CH5540 Multivariate Data Analysis - Assignment 1

AE14B050

January 28, 2018

Solution 1:

Given ,

$$\min_{m,c,\hat{x}_i} \sum_{i=1}^{N} (y_i - m\hat{x}_i - c)^2/\sigma_\epsilon^2 + (x_i - \hat{x}_i)^2/\sigma_\delta^2$$

$$\lambda = \frac{\sigma_\epsilon^2}{\sigma_\delta^2}$$

Simplifying we get our objective function for WTLS which we denote with Z

$$Z = \min_{m,c,\hat{x}_i} \sum_{i=1}^{N} (y_i - m\hat{x}_i - c)^2 + \lambda(x_i - \hat{x}_i)^2 \tag{1}$$

To find the minima we shall now follow standard procedure and take partial differentials over all the $N+2$ variables i.e. $m, c, \hat{x}_i$.

$$\frac{\partial Z}{\partial c} = \sum_{i=1}^{N} (y_i - m\hat{x}_i - c)(-2) = 0 \implies \sum_{i=1}^{N} (y_i - m\hat{x}_i - c) = 0 \tag{2}$$

$$\frac{\partial Z}{\partial m} = \sum_{i=1}^{N} (y_i - m\hat{x}_i - c)(-2\hat{x}_i) = 0 \implies \sum_{i=1}^{N} \hat{x}_i(y_i - m\hat{x}_i - c) = 0 \tag{3}$$

$$\frac{\partial Z}{\partial \hat{x}_i} = (y_i - m\hat{x}_i - c)(-2m) + \lambda(x_i - \hat{x}_i)(-2) = 0 \implies m(y_i - m\hat{x}_i - c) + \lambda(x_i - \hat{x}_i) = 0 \tag{4}$$

Where (4) is by term basis as all other terms vanish except for the i^{th} variable .

Let us denote the mean of any variable x by $\overline{x}$.

Simplifying (2) we get

$$\overline{y} - m\overline{\hat{x}} = c \tag{5}$$

Summation of all the terms in (4) yeilds us,

$\sum_{i=1}^{N} m(y_i - m\hat{x}_i - c) + \lambda(x_i - \hat{x}_i) = m\sum_{i=1}^{N}(y_i - m\hat{x}_i - c) + \lambda\sum_{i=1}^{N}(x_i - \hat{x}_i) = 0$

$\implies m\frac{\partial Z}{\partial c} + \lambda N(\overline{x} - \overline{\hat{x}}) = 0 \implies \overline{x} = \overline{\hat{x}}$

Substituting this result back in (5),

$$\overline{y} - m\overline{x} = c \tag{6}$$

From (4) we can write the unknown terms $\hat{x}_i$ in other known terms,

$$\hat{x}_i = \frac{m(y_i - c) + \lambda x_i}{m^2 + \lambda} \tag{7}$$

Substituting (7) in (3),

$\sum_{i=1}^{N} \hat{x}_i(y_i - m\hat{x}_i - c) = \sum_{i=1}^{N}[\frac{m(y_i-c)+\lambda x_i}{m^2+\lambda}][y_i - m(\frac{m(y_i-c)+\lambda x_i}{m^2+\lambda}) - c] = 0$

$= \sum_{i=1}^{N}[m(y_i - c) + \lambda x_i][(y_i - c)(m^2 + \lambda) - m(m(y_i - c) + \lambda x_i)] = 0$

$= \sum_{i=1}^{N}[m(y_i - c) + \lambda x_i][(y_i - c)(\lambda) - m\lambda x_i)] = \sum_{i=1}^{N}[m(y_i - c) + \lambda x_i]\lambda[(y_i - c) - mx_i)] = 0$

$= \sum_{i=1}^{N} m(y_i - c)^2 + (\lambda - m^2)x_i(y_i - c) - \lambda m x_i^2 = 0$

$= \sum_{i=1}^{N} -m^2[x_i(y_i - c)] + m[(y_i^2 + c^2 - 2cy_i) - \lambda x_i^2] + \lambda[x_i(y_i - c)] = 0$

Substituting(6) into this equation and using the well known summation result $\sum(X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_i(Y_i - \overline{Y}) = \sum(X_i - \overline{X})Y_i$ as required,

$\sum_{i=1}^{N}[-m^2(x_i - \overline{x})(y_i - \overline{y})] + [-m^3 x_i] + [my_i^2] + [m^3 x_i] + [-2m^2\overline{xy}] + [-2my_i\overline{y}] + [2m^2\overline{x}y_i] + [-m\lambda x_i^2] + [\lambda x_i(y_i - \overline{y})] + [m\lambda x_i\overline{x}] = 0$

