Application of Multiple Linear Regression on Real Estate Market Data of Tallahassee

Rufeng Liu rl19o@my.fsu.edu

Ting Hu th19d@my.fsu.edu

Department of Statistics, Florida State University

April 30, 2020

Abstract

This report applies multiple linear regression model to study a dataset from real estate market of Tallahassee. To improve our regression model, we apply various methods such as variables transformation, adding quadratic terms, eliminating outliers, and implementing variable selection. We also use cross validation to compare performance of different models. In this analysis, we find that, for sold price of a house, the most important numerical predictor is floor size and the most important categorical predictor is ZIP code.

1. Introduction

Evaluating value of a house is a key problem in real estate market. It is hard for both buyers and sellers to decide a correct price of a house. We will try to make a model to solve this problem according to the data collected by an estate company. In this report, our main purpose is to figure out how the sold price of a house is influenced by different predictors.

Our dataset contains information of houses sold between Jan 7th and Apr 7th 2020 in Tallahassee. We implement multiple linear regression and residual diagnosis on this dataset. This paper proceeds as follows. In section 2 we show the methodologies we use. In section 3 we discuss the data and model we use, and section 4 shows the process of model fitting. In section 5, numerical result is showed to illustrate our findings. Finally, we close with a summary in section 6.

2. Methodology

2.1. Box-Cox Transformation

Box-Cox transformation can improve the quality of residuals of a certain model by introduce a transformation to predictors or response as:

$$\mathrm{y} \rightarrow y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

Suppose there exists a λ makes transformed response $y^{(\lambda)}$ with normal assumption followed, the likelihood function is:

$$L(\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left\|Y^{\lambda} - X\beta\right\|^2\right) \times \prod_{i=1}^{n} y_i^{\lambda-1}$$

The log-likelihood function is:

$$logL(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\|Y^{\lambda} - X\beta\right\|^2 + (\lambda - 1)\sum \log(y_i)$$

Then MLE is given by:

$$l_p(\lambda) = l\left(\lambda, \hat{\beta}(\lambda), \hat{\sigma}^2(\lambda)\right) = C - \frac{n}{2} logRSS(\lambda) + (\lambda - 1)\sum \log(y_i)$$

The best estimation of λ is given by $\widehat{\lambda} = argmax(l_p(\lambda))$.

Refitting the model with transformed variables will give us more normal residuals.

2.2. Stepwise Regression

Stepwise regression includes forward selection, backward selection and combination of both. For forward selection, we start with adding a predictor to the model that contains no variable, and choose the predictor improves RSS most. We add it into our model if AIC becomes smaller with this predictor added, otherwise we stop selection process.

For backward selection we start with the whole model that contains all variables, remove the predictor that cause least change of RSS during each step, and make sure AIC becomes smaller, stop selection process otherwise. Combination of both approaches is based on forward selection, except check if there is any predictor we need to delete after adding a predictor in our model.

2.3. AIC & BIC

Akaike Information Criterion and Bayesian Information Criterion are two criteria using in variable selection and model selection. Their formulas are given by:

$$AIC = -2logL(\hat{\theta}) + 2k; \qquad BIC = -2logL(\hat{\theta}) + klog(n)$$

where n is sample size, k is number of parameters, $L(\hat{\theta})$ is likelihood function, $\hat{\theta}$ is the maximum likelihood estimation.

In the case of multiple linear regression, they are given by:

$$AIC = log\ ((RSS_k)/n) + 2k + C; \qquad BIC = log\ ((RSS_k)/n) + klog(n) + C$$

where C is an irrelevant constant, RSS_k is the residual sum of squares with k predictors in our linear model.

We can see AIC and BIC are similar except that BIC penalizes complexity of model harder.

2.4. Cross Validation

We use the simplest validation method in this report. We reserve a small sample of the dataset. Then we train the model on the training dataset. We apply the model to the test dataset to predict the outcome of new unseen observations.

