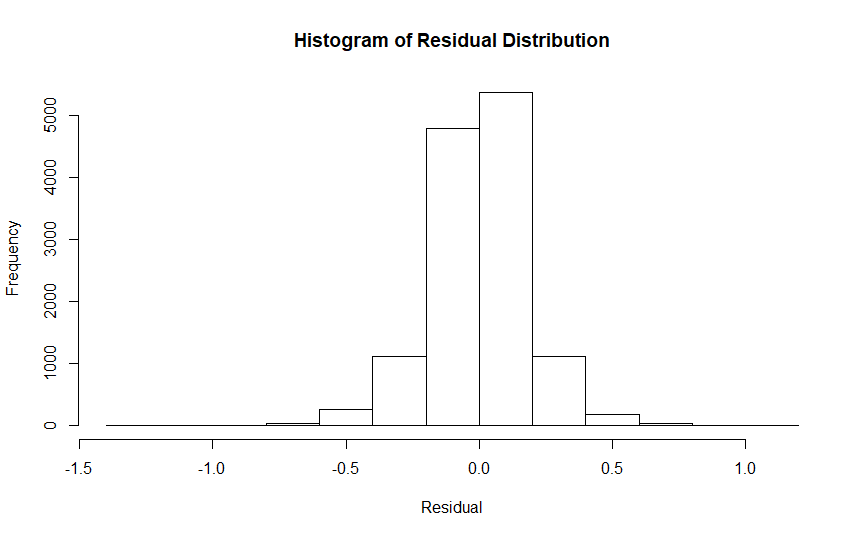
A modified regression model was fit to the data excluding variables with residuals outside of 4 standard deviations of the mean. The model was tested against a test sample which included the omitted residuals. The RMSE came out to 137,120.3. On average this predictive model deviated from the actual test set by $88,883.16. This is an improvement on the baseline regression model which was fit with all regressors and no pre-processing. The baseline model demonstrated an RMSE of 151,035.4 with average price deviation of $97,386.32. This makes clear that the model trained with outliers was influenced to some degree.

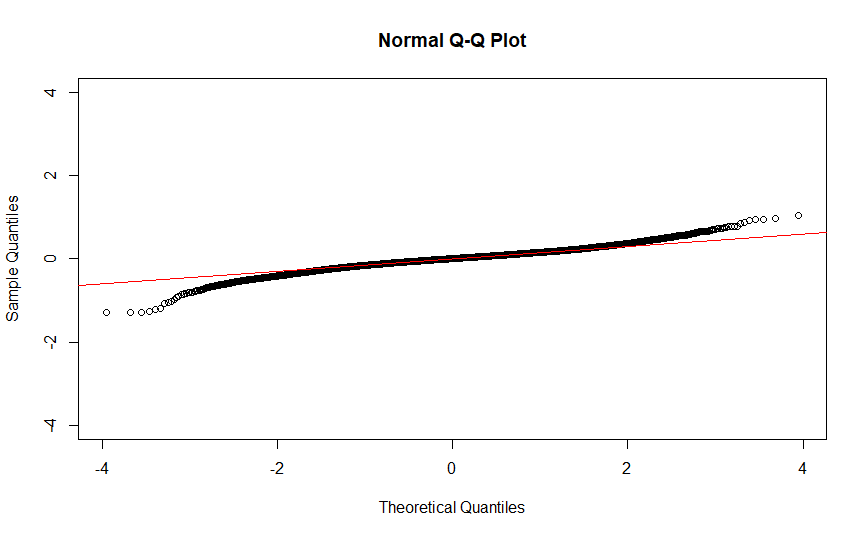
The next logical step is to perform subset selection. The backward-selection subset reduced the predictive variables from 84 to 67 but had little significant impact on prediction with an RMSE of 137460.4 and average predicted deviation of $89,042.93. The forward selection model pushed all variables into the model. We felt that traditional subset selection would not provide notable improvement to our model. Surprisingly, the adjusted R2 is almost the same in the summary: 0.8085 for the unreduced model and 0.8080 for the reduced model. This result is surprising since some predictors appeared very insignificant and the adjusted R2 did not punish the model excessively.

Following initial subset selection, investigation was conducted on the form of the data. Of principal interest was the relationship of individual predictors on the price of homes. Naturally this could not be graphically investigated for the zip code variables but we did research the quantitative variables and discovered certain patterns that suggested potential benefit from conducting a data transformation. One of the more telling variables was the plot of grade to price shown below. The data shows a potentially nonlinear pattern with non-constant variance. This model inadequacy did not stop at grade (a subjective determination of home quality). We discovered this pattern even with the number of bedrooms plotted against price. The result found was very surprising. We found that the minimum price of homes in our dataset increased slightly but persistently with additional bedrooms but overall the pattern did not persist and peaked at 8 bedrooms even though there were many homes with more bedrooms. Notice the outlier in the plot and the described pattern as well.



