Using Maple to Perform Least Squares Fit (or Regression)

CURM Mathematical Background, Fall 2014

Linear Regression

Look at an example of height vs. age. The data is copied from *An Introduction to the Mathematics of Biology, with Computer Algebra Models*

by Yeargers, Shonkwiler, & Herod.

$restart$

$ht := [75, 92, 108, 121, 130, 142, 155]$

$$[75, 92, 108, 121, 130, 142, 155] \tag{1.1}$$

$age := [1, 3, 5, 7, 9, 11, 13]$

$$[1, 3, 5, 7, 9, 11, 13] \tag{1.2}$$

We can actually find the linear least squares fit using the formulas determined in class:

$sumy := sum(ht[n], n = 1 ..7)$

$$823 \tag{1.3}$$

$sumx := sum(age[n],\ n = 1 ..7)$

$$49 \tag{1.4}$$

$sumx2 := sum\left(age[n]^2, n = 1 ..7\right)$

$$455 \tag{1.5}$$

$sumxy := sum(age[n]\cdot\ ht[n],\ n = 1 ..7)$

$$6485 \tag{1.6}$$

$$m := evalf\left(\frac{(7\cdot sumxy - sumx\cdot sumy)}{7\cdot\ sumx2 - sumx^2}\right)$$

$$6.464285714 \tag{1.7}$$

$$b := evalf\left(\frac{(sumx2\cdot sumy - sumx\cdot sumxy)}{7\cdot\ sumx2 - sumx^2}\right)$$

$$72.32142857 \tag{1.8}$$

Alternately, we can let Maple find the least squares fit for us.

```
m :='m':  b :='b':
```

Set up a sequence of points for plotting purposes:

```
pts :=[seq([age[i], ht[i]], i = 1 ..7 )];
                [[1, 75], [3, 92], [5, 108], [7, 121], [9, 130], [11, 142], [13, 155]]          (1.9)
```

Include the plots and stats libraries.

```
with(plots):  with(stats):
```

Set up the graph of the points, and then determine the least squares fit. This is done using the fit function, which has as a category leastsquare, which we can further define as the line. The arguments of the fit function are the age and ht sequences.

```
Data := pointplot(pts, symbol = cross) :
fit[leastsquare[[x, y], y = m·x + b]]([age, ht])
```

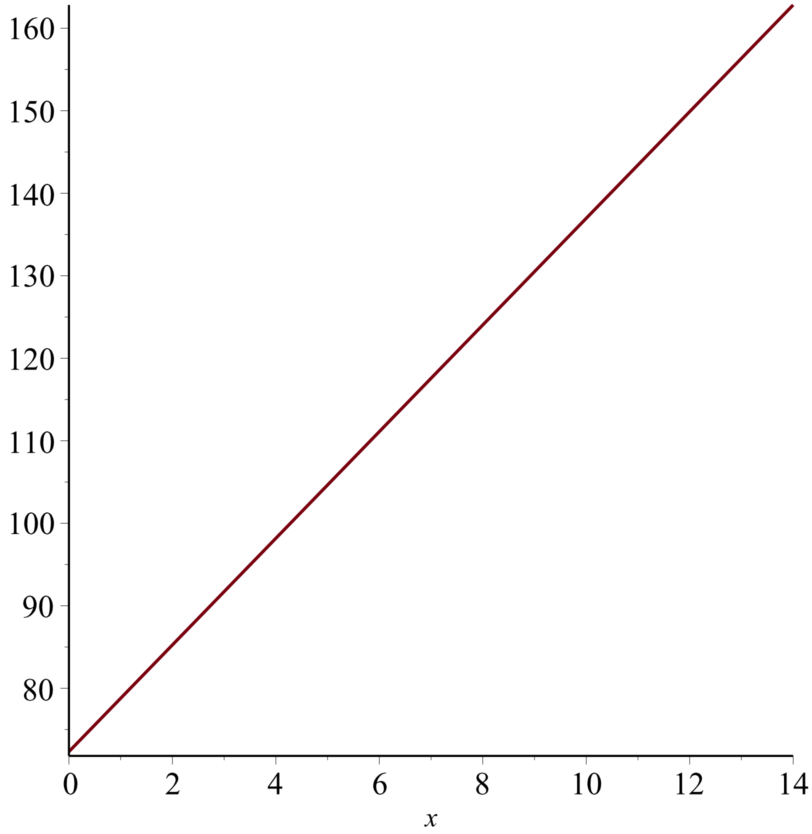
$$y = \frac{181}{28}\, x + \frac{2025}{28} \qquad \textbf{(1.10)}$$

We can recapture the values of the slope, m, and the y-intercept, b, as follows:

```
m := evalf(op(1, op(1, rhs(%))))
                                   6.464285714                                  (1.11)
b := evalf(op(2, rhs(%%)))
                                   72.32142857                                  (1.12)
```

Finally, store the plot of the line thus determined into another variable, Fit, and display the two plots together.

```
Fit := plot(m · x + b, x = 0 ..14 ):
display({Fit, Data});
```

