Some common relationships between data

Most common:

- linear $y = ax + b$
- exponential $y = a\exp(bx)$
- power law $y = ax^b$
- logarithmic $y = a\log(bx)$

Note that the power law includes

- inverse proportionality $y = a/x$
- proportionality to square root $y = a\sqrt{x}$
- simple quadratic dependence $y = ax^2$
- simple cubic dependence $y = ax^3$

Often we want to fit one of the above curves, that is, find a and b parameters.

Linear least squares method

The problem of fitting of $y = ax + b$ to data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is rather straightforward. We minimize the sum of squared residuals, that is,

$$S(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2.$$

From multivariable calculus, we know that the necessary condition for function $S(a, b)$ to reach minimum is

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0,$$

hence

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2(ax_i + b - y_i)x_i = 0$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2(ax_i + b - y_i) = 0$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0$$

$$a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} b - \sum_{i=1}^{n} y_i = 0$$

Treating the above as a system of two equations with two unknowns, and solving for a, b, one obtains

$$a = \frac{S_x S_y - nS_{xy}}{S_x^2 - nS_{xx}}, \quad b = \frac{S_x S_{xy} - S_y S_{xx}}{S_x^2 - nS_{xx}},$$

where

$$S_x = \sum_{i=1}^{n} x_i,\ S_y = \sum_{i=1}^{n} y_i,\ S_{xx} = \sum_{i=1}^{n} x_i^2,\ S_{xy} = \sum_{i=1}^{n} x_i y_i.$$

The values of a, b calculated this way define the line of the best fit. This is called *simple linear regression*.

What if we need to fit nonlinear function?

One possibility is to *linearize*, that is, to change coordinates to obtain linear dependence. Example of linearization:

$$y = a \exp(bx)$$

Take log of both sides

$$\log y = \log a + bx$$

Define $Y = \log y$, $A = \log a$, $B = b$, $X = x$, then we obtain linear function

$$Y = A + BX$$

Another example, for power law:

$$y = ax^b$$

Take log of both sides

$$\log y = \log a + b \log x$$

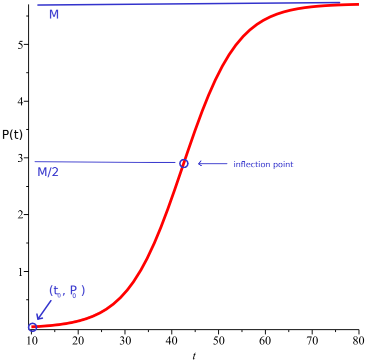
Define $Y = \log y$, $A = \log a$, $B = b$, $X = \log x$, then we obtain linear function

$$Y = A + BX$$

Note that here slope of the straight line B is the same as the exponent b in $y = ax^b$.

Logistic curve

Consider now a more complicated curve, for example, the so-called *logistic curve* with four parameters, t_0, P_0, M and r:



$$P(t) = \frac{MP_0}{P_0 + (M - P_0)e^{-rM(t-t_0)}},$$

History

Logistic curve has been discovered by Pierre François Velhulst (1804–1849) in 1825, as a solution of a differential equation describing limited growth (we will discuss it later).In 1920s, biologist Raymond Pearl (1879–1940) and biostatistician Lowell Jacob Reed (1886–1966) rediscovered this curve and promoted its applications.

