Example of a "Least Squares Fitting" calculation for simple LR

Find the "least square fit" *y = b0 + b1x* for the experimental data points: {(1 , 2) , (3 , 4) , (2 , 6) , (4 , 8) , (5 , 12) , (6 , 13) , (7 , 15)}

Solution:

Number of Point N=7 Mean-Values ("Mittelwerte"): [M(x),M(y)] = [28/7; 60/7] ~ [4; 8,5714]

4 8,571428571

Set up a table with the quantities included in the above formulas for b0 and b1 and also the quantities for the calculation of R^2:

	needed for calculation of b0 and b1				needed for calculation of R^2			SST = SSE + SSR ?
i	***x_i***	***y_i***	***x_i*y_i***	***x_i^2***	***$y(x_i)$***	***SSE=sum(yi-y(x_i))²***	***SST=sum(yi-M(y))²***	***SSR=sum(y(xi) - M(y))²***
1	1	2	2	1	2,0357	0,001274	43,1833	42,7101
2	3	4	12	9	6,3929	5,726	20,8977	4,7459
3	2	6	12	4	4,2143	3,1887	6,6121	18,9843
4	4	8	32	16	8,5715	0,3266	0,3265	0
5	5	12	60	25	10,7501	1,5623	11,7553	4,7467
6	6	13	78	36	12,9287	0,00661	19,6125	18,9861
7	7	15	105	49	15,1073	0,01151	41,3269	42,718
sum	**28**	**60**	**301**	**140**		**10,822994**	**143,7143**	**132,8911**

xi - M(x)	***yi - M(y)***	***[xi - M(x)]*[yi-M(y)]***
-3	-6,571428571	19,71428571
-1	-4,571428571	4,571428571
-2	-2,571428571	5,142857143
0	-0,571428571	0
1	3,428571429	3,428571429
2	4,428571429	8,857142857
3	6,428571429	19,28571429
0	**0**	**61**

Substitute these values into Formula I and II:

b_0 = (140*60/7 -4*301)/(140 -7*16) = -28/196
= -1/7 ~ -0,14286

b_1 = (301 -7*4*(60/7))/28 = 61/28 ~ 2,1786

----> Regression-Line: y = -1/7 + (61/28)*x

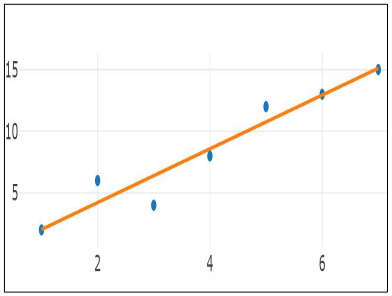
R^2 = 1 - Sum((yi-y(xi))²)/Sum((yi-M(y))²) = 1 - (10,822994 / 143,7143) ~ 0,9247

Compare with Python

intercept:
- 0.14285714285714057

slope: [2.17857143]

coeff. determination:
0.9247017892644135



Check of Proposition (P5.1): f(Mean(x)) = -1/7 + (61/28)*4 = -1/7 + 61/7 = 60/7 = Mean(y) q.e.d.

Example of a "Least Squares Fitting (LSF)" calculation for multiple LR (mLR)

Find the "least square fit" for z = a + b*x + c*y for the following training-set:[x,y] = [[0, 1], [5, 1], [15, 2], [25, 5], [35, 11], [45, 15], [55, 34], [60, 35]]; z = [4,

Solution:

Number of Point N=8 Mean-Values ("Mittelwerte") = : [M(x),M(y),M(z)] ~ [240/8, 104/8, 178/8] = [30; 13; 22,25]

Set up a table with the quantities included in the above LSF formulas (I) and (II) for simple LR:

	nedded for the calculation of a, b and c										
i	*xi*	*y_i*	*z_i*	*Xi:=xi-M(x)*	*Yi:=yi-M(y)*	*Zi:=zi-M(z)*	*Xi*Yi*	*Xi*Zi*	*Yi*Zi*	*Xi^2*	*Yi^2*
1	0	1	4	-30	-12	-18,25	360	547,50	219,00	900	144
2	5	1	5	-25	-12	-17,25	300	431,25	207,00	625	144
3	15	2	20	-15	-11	-2,25	165	33,75	24,75	225	121
4	25	5	14	-5	-8	-8,25	40	41,25	66,00	25	64
5	35	11	32	5	-2	9,75	-10	48,75	-19,50	25	4
6	45	15	22	15	2	-0,25	30	-3,75	-0,50	225	4
7	55	34	38	25	21	15,75	525	393,75	330,75	625	441
8	60	35	43	30	22	20,75	660	622,50	456,50	900	484
Sum	**240**	**104**	**178**	**0**	**0**	**0,00**	**2070**	**2115,00**	**1284,00**	**3550**	**1406**