$\sum_{i=1}^{N}[-m^2(x_i - \overline{x})(y_i - \overline{y})] + [m(y_i - \overline{y})^2] + [-m\lambda(x_i - \overline{x})^2] + [\lambda(x_i - \overline{x})(y_i - \overline{y})] = 0$

$$-m^2 S_{xy} + m[S_{yy} - \lambda S_{xx}] + \lambda S_{xy} = 0 \tag{8}$$

Which is a quadratic equation whose positive root gives the minima condition on m(the other root is the orthogonal line which gives the maximum).

$$m = \frac{(S_{yy} - \lambda S_{xx}) + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}} \tag{9}$$

Where,

$$S_{xy} = \sum_{i=1}^{N}[(x_i - \overline{x})(y_i - \overline{y})] \quad S_{xx} = \sum_{i=1}^{N}(x_i - \overline{x})^2 \quad S_{yy} = \sum_{i=1}^{N}(y_i - \overline{y})^2 \tag{10}$$

(a)Standard OLS:

$$m = \lim_{\lambda \to \infty} \frac{(S_{yy} - \lambda S_{xx}) + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}}$$

We solve this by substituting $t = 1/\lambda$ and using bionomial expansion,

$m = \lim_{t \to 0} \frac{(tS_{yy} - S_{xx}) + \sqrt{(tS_{yy} - S_{xx})^2 + 4tS_{xy}^2}}{2tS_{xy}} = \lim_{t \to 0} \frac{(tS_{yy} - S_{xx}) \pm (tS_{yy} - S_{xx})(1 + \frac{1}{2}\frac{4tS_{xy}^2}{(tS_{yy} - S_{xx})^2})}{2tS_{xy}}$

only the negative expansion gives us a satisfactory result the other being $m = \infty$

$= \lim_{t \to 0} \frac{-(\frac{1}{2}\frac{4tS_{xy}^2}{(tS_{yy} - S_{xx})^2})(tS_{yy} - S_{xx})}{2tS_{xy}} = \lim_{t \to 0} \frac{-S_{xy}}{(tS_{yy} - S_{xx})} = \frac{S_{xy}}{S_{xx}}$

$$m_{OLS} = \frac{S_{xy}}{S_{xx}} \tag{11}$$

(a)Inverse OLS:

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

Which is straight forward to solve by substituting the value of the limit directly,

$$m_{IOLS} = \frac{S_{yy}}{S_{xy}} \tag{12}$$

(b)Case :$c = 0$ For this case we rewrite our objective function Z ,

$$Z^c = \min_{m,c,\hat{x}_i} \sum_{i=1}^{N} (y_i - m\hat{x}_i)^2 + \lambda(x_i - \hat{x}_i)^2 \tag{13}$$

We now have only N+1 variables and equations to solve for. Following the same procedure as in the normal case we have ,

$$\frac{\partial Z^c}{\partial m} = \sum_{i=1}^{N} (y_i - m\hat{x}_i)(-2\hat{x}_i) = 0 \implies \sum_{i=1}^{N} \hat{x}_i(y_i - m\hat{x}_i) = 0 \tag{14}$$

$$\frac{\partial Z^c}{\partial \hat{x}_i} = (y_i - m\hat{x}_i)(-2m) + \lambda(x_i - \hat{x}_i)(-2) = 0 \implies m(y_i - m\hat{x}_i) + \lambda(x_i - \hat{x}_i) = 0 \tag{15}$$

From (15) we can write the unknown terms $\hat{x}_i$ in other known terms,

$$\hat{x}_i = \frac{my_i + \lambda x_i}{m^2 + \lambda} \tag{16}$$

Substituting (16) in (14),

$\sum_{i=1}^{N} \hat{x}_i(y_i - m\hat{x}_i) = \sum_{i=1}^{N}[\frac{my_i+\lambda x_i}{m^2+\lambda}][y_i - m(\frac{my_i+\lambda x_i}{m^2+\lambda})] = 0$

$= \sum_{i=1}^{N}[my_i + \lambda x_i][y_i(m^2+\lambda) - m(my_i + \lambda x_i)] = \sum_{i=1}^{N}[my_i + \lambda x_i]\lambda[y_i - mx_i] = 0$

