Faculty of EEMCS

Fundamentals of Data Analytics Assignment 1

Authors :
Yiran Liu 4519140 Y.Liu-25@student.tudelft.nl
Lu Dai 4506677 L.Dai-1@student.tudelft.nl
Manoj Krishnaraj 4485742 m.krishnaraj@student.tudelft.nl

May 1, 2016

Indroduction

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. Organisation and description of data is the subject area of descriptive statistics, and how to draw conclusions from data is the subject of statistical inference.

Usually, a nice approach to describe and analyze data is to generate model. In this situation, Bias - Variance dilemma is of importance to be discussed to find the best data model. The following formula shows that mean square error equals the sum of squared bias and variance, which is used for evaluating model and analyzing Bias-variance dilemma in this work.

$$\mathrm{E}_T[(f(\mathbf{x}) - \hat{f}(\mathbf{x}\,|\,T))^2] = (f(\mathbf{x}) - \mathrm{E}_T[\hat{f}(\mathbf{x}\,|\,T)])^2 + \mathrm{E}_T[(\hat{f}(\mathbf{x}\,|\,T) - \mathrm{E}_T[\hat{f}(\mathbf{x}\,|\,T)])^2]$$

Experiment Set Up

The best way to test the effect of regression is to apply functions that are not easy to predict. That's why we need to generate proper datasets in order to perfom simulation. If we use the function like F(x) = x^2, it will probably be fitted by quadratic regression so we will be not able to compare it with other regression functions. Based on these considerations, we define our function as:

F(x) = ax^2+bx+xcos(x) (a = 0.6, b = 0.3)

Next, on purpose of making more proper sampling data, we added Gauss Noise to the original function points, using the following formula:

$$Y \sim N(\mu = F(X), \sigma^2 = 1)$$

In the experiments, we generate dataset between 0 to 10 with different sample sizes 10,100,1000. Then performing linear, quadratic, cubic, 4-th power and 5-th power regressions with different samples respectively. We want to show the fitting models and analyze how they perform on predicting from 10 to 30. To make the experiment results more intuitive, different levels of regressions are separated by different colors, and results data are demonstrated in the table for comparing expected value, bias, variance and mean square error.

Experiment Results

1. Fitting

Table 1 shows the fitting results with detailed values of squared bias, variance and mean square error during different simulations with data sample (from 0 to 10). A more detailed fitting situation (average of all fittings and variance) can be found in the appendix.

Sample Size	10			100			1000		
Metric	Squared Bias	Variance	Error	Squared Bias	Variance	Error	Squared Bias	Variance	Error
Linear	16.926	4.64	21.566	16.783	0.332	17.115	16.774	0.002	16.777
Quadratic	13.433	19.816	33.248	11.76	0.362	12.121	11.754	0.003	11.757
Cubic	5.771	22.941	28.712	4.618	0.251	4.868	4.593	0.004	4.597
4-th Power	41.504	249.822	291.326	4.61	0.461	5.071	4.543	0.005	4.548
5-th Power	4.486	3922.75	3927.236	0.443	0.114	0.557	0.438	0.006	0.444

Table 1

It can be easily observed that:

5-th power regression always have the smallest bias while linear regression are usaully tend to have bigger bias.

With relative large sample size, 5-th power regression has the smallest mean square error.

The bias value does not show obvious changes for linear, quadratic and cubic regression models.

The variance value can be largely reduced with larger sample size. And it always shows decreasing trend with increasing of size of data sample.

When the sample size turns to 100 or even larger, the value of squared bias, variance and mean square error don’t change a lot.

When the sample size is 10, the value of squared bias, variance and error of different regressions are huge and of big differences. This may be caused by the fact that there are only 10 sample points available for different regressions. The existing of gaussian white noise is also a reasonable explaination to this situation.

2. Prediction

The four figures below show the simulations with data sample size of 10,100,1000 and 10000 respectively. Each figure contains five levels of regression. From the figures, we can not only see their fitting situation and but also observe their prediction performance on the later range (10 to 30).

