Data Fitting With Regression

Introduction

Procedures to minimize the error between given data points and the calculated values from an equation. The value that is to be minimized (the value of the objective function) is typically the sum of the errors squared.

Linear Regression

All functional forms of the independent variables must be linearly combined:

$$y = \sum_{i=1}^{N} a_i \cdot \mathrm{u}_i\left(x_1, x_2, \ldots, x_M\right) = \sum_{i=1}^{N} a_i u_i$$

With linear regression we would like to minimize the square error of K data points (where $K \geq N$). The objective function is:

$$F \equiv \sum_{k=1}^{K}\left(y_k - \sum_{i=1}^{N} a_i u_{i,k}\right)^2 .$$

The unconstrained minimum will occur at the extremum point – all partial derivatives with respect to the regression parameters (a_1, a_2, etc.) will be zero. The partial derivatives are:

$$\left(\frac{\partial F}{\partial a_j}\right) \equiv \sum_{k=1}^{K} 2\left(y_k - \sum_{i=1}^{N} a_i u_{i,k}\right)\left(-\sum_{i=1}^{N} \frac{\partial a_i}{\partial a_j} u_{i,k}\right)$$
$$= -2\sum_{k=1}^{K}\left(y_k - \sum_{i=1}^{N} a_i u_{i,k}\right)\left(u_{j,k}\right)$$

This means there is a system of N equations such that:

$$\sum_{k=1}^{K}\left(y_k - \sum_{i=1}^{N} a_i u_{i,k} \right)\left(u_{j,k} \right) = 0$$

$$\sum_{k=1}^{K}\left(y_k \cdot u_{j,k} - \sum_{i=1}^{N} a_i u_{i,k} \cdot u_{j,k} \right) = 0$$

$$\sum_{k=1}^{K}\left(y_k \cdot u_{j,k} \right) - \sum_{k=1}^{K}\sum_{i=1}^{N}\left(a_i \cdot u_{i,k} \cdot u_{j,k} \right) = 0$$

$$\sum_{k=1}^{K}\left(y_k \cdot u_{j,k} \right) - \sum_{i=1}^{N}\sum_{k=1}^{K}\left(a_i \cdot u_{i,k} \cdot u_{j,k} \right) = 0$$

$$\sum_{k=1}^{K}\left(y_k \cdot u_{j,k} \right) - \sum_{i=1}^{N} a_i \left[\sum_{k=1}^{K}\left(u_{i,k} \cdot u_{j,k} \right) \right] = 0$$

$$\therefore \quad \sum_{i=1}^{N} a_i \left[\sum_{k=1}^{K}\left(u_{i,k} \cdot u_{j,k} \right) \right] = \sum_{k=1}^{K}\left(y_k \cdot u_{j,k} \right).$$

This is a system of N equations that can be expressed in the following matrix form:

$$\bar{\bar{\mathbf{U}}}\bar{\mathbf{a}} = \bar{\bar{\mathbf{Y}}}$$

where the elements of the coefficient matrix are:

$$U_{i,j} = \sum_{k=1}^{K}\left(u_{i,k} \cdot u_{j,k} \right)$$

the elements of the unknown parameter vector are a_i, and the elements of the right-hand-side vector are:

$$Y_i = \sum_{k=1}^{K}\left(y_k \cdot u_{i,k} \right)$$

Forms for Polynomial Regression

Simple forms if the equation is a polynomial. Then the regression equation is:

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_N x_N$$

and the system of equations to find the parameters in matrix form is:

$$\begin{bmatrix} K & \sum_{k=1}^{K} x_k & \sum_{k=1}^{K} x_k^2 & \cdots & \sum_{k=1}^{K} x_k^N \\ \sum_{k=1}^{K} x_k & \sum_{k=1}^{K} x_k^2 & \sum_{k=1}^{K} x_k^3 & \cdots & \sum_{k=1}^{K} x_k^{N+1} \\ \sum_{k=1}^{K} x_k^2 & \sum_{k=1}^{K} x_k^3 & \sum_{k=1}^{K} x_k^4 & \cdots & \sum_{k=1}^{K} x_k^{N+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{K} x_k^N & \sum_{k=1}^{K} x_k^{N+1} & \sum_{k=1}^{K} x_k^{N+22} & \cdots & \sum_{k=1}^{K} x_k^{2N} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} y_k \\ \sum_{k=1}^{K} y_k x_k \\ \sum_{k=1}^{K} y_k x_k^2 \\ \vdots \\ \sum_{k=1}^{K} y_k x_k^N \end{bmatrix}.$$