We then decided to investigate the effect of a logarithmic transformation on the price and tested the model’s performance on the test set. This offered substantial improvement over what we had previously. We achieved an RMSE of 116,284.7 and an average deviation from true price of $69,325.07. This was close to $20,000 better average accuracy in prediction! This also improved the adjusted R2 to 0.8761 With the updated model we repeated the backward selection algorithm to determine if noteworthy improvement can be had. Unfortunately no significant improvement was found with an RMSE of 116,144.71 and average deviation from true price of $69,314.71. Further, concerns over the robustness of this algorithm arose when it was realized that the algorithm was omitting a few of the zip code dummy variables. The only non-significant variable was sqft\_lot15 [WHAT WAS THIS VARIABLE AGAIN?]. Thus we determined a linear fit on all predictors (forcing every available zip code into the model) except for sqft\_lot15 and acquired the same adjusted R2 but a slight improvement to overall accuracy. RMSE worsened to 116,147.5 but average deviation in prediction improved to $69,302.77. It is not likely that much improvement can be derived from this.

 The final model can be observed in the plots below. Rather than solely rely on the normal Quantile-Quantile plot we include a histogram of the residuals as well to really clarify what is happening.



Overall we can see that the skew from the initial model is gone and we are left with some more extreme values than expected but the overall shape still appears relatively acceptable. Ultimately, we feel that further work on this regression model will not lead to substantial benefit.

Artificial Neural Network Process

The usage of artificial neural networks has demonstrated substantial strength in its ability to predict in various scenarios. It is one of the top methods for modern classification problems, often beating out random forest and support vector machine models. Utilizing a strong package for this task is imperative. Initially we considered utilizing NeuralNet but found the progress quite slow. For this reason we decided to work with H2O which is one of the most utilized packages for Fortune 500 companies. One major benefit of H2O is its ability to select automatic loss functions. Another Pursuing a method with automatic loss selection and a single hidden layer with 25 nodes showed some degree of promise. This is one of the simplest models we could have specified and the result was a RMSE of 108,143.14 (a discernable improvement from before) and prediction deviation of $68,335.71. Note that this result did not require a logarithmic transformation. This result could be seen as a baseline from which to improve. Further improvement was certainly expected.

The two loss functions we considered were traditional squared error and absolute loss. This selection is a debated subject. Our initial thought was that the absolute deviation loss function is superior, given our goal of predicting most homes with maximal accuracy. However, using trimmed data for the model creation also prevents issues of outliers pulling data excessively in the wrong direction. Nonetheless, we determined that testing both loss functions would be superior. We were then left with a few more decisions: Activation function, number of epochs, and overall design of the neural network.

Naturally we would seek to begin with the hyperbolic tangent function which has been used extensively for many years and demonstrated notable improvement over the Sigmoid function. We then decided to compare this performance to that of the Rectified Linear Units (ReLU) which has gained extensive attention in recent years. Number of epochs is a unique task that we decided to set as high as computationally possible. The goal is to achieve convergence before the pre-specified epochs are reached. We purposely kept it high and chose to monitor for overfitting. Thus epochs was set to 500 with a stopping criterion of 10,000 for improvement to the Mean Squared Error. Finally, neural network design was a work in progress of adding neurons then adding hidden layers. Some arguments have been made for the power of a single hidden layer so we decided to test the results. This selection is not an exact science but we decided to test with 200 neurons which led to immediate results. RMSE improved substantially to 99,338.08 and average deviation dropped to $62,022.73. Note that this was merely a test and a proper loss function had not been specified. However, it was valuable to show how much improvement could be derived from a model with so little specification on our part.

The experimentation continued as described but we found that more complex models proved to be very effective at accurately predicting home prices. For example, one of the top models achieved a RMSE of 85,484.05 and average deviation of $46,009.31. This essentially cut the original deviations in half! The model with this performance had three hidden layers of 200, 100, and 50 nodes respectively. All layers used the rectifier activation function and loss was computed as absolute loss. So where does this model fall short? We found that for some very highly priced homes our model demonstrated very large deviations from actual price. For example, the most expensive home in the test set was sold for almost $4,000,000 but our model predicted it at close to $2,500,000. An improvement on this specific instance was made when quadratic error was used leading to a model prediction of just under $3,000,000 but this also led to a decrease in RMSE and average deviation which is not the goal for this model.