Substitute the values to the formulas (I) and (II) of LSF for mLR:

det = sum(Xi^2)*sum(Yi^2)-(sum(Xi*Yi))^2=3550*1406-(2070) 2=706400

a = Mean(z)-b*Mean(x)-c*Mean(y) ~ 22,25-0,4471*30 + 0.25500 *13* ~ 5, 522

b = (1/det)*(sum(Yi^2)*sum(XiZi) sum(XiYi)*sum(YiZi))=(1/det)*(1406*2115-2070*1284) ~ 0,4471

c = (1/det)*(sum(Xi^2)*sum(YiZi)- sum(XiYi)*sum(XiZi))=(1/det)*(3550*1284-2070*2115) ~ 0,2550

Compare with Python-Pgm (next slides):
intercept: 5.52257927519819
coefficients: [0.44706965 0.25502548]

So we get the optimal mLR line: z = 5,522 + 0,4471*x + 0,255*y

R^2 = 1 - SSE/SST ~ 0,86159   (details see notes-page)

--> Adj.R^2= 1 -(1-R^2)*(7/5) ~ 0,8062 (details see notepage)

coefficient of determination: 0.8615939258756776

Set up a table with the quantities included in the formulas for R^2 and Adj.R^2:

needed for calculation of R^2			SST = SSE + SSR ?
z(xi,yi)	*SSE=sum(zi-z(xi,yi))^2*	*SST=sum(zi-M(z))^2*	*SSR=sum(z(xi,yi)-M(z))^2*
5,777	3,1577	333,0625	271,3597
8,0125	9,0752	297,5625	202,7064
12,7385	52,7294	5,0625	90,4686
17,9745	15,7967	68,0625	18,2799
23,9755	64,3926	95,0625	2,9774
29,4665	55,7486	0,0625	52,0779
38,7825	0,6123	248,0625	273,3236
41,273	2,9825	430,5625	361,8745
	204,4950	**1477,5000**	**1273,0680**

sse+ssr: 1477,5630

R^2 = 1-SSE/SST = 1 - Sum((zi-z(xi,yi))^2)/Sum((zi-M(z))^2) = 1 -(204,495/1477,5)~ 0,86159

--> Adj.R^2= 1 -(1-R^2)*(7/5) ~ 0,8062

*xi*yi*	*sum[(xi - M(x))*(yi - M(y))]*
0	360
5	300
30	165
125	40
385	-10
675	30
1870	525
2100	660
5190	**2070**

Check Corollary (C5.2) -*"center of mass"*

f(Mean(x), Mean(y))=5,522 + 0,4471*30 + 0,255*13 = 89/4 = 22,25= Mean(z)   q.e.d.

Check Proposition (P5.1-(iii)):

sum(xi*yi) = sum[(xi - M(x))*(yi - M(y))] + n*M(x)*M(y)

5190=2070 + 8*30*13=2070+3120   q.e.d.

Example of a "Least Squares Fitting" calculation for simple LR: "Student Examination"

Find "least square fit" $y = b0 + b1x$ with {(exam prep.[h], score[pt.])}= {(7, 41), (3, 27), (5, 35), (3, 26), (8, 48), (7, 45), (10, 46), (3, 27), (5, 29) , (3, 19)}

Solution:

Number of Point N=10 Mean-Values ("Mittelwerte"): [M(x),M(y)] = [54/10; 343/10] = [5,4; 34,3]

