California Housing Price Prediction

Huiyu Xie

April 30, 2020

Contents

1 Prerequisite

In this report, we focus on the topic of California Housing Prediction. In the data set, We are given 1 output variable and 8 input variables.

Output ($\boldsymbol{y}$)	Inputs (X)
ln(median house value)($\boldsymbol{y}$)	longitude($\boldsymbol{x_1}$), latitude($\boldsymbol{x_2}$), housing median age($\boldsymbol{x_3}$), total rooms($\boldsymbol{x_4}$), total bedrooms($\boldsymbol{x_5}$), population($\boldsymbol{x_6}$), households($\boldsymbol{x_7}$), median income($\boldsymbol{x_8}$)

Table 1: Output and Input Variables

Note that the data set contains $n = 20640$ observations on 9 variables. We randomly divide them into training data set $\mathcal{D}_{train} = \{X_{train}, \boldsymbol{y}_{train}\}$ and test data set $\mathcal{D}_{test} = \{X_{test}, \boldsymbol{y}_{test}\}$, where $n_{train} = 12384$ and $n_{test} = 8256$. We use different regression techniques to get accurate housing price prediction.

1.1 Data Characteristic

To start with, we try to describe the characteristic of data set. Here we focus on each variable to study the data characteristic.

The direct way is to draw normal QQ plot for each data set. By comparing theses QQ plots to QQ plots of some common distribution, we can easily get the rough distribution of each data set. According to kinds of distributions, we divide the data set into different groups.

The following contents conclude the characteristic of each data set.

- longitude($\boldsymbol{x_1}$), latitude($\boldsymbol{x_2}$): it seems that no specific distribution can be matched, kind of like being free distributed.
- housing median age($\boldsymbol{x_3}$): it is roughly close to uniform distribution, but with more concentration on median.
- total rooms($\boldsymbol{x_4}$), total bedrooms($\boldsymbol{x_5}$), population($\boldsymbol{x_6}$), households($\boldsymbol{x_7}$): it is closed to chi-square distribution, the distribution is skewed to the left.
- median income($\boldsymbol{x_8}$): it is roughly close to chi-square distribution, but with slight heavy tail on the right side.
- ln(median house value)($\boldsymbol{y}$): it is roughly close to normal distribution, but with slight heavy tail on the right side.

Note that in order to double check the characteristic of data set, we also plot histograms for each data set. All the histograms and QQ plots are in Appendix.

1.2 Multicollinearity Problem

Then we have to check the multicollinearity of the data set. Since the original data indicates a serve multicollinearity problem, for avoiding misjudgment, we use unit length normalization.

After standardization of data set, we can compute the eigenvalues of $X^T X$.

λ_{max}	λ_{min}
3.9159	0.0151

Table 2: Maximum and Minimum Eigenvalues

Note that for simplicity, we only show maximum and minimum eigenvalues in Table 2. Since the condition number $K < 1000$, we can declare that no serve multicollinearity is indicated.

Besides, we can also check multicollinearity by computing VIF of each input variables.

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$	$\boldsymbol{x_6}$	$\boldsymbol{x_7}$	$\boldsymbol{x_8}$
VIF	8.7112	8.8851	1.2512	12.6790	35.1399	6.9181	35.2487	1.7733

Table 3: VIF of Input Variables

Since in Table 3, all $VIF < 100$, we can get the same result that no serve multicollinearity is indicated.

2 Multiple Linear Regression

First, we try to fit the multiple linear regression model to training data set. The regression model is

$$y = \beta_0 + \sum_{i=1}^{8} \beta_i x_i + \varepsilon \tag{1}$$

After the LS fitting, we can get the value of all the parameters. Then we construct t-test to test the significance of each parameter. Since all the p-value are close to 0, we preserve all the parameters. The values of parameters are in Appendix A.

In this model, $R^2 = 0.6520$, $R^2_{adj} = 0.6518$ and $MSE = 0.1214$. Besides, we can draw R-student residuals plot and normal QQ plot of residuals.

