Using Linear Regression to Predict House Prices in Washington

Katherine Botz, Jessica Páez Bonilla, Manuel Jordán Expósito

Abstract

The aim of this Predictive Modeling Project is to find an accurate model for predicting house sale prices in King County, Washington. We build several multivariate linear regression models to predict the Price response variable and select the best fitting model. To do this, we split the dataset into Train and Test data, then train the model using the former and calculate the model accuracy using the latter. To find each model we first fit a Linear Regression model with all predictor variables, then perform a stepwise model selection by minimizing the Bayesian Information Criterion. To adhere to the assumptions of the model we perform nonlinear transformations on certain variables while also balancing the fit and accuracy of the model. We select the best model in the end by maximizing both the Adjusted R-Squared and calculated Accuracy of the Price prediction in the Test data. Our final model consists of 15 predictor variables (which are all significant based on the t-test p-values) giving an Adjusted R-Squared of 0.7687 and 87.47% Accuracy.

Introduction

We use an open source dataset containing 21,613 instances of house sales in King County, Washington, USA between May 2014 and May 2015. The dataset includes descriptive information about the house itself (bedrooms, bathrooms, square feet, etc.), as well as location (in latitude and longitude, and zip code), grade, condition, and information about when it was built and/or renovated. With Price as the response variable, the goal is to build a Linear Regression Model to predict future prices of house sales based on these predictor variables.

Methods

We first perform an exploratory descriptive analysis on the data, and by means of a correlation matrix we find out which variables may have more influence on the sale price of the houses (high correlation) and which ones do not have any predictive value (low correlation). The correlation coefficient measures the sample covariance on a scale from -1 to 1, where the a number closer to 1 indicates a strong positive linear correlation between the variables. Variables with significant correlation mean that it could have predictive capacity over the response variable. A Scatterplot Matrix is also a good tool to visualize this relationship between the indicator variables and the response variable.

After this preliminary analysis is done we use the `lm` function to fit different linear regression models, which approximates a linear relationship between the response variable, Y (price), and the other predictor variables, $X_1, \ldots, X_p$, by finding the coefficients $\beta_0, \ldots, \beta_p$ such that: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

The linear model generated by `lm` is the one that minimizes the vertical distance of each data point projected on the line, or the least squares approximation. To select the best model we balance a few important qualities: the highest adjusted R-squared, significance of the variables, all while ensuring the normality assumption of the model is fulfilled.

Multiple R-Squared is a measure of the fit of a model; the higher the R-Squared, the lower the average variance of the data around the multivariate linear model. However, the Multiple R-Squared increases with complexity of the model, so we use the Adjusted R-Squared which does not favor more complex models.

In addition to the Adjusted R-Squared, we analyze the significance of each predictor variable on the response variable. We measure the significance of each variable by the p-value of the t-test; the Null Hypothesis in this test is that each $\beta_j = 0$. When the p-value is sufficiently low we can reject the Null Hypothesis and conclude that the variable has a significant effect on the response variable. A good model will have variables that all have a significant effect on Y.

To start with the model selection and analysis, we decide to use the stepAIC function to perform a stepwise model selection with different combinations of variables. This function analyzes each stepwise model for the best fit by finding the one with the smallest Bayesian Information Criterion when using k=log(n) as a parameter. This function (stepAIC) give us a set of variables with the most influence on the house price while minimizing the multivariate noise.

The Adjusted R-Squared and significance of variables are only a valid test of model fitness when the assumptions of the model are fulfilled. The assumptions of the model are: Linearity, Homoscedasticity, Normality, and Independence of the errors. During the analysis we notice that the data may not have a Normal distribution for this model. To judge Normality, we plot the QQ-plot to check that the standardized residuals more or less follow the diagonal line. If there are extreme departures from the diagonal line, we know that there may not be Normality in the data.

To fix the issue of Normality, we use the BoxCox function to find a nonlinear transformation and transform the variables that need the transformation to find a more normal distribution. However, we lose some fitness in our model by transforming too many variables. In the end, we balance model fitness in terms of Adjusted R-Squared and Normality and find a better model with a logarithmic transformation of Price and sqft_living.