$= \sum_{i=1}^{N}[my_i^2 + (\lambda - m^2)x_iy_i - m\lambda x_i^2] = \sum_{i=1}^{N}[-m^2x_iy_i + my_i^2 + \lambda x_iy_i - m\lambda x_i^2] = 0$

We shall use the summation identity $\sum(X_iY_i - \overline{XY}) = \sum(X_i - \overline{X})(Y_i - \overline{Y})$ and notations from (10) into all the terms above

$= \sum_{i=1}^{N}[-m^2(x_iy_i - \overline{xy} + \overline{xy}) + m(y_i^2 + \overline{y}^2 - \overline{y}^2) + \lambda(x_iy_i + \overline{xy} - \overline{xy}) - m\lambda(x_i^2 + \overline{x}^2 - \overline{x}^2)] = 0$

$= [-m^2S_{xy} + mS_{yy} + \lambda S_{xy} - m\lambda S_{xx}] + N[-m^2(\overline{xy}) + m(\overline{y}^2) + \lambda(\overline{xy}) - m\lambda(\overline{x}^2)] = 0$

$= [-m^2K_{xy} + mK_{yy} + \lambda K_{xy} - m\lambda K_{xx}] = 0$

Which is a quadratic equation whose positive root gives the minima condition on m(the other root is the orthogonal line which gives the maximum).

$$m = \frac{(K_{yy} - \lambda K_{xx}) + \sqrt{(K_{yy} - \lambda K_{xx})^2 + 4\lambda K_{xy}^2}}{2K_{xy}} \tag{17}$$

Where,

$$K_{xy} = S_{xy} + N\overline{xy} \quad K_{xx} = S_{xx} + N\overline{x}^2 \quad K_{yy} = S_{yy} + N\overline{y}^2 \tag{18}$$

(b)Estimate solutions:

OLS($\lambda = \infty$):

$$y_i = \frac{S_{xy}}{S_{xx}}x_i + (\overline{y} - \frac{S_{xy}}{S_{xx}}\overline{x}) \tag{19}$$

IOLS($\lambda = 0$):

$$y_i = \frac{S_{yy}}{S_{xy}}x_i + (\overline{y} - \frac{S_{yy}}{S_{xy}}\overline{x}) \tag{20}$$

TLS($\lambda = 1$):

$$y_i = \frac{(S_{yy} - S_{xx}) + \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}}{2S_{xy}} x_i + (\overline{y} - \frac{(S_{yy} - S_{xx}) + \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}}{2S_{xy}} \overline{x}) \tag{21}$$

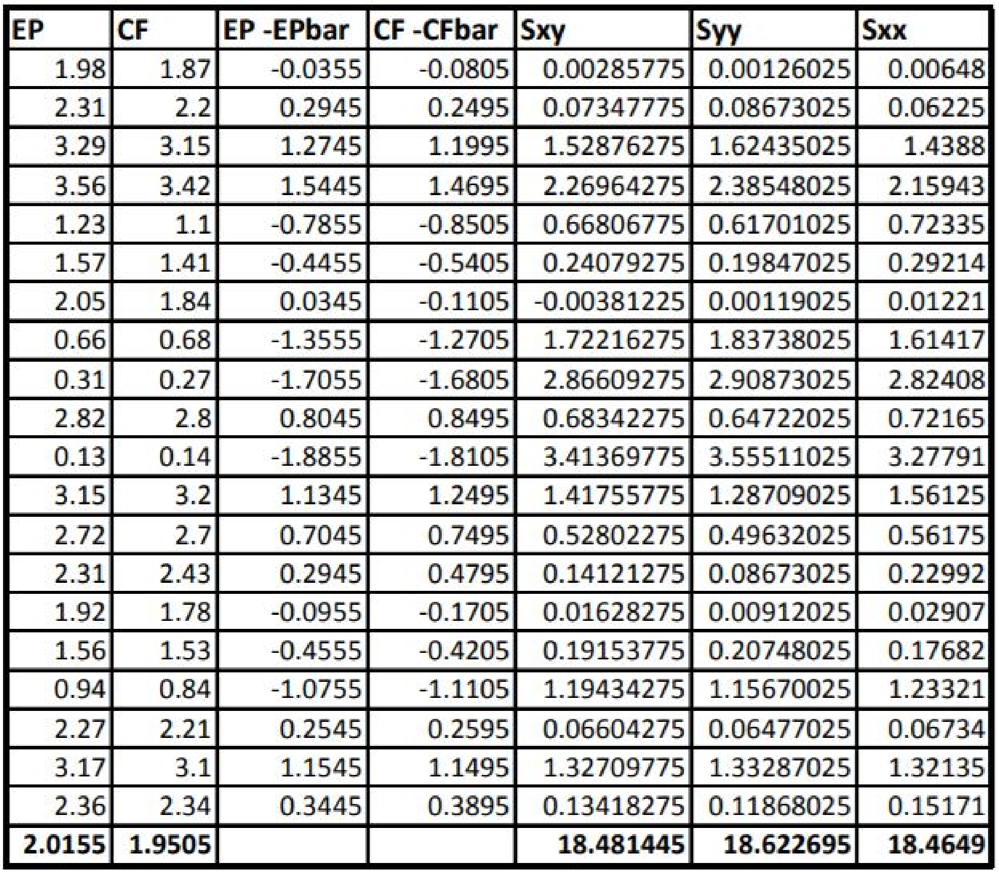
Solution 2:

