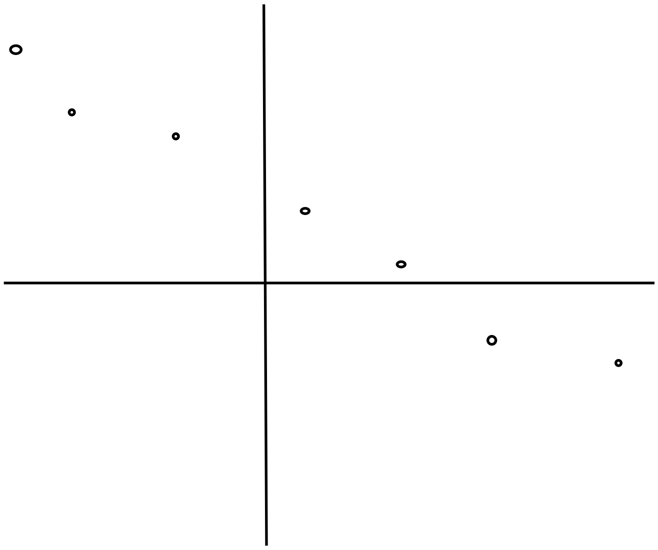
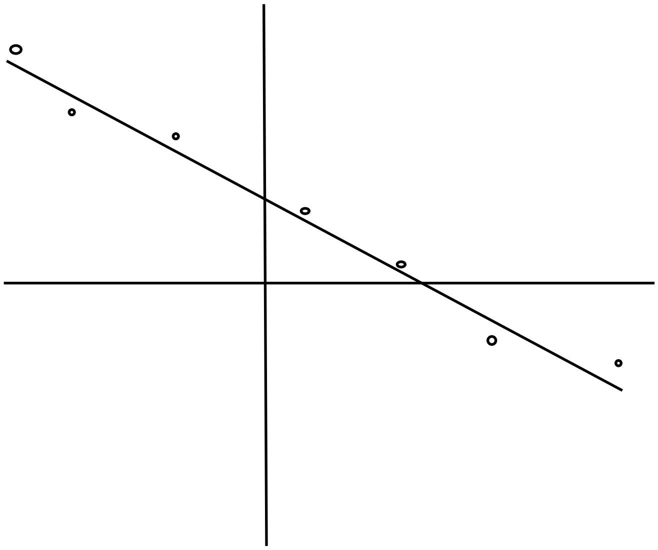
Math 181: The regression line

Sometimes when one is given data, the data have a linear tendency. For example, suppose we look at the following "scatterplot" of points:

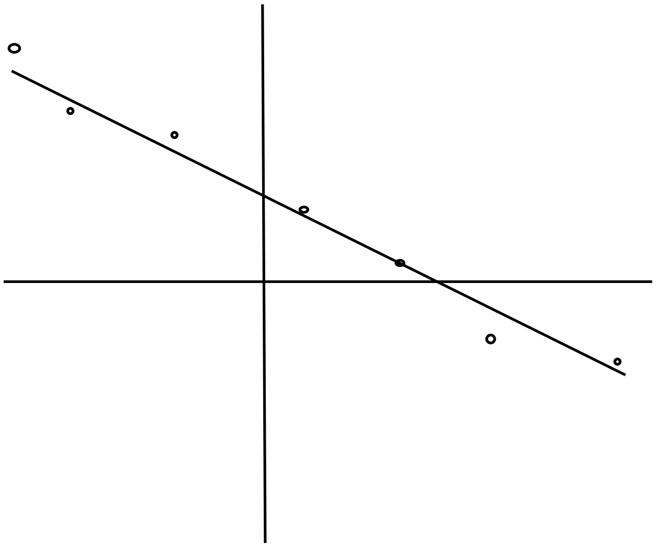


We see that the points are approximately on a straight line, but there is no true line that contains all of them. Perhaps, such data might arise from an experiment in which some small effects or random errors perturb results from a true linear relationship. We wish to find an overall line that describes the linear tendency.