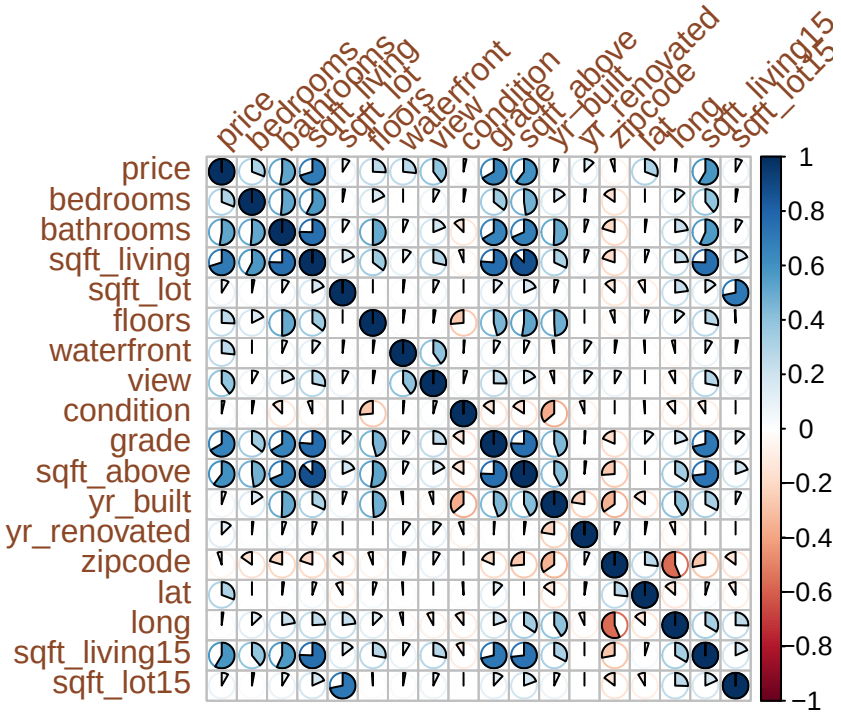
The goal of fitting our model is to make accurate predictions of Price, so we select three final models and use the Test data to predict the Price and then compare to the actual values to calculate the model accuracy. In this way, based on accuracy, best fit, and significant variables we are able to select the best model.

Statistical Analysis

We exclude the ID and Date variable, as these do not offer any predictive value for the price of the house. We also exclude sqft_basement, as it has a perfect dependent relationship with sqft_living. Finally we train the model with 80% of the data and reserve 20% to compare the predicted values to the actual values with our final selected model.

We first obtain the sample Correlation Matrix to find the variables with the greatest Correlation with Price. We can inspect the Correlation Heat Map to visualize this correlation:

```
cor.housing <- cor(housing)
corrplot(cor.housing,method="pie",tl.col = "sienna4",tl.srt = 45)
```



We find that the biggest Correlation Coefficients compared to Price are as follows:

Table 1: Top 5 Variables with Highest Correlation Coefficient Compared to Price

Variable	Correlation Coefficient
sqft_living	0.7020350
grade	0.6674342
sqft_above	0.6055673
sqft_living15	0.5853789
bathrooms	0.5251375

To further visualize the relationship between these particular variables and the response variable, we generate a Scatterplot Matrix:

```
#scatterplotMatrix(~ price + sqft_living + grade + sqft_above + sqft_living15 + bathrooms,
                   #reg.line = lm, smooth = FALSE, spread = FALSE, span = 0.5, ellipse = FALSE,
                   #levels = c(.5, .9), id.n = 0, diagonal = 'density', data = housing)
```

Indeed there does seem to be a positive linear relationship between Price and these specific variables

with the largest Correlation. However, it remains to be seen whether these variables will play a significant role in fitting a multivariate linear model to predict Price.

