CH5440 Assignment 1

Prem Sagar S - AE14B021

January 30, 2017

Problem 1

The WTLS is the minimization of J, ($\hat{y_i} = m\hat{x_i} + c$)

$$J = \sum (y_i - m\hat{x_i} - c)^2 + \gamma \sum (x_i - \hat{x_i})^2$$

The decision variables are m, c and x_i^*s. The partial derivative of J w.r.t to these variables is zero.

$$\frac{\partial J}{\partial m} = \sum 2(-y_i\hat{x_i} + m\hat{x_i}^2 + c\hat{x_i}) = 0$$

$$\sum -y_i\hat{x_i} + m\sum \hat{x_i}^2 + c\sum \hat{x_i} = 0 \tag{1}$$

$$\frac{\partial J}{\partial c} = \sum c(-y_i + m\hat{x_i} + c) = 0 \tag{2}$$

$$-\bar{y} + m\bar{\hat{x}} + c = 0$$

$$\frac{\partial J}{\partial \hat{x_j}} = 0 \Rightarrow m(-y_j + m\hat{x_j} + c) - \gamma(x_j - \hat{x_j}) = 0$$

$$\Rightarrow \hat{x_i} = \frac{y_i m + \gamma x_i - mc}{m^2 + \gamma}$$

On taking summation to N samples,

$$\bar{\hat{x}} = \frac{\bar{y}m + \bar{\gamma x} - mc}{m^2 + \gamma}$$

Using this in (2),

$$\bar{y} = m\bar{x} + c \tag{3}$$

Eq (1) can be rewritten as,

$$\sum y_i(\frac{y_i m + x_i - mc}{m^2 + \gamma}) - m\sum(\frac{y_i m + x_i - mc}{m^2 + \gamma})^2 - c\sum(\frac{y_i m + x_i - mc}{m^2 + \gamma}) = 0$$

which on simplifying becomes,

$$\sum m^2(\gamma c x_i - \gamma x_i y_i) + m(\gamma y_i^2 - 2c\gamma y_i + c^2\gamma - \gamma^2 x_i^2) + \gamma x_i y_i - c\gamma^2 x_i = 0$$

Using $c = \bar{y} - m\bar{x}$, and $NS_{xy} = \sum x_i y_i - N\bar{x}\bar{y}$, the above equation becomes,

$$m^2[\gamma\bar{x}(\bar{y}-m\bar{x})-\gamma(S_{xy}+\bar{x}\bar{y})]+m[\gamma(S_{yy}+\bar{y}^2)-2\gamma\bar{y}(\bar{y}-m\bar{x})+\gamma(\bar{y}-m\bar{x})^2-\gamma^2(S_{xx}+\bar{x}^2)]+\gamma^2(S_{xy}+\bar{x}\bar{y})-\gamma^2\bar{x}(\bar{y}-m\bar{x}) = 0 \tag{4}$$

which on further simplification becomes,

$$S_{xy}m^2 - (S_{yy} - \gamma S_{xx})m - \gamma S_{xy} = 0$$

The roots of this equation are,

$$m = \frac{S_{yy} - \gamma S_{xx} \pm \sqrt{(S_{yy} - \gamma S_{xx})^2 + 4\gamma S_{xy}^2}}{2S_{xy}}$$

Assuming the given samples are positive numbers, the highest slope is chosen,

$$m = \frac{S_{yy} - \gamma S_{xx} + \sqrt{(S_{yy} - \gamma S_{xx})^2 + 4\gamma S_{xy}^2}}{2S_{xy}}$$

Having known **m,**

$$c = \bar{y} - m\bar{x}$$

Inverse OLS

Taking $\gamma = 0$, the estimate for m can be written as,

$$m = \frac{S_{yy} - (0)S_{xx} + \sqrt{(S_{yy} - (0)S_{xx})^2 + 4(0)S_{xy}^2}}{2S_{xy}} = \frac{2S_{yy}}{2S_{xy}} = \frac{S_{yy}}{S_{xy}}$$

The result is same as expected. Setting $\gamma = 0$ was equivalent to considering $\sigma_\delta^2 = 0$ while $\sigma_\epsilon^2 \neq 0$.