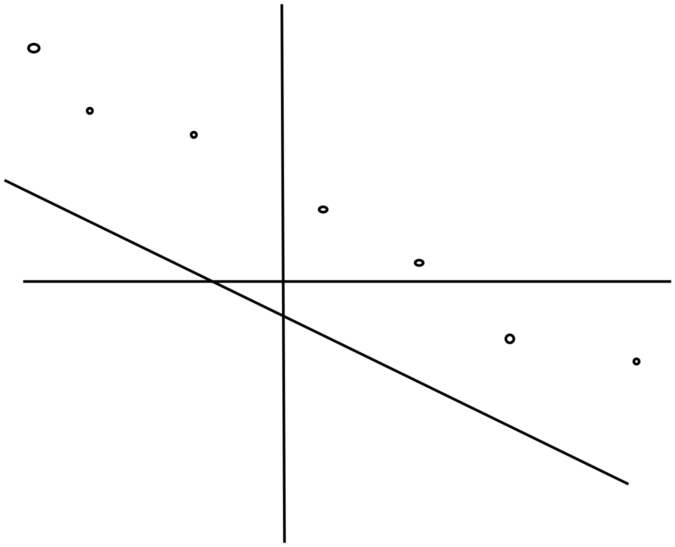
By eye, we can try to draw a line. For example, the following line



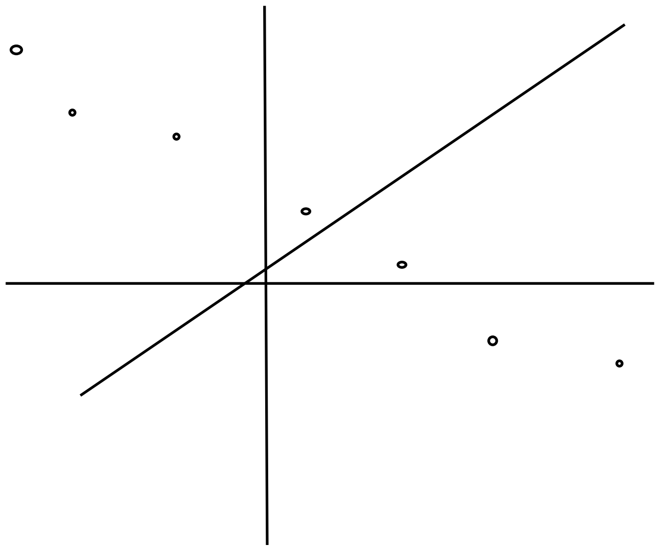
does pretty well. The line passes below some of the data points and above others. Another attempt might lead to



which differs slightly but also seems plausible. On the other hand,



is clearly not as good since all the data points are on the same side of the line, and



is not very good since its slope does not correspond to the linear tendency that is clear in the data, even though approximately half the points are above the line and half are below the line.

Once we have drawn such a "line of best fit" by eye, we can estimate the coordinates of two points on the line and find its equation.

But which line is the "right" line? The first two above were clearly pretty good, and the last two were clearly pretty bad. But there are lots of pretty good lines. How do we choose between the first and second line?

Using the eye alone to find the line can lead to fairly good answers, but there is a mathematically precise method that leads to a line that frequently is accepted as "best." This line is often called the "regression" line or the "least squares" line. There is a procedure in most scientific calculators and many software packages that finds this mathematically precise regression line when the data are entered.

Example. Find the regression line (or line of best fit) for the data

x	y
3	5.00
4	7.10
5	9.40
4.5	8.00

I'll demonstrate the procedure using the TI85 or TI86 calculator. Most other calculators have similar procedures.

Step 1: On either calculator, prepare a "list" for the x-data and another list for the y-data:

```
2nd LIST
{3,4,5,4.5}
STO    L1

{5,7.1,9.4,8.0}
```

