Assignment 1

Problem 1

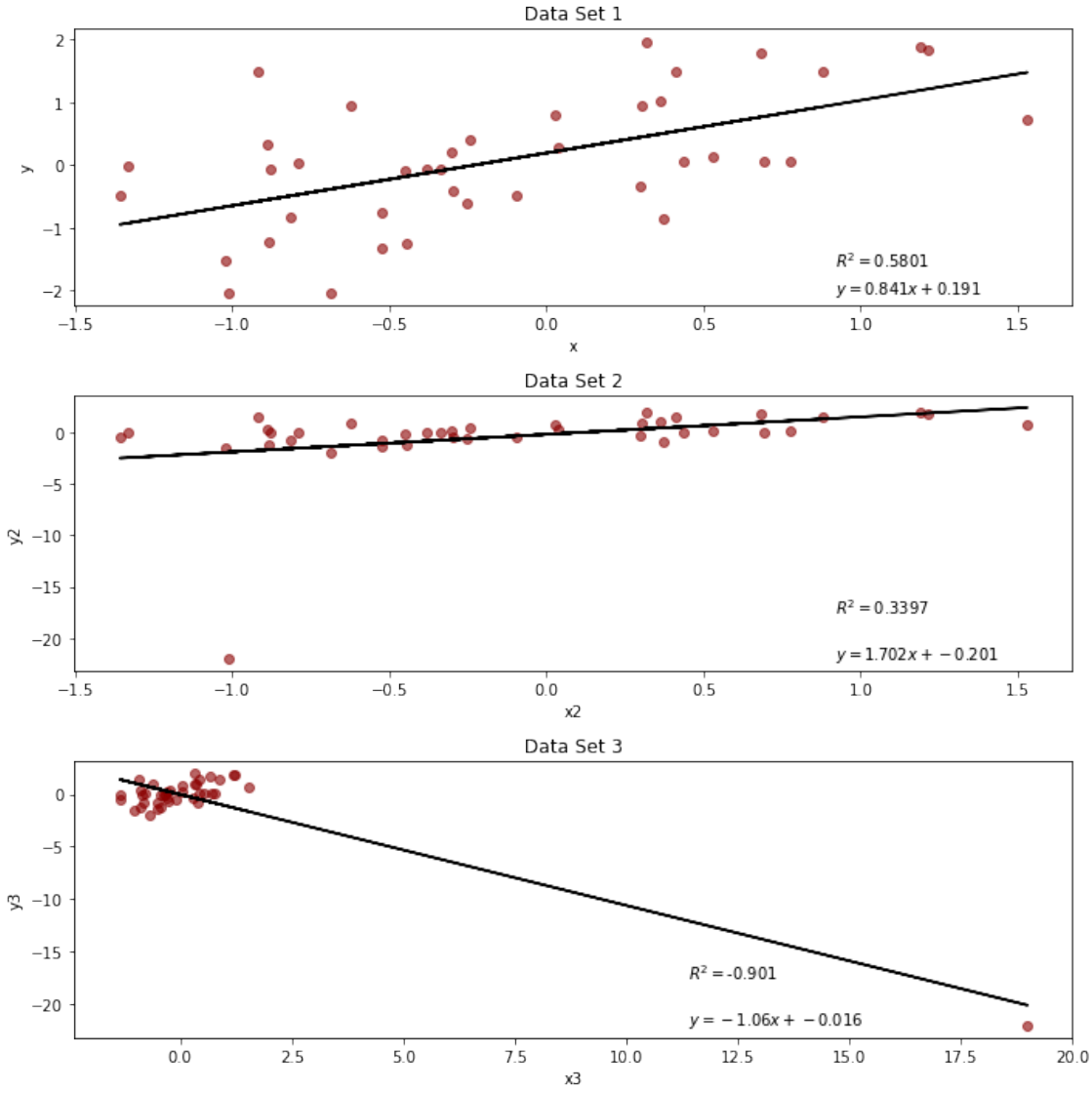
1. **Calculate Pearson correlation for each of the following pair of variables: x and y, x2 and y2, and x3 and y3. This should produce three correlation coefficients.**

   x,y) $R^2 = 0.5801$

   x2,y2) $R^2 = 0.3397$

   x3,y3) $R^2 = 0.9010$

2. **Make scatterplots of the data points in the xy space, the x2y2 space and the x3y3 space. Also plot the linear regression line in the scatterplots.**



3. **For each scatter plot answer the following: are there any outliers and if yes, is the correlation resistant to outliers? If not resistant to outliers, what resistant statistical measure would you need to use as an alternative to Pearson correlation?**

