COMP 551
Assignment 1

Richie Piyasirisilp, 260624968

Q1. Sampling

1. Pseudo code to sample from the given multinomial distribution

For each sample used, the algorithm will create a random number and see which distribution value it is nearest to. Once it has found the nearest value, a count will increment in the output array to show the probability of random values to appear as the activity distribution.

```
import numpy as np
import math
import matplotlib.pyplot as plt

def dailyRoutine(samples) :
    activityDist = [0.2, 0.4, 0.1, 0.3]
    computedDist = [0.0, 0.0, 0.0, 0.0]

    for i in range(samples):
        r = np.random.rand()

        for j in range(len(activityDist)):
            # Get distance from each random sample to the activity distribution
            if ((r-activityDist[j]) < 0.05):
                # If close value, increment occurence of sample
                computedDist[j] += 1/samples

    return computedDist
```

2. Multinomial distribution given 100 and 1000 samples

Compared to the underlying distribution, the function performs well and shows similar values to what was expected.

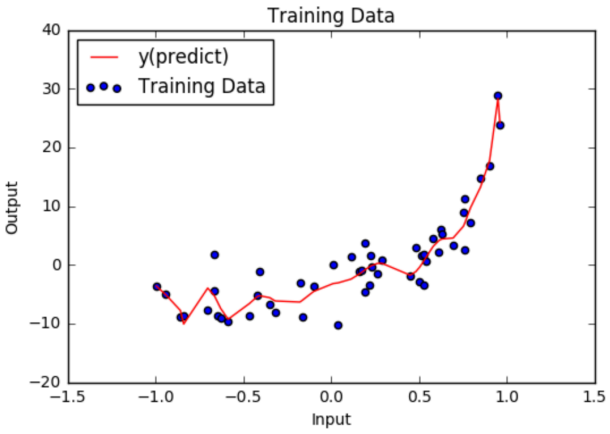
```
Distribution with 100 samples:  [0.24000000000000007, 0.4400000000000002, 0.12999999999999998, 0.34000000000000014]
Distribution with 1000 samples:  [0.23600000000000018, 0.43600000000000033, 0.1320000000000001, 0.33600000000000024]
```