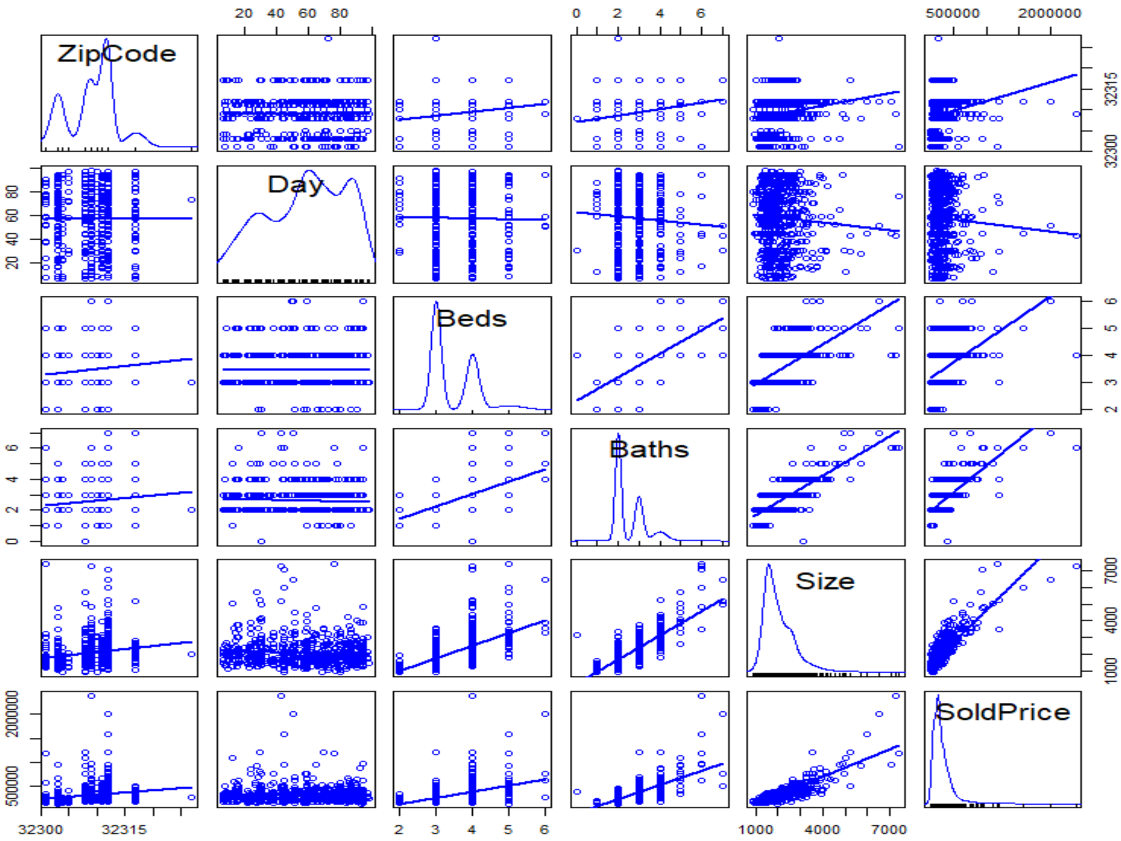
We quantify the prediction error as the mean squared difference between the observed and the predicted outcome values.

3. Data and Approach

3.1. Data preprocessing

Our dataset is collected by Eden Company. It contains 630 observations from the beginning of this year till April 7th. It contains 9 variables. We remove **DOM** as the values of this variable are unknown. We are not interested in **List Price** and **Cost per ft2**, as **List Price** and **Sold Price** are almost the same, and **Cost per ft2** can be calculated by $\frac{\textbf{Sold Price}}{\textbf{Living ft2}}$. We will treat $\textbf{Living ft2}$ and **Sold Date** as numerical variables while **Address**, **Beds** and **Baths** as categorical variables. Specially, for **Address**, we transform it into a categorical variable "**ZipCode**" with 9 levels to indicate the different area of Tallahassee. For **Sold Date**, we transform it into **Day** which is the date from the beginning of this year (2020).