STO L2

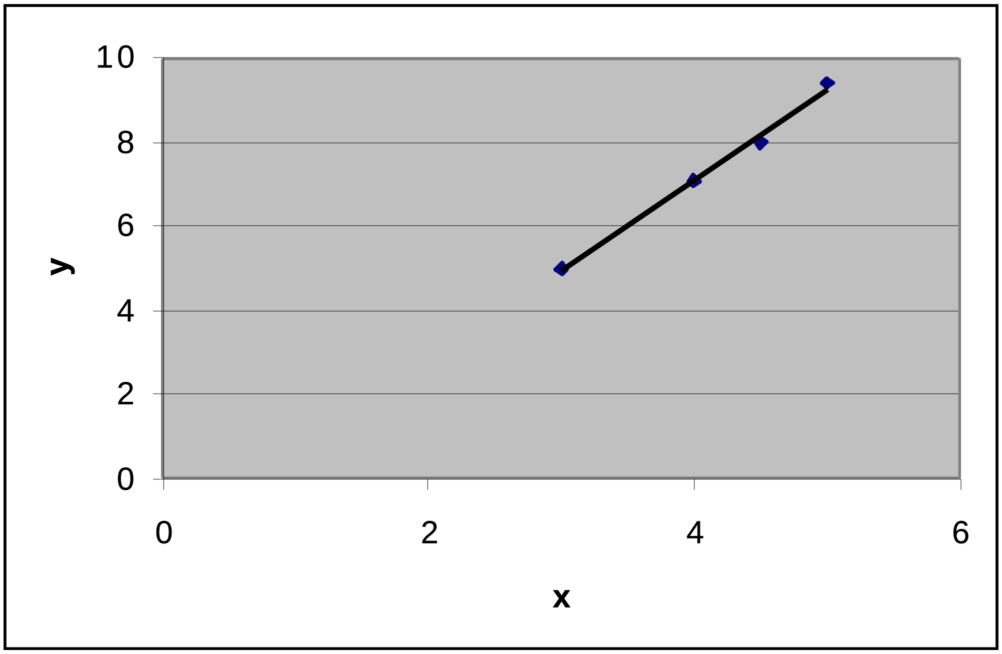
Step 2: Go to the STAT menu to do the calculations:
STAT
CALC

Step 3: (a) On TI85:
xlist Name = L1 <ENTER>
ylist Name = L2 <ENTER>
LINR
Read off the line a = -1.51142857143 b=2.15428571429

(b) On the TI86
LinR L1,L2 <ENTER>
Read off the line: y = a+bx
a = -1.5114286 b=2.15428571

Step 4: Write the answer. Since our data only have accuracy to 3 significant figures, we give our answer only to comparable accuracy:
$y = -1.51 + 2.15x$
is the equation of the regression line.

How do we tell whether the fit is good? There are two checks:
(Check1) Draw a graph in which you have the data and the line. The points should be close to the line. For example, we obtain



Alternatively, on the graphing calculator locate the various points by moving the cursor using the arrow keys: Is (3,5) on the graph? Move the cursor to x = 3, y = 5 to check.

(Check 2) We form a table to ask whether the values are reasonable.
When we have a regression line, then we can use the regression line to predict a value for y when each value of x is given. We might call this the predicted value.

For each data point x, the **residual** at x is
residual = data value - predicted value

In this example, the residual at 3 is computed as follows:

data value when x = 3 is 5
predicted value when x = 3 is y = -1.51 + 2.15x = -1.51+2.15(3) = 4.94
Hence the residual is 5 - 4.94 = 0.06

Example. Find the residual when x = 4.
Solution. When x = 4, the predicted value is -1.51+2.15(4) = 7.09
Hence the residual is 7.1-7.09 = 0.01

In this manner we form a table containing for each data point the numbers x, y, predicted y, and the residual:

x	y	predicted	residual
3	5	4.94	0.06
4	7.1	7.09	0.01
5	9.4	9.24	0.16
4.5	8	8.165	-0.165

If the fit is good , then all the residuals should be small. There should not be a tendency to be bigger on one end and smaller on another. In this example, the fit is pretty good.

To find these predicted values quickly on the calculator, type Y = -1.51 +2.15*X <ENTER>
Now 3 STO X <ENTER>
Y <ENTER> leads to 4.94
4 STO X <ENTER>
Y <ENTER> leads to 7.09.
In this manner we quickly find the values predicted for the data on the regression line.

Now I'll demonstrate the procedure on the computer, using EXCEL:
Step 1. Place the data into two adjacent columns: For example, the following might be in columns B1:B5 and C1:C5:

x	y
3	5.00
4	7.10
5	9.40
4.5	8.00

Step 2. Find the coefficients of the regression line.

Suppose that the numerical values for x are in B2:B5 while those for y are in C2:C5. Then in a new cell (say B7) you can find the slope of the regression line as