In Figure 1, the R-student residuals are not randomly distributed around 0, instead, it seems that residuals follows a linear pattern. That is, when $\hat{y}$ becomes smaller, the r_i becomes larger. Also, in Figure 2, the distribution of residuals is not a normal distribution. It is clear the actual distribution is heavy-tailed, and it is more like a student's t-distribution.

Thus we deny the assumption that the random error terms, ε, are Gaussian i.i.d with zero mean.

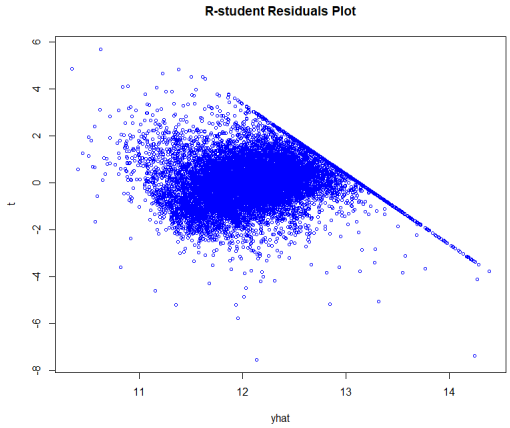

Figure 1: R-student Residuals Plot

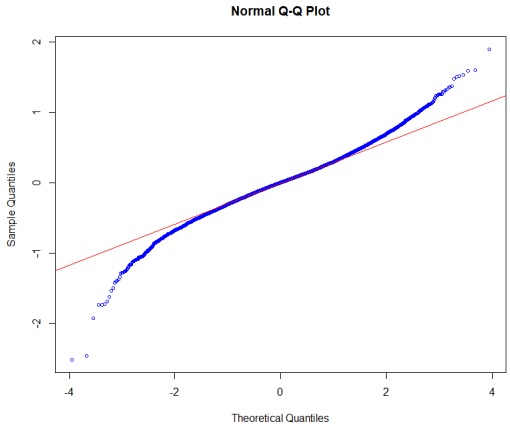

Figure 2: Normal QQ Plot

3 Polynomial Regression

Then, we try to fit the polynomial regression model to training data set. Here we fix $K = 2$ in the polynomial regression model. The regression model is

$$y = \beta_0 + \sum_{i=1}^{8} \beta_i x_i + \sum_{i=1}^{8} \beta_{ii} x_i^2 + \sum_{j=1,k=1,j<k}^{8} \beta_{jk} x_j x_k + \varepsilon \tag{2}$$

After the LS fitting, we can get the value of all the parameters. Then we construct t-test to test the significance of each parameter. Since some p-value are not close to 0, we eliminate those parameters in the model. We list them in the following table.

β_5	β_6	β_8	β_{16}	β_{33}	β_{35}	β_{38}	β_{47}	β_{56}	β_{67}	β_{68}

Table 4: Eliminated Parameters in Polynomial Regression

In this model, $R^2 = 0.7235$, $R^2_{adj} = 0.7227$ and $MSE = 0.1467$. Now we conclude the merits and demerits of using the polynomial regression model.

- Merits: consider the factors that combine effect of two or more variables, bring more functions which can be fit under it, make the prediction more precise.
- Demerits: may cause the problem of overfitting when the order is large; raise the probability of multicollinearity problem; model are quite sensitive to the outliers, which causes that few outliers in data can seriously affect results.

Then we try to fit the ridge regression model, that is, adding an L_2 regularization term to the cost function.

$$S(\beta) = \|X\beta - \boldsymbol{y}\|^2 + \lambda\|\beta\|^2 \tag{3}$$

The best choice of parameter λ is the one that makes the value of MSE to be the smallest. We can compare them through GCV. Here we get the graph of the relation

between GCV and λ. The graph is in Appendix.

We can know that MSE increases when λ becomes larger. Thus adding an L_2 regularization term to the cost function dose not help decrease MSE. The probable reason is that the value of β is not large in the original LS estimate, adding an L_2 term affect the precision of β estimate.

4 Nonlinear Regression

Then, we try to fit the polynomial regression model, namely the neural network (NN) model, to training data set.

The neural network contains input layer, output layer, and hidden layers. Each layer contains some neurons.

