UNIVERSITY OF SAN FRANCISCO

Case Study: Housing Price Analysis with a Regression Approach

by

Bingyi Li
Fang Wang
Guoqiang Liang

A report submitted in partial fulfillment for the
course of Linear Regression Analysis

in the
Master of Analytics Program

October 2017

Contents

Chapter 1

Introduction

Every home buyer places a certain amount of significance on the many different aspects of a particular property that give that property its inherent value. For instance, some houses are worth more because of the year the house was built, square footage of the home, the number of bedrooms or bathrooms, their proximity to a desirable community, or even the height of the basement ceiling. Some of these aspects seem to be trivial at first glance, however they can actually have much more influences on price negotiation.

This research aims to explain if there are certain identifiable characteristics of housing that tend to make the greatest impact in terms of which houses have a higher price in a market and which houses have a lower price. We also attempted to accurately predict the sale price of a house based on those significant features. Our research is focused specifically on residential homes in Ames, Iowa. The method we use for analysis is linear regression.

The housing data has 82 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. However, some of them are not stored in appropriate types or include missing values. So we need to perform data cleaning and imputation before regression analysis, which is described in Chapter 2.

After data manipulation, we attempt to build two regression models: an explanatory model and a predictive model. The explanatory model is applied to show relationships between sale prices and aspects of houses. By applying this model, we will be able to explain which features have significant influences on prices. Our aim is to offer simple and clear explanations with as few parameters as possible. For predictive models, we intend to focus less on parsimony or simplicity, but more on their ability to predict sale prices.

Chapter 2

Data

2.1 Descriptive Statistics

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. The dataset contains information from the Ames Assessors Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. Tab characters are used to separate variables in the data file. The data has 82 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers).

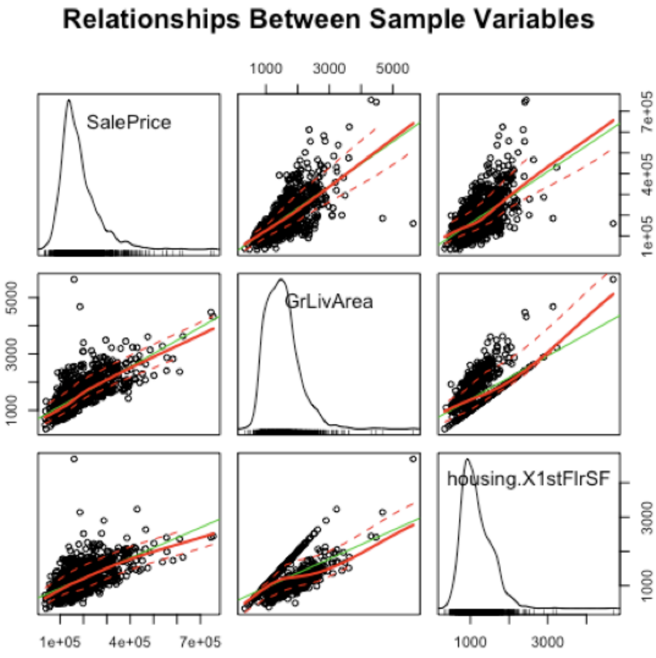


Figure 2.1: Relationships Between Sample Variables

Among those 20 continuous numerical variables, some of them might be highly correlated with sale prices, and they are very likely to be significant in our regression model. For example, figure 2.1 shows the relationship between sale price, above ground living area and first-floor square feet of all houses.

From the diagram, it is clear to see that both above ground living area and first-floor square feet are strongly positively correlated with the sale price, so it is reasonable to expect that these two factors will be significant in our regression models. However, we can also notice that there is also a strong correlation between above ground living area and first-floor square feet, which indicates collinearity problem among independent variables. Such problem will be handled by feature selection, which will be discussed later in section 3.3.

There are also 46 categorical variables In the housing datasets. In order to have more insight into the data, we created bar plots for categorical features which might be correlated with the sale price.

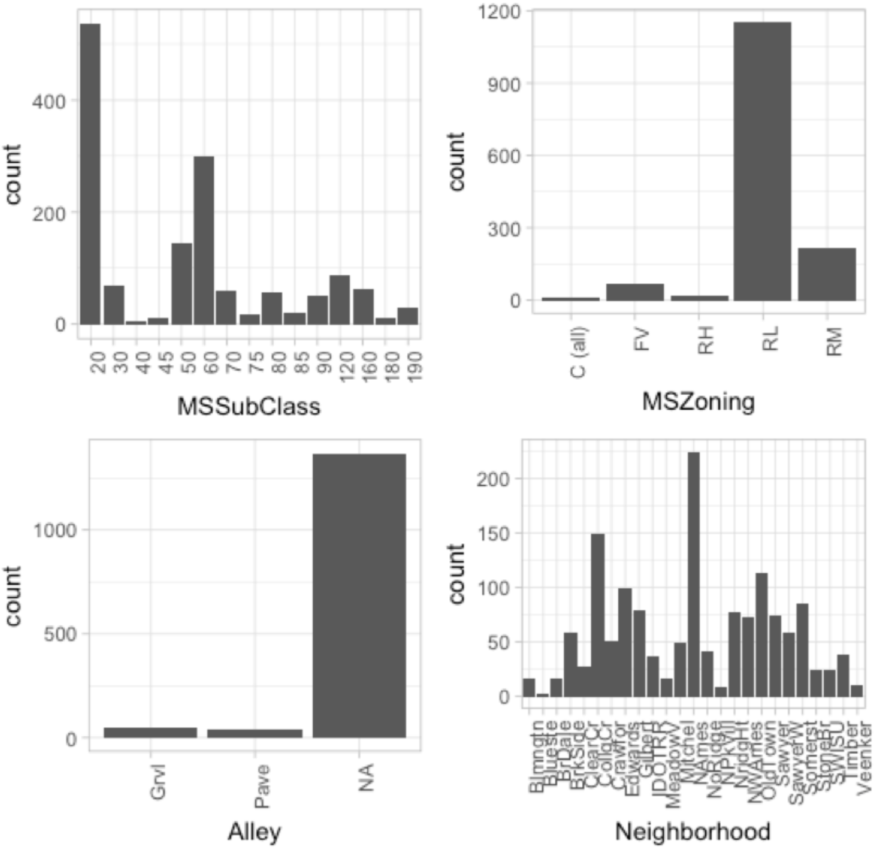


FIGURE 2.2: Bar Plots for Sample Categorical Features

For example, figure 2.2 indicates the following:

1. The most common MSSubClass is 20, which means that most houses have 1 story and are built after 1946.
2. Majority of the houses are located in low density residential areas and medium density residential area.
3. Most houses have no information about alley.
4. NAmes and CollgCr are the neighborhoods which have the most amount of houses.

