The Chinese University of Hong Kong, Shenzhen

MAT 2040

Report for Project

Author:
Zhang Qihang

Student Number:
119010434

December 29, 2020

1.Introduction

The purpose of this project is to build a model to predict the relationship between time and price. We use three different models to predict. All data is divided into two groups. The first part of the data is used to train the program and get the coefficients of the prediction model. The second part of the data needs to be compared with the results of the fitting model, and finally the corresponding error is obtained. According to the selected standard, we can determine the optimal model, and use the best model to further forecast the five companies' products price.

2.Theory

The first part of the idea is to use the constructed matrix and use the least square method to get the corresponding coefficient matrix, the most important formula is:

$$Ax = b$$

According to the principle of least square method, the best result of the above formula can be obtained by the following formula:

$$A^T Ax = A^T b$$

Therefore, the result of the coefficient matrix required for the prediction result is:

$$x = (A^T A)^{-1} A^T b$$

After getting the matrix used for prediction, we can calculate the data we predicted based on x that based on the test data set. But on this basis, we have two different prediction methods.

In the first method, after predicting the data for each day, we will use the real data (values from the test data set) of the day as the data for subsequent predictions. The corresponding error calculated by this method will be smaller, but because of this This method has lost the meaning of future prediction, so we abandoned this prediction method.

In the second method, we will use the forecast data every day as the data used for subsequent forecasts, so that we can ensure that the entire process is predicted by us. However, the error value corresponding to this method may be larger, so we need to choose the appropriate N to reduce the error.

In order to reduce the error, we have selected three different models, and obtained the optimal model we need by comparing the prediction errors of each model.

The first model is a simple linear regression model. The selected N represents that the data of each day may be related to the previous few days. If the selected N is too small, it may cause the fitting curve to be insufficiently accurate and the difference between the prediction value and the actual value is too large. If the selected N is too large, the result may be over-fitted. The fit curve fits the training data to a high degree, but the difference between the prediction value and the test value is too large.

Selection of good or bad fitting standard

We choose MSE to quantify how well a model fits, and choose the model with the smallest MSE compared with the test data and the corresponding N as the best model.

According to the calculation method of MSE:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Here y_i and $\hat{y}_i$ respectively represent the predicted value and actual value.

Through this formula, we can easily obtain the best model needed by comparing the sizes.

Model 1:Autoregressive model

In this model, the matrix required for training and prediction is:

$$A = [X(t-N) \quad .... \quad X(t-1)]$$

The size of matrix A is (200-N)×N

$$b = [y_{t-N} \quad .... \quad y_{t-1}]\hat{}(-1)$$

The size of matrix b is N×1

By formula:

$$x = (A^T A)^{-1} A^T b$$

We can get the coefficient matrix and use the coefficient matrix to predict.

Then we can calculate the MSE corresponding to all N and choose the most suitable N through comparison.

Model 2:Fourier series

Model 2 is similar to Model 1, but the expected result of x will be changed to:

$$x = [a_1 \quad b_1 \quad \dots \quad a_N \quad b_N]\hat{}(-1)$$

The corresponding size is 2N×1. On this basis, we also need to make corresponding modifications to A and b, and finally get a process similar to model 1 and get MSE.

Model 3:Taylor formula

Model 3 is calculated similarly to Model 1 and 2, so I won’t repeat it.

After obtaining the results according to MSE, we can determine which model corresponds best. After we have selected the model, proceed to part2.