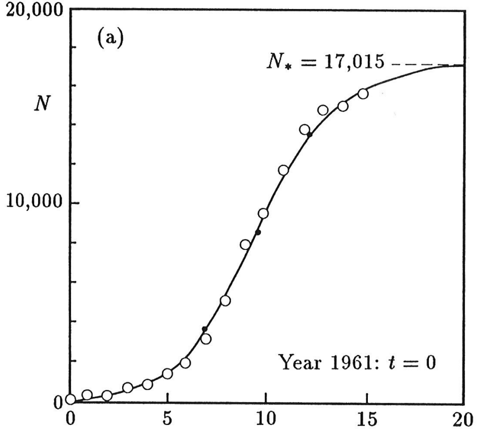
Verhulst

Pearl

Reed

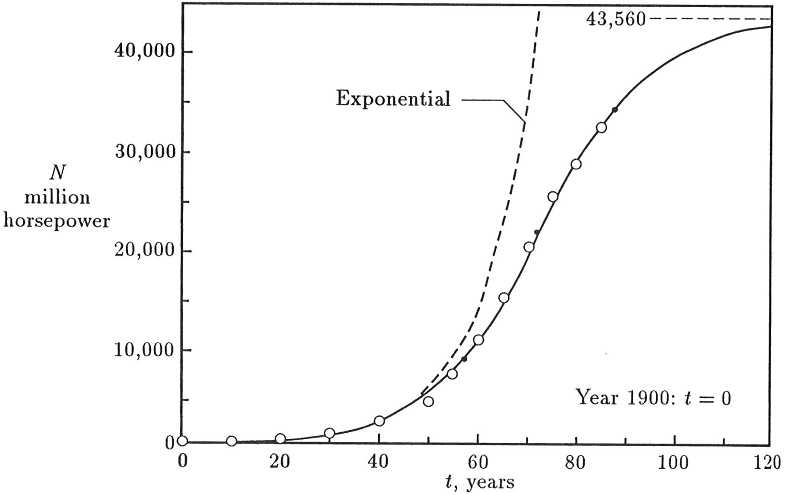
Examples of logistic curve

Growth and Diffusion Phenomena,Robert B. Banks, Springer-Verlag, Berlin 1994.



Number of ranchers adopting new pasture technology in Uruguay; Data of Jarvis (1981)

Examples of logistic growth - continued

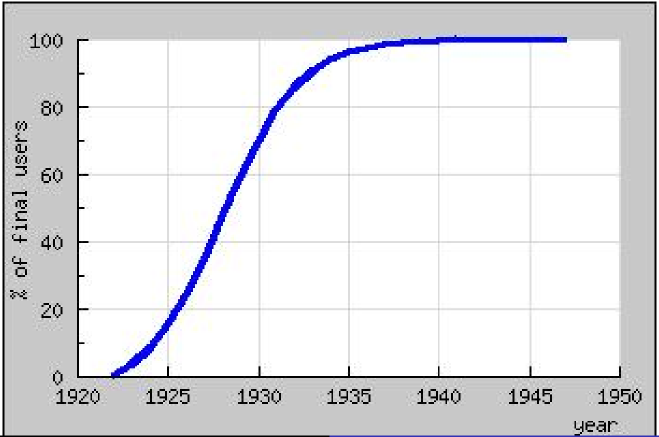


Total horsepower of all prime movers in the United States, 1900–1985

Examples of logistic curve - radio

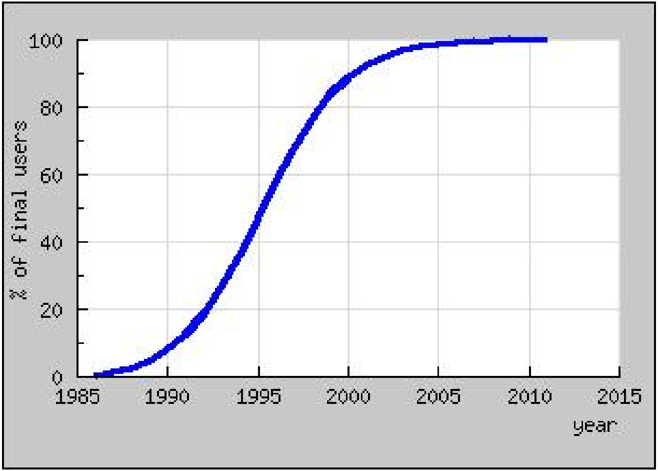
Source: “Bias and Systematic Change in the Parameter Estimates of Macro-Level Diffusion Models”, G. Lilien and C. Van den Bulte), Marketing Science, Vol. 16, No. 4, 1997, pp. 338-353. Data available at

`http://andorraweb.com/bass/index.php?show[examples]=1`

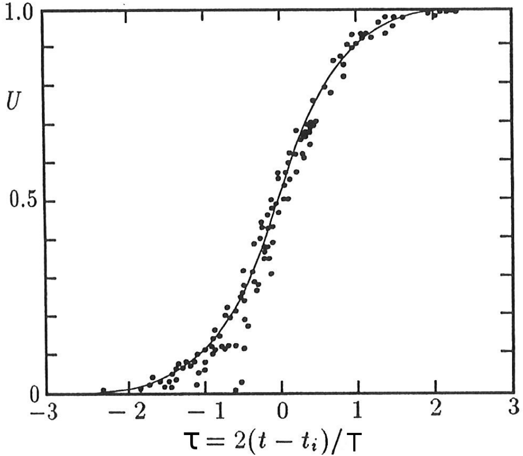


Examples of logistic curve - cell phones

Source: ibid.