(a): It is mentioned the variance of both the methods is same. If we estimate a linear regression relation between the two measuring equipment we shall now end up with a TLS problem.

For TLS, the slope equation for best regression fit is given by,

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

Let us take the established method "EP" measurements as the y-vector of measurements while the "CF" measurements as the x-vector we are trying to map. We tabuate the data below



EP	CF	EP -EPbar	CF -CFbar	Sxy	Syy	Sxx
1.98	1.87	-0.0355	-0.0805	0.00285775	0.00126025	0.00648
2.31	2.2	0.2945	0.2495	0.07347775	0.08673025	0.06225
3.29	3.15	1.2745	1.1995	1.52876275	1.62435025	1.4388
3.56	3.42	1.5445	1.4695	2.26964275	2.38548025	2.15943
1.23	1.1	-0.7855	-0.8505	0.66806775	0.61701025	0.72335
1.57	1.41	-0.4455	-0.5405	0.24079275	0.19847025	0.29214
2.05	1.84	0.0345	-0.1105	-0.00381225	0.00119025	0.01221
0.66	0.68	-1.3555	-1.2705	1.72216275	1.83738025	1.61417
0.31	0.27	-1.7055	-1.6805	2.86609275	2.90873025	2.82408
2.82	2.8	0.8045	0.8495	0.68342275	0.64722025	0.72165
0.13	0.14	-1.8855	-1.8105	3.41369775	3.55511025	3.27791
3.15	3.2	1.1345	1.2495	1.41755775	1.28709025	1.56125
2.72	2.7	0.7045	0.7495	0.52802275	0.49632025	0.56175
2.31	2.43	0.2945	0.4795	0.14121275	0.08673025	0.22992
1.92	1.78	-0.0955	-0.1705	0.01628275	0.00912025	0.02907
1.56	1.53	-0.4555	-0.4205	0.19153775	0.20748025	0.17682
0.94	0.84	-1.0755	-1.1105	1.19434275	1.15670025	1.23321
2.27	2.21	0.2545	0.2595	0.06604275	0.06477025	0.06734
3.17	3.1	1.1545	1.1495	1.32709775	1.33287025	1.32135
2.36	2.34	0.3445	0.3895	0.13418275	0.11868025	0.15171
2.0155	**1.9505**			**18.481445**	**18.622695**	**18.4649**

Figure 1: Tabulated data for problem 2

From the tabulation we calculate,

$\overline{x} = 1.9505$

$\overline{y} = 2.0155$

$S_{xy} = 18.481445$

$S_{xx} = 18.4649$

$S_{yy} = 18.622695$

substituting all the above terms back into (22)

$$m = 1.004278 \approx 1$$

$$c = \overline{y} - m\overline{x} = 0.0566 \approx 0 \qquad (23)$$

Inverting the x and y vectors to EP and CF data respectively we obtain,

$\overline{y} = 1.9505$

$\overline{x} = 2.0155$

$S_{xy} = 18.481445$

$S_{yy} = 18.4649$

$S_{xx} = 18.622695$

and

$$m^{'} = 0.99574 \approx 1$$

$$c^{'} = \overline{y} - m\overline{x} = -0.05641397 \approx 0 \qquad (24)$$

Which state that both methods are statistically similar if not the same and hence we will able to use the new method (CF) as a good substitute for the established (EP) method.