4.1 Introduction of Neural Network

Suppose the neural network contains n layers, and thus the input layer is the 1^{st} layer, the output layer is the n^{th} layer, the rest are hidden layers. The value of each neuron of the i^{th} layer is computed by all the neurons of $(i-1)^{th}$ layer.

Here are the equation for computation of each layer.

$$f_j^{(i+1)} = f\left(\sum_{k=1}^{m} \omega_k f_k^{(i)} + \theta_j\right) \tag{4}$$

Note that in equation (4), $f_j^{(i+1)}$ is the value of the j^{th} neuron from the $(i+1)^{th}$ layer, m is the number of neurons from the i^{th} layer, ω_k is the weight of $f_k^{(i)}$, and θ_j is the bias. Besides, f is known as the activation function.

When consider all the layers, the equation can be rewrite.

$$f(X;\theta) = f^{(n-1)}\left(\cdots f^{(3)}\left(f^{(2)}\left(f^{(1)}\left(X;W_1\right);W_2\right);W_3\right);W_{(n-1)}\right) \tag{5}$$

Note that in equation (5), $f^{(i)}$ is different from equation (4), for the reason that both input and output are vectors. Here $X = (x_1, x_2, \ldots, x_n)$, which is the combination of all input variables. Here $W_i = (w_{i1}, w_{i2}, \ldots, w_{in})$, where n is the number of neurons from the i^{th} layer. Besides, θ is the parameter in this model.

4.1.1 Reason for Nonlinear Model

Since the neural network contains several layers, even though the relation between adjacent layers is linear, after the combination of more than one hidden layers, the relation between input and output variables are nonlinear. Thus, the neural network is a kind of nonlinear model.

4.1.2 Procedure of Training Data

Then we illustrate the mechanism of the neural network. Giving the initial value of the parameter, model can give out the prediction of output, then through modifying

the value of parameter, model can get the prediction of output more precisely, thus get the final training result.

4.2 Application of Neural Network

Since it is better to normalize the data before applying model, we use min-max normalization here. By setting the layer neurons number as 8, 5, 3, 1, we can get the training result.

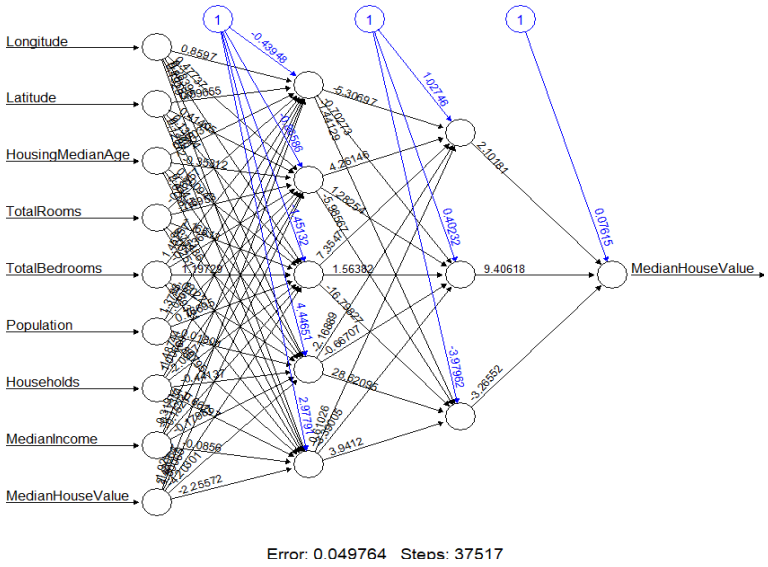

Figure 3: The Result of Neural Network Training

Note that the median house value is transformed to ln(median house value) in the training procedure.

In this model, $R^2 = 0.99997$, where R^2_{adj} is close to R^2, and $MSE = 9.9235 \times 10^{-6}$. Both R^2_{adj} and MSE are much smaller than those of linear models. Now we conclude the merits and demerits of using the polynomial regression model.