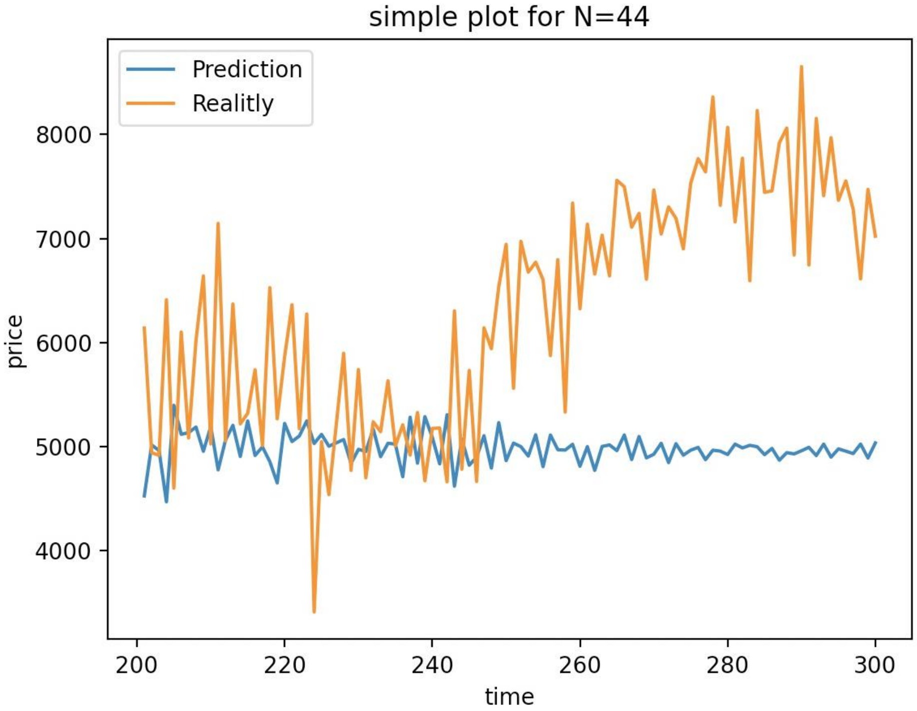
3.Data

According to the conclusion of model 1, we can find that the calculated MSE corresponding to each different N is:

```
[[3168255.1326226834, 44], [3210691.265273016, 40], [3227799.1627488034, 41],
[3232772.9698220454, 42], [3249769.1811348726, 55], [3251419.7214652584, 43],
[3278911.5649563195, 57], [3283857.1286048396, 39], [3312720.2731343806, 56],
[3316465.403011871, 58], [3369210.6153152348, 38], [3411042.0827080254, 47],
[3412668.403119707, 54], [3422737.779861139, 46], [3433020.0846895683, 37],
[3436493.370986925, 45], [3442134.381895902, 53], [3447532.796582937, 48],
[3473758.0123654185, 66], [3488028.0155638065, 34], [3492574.833153273, 59],
[3506267.49696847, 67], [3512889.44967273, 36], [3514720.7261529462, 51],
[3515235.0651288377, 60], [3523593.6384670483, 35], [3551195.4091445184, 65],
[3586713.5517389523, 49], [3602678.536336328, 52], [3628888.162538811, 50],
[3686512.2254987983, 62], [3701768.989143751, 68], [3730125.8736318117, 64],
[3733524.115002407, 69], [3739809.5824964275, 61], [3770433.1367127215, 33],
[3774345.4342960697, 63], [3791832.5511140297, 32], [3951374.017707773, 31],
[4086664.3518431825, 30], [4129669.2196861766, 70], [4228539.419556303, 17],
[4232501.322568756, 73], [4239669.278153608, 72], [4249786.873095131, 71],
[4279831.75413931, 29], [4355817.743685637, 74], [4476383.879949895, 16],
[4477827.48315162, 20], [4495417.601515149, 76], [4561914.896115844, 21],
[4596394.954701358, 27], [4608272.580161787, 18], [4609024.546339747, 13],
[4628621.24082852, 75], [4649384.32061314, 28], [4678214.294536189, 23],
[4699997.166446224, 22], [4711863.23775047, 24], [4726954.364782227, 11],
[4752887.406950365, 77], [4759692.728625653, 19], [4766623.7766917255, 14],
[4771349.669214875, 15], [4783076.579730525, 25], [4811562.251491701, 12],
[4879389.5792252, 26], [4881579.880078711, 10], [4948543.787703514, 9],
[5044990.509693963, 78], [5275601.770733861, 8], [5439515.06673597, 79],
[5455244.131582919, 7], [5579035.747035053, 80], [5650702.613280505, 4],
[5918196.338362854, 5], [6035382.555755359, 84], [6190239.3346680775, 86],
[6648502.250353942, 3], [6901988.641367326, 82], [6945678.320233654, 6],
[7002084.570262433, 85], [7125227.901055637, 81], [7301976.698628206, 83],
[8173991.404497831, 88], [8719272.8767945, 87], [10168202.461333519, 2],
[12391048.256924393, 90], [21278674.080923676, 89], [23112771.588239733, 1],
[44276338.154654324, 91], [338649612.5045533, 95], [634319646.7812119, 92],
[717744770.9694191, 94], [924016306.6847337, 93], [2471873682.337752, 96],
[1608112427507.4182, 97], [10631058599090.334, 98], [10799561123488.627, 99],
[668287815102119.2, 100]]
```

According to the above data, we can clearly see that the fitting result we get is closest to the true value when N=44

Similarly, we can draw the image when N=44:



Similarly, we can get the MSE corresponding to the results fitted by model 2 and model 3.

Model 2: MSE=13005064.520469453

At this time t=300, N=5

Model 3: MSE=8206075.631570362

At this time N=3

So we need to use the model 1 corresponding to the smallest MSE, N=44 as the parameters of the model we need to proceed to the second part.

4.Conclusion

We can find that the image fits well at the beginning and deviates from the true value continuously. This shows that our model is not perfect enough. We cannot rely on the data of the first few days to infer all the subsequent data, but if we can The real value is added to the modeling process, and the model is constantly revised, and we can get a better model.