Predictive models

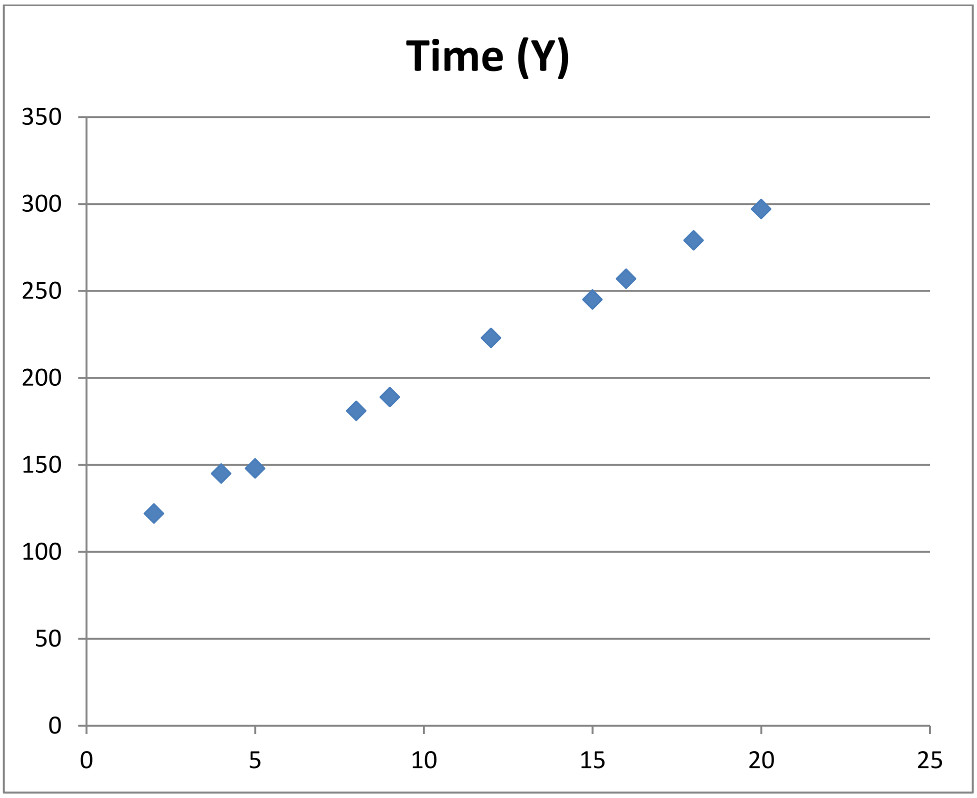
Linear regression and the least squares method

Assume we want to make a prediction on a certain value of interest: for instance, the time needed to prepare goods to be shipped by train. Assume we also have some intuition about something from which this value is likely to depend: for instance, the number of items to be shipped. Intuitively, it looks reasonable to assume that it takes more time to prepare a shipment of many items than the shipment of few items. Then, we call these two quantities *dependent variable* (the preparation time) and *independent variabile* (the number of items). We will indicate the dependent variable with Y and the independent variable with X.

Now, assume we have some *historical data* available: for several times in the past shipments have been prepared and every time someone has recorded the number of shipped items and the time taken to prepare the shipment. These records are the *input data* for our analysis. Assume we have imported them in a spreadsheet.

	A	B	C
1			
2		**Colli (X)**	**Tempo (Y)**
3		2	122
4		4	145
5		5	148
6		8	181
7		9	189
8		12	223
9		15	245
10		16	257
11		18	279
12		20	297
13			

If we visualize the X and Y values in a scatter plot, we can have an intuition about the most suitable function for a regression analysis.

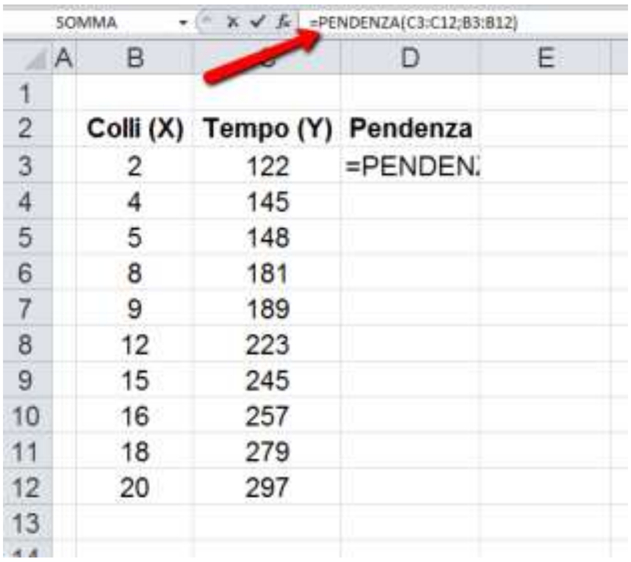


In the plot, the points are rather well aligned along a *straight line*. This suggests to make a *linear regression*, i.e. to compute the line interpolating the given points in the best way.

The line we are searching for is defined by two coefficients:

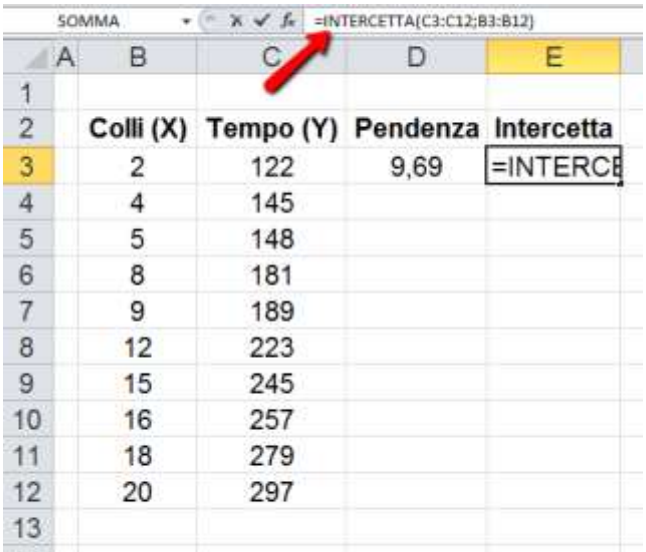
- the *slope*, that is the increase of Y corresponding to a unit increase of X;
- the *intercept*, that is the value of Y corresponding to the null value of X.

Spreadsheets include pre-defined functions that allow to execute linear regression immediately. In particular, it is possible to obtain the slope and the intercept of the line, using respectively the functions SLOPE() and INTERCEPT(), and passing them as arguments the vector of the Y values and the vector of the X values, separated by a semicolon.



SOMMA =PENDENZA(C3:C12;B3:B12)

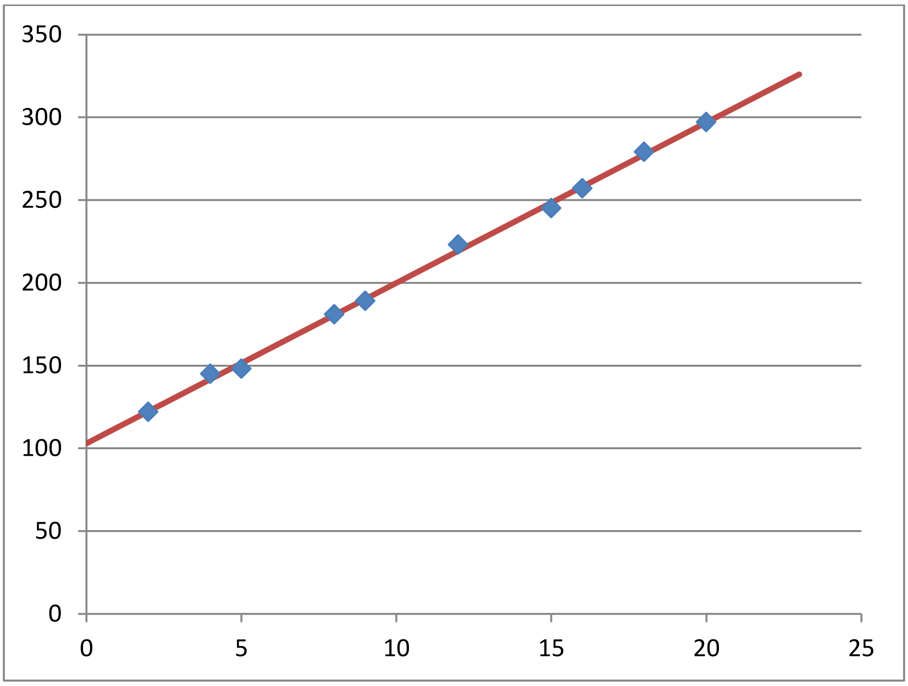
	A	B	C	D	E
1					
2		Colli (X)	Tempo (Y)	Pendenza	
3		2	122	=PENDEN	
4		4	145		
5		5	148		
6		8	181		
7		9	189		
8		12	223		
9		15	245		
10		16	257		
11		18	279		
12		20	297		
13					



SOMMA =INTERCETTA(C3:C12;B3:B12)

	A	B	C	D	E
1					
2		Colli (X)	Tempo (Y)	Pendenza	Intercetta
3		2	122	9,69	=INTERCE
4		4	145		
5		5	148		
6		8	181		
7		9	189		
8		12	223		
9		15	245		
10		16	257		
11		18	279		
12		20	297		
13					

Indicating the slope with *m* and the intercept with *q*, the line has equation $Y = m\ X + q$. In this example we have $m = 9{,}69$ e $q = 102{,}97$.



After having computed the line, it is very useful to check the value of the *linear correlation coefficient*, also called *Pearson index*. Also this value can be obtained directly with the predefined function PEARSON() with the same arguments as above.

F3 =PEARSON(C3:C12;B3:B12)

	A	B	C	D	E	F
1						
2		**Colli (X)**	**Tempo (Y)**	**Pendenza**	**Intercetta**	**Pearson**
3		2	122	9,69	102,97	0,999
4		4	145			
5		5	148			
6		8	181			
7		9	189			
8		12	223			
9		15	245			
10		16	257			
11		18	279			
12		20	297			
13						

Since the value of this index is positive, the line is increasing. Since its absolute value is very close to 1, the points show a very high linear correlation, that is they are very close to the regression line.

Another useful indicator of the quality of the model (and of the forecasts it provides) is a measure of the distance between the line and the data. The regression line minimizes the *mean square error,* i.e. the mean value of the squared differences. The differences are as many as the (X,Y) pairs in input. For each pair (X_i,Y_i) the error is defined as the difference between the input value Y_i and the value computed on the regression line for X_i, i.e. $mX_i + q$. There is no predefined function in the spreadsheet to show the mean square error. However, a predefined function provides the *standard error*. This is the square root of the mean square error.

