Simulation Example on Designing and Assessing a Regression Function

Soh Ornella Lucresse

May 18, 2020

0.1 Introduction

Nowadays, adressing problems such as regression, classification have became a big challenge due to the availability of real data. Then, simulated data seems to be the appropriate way of addressing such challenge. This method is presented as a way for designing and assessing regression models.
However, this work aims to use Monte Carlo on simulated data from binormal distribution in order to investigate the variation of the model with respect to the training dataset's size.

0.2 Simulated data from binormal distribution

1. **Small sample of 10 observation**
   In this section, we start by simulating a small training dataset from the binormal distribution, then fit it into a linear model. Fitting the model allowed us to make prediction and compute the Mean Squared Error (MSE) using the Residual Sum Squared(RSS). Finally in the same graph, we plotted the data, linear model and the best regression function(also known as the conditional expectation of a bivariate normal). The visualization led us to the following figure describing these three differents components.

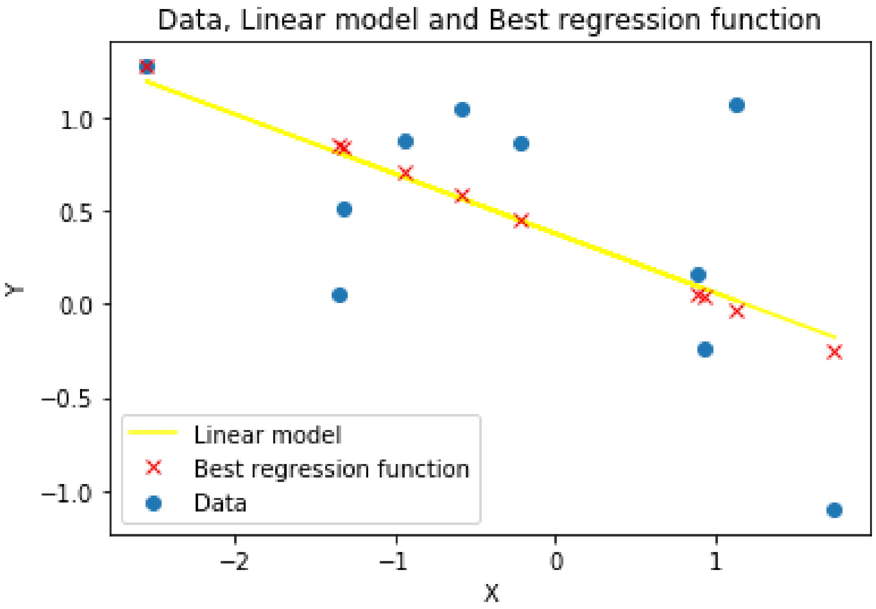

Figure 1: Plot of the linear model and the best regression function

   Observation
   From figure 1, we observed that the linear model does not fit well the data and this can be seen as an **underfitting** problem because we have high variance. However, with data generated from the best regression function, the linear model fit them well.

2. **Large sample of 1000 observations**
   Next, we generate a large training dataset from the same distribution and calculate both the true error rate and the performance of best regression function. It turns out that, as expected, the mean squared error obtained with the best regression function is smaller than the one obtained with the linear model.

0.3 Monte Carlo Simulation

Monte Carlo simulator in defined as an important technique to visualize most or all of the potential outcomes to have a better idea regarding the risk of a decision. To simulate Monte Carlo, we simulated a 500 training dataset of 10 observations each and a large testing data only once using the binormal distrution with mean 0 and unit variance. Then we plot all of the 500 linear models from the 500 training datasets and also the best regression function from the test dataset as shown in figure 2.