Standard OLS

If the measurements in x are perfect, $\sigma_\epsilon^2 = 0$. **m** can be rewritten as,

$$m = \frac{\sigma_\epsilon^2 S_{yy} - \sigma_\delta^2 S_{xx} + \sqrt{(\sigma_\epsilon^2 S_{yy} - \sigma_\delta^2 S_{xx})^2 + 4\sigma_\delta^2 S_{xy}^2}}{2\sigma_\epsilon^2 S_{xy}} = \frac{S_{xy}}{S_{xx}}$$

....

If it was known apriori that c=0, $\bar{y} = m\bar{x}$, equation (4) becomes,

$$m^2[-\gamma(S_{xy} + m\bar{x}^2)] + m[\gamma(S_{yy} + m^2\bar{y}^2)] - \gamma^2(S_{xx} + \bar{x}^2) + \gamma^2(S_{xy} + m\bar{x}^2) = 0$$

which on simplifying gives,

$$S_{xy}m^2 - (S_{yy} - \gamma S_{xx})m - \gamma S_{xy} = 0$$

This is the same equation in **m** obtained before. Thus the structure of **m** doesn't change with **c** being zero, but it simplifies the computation.

Problem 2

The covariane matrix can be written as,

$$\underline{S}_{ZS} = \underline{Z}_S \underline{Z}_S^T$$

If the weight of x is γ times y, the covariance matrix becomes,

$$\underline{S}_{ZS} = \begin{bmatrix} S_{yy} & S_{xy} \\ \gamma S_{xy} & \gamma S_{xx} \end{bmatrix}$$

The eigen values can be computed by,

$$\left|\underline{S}_{ZS} - \lambda I\right| = 0$$

$$(S_{yy} - \lambda)(\gamma S_{xx} - \lambda) - \gamma S_{xy}^2 = 0$$

$$\lambda^2 - (S_{yy} + \gamma S_{xx})\lambda + \gamma S_{xx} S_{yy} - \gamma S_{xy}^2 = 0$$

$$\lambda = \frac{S_{yy} + \gamma S_{xx} \pm \sqrt{(S_{yy} - \gamma S_{xx})^2 + 4\gamma S_{xy}^2}}{2}$$

Taking the smaller eigen value,

$$\lambda_2 = \frac{S_{yy} + \gamma S_{xx} - \sqrt{(S_{yy} - \gamma S_{xx})^2 + 4\gamma S_{xy}^2}}{2}$$

Now the corresponding eigen vector is given by,

$$\begin{bmatrix} S_{yy} - \lambda & S_{xy} \\ \gamma S_{xy} & \gamma S_{xx} - \lambda \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0$$

$$\frac{v_{22}}{v_{21}} = \frac{\lambda - S_{yy}}{S_{xy}} = \frac{\frac{S_{yy} + \gamma S_{xx} - \sqrt{(S_{yy} - \gamma S_{xx})^2 + 4\gamma S_{xy}^2}}{2} - S_{yy}}{S_{xy}} = \frac{-S_{yy} + \gamma S_{xx} - \sqrt{(S_{yy} - \gamma S_{xx})^2 + 4\gamma S_{xy}^2}}{2S_{xy}}$$

which gives,

$$m = -\frac{v_{22}}{v_{21}} = \frac{S_{yy} - \gamma S_{xx} + \sqrt{(S_{yy} - \gamma S_{xx})^2 + 4\gamma S_{xy}^2}}{2S_{xy}}$$

Thus, both the methods give the right estimate for **m**, the covariance method being more easier.
c can be calculated as usual, $c = \bar{y} - m\bar{x}$