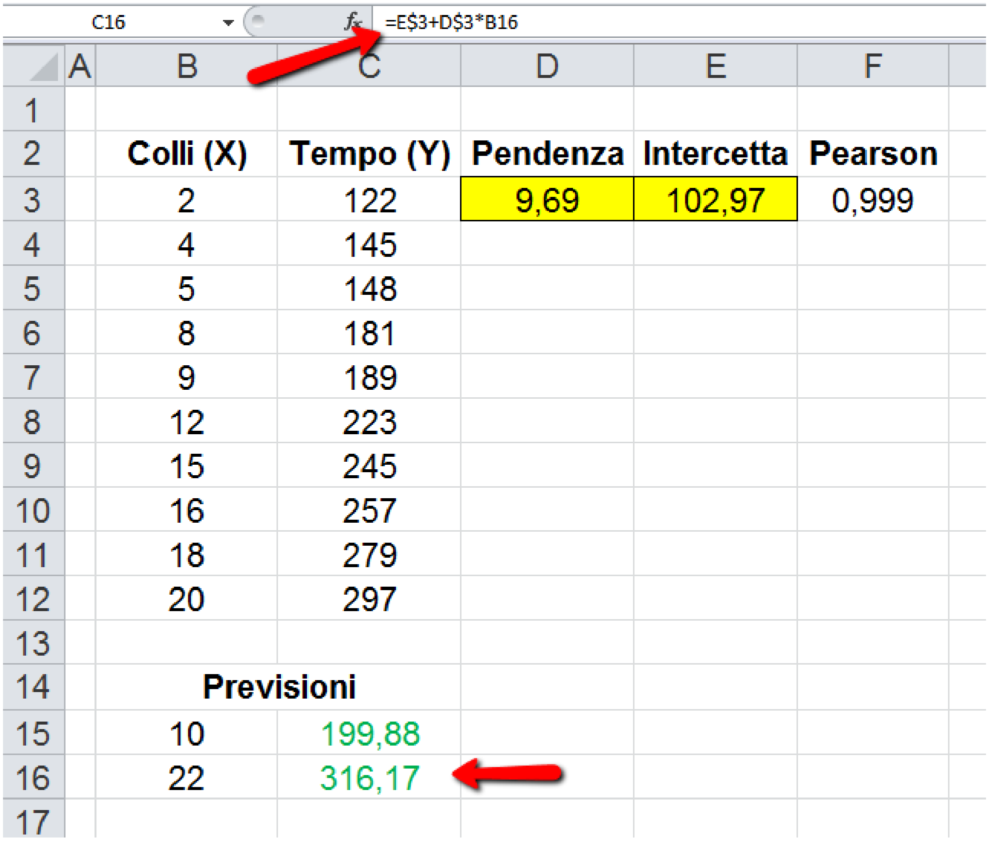
F6 =ERR.STD.YX(C3:C12;B3:B12)

	A	B	C	D	E	F
1						
2		**Colli (X)**	**Tempo (Y)**	**Pendenza**	**Intercetta**	**Pearson**
3		2	122	9,69	102,97	0,999
4		4	145			
5		5	148			**Err. Std**
6		8	181			2,5718
7		9	189			
8		12	223			
9		15	245			
10		16	257			
11		18	279			
12		20	297			
13						

In our example the standard error is 2.57. Compared with the Y values (around 200) it corresponds to about 1%.

A third check is on the slope of the line. Should the line be horizontal, this would indicate that the dependent variable Y actually weakly depends on the value of the independent variable X, which would put some doubt on our initial assumption "Y depends on X". In our example, this does not happen: for different values of X we have definitely different values of Y; in particular, for each additional items to be shipped about ten additional minutes are required (*m = 9.69*).

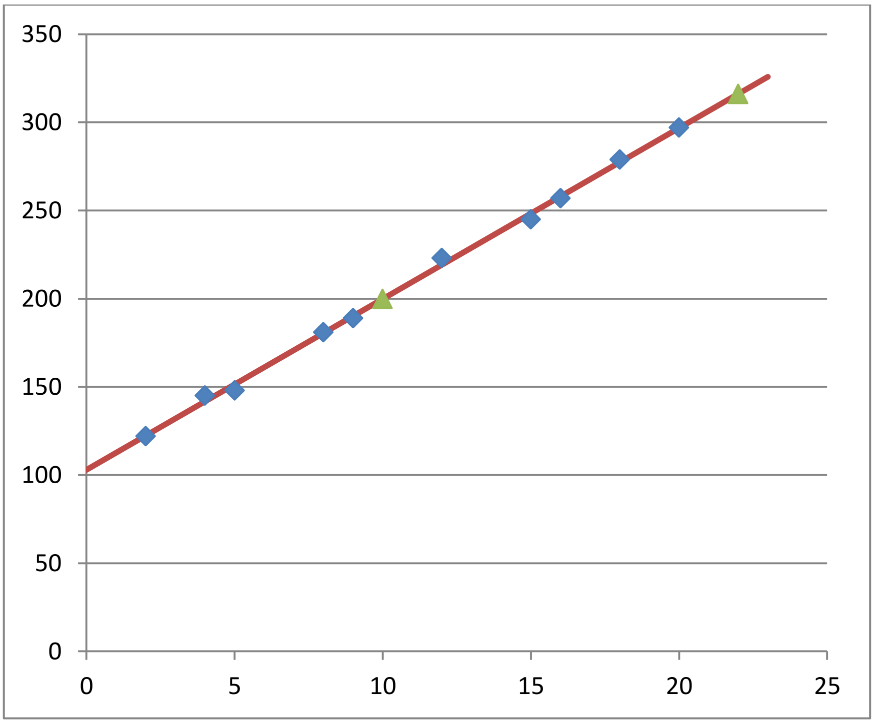
Hence we can reasonably assume that the model we have obtained from linear regression in this example be reliable and we can use it to make *predictions*. For instance, if we want to predict how i twill take to prepare two shipments, one with ten 10 items and the other with 22 items, we can compute the corresponding values of the required time (Y) using the equation of the regression line setting X=10 in one case and X=22 in the other.



C16 =E$3+D$3*B16

	A	B	C	D	E	F
1						
2		**Colli (X)**	**Tempo (Y)**	**Pendenza**	**Intercetta**	**Pearson**
3		2	122	9,69	102,97	0,999
4		4	145			
5		5	148			
6		8	181			
7		9	189			
8		12	223			
9		15	245			
10		16	257			
11		18	279			
12		20	297			
13						
14		**Previsioni**				
15		10	199,88			
16		22	316,17			
17						

If we visualize the two points corresponding to these two forecasts, we can verify that they are exactly on the regression line.



Now we repeat the construction of this simple predictive model with a more general procedure that does not use the predefined functions SLOPE() and INTERCEPT(). This second method consists of formulating and solving an *optimization problem*, that is a problem in which we want to find the values of the *variables* that allow to achieve the maximum or minimum value of an *objective* we want to optimize.

Let us restart from the data, that we have observed to be almost aligned along a straight line.

	A	B	C
1			
2		**Colli (X)**	**Tempo (Y)**
3		2	122
4		4	145
5		5	148
6		8	181
7		9	189
8		12	223
9		15	245
10		16	257
11		18	279
12		20	297
13			

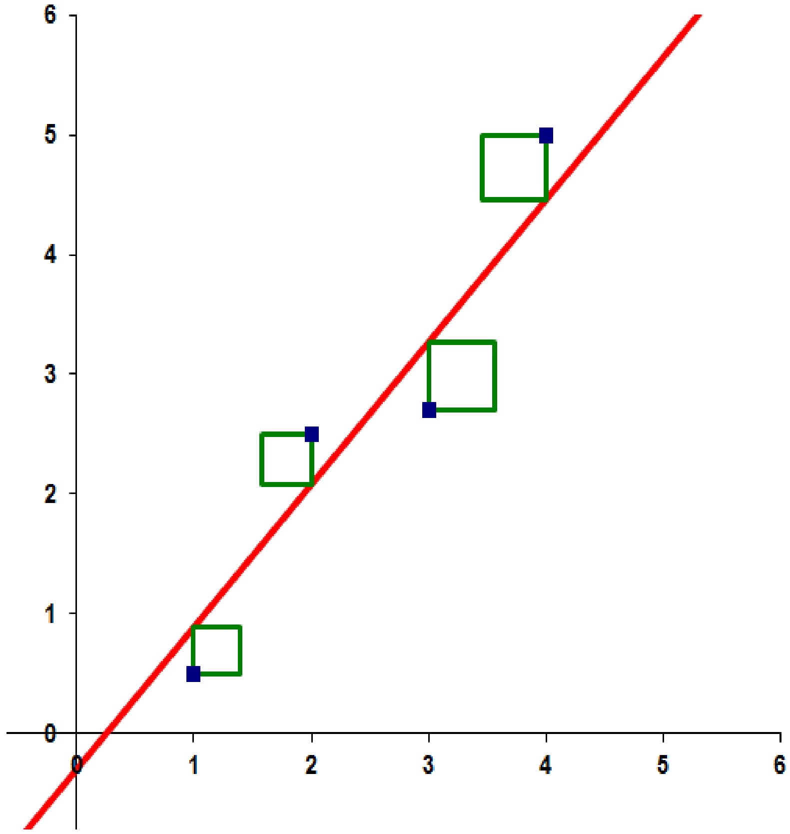
Now we indicate on the spreadsheet the two coefficients, *m* and *q*, we want to determine. These are the variables in our optimization problem.

	A	B	C	D	E	F	G	H
1								
2		**X**	**Y**			**Pendenza**	**Intercetta**	
3		2	122					
4		4	145					
5		5	148					
6		8	181					
7		9	189					
8		12	223					
9		15	245					
10		16	257					
11		18	279					
12		20	297					
13								

Now we must put also the objective somewhere in the spreadsheet. In linear regression we want to minimize the sum (or the average) of the squares of all errors, i.e. the quantity

$$Q = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - (mX_i + q)\right)^2$$

In our example $N=10$ and X_i and Y_i indicate the values of the data for each $i=1,\dots,10$. The meaning of Q is shown in the figure: Q is the average value of the area of the squares whose edge is the difference between the value Y_i (little blue square) and the value of the line for $X=X_i$.