(b): $EP = 2.31mg/l$

using (24) we estimate in CF as

$CF = (m^{'}x) + c^{'} = (0.99574 * 2.31) - 0.05641397 = 2.24374543mg/l$

$CF^{'} = 2.20mg/l$

using (23) we estimate in EP as

$EP = (mx) + c = (1.004278 * 2.20) + 0.0566 = 2.2660116mg/l$

Solution 3 :

We use the following matlab code to plot the linear regression fit for each data set. We also visualize the fit to answer the question.

```
%Linear regression fit
D =[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];
x=D(:,1);
y=D(:,2);
x_mean=mean(x);
y_mean=mean(y);
Sxy=0;
Sxx=0;
for i=1:11
            Sxy = Sxy + (x(i)-x_mean)*(y(i)-y_mean);
            Sxx = Sxx + (x(i)-x_mean)*(x(i)-x_mean);
end
m=Sxy/Sxx;
c=y_mean - m*x_mean;
fprintf('\n Slope %f',m);
fprintf('\n intercept %f',c);
y1 = m*x + c;
plot(x,y,'*',x,y1,'-');
```

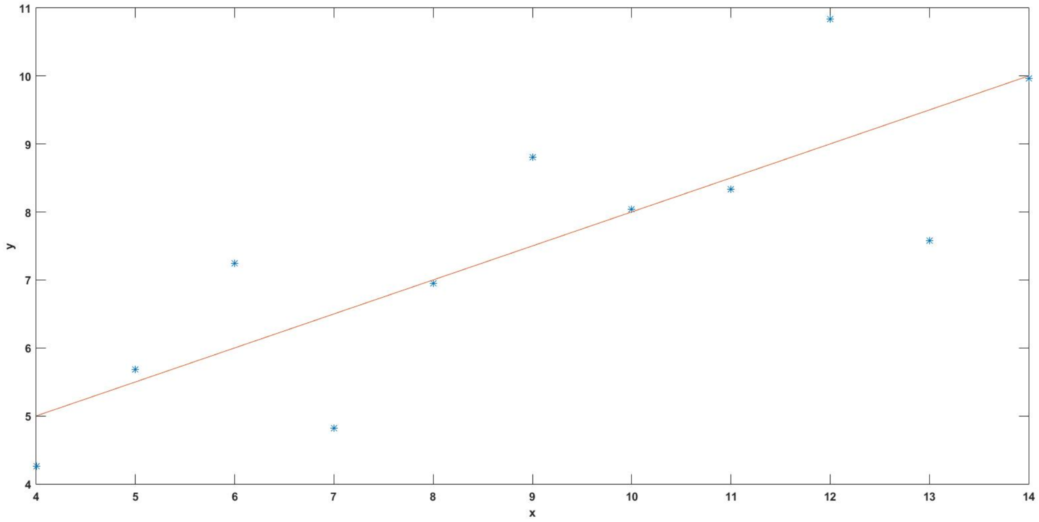


Figure 2: Regression fit of data I

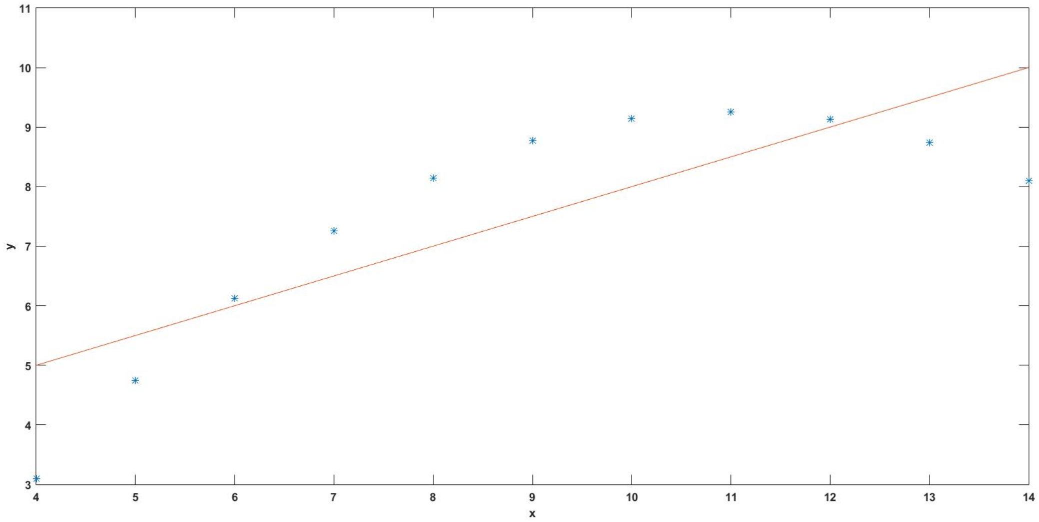


Figure 3: Regression fit of data II

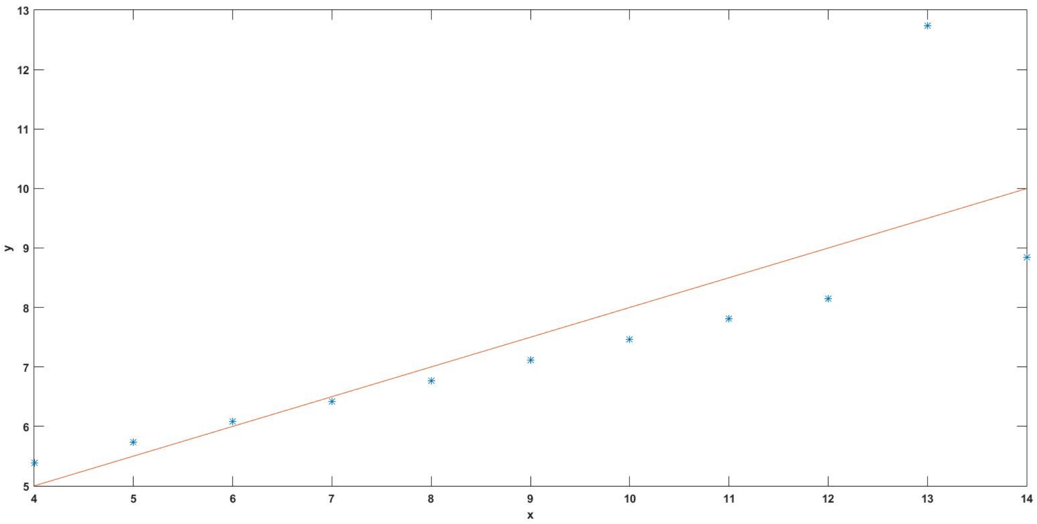