3.2. Exploratory data analysis



From the Scatterplot Matrix, we find there is a data with **Baths** = 0. In common sense, the number of bedrooms cannot be 0, we remove this data from our dataset. **Size**, **Beds** and **Baths** are positive related with **Sold Price**, and they are positive related with each other. In **Size** VS **Sold Price** plot, there is a heavy tail indicates that transformation is required.

3.3. Model specification

We use a multiple linear regression: $Y = X\beta^T + \epsilon, \varepsilon \sim N(0, \sigma^2 I)$ where Y is **Sold Price**, X represent other variables which we treat as predictors, and β denotes coefficients.

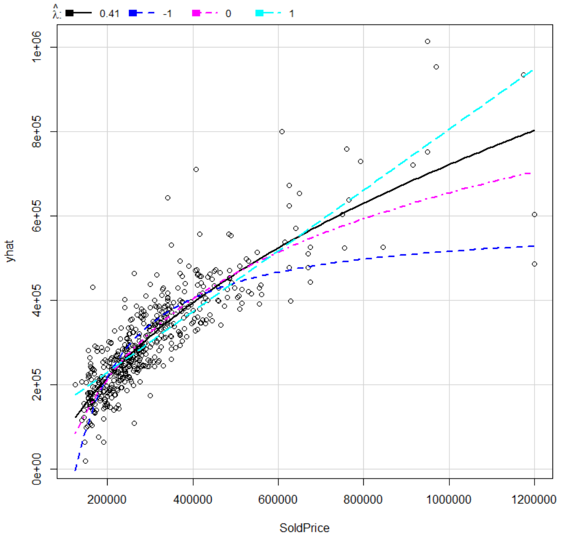
4. Model Fitting

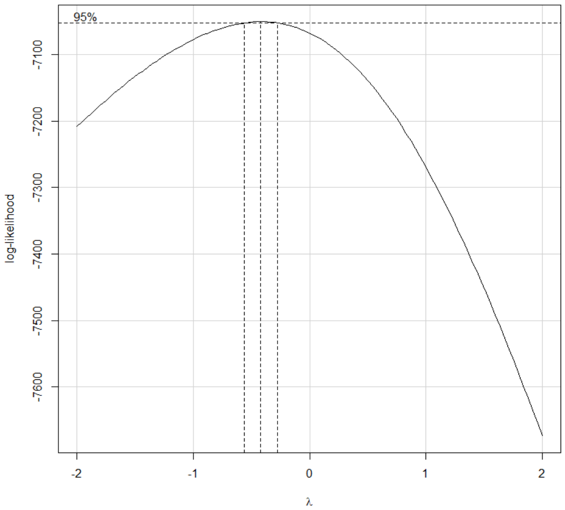
Before we do the model fitting, we first split the dataset into two part. 80% of the data are randomly selected for training while 20% are reserved for validation.

4.1. Transform X and Y

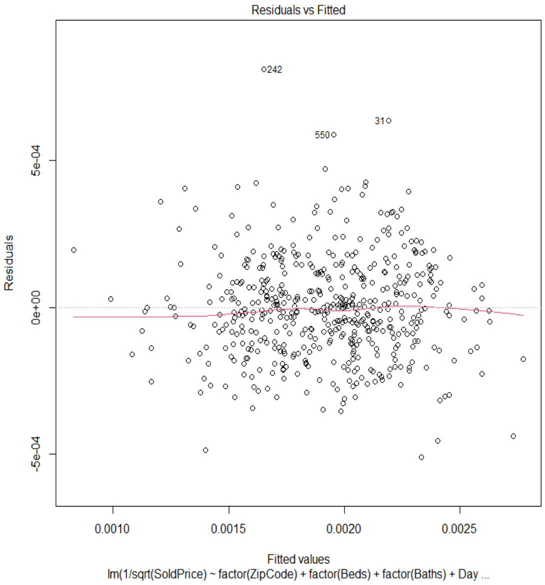
Power transformation shows that we need $\lambda = -0.5$ for **Size** and no transformation for **Day**. For categorical variables, we do not need to do transformation.