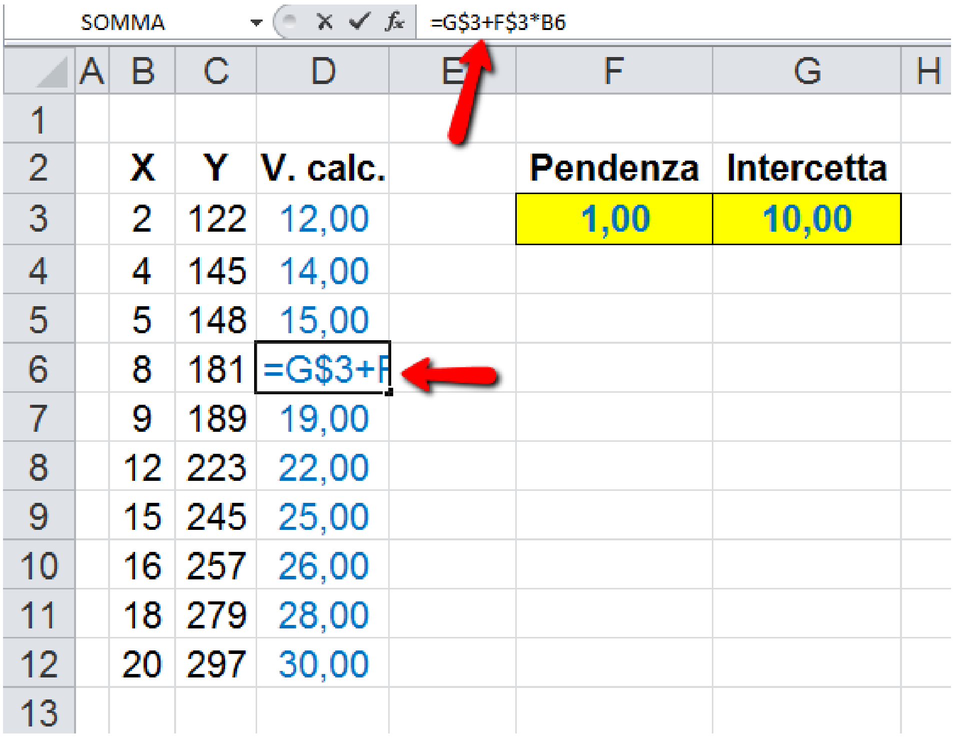


This is why the method is known as “least squares method”.

In an ideal case, if the line passed through all the given points exactly, the resulting value of Q would be zero.

To be able to express Q, we must indicate the values mX_i+q and the corresponding squared errors.

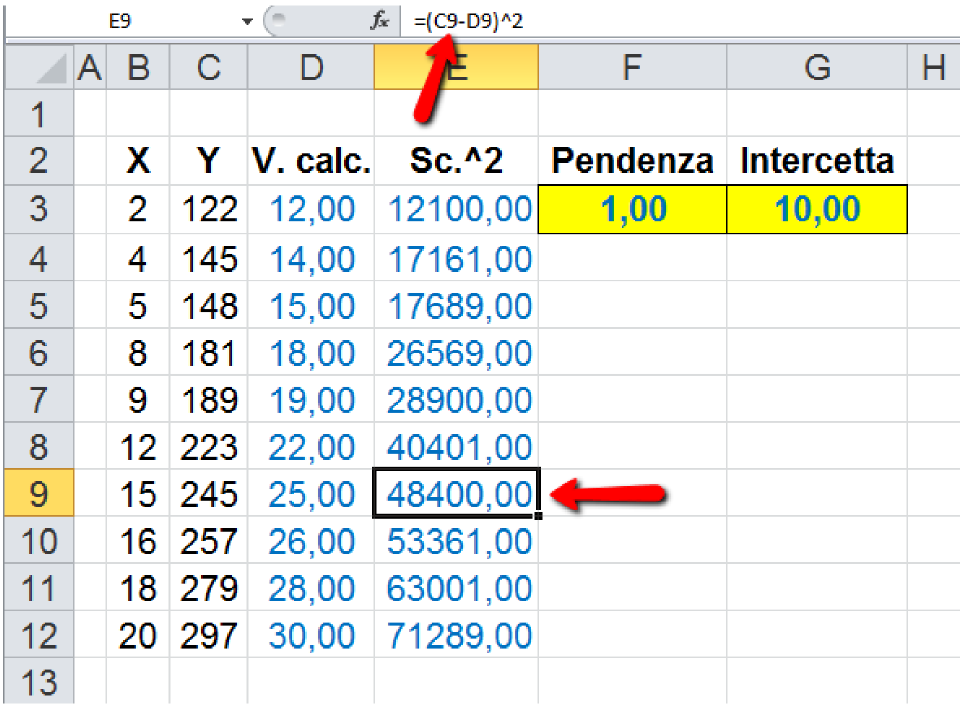
Hence, aside of the column of the *observed values* Y, we insert a column of *computed values*. The values in this column are computed according to m and q, that are initially set to arbitrary values.



SOMMA | =G$3+F$3*B6

	A	B	C	D	E	F	G	H
1								
2		X	Y	V. calc.		Pendenza	Intercetta	
3		2	122	12,00		1,00	10,00	
4		4	145	14,00				
5		5	148	15,00				
6		8	181	=G$3+F				
7		9	189	19,00				
8		12	223	22,00				
9		15	245	25,00				
10		16	257	26,00				
11		18	279	28,00				
12		20	297	30,00				
13								

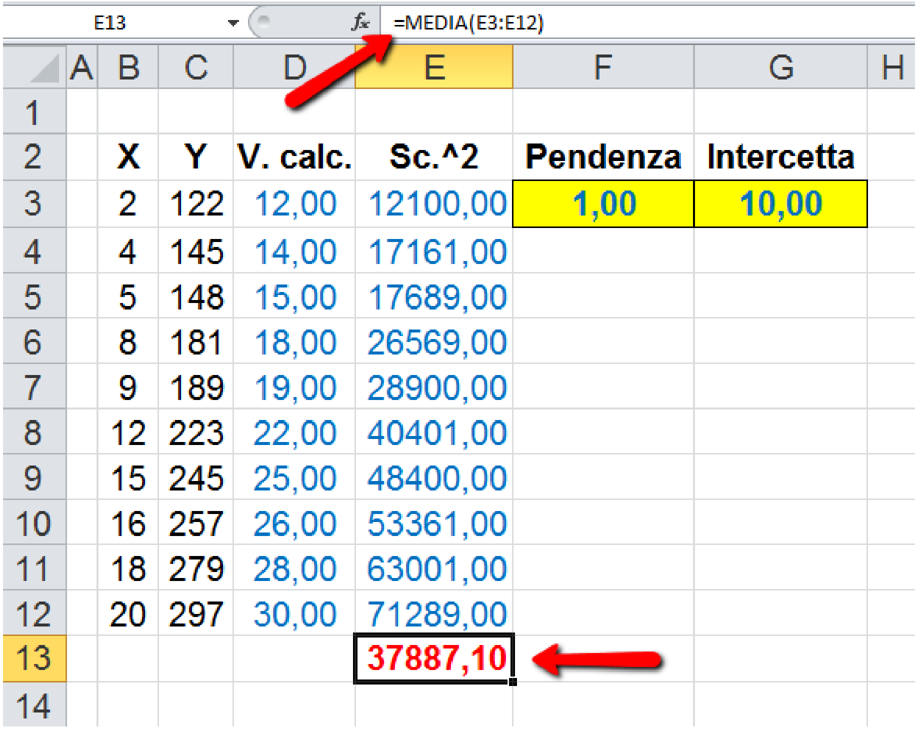
In the next column we can now indicate the squared errors.



E9 | =(C9-D9)^2

	A	B	C	D	E	F	G	H
1								
2		X	Y	V. calc.	Sc.^2	Pendenza	Intercetta	
3		2	122	12,00	12100,00	1,00	10,00	
4		4	145	14,00	17161,00			
5		5	148	15,00	17689,00			
6		8	181	18,00	26569,00			
7		9	189	19,00	28900,00			
8		12	223	22,00	40401,00			
9		15	245	25,00	48400,00			
10		16	257	26,00	53361,00			
11		18	279	28,00	63001,00			
12		20	297	30,00	71289,00			
13								

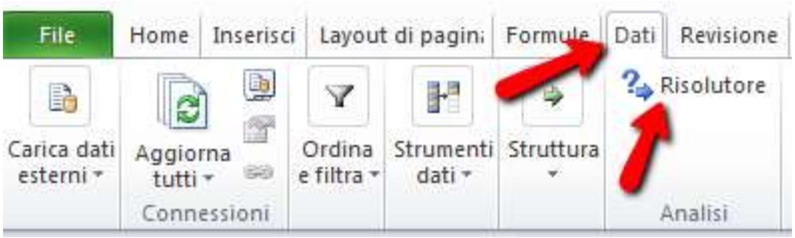
Now we can compute *Q* as the average of the values in the last column.. This is the objective to be minimized.



E13 =MEDIA(E3:E12)

	A	B	C	D	E	F	G	H
1								
2		X	Y	V. calc.	Sc.^2	Pendenza	Intercetta	
3		2	122	12,00	12100,00	1,00	10,00	
4		4	145	14,00	17161,00			
5		5	148	15,00	17689,00			
6		8	181	18,00	26569,00			
7		9	189	19,00	28900,00			
8		12	223	22,00	40401,00			
9		15	245	25,00	48400,00			
10		16	257	26,00	53361,00			
11		18	279	28,00	63001,00			
12		20	297	30,00	71289,00			
13					37887,10			
14								

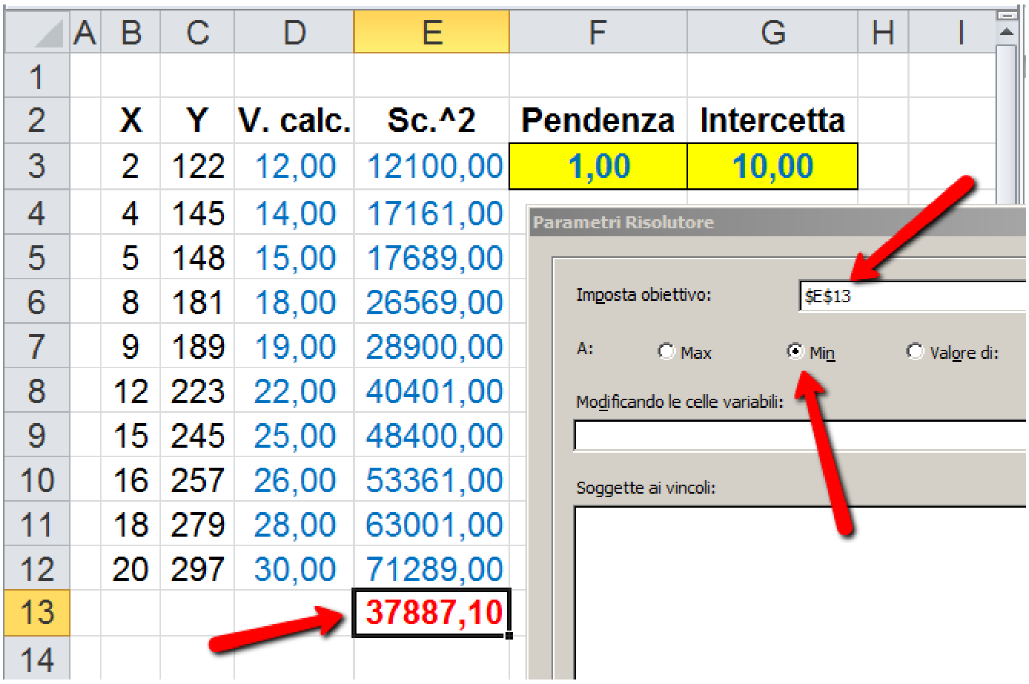
Now, modifying the values of the variables, i.e. the slope *m* and the intercept *q*, the spreadsheet automatically updates the value of the objective, that is *Q*. However, working by trial and error would be boring, time-consuming and would providea approximate results. The optimization problem we have formulated can be solved with the *Solver* add-in. It is included in Microsoft Excel and in LibreOffice, but not in OpenOffice. By default it is inactive; first of all it is necessary to search for it among the add-ins and to activate it. When the Solver is active, it is indicated by a blue icon in the “Data” tab.



