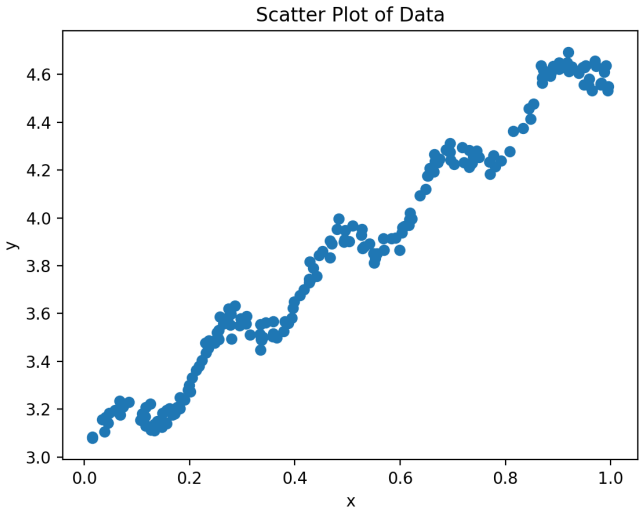
2 Linear Regression Model Fitting: Written Portion

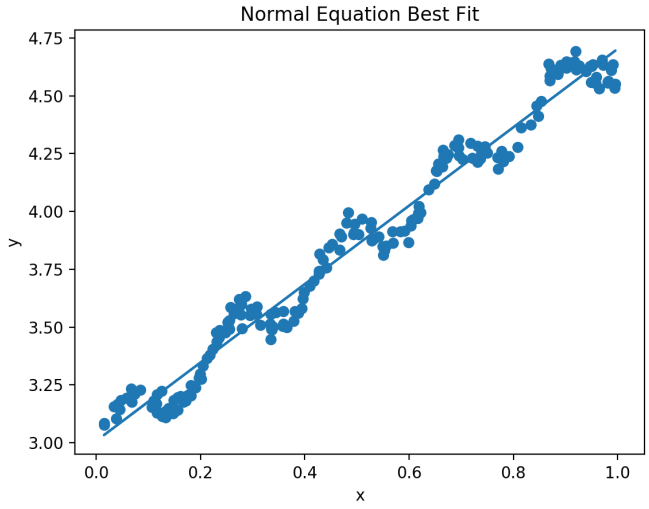
1. Scatter Plot of Data



2. Regression Models Discussion

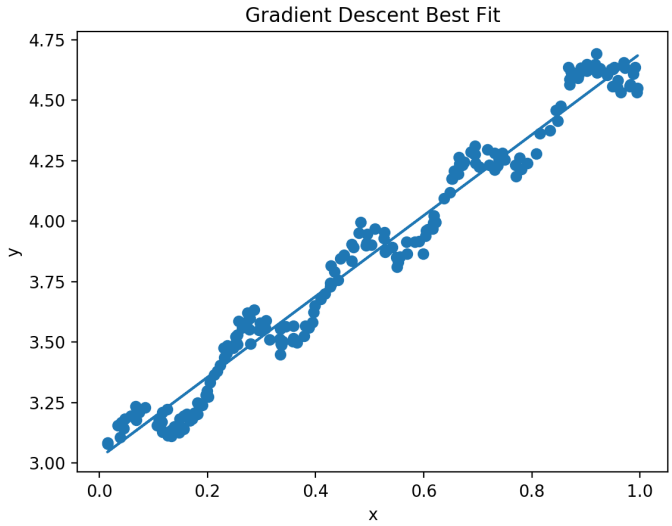
2.1 Normal Equation

concrete value of the derived theta : [3.00774324, 1.69532264]



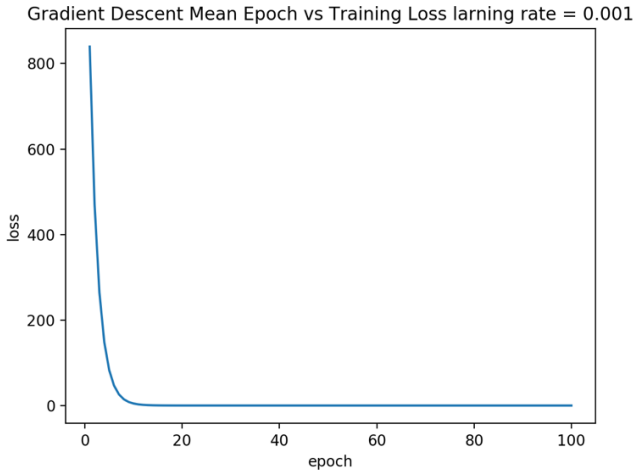
2.2 Gradient Descent

concrete value of the derived theta: [3.02023057, 1.67148924]

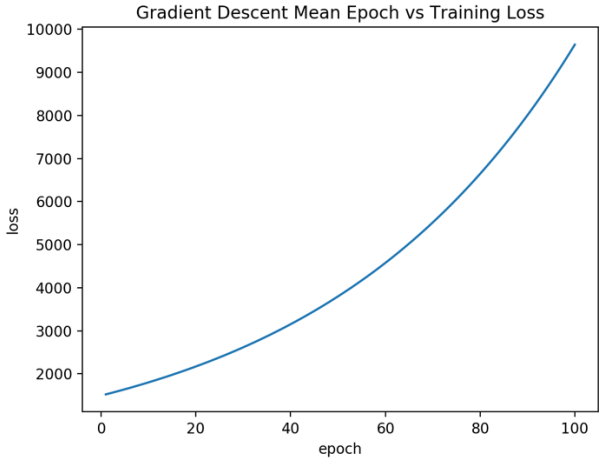


When the learning rate grows from 0.001 to 0.007, the model reaches the best performance with less epochs.