To start, we will fit a linear model with all 18 variables with the response set to the logarithm of Price. We find that the Adjusted R-Squared is a strong 0.7672, but certain variables such as sqft_lot15 do not have a significant linear effect on Price, due to the large p-value of the t-test.

```
modHouse1 <- lm(log(price) ~ ., data = train)
summary(modHouse1)
```

```
##
## Call:
## lm(formula = log(price) ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39950 -0.16214  0.00328  0.15987  1.19437
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.571e+00  4.134e+00  -0.864  0.38766
## bedrooms      -1.516e-02  2.793e-03  -5.428 5.79e-08 ***
## bathrooms      7.279e-02  4.605e-03  15.808  < 2e-16 ***
## sqft_living    1.570e-04  6.246e-06  25.138  < 2e-16 ***
## sqft_lot       4.056e-07  7.308e-08   5.551 2.89e-08 ***
## floors         7.693e-02  5.040e-03  15.263  < 2e-16 ***
## waterfront     3.631e-01  2.470e-02  14.699  < 2e-16 ***
## view           5.998e-02  3.049e-03  19.671  < 2e-16 ***
## condition      5.985e-02  3.345e-03  17.892  < 2e-16 ***
## grade          1.587e-01  3.033e-03  52.314  < 2e-16 ***
## sqft_above    -1.609e-05  6.153e-06  -2.614  0.00895 **
## yr_built      -3.487e-03  1.024e-04 -34.065  < 2e-16 ***
## yr_renovated   3.688e-05  5.136e-06   7.181 7.22e-13 ***
## zipcode       -6.638e-04  4.651e-05 -14.273  < 2e-16 ***
## lat            1.393e+00  1.513e-02  92.114  < 2e-16 ***
## long          -1.651e-01  1.850e-02  -8.924  < 2e-16 ***
## sqft_living15  9.490e-05  4.905e-06  19.347  < 2e-16 ***
## sqft_lot15    -1.840e-07  1.051e-07  -1.752  0.07985 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2536 on 17272 degrees of freedom
## Multiple R-squared:  0.7674, Adjusted R-squared:  0.7672
## F-statistic:  3352 on 17 and 17272 DF,  p-value: < 2.2e-16
```

In order to find a better fitting model for Price, we perform a stepwise model selection with different combinations of variables using the stepAIC function. For this dataset, the stepAIC function finds a model with Adjusted R-Squared = 0.7671. Although this is marginally smaller than the model with all variables, the variables all seem to be a good predictor of our response variable Price, as

shown by the respective t-test p-values.

```
summary(modBIC1)
```

```
##
## Call:
## lm(formula = log(price) ~ bedrooms + bathrooms + sqft_living +
##     sqft_lot + floors + waterfront + view + condition + grade +
##     yr_built + yr_renovated + zipcode + lat + long + sqft_living15,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39883 -0.16300  0.00334  0.15975  1.17852
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.432e+00  4.087e+00  -1.329    0.184
## bedrooms      -1.484e-02  2.791e-03  -5.316 1.07e-07 ***
## bathrooms      7.486e-02  4.551e-03  16.450  < 2e-16 ***
## sqft_living    1.461e-04  4.770e-06  30.637  < 2e-16 ***
## sqft_lot       3.113e-07  5.368e-08   5.800 6.76e-09 ***
## floors         7.140e-02  4.534e-03  15.749  < 2e-16 ***
## waterfront     3.600e-01  2.468e-02  14.587  < 2e-16 ***
## view           6.133e-02  3.004e-03  20.419  < 2e-16 ***
## condition      6.039e-02  3.335e-03  18.110  < 2e-16 ***
## grade          1.577e-01  3.002e-03  52.519  < 2e-16 ***
## yr_built      -3.485e-03  1.023e-04 -34.052  < 2e-16 ***
## yr_renovated   3.689e-05  5.136e-06   7.184 7.05e-13 ***
## zipcode       -6.612e-04  4.650e-05 -14.220  < 2e-16 ***
## lat            1.399e+00  1.502e-02  93.143  < 2e-16 ***
## long          -1.762e-01  1.816e-02  -9.705  < 2e-16 ***
## sqft_living15  9.246e-05  4.843e-06  19.092  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2537 on 17274 degrees of freedom
## Multiple R-squared:  0.7673, Adjusted R-squared:  0.7671
## F-statistic:  3797 on 15 and 17274 DF,  p-value: < 2.2e-16
```

Although this model does seem to be a good fit in terms of Adjusted R-Squared and t-test p-values, we find that we may not have Normality in the data.

```
qqPlot(modBIC1, main="QQ Plot")
```

This violates our assumptions for the model selection, so we will use the BoxCox function from the Caret library to find a proper transformation for each variables in order to come closer to Normality.