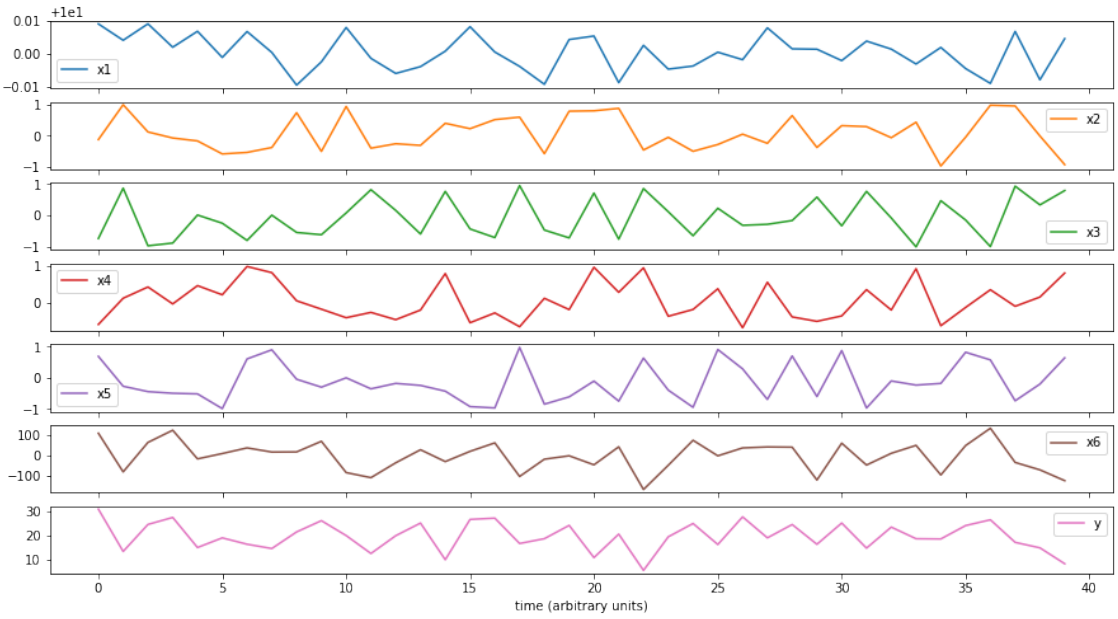
   Data Set 1) No major outliers

   Data Set 2) One outlier at $\approx (-1, -20)$, but there are enough data points that the effect is not massive. The fit line is still obviously deflected by the outlier, so a weighted linear fit might better represent this data

   Data Set 3) One outlier completely changes the fit. Either a weighted fit could solve this problem, or investigate the outlier to see if it is in fact "real" data

Problem 2

a) **plot the data as timeseries in one plot**

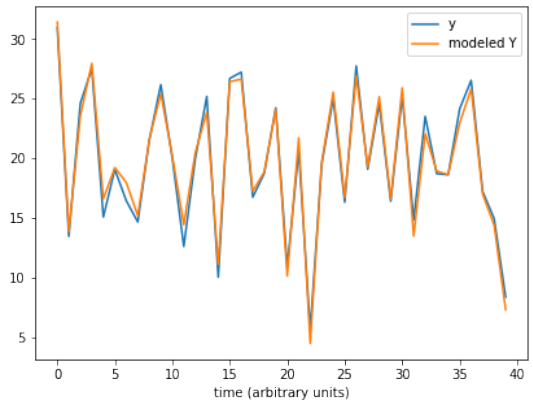


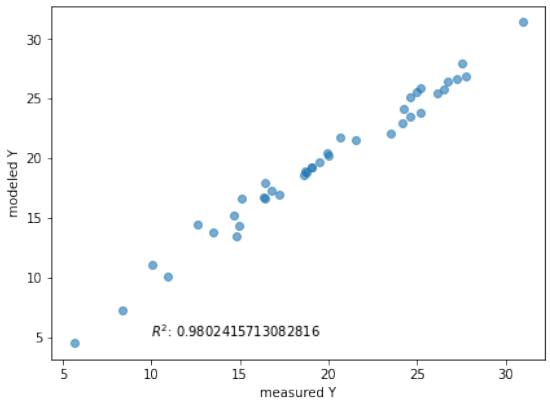
b) **perform multiple linear regression (MLR) and show the values of regression coefficients. Plot modelled y against true y in a scatterplot, and calculate multiple correlation coefficient (R2)**

Fitted equation:

$$y_{MLR} = -820.359 + 84.0383x_1 + 0.393594x_2 - 3.34462x_3 - 6.69057x_4 + 0.182755x_5 + 0.0382950x_6$$

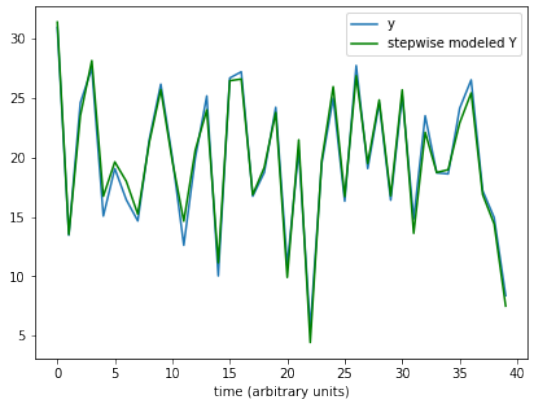
Assignment 1

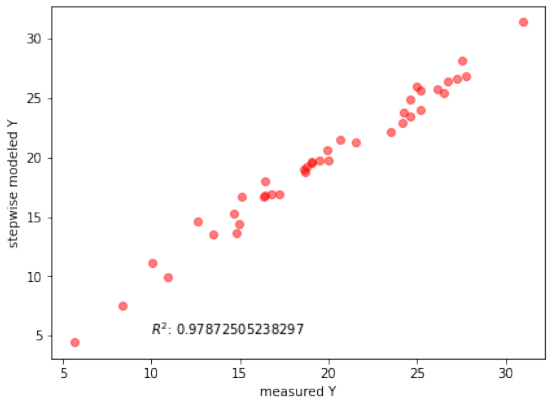




c) **perform stepwise regression, and show the final model equation with coefficients' values. Plot modelled y against true y in a scatterplot, and calculate multiple correlation coefficient (R^2). Rank the importance of the individual predictors in their influence on y.**

$$y_{SW} = -796.548 - 81.6604x_1 - 3.15179x_3 - 6.69257x_4 + 0.0401516x_6$$





Rank predictors x_i from most to least important:

$$x_1, x_4, x_3, x_6$$