However, when the learning rate equals to or bigger than 0.008, the loss will increase with iterations. Because the loss function is a quadratic function, too large learning rate will make the descent process exceed the low loss point and be farther away from the minimum point in an iteration. Thus, the absolute value of gradient will be larger. In the next iteration, it will go to a point which is farther away from the minimum point than the last iteration.



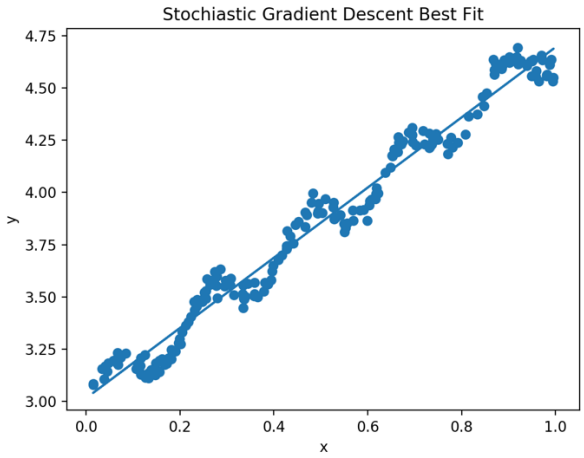
Gradient Descent, learning rate = 0.001



Gradient Descent, learning rate = 0.008

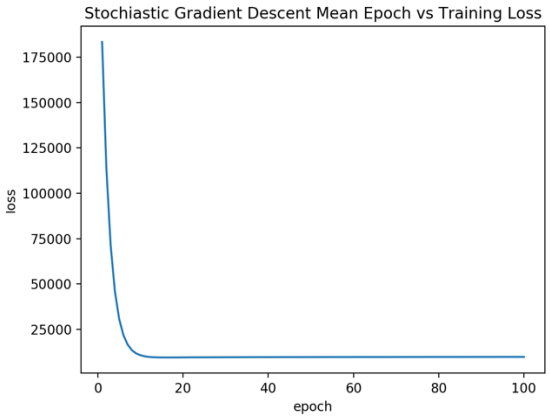
2.3 Stochiastic Gradient Descent

concrete value of the derived theta: [3.02035715, 1.67155705]

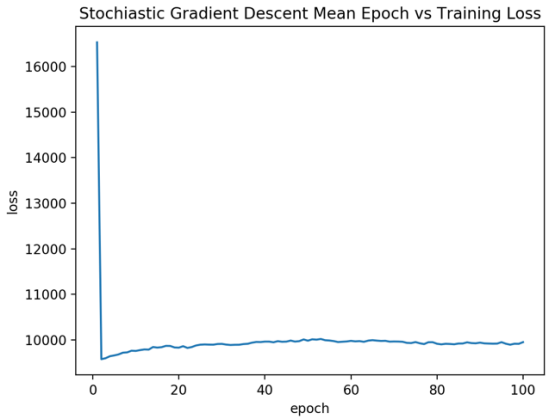


When learning rate increases from 0.001 to 0.003, the loss decreases faster, however, when it grows to a relatively large number, the loss starts to fluctuate and increase.

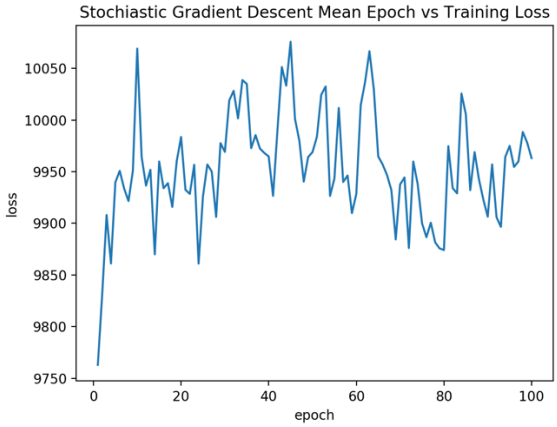
Too large learning rate will make the model exceed the best performance point. Comparing to Gradient descent, Stochiastic Gradient Descent's gradient function is related to the sampled data at each iteration, so it tends to fluctuate at large learning rates.