This is a system of N equations that can be expressed in the following matrix form:

$$\bar{\bar{\mathbf{U}}}\bar{\mathbf{a}} = \bar{\bar{\mathbf{Y}}}$$

where the elements of the coefficient matrix are:

$$U_{i,j} = \sum_{k=1}^{K} \left(u_{i,k} \cdot u_{j,k} \right)$$

the elements of the unknown parameter vector are a_i, and the elements of the right-hand-side vector are:

$$Y_i = \sum_{k=1}^{K} \left(y_k \cdot u_{i,k} \right)$$

Putting an Equation Into Linear Form

Even equations that do not look linear can be algebraically manipulated into a linear form. For example, the following equation is for a response of a 1st order system with dead time to a step forcing function:

$$y' = y'_{\infty} \left[1 - \exp\left(-\frac{t - \theta}{\tau} \right) \right] \quad t \geq \theta$$

where τ, θ, and possibly y'_{∞} are the unknown parameters that are to be determined from a fit to data. The equation in this form is obviously non-linear – the unknowns τ & θ are in the exponential term. However, if we algebraically manipulate the equation we can get:

$$\frac{y'}{y'_\infty} = 1 - \exp\left(-\frac{t-\theta}{\tau}\right) \quad t \geq \theta$$
$$1 - \frac{y'}{y'_\infty} = \exp\left(-\frac{t-\theta}{\tau}\right) \quad t \geq \theta$$
$$\ln\left(1 - \frac{y'}{y'_\infty}\right) = \left(\frac{\theta}{\tau}\right) - \left(\frac{1}{\tau}\right)t \quad t \geq \theta .$$

Now this is in a linear form so long as y'_∞ is not considered an unknown parameter:

$$Y = a_0 + a_1 t .$$

Where:

$$Y \equiv \ln\left(1 - \frac{y'}{y'_\infty}\right)$$
$$a_1 \equiv -\left(\frac{1}{\tau}\right) \quad \Rightarrow \quad \tau = -\frac{1}{a_1}$$
$$a_0 \equiv \left(\frac{\theta}{\tau}\right) \quad \Rightarrow \quad \theta = \tau \cdot a_0 = -\frac{a_0}{a_1}.$$

The following Figures 1 & 2 show how the data points can be linearized. The curve in Figure 2 shows that a straight line can be obtained, but this requires that we have a good value for y'_∞. From the data it is pretty easy to see that $y'_\infty = 1$. However, it is not always possible to come up with a good value for y'_∞ before the regression is done. If a poor value of y'_∞ is used then the resulting transformed data will not form a straight line – Figure 3 shows this.