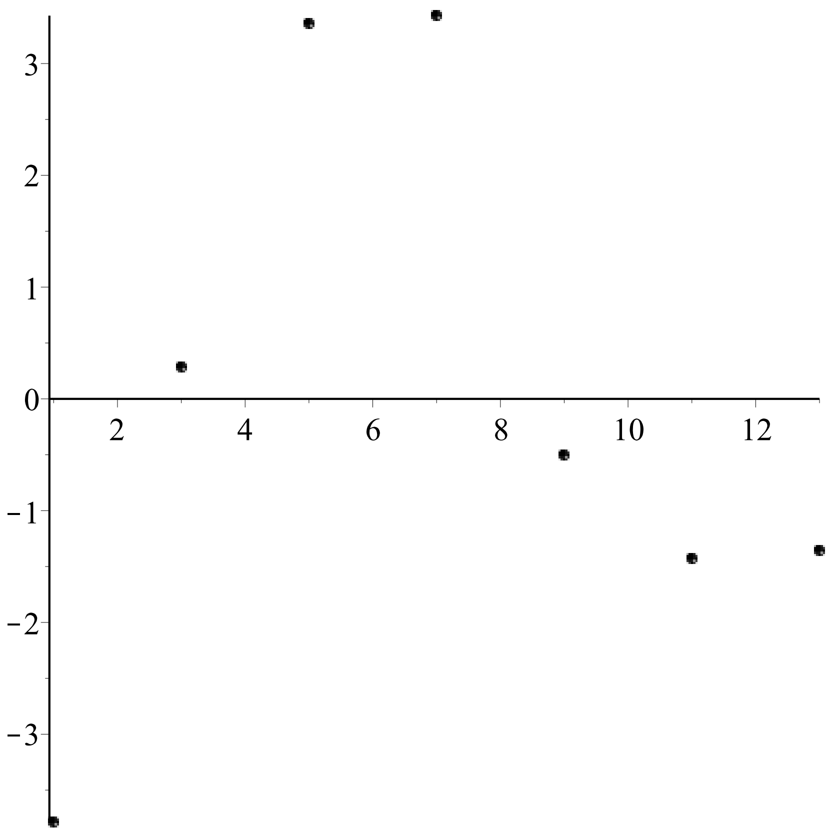


It appears that our model is a close fit. Let's look at the residuals, and see if our model appears to be acceptable.

$resid := seq(ht[i] - (m \cdot age[i] + b), i = 1 ..7)$

```
-3.78571428, 0.28571429, 3.35714286, 3.42857143, -0.50000000, -1.42857142,          (1.13)
    -1.35714285
```

$pointplot(\{seq([age[i], resid[i]], i = 1 ..7)\}, symbol = solidcircle)$



Since the residual does decrease in absolute value, and is nowhere very large, it apperas that the model is acceptable.

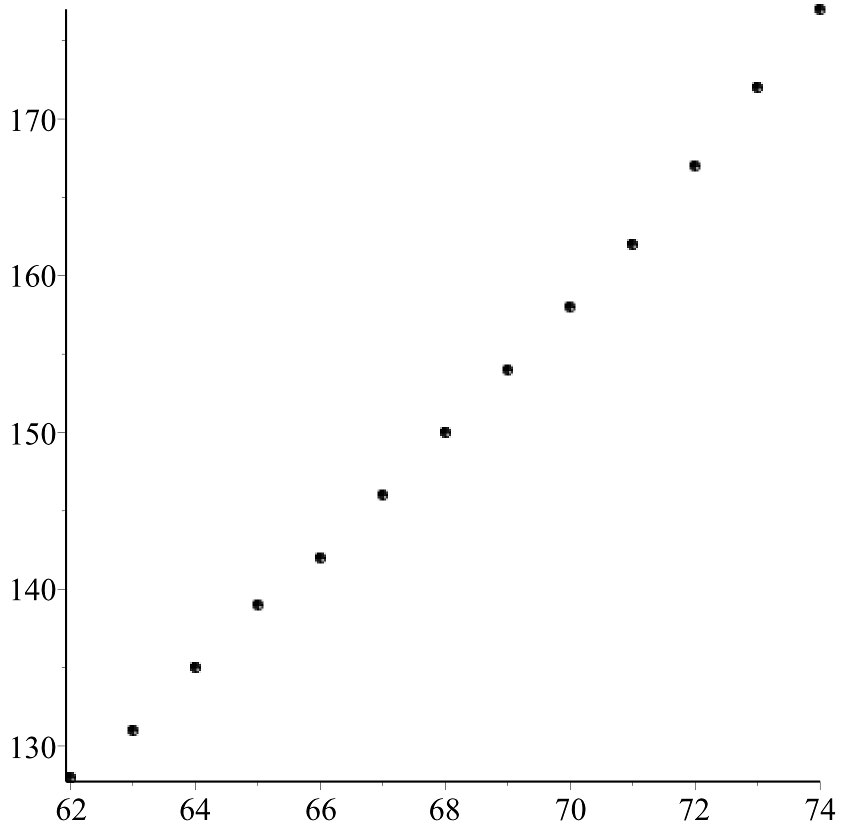
Fitting a Power Curve

Look at an example of ideal weight vs. height for medium built males. The data is copied from *An Introduction to the Mathematics of Biology, with Computer Algebra Models*

by Yeargers, Shonkwiler, & Herod.

```
restart
ht ≔ [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74]
                    [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74]                (2.1)
wt ≔ [128, 131, 135, 139, 142, 146, 150, 154, 158, 162, 167, 172, 177];
          [128, 131, 135, 139, 142, 146, 150, 154, 158, 162, 167, 172, 177]             (2.2)
```

```
with(plots) :
pointplot( {seq( [ht[i], wt[i]], i = 1 ..13 ) }, symbol = solidcircle)
```