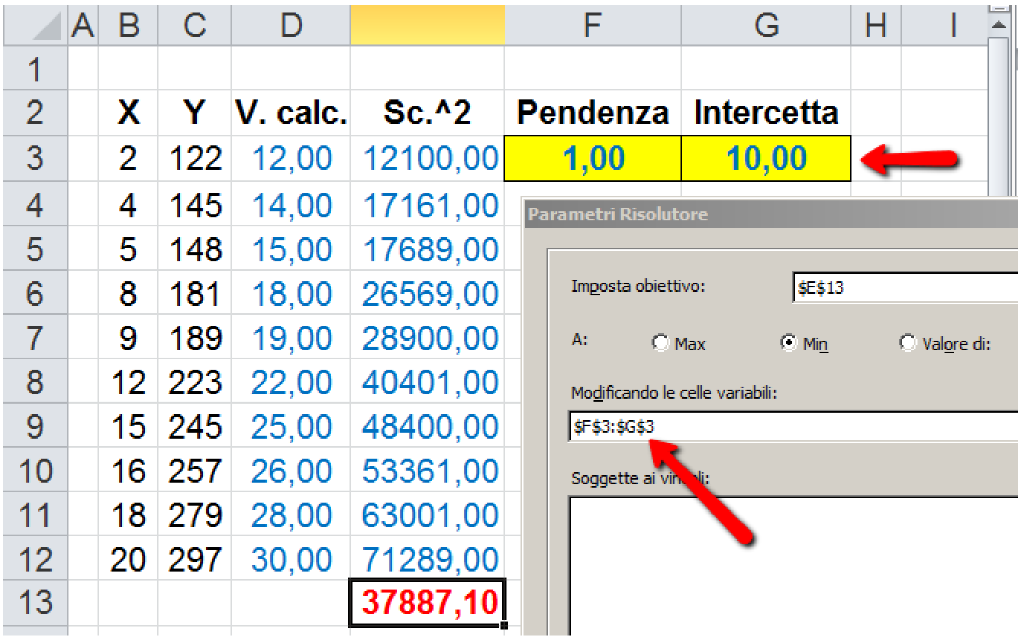
The Solver shows a window in which we can insert the model of the optimization problem we want to solve, specifying its variables, possibly its constraints and its objective.

In our example there are no constraints, but only two variables, the slope *m* and the intercept *q*, and an objective, the cell with *Q*.

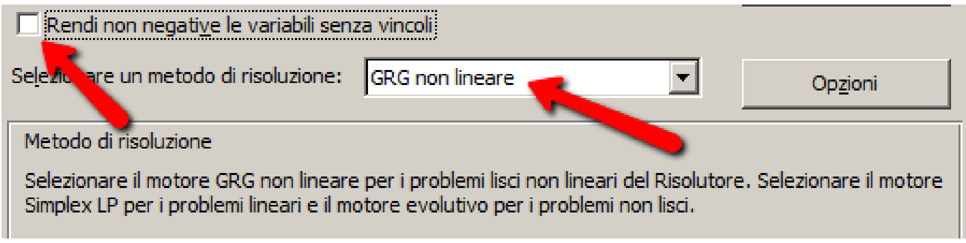
First of all, let us indicate the objective, clicking on the cell with the mean square error.



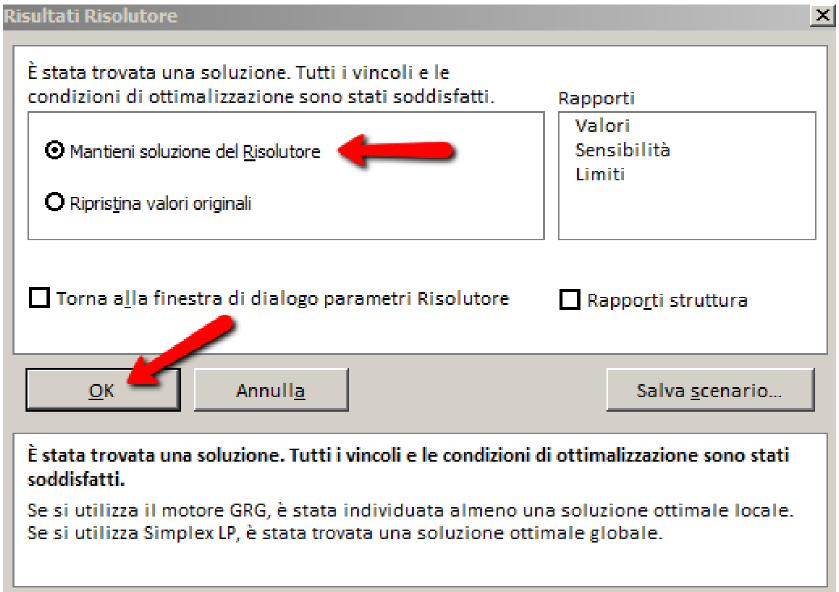
Then we indicate the variable cells (by selecting them with the mouse).



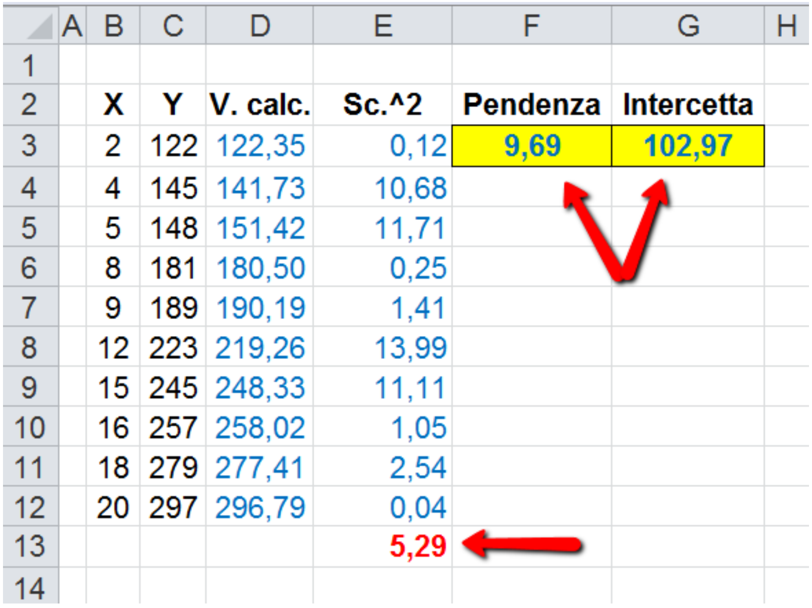
Finally we specify that the variables are not restricted to take on non-negative values and that the problem is non-linear (because the objective depends on the squares of the variables).



Now, clicking on the “Solve” button, the Solver runs an optimization algorithm and provides the result. We accept it by clicking on “OK”.



He solution computed by the Solver is better than the starting one (*Q* is smaller) and the values of slope and intercept correspond to those we had found using the predefined functions SLOPE() and INTERCEPT().



	A	B	C	D	E	F	G	H
1								
2		X	Y	V. calc.	Sc.^2	Pendenza	Intercetta	
3		2	122	122,35	0,12	9,69	102,97	
4		4	145	141,73	10,68			
5		5	148	151,42	11,71			
6		8	181	180,50	0,25			
7		9	189	190,19	1,41			
8		12	223	219,26	13,99			
9		15	245	248,33	11,11			
10		16	257	258,02	1,05			
11		18	279	277,41	2,54			
12		20	297	296,79	0,04			
13					5,29			
14								

As a further check, we can compute the value of the standard error, which is the square root of the variance. The variance in turn is computed like the mean square error, but considering a number of points equal to *N-2*, i.e. 8 instead of 10 in our example.

H6 f_x =RADQ(E13*10/8)

	A	B	C	D	E	F	G	H
1								
2		**X**	**Y**	**V. calc.**	**Sc.^2**	**Pendenza**	**Intercetta**	
3		2	122	122,35	0,12	**9,69**	**102,97**	
4		4	145	141,73	10,68			
5		5	148	151,42	11,71			**Err. Std**
6		8	181	180,50	0,25			2,57
7		9	189	190,19	1,41			
8		12	223	219,26	13,99			
9		15	245	248,33	11,11			
10		16	257	258,02	1,05			
11		18	279	277,41	2,54			
12		20	297	296,79	0,04			
13					**5,29**			

In this way we find the same value of the standard error given by the predefined function of the spreadsheet.

