Nha Do
University of California, Los Angeles (UCLA)
This note is created and shared for the purpose of summarizing my work and understand.
Any typos or suggestions should be emailed to nhado401@gmail.com

Linear Regression and Application in House Prices Prediction

1 Linear Regression

1.1 Mathematical Presentation

Linear Regression is an attractive, simple but quite useful model in Machine Learning. It's simply the line of best fit in which we can think of a basic math equation:

$$\hat{y} = mx + b \tag{1}$$

Where:

- x is an input feature
- m is slope
- b is y-intercept

In 2-D, it is a line, but it would be a plane in 3-D and would be more complicated shapes in higher dimension. For the simple application that I worked on, a line is enough.

In fact, we might have more than one input feature and let's consider m and x above are row vectors $\mathbf{w} = [w_1,w_2,w_3,..]$ and $\mathbf{x} = [x_1,x_2,x_3,...]$. Equation (1) can be re-written as:

$$\hat{y} = \sum_{d=1}^{D} w_d x_d + b = w^T x + b \tag{2}$$

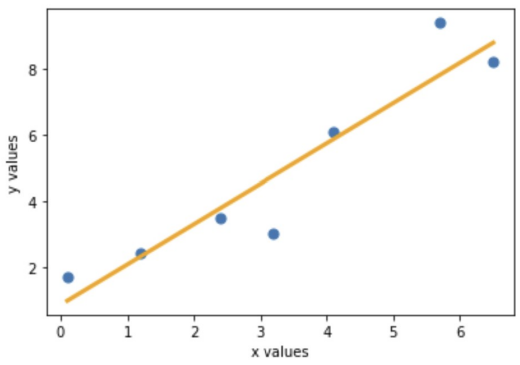


Figure 1: *A Simple Example of Best Fit Line*

1.2 Loss Function

Loss function or cost function is the difference between the prediction and the actual value. Obviously, we always want to minimize this and if the error is 0 which means the prediction and the true target are exactly correct.

The loss function that is used here is called Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y - \hat{y}\right)^2 \tag{3}$$

Where:

- n is the number of samples or data points
- y is the actual value
- $\hat{y}$ is the prediction

We divide by the number of sample for the convenience of calculation and take a square since the difference between y and $\hat{y}$ could be negative. Our final goal is to optimize vector **w** such that MSE is minimal.

The interesting here is why don't we use the absolute value? We know that the simple way to find the root of an optimization problem is to solve the equation of the derivative equals to 0. That answers the question above. The square function is differentiable everywhere while the absolute function is discontinuous at 0.

1.3 Data Transformation

Most real world dataset are non-linear, but in some cases we can turn it into a model which is "likely" linear. For example, if we have a set of data points which is distributed like this:

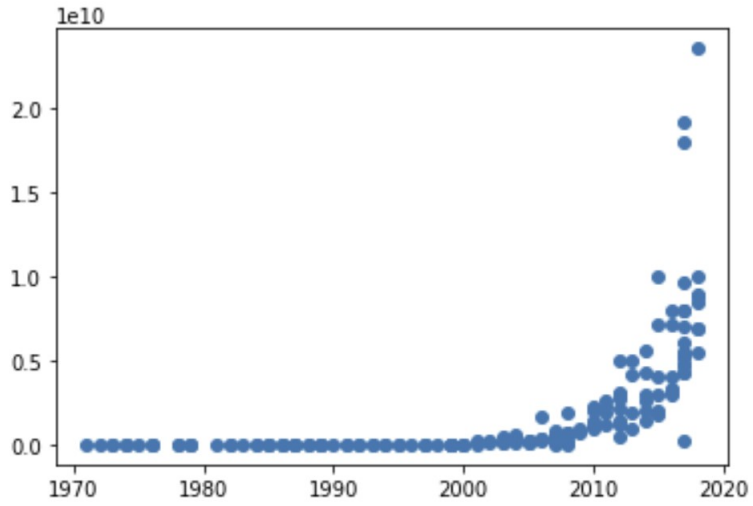


Figure 2: *Number of Transistors in IC (Moore's Law) - Source: lazyprogrammer'sGithub[1]*

We don't tend to get a good fit with these points by a straight line, but if we transform this data using the log transformation and the new data points will lie up in a straight line.

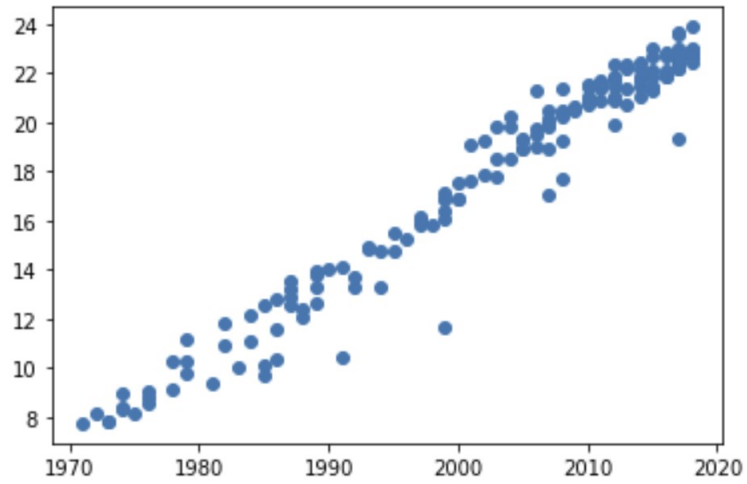


Figure 3: *Log Transformation - Source: lazyprogrammer'sGithub[1]*

This is simply taking the log value on all the data points and by doing so, the linear regression model would be good to be applied.