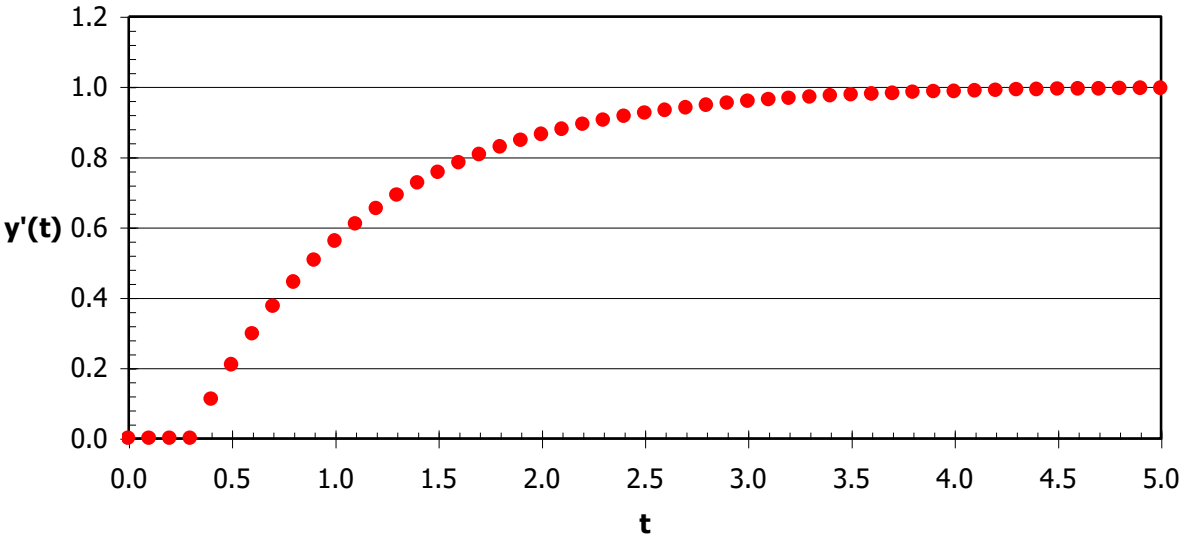


Figure 1. Data Points for FOPDT Process With Step Change Forcing Function

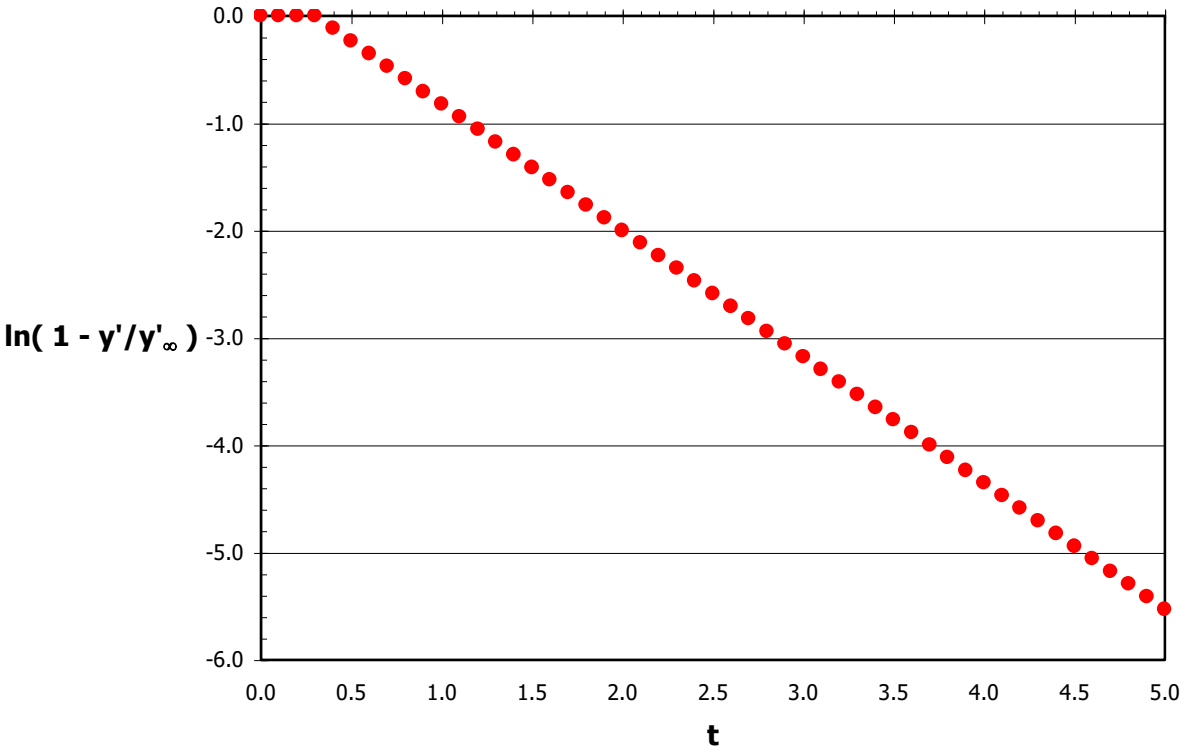


Figure 2. Data Points After Transforming FOPDT Response Equation

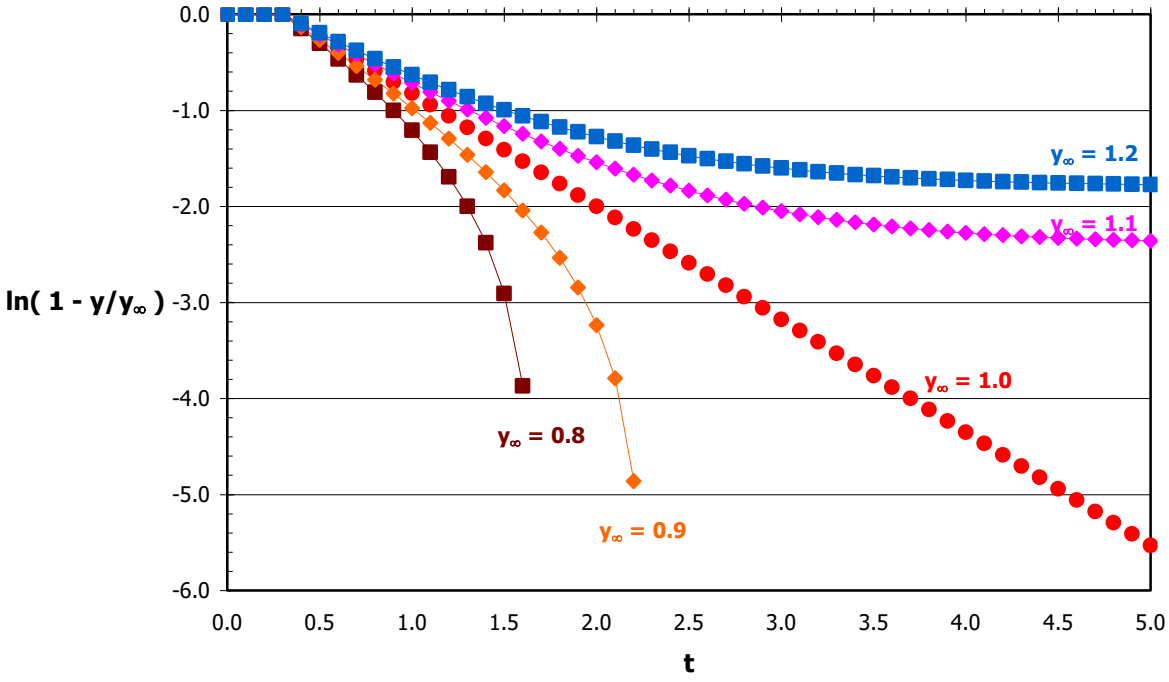


Figure 3. Data Points After Transforming FOPDT Response Equation

Non-Linear Regression

Much more general because the regression equation does not have to be in a specific form. Set up an objective function of errors that should be minimized. If the regression equation has a form of:

$$y = \mathrm{f}\left(x_1, x_2, \ldots, x_M, a_0, a_1, \ldots, a_N\right)$$

then to be consistent with linear regression the objective function should be of the form:

$$F \equiv \sum_{k=1}^{K} \varepsilon_k^2 = \sum_{k=1}^{K} \left[y_k - \mathrm{f}\left(x_{1,k}, x_{2,k}, \ldots, x_{M,k}, a_0, a_1, \ldots, a_N\right) \right]^2 .$$

Here the parameters a_0, a_1, ..., a_N are the model parameters & they are to be adjusted to minimize the value of this objective function.