Logistic curve - universality



Universal plot with normalized time for 17 cases of technology substitution. From Fisher and Pry (1971)

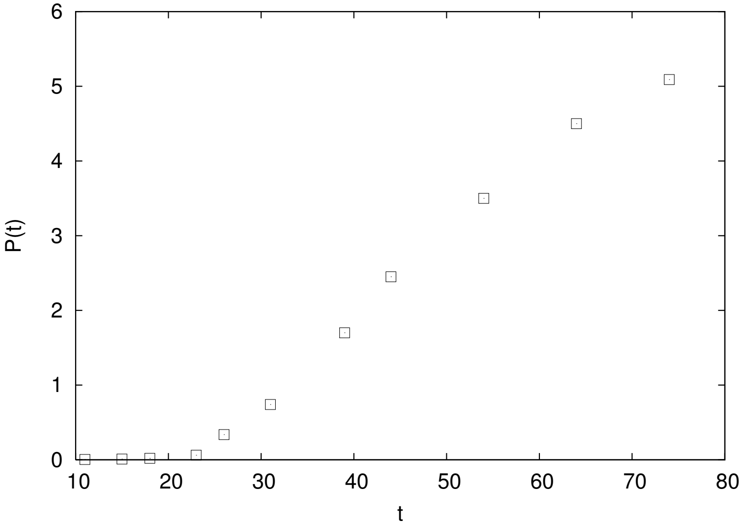
Algae growth data

The table below shows the growth of an algal sample taken from the Adriatic Sea. The biomass is measured as as the area of a microscopic slide covered by the sample (in mm^2).

Time (days)	Biomass (mm^2)
11	0.00476
15	0.0105
18	0.0207
23	0.0619
26	0.337
31	0.74
39	1.7
44	2.45
54	3.5
64	4.5
74	5.09

Can we fit logistic curve to it, by linearization?

Scatter plot of data



Fitting by linearization

We will first try to fit by “pedestrian” way. The first pair of parameters, t_0 and P_0, can be obtained directly form the data table, $t_0 = 11$ and $P_0 = 0.00476$. Similarly, it is reasonable to assume that M is slightly above the largest value of P_i in the data set. Therefore, we will take $M \approx 5.1$. The remaining parameter r will have to be estimated using linearization.

From the definition of logistic curve we obtain

$$e^{-rM(t-t_0)} = \frac{(M - P(t))P_0}{(M - P_0)P(t)},$$

hence

$$\ln \frac{(M - P(t))P_0}{(M - P_0)P(t)} = -rM(t - t_0),$$

Introducing a new variable

$$y(t) = \ln \frac{(M - P(t))P_0}{(M - P_0)P(t)}$$

we obtain

$$y(t) = -rMt + rMt_0,$$

or

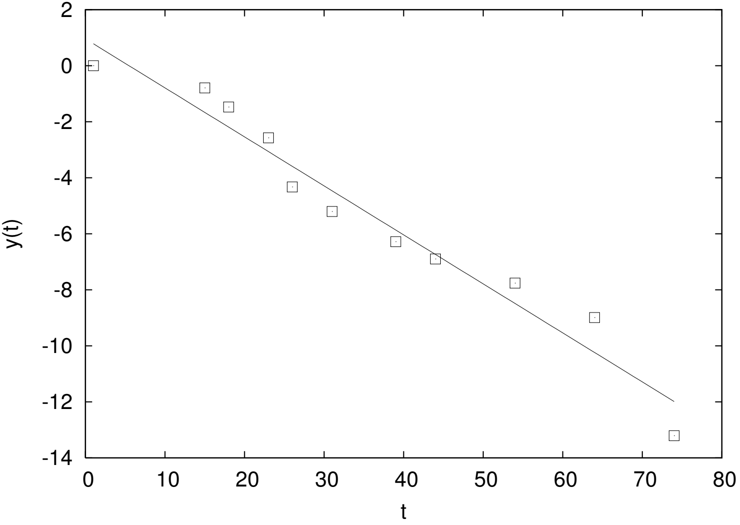
$$y(t) = At + B,$$

where $A = -rM$, $B = rMt_0$. This means that the plot of (t_i, y_i) should be linear, assuming that