Figure 4: Regression fit of data III

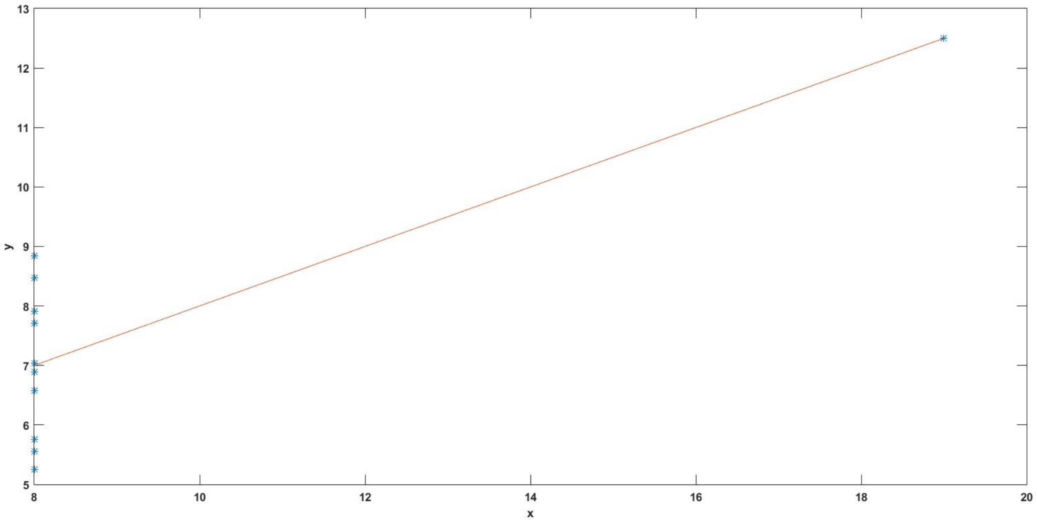


Figure 5: Regression fit of data IV

We have the following observations:

- The linear fit looks best for both data sets I and III, I is just larger than normal deviation (or error) that still maps onto the linear fit while III has outlier point skewing the data, without which we would be getting a close to perfect regression fit.
- For data set II we observe a curve being formed by the data points instead of a straight. A quadratic or higher order fit will give better results compared to the line fit we are using.
- Data set IV is highly errenous as the only reason for non infinity slop for the regression is persumably the outlier point, which we don't even consider into the regression fit generally. If we do know that this point

of data is true and valid we are basically trying to fit with just two points of data, one of them being the mean in x. It is possible that this fit still accurate but the chance of it being errenous is much higher.