For example, non-linear regression could be performed directly on the response curve of an FOPDT driven by a step change by creating an objective function of the form:

$$F = \sum_{k=1}^{K} \left\{ y'_k - y'_\infty \left[1 - \exp\left(-\frac{t_k - \theta}{\tau} \right) \right] \right\}^2$$

and then directly iterating on the values of τ, θ, and even y_∞.

Numerical Example Using Excel

The following table presents data for fitting to a FOPDT process response to a step change. The form of the response curve is:

$$y(t) = \begin{cases} y^* & t < \theta \\ y^* + y'_\infty \left[1 - \exp\left(-\frac{t - \theta}{\tau} \right) \right] & t \geq \theta \end{cases}$$

where τ, θ, and y'_∞ are the unknown parameters that are to be determined from a fit to data.

Note that this data was generated from a 3rd order process model's response to a step change:

$$\bar{y}' = \frac{10}{s(5s+1)(4s+1)(0.4s+1)}$$

Note that the Taylor series approximation FOPDT model would be:

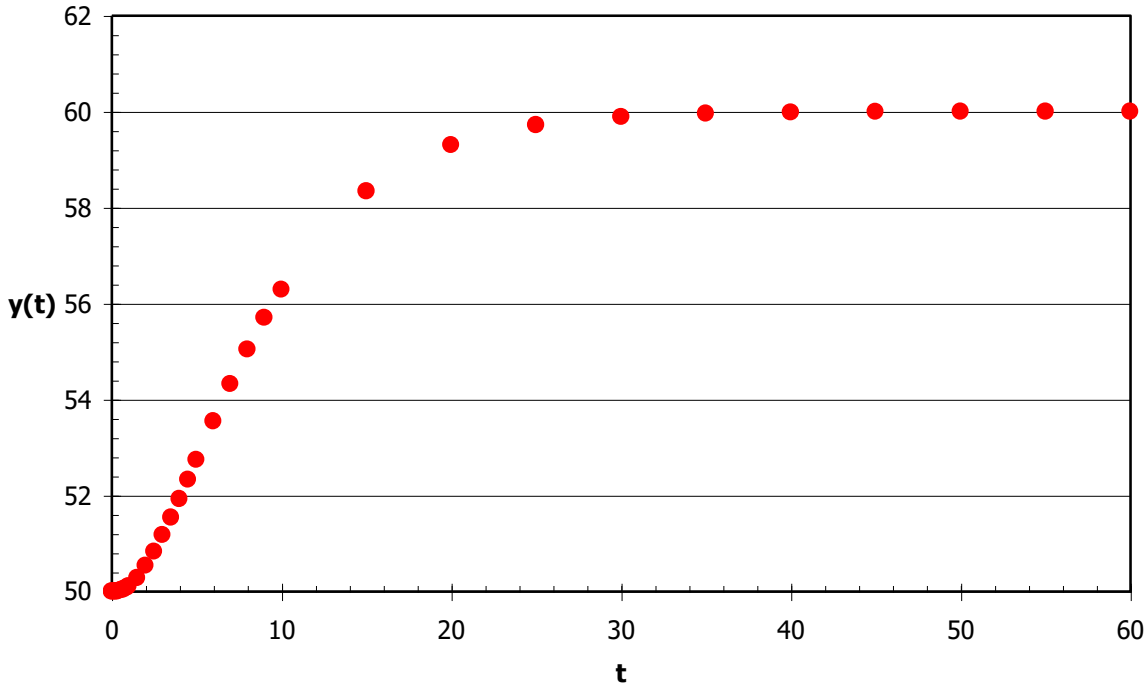
$$\bar{y}' = \frac{1}{s} \cdot \frac{10}{(5s+1)(4s+1)(0.4s+1)} = \frac{10e^{-4.4s}}{5s+1}.$$