2 House Prices Prediction

2.1 Data

Scikit-learn is a free software machine learning library for Python and we can take advantage of this for our linear regression model. With that being said, we won't need to calculate any gradients manually but the library itself does that for us. We can also load the training data set from Scikit-learn library for practice. The dataset that I used is from load_boston which was collected in 1970s. The prices are different now because of inflation, but we can always convert it into today's dollar relatively.

The dataset for the first couple of rows looks like:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	PRICE
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

Figure 4: *Boston Dataset*

2.2 Data Analysis

2.2.1 Correlation and The Problem of Multicollinearity

Correlation is a statistical measure that expresses the relationship between two variables. In statistic, correlation is between -1 and 1 in which -1 represents for perfect negative correlation and 1 represents for perfect positive correlation.

$$-1 \leq \rho_{XY} \leq +1 \tag{4}$$

When it comes to data exploration, any values close to 1 will make a strong movement to predicted value. However, the closer value of correlation to 1 or -1, the more careful that we have to inspect because it could be a sign of multicollinearity problem. An independent variable which is very highly correlated to one or more other variables could imply a relative large standard error. Another word, it's difficult to see the individual contribution and one of them could be redundant.

To determine which values are redundant, we can do a test by calculating p-value and drop the columns that are not statistically significant (> 0.05). In this case, they are INDUS and AGE.