Based on the results, we can see that there are numbers of problems with the original data. So before performing regression analysis, we need to do some imputations and transformations.

2.2 Data Imputation and Transformation

As discussed in section 2.1, there are numbers of problems that needed to be fixed before we can perform regression analysis. In this section, we will identify each type of problems with the original dataset, and then perform data imputation and transformation to fix those issues.

The first problem is about missing values. In the original dataset, there are large numbers of missing values stored as NA. Since NAs are not calculable, it is necessary to replace them with something else. For categorical variables, we replaced all of them with None. And for numerical variables, we used various means to replace missing values. This method will work because average values are always in the middle of the range and thus will not have any strong effect on regression results. By replacing NAs with either None or mean values, we are able to do calculations with missing values. The only exception here is that we decide to delete the single row with missing value in the feature 'Electrical'.

The second change we made is to change 'year' to 'year difference'. Since values of years are between 1872 and 2017, which are relatively large numbers comparing to other numerical values, we decided to replace them with the difference between 2017 and each value. For example, if a house was built in 2010, we replaced 2010 with the difference between 2017 and 2010, which is 7. The advantage of this method is that we can avoid coefficients of certain variables being too small compared to others, thus avoid underestimating the effects on the sale price of those variables.

The third change we made to the original dataset is log transformations of some independent variables. As Figure 2.3 shows, distributions of some numerical variables in the original dataset are highly skewed, but with a log transformation, their distributions can be well normalized. After testing all numerical variables, we chose 4 of them to do log transformation:

- Lot Area, MasVnrArea, 1stFlrSF, GrLivArea

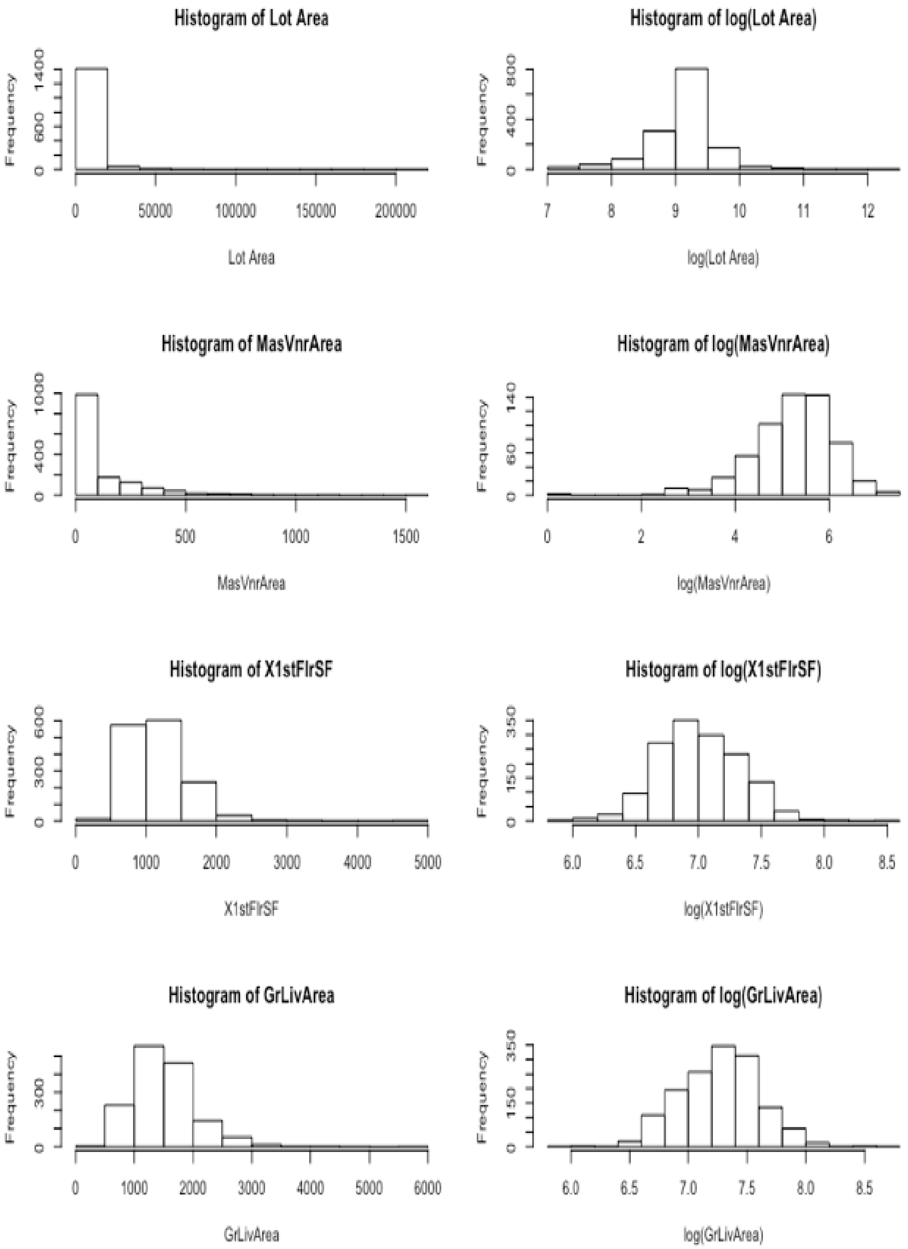


Figure 2.3: Log Transformation of Features

Apart from the changes described above, we determined ordinal factors in the dataset and changed their levels into numerical values. For example, variable HeatingQC originally has 5 levels: Po, Fa, TA, Gd and Ex. It is obvious that these 5 levels are ordinal since ‘Excellent’ is the most desirable, and ‘Poor’ is the least desirable. So we change those 5 levels into integers 1, 2, 3, 4, 5, which have the same explanatory and predictive power as before, but has less degree of freedom. Table 2.1 includes all the factors that we changed.