$$y_i = \ln \frac{(M - P_i)P_0}{(M - P_0)P_i}.$$

The table below shows values of (t_i, y_i) obtained from original data.

t_i	y_i
1	0.0
15	-.7922547655
18	-1.473019346
23	-2.576550255
26	-4.327264311
31	-5.202237231
39	-6.282667201
44	-6.897342766
54	-7.758573721
64	-8.990717402
74	-13.20826240


y(t)
t
2
0
-2
-4
-6
-8
-10
-12
-14
0
10
20
30
40
50
60
70
80

The slope of this line is $A = -0.17494$. Since $A = -rM$, we obtain

$$r = \frac{-A}{M} = \frac{0.17494}{5.1} \approx 0.0343$$

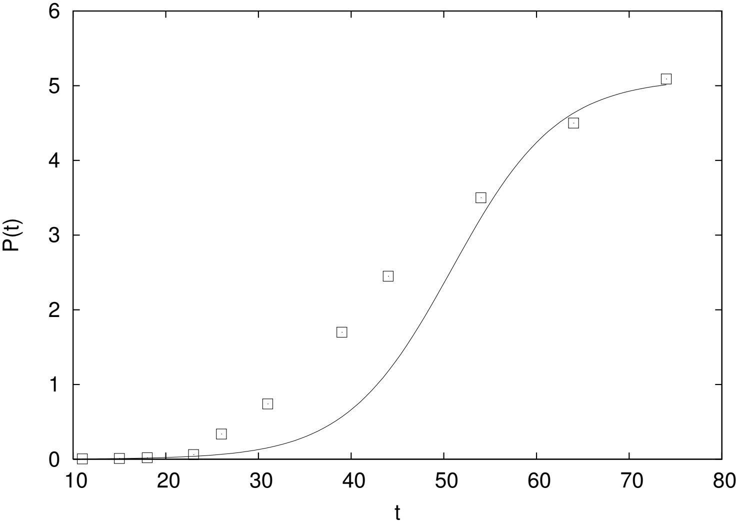
We thus obtained the following set of parameters

$$\begin{aligned} M &= 5.1 && (1) \\ r &= 0.0343 && (2) \\ P_0 &= 0.00476 && (3) \\ t_0 &= 11.0 && (4) \end{aligned}$$

This yields

$$P(t) = \frac{35.7}{7 + 51326.09e^{-0.17493t}}.$$

One can now use this equation to predict (extrapolate), for example, the biomass after 90 days, $P(90) = 5.0946$.



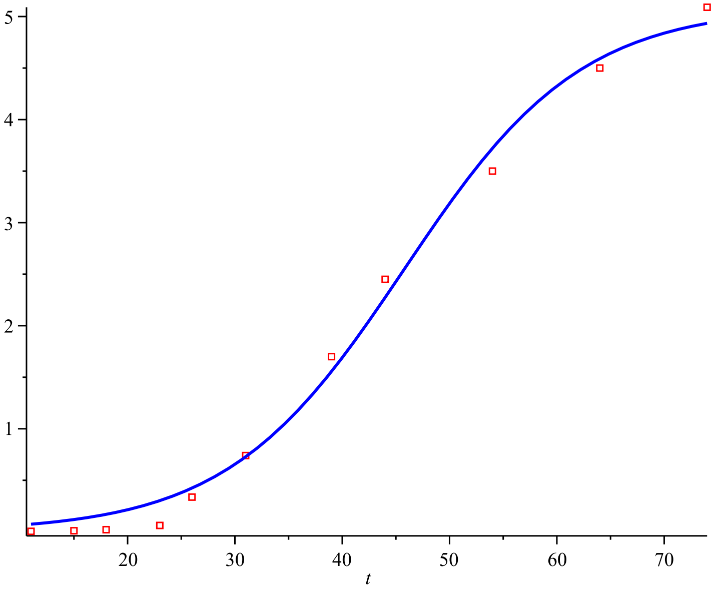
If we plot the fitted curve with original data, the fit does not seem to be too good.