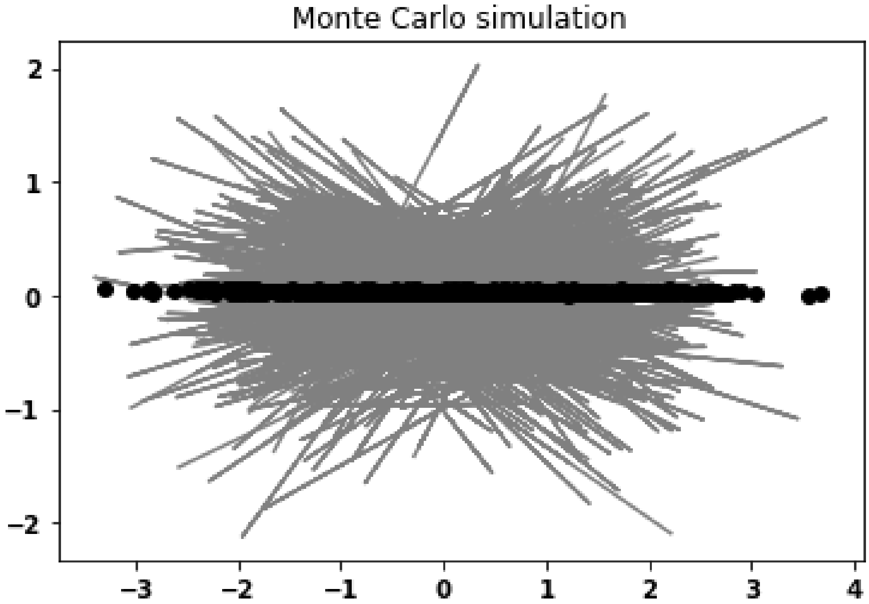

Figure 2: Monte Carlo Simulation

In order to see the effect of the training dataset size on the models, we apply Monte Carlo simulation 10 more times by varying the size of data then compute the mean and variance of the errors. For each trained model, error is obtained by taking the mean of the 500 computed error. In figure 3, we are visualizing the error of the best regression, the training error and finally the variance of the error with respect to the size of the data.

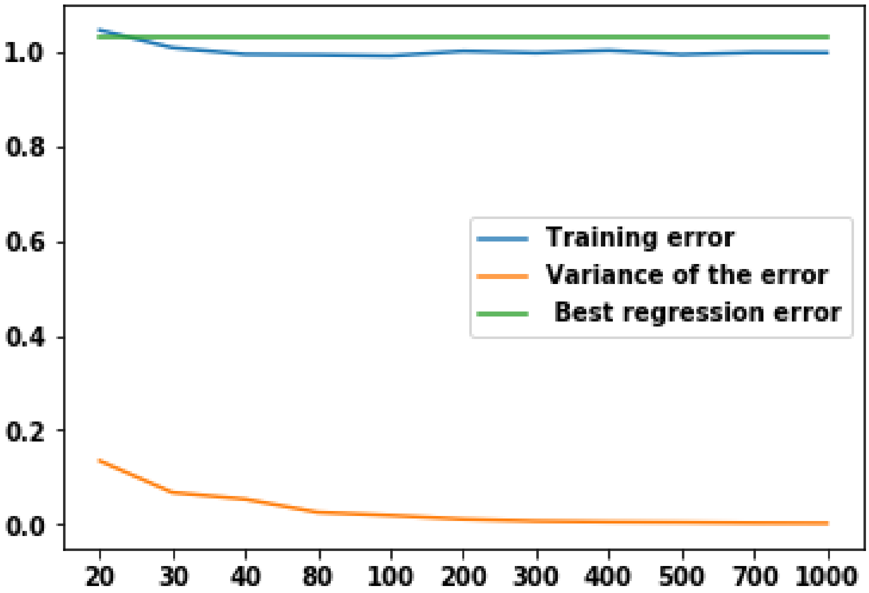

Figure 3: Error visualization

Observation
From figure 3 we can observe that, as the size of training data increase the variance of the error decrease significantly to 0 which implies that the model's error tend to be concentrated around 1. For the test dataset, the error is constant.

0.4 Conclusion

Through this work, we were able to apply Monte Carlo in order to design and assess regression models by simulating from binormal distribution. Then, study the effect of the data size through the error of the models.