Factors	**Original Levels**	**Integers**
ExterQual	Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5
ExterCond	Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5
HeatingQC	Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5
KitchenQual	Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5
BsmtQual	None, Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5, 6
BsmtCond	None, Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5, 6
FireplaceQu	None, Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5, 6
GarageQual	None, Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5, 6
GarageCond	None, Po, Fa, TA, Gd, Ex	1, 2, 3, 4, 5, 6
BsmtExposure	None, No, Mn, Av, Gd	1, 2, 3, 4, 5
Functional	Sal,Sev,Maj2,Maj1,Mod,Min2,Min1,Typ	1, 2, 3, 4, 5, 6, 7, 8
GarageFinish	None, Unf, RFn, Fin	1, 2, 3, 4
PavedDrive	N, P, Y	1, 2, 3
PoolQC	None, Fa, TA, Gd, Ex	1, 2, 3, 4, 5
Fence	None, MnWw, GdWo, MnPrv, GdPrv	1, 2, 3, 4, 5

Table 2.1: Factor Change Table

Lastly, categorical feature columns ‘Condition1’ and ‘Condition2’ are essentially representing the same feature of the house. Manually combining them before they are transformed into dummy variables would reduce the degree of freedom. The same solution also works for feature columns ‘Exterior1st’ and ‘Exterior2nd’.

Chapter 3

Explanatory Modeling

Statistical modeling is generally categorized into three spectrums, based on the motivation and intention. Explanatory modeling is usually applied if the target is testing potential causal relations between underlying factors measured by variables X and the assumed underlying effect measured by variable Y. Predictive modeling, which will be discussed in detail in the next chapter, emphasizes on making predictions for variables X that are not available at the time of fitting the model. The last one is descriptive modeling, which is aimed at summarizing the data structure. Fitting a regression model can be descriptive if the main idea is capturing the association between variables rather than inference or prediction.

3.1 Definition of Explanatory Modeling

In this chapter, we will focus on Explanatory modeling. Given that the main target is not making predictions, all the data entries will make their way to the fitting of the model, and there is no hold-out dataset. Due to the fact that explanatory modeling cares the most about interpretability, the fitted model cannot be biased, otherwise, it could not be statistically explained or theoretically verified.

Considering the different models in linear regression, ordinary least square would be the most suitable model for explanatory modeling. Recall the fact that the Gauss-Markov Theorem proved the least square solution is the best linear unbiased estimator (BLUE). As for other regression models such as Ridge and Lasso, they are biased because of the tuning parameter and the penalty term in the model. However, Lasso could potentially be used for variable selection during the phase of feature engineering, as long as the final model will be fitted with ordinary least square.

3.2 Full Model Diagnositcs

In the first trial, we tried to put all variables in an Ordinary Least Square model. Unsurprisingly, the model is not working well. From figure 3.1 we can see problems in several folds.

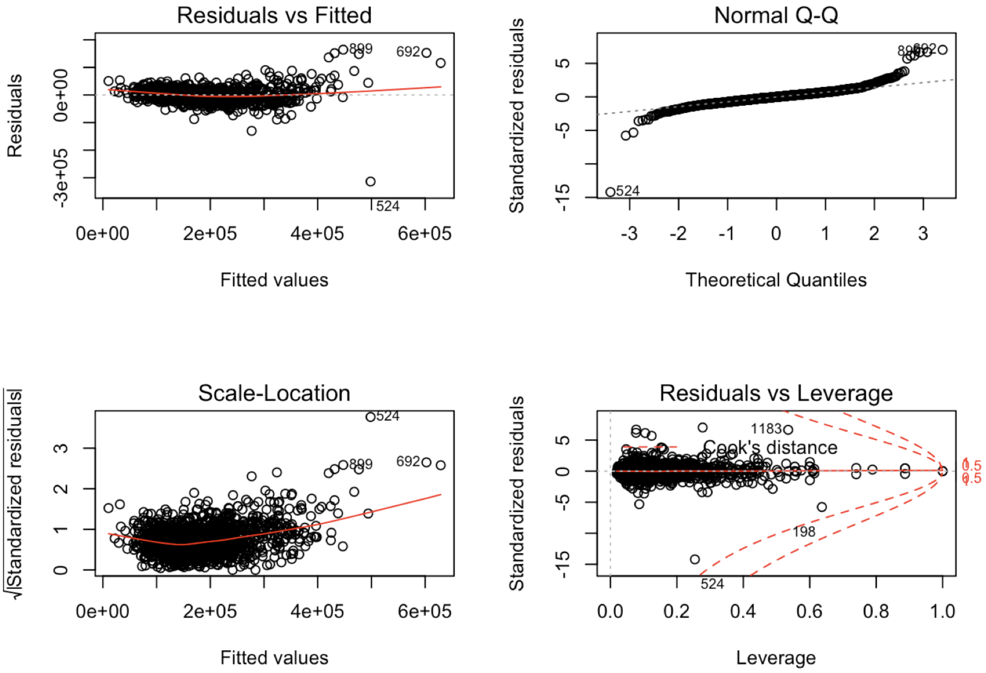


FIGURE 3.1: Diagnosis of Full Model

First, the model suffers from collinearity. For example, totalBsmtSF and BldgtypeDuplexs parameters would be turned into NA because R detected collinearity automatically. Besides that, we think there will be more cases of multicollinearity to be spotted. For example, LotArea should be related with 1stFlrSF and 2ndFlrSF; Heating (a type of heat) should be closely related with YearBuilt.

Second, homogeneity of variance doesnt hold in this estimation. When the fitted value is smaller, the residual has a 'taller' distribution; while the distribution gets 'wider' when the fitted value gets larger.

Third, as we can see in Normal Q-Q, while this model gives a satisfying model in the middle part, it was biased at the head and tail quantiles of the data. This means we should conduct xx test to see if we should adapt logistic form or square form of a dependent variable.

Fourth, there are outliers and influential points that should be removed. Given the above problems in full model, we proposed several feature-selected models.

3.3 Feature Selection

We used three methods to select features: Lasso, forward selection and backward selection.

LASSO (least absolute shrinkage and selection operator) is a regression model widely used in variable selection. Through minimizing the sum of squared error with a limited maximum value on the sum of coefficients' absolute value, it has an outstanding performance in variable selection and shrinkage. The number of variables to select can be easily calculated. However, the process of selecting variables is not interpretable.

Forward selection and backward selection are both stepwise regression. The first begins with no variable, and add variables that have the greatest statistical significance until the stopping rule is satisfied. The later begins with the full model and remove variables that has the lowest statistical significance until stop rule is satisfied. There are several stop rules that we can use in the forward/backward selection. In this model, we applied AIC (Akaike Information Criterion) failing to decease as stop rules. Compared with AIC, BIC (Bayesian Information Criterion) tends to have larger size penalty, thus works better on small subsets. Since forward selection and backward selection is very similar with each other, we will only introduce backward selection result here.