Can we do better? Let us use Maple **Fit** command, using previously obtained values as initial “guesses”:

```
Fit(M*P0/(P0 +(M-P0)*exp(-r*M*(t-t0))), X, Y, t,
initialvalues=[M=5.1, r=0.0343,P0=0.00476,t0=11]);
```

This yields parameter values (rounded to 3 significant digits)

$$M = 5.0949,\ P_0 = 0.00648,\ r = 0.0238,\ t_0 = -9.175$$

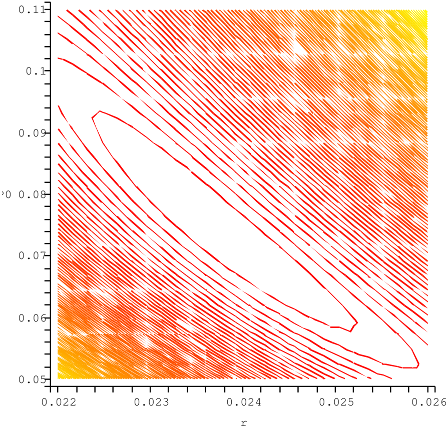

5
4
3
2
1
20
30
40
50
60
70
t

How does Maple do this? In order to understand this, let us consider simplified problem, where we try to fit only two parameters instead of four. Suppose we take our initial guess $M := 5.1$, $t0 := 11$.Construct the sum of squared residuals as follows:

```
 restart: with(plots):
M:=5.1;
t0:=11;
tdata:=Vector([11,15,18,23,26,31,39,44,54,64,74]);
Pdata:=Vector([0.00476,0.0105,0.0207,0.0619,0.337,0.74,
            1.7,2.45,3.5,4.5,5.09]);
P:=t->(M*P0)/(P0+(M-P0)*exp(-r*M*(t-t0)));
S:=add((P(tdata[i])-Pdata[i])^2,i=1..11);
```

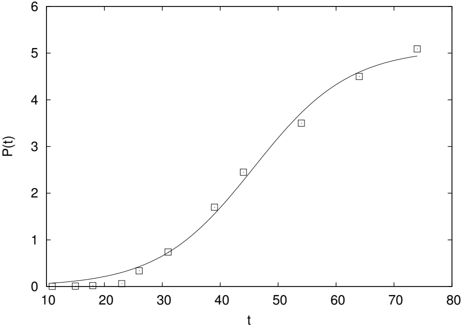
Contour plot of S , obtained using

```
contourplot(S,r=0.022..0.026,P0=0.05..0.11,contours=100);
```



Existence of minimum is clearly visible in the centre of the innermost curve.

From this plot, one can determine the location of the minimum as the centre of the smallest contour. This yields $r = 0.024$, $P_0 = 0.074$. Using these values together with $M = 5.1$ and $t_0 = 11$ we obtain reasonably good fit,



Maple performs similar procedure of minimization in 4-dimensional space, using clever algorithm. You must tell it, however, where to search for minimum!

Directional change

Function S can be used to find out how sensitive is our fit to small parameter changes.

$$S(t_0, P_0, M, r) = \sum_{i=1}^{11} (P(t_i) - P_i)^2.$$

Imagine that we keep all parameters constant except M, which increases.How rapidly will S change in response to changes in M? Of course, this rate of change of S is given by $\partial S/\partial M$. Let us not compute all four partial derivatives of S and evaluate them at $r = 0.024$, $P_0 = 0.074$, $M = 5.1$, $t_0 = 11$. The results are:

$$\begin{aligned}\frac{\partial S}{\partial M} &= -7.429770225, \ \frac{\partial S}{\partial P_0} = -89.53277367 \\ \frac{\partial S}{\partial r} &= -1349.803928 \ \frac{\partial S}{\partial t_0} = 0.6659877466\end{aligned}$$

Clearly, S is most sensitive to changes of r.