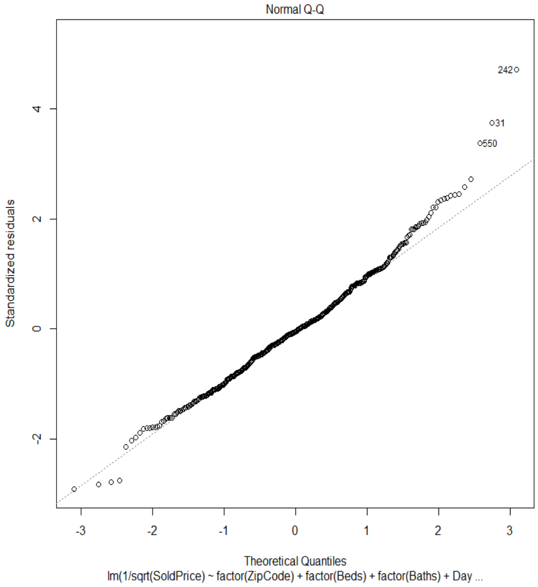
As Inverse Response plot and Box-Cox plot indicate, we will do transformation for response with $\lambda = -0.5$





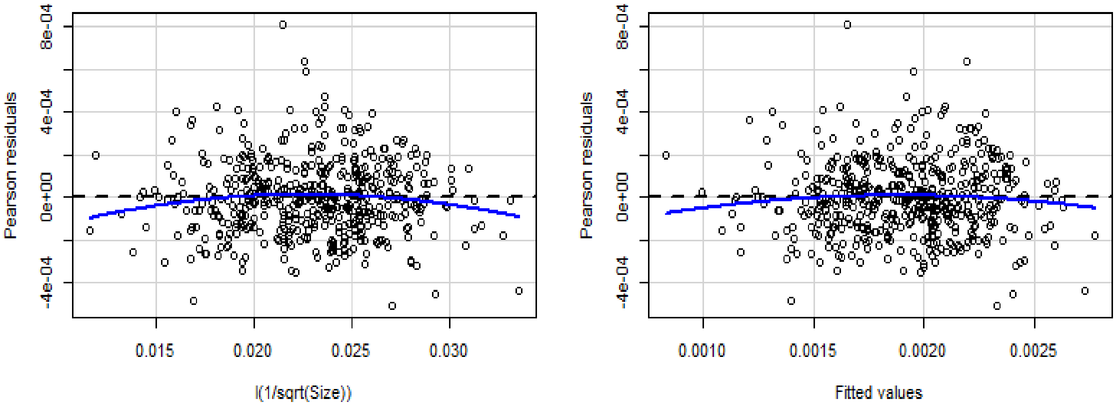
After refitting the model with transformed response and predictors, our residual plot and normal Q-Q plot show that the normality assumption is followed, and Non-constant Variance Score Test has p-value= 0.69011.