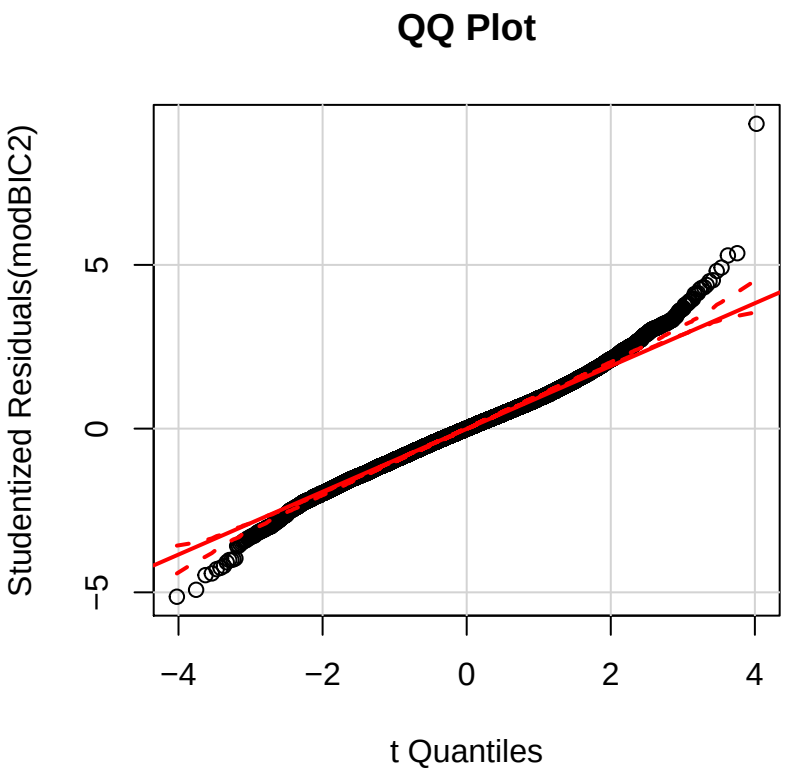
```
modBIC2 <- lm(formula = I(log(price)) ~ bedrooms + bathrooms + I(log(sqft_living)) +
                sqft_lot + floors + waterfront + view + I(1/sqrt(condition)) +
                I(log(grade)) + I((yr_built)^2) + yr_renovated + I(1/(zipcode)^2) +
                I((lat)^2) + long + I(log(sqft_living15)), data = train)
summary(modBIC2)
```

```
##
## Call:
## lm(formula = I(log(price)) ~ bedrooms + bathrooms + I(log(sqft_living)) +
##     sqft_lot + floors + waterfront + view + I(1/sqrt(condition)) +
##     I(log(grade)) + I((yr_built)^2) + yr_renovated + I(1/(zipcode)^2) +
##     I((lat)^2) + long + I(log(sqft_living15)), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31083 -0.16372 -0.00026  0.15640  2.33894
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -7.324e+01  3.927e+00 -18.650  < 2e-16 ***
```

```
## bedrooms              -3.088e-02  2.927e-03 -10.549  < 2e-16 ***
## bathrooms              8.352e-02  4.533e-03  18.428  < 2e-16 ***
## I(log(sqft_living))    3.227e-01  1.040e-02  31.022  < 2e-16 ***
## sqft_lot               4.162e-07  5.384e-08   7.731 1.13e-14 ***
## floors                 7.184e-02  4.571e-03  15.717  < 2e-16 ***
## waterfront             3.802e-01  2.486e-02  15.291  < 2e-16 ***
## view                   6.818e-02  3.011e-03  22.647  < 2e-16 ***
## I(1/sqrt(condition))  -6.099e-01  4.324e-02 -14.104  < 2e-16 ***
## I(log(grade))          1.212e+00  2.259e-02  53.650  < 2e-16 ***
## I((yr_built)^2)       -1.026e-06  2.596e-08 -39.534  < 2e-16 ***
## yr_renovated           2.497e-05  5.161e-06   4.838 1.32e-06 ***
## I(1/(zipcode)^2)       2.942e+11  2.212e+10  13.304  < 2e-16 ***
## I((lat)^2)             1.465e-02  1.595e-04  91.898  < 2e-16 ***
## long                  -1.647e-01  1.827e-02  -9.017  < 2e-16 ***
## I(log(sqft_living15))  2.184e-01  9.850e-03  22.173  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2556 on 17274 degrees of freedom
## Multiple R-squared:  0.7637, Adjusted R-squared:  0.7635
## F-statistic:  3723 on 15 and 17274 DF,  p-value: < 2.2e-16
```

The latest model has a lower Adjusted R-Squared of 0.7635, but the data distribution is getting further from Normality.

```
qqPlot(modBIC2, main="QQ Plot")
```