- Merits: optimize the conventional model, enable the model become robust, make the prediction much more precise.
- Demerits: much more data have to be trained in order to construct the model; the complexity of algorithm increases; costs more time to get the final model; depends on computer to solve out the model.

5 Alternating Conditional Expectation

Now we try to use alternating conditional expectation (ACE) algorithm, which applies optimal transformation on both output and input variables of the original data set.

In the ACE model, $R^2 = 0.7099$, where R^2_{adj} is close to R^2, and $MSE = 0.1493$.

We can find that the R^2_{adj} and MSE are close to multiple and polynomial regression model, but differ quite a lot with NN model.

Then we try to compare the residual plots of ACE model with those of the ordinary LS model which is constructed without performing data transformation.

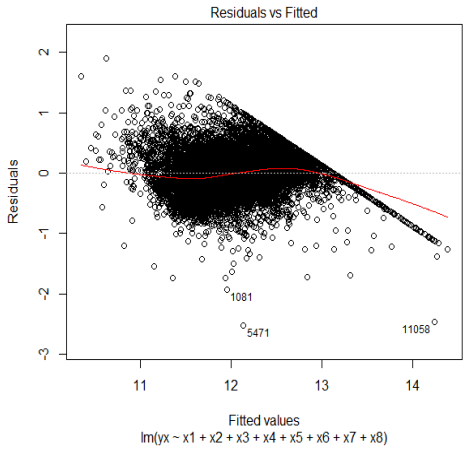

Figure 4: Residual Plot of Multiple Regression

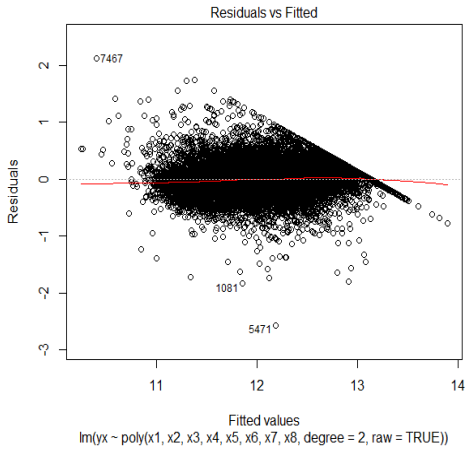

Figure 5: Residual Plot of Polynomial Regression

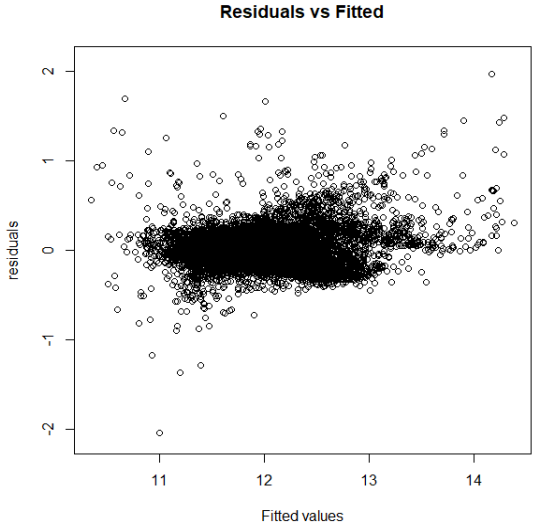

Figure 6: Residual Plot of ACE Model

In figure 4, 5 and 6, we can see that the residuals are concentrated around the line $y = 0$, and the bound line becomes vague. In this question, we can not conclude that the ACE model is superior to multiple linear regression model, for the reason that they are close in terms of performances. But we can not deny that ACE model may be superior in solving other questions. The judgment depends on case being studied.

Now we discuss the merits and demerits of using ACE model.

- Merits: optimize the conventional model , make the training and prediction precise.
- Demerits: the algorithm complexity increases; costs much more time to construct model; sensitive to outliers; depends on computer to solve out the model.

6 Outliers

In this section, we focus on the outliers in data set. Here we apply Cook's measure to training data to identify the outliers, or more precisely, influential points.

We choose the models mentioned before to help identify outliers, including multiple linear regression model and polynomial regression model. Note that we preserve all the parameters in polynomial regression here.