and the Skogestad FOPDT model would be:

$$\bar{y}' = \frac{1}{s} \cdot \frac{10}{(5s+1)(4s+1)(0.4s+1)} = \frac{1}{s} \cdot \frac{10e^{-(4/2+0.4)s}}{\left(5+\frac{1}{2}\cdot 4\right)s+1} = \frac{10e^{-2.4s}}{7s+1}.$$

t	$y(t)$	t	$y(t)$	t	$y(t)$	t	$y(t)$
0	50.00	0.9	50.08	5	52.74	30	59.89
0.1	50.00	1.0	50.11	6	53.55	35	59.96
0.2	50.00	1.5	50.28	7	54.32	40	59.98
0.3	50.00	2.0	50.53	8	55.04	45	59.99
0.4	50.01	2.5	50.83	9	55.70	50	60.00
0.5	50.02	3.0	51.17	10	56.29	55	60.00
0.6	50.03	3.5	51.54	15	58.34	60	60.00
0.7	50.04	4.0	51.93	20	59.30		
0.8	50.06	4.5	52.33	25	59.72		

The following figure plots this data. Note that the values at the higher times have steadied out to a value of 60. So, a reasonable ultimate value is $y_\infty = 60 \Rightarrow y'_\infty = 10$.



Linear Regression Using the Regression Data Analysis Add-In

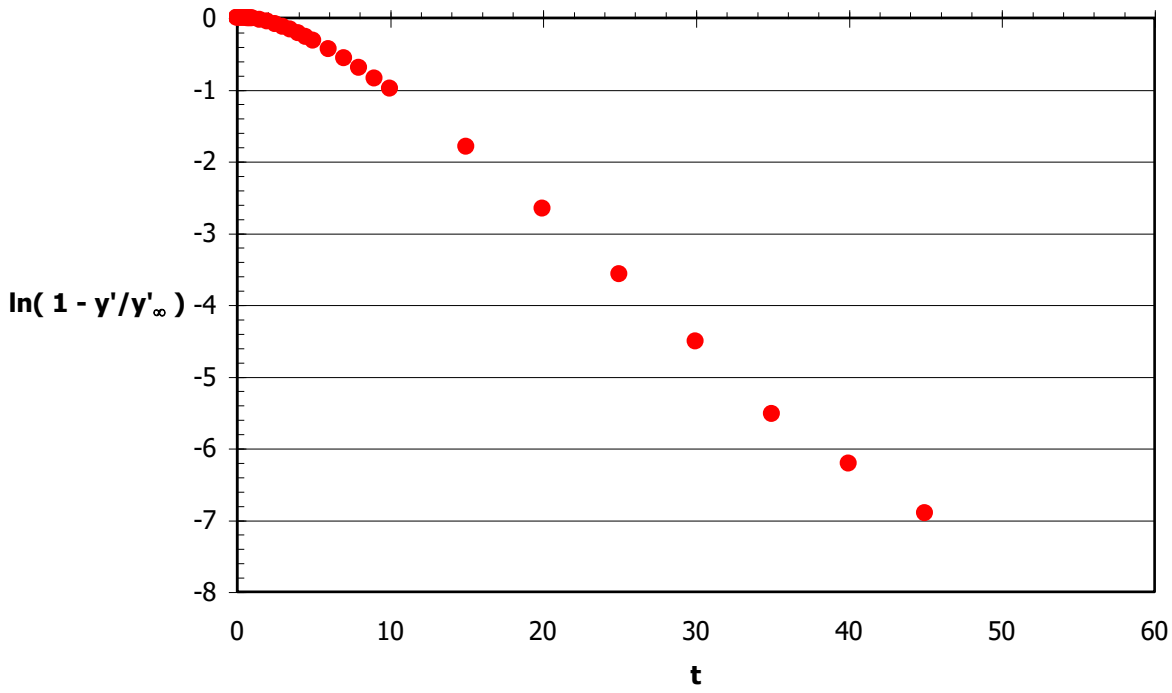
There are several ways in which linear regression can be performed in Excel. Some of these ways are:

- Using the Regression Data Analysis add-in.
- Using the curve fitting option directly on a chart.
- Using the built-in function LINEST.