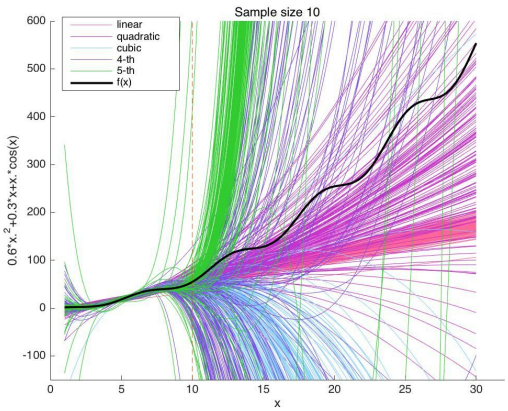

Figure 1

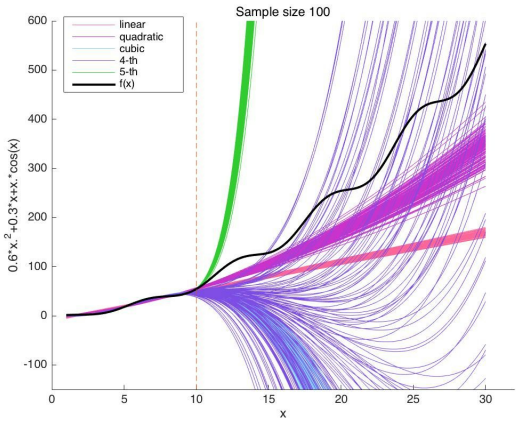

Figure 2

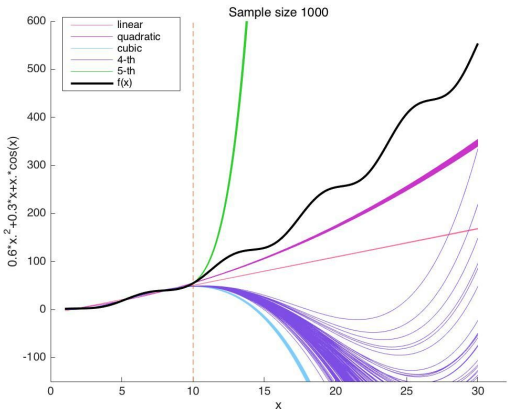

Figure 3

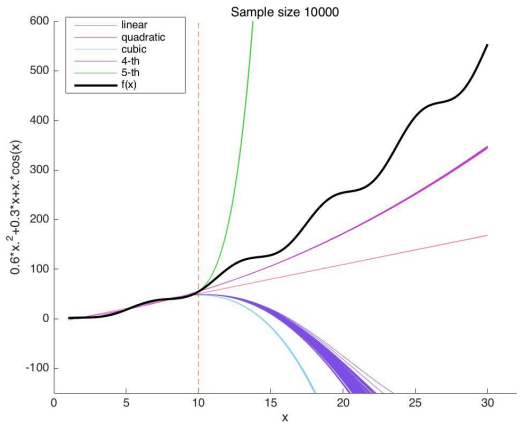

Figure 4

Some trends show up from the figures:
With larger sample size the regression curves are more intensive which indicate lower variance.
Among them, the 4-th power regression always have the biggest variance comparing with other regression levels.
The qudratic regression shows the best fittings to the original function. Especially with the sample size of 10, some quadratic regression fitting curves are almost the same trend with the original curves. This may due to the reason that we adopt the quadratic formula *ax^2* as part of our function which has strong influence to the regression. The xcos(x) part in original function makes for the differences between the quadratic regression and the original function.
Through comparing all 4 figures above, we can draw the conclusion that the regressions become more concentrated by the increasing of sample sizes.

Conclusions

In this assignment, we adopt Matlab as our tool, define a function to generate data samples, then perfom simulation on sample sets and finally anaylze the results. Generally, the result of the experiment shows that the 5-th power regression model can fit the sample data with lowest error while the quadratic regression grab the best prediction performance. The good performance of quadratic regression can be explained as the existing of quadratic part in our original function. From this assignment, we have obtained a better understanding of bias, variance dilemma and how to find the regression which can best fit the function. We all believe that this assignment will help us a lot in the futher study of Foundamentals Data Analytics.