Set up a table with the quantities included in the above formulas for b0 and b1 and also the quantities for the calculation of R^2:

	needed for calculation of b0 and b1				needed for calculation of R^2			SST = SSE + SSR ?
student i	**exam prep. x_i**	**points y_i**	**x_i*y_i**	**x_i^2**	**$y(x_i)$**	**SSE=sum(yi-y(x_i))²**	**SST=sum(yi-M(y))²**	**SSR=sum(y(xi) - M(y))²**
1	7	41	287	49	40,269	0,534361	44,89	35,628961
2	3	27	81	9	25,325	2,805625	53,29	80,550625
3	5	35	175	25	32,797	4,853209	0,49	2,259009
4	3	26	78	9	25,325	0,455625	68,89	80,550625
5	8	48	384	64	44,005	15,960025	187,69	94,187025
6	7	45	315	49	40,269	22,382361	114,49	35,628961
7	10	46	460	100	51,477	29,997529	136,89	295,049329
8	3	27	81	9	25,325	2,805625	53,29	80,550625
9	5	29	145	25	32,797	14,417209	28,09	2,259009
10	3	19	57	9	25,325	40,005625	234,09	80,550625
sum	**54**	**343**	**2063**	**348**		**134,217194**	**922,10**	**787,214794**

sse+ssr
921,431988

Substitute these values into Formula I and II:

b_0 = (348*34,3 -5,4*2063)/(348-10*5,4²) =
(3981/5)/(282/5) = 1327/94 ~14,117
b_1 = (2063-10*5,4*34,3)/(282/5)= (1054/5)/(282/5)=527/141~3,7376

----> Regression-Line: y = 14,117 + 3,7376*x

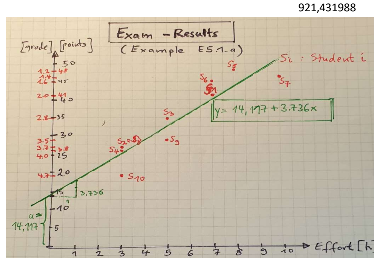
R^2 = 1 - Sum((yi-y(xi))²)/Sum((yi-M(y))²) = 1 - (134,2172/922,1) ~ 0,8544

Compare with Python

intercept:
14.11702127659574
slope: [3.73758865]

coeff. determination:
0.8544449495100991

51.4929078



Question 1: 14,117 points.
Question 2: 14,117 + 37,38 = 51,497 = 50 points (grade= 1,0).
Question 3: x= (25-14,117)/3,7376=2,91[h]

Check of Proposition (P5.1): f(Mean(x)) = 14,117 + 3,736*5,4 = 34,2914 ~ Mean(y) q.e.d.

	needed for calculation of b0 and b1				
student i	**exam. prep. x_i**	**points y_i**	**x_i*y_i**	**x_i^2**	**grade**
1	7	41	287	49	2,0
2	3	27	81	9	3,7
3	5	35	175	25	2,8
4	3	26	78	9	3,8
5	8	48	384	64	1,2
6	7	45	315	49	1,6
7	10	46	460	100	1,4
8	3	27	81	9	3,7
9	5	29	145	25	3,5
10	3	19	57	9	4,7
sum	**54**	**343**	**2063**	**348**	**2,8**

Find "least square fit" $y = b0 + b1x$ with {(homework[h], score[pt.])}= {(5, 41), (4, 27), (5, 35), (3, 26), (9, 48), (8, 45), (10, 46), (5, 27), (3, 29), (3, 19)}

Solution:

Number of Point N=10 Mean-Values ("Mittelwerte"): [M(x),M(y)] = [55/10; 343/10] = [5,5; 34,3]

Set up a table with the quantities included in the above formulas for b0 and b1 and also the quantities for the calculation of R^2:

	needed for calculation of b0 and b1				needed for calculation of R^2			SST = SSE + SSR ?
student i	**exam prep. x_i**	**points y_i**	**x_i*y_i**	**x_i^2**	**$y(x_i)$**	**SSE=sum(yi-y(x_i))²**	**SST=sum(yi-M(y))²**	**SSR=sum(y(xi) - M(y))²**
1	5	41	205	25	32,559	71,250481	44,89	3,031081
2	4	27	108	16	29,08	4,3264	53,29	27,2484
3	5	35	175	25	32,559	5,958481	0,49	3,031081
4	3	26	78	9	25,601	0,159201	68,89	75,672601
5	9	48	432	81	46,475	2,325625	187,69	148,230625
6	8	45	360	64	42,996	4,016016	114,49	75,620416
7	10	46	460	100	49,954	15,634116	136,89	245,047716
8	5	27	135	25	32,559	30,902481	53,29	3,031081
9	3	29	87	9	25,601	11,553201	28,09	75,672601
10	3	19	57	9	25,601	43,573201	234,09	75,672601
sum	**55**	**343**	**2097**	**363**		**189,699203**	**922,10**	**732,258203**

sse+ssr
921,957406

Substitute these values into Formula I and II:

b_0 = (363*34,3 -5,5*2097)/(363-10*5,5²) = (4587/5)/(121/2) =
834/55 ~15,164
b_1 = (2097 -10*5,5*34,3)/(121/2) = 421/121~3,479