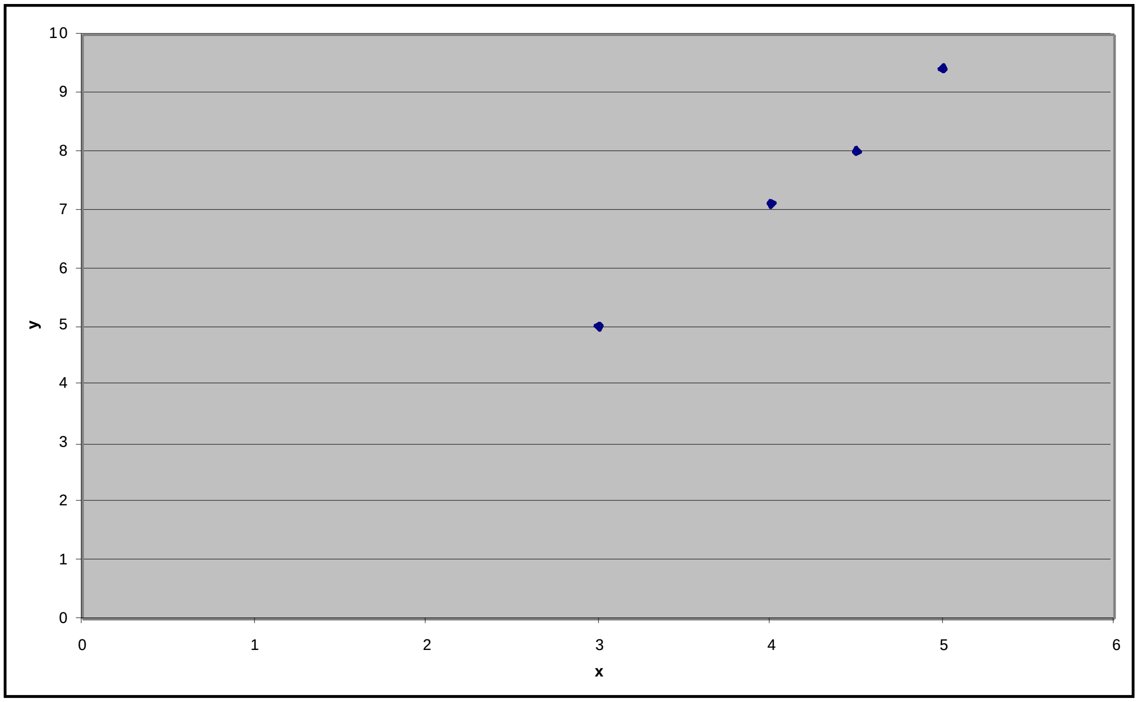
= SLOPE(C2:C5,B2:B5)

and in another new cell (say B8) you can find the y-intercept of the regression line as

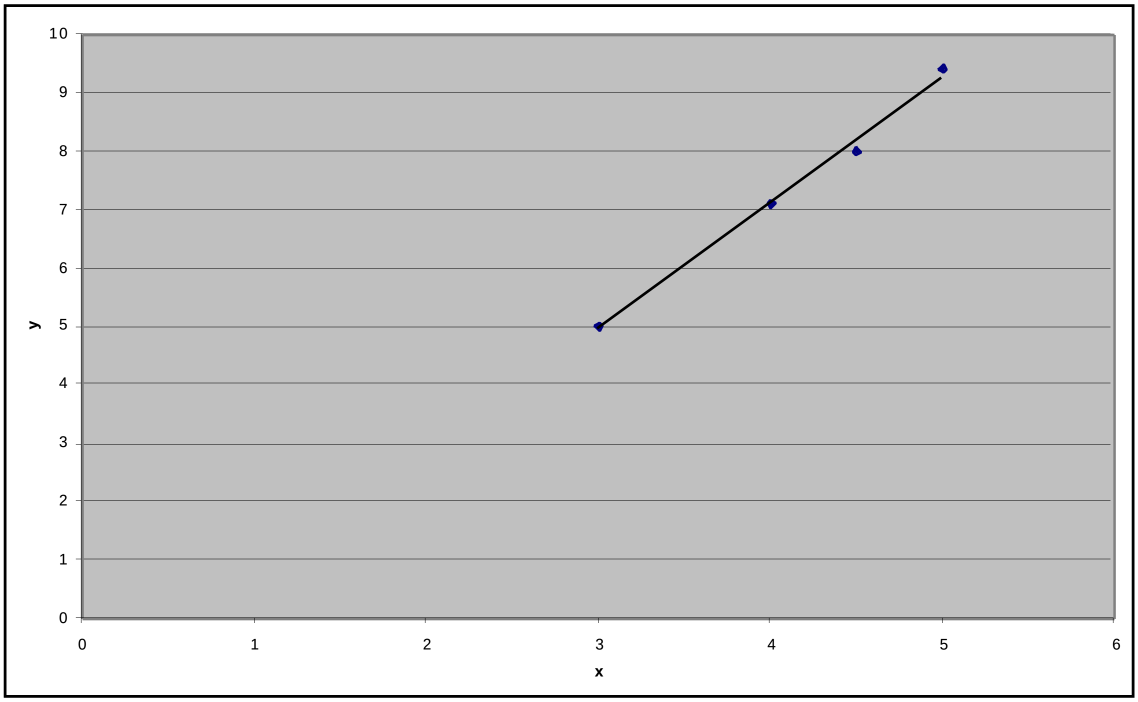
=INTERCEPT(C2:C5,B2:B5)

Note that in both cases, the y-values are listed before the x-values.