We will only discuss the first way here.

All of these procedures require the data in its manipulated form. Here we've used a couple values directly observed from the data: $y^* = 50$ & $y'_\infty = 60 - 50 = 10$. The following table & figure show this. Notice that transformed values of the values at the largest times can not be calculated – this is not really a problem since these points are not significant to the dynamic response of the system.

	A	B	C	D
4	t	y	y'	ln(1 -y'/y'∞)
5	0	50.00	0.00	0.000
6	0.1	50.00	0.00	0.000
7	0.2	50.00	0.00	0.000
8	0.3	50.00	0.00	0.000
9	0.4	50.01	0.01	-0.001
10	0.5	50.02	0.02	-0.002
11	0.6	50.03	0.03	-0.003
12	0.7	50.04	0.04	-0.004
13	0.8	50.06	0.06	-0.006
14	0.9	50.08	0.08	-0.008
15	1	50.11	0.11	-0.011
16	1.5	50.28	0.28	-0.028
17	2	50.53	0.53	-0.054
18	2.5	50.83	0.83	-0.087
19	3	51.17	1.17	-0.124
20	3.5	51.54	1.54	-0.167
21	4	51.93	1.93	-0.214
22	4.5	52.33	2.33	-0.265
23	5	52.74	2.74	-0.320
24	6	53.55	3.55	-0.439
25	7	54.32	4.32	-0.566
26	8	55.04	5.04	-0.701
27	9	55.70	5.70	-0.844
28	10	56.29	6.29	-0.992
29	15	58.34	8.34	-1.796
30	20	59.30	9.30	-2.659
31	25	59.72	9.72	-3.576
32	30	59.89	9.89	-4.510
33	35	59.96	9.96	-5.521
34	40	59.98	9.98	-6.215
35	45	59.99	9.99	-6.908
36	50	60.00	10.00	
37	55	60.00	10.00	
38	60	60.00	10.00	

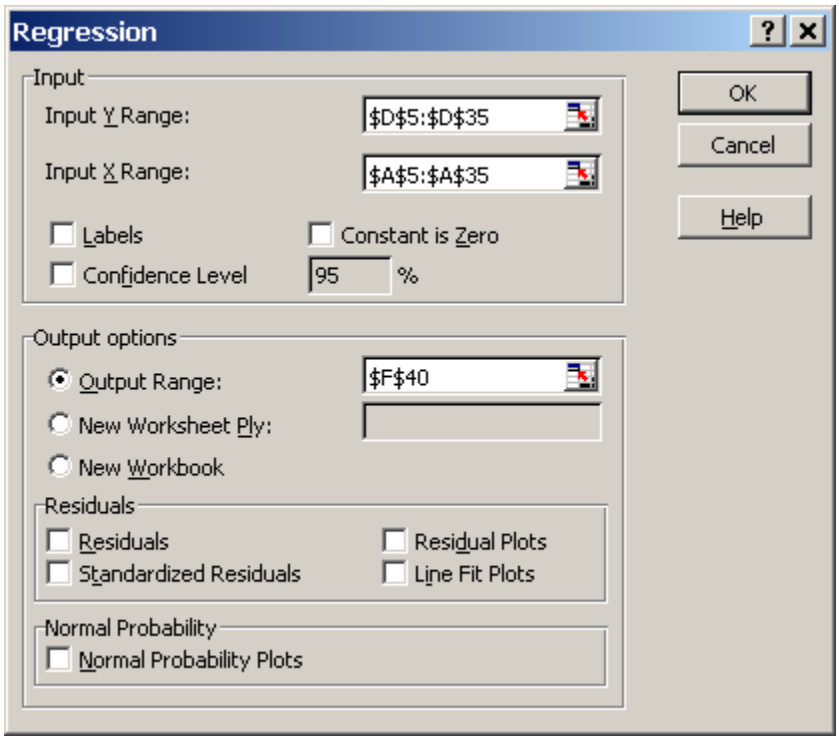


To use the Regression add-in it must first be installed. This can be done by the commands:

- Tools, Add-Ins …
- Place a check mark in front of the Analysis TookPak in the list presented.