----> Regression-Line: y = 15,164 + 3,479*x

R^2 = 1 - Sum((yi-y(xi))²)/Sum((yi-M(y))²) = 1 -(189,699 / 922,1) ~ 0,7943

Compare with Python

intercept:
15.163636363636353
slope: [3.47933884]

coeff. determination:
0.7942748361851003

49.95702479

Question 1: 15,164 points.
Question 2: 15,164 + 34,79 = 49,954 = 50 points (grade= 1,0).
Question 3: x= (25-15,164)/3,479=2,83[h]

Check of Proposition (P5.1): f(Mean(x)) = 15,164 + 3,479*5,5 = -34,2985 ~ Mean(y) q.e.d.

Example of a "Least Squares Fitting (LSF)" calculation for multiple LR (mLR)

Find "least square fit" z = *a + b*x + c*y* with *Training Set TS* ={(x, y; z) | (exam prep.[h], homework[h]; score[pt.])}=
{(7,5;41), (3,4;27), (5,5;35), (3,3;26), (8,9;48), (7,8;45), (10,10;46), (3,5;27), (5,3;29), (3,3;19)}

Solution:

Number of Point N=10 Mean-Values ("Mittelwerte") = : [M(x); M(y); M(z)] ~ [5,4 ; 5,5 ; 34,3]

Set up a table with the quantities included in the above LSF formulas (I) and (II) for simple LR:

	nedded for the calculation of a, b and c										
i	*xi*	*y i*	*z i*	*Xi:=xi-M(x)*	*Yi:=yi-M(y)*	*Zi:=zi-M(z)*	*Xi*Yi*	*Xi*Zi*	*Yi*Zi*	*Xi^2*	*Yi^2*
1	7	5	41	1,6	-0,5	6,7	-0,80	10,7	-3,4	2,56	0,25
2	3	4	27	-2,4	-1,5	-7,3	3,60	17,5	11,0	5,76	2,25
3	5	5	35	-0,4	-0,5	0,7	0,20	-0,3	-0,4	0,16	0,25
4	3	3	26	-2,4	-2,5	-8,3	6,00	19,9	20,8	5,76	6,25
5	8	9	48	2,6	3,5	13,7	9,10	35,6	48,0	6,76	12,25
6	7	8	45	1,6	2,5	10,7	4,00	17,1	26,8	2,56	6,25
7	10	10	46	4,6	4,5	11,7	20,70	53,8	52,7	21,16	20,25
8	3	5	27	-2,4	-0,5	-7,3	1,20	17,5	3,7	5,76	0,25
9	5	3	29	-0,4	-2,5	-5,3	1,00	2,1	13,3	0,16	6,25
10	3	3	19	-2,4	-2,5	-15,3	6,00	36,7	38,3	5,76	6,25
Sum	**54**	**55**	**343**	**0,0**	**0,0**	**0,0**	**51,00**	**210,8**	**210,5**	**56,40**	**60,50**

Substitute the values to the formulas (I) and (II) of LSF for mLR:

det = sum(Xi^2)*sum(Yi^2)-(sum(Xi*Yi))^2=56,40*60,50-(51*51)=811,20

a = Mean(z)-b*Mean(x)-c*Mean(y) ~ 34,3-2,4875*5,4 - 1.3824*5,5 ~ 13,2643

b = (1/det)*(sum(Yi^2)*sum(XiZi)-sum(XiYi)*sum(YiZi))=(1/det)*(60,50*210,8-51*210,5)~ 2,4875

c = (1/det)*(sum(Xi^2)*sum(YiZi)- sum(XiYi)*sum(XiZi))=(1/det)*(56,4*210,5-51*210,8) ~ 1,3824

Compare with Python-Pgm (next slides):
intercept: 13,2641
coefficients: [2.4875 1.3824]

So we get the optimal mLR line: z = 13,264 + 2,488*x + 1,382*y

R^2 = 1 - SSE/SST ~ 0,8843 (details see notes-page)

coefficient of determination: 0.8843

--> Adj.R^2= 1 -(1-R^2)*(9/7) ~ 0.8512 (details see notepage)