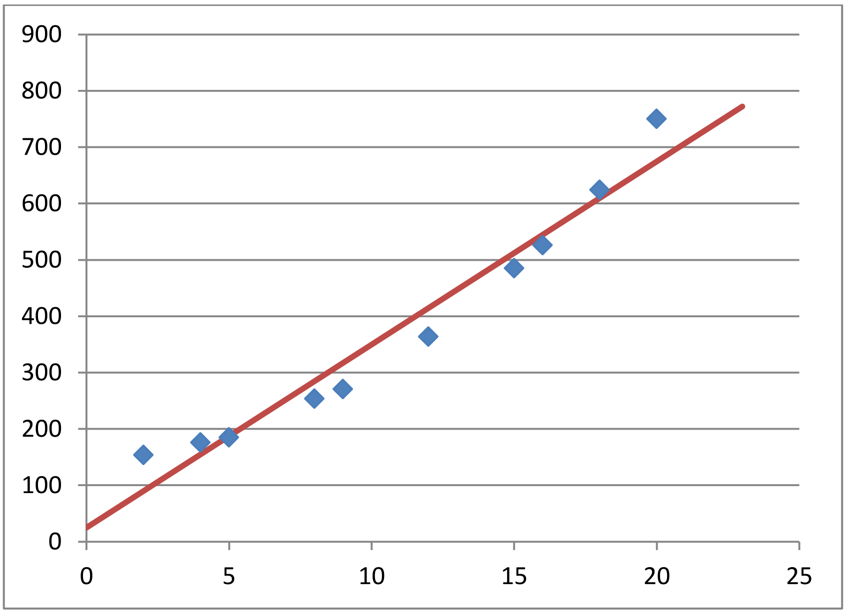
Let us repeat the same analysis on a different data set.

	A	B	C
1			
2		**Colli (X)**	**Tempo (Y)**
3		2	154
4		4	176
5		5	185
6		8	254
7		9	271
8		12	364
9		15	485
10		16	526
11		18	624
12		20	750
13			

Following the same procedure as before, using the predefined functions, we get this model.

	A	B	C	D	E	F
1						
2		**Colli (X)**	**Tempo (Y)**	**Pendenza**	**Intercetta**	**Pearson**
3		2	154	32,48	24,82	0,978
4		4	176			
5		5	185			**Err. Std**
6		8	254			46,37
7		9	271			
8		12	364			
9		15	485			
10		16	526			
11		18	624			
12		20	750			
13						

The standard error turns out to be much larger than before and Pearson index is slightly smaller. This puts some doubts about the model. Visualizing the data in a scatter plot, we see that in this case they do not lie along a line.



The data points are set along a curve. This suggests us to try to interpolate them with a parabola, instead of a line. The procedure is the same; also in this case we want to minimize the mean square error. However we cannot rely on predefined functions of the spreadsheet. We must explicitly represent variables and the objective and then we must use the Solver. The only substantial difference with the previous example is that

- now the variables are the coefficients defining a parabola and hence they are three, not two;
- the computed values are computed according to the equation of a parabola, instead of the equation of a line.

The equation of a parabola is $y = ax^2 + bx + c$. Therefore the variables are now the coefficients a, b and c. Furthermore the computed values are computed as $aX_i^2+bX_i+c$ for each $i=1,\ldots,N$.

As before we initially set the variables with arbitrary values and we define two columns with the computed values and the square errors. Then we compute the average value of the square errors (the cell in red).

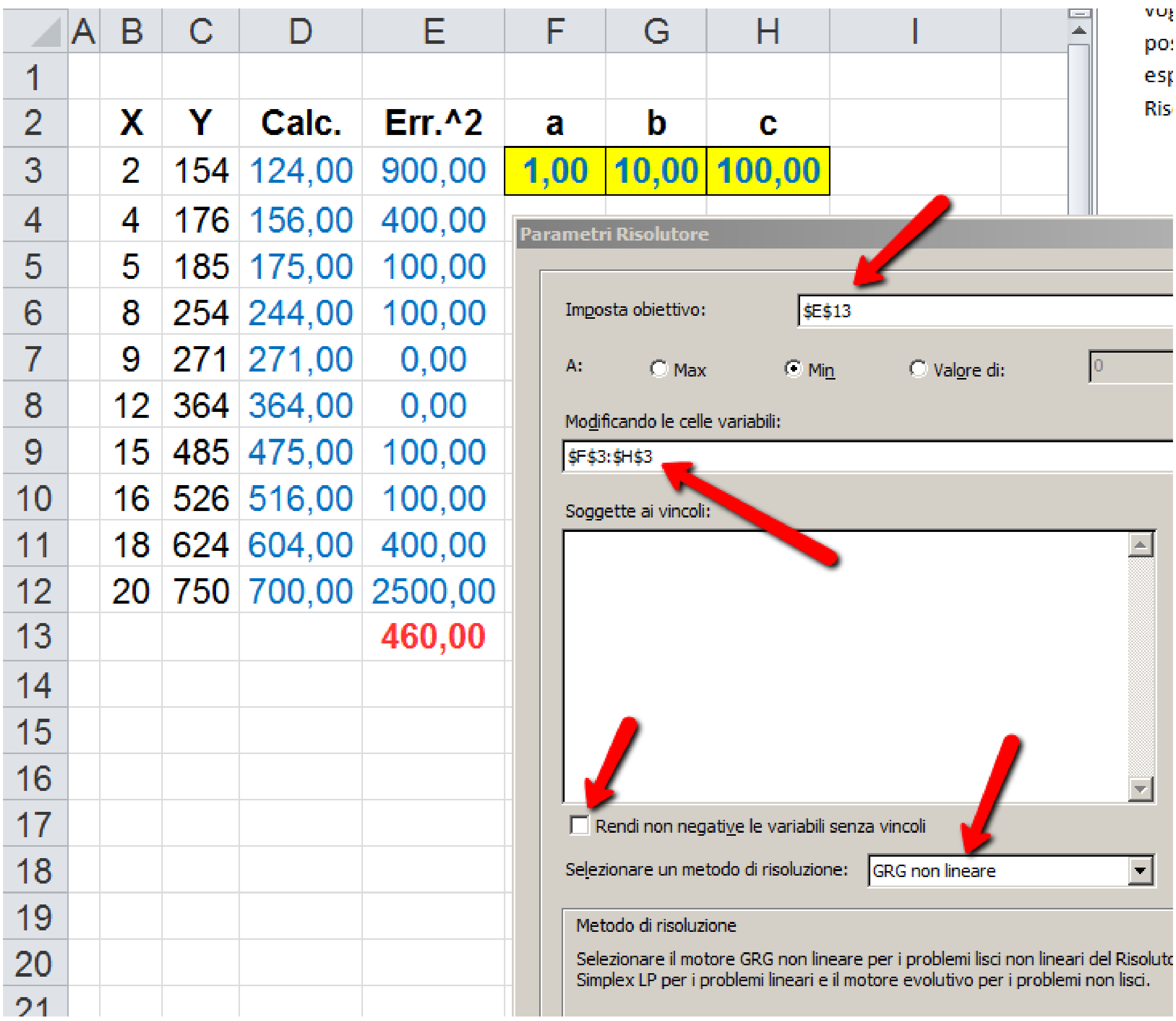
D7 =F$3*B7^2+G$3*B7+H$3

	A	B	C	D	E	F	G	H
1								
2		X	Y	Calc.		a	b	c
3		2	154	124,00		1,00	10,00	100,00
4		4	176	156,00				
5		5	185	175,00				
6		8	254	244,00				
7		9	271	271,00				
8		12	364	364,00				
9		15	485	475,00				
10		16	526	516,00				
11		18	624	604,00				
12		20	750	700,00				
13								

E11 =(C11-D11)^2

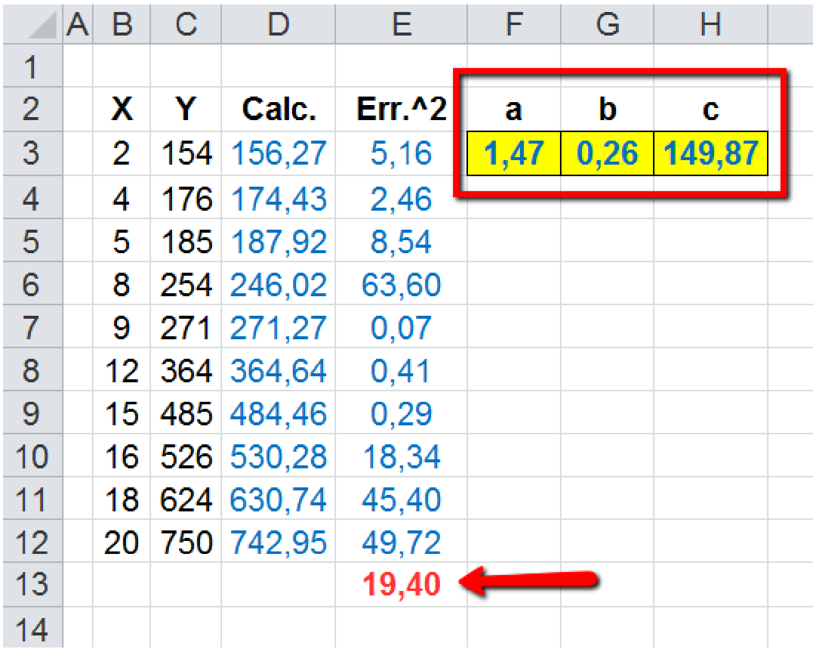
	A	B	C	D	E	F	G	H
1								
2		X	Y	Calc.	Err.^2	a	b	c
3		2	154	124,00	900,00	1,00	10,00	100,00
4		4	176	156,00	400,00			
5		5	185	175,00	100,00			
6		8	254	244,00	100,00			
7		9	271	271,00	0,00			
8		12	364	364,00	0,00			
9		15	485	475,00	100,00			
10		16	526	516,00	100,00			
11		18	624	604,00	400,00			
12		20	750	700,00	2500,00			
13					460,00			
14								

Now we can use the Solver, as before, indicating the three variable cells and the objective cell.



	A	B	C	D	E	F	G	H	I
1									
2		X	Y	Calc.	Err.^2	a	b	c	
3		2	154	124,00	900,00	1,00	10,00	100,00	
4		4	176	156,00	400,00				
5		5	185	175,00	100,00				
6		8	254	244,00	100,00				
7		9	271	271,00	0,00				
8		12	364	364,00	0,00				
9		15	485	475,00	100,00				
10		16	526	516,00	100,00				
11		18	624	604,00	400,00				
12		20	750	700,00	2500,00				
13					460,00				

From the Solver we obtain the optimal solution.



	A	B	C	D	E	F	G	H
1								
2		X	Y	Calc.	Err.^2	a	b	c
3		2	154	156,27	5,16	1,47	0,26	149,87
4		4	176	174,43	2,46			
5		5	185	187,92	8,54			
6		8	254	246,02	63,60			
7		9	271	271,27	0,07			
8		12	364	364,64	0,41			
9		15	485	484,46	0,29			
10		16	526	530,28	18,34			
11		18	624	630,74	45,40			
12		20	750	742,95	49,72			
13					19,40			
14								

The values of *a*, *b* and *c* are the coefficients of the parabola that interpolates the data points in the best way.