3.3.1 Lasso

When mean squared error drop to the lowest point, LASSO selected 96 variables from about 200 total features (with λ equals 600.03).

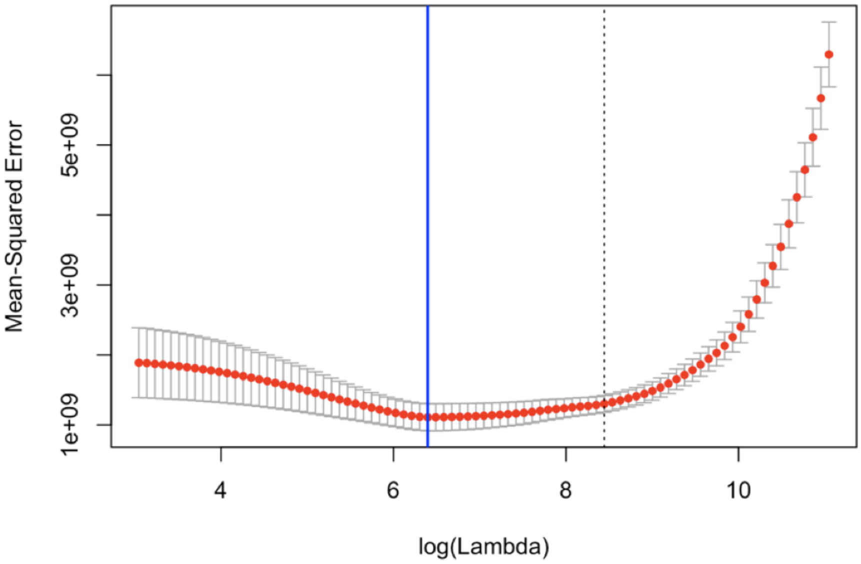


Figure 3.2: Parameter Selection for Lasso

Those features are related to the basement, living area, house decoration, style and age, functional areas, neighborhood features and others. In the lasso model that we run, best λ is about 600.

We then used DFFITS plot 3.3 to filter influential points in the dataset.

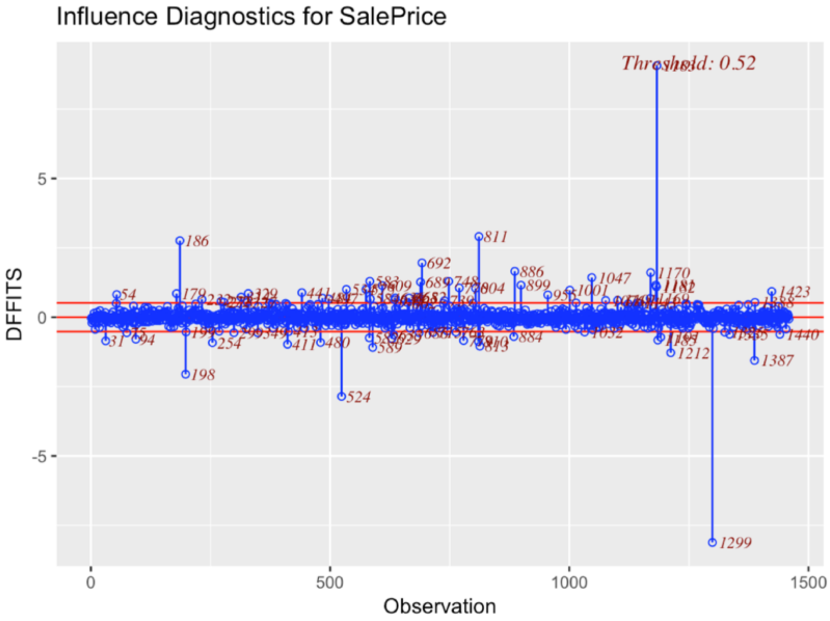


FIGURE 3.3: Influence Diagnositcs for SalePrice

After filtering the influential points, we rerun the model for box-Cox test and KS-test. The p-value of KS-test is larger than 0.05, which means residual of y is normally distributed. Although boxcox shows that λ is very near to 0, indicating that we should take the log or square root form of y.

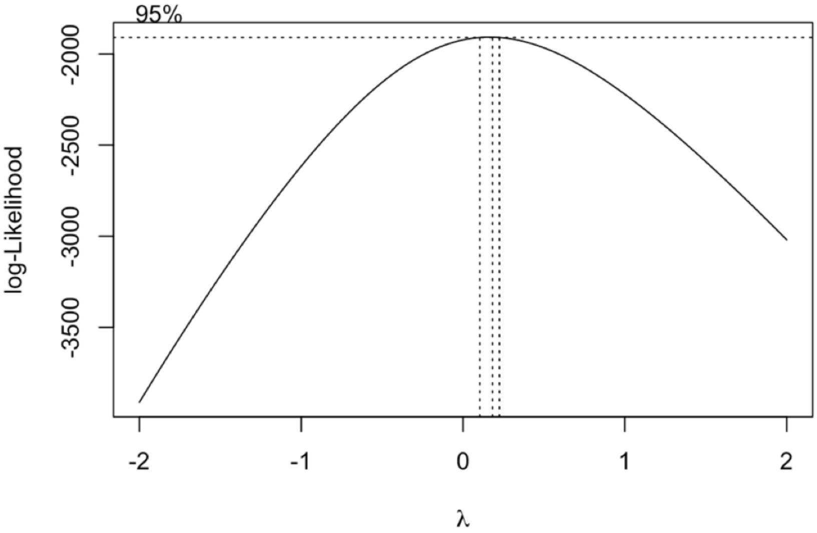


FIGURE 3.4: Box-Cox Test

Although many features can significantly affect the price of the house, there are still variables automatically ignored by R, which means there are multicollinearity issues in this model, which is not picked out by LASSO method.

The last diagram 3.5 is the diagnosis plot of this model. As we can see, the residual has a distribution of heteroscedasticity.