Step 3. Generate a graph containing the data as individual points and the regression line (without markers on the line).
The data are in B2:B5, and C2:C5. Highlight these. Try to insert a new chart. The kind of chart is "Scatter" since we wish to plot the first values versus the second values. Select the option that draws no lines. Successive clicks permit one to give names to the axes. Call them the x and y axes. Select not to give a legend. The chart now may look like



Now from the Chart menu select "Add Trendline" and choose the linear option. The graph now looks like:



Note that it shows the data points specifically and also has a trend line. Note that the trend line is quite good but does not in fact pass through each data point.

Remarks:

Sometimes we want only the fit to some of the data. If so, then we only use the relevant data in the lists.

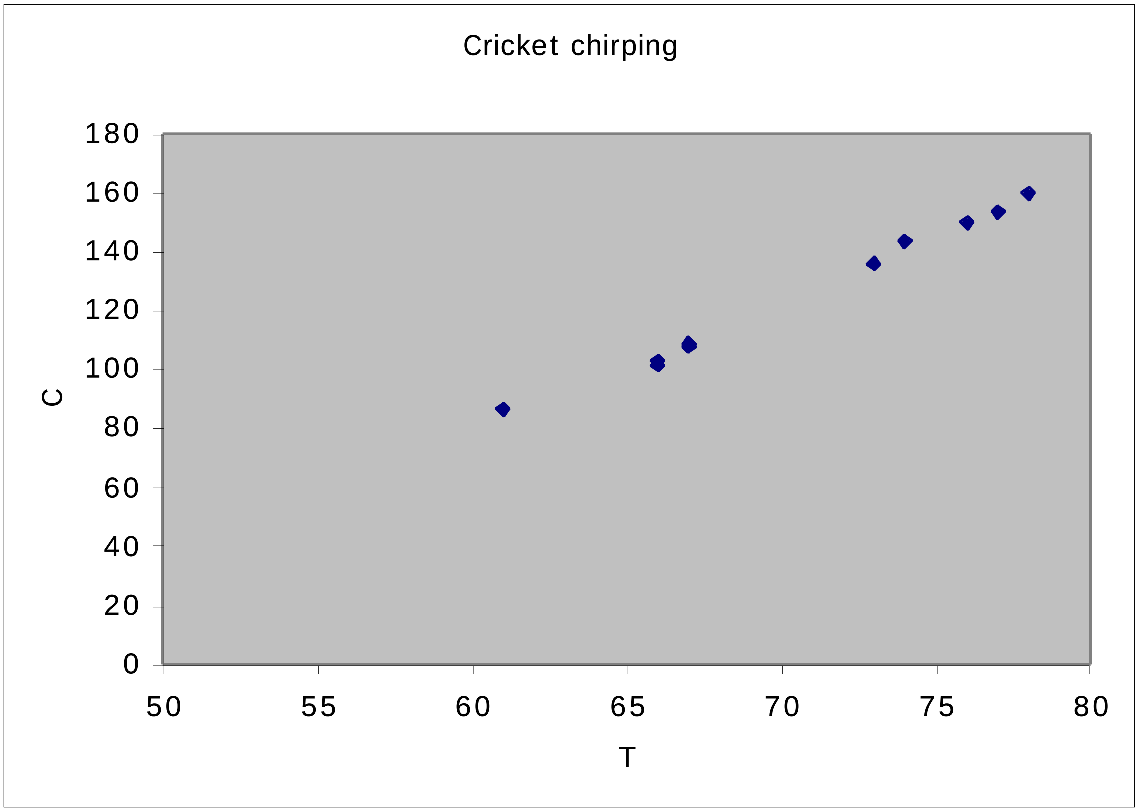
Sometimes the line of best fit doesn't fit the data very well. If so, we should say so.