I6 f_x =RADQ(E13*10/8)

	A	B	C	D	E	F	G	H	I
1									
2		**X**	**Y**	**Calc.**	**Err.^2**	**a**	**b**	**c**	
3		2	154	156,27	5,16	**1,47**	**0,26**	**149,87**	
4		4	176	174,43	2,46				
5		5	185	187,92	8,54				**Err. Std**
6		8	254	246,02	63,60				4,924
7		9	271	271,27	0,07				
8		12	364	364,64	0,41				
9		15	485	484,46	0,29				
10		16	526	530,28	18,34				
11		18	624	630,74	45,40				
12		20	750	742,95	49,72				
13					**19,40**				
14									

The standard error, that we can compute as before, turns out to be much smaller compared to the one obtained from a straight line (about 5 instead of about 46: almost one tenth). This confirms that the model based on a parabola is more precise than the model based on a line with these data; hence it is likely to provide more reliable forecasts.

Predictions can be made as before: given a value for X, using the coefficients of the parabola we have found, we can compute the corresponding value for Y.

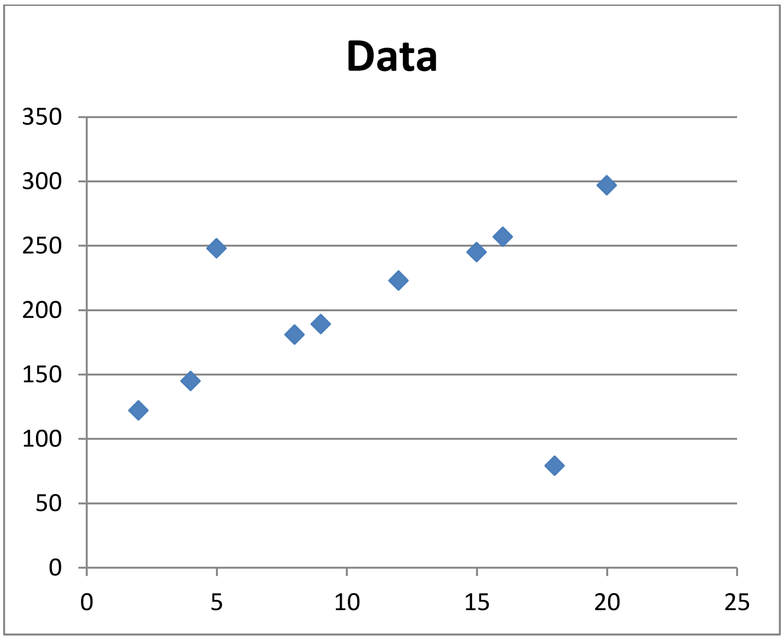
C17 f_x =F$3*B17^2+G$3*B17+H$3

	A	B	C	D	E	F	G	H	I
1									
2		**X**	**Y**	**Calc.**	**Err.^2**	**a**	**b**	**c**	
3		2	154	156,27	5,16	**1,47**	**0,26**	**149,87**	
4		4	176	174,43	2,46				
5		5	185	187,92	8,54				**Err. Std**
6		8	254	246,02	63,60				4,924
7		9	271	271,27	0,07				
8		12	364	364,64	0,41				
9		15	485	484,46	0,29				
10		16	526	530,28	18,34				
11		18	624	630,74	45,40				
12		20	750	742,95	49,72				
13					**19,40**				
14									
15		**Previsioni**							
16		10	299,46						
17		22	866,92						
18									

Formulating an optimization problem and then solving it with a Solver, is a more general method with respect to the use of pre-defined formulae, not only when one uses a regression function different from a straight line, but also for other reasons: for instance, when we want to automatically identify *outliers*, i.e. data points that must be considered as exceptions and therefore must not be used to define the predictive model because they would harm it.

Let us go back to the first data-set and let assume that a couple of Y data had been recorded incorrectly.

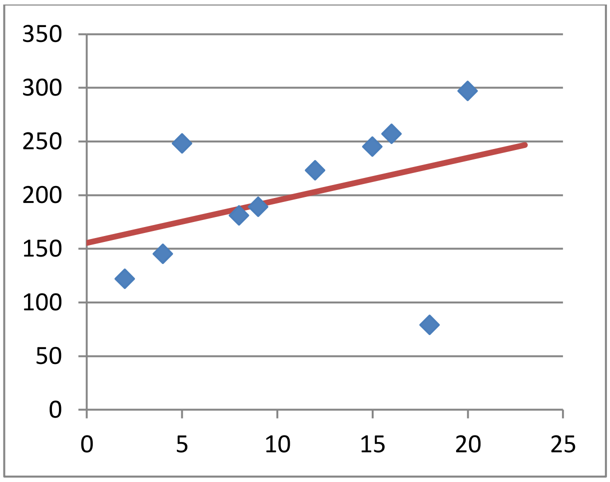
X	Y
2	122
4	145
5	248
8	181
9	189
12	223
15	245
16	257
18	79
20	297



Obviously we assume not to know how many and which data are “wrong” (they can be easily detected from the scatter plot).

Doing the regression analysis with the predefined functions of the spreadsheet, we get this solution. We note that the model is of bad quality and it does not allow for reliable forecasts: the Pearson index has a very small value and the standard error is large compared to the data (a 67 minutes error in forecasting the duration of an operation that takes about 200 minutes corresponds to a 33% error).

X	Y	Pendenza	Intercetta	Pearson
2	122	3,96	155,40	0,363
4	145			
5	248			Err. Std
8	181			67,2844
9	189			
12	223			
15	245			
16	257			
18	79			
20	297			



Then we have recourse to the Solver, but we modify the model we used so far, inerting some additional variables. In particular, we want to express the decision of considering or neglecting each point. The neglected data points are labelled as outliers that “disturb” the analysis and the construction of a reliable model. For this purpose we use a binary variable for each pair (X_i,Y_i) of data and we indicate it with z_i. Initially the values of all these variables are set arbitrarily to 0 or 1.

If $z_i=1$ the datum is used; if $z_i=0$ the datum is considered as an outlier and neglected. When z=0, the outliers yield a null effect to eht computation of the mean square error (see lines 8, 9 and 10 in the figure). On the next column in the spreadsheet we report the square errors actually used to compute the objective.

G6 =E6*F6

	A	B	C	D	E	F	G	H	I	J
1										
2		X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta	
3		2	122	163,33	1708,13	1	1708,13	3,96	155,40	
4		4	145	171,26	689,35	1	689,35			
5		5	248	175,22	5297,16	1	5297,16			
6		8	181	187,11	37,30	1	37,30			
7		9	189	191,07	4,29	1	4,29			
8		12	223	202,96	401,64	0	0,00			
9		15	245	214,85	909,14	0	0,00			
10		16	257	218,81	1458,40	0	0,00			
11		18	79	226,74	21826,17	1	21826,17			
12		20	297	234,66	3885,93	1	3885,93			
13					3621,75					
14										

Now the objective to be optimized is till the mean square error but the average must be computed only on the meaningful termsm negleting those considered as outliers. Then we need to know how many data points are considered. For this it is enough to sum up the *z* variables.

F13 =SOMMA(F3:F12)

	A	B	C	D	E	F	G	H	I	J
1										
2		X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta	
3		2	122	163,33	1708,13	1	1708,13	3,96	155,40	
4		4	145	171,26	689,35	1	689,35			
5		5	248	175,22	5297,16	1	5297,16			
6		8	181	187,11	37,30	1	37,30			
7		9	189	191,07	4,29	1	4,29			
8		12	223	202,96	401,64	0	0,00			
9		15	245	214,85	909,14	0	0,00			
10		16	257	218,81	1458,40	0	0,00			
11		18	79	226,74	21826,17	1	21826,17			
12		20	297	234,66	3885,93	1	3885,93			
13					3621,75	7				
14										

The mean square error to be minimized is now given by the sum of all square errors divided by the number of significant terms in the sum.

G13 =SOMMA(G3:G12)/F13

	A	B	C	D	E	F	G	H	I	J
1										
2		X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta	
3		2	122	163,33	1708,13	1	1708,13	3,96	155,40	
4		4	145	171,26	689,35	1	689,35			
5		5	248	175,22	5297,16	1	5297,16			
6		8	181	187,11	37,30	1	37,30			
7		9	189	191,07	4,29	1	4,29			
8		12	223	202,96	401,64	0	0,00			
9		15	245	214,85	909,14	0	0,00			
10		16	257	218,81	1458,40	0	0,00			
11		18	79	226,74	21826,17	1	21826,17			
12		20	297	234,66	3885,93	1	3885,93			
13					3621,75	7	4778,33			
14										

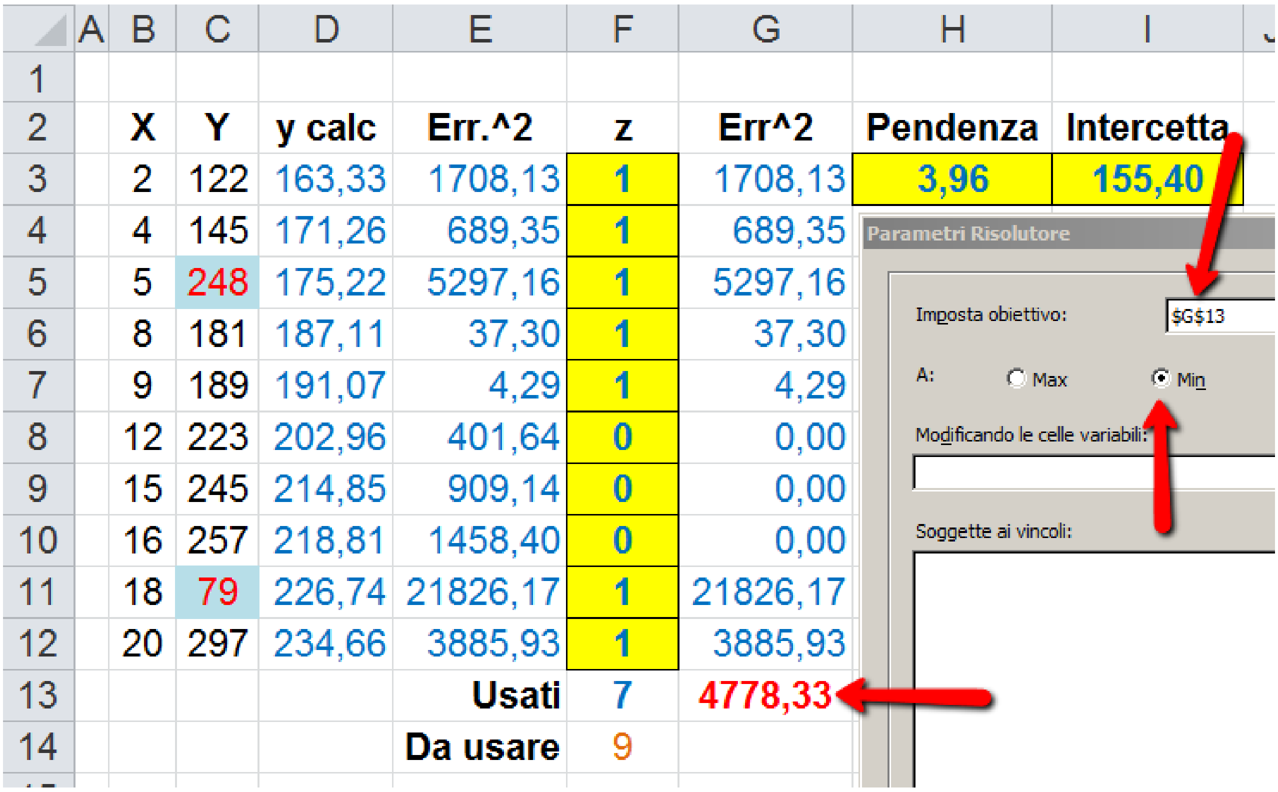
As a last step we must decide the number of allowed outliers. Let us try first with one allowed outlier; therefore we impose to consider at least 9 of the 10 available data points. We insert this parameter in the spreadsheet.