The following graphs show the leverage of each residual points. The left side is from multiple linear regression, and the right side is from polynomial regression.

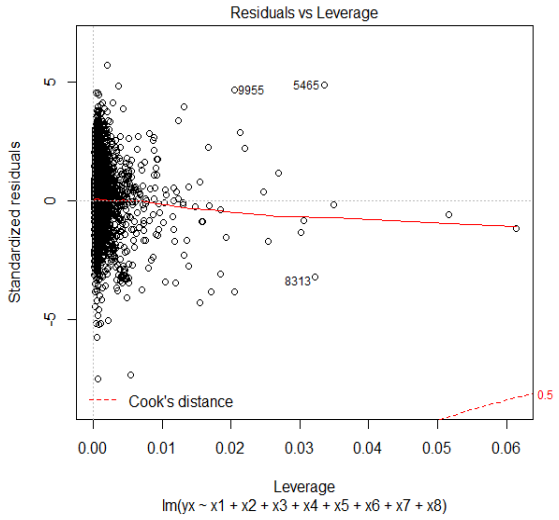

Figure 7: QQ Plots of Variables

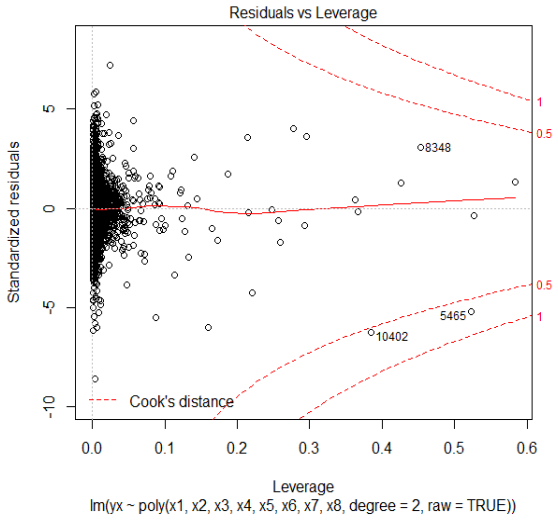

Figure 8: Histograms of Variables

In Figure 4, we can see that all the residual points are within Cook's distance, which roughly implies that no influential points in this model. In Figure 5, we can see that 2 residual points are outside Cook's distance, which implies these data points are influential in the regression procedure of this model. Also, most of the residual points are within the range (-2,2). So far, we can conclude that only a small fraction of data are outliers.

Now we discuss the way to achieve robustness against outliers. We can just follow the steps to build robust regression model. (1) delete the data points which are within Cook's distance; (2) consider of the data points which are outside Cook's distance, and think about whether they are subject to "men-made" errors, if so, then just delete, otherwise, keep them in the data set; (3) rebuild the regression model. However, we can also use other way to achieve robustness, like applying NN model and ACE model which are more complex.

7 Comparison

In this section, we try to make comparison of different regression models in terms of some kinds of indicators. In order to see the comparison visually, here we make a table for the comparison results from this question.

	Multiple Model	Polynomial Model	NN Model	ACE Model
Training Performance	✓	✓✓	✓✓✓✓✓	✓✓
Prediction Performance	✓✓	✓	✓✓✓✓✓	✓✓
Overfitting	\	✓	✓	✓
Time Complexity	✓	✓✓	✓✓✓✓	✓✓✓✓✓
Outliers Sensitivity	✓✓✓	✓✓✓	✓	✓✓✓

Table 5: Comparison of Regression Models

Note that in Table 5, more ticks means the certain model achieve more on the indicator which is labeled on the left, like better training performance, better prediction performance, consuming more time, and more sensitive to outliers. Besides, we can see that, except for multiple linear regression model, it is possible for all the other models to have the problem of overfitting.

8 Summary

In this section, we make a summary of the main findings obtained from solving the above regression models.