The bulk of the data appears linear, so let's try a linear least squares fit.

```
with(stats) :
fit[leastsquare[ [x, y], y = m·x + b]]( [ht, wt])
```

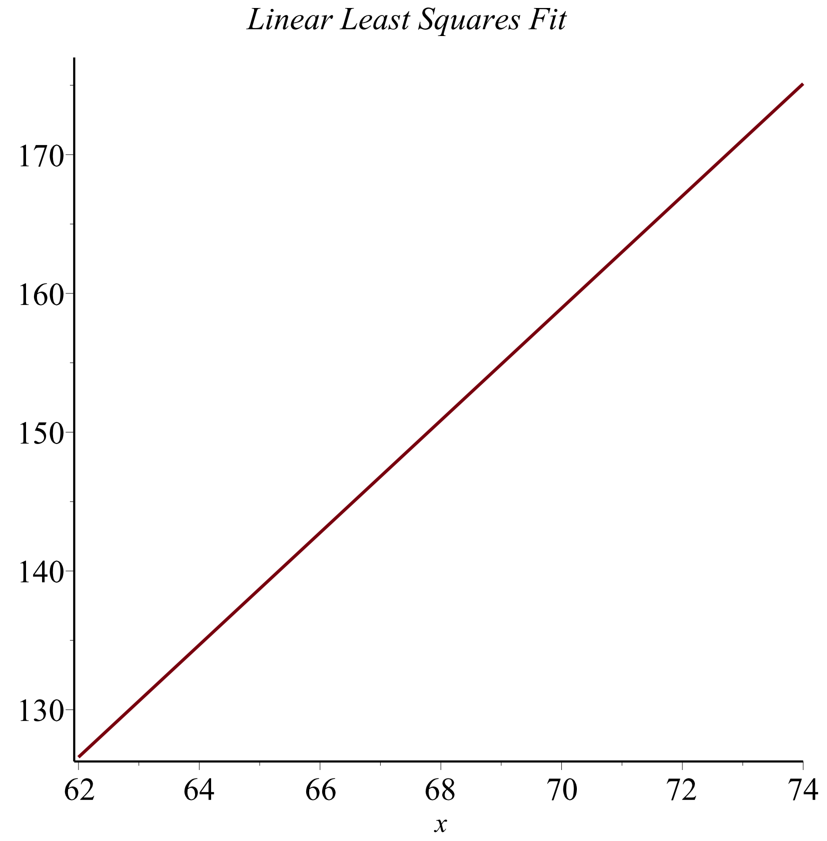
$$y = \frac{368}{91}\,x - \frac{869}{7} \tag{2.3}$$

```
m := op(1, op(1, rhs(%) ) );
b := op(2, rhs(%%) );
```

$$\frac{368}{91}$$

$$-\frac{869}{7} \qquad (2.4)$$

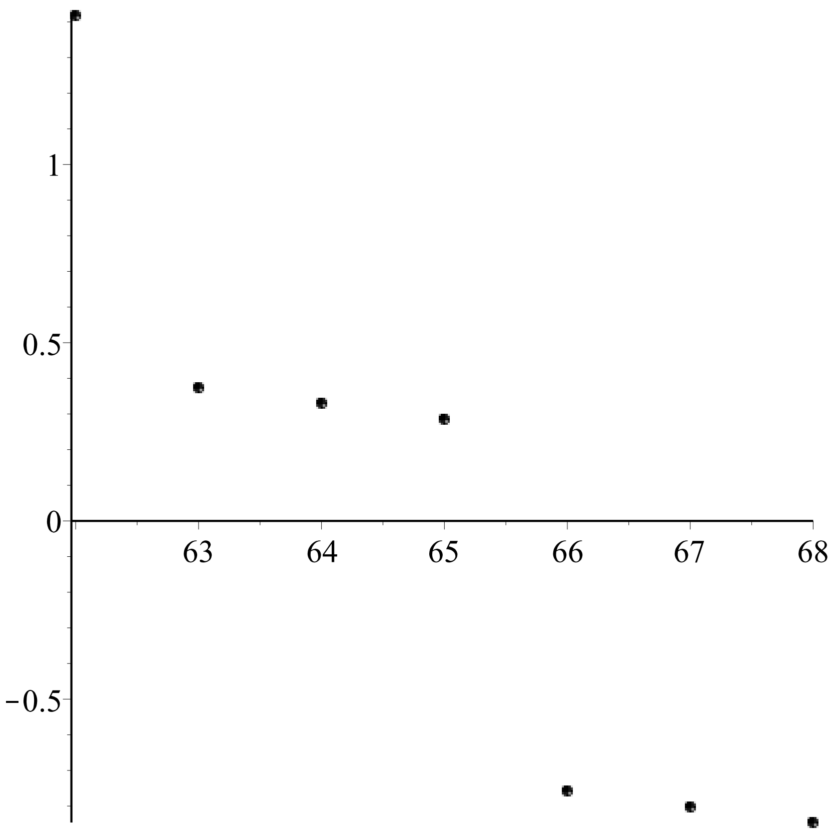
```
Data := pointplot( {seq( [ht[i], wt[i]], i = 1 ..13 ) }, symbol = cross ) :
Fit := plot(m·x + b, x = 62 ..74 ) :
display( {Data, Fit}, title = `Linear Least Squares Fit`)
```



```
resid := seq(wt[i] − (m·ht[i] + b ), i = 1 ..13 )
```