F14 9

	A	B	C	D	E	F	G	H	I	J
1										
2		X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta	
3		2	122	163,33	1708,13	1	1708,13	3,96	155,40	
4		4	145	171,26	689,35	1	689,35			
5		5	248	175,22	5297,16	1	5297,16			
6		8	181	187,11	37,30	1	37,30			
7		9	189	191,07	4,29	1	4,29			
8		12	223	202,96	401,64	0	0,00			
9		15	245	214,85	909,14	0	0,00			
10		16	257	218,81	1458,40	0	0,00			
11		18	79	226,74	21826,17	1	21826,17			
12		20	297	234,66	3885,93	1	3885,93			
13					Usati	7	4778,33			
14					Da usare	9				

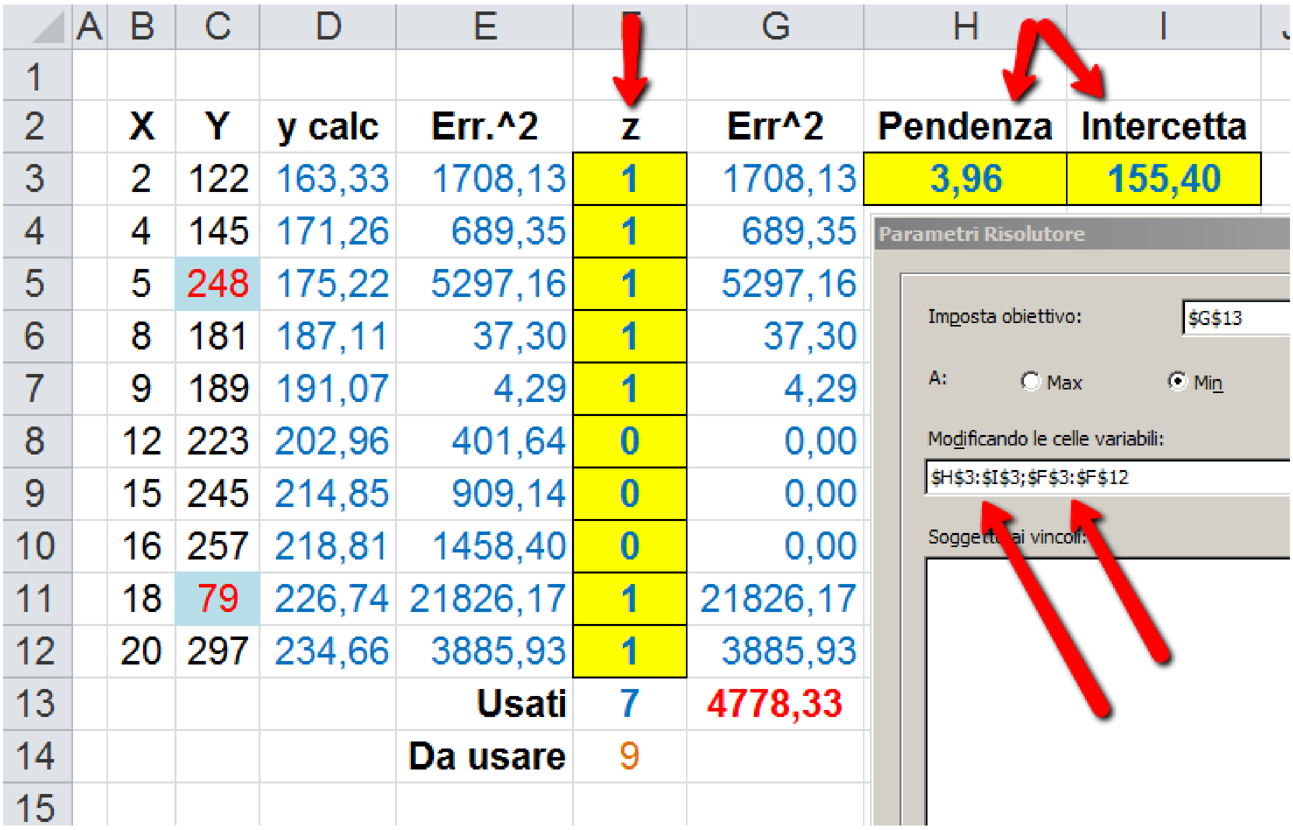
Now we can call the Solver and we can describe this new optimization problem. It has two continuous variables, slope and intercept, and ten binary variables z; its objective if the mean square error (cell G13) and there is a contraint on the minimum number of data points to consider (at least 9).

We insert all these pieces of information in the Solver mask. Let us start with the objective.



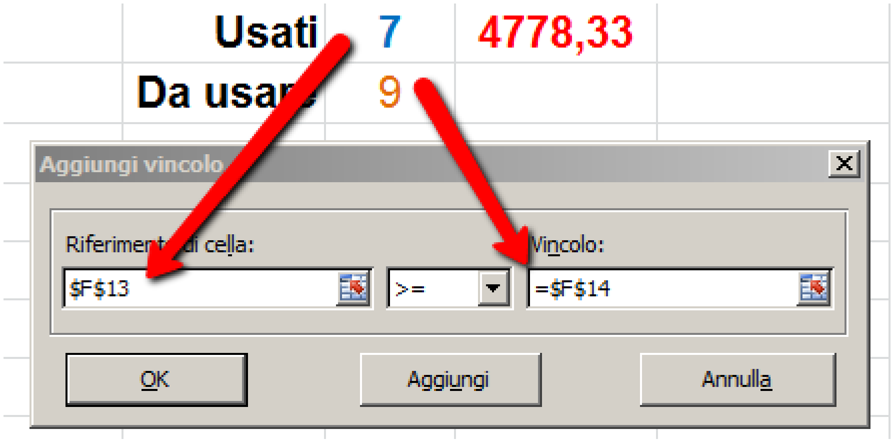
	X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta
3	2	122	163,33	1708,13	1	1708,13	3,96	155,40
4	4	145	171,26	689,35	1	689,35		
5	5	248	175,22	5297,16	1	5297,16		
6	8	181	187,11	37,30	1	37,30		
7	9	189	191,07	4,29	1	4,29		
8	12	223	202,96	401,64	0	0,00		
9	15	245	214,85	909,14	0	0,00		
10	16	257	218,81	1458,40	0	0,00		
11	18	79	226,74	21826,17	1	21826,17		
12	20	297	234,66	3885,93	1	3885,93		
13				Usati	7	4778,33		
14				Da usare	9			

The, we indicate the variables (the two vectors are separated by a semicolon).

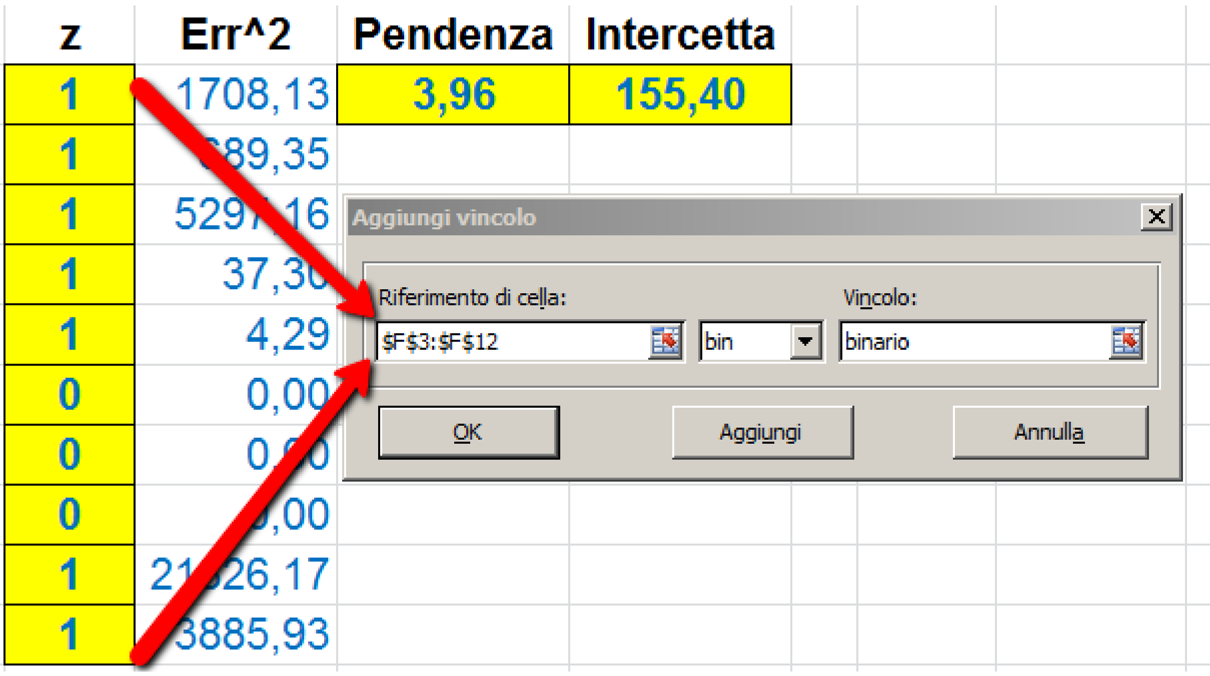


	X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta
3	2	122	163,33	1708,13	1	1708,13	3,96	155,40
4	4	145	171,26	689,35	1	689,35		
5	5	248	175,22	5297,16	1	5297,16		
6	8	181	187,11	37,30	1	37,30		
7	9	189	191,07	4,29	1	4,29		
8	12	223	202,96	401,64	0	0,00		
9	15	245	214,85	909,14	0	0,00		
10	16	257	218,81	1458,40	0	0,00		
11	18	79	226,74	21826,17	1	21826,17		
12	20	297	234,66	3885,93	1	3885,93		
13				Usati	7	4778,33		
14				Da usare	9			