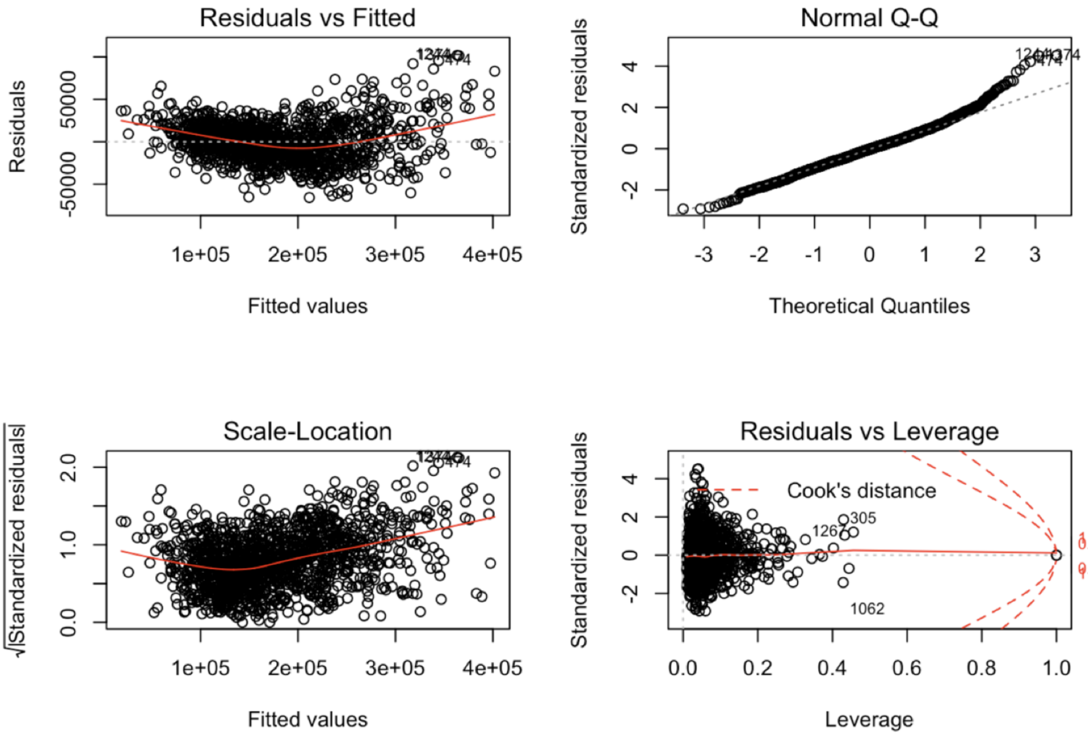


FIGURE 3.5: Diagnosis of Lasso

After that, we tried to replace y with $log(y)$. However, the result is not satisfying.

3.3.2 Backward Selection

Backward selection has the best performance among the three methods. Firstly, we used normal y as dependent variables. It selected 109 features. After the selection, we made an OLS estimation to filter the influential points in DFFITS diagram 3.6. As what we can see, there are approximately 100 data points that have a threshold higher than 0.55.

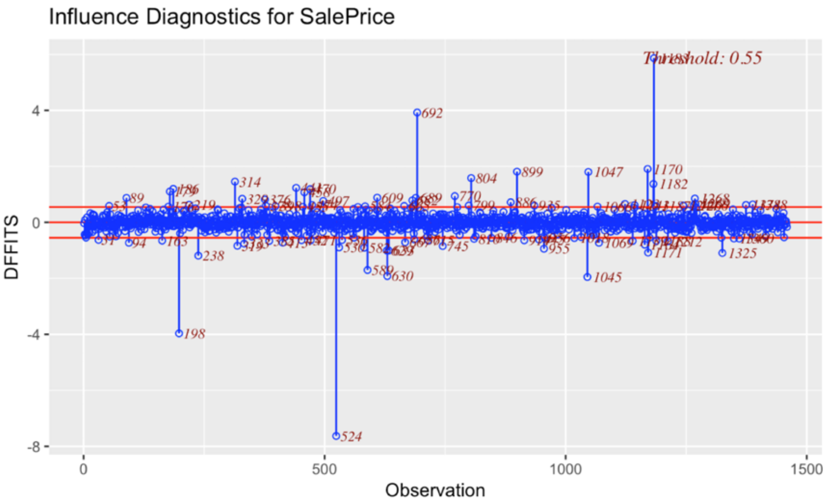


FIGURE 3.6: Influence Diagnositcs for SalePrice

After getting rid of those points, we run another OLS estimation. We calculated the average value of 100 KS-test's p-value, and it's around 0.1. However, if we look at the normal Q-Q diagram, it is still not a strict line.

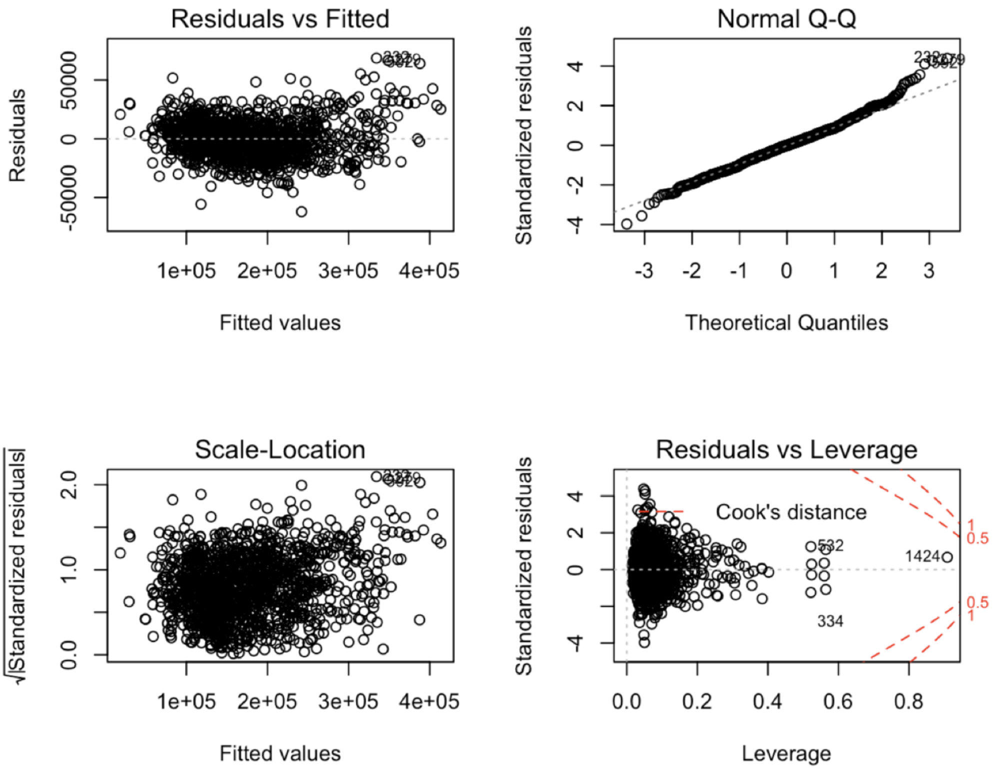


FIGURE 3.7: Diagnosis of Backward Selection without Log Transformation

According to Box-Cox test, we think it is a good idea to transform y into $log(y)$[1].