$$\frac{129}{91}, \frac{34}{91}, \frac{30}{91}, \frac{2}{7}, -\frac{69}{91}, -\frac{73}{91}, -\frac{11}{13}, -\frac{81}{91}, -\frac{85}{91}, -\frac{89}{91}, -\frac{2}{91}, \frac{85}{91}, \frac{172}{91} \qquad (2.5)$$

```
pointplot( {seq( [ht[i], resid[i]], i = 1 ..7 ) }, symbol = solidcircle)
```



The residual appears to be increasing towards the end of the height range. This tells us that the model needs to be refined. Since we know that in many geometric solids, the volume changes with the cube of the height, we will try to find a cubic fit for the data.

$restart;$
$with(stats) : with(plots) :$
$ht := [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74]$

$$[62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74] \tag{2.6}$$

$wt := [128, 131, 135, 139, 142, 146, 150, 154, 158, 162, 167, 172, 177]$

$$[128, 131, 135, 139, 142, 146, 150, 154, 158, 162, 167, 172, 177] \tag{2.7}$$

$fit[leastsquare[[x, y], y = a\cdot x^3 + b\cdot x^2 + c\cdot x + d]]([ht, wt])$

$$y = \frac{19}{3432}\,x^3 - \frac{163}{154}\,x^2 + \frac{1707059}{24024}\,x - \frac{3060027}{2002} \tag{2.8}$$

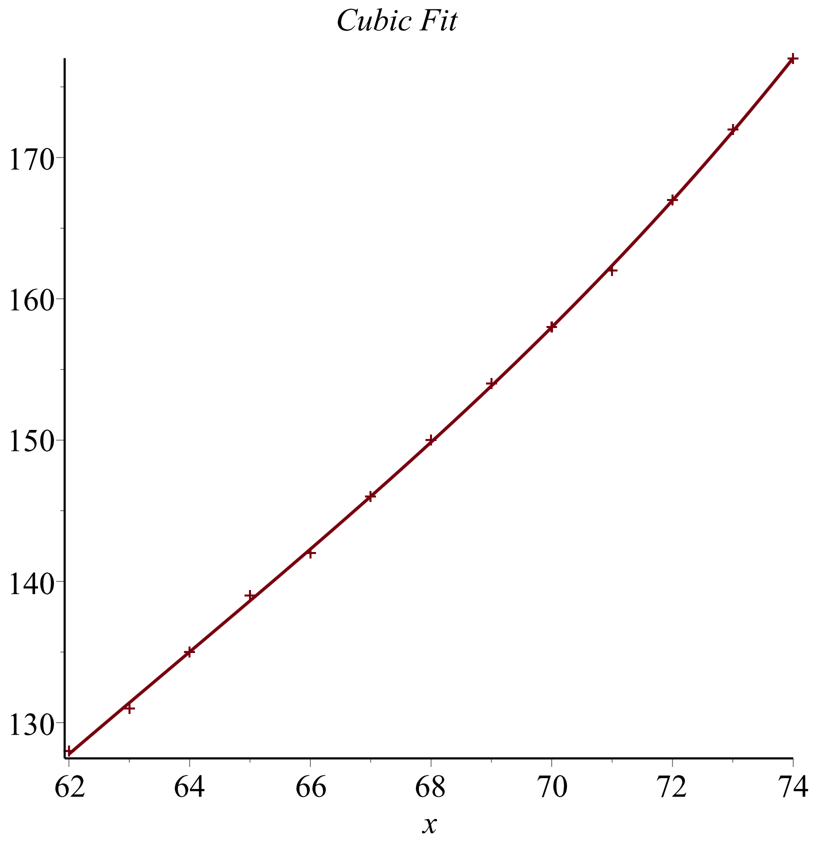
$y := unapply(rhs(\%), x)$

$$x \rightarrow \frac{19}{3432}\,x^3 - \frac{163}{154}\,x^2 + \frac{1707059}{24024}\,x - \frac{3060027}{2002} \tag{2.9}$$

$pts := [seq([ht[i], wt[i]], i = 1..13)]$

$$[[62, 128], [63, 131], [64, 135], [65, 139], [66, 142], [67, 146], [68, 150], [69, 154], [70, 158], [71, 162], [72, 167], [73, 172], [74, 177]] \tag{2.10}$$

```
Data := plot(pts, style = point, symbol = cross) :
Fit := plot(y(x), x = 62 ..74) :
display( {Data, Fit}, title = `Cubic Fit`)
```