- In general, nonlinear models have better performances than linear models, that is, nonlinear models are more useful, which can be reflected in training and prediction performance.
- Overfitting problem exits in most regression models, remember to be avoid of making the model too complex, which can help keep from overfitting.
- Outliers problem have an effect on all the models, however, a compromise between deleting outliers and retaining outliers can leads to robust regression.
- Data normalization enable the procedure of dealing with data to be more easy, it is better to normalize the data set before applying regression.
- The training performance is not certainly related with prediction performance, due to the uncertainty of data, model with good training performance may have bad prediction performance and vice versa.
- In the above models, there exists a balance between model performance and model time complexity, that is, achieving better results needs more time to training the data.

A Appendix

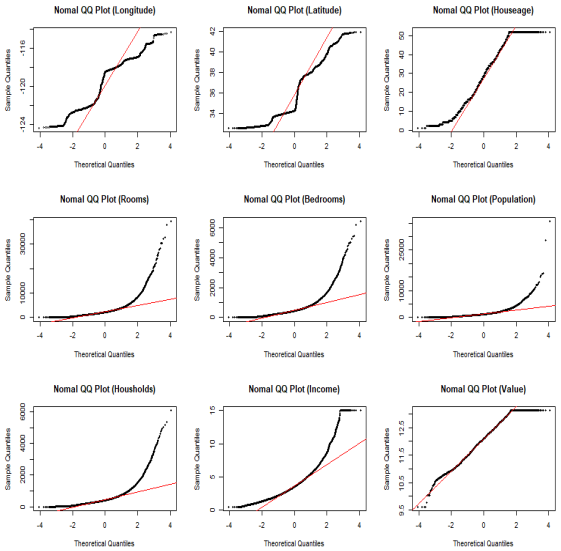

Figure 9: QQ Plots of Variables

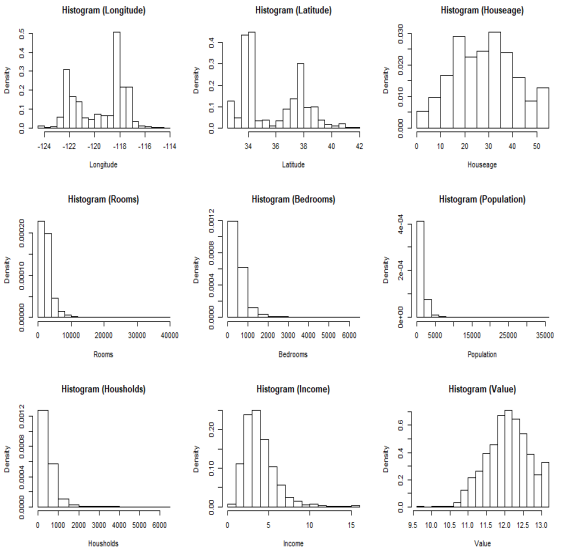

Figure 10: Histograms of Variables

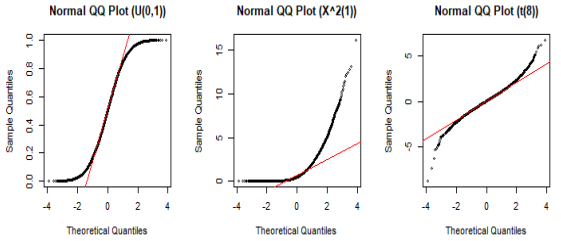

Figure 11: QQ Plots of Common Distributions

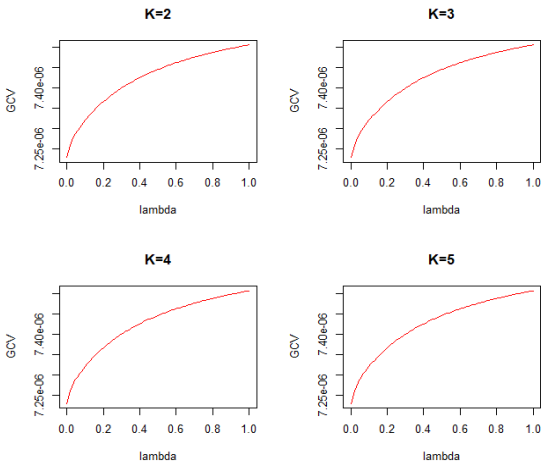

Figure 12: Relation of GCV and λ