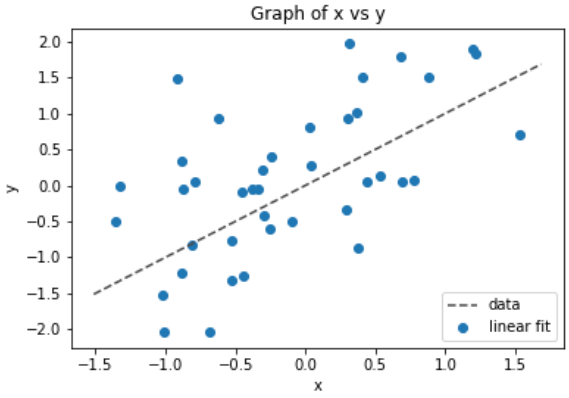
ASSIGNMENT 1,EOSC 410 (ANGELENE LEOW, 23162167)

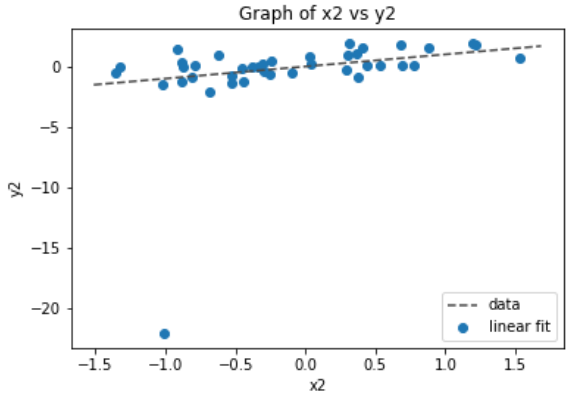
Problem 1:

1. Pearson correlation of x and y = 0.5800975391211216
2. Pearson correlation of x2 and y2 = 0.3397211855203799
3. Pearson correlation of x3 and y3 = -0.9010291351446986
4. Spearman correlation of x and y = 0.5724202626641651
5. Spearman correlation of x2 and y2 = 0.5724202626641651
6. Spearman correlation of x3 and y3 = 0.4318949343339587

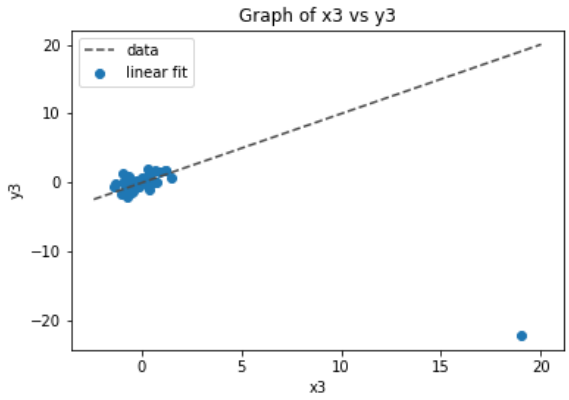
Plot 1:



Plot 2:



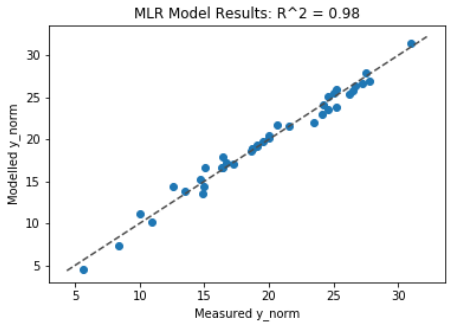
Plot 3:



The outliers in plot 2 and plot 3 can be easily observed. Spearman correlation is more resistant to outliers than Pearson correlation. The Spearman coefficient of x and y is similar with x2 and y2 as the formula is biased against outliers. The 5th point (outlier) in plot 2 was at the lowest value in plot 1, hence there is no change in the ranking of the y data when the data point became an outlier in plot 2.

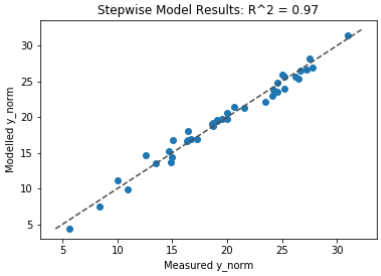
Problem 2:

Using Multiple Linear Regresssion:




```
intercept = -820.3594995596047
coefficient of x1 = 84.0383574070665
coefficient of x2 = 0.39359447394793784
coefficient of x3 = -3.3446220660735007
coefficient of x4 = -6.690571786532005
coefficient of x5 = 0.18275533513346662
coefficient of x6 = 0.18275533513346662
```

Using Stepwise:




```
   1. Add  x6                 with p-value 1.01879e-09
   2. Add  x4                 with p-value 3.45675e-17
   3. Add  x3                 with p-value 7.44399e-07
   4. Add  x1                 with p-value 0.00532758
resulting features that are in:
['x6', 'x4', 'x3', 'x1']
```

features that are out:
['x2','x5']

```
Stepwise coefficient results:
intercept = -796.5486994468267
coefficient of x1 = 0.040151625378134506
coefficient of x3 = -6.692578926452557
coefficient of x4 = -3.1517990772981745
coefficient of x6 = 81.66048798250597
```

Results:

A stepwise regression only takes into account significant features whereas multiple linear regression (MLR) takes

all features/ predictors into account. Hence 4 predictors are chosen out of 6. For stepwise, the smaller p-value show s a higher significance. Therefore the 4 predictors in order of significance are x4, x6, x3 and x1.

The regression coefficients for both MLR and stepwise are different as stepwise only takes into account 4 predictors (x1,x3,x4,x6) whereas the coefficient of MLR is a result of all 6 predictors.

By using a standardized predictor, where

$$x = \frac{mean(x)}{\sqrt{Var(x)}}$$

Using Multiple Linear Regression:

```
intercept = 19.722171524375096
coefficient of x1 = 0.4585615567584299
coefficient of x2 = 0.21761556761840684
coefficient of x3 = -2.1073752554131877
coefficient of x4 = -3.3863144555123483
coefficient of x5 = 0.11391349361394731
coefficient of x6 = 0.11391349361394731
```

Using Stepwise:

```
Add  x6                         with p-value 1.01879e-09
Add  x4                         with p-value 3.45675e-17
Add  x3                         with p-value 7.44399e-07
Add  x1                         with p-value 0.00532758
Resulting features:
['x6', 'x4', 'x3', 'x1']
```

features that are out:
['x2','x5']

```
Stepwise coefficient results:
intercept = 19.722171524375096
coefficient of x1 = 2.8962539950672204
coefficient of x3 = -3.3873303338474217
coefficient of x4 = -1.985881589703753
coefficient of x6 = 0.44558653512856405
```

Results:

After standardizing the x values , the sequence of order of significant predictors for the stepwise regression remains the same. However, the regression coefficients in MLR are higher which shows the increased importance of the predictors.