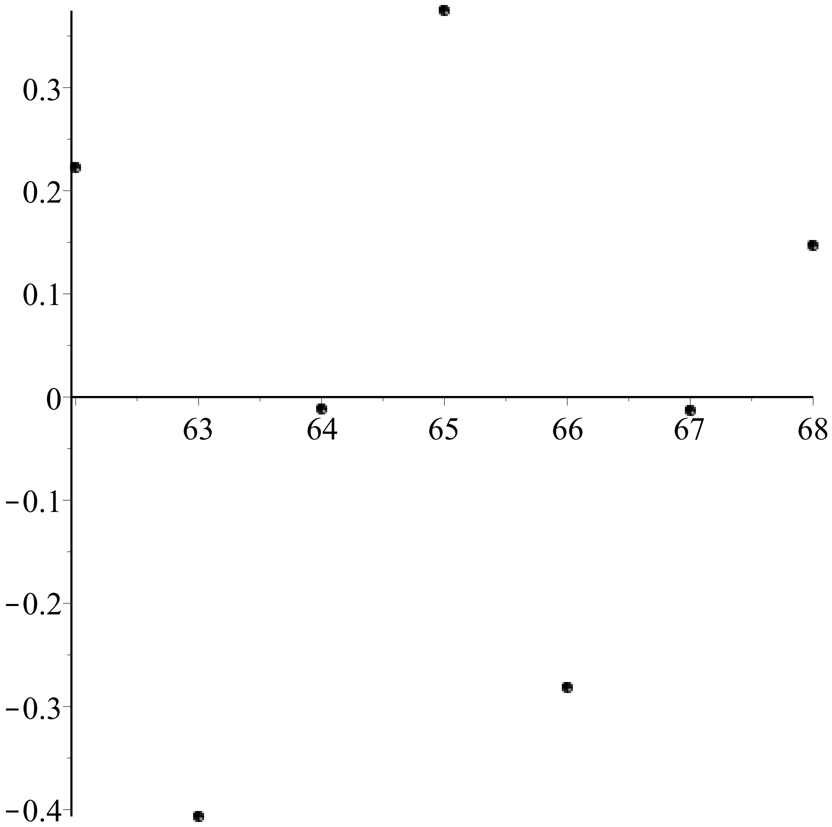


This appears much better. Let's look at the residual.

```
resid := seq(wt[i] − y(ht[i]), i = 1 ..13)
```

$$\frac{81}{364}, -\frac{37}{91}, -\frac{23}{2002}, \frac{375}{1001}, -\frac{161}{572}, -\frac{1}{77}, \frac{21}{143}, \frac{15}{91}, \frac{31}{4004}, -\frac{358}{1001}, \frac{71}{2002}, \frac{2}{13}, -\frac{1}{28} \tag{2.11}$$

```
pointplot( {seq( [ht[i], resid[i]], i = 1 ..7 ) }, symbol = solidcircle)
```



The residual is much better now, and appears to be decreasing as the height incrases.

Can we do better? Suppose we try to find a fit of the form $wt = A \cdot (ht - 60)^n$?

```
restart;
with(stats) : with(plots) :
ht := [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74]
                [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74]              (2.12)

htm60 := [seq(ht[i] - 60, i = 1 ..13)]
                    [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]                  (2.13)
wt := [128, 131, 135, 139, 142, 146, 150, 154, 158, 162, 167, 172, 177]
          [128, 131, 135, 139, 142, 146, 150, 154, 158, 162, 167, 172, 177]       (2.14)
LnWt := map(ln, wt·1.0)
[4.852030264, 4.875197323, 4.905274778, 4.934473933, 4.955827058, 4.983606622,    (2.15)
    5.010635294, 5.036952602, 5.062595033, 5.087596335, 5.117993812, 5.147494477,
    5.176149733]
```

```
LnHt ≔ map(ln, htm60·1.0)
[0.6931471806, 1.098612289, 1.386294361, 1.609437912, 1.791759469, 1.945910149,     (2.16)
    2.079441542, 2.197224577, 2.302585093, 2.397895273, 2.484906650, 2.564949357,
    2.639057330]
```

```
fit[leastsquare[[x, y], y = m·x + b]]([LnHt, LnWt])
                    y = 0.1681952344 x + 4.685291069                                  (2.17)
```