Problem 3

Data set 1

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

$\bar{y} = 7.5009090909$, $\bar{x} = 9.00$, $S_{xx} = 10.00$ $S_{yy} = 3.75$, $S_{xy} = 5.0009090909$

$$m = \frac{S_{xy}}{S_{xx}} = \frac{5.00909091}{10.00} = 0.5000909091$$

$$c = \bar{y} - m\bar{x} = 3.001$$

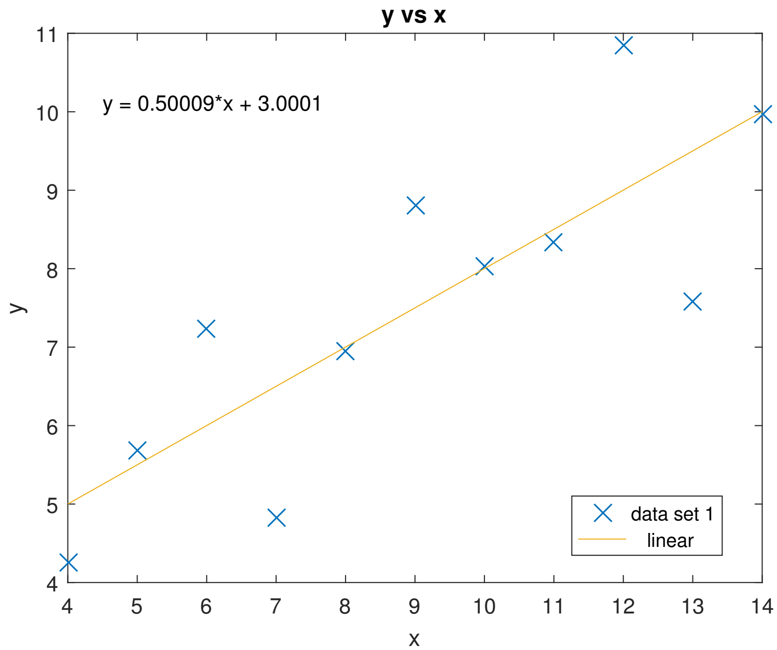

Figure 1: Data set 1

The data points are scattered across the linear fit, the model might be an approximate but not highly accurate.

Data set 2

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

$\bar{y} = 7.5009090909$ $\bar{x} = 9$, $S_{xx} = 10.00$ $S_{yy} = 3.7523900826$, $S_{xy} = 5$

$$m = \frac{S_{xy}}{S_{xx}} = \frac{5}{10.00} = 0.5$$

$$c = \bar{y} - m\bar{x} = 3.0009090909$$

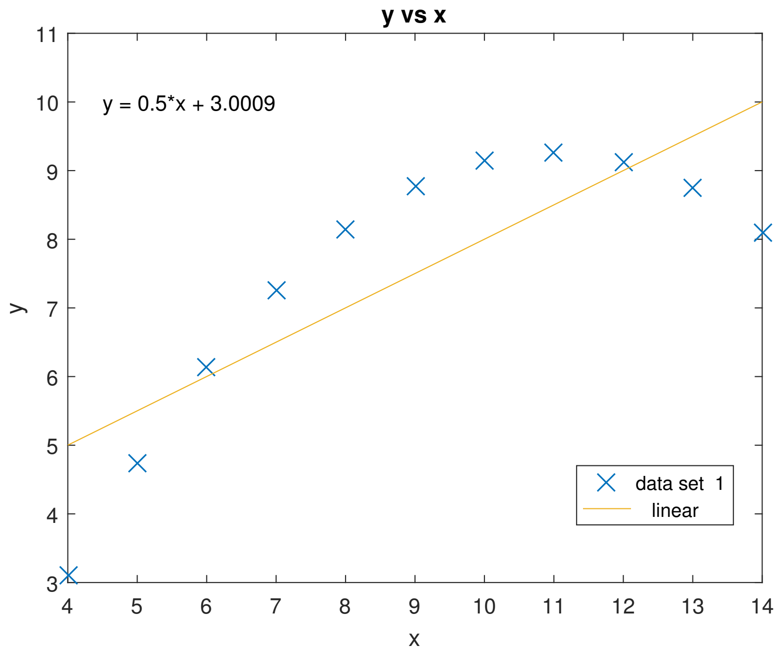

Figure 2: Data set 2

The data points seem to follow a non linear curve. The model doesn't estimate it acurately, but is only an approximation.