[1] If $y \leq 0$, then $log(y + 1)$ will work better. Since price of house is not likely to be 0, we adopted $log(y)$

We repeated the above process with $log(y)$. The average of 100 KS-test is about 0.35 in this situation, which means the residual is normally distributed. Diagnositic plot 3.8 tells the same story. In the diagram on the left-up corner, the residual is similarly distributed around different fitted values. In the diagram on the right-up corner, normal Q-Q is a beautiful line, which means the transformation of y works good.

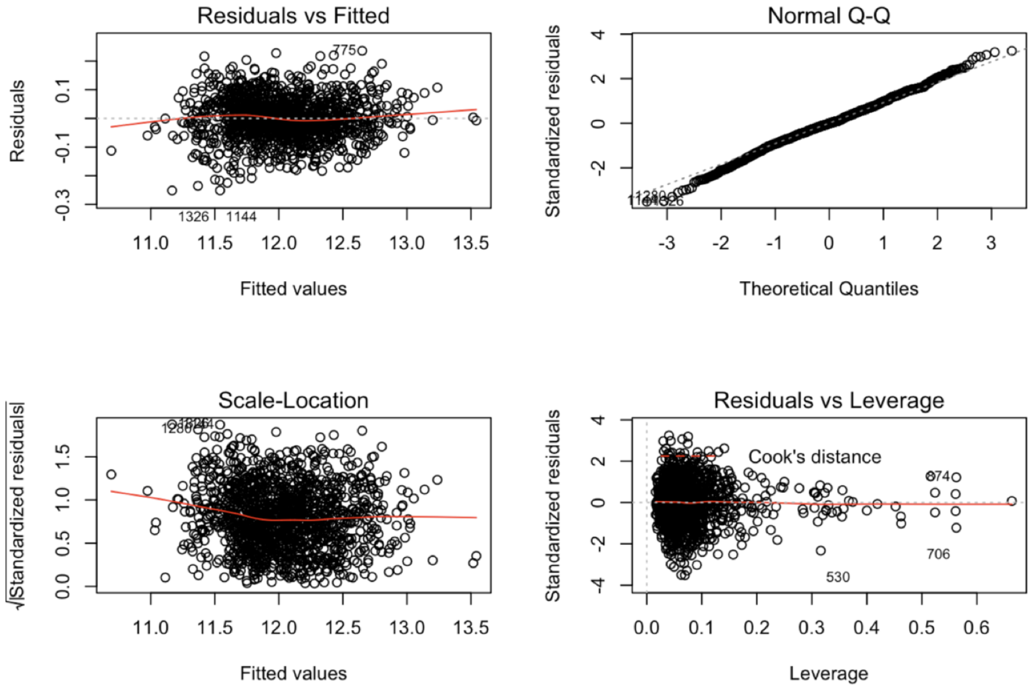


FIGURE 3.8: Diagnosis of Backward Selection with Log Transformation

3.4 Summary of the Models

	LASSO	**Backward Selection**
# of variables selected	96	109
# of observations filtered out	88	99
Boxcox result	close to 0	close to 0
KS test result	didn't pass	pass
Take $log(y)$ is better	no	yes
NA in parameters	some	no
Homoscedasticity	no	yes
Shape of Normal Q-Q	not a line	a nearly perfect line

TABLE 3.1: Advantages and Disadtanges of LASSO/Backward Selection

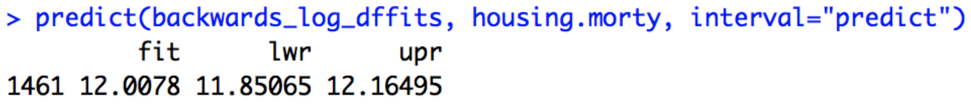
As what we have stated, the full model has all sorts of problems in explaining the price of the house. Thus, we used LASSO, forward and backward to select rightful features for OLS regression model. Since forward and backward selection is very similar to each other, we only introduced backward selections result, who has better performance in Mean Squared Error (MSE).

3.5 Case Study: Exploration of Home Value

To test the reliability of our explanatory model, we have tested the model on a new observation. Morty's house had been evaluated by another firm, stating that his house will sell for $143,000. We attempted to evaluate the same house, which was then compared with the price given by that firm. Moreover, a few recommendations will be provided to potentially promote the value of the house.

3.5.1 Predictive Interval

Morty's data was cleaned in the same manner as described in chapter 2. It follows that the 95% prediction interval is given in figure 3.9.



```
> predict(backwards_log_dffits, housing.morty, interval="predict")
          fit      lwr      upr
1461 12.0078 11.85065 12.16495
```

FIGURE 3.9: Pretictive Interval for Morty's House

Given that we are estimating the log-transformed value of the sale price, our predicted sale price should be:

$$\begin{aligned} Sale\_Price &= e^{fit} = e^{12.0078} = 164029.2 \\ Low\_Price &= e^{lwr} = e^{11.85065} = 140175.4 \\ High\_Price &= e^{upr} = e^{12.16495} = 191942.3 \end{aligned} \tag{3.1}$$

Based on our estimation, Morty's house was under-valued by the other agency. We believe Morty can sell his house at a price as high as $191,942.

3.5.2 Improvement Recommendations

In terms of improvement recommendations, we have considered following criteria:

- The coefficient of the feature in the model should be statistically significant.
- The coefficient should also be positively correlated with the sale price.
- The feature should be changeable. We consider features such as the area of the house to be a intrinsic property of the house, which is unrealistic to be changed.

The 3 most important features to be changed are:

- Roof Material: change from ‘Standard (Composite) Shingle’ to ‘Wood Shingles’.
- Exterior covering the house: change from ‘Vinyl Siding’ to ‘Brick Face’.
- Overall condition: this can be potentially improved by refurbishing.

Chapter 4

Predictive Modeling

In the previous chapter, we introduced three types of statistical modeling and then how explanatory model could be constructed on the housing price data. Furthermore, we looked into Morty's case by calculating a prediction interval based on the explanatory model. Now we have changed the motivation of modeling and would like to figure out how to maximize the accuracy of making predictions.