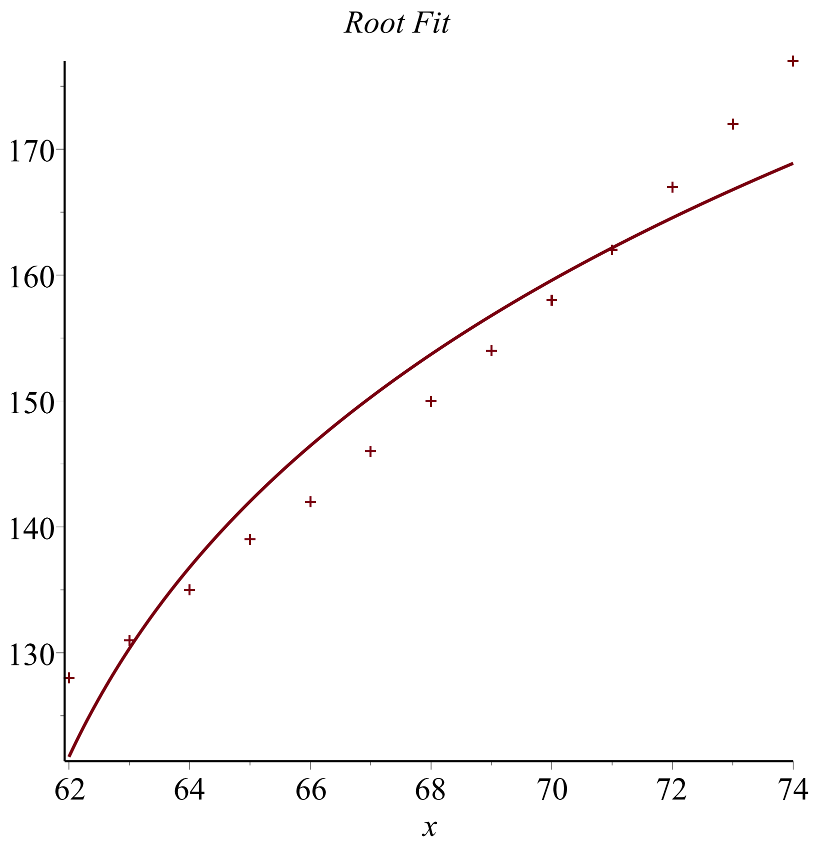
```
n ≔ evalf(op(1, op(1, rhs(%))));
LnA ≔ evalf(op(2, rhs(%%)));
 A ≔ exp(LnA)
                              0.1681952344
                              4.685291069
                              108.3418027                                             (2.18)
```

Problem: *n* should be an integer. But, let's allow non-integer powers, since the mathematics behind this approach does not require an integer (we just took the natural log of both sides to get a linear equation).

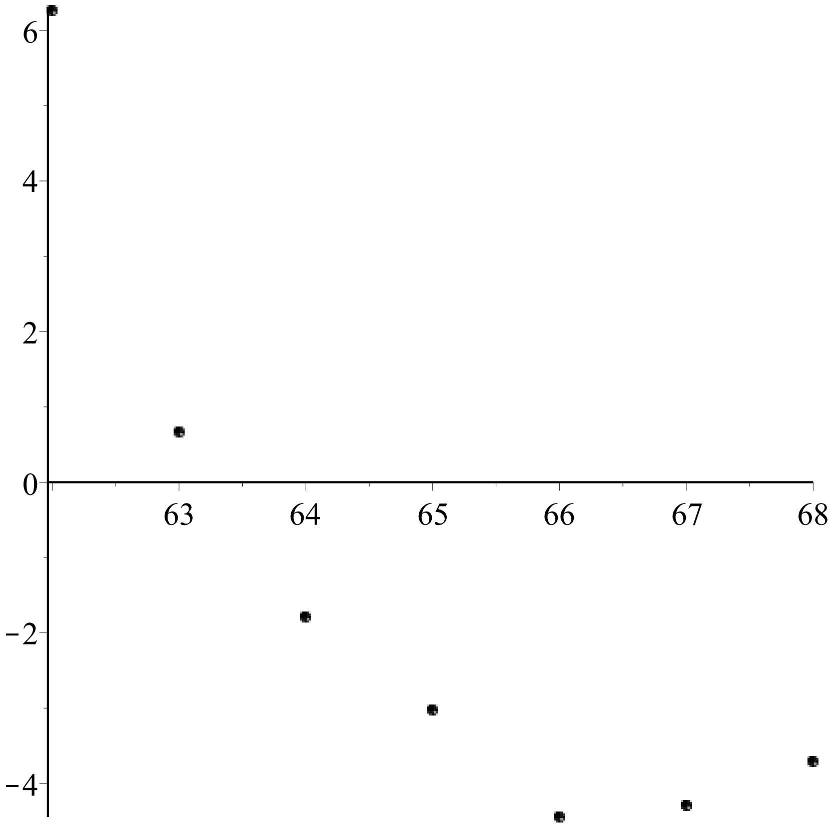
```
y ≔ x→A·(x − 60)^n
                              x→A (x − 60)^n                                          (2.19)
```

```
pts ≔ [seq([ht[i], wt[i]], i = 1 ..13)]
[[62, 128], [63, 131], [64, 135], [65, 139], [66, 142], [67, 146], [68, 150], [69, 154],   (2.20)
    [70, 158], [71, 162], [72, 167], [73, 172], [74, 177]]
```

```
Data ≔ plot(pts, style = point, symbol = cross) :
Fit ≔ plot(y(x), x = 62 ..74) :
display({Data, Fit}, title = `Root Fit`)
```



```
resid := seq(wt[i] − y(ht[i]), i = 1 ..13)
6.2615219, 0.6696448, -1.7916786, -3.0232614, -4.4459580, -4.2925733, -3.7062367,   (2.21)
    -2.7816029, -1.5847148, -0.1635878, 2.4457133, 5.2153720, 8.1234566
pointplot( {seq( [ht[i], resid[i]], i = 1 ..7 ) }, symbol = solidcircle)
```