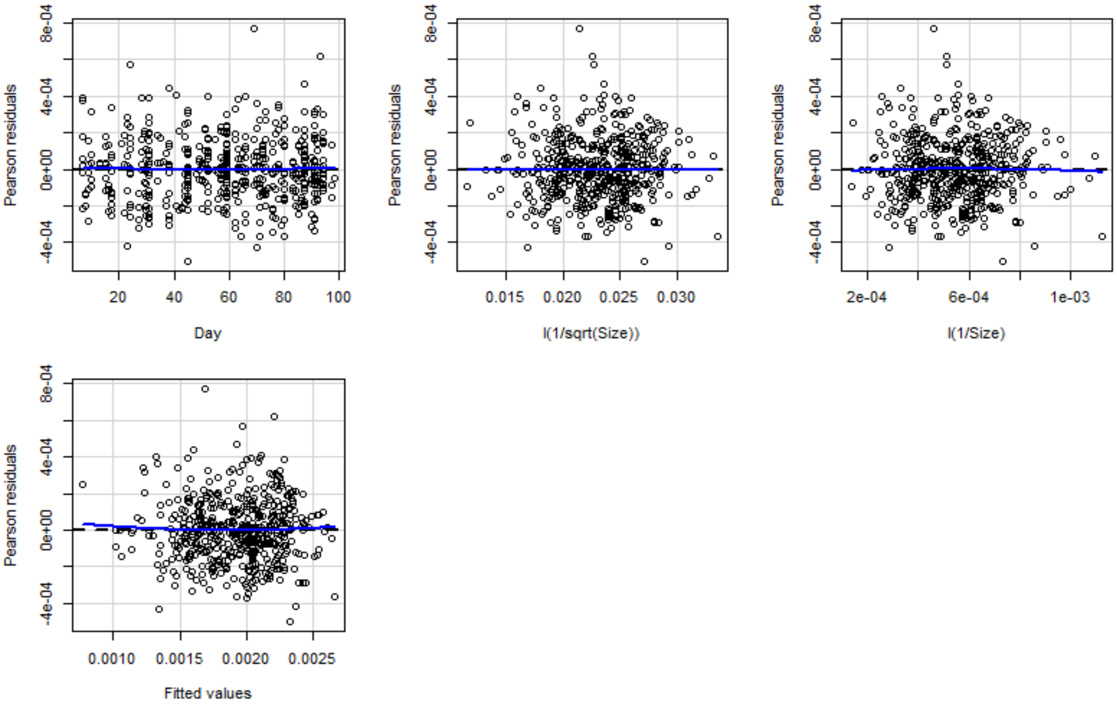


4.2. Add quadratic terms

With Tukey test has a p-value less than 0.05 and residual plots show curvature, we need to add some quadratic terms.



By adding a quadratic term for $Size^{-0.5}$, Multiple R^2goes up to 0.7748 (Adjusted R^2: 0.7646). Tukey test shows no significance and our residual plots shows no curvature.



4.3. Remove Outliers

The residual plots above 4.2 and the function outlierTest() in R identify conspicuous outliers, whose observation numbers are 242,31,550 and 429. Sold prices of these houses are much higher or lower than other houses which have similar conditions, we can see from our original table that they have quite different **Cost per ft2**. By removing these outliers, our Multiple R^2goes up to 0.7946 (Adjusted R^2: 0.7852).