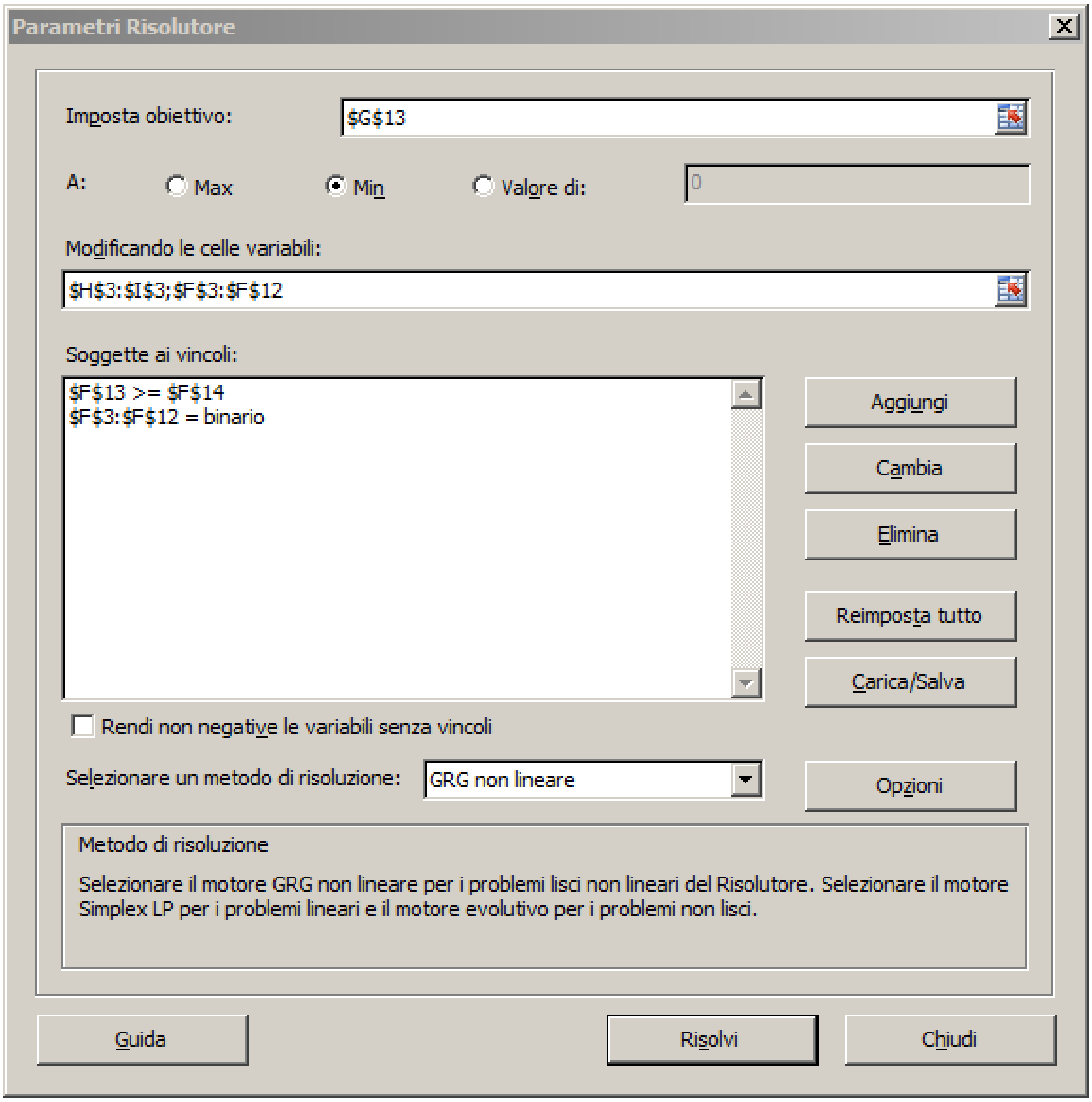
Then, clicking on "Insert/Add" we can insert the constraint on the number of data points to be used. The constraint is expressed by a simple inequality: the content of cell F13 (number of data points used) must be greater than or equal to the content of cell F14 (minimum required number of data points).



Besides this constraint, we must also specify that all z variables are binary. This requirement is inserted as a constraint.



The complete model is the following.



The Solver computes this optimal solution.

	A	B	C	D	E	F	G	H	I	J
1										
2		X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta	
3		2	122	147,69	659,85	1	659,85	7,92	131,86	
4		4	145	163,52	342,92	1	342,92			
5		5	248	171,43	5862,46	1	5862,46			
6		8	181	195,18	201,04	1	201,04			
7		9	189	203,09	198,65	1	198,65			
8		12	223	226,84	14,74	1	14,74			
9		15	245	250,59	31,20	1	31,20			
10		16	257	258,50	2,25	1	2,25			
11		18	79	274,33	38154,31	0	0,00			
12		20	297	290,16	46,76	1	46,76			
13					Usati	9	817,765			
14					Da usare	9				
15										

The Solve has used the minimum number of data opints, i.e. 9, and it has identified as an outlier one of the two data points one of the two points affected by an error. Hwever, we are not satisfied with this model because the mean square error is relatively large. We can compute the standard error. In this example it must be computed dividing by 7, because 9 data points are used.

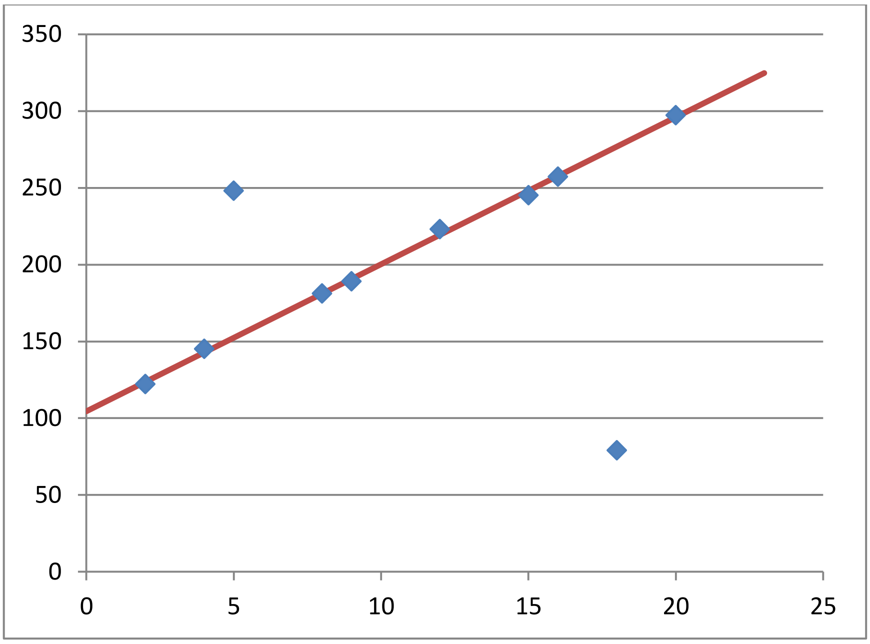
J6 =RADQ(G13*F13/(F13-2))

	A	B	C	D	E	F	G	H	I	J
1										
2		X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta	
3		2	122	147,69	659,85	1	659,85	7,92	131,86	
4		4	145	163,52	342,92	1	342,92			
5		5	248	171,43	5862,46	1	5862,46			Err. Std
6		8	181	195,18	201,04	1	201,04			32,425
7		9	189	203,09	198,65	1	198,65			
8		12	223	226,84	14,74	1	14,74			
9		15	245	250,59	31,20	1	31,20			
10		16	257	258,50	2,25	1	2,25			
11		18	79	274,33	38154,31	0	0,00			
12		20	297	290,16	46,76	1	46,76			
13					Usati	9	817,765			
14					Da usare	9				
15										

The standard error is definitely bad (too large). Then we try again, allowing the Solver to neglect two outiers. So, we insert 8 instead of 9 in cell F14 and we solve the model again (the model remains the same: we must simply click the “Solve” button again). The solution we obtain is like this.

	A	B	C	D	E	F	G	H	I	J
1										
2		X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta	
3		2	122	123,61	2,58	1	2,58	9,57	104,46	
4		4	145	142,75	5,04	1	5,04			
5		5	248	152,33	9153,20	0	0,00			Err. Std
6		8	181	181,05	0,00	1	0,00			2,400
7		9	189	190,62	2,63	1	2,63			
8		12	223	219,34	13,38	1	13,38			
9		15	245	248,06	9,38	1	9,38			
10		16	257	257,64	0,40	1	0,40			
11		18	79	276,78	39117,90	0	0,00			
12		20	297	295,93	1,15	1	1,15			
13					Usati	8	4,32097			
14					Da usare	8				
15										

Now the standard errori s very small: the model interpolates the data very well. We have got this result at the price of neglecting two data points, namely those for which the variable z has been set to 1 by the Solver. But they are just the two data that were affected by an error! The Solver identified them and , neglecting them, it could find a regression line that interpolates the remaining data very well. As expected, this line is quite similar to the regression line we had computed in the case with no errors in the data.



Now we have obtained a reliable model, we can use it to make predictions, as we have already done before.

C17 =I$3+H$3*B17

	A	B	C	D	E	F	G	H	I
1									
2		X	Y	y calc	Err.^2	z	Err^2	Pendenza	Intercetta
3		2	122	123,61	2,58	1	2,58	9,57	104,46
4		4	145	142,75	5,04	1	5,04		
5		5	248	152,33	9153,20	0	0,00		
6		8	181	181,05	0,00	1	0,00		
7		9	189	190,62	2,63	1	2,63		
8		12	223	219,34	13,38	1	13,38		
9		15	245	248,06	9,38	1	9,38		
10		16	257	257,64	0,40	1	0,40		
11		18	79	276,78	39117,90	0	0,00		
12		20	297	295,93	1,15	1	1,15		
13					Usati	8	4,32097		
14					Da usare	8			
15		Previsioni							
16		10	200,19						
17		22	315,08