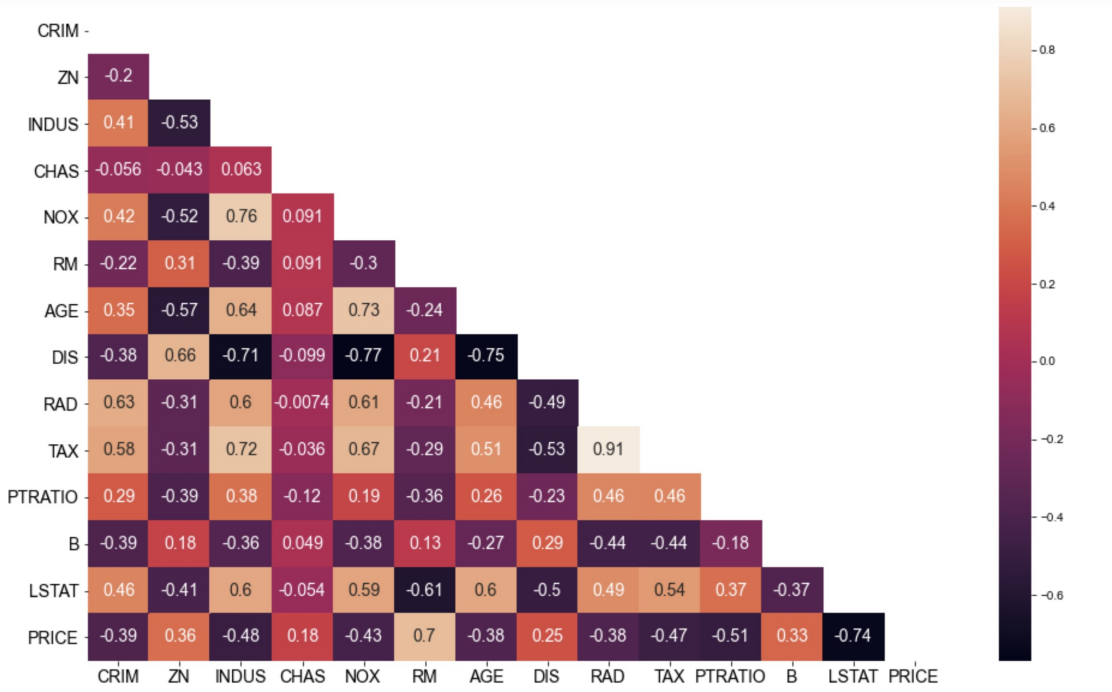


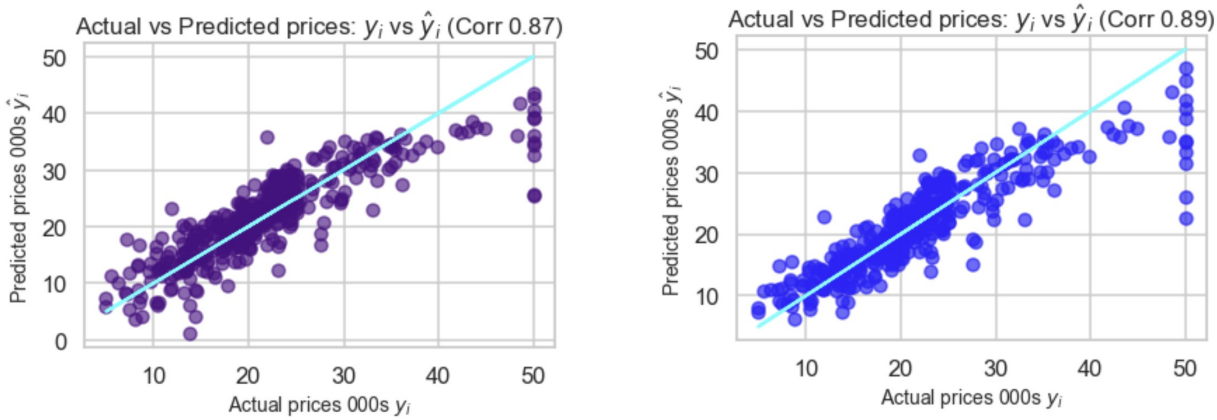
Figure 5: *Correlation Across Different Variables*

	coef	p-value
const	4.059944	0.000
CRIM	-0.010672	0.000
ZN	0.001579	0.009
INDUS	0.002030	0.445
CHAS	0.080331	0.038
NOX	-0.704068	0.000
RM	0.073404	0.000
AGE	0.000763	0.209
DIS	-0.047633	0.000
RAD	0.014565	0.000
TAX	-0.000645	0.000
PTRATIO	-0.034795	0.000
B	0.000516	0.000
LSTAT	-0.031390	0.000

Figure 6: *Testing on P values*

2.2.2 Log Transformation on Linear Regression

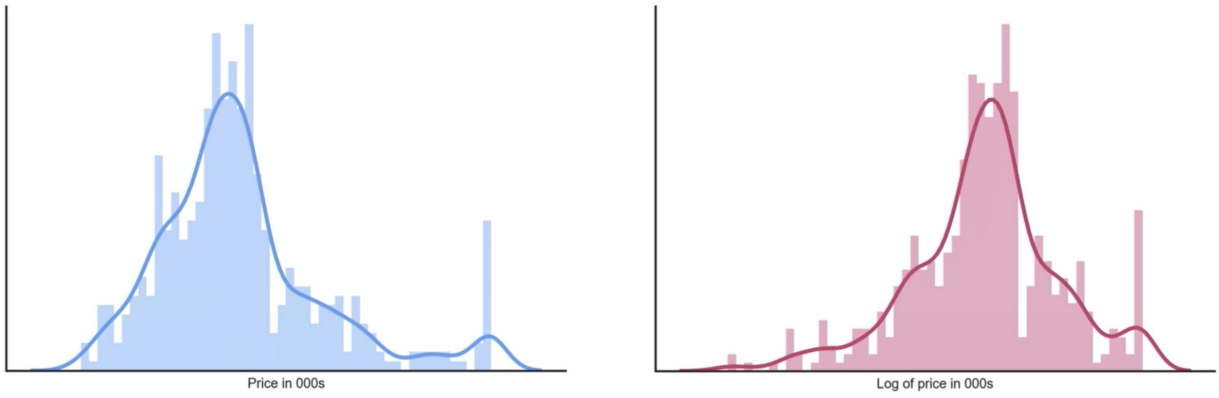
By taking the log price and dropping 2 input features mentioned above, we can compare how good the model has been improved.



(a) *Normal Price* (b) *Log Price*

Figure 7: *Normal Price vs Log Price*

Another good thing about log transformation is that it can transform the model closer to the normal distribution (Skew = 0). Another word, it will be more symmetrical. That happens in this dataset because of the skewness in one end of the plot, which causes the skew = 1.1 on the original dataset but reduced to -0.33 by Log Transformation.



(a) *Skew = 1.1* (b) *Skew = -0.33*

Figure 8: *Skewness*

3 Discussion

3.1 Advantages

Linear Regression Model is simple to implement and easy to understand.

As long as the model is linear or "likely" linear, this algorithm is the best because of the less complexity compared to other Machine Learning Algorithms.

There are some other simple techniques, like the data transformation above, can be used to improve the model.

3.2 Disadvantages

Linear Regression Model is very noise sensitive. If there exists any data points located far outside the fit line, the error will be very large. If we look at the plots above, there are outliers occurred at one end, which cause huge negative effects on the regression.

Most real world dataset are non-linear, therefore, assuming the model is linear or try to make it "likely" linear is impossible and dangerous.

4 References

[1]Lazyprogrammer's Github.$https://github.com/lazyprogrammer/machine_learning_examples$

[2]Tiep Vu's blog.$https://machinelearningcoban.com/2016/12/28/linearregression/$