Stochiastic Gradient Descent, Learning rate = 0.001



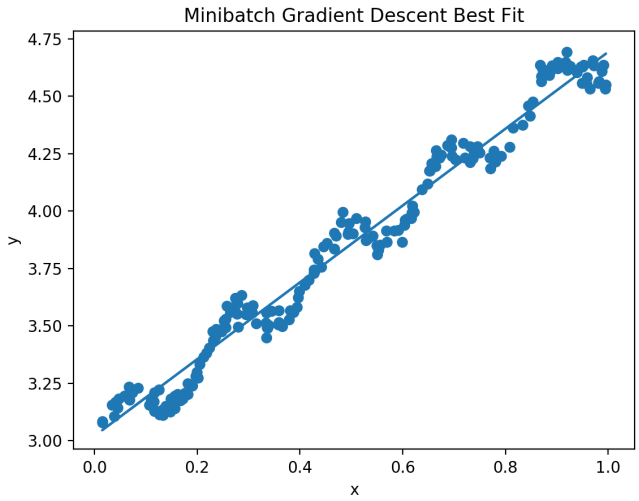
Stochiastic Gradient Descent, Learning rate = 0.007



Stochiastic Gradient Descent, Learning rate = 0.02

2.4 MiniSGD

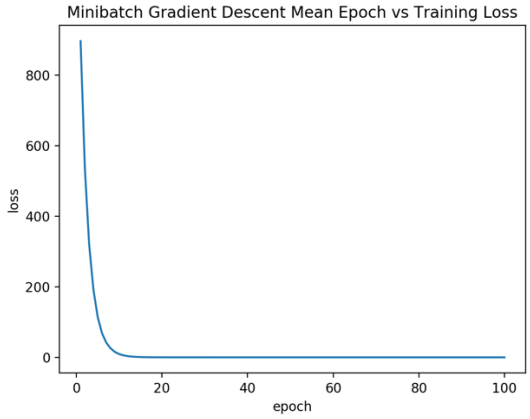
concrete value of the derived theta: [3.02395477, 1.65672719]



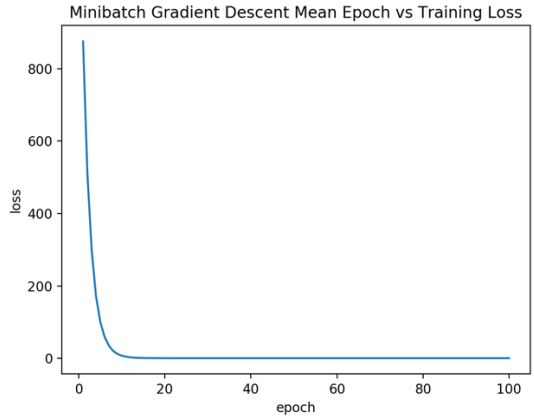
When using a batch size of 20, loss decreases faster when using larger learning rate until learning rate reaches a relatively large rate. Then the loss starts to fluctuate and increase.

When using the same effective learning rate, for instance 0.001, increasing the batch size (batch size < 200) will make the loss decreases and model converges faster.

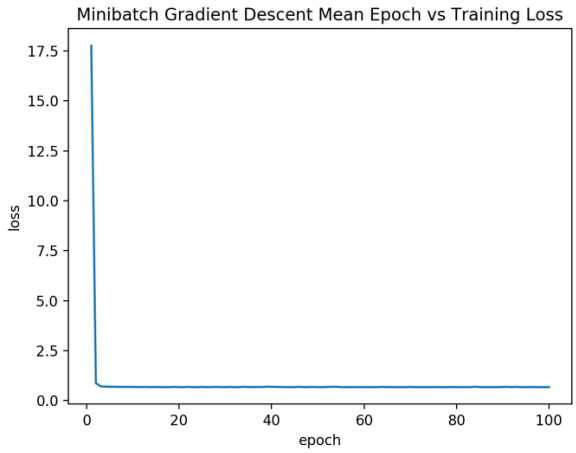
When using a larger learning rate, like 0.008, a smaller batch size (like 20) will be effective. However, larger batch size will make the loss fluctuate and increase, since increasing the batch size will make the descent more like gradient descent that uses the whole training set as sample.



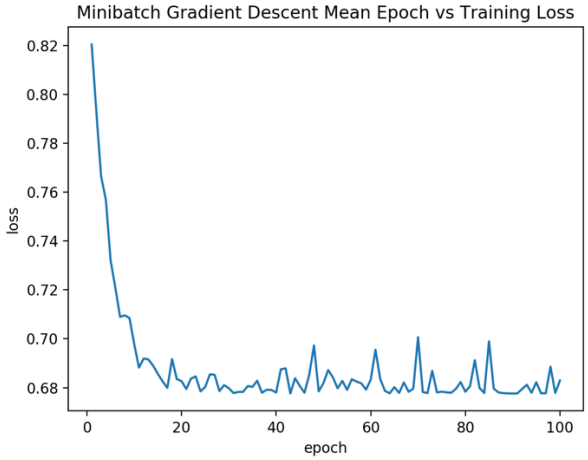
MiniSGD, Learning rate = 0.001, batch size = 20



MiniSGD, Learning rate = 0.001, batch size = 100



MiniSGD, Learning rate = 0.008, batch size = 20



MiniSGD, Learning rate = 0.008, batch size = 100