Set up a table with the quantities included in the formulas for R^2 and Adj.R^2:

needed for calculation of R^2			SST = SSE + SSR ?
z(xi,yi)	*SSE=sum(zi-z(xi,yi))^2*	*SST=sum(zi-M(z))^2*	*SSR=sum(z(xi,yi)-M(z))^2*
37,59	11,6281	44,8900	10,8241
26,256	0,5535	53,2900	64,7059
32,614	5,6930	0,4900	2,8426
24,874	1,2679	68,8900	88,8495
45,606	5,7312	187,6900	127,8256
41,736	10,6537	114,4900	55,2941
51,964	35,5693	136,8900	312,0169
27,638	0,4070	53,2900	44,3822
29,85	0,7225	28,0900	19,8025
24,874	34,5039	234,0900	88,8495
	106,7302	**922,1000**	**815,3930**

sse+ssr: 922,1231

*xi*yi*	*sum[(xi - M(x))*(yi - M(y))]*
35	-0,80
12	3,60
25	0,20
9	6,00
72	9,10
56	4,00
100	20,70
15	1,20
15	1,00
9	6,00
348	**51,00**

R^2 = 1-SSE/SST = 1 - Sum((zi-z(xi,yi))^2)/Sum((zi-M(z))^2) = 1 -(106,7302/922,1) ~ 0,88425

--> Adj.R^2= 1 -(1-R^2)*(9/7) ~ 0,8512

Check Corollary (C5.2) -*"center of mass"*

f(Mean(x), Mean(y))=13,264 + 2,488*5,4 + 1,382*5,5 ~ 34,30 = Mean(z) q.e.d.

Step 4: Get results

You can obtain the properties of the model the same way as in the case of simple linear regression:

```
In [4]: r_sq = model.score(x, y)
        print('coefficient of determination:', r_sq)
        print('intercept:', model.intercept_)
        print('coefficients:', model.coef_)

coefficient of determination: 0.8842531690055735
intercept: 13.264053254437865
coefficients: [2.48754931 1.38239645]
```

Check Proposition (P5.1-(iii)):

sum(xi*yi) = sum[(xi - M(x))*(yi - M(y))] + n*M(x)*M(y)

348 = 51 + 10*5,4*5,5 = 51 + 297 q.e.d.

Manuel calculation of two sLR-lines (green ,red) (Homework (H5.1_a) + Compare with optimal sLR-line (homewrk (H5.1_b) + Check Results with the new metric R^2=SSR/SST

Decide what is the "better" sLR-Line: *y = 1,5 + 0,5*x* *or* *y = 1,25+0,5*x* ?

Solution:

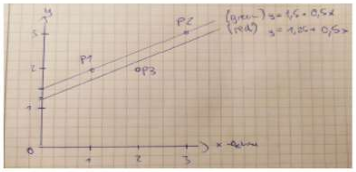
Number of Point N=3 Mean-Values ("Mittelwerte"): [M(x),M(y)] = [2; (7/3)]

Set up a table with the quantities included in the above formulas for a and b and also the quantities for the calculation of R^2:

With the defintion R^2:=SSR/SST we get as an result that the green line is the best sLR-line of the three --> With R^2=1-SSE/SST it was the yellow line

	needed for calculation of a and b				Needed for calculation of R^2				SST = SSE + SSR ?	
i	*xi*	*yi*	x_i*y_i	x_i^2	$y(x_i)$	SSE=sum(yi-y(x_i))²	SST=sum(yi-M(y))²	R^2	SSR=sum(y(xi) - M(y))²	R^2
1	1	2	2	1	2,00	0,0000000	0,1111111	1,0000000	0,1111111	
2	3	3	9	9	3,00	0,0000000	0,4444444	1,0000000	0,4444444	
3	2	2	4	4	2,50	0,2500000	0,1111111	0,9000000	0,0277778	
sum	**6**	**7**	**15**	**14**		**0,2500000**	**0,6666667**	0,6250000	**0,5833333**	0,8750000
									0,8333333	<--SSR + SSE

- y = 1,5 + 0,5*x
- y = 1,25 + 0,5*x



Which estimation (red or green) is better?