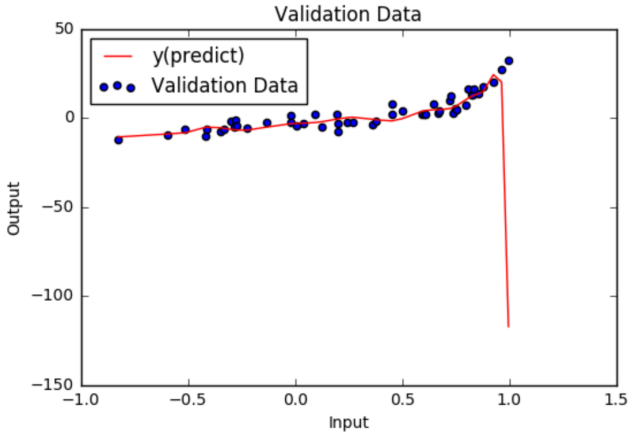
Q2. Model Selection

1(a) Training and validation MSE (Mean-Square Error) w/o regularization

Training MSE: 7.15251895354952
Validation MSE: 459.2411569614929

1(b) Visualize the fit

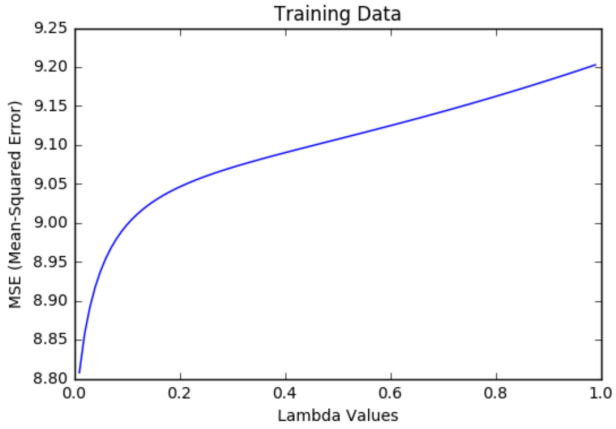


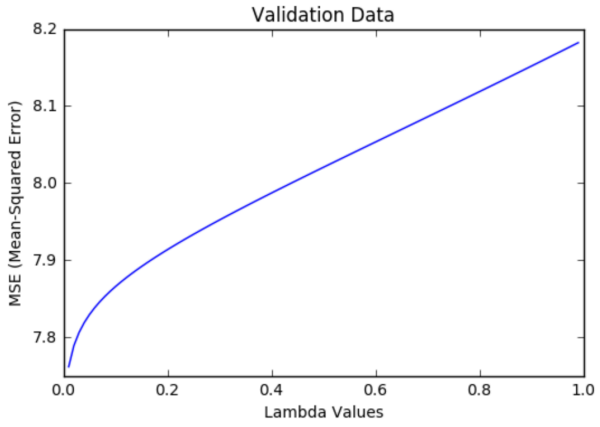


1(c) Is the model overfitting or underfitting?

The model is overfitting. This can be seen with the validation data since the predicted y has a very large error near the end of the dataset. The cause of the overfitting is because the model fit too well with the training data. This can be seen by the constant changes in y(predict) to fit data points in the Training data diagram. As a result, the model had low bias and high variance.

2(a) Plot training and validation MSE w/ regularization





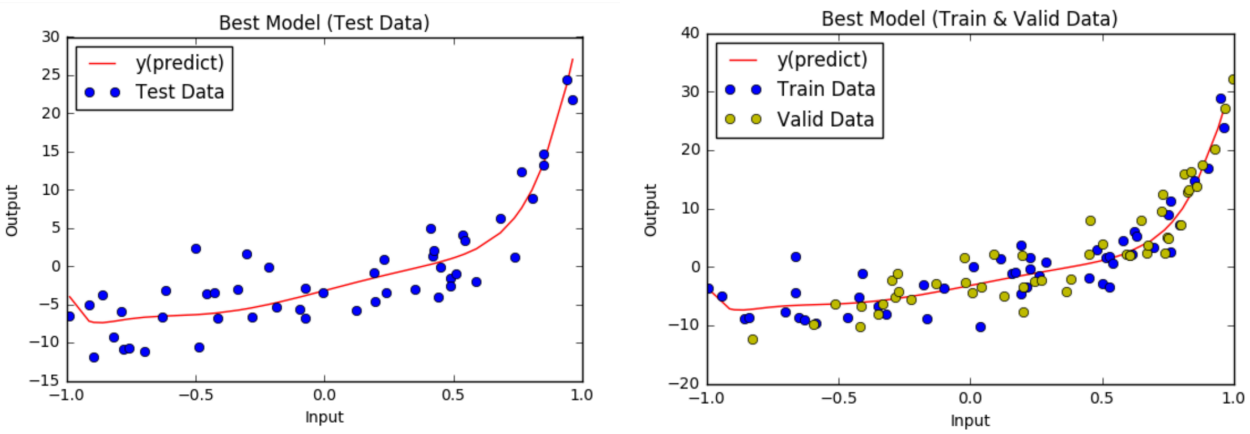
2(b) Best λ and MSE values for training and validation data & MSE of test data

Best training value of λ: 0.01
Best training MSE: 8.807610880442795

Best validation value of λ: 0.01
Best validation MSE: 7.7615051714404375

MSE of test data: 10.822349630572226

2(c) Visualize the fit



2(d) Is the model overfitting or underfitting?

The model is not overfitting since there is low variance. This can be seen in the right diagram because the model does not vary and change course greatly to try and fit data points. The model however might be underfitting but we do not know the bias. If the bias is high the model will make more assumptions about the shape of the test data but this cannot be confirmed since the test data is very similar in shape to the training and validation data.

3. Degree of source polynomial?

The degree of the source polynomial could be around 4. This can be seen in the shape of the models with and without regularization. With a 2-degree polynomial there is a high chance the model would be underfitting but with a 4-degree polynomial there would be enough bias without increasing the variance too much.