Start the regression process with the command Tools, Data Analysis … Chose the option Regression from the list presented. You will then get a window like the one shown below. The important items to fill out are:

- What cells have the transformed function data? Here the values are in the cells denoted as Input Y Range.
- What cells have the time data? Here the values are in the cells denoted as Input X Range.
- Where do you want your calculated values reported? The upper right hand cell is specified as Output Range.



The following shows the report that is generated by this procedure. There are various statistics reported (most of which we'll ignore for right now) – what we're most interested in are the values in the Coefficients column of the last table. From these values the regression equation will be:

$$\ln\left(1-\frac{y'}{y'_\infty}\right)=a_0+a_1 t=0.263884-0.15702\cdot t\,.$$

From this we can determine the time constant & the time delay:

$$\tau=-\frac{1}{a_1}=\frac{1}{0.15702}=6.37$$
$$\theta=-\frac{a_0}{a_1}=\frac{0.263884}{0.15702}=1.67\,.$$

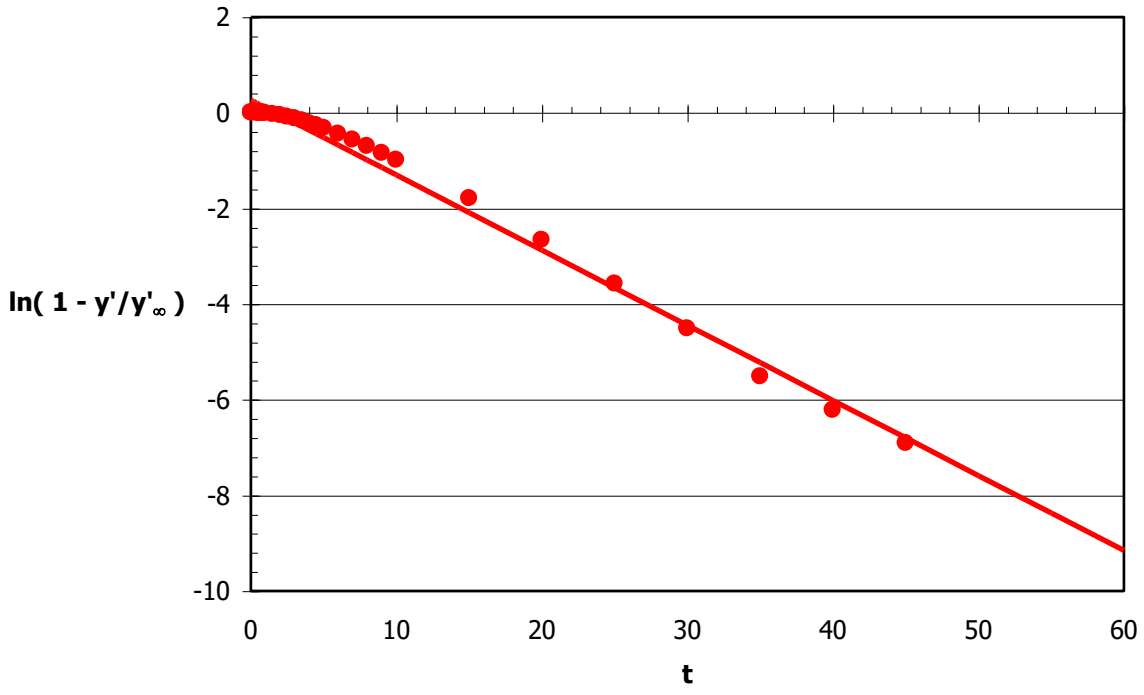
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.994951
R Square	0.989928
Adjusted R	0.989581
Standard E	0.205303
Observatio	31