With some trial and error using various combinations of nonlinear transformations, we find a better fitting model with Adjusted R-Squared = 0.7687 and with a more Normal distribution.

```
modBIC3 <- lm(formula = log(price) ~ bedrooms + bathrooms + log(sqft_living) +
     sqft_lot + floors + waterfront + view + condition + grade +
     yr_built + yr_renovated + zipcode + lat + long + sqft_living15, data = train)
summary(modBIC3)
```

```
##
## Call:
## lm(formula = log(price) ~ bedrooms + bathrooms + log(sqft_living) +
##     sqft_lot + floors + waterfront + view + condition + grade +
##     yr_built + yr_renovated + zipcode + lat + long + sqft_living15,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31072 -0.16149  0.00155  0.15686  1.25640
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.057e+01  4.074e+00  -2.595  0.00947 **
## bedrooms        -2.651e-02  2.902e-03  -9.137  < 2e-16 ***
## bathrooms        7.495e-02  4.492e-03  16.686  < 2e-16 ***
```

```
## log(sqft_living)  3.286e-01  1.007e-02  32.648  < 2e-16 ***
## sqft_lot          3.720e-07  5.325e-08   6.986 2.92e-12 ***
## floors            6.837e-02  4.516e-03  15.138  < 2e-16 ***
## waterfront        3.722e-01  2.459e-02  15.137  < 2e-16 ***
## view              6.357e-02  2.990e-03  21.263  < 2e-16 ***
## condition         5.544e-02  3.328e-03  16.660  < 2e-16 ***
## grade             1.601e-01  2.944e-03  54.405  < 2e-16 ***
## yr_built         -3.714e-03  1.011e-04 -36.743  < 2e-16 ***
## yr_renovated      3.215e-05  5.120e-06   6.280 3.48e-10 ***
## zipcode          -6.331e-04  4.630e-05 -13.672  < 2e-16 ***
## lat               1.399e+00  1.496e-02  93.476  < 2e-16 ***
## long             -1.818e-01  1.810e-02 -10.044  < 2e-16 ***
## sqft_living15     9.655e-05  4.738e-06  20.378  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2528 on 17274 degrees of freedom
## Multiple R-squared:  0.7689, Adjusted R-squared:  0.7687
## F-statistic:  3832 on 15 and 17274 DF,  p-value: < 2.2e-16
```

```
qqPlot(modBIC3, main="QQ Plot")
```

The latest model has an improved Adjusted R-Squared with all significant variables, while also improving the Normal distribution and adhering to the assumptions of a multivariate linear model.

We can analyze all models for best Adjusted R-Squared, Normality, and significant variables, but the model with the best fit will have the most accurate prediction of the response variable, Price. We will predict the Price using our reserved Test data for our top three models, then compare to the actual reported prices to compute the model accuracy.

Table 2: Top 3 Models for Fitness and Accuracy

Model	Description	Adjusted R-Squared	Correlation	Accuracy
modBIC1	Model Selected by StepAIC	0.7671	79.68%	80.3%
modBIC2	Previous Model Modified by Transformations	0.7635	86.95%	80.4%
modBIC3	Previous Model With Limited Transformations	0.7687	87.47%	80.5%

NOTE: Although there are 15 variables in each of the best models, an attempt to reduce the dimensions using Principal Component Analysis was not fruitful; the best model using Standardized Principal Components consisted of 16 elements. Therefore we will not include that analysis in this report.

Conclusions

Based on the previous table we come to the conclusion that the best model for predicting the house prices is the third one (modBIC3), as it has the highest Adjusted R-Squared (0.7687) and the highest accuracy rate (80.5%) in the prediction of the response variable.

In this model we have 15 variables that all significantly affect and offer predictive value to the house price. We use logarithmic transformations on the response variable and the variable sqft_living in the model. We also have normality in our data, therefore meeting the assumptions of the model and achieving the aim of our project to create a model to accurately predict Price. We can say with 80.5% certainty that this model will accurately predict the price of a house sold in King County, Washington, USA between May 2014 and May 2015 with the predictor variables in our model.

References

Harlfoxem. "House Sales in King County, USA." *Kaggle*, 25 Aug. 2016, www.kaggle.com/harlfoxem/housesalespredi

Portugués, Eduardo García. *Notes for Predictive Modeling.* 29 Dec. 2017, bookdown.org/egarpor/PM-UC3M/.

Prabhakaran, Selva. "Tutorials on Advanced Stats and Machine Learning With R." *r-Statistics.co*, 2016, r-statistics.co.