Data set 3

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

$\bar{y} = 7.5$ $\bar{x} = 9$, $S_{xx} = 10.00$ $S_{yy} = 3.7478371901$, $S_{xy} = 4.9972727273$

$$m = \frac{S_{xy}}{S_{xx}} = \frac{4.9972727273}{10.00} = 4.9972727273$$

$$c = \bar{y} - m\bar{x} = 3.0024545455$$

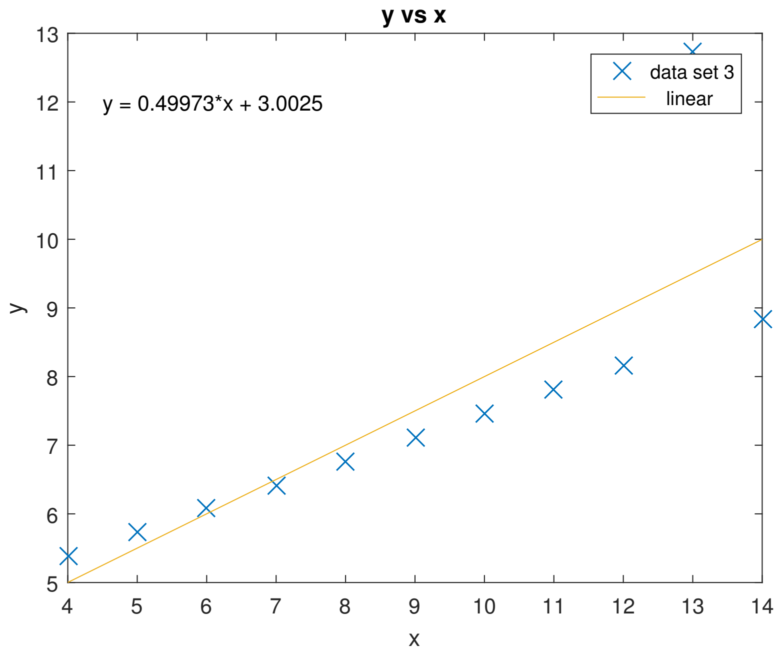

Figure 3: Data set 3

The scatter of the data is biased to one side of the linear fit in certain sections of the curve. The model can be split to both the sections to get a better estimation.

Data set 4

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

$\bar{y} = 7.5009090909$ $\bar{x} = 9$, $S_{xx} = 10.00$ $S_{yy} = 3.7484082645$, $S_{xy} = 4.9990909091$

$$m = \frac{S_{xy}}{S_{xx}} = \frac{4.9990909091}{10.00} = 0.4999090909$$

$$c = \bar{y} - m\bar{x} = 3.0017272727$$

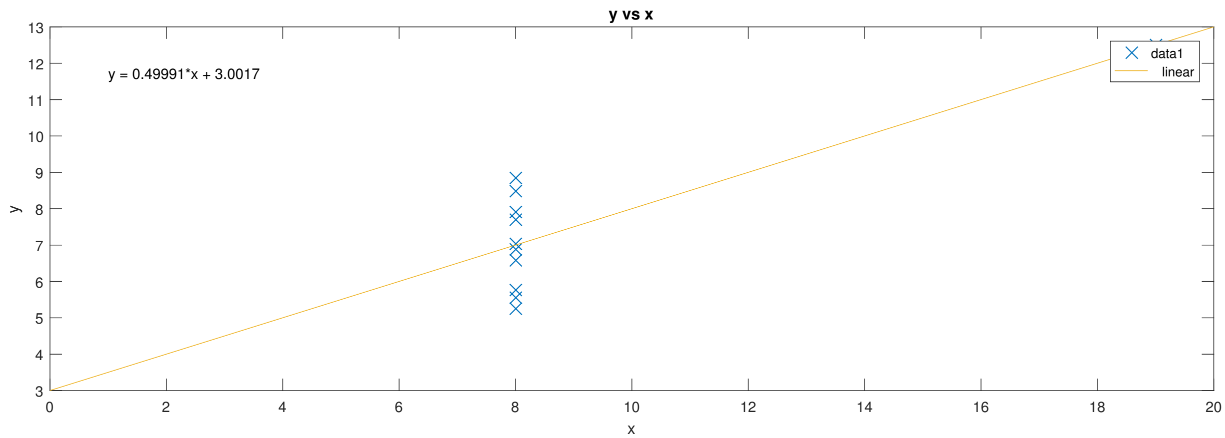

Figure 4: Data set 4

The model fails totally here because for the same x values, the model just gives the average value of y.
The model was most appropriate for Data set 1.