4.1 Definition of Predictive Modeling

In prediction modeling, the only ultimate goal is improving the quality of the model in terms of result prediction. To be more specific, we need to minimize the mean square prediction error of the predicted housing prices with respect to the housing dataset. Having said that, unlike explanatory modeling, the dataset should be split into two sections — one for training, and the other for testing. As a result, a random selection of the data entries (30%) went to the test set and the remaining are for model fitting.

Since interpretability is not the major concern in prediction modeling, the regression methods that could be employed here are no longer constrained to ordinary least square only. Regardless of the bias-ness, the method achieving a minimum mean square prediction error wins.

4.2 Principal Component Analysis

Given the difference in the motivation of modeling, we no longer use the features selected in the previous chapters for prediction modeling. Instead, we will start from the place where the data has been cleaned and transformed appropriately as discussed in chapter 2. Based on previous analysis, we know that potentially there are some outliers and influential points in the dataset. However, we are not going to remove these data points in the prediction modeling because these kinds of anomaly situations may happen when making predictions, and we don't want to see our models fail to produce sensible predictions for similar cases. What we do want to handle is the collinearity issues in the dataset.

Apart from Ridge Regression, there are many other techniques could be utilized to tackle multicollinearity, and one of them is principal component analysis (PCA). PCA is a statistical procedure that conducts an orthogonal linear transformation to transform the data into a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on.

After PCA transformation, the number of features won't change, but each new feature becomes a linear combination of old features. Since new features are ordered decreasingly with respect to the variance explained by that feature, we can select the top N new features such that the total variance explained is larger than a threshold. Essentially, this is the other reason why we use principal component analysis. There were about 80 features in the raw data, and the number increases to over 200 when the categorical variables were transformed into dummy variables. By conducting PCA, not only the collinearity problems are addressed, but also the degree of freedom could be decreased significantly.

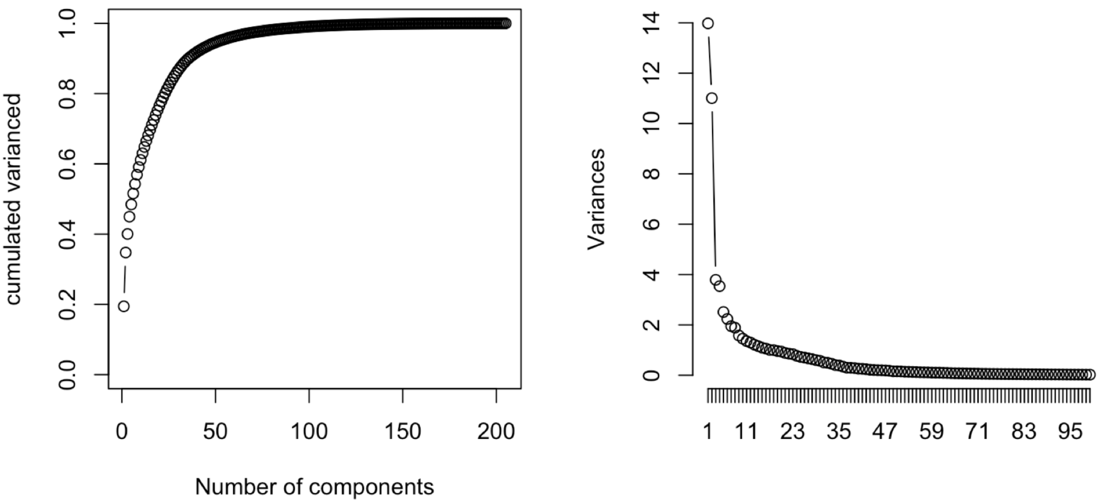


Figure 4.1: Variance Explained By PCA

In the case of housing price, the cumulative variance being explained by PCA is shown in figure 4.1. We decided to use only the first 54 pricipal components, because over 95% of the variance could be explained by these components. Adding additional components won't significantly affect the models, but only add computational complexity to the problem.

It should be noted that the data are normalized before sent to PCA because the new features would only capture columns with large variances otherwise. In certain cases, depending on how the data were split, some features might have zero variance. This is particularly in the case of sparse dummy variables where all entries could have zero value in the training dataset. Under such circumstances, we have to drop those features in the test dataset as well. For all other dummy variables, we choose to leave them as they are, because it makes no sense to scale "boolean" variables. Another thing to point out is that the normalization of test dataset should base on the center and scale of training data.

4.3 Fitted Models

In order to find the best linear model for housing price prediction, we will consider four linear regression methods and compare the mean square prediction error (MSPE) of each. For each method, we will also discuss how to choose parameters where applicable. Based on previous analysis, we have decided to perform log transformation on the housing price. Recall that MSPE is defined as:

$$MSPE(\hat{f}) = \frac{1}{|\Theta|}\sum_{j\in\Theta}(y_j - \hat{f}(\mathbf{x}_j))^2 \tag{4.1}$$

where Θ is the set of indices that represents the test dataset, $\hat{f}$ is the fitted model, and y_j is the real label of j^{th} data entry. With respect to the case of housing price, now we have $y_j = log(price_j)$, thus we would expect the MSPE be much smaller than the case without log transformation.

4.3.1 Ordinary Least Square

In Ordinary Least Square regression model, there is no additional parameter to be selected, so we can simply fit the model with all training data. The calculated MSPE is 0.01449555 in this case.

4.3.2 Ridge Regression

In order to tune the parameter of Ridge Regression, we have conducted 10-fold cross-validation on the training set and have selected the λ resulting in minimum MSPE on the validation set.

