**README**

The zip file contains the following files:

Linreg.py- which contains the source code for multiple linear regression

yxlin.out- which contains the output derived by running linreg.py on yxlin.csv

yxlin2.out-which contains the output derived by running linreg.py on yxlin2.csv

GradientDescent.py

yxlin\_gd.out

yxlin\_gd.out

Steps to run on local cluster

In home/cloudera place the linreg.py file

In home/cloudera/input place the input csv files

On the terminal:

$ hadoop fs -mkdir /user/cloudera/linreg /user/cloudera/linreg/input

$ hadoop fs -put input/\* /user/cloudera/linreg/input

$ spark-submit linreg.py /user/cloudera/linreg/input/yxlin.csv >yxlin.out

Command on cluster

$ spark-submit <mysparkprogram.py> <inputdatafile>

Extra Credit

For linear regression using gradient descent

The value of alpha and number of iterations was taking as command line arguments.

Beta was calculated using the formula:

Beta=Beta +alpha\* x\_transpose \*(Y-(X\*beta))

The value used for alpha was 0.01 and number of iterations was 10 for both the input files.