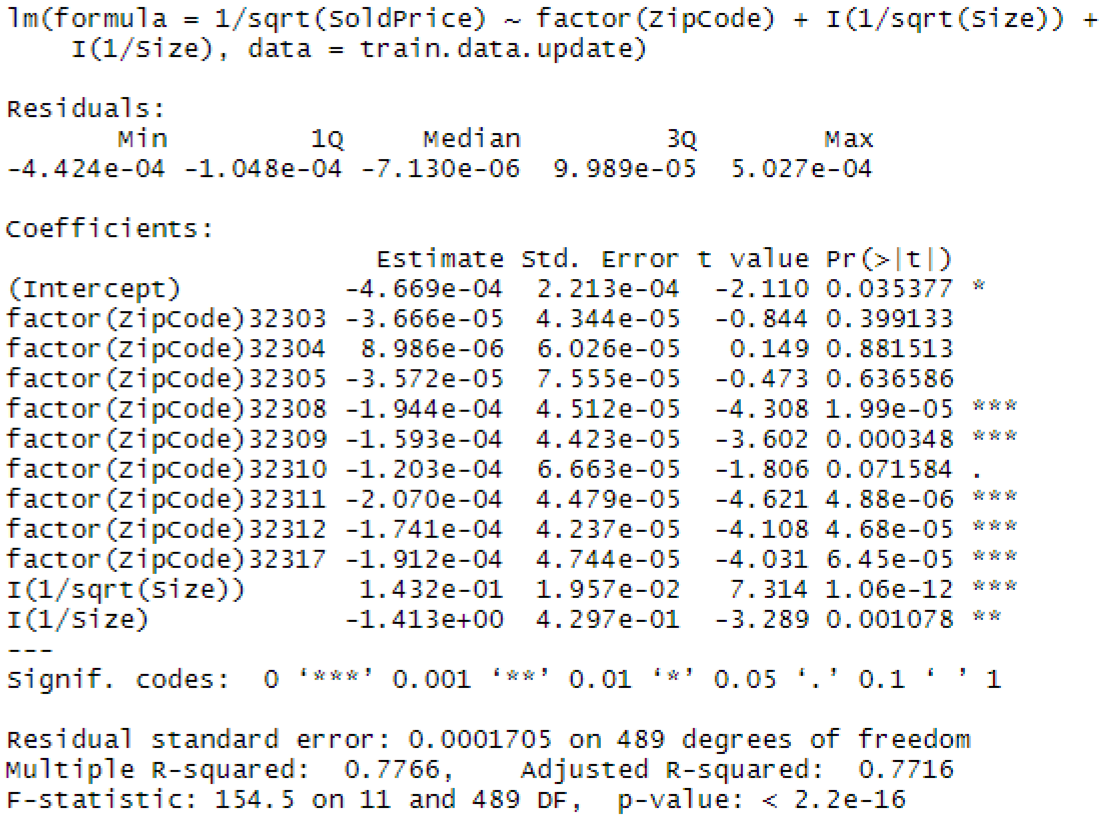
4.4. Variable selection

As unnecessary predictors can cause multicollinearity, we apply stepwise regression which run forward and backward with two information criteria, AIC and BIC. Backward stepwise with BIC gives the smallest model, which remove the non-significant predictors **Beds** and **Baths**. We decide the model from backward selection with BIC as our result.

	FULL	FORWARD/AIC	BACKWARD/AIC	FORWARD/BIC	BACKWARD/BIC
# OF COEFFICIENTS	21	20	20	20	11
P-VALUE IN NCV TEST	0.89723	0.89723	0.87534	0.89723	0.7748
ADJUSTED *R*	0.7852	0.7852	0.7728	0.7852	0.7716

5. Result and Interpretation

Our final model is:



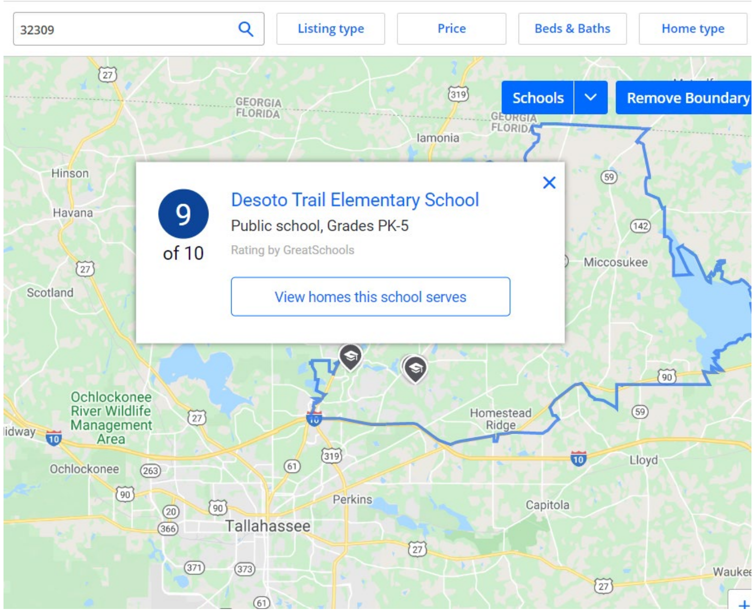
```
lm(formula = 1/sqrt(SoldPrice) ~ factor(ZipCode) + I(1/sqrt(Size)) +
    I(1/Size), data = train.data.update)

Residuals:
       Min         1Q     Median         3Q        Max
-4.424e-04 -1.048e-04 -7.130e-06  9.989e-05  5.027e-04

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -4.669e-04  2.213e-04  -2.110 0.035377 *
factor(ZipCode)32303 -3.666e-05  4.344e-05  -0.844 0.399133
factor(ZipCode)32304  8.986e-06  6.026e-05   0.149 0.881513
factor(ZipCode)32305 -3.572e-05  7.555e-05  -0.473 0.636586
factor(ZipCode)32308 -1.944e-04  4.512e-05  -4.308 1.99e-05 ***
factor(ZipCode)32309 -1.593e-04  4.423e-05  -3.602 0.000348 ***
factor(ZipCode)32310 -1.203e-04  6.663e-05  -1.806 0.071584 .
factor(ZipCode)32311 -2.070e-04  4.479e-05  -4.621 4.88e-06 ***
factor(ZipCode)32312 -1.741e-04  4.237e-05  -4.108 4.68e-05 ***
factor(ZipCode)32317 -1.912e-04  4.744e-05  -4.031 6.45e-05 ***
I(1/sqrt(Size))       1.432e-01  1.957e-02   7.314 1.06e-12 ***
I(1/Size)            -1.413e+00  4.297e-01  -3.289 0.001078 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0001705 on 489 degrees of freedom
Multiple R-squared:  0.7766,    Adjusted R-squared:  0.7716
F-statistic: 154.5 on 11 and 489 DF,  p-value: < 2.2e-16
```