Needed for calculation of R^2				SST = SSE + SSR ?	
$y(x_i)$	SSE=sum(yi-y(x_i))²	SST=sum(yi-M(y))²	R^2	SSR=sum(y(xi) - M(y))²	R^2
1,75	0,0625000	0,1111111		0,3402778	
2,75	0,0625000	0,4444444		0,1736111	
2,25	0,0625000	0,1111111		0,0069444	
	0,1875000	**0,6666667**	0,7187500	**0,5208333**	0,7812500
				0,7083333	<--SSR + SSE

From Homework (H5.1_b) we get the data for the "**optimal" sLR-line:**

needed for calculation of R^2				SST = SSE + SSR ?	
y(xi)	*SSE=sum(yi-y(xi))²*	*SST=sum(yi-M(y))²*	R^2	*SSR=sum(y(xi) - M(y))²*	R^2
11/6	(1/6)²=1/36	(-1/3)²=1/9		1/4	
17/6	(1/6)²=1/36	(2/3)²=4/9		1/4	
14/6	(-2/6)²=4/36	(1/3)²=1/9		0	
42/6=7	**6/36=1/6**	**2/3**	0,7500000	**1/2**	0,7500000
				0,6666667	<--SSR + SSE

Example of a "Least Squares Fitting" calculation for simple LR: Homework 5.1_b

Coefficients a, b calculated with formulas of Theorem: "(sst=sse+ssr)=??=> optimal"

Find "least square fit" $y = a + b*x$ with Trainingset TS :={(x, y))}= {(1,2), (3, 3), (2, 2)}

Solution:

Number of Point N=3 Mean-Values ("Mittelwerte"): [M(x),M(y)] = [2; (7/3)]

Set up a table with the quantities included in the above formulas for b0 and b1 and also the quantities for the calculation of R^2:

2 7/3

	needed for calculation of b0 and b1				needed for calculation of R^2			SST = SSE + SSR ?
i	***xi***	***yi***	x_i*y_i	x_i^2	$y(x_i)$	$SSE=sum(yi-y(x_i))^2$	$SST=sum(yi-M(y))^2$	$SSR=sum(y(xi) - M(y))^2$
1	1	2	2	1	11/6	$(1/6)^2=1/36$	$(-1/3)^2=1/9$	1/4
2	3	3	9	9	17/6	$(1/6)^2=1/36$	$(2/3)^2=4/9$	1/4
3	2	2	4	4	14/6	$(-2/6)^2=4/36$	$(1/3)^2=1/9$	0
sum	**6**	**7**	**15**	**14**	**42/6=7**	**6/36=1/6**	**2/3**	**1/2**

0,25	***xi - M(x)***	***yi - M(y)***	***[xi - M(x)]*[yi-M(y)]***
0,25	-1	-0,333333333	0,333333333
	1	0,666666667	0,666666667
	0	-0,333333333	0
0,50	**0**	**-4,44089E-16**	**1**

Substitute these values into Formula I and II:

Compare with Python

b_0 = ((7/3)*14 -2*15)/(14-12) = (8/3)/2= 4/3

b_1 = (15 -3*2*(7/3))/2 = 1/2 = 0.5

intercept: 1.3333333333333334

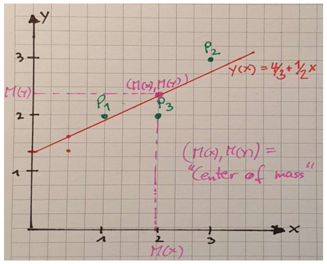
slope: [0.5]

----> Regression-Line: y = 4/3 + 1/2*x

R^2 = 1 - Sum((yi-y(xi))²)/Sum((yi-M(y))²) = 1 -(1/6)/(2/3)
= 1 -(1*3)/(6*2)=1-3/12= 1-1/4=3/4

coefficient of determination:
0.7499999999999999

Check of Proposition (P5.1): f(Mean(x)) = (4/3) + (1/2)*2 = 7/3 = Mean(y) q.e.d.



Coefficients a, b calculated with formulas of Theorem: "(sst=sse+ssr)=??=> optimal"

a= [(7/3)*14-(14/3)*6]/(14-3*4)= 14/(3*2) = 14/6=7/3 = M(y)

b=[(7/3)*6 - 3*2*(7/3)]/2 = [14-14]/2=0 ===> y(x) = 7/3 + 0*x ===>y(xi)=M(y) => SSR= 0 ---> unclear ??!

SSR=0

SSE = (2-7/3)²+ (3-7/3)²+0=1/9+4/9=5/9 ; SST =6/9 > 5/9 = SSR + SSE

SST = 2/3 => R^2=1-(5/9)/(2/3)= 1-15/18=3/18=1/6 not max. ! (SST<>sse+sst --> not counter-example !!!