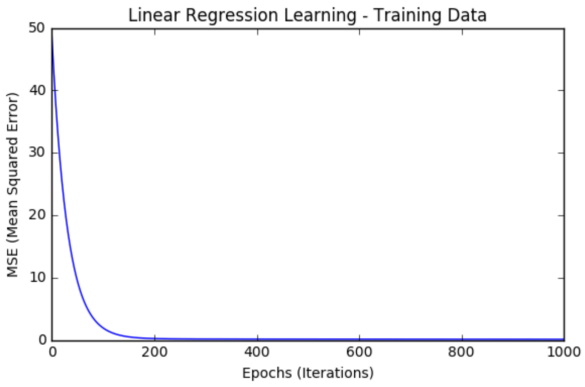
Q3. Gradient Descent for Regression

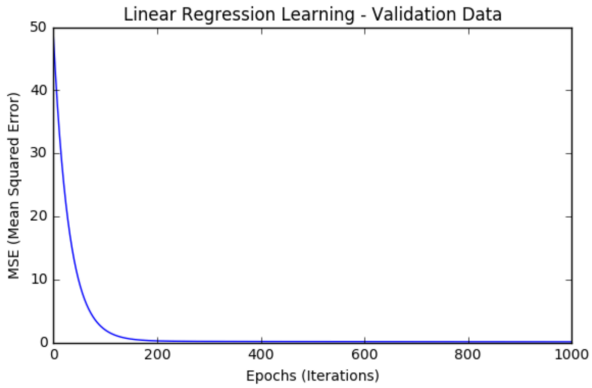
1(a) Compute MSE on validation set

Final m/w0 value: 4.06267983717
Final b/w1 value: 3.79549738172
Therefore, w=[4.06267983717 , 3.79549738172]