Beds and **Baths** are insignificant probably because it is highly correlated with **Size**. While, **Day** is insignificant because our dataset only contains the first four months of this year, which can not show whether **Sold Price** changes seasonally or yearly. Some level of **ZipCode** are significant while some are not. By checking Zillow, we find areas with high rating schools are always significant, while areas with low rating schools are insignificant.

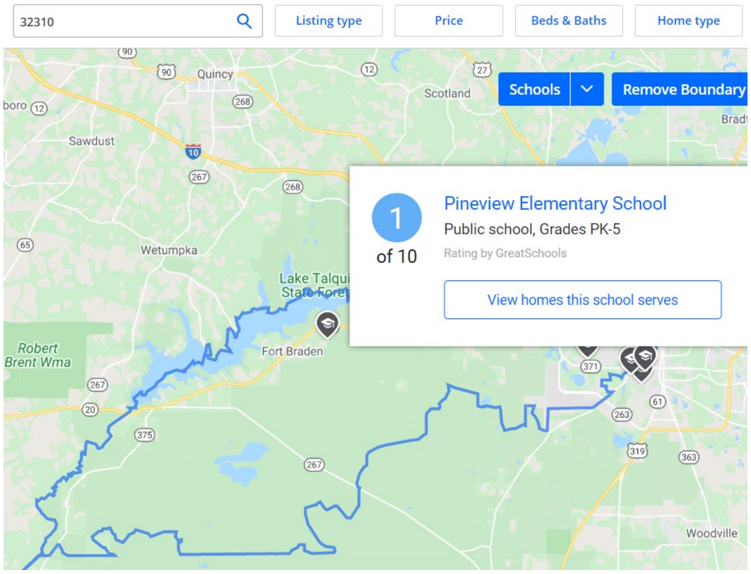

32309

9 of 10

Desoto Trail Elementary School

Public school, Grades PK-5

Rating by GreatSchools


32310

1 of 10

Pineview Elementary School

Public school, Grades PK-5

Rating by GreatSchools

6. Conclusion

This project applies multiple linear regressions to study the datasets from Eden Company. By using approaches including Box-Cox transformation, adding quadratic terms, removing outliers and variable selection, our final model for the home price has adjusted coefficient of determination is 77.16%. Constant variance assumption and normality assumption are hold. Our final model can explain our data well enough.

For further study, we may need a larger dataset which contains the last three- or five-years' data. We will check the robustness of our estimation by time and find whether the house sold price changes seasonally or yearly.