As mentioned above, ideally n is an integer, or at least a rational number. The rational number closest to n is $\frac{1}{6}$. So, let's re-do the above using that value of n.Error, missing operator or `;`

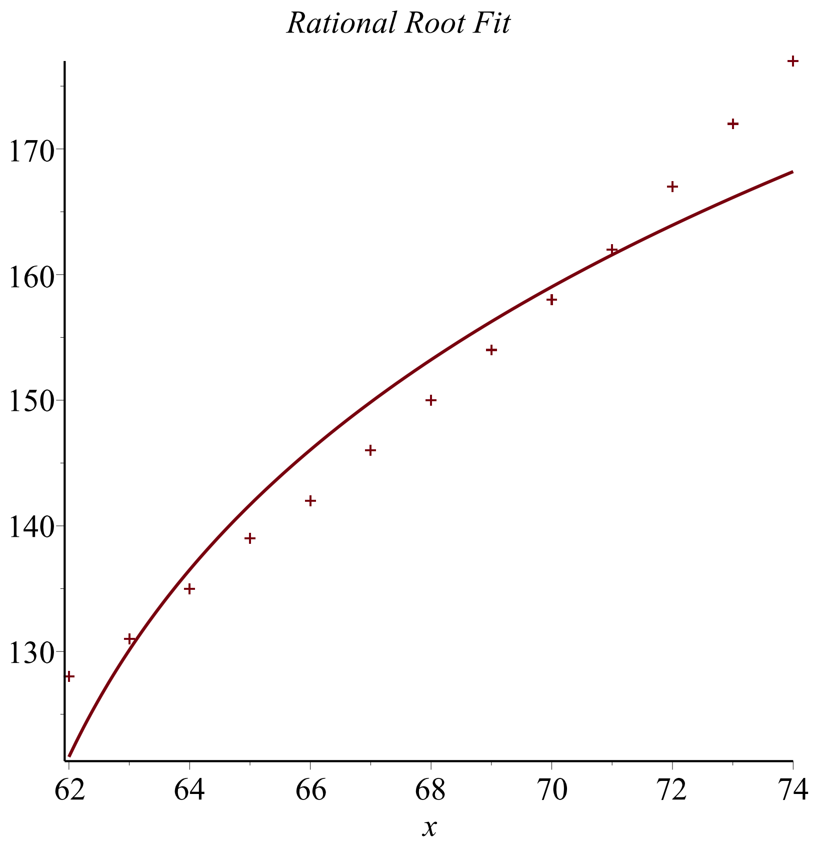
$n := \frac{1}{6}$

$$\frac{1}{6} \tag{2.22}$$

$y := x \rightarrow A \cdot (x - 60)^n$

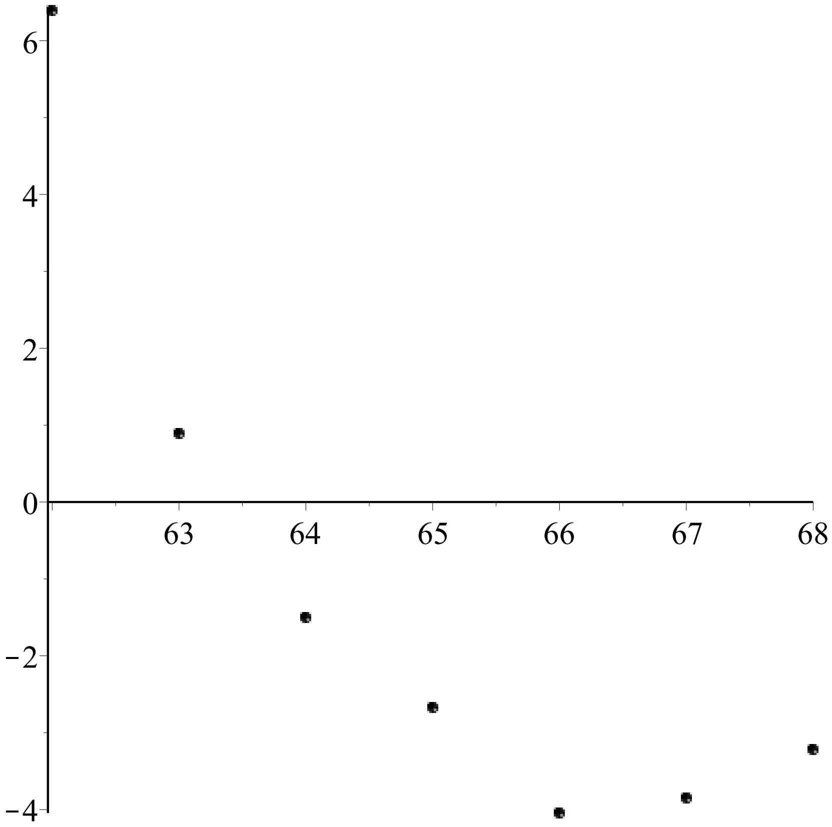
$$x \rightarrow A\,(x - 60)^n \tag{2.23}$$