A biological example: Cricket chirps

Some students of Professor Jim Cornette listened to crickets chirping on several nights during August and September 1997. They counted the number of chirps in a minute and also recorded the ambient (air) temperature for the night. The data were collected between 9:30 and 10:00 pm each night. The data they recorded and a scatterplot for the data are shown as follows:

T = Temperature in ˚F
C = rate of chirping in chirps per minute

T	C
67	109
73	136
78	160
61	87
66	103
66	102
67	108
77	154
74	144
76	150



It is easy to see that the points actually have a linear trend, although no single line passes through all the data points. Since there are 10 data points and no single line through them the methods used before we studied regression lines are inappropriate and don't work. We now try to solve similar problems related to these data, using all the data:
(a) Compute the equation of the regression line.
(b) If the ambient temperature is 70 ˚F, what rate of chirping do you predict?
(c) If on some evening the rate of chirping is 140 chirps / minute, what do you predict is the ambient temperature?
(d) Sketch the regression line together with the data. Comment on the fit.

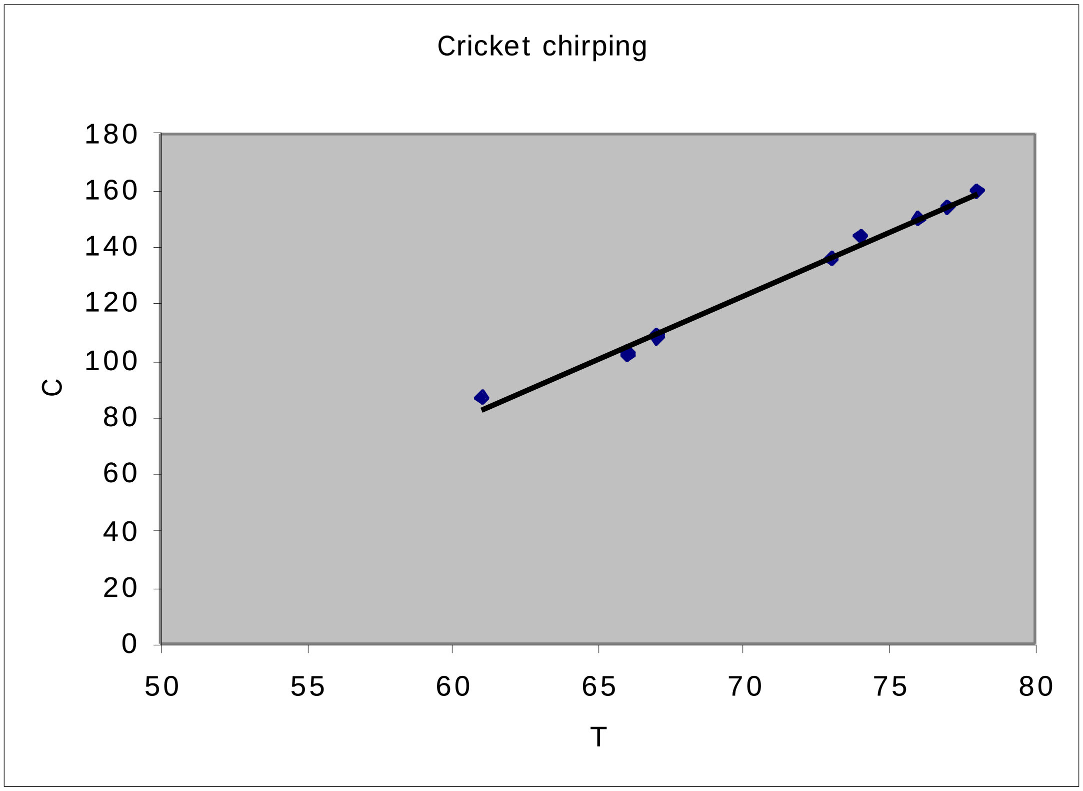
Solution. (a) Using either calculator or computer, we enter the data and find the regression line. We remember to use T in place of x since it is the variable on the horizontal axis, while we use C in place of y since it is the variable on the vertical axis. We find that the slope of the regression line is 4.501 and the intercept is -192. Hence the equation of the regression line is

$$C = 4.501\ T - 192 \text{ chirps/min}$$

(b) If T = 70 ˚F, then C = 4.501(70) - 192 = 123 chirps/minute.
(c) If C = 140, then we solve the equation

$$140 = 4.501\ T - 192$$
$$4.501\ T = 140 + 192$$
$$4.501\ T = 332$$
$$T = 332/4.501 = 73.8\ ^\circ\text{F}.$$

(d) Using the computer and Excel we find the graph with regression line is



The fit appears by eye to be quite good. The line follows the points very reasonably.