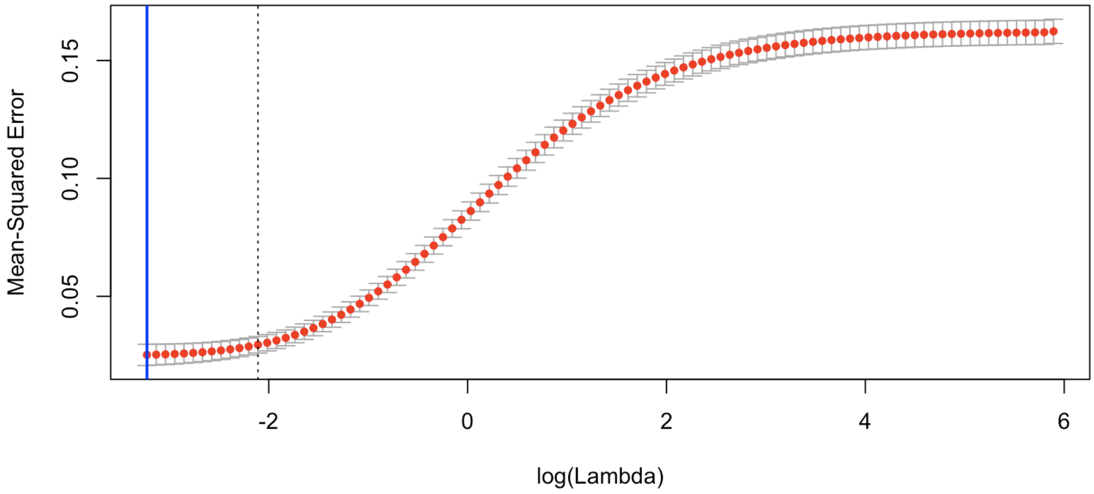


Figure 4.2: Parameter Selection for Ridge Regression

From figure 4.2 it is clear to see that the best $log(\lambda)$ is actually the smallest possible value, which means λ is close to 0. In this case, the regularization term does not have much effect on the fitted model, and the final model is quite close to OLS.

4.3.3 Lasso

The same approach was applied to parameter selection in Lasso.

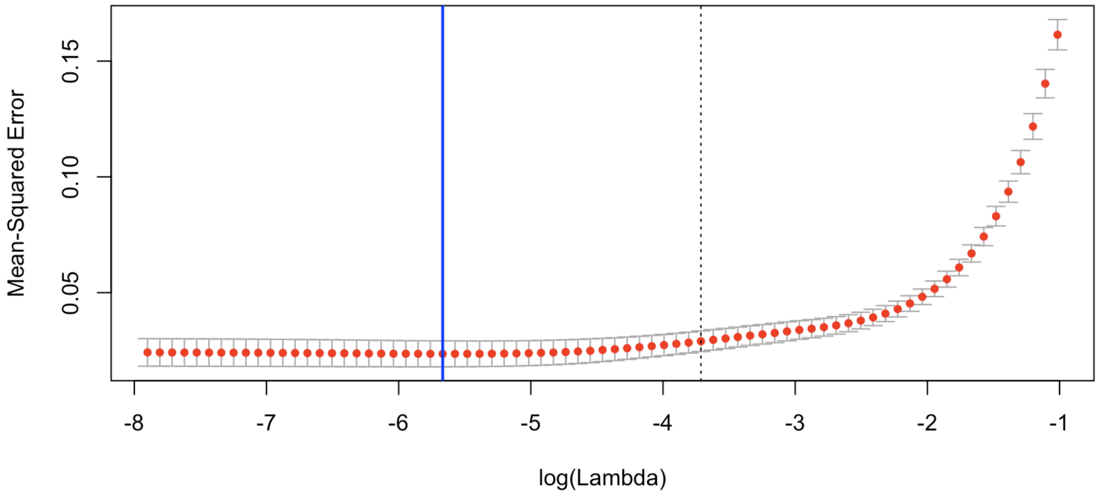


Figure 4.3: Parameter Selection for Lasso

4.3.4 Elastic Net

Elastic Net is essentially a combination of Ridge and Lasso. For simplicity, we have set $\alpha = 0.5$ and then used cross-validation to select the best λ that minimizes MSPE on the validation set.

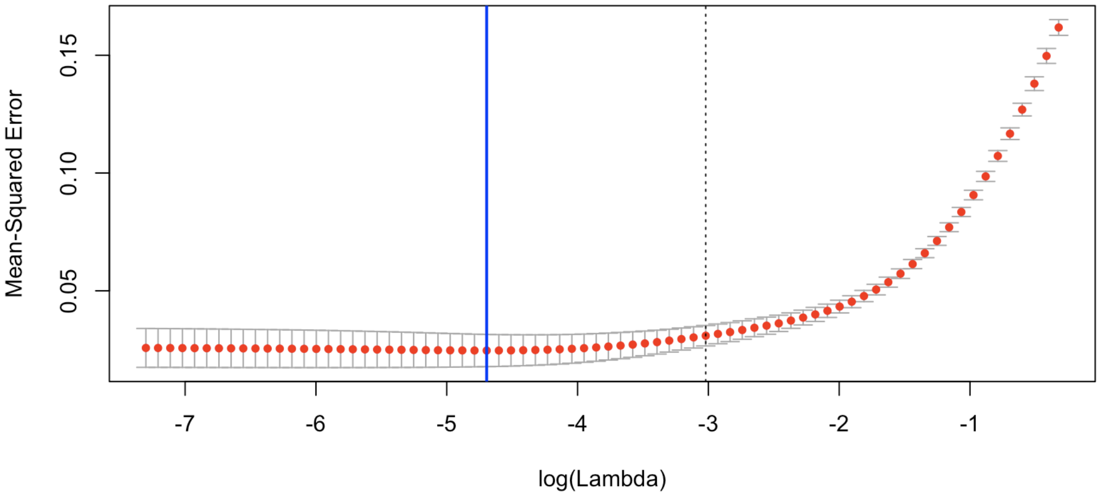


Figure 4.4: Parameter Selection for Elastic Net

4.3.5 Model Comparison

Table 4.1 is a summary of MSPE of the models described above.

Model	MSPE
OLS	0.01449555
Ridge	0.01550225
Lasso	0.01447140
Elastic Net	0.01471138

Table 4.1: Summary of MSPE

From this table, it seems that all the models have similar performance in terms of making predictions. Based on the plots of λ, Ridge, Lasso and Elastic Net all selected some small number for the regularization parameters, thus they should have similar ability level for prediction theoretically.

The similarity can be largely attributed to feature engineering and principal component analysis. Since PCA has addressed both multicollinearity issues and feature selection, penalized terms cannot have much effect on the fitting. As a result, all these models would be similar to OLS.

Chapter 5

Conclusion

In this project, we have constructed effective explanatory and predictive linear models to estimate the housing price in Iowa. In the explanatory modeling part. both Lasso and backward selection had been applied to conduct feature selection, while backward selection algorithm outperformed in all aspects.

As for the part of predictive modeling, we have employed principal component analysis to extract the most valuable information in the dataset, which helped us to fit various linear regression models.

Since this dataset is categorical-dominated, linear regression models may not be the most appropriate methods to unveil the relationship between housing price and all those features. Tentatively, one can approach the problem using tree-based models or more advanced machine learning algorithms.