Fit := *plot*(*y*(*x*), *x* = 62 ..74) :
display({*Data*, *Fit*}, *title* = \`*Rational Root Fit*\`)



Is this result any better, in terms of the residual? Let's check.

```
resid := evalf(seq(wt[i] − y(ht[i]), i = 1 ..13))
6.3904383, 0.8883254, -1.5021178, -2.6742944, -4.0454169, -3.8461989, -3.2184467,   (2.24)
    -2.2559184, -1.0240187, 0.4297098, 3.0695623, 5.8680040, 8.8033286
pointplot({seq([ht[i], resid[i]], i = 1 ..7)}, symbol = solidcircle)
```



The results are virtually identical (not surprisingly).

The residual is very high for some of the values. Why do you think this is? What do you think would happen if we looked at the residual of the results using the logarithm of the values?

Multiple Regression

Let us look at Example (3) from the class notes. Again, the data is copied from *An Introduction to the Mathematics of Biology, with Computer Algebra Models*

by Yeargers, Shonkwiler, & Herod.

First, set up the data sequences and include the stats library. BMI is defined as the weight (in kg) divided by the square of the height (in m). The data for the height and weight below is given in inches and pounds, respectively. Therefore, in order to determine the BMI, we need to convert the height and weight to their metric equivalents and do the division.

$restart;$

$ht := [63, 65, 61.7, 65.2, 66.2, 65.2, 70.0, 63.9, 63.2, 68.7, 68, 66]$

$$[63, 65, 61.7, 65.2, 66.2, 65.2, 70.0, 63.9, 63.2, 68.7, 68, 66] \tag{3.1}$$

$wt := [109.3, 115.6, 112.4, 129.6, 116.7, 114.0, 152.2, 115.6, 121.3, 167.7, 160.9, 149.9]$

$$[109.3, 115.6, 112.4, 129.6, 116.7, 114.0, 152.2, 115.6, 121.3, 167.7, 160.9, 149.9] \tag{3.2}$$

$convert\left(\left[seq\left(\frac{wt[i]\cdot lbs}{\left(\frac{ht[i]}{12}\cdot feet\right)^2}, i=1..12\right)\right], metric\right)$

$$\left[\frac{19.36142733\,kg}{m^2}, \frac{19.23664933\,kg}{m^2}, \frac{20.75841980\,kg}{m^2}, \frac{21.43424139\,kg}{m^2}, \frac{18.72204068\,kg}{m^2}, \frac{18.85419381\,kg}{m^2}, \frac{21.83820205\,kg}{m^2}, \frac{19.90464448\,kg}{m^2}, \frac{21.35133092\,kg}{m^2}, \frac{24.98146457\,kg}{m^2}, \frac{24.46451024\,kg}{m^2}, \frac{24.19424472\,kg}{m^2}\right] \tag{3.3}$$

$BMI := [19.36, 19.24, 20.76, 21.43, 18.72, 18.85, 21.84, 19.90, 21.35, 24.98, 24.46, 24.19]$

$$[19.36, 19.24, 20.76, 21.43, 18.72, 18.85, 21.84, 19.90, 21.35, 24.98, 24.46, 24.19] \tag{3.4}$$

$SF := [86.0, 94.5, 105.3, 91.5, 75.2, 93.2, 156.0, 75.1, 119.8, 169.3, 170.0, 148.2]$

$$[86.0, 94.5, 105.3, 91.5, 75.2, 93.2, 156.0, 75.1, 119.8, 169.3, 170.0, 148.2] \tag{3.5}$$

$PBF := [19.3, 22.2, 24.3, 17.1, 19.6, 23.9, 29.5, 24.1, 26.2, 33.7, 36.2, 31.0]$

$$[19.3, 22.2, 24.3, 17.1, 19.6, 23.9, 29.5, 24.1, 26.2, 33.7, 36.2, 31.0] \tag{3.6}$$

$with(stats):$

We will now compute the least squares fit. Notice that here we don't specify a linear least squres fit. This is because Maple computes a linear least squares fit by default, with the last variable listed (here, c) being the dependent variable.

$fit[leastsquare[[bdymass, sfld, bf]]]([BMI, SF, PBF])$

$$bf = 0.006561278951\,bdymass + 0.1506644621\,sfld + 8.074305581 \tag{3.7}$$

$bdft := unapply(rhs(\%), (bdymass, sfld))$

$$(bdymass, sfld) \rightarrow 0.006561278951\,bdymass + 0.1506644621\,sfld + 8.074305581 \tag{3.8}$$

Is this a good fit? Let's check it out using sample data not used in the calculations. The subject is 64.5 inches tall, weighs 135 pounds, and has skin-fold that measures 159.9 millimeters. Her true body fat percentage is 30.8. The predicted value is

$$convert\left(\frac{135 \cdot lbs}{\left(\frac{64.5 \cdot ft}{12}\right)^2},\ metric\right)$$

$$\frac{22.81458885\ kg}{m^2} \tag{3.9}$$

$$bdft(22.815,\ 159.9)$$

$$32.31524865 \tag{3.10}$$