Training MSE: 5.84551331137e-13
Validation MSE: 2.85012327771e-12

1(b) Training and validation MSE learning curve

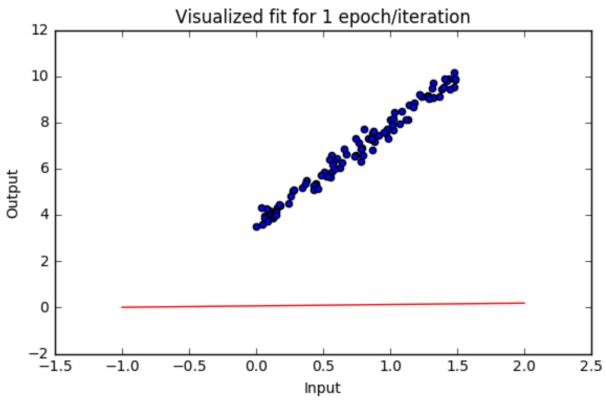


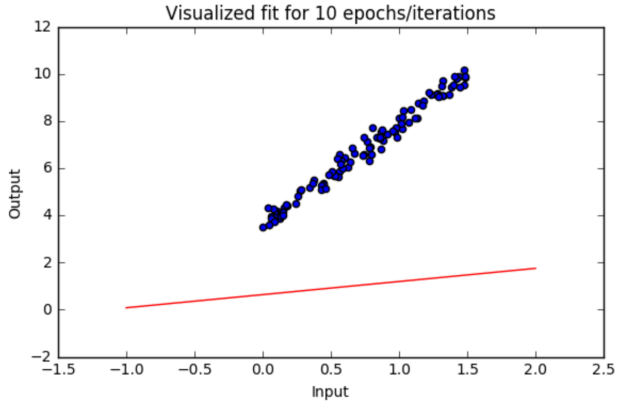


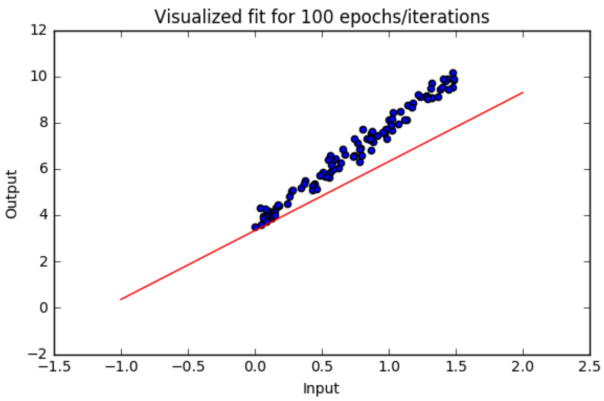
2. Use different step sizes

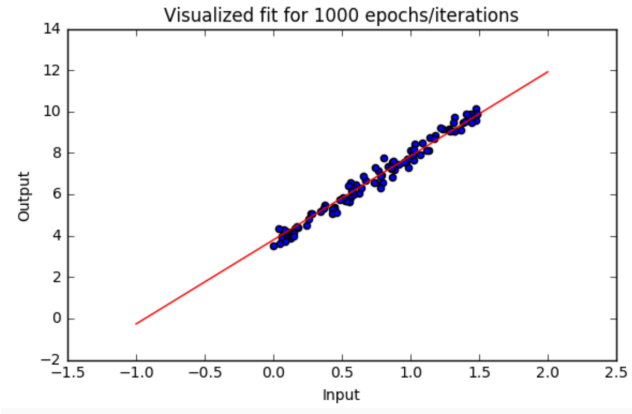
By increasing the step size, the model scans downhill faster. This is done so that the model doesn't take too long to finish however if the step size is too large, the model might miss the lowest loss value and be stuck jumping between false positives.

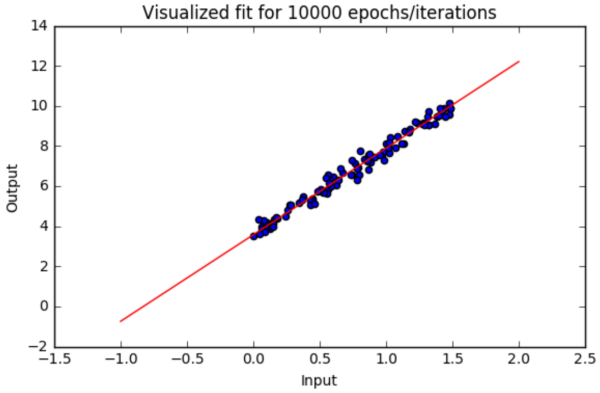
3. Visualize the fit (5 visualizations)











Q4. Real Life Dataset

1(a) Use sample mean to fill in missing data. Is this a good choice?
This is not a good choice because the mean that is calculated from each column includes all the outliers in the data.

1(b) What else could fill in the missing data?
Other ways to fill in the missing data could include using the median value or a random value from each column.

1(c) Describe a better method
A better method would be to use the median value of each column since this excludes all the outliers in the data.

1(d) Turn in completed data set
The completed data set can be found as “Communities_and_Crime_Cleaned.csv” under the “data/” directory.

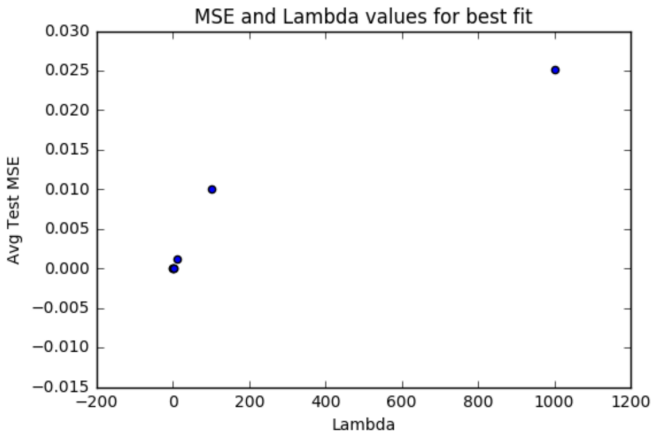
2. Fit data using linear regression and report 5-fold cross-validation error

The w values of each of the 5 sets can be found in the Jupyter Notebook files.
The 5-fold cross-validation error for the data sets used was 41.4077605004

3(a) λ vs MSE Plot

λ= 0.0, MSE = 1.311792e-13
λ= 0.1, MSE = 2.744916e-07
λ= 0.5, MSE = 0.000006
λ= 1.0, MSE = 0.000023
λ= 10.0, MSE = 0.001144
λ= 100.0, MSE = 0.01009
λ= 1000.0, MSE = 0.025088

Therefore, a λ value of 0.0 is used to get the best fit



3(b) Can we use this information for feature selection?
This information can be used to filter out features based on their parameters.

3(c) Show results of best fit
The w values of each of the λ values can be found in the Jupyter Notebook files.

3(d) Difference in model performance with reduced features and all features
If a model were to use all the features there will most likely be features that aren't useful that will cause noise. This will increase the overall MSE. By using reduced features the model will only use features that are useful which will improve the overall results and decrease the total MSE.