ANOVA

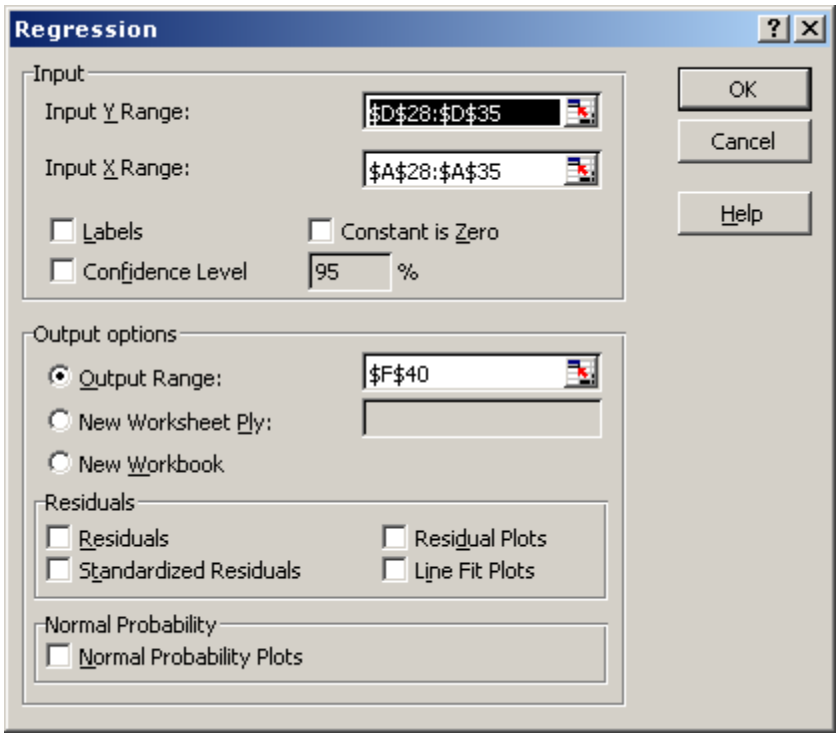
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	120.1397	120.1397	2850.33	1.64E-30
Residual	29	1.222333	0.042149		
Total	30	121.3621			

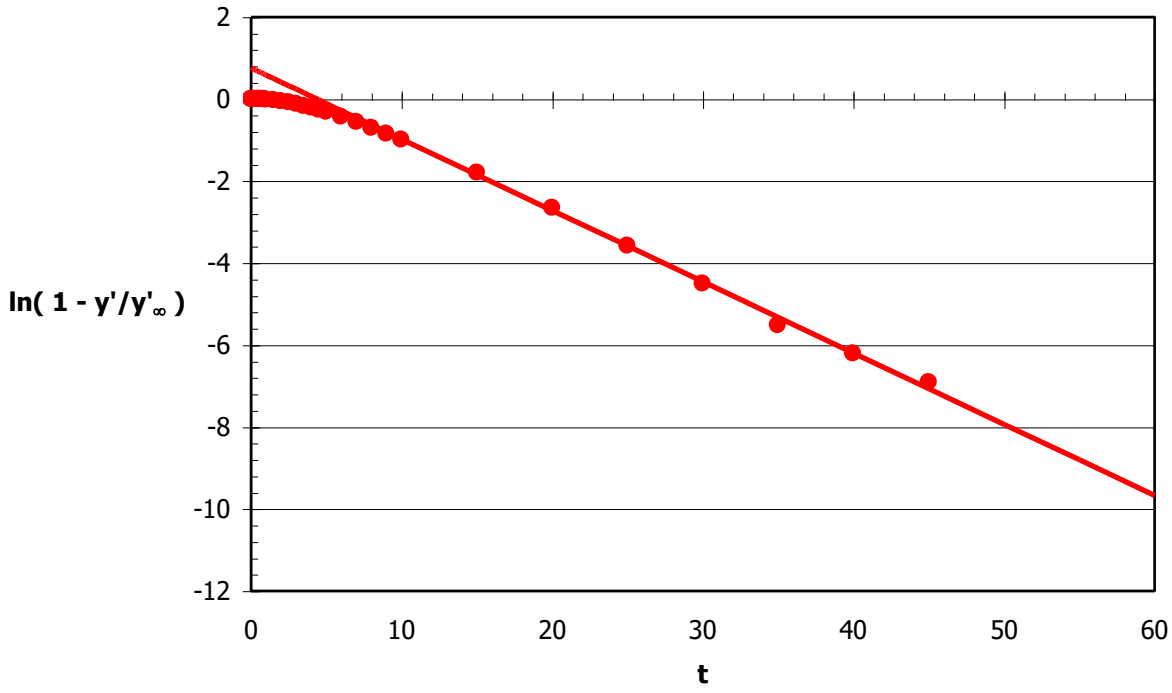
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0.263884	0.04553	5.795897	2.79E-06	0.170766	0.357003	0.170766	0.357003
X Variable	-0.15702	0.002941	-53.3885	1.64E-30	-0.16304	-0.15101	-0.16304	-0.15101



The chart shows the regressed line compared to the original data. It does not look very good. Why? All of the points were specified, including the early time values that are going to be approximated with the time delay. It is probably better to not use those points in the

regression. So, let's re-run the regression only using the points for $t \geq 10$. When you re-open the regression window it remembers what you had done before. Let's change the range of data points to be used for the regression and re-run.





Now the fit looks much better. Now we get the results:

$$\ln\left(1-\frac{y'}{y'_\infty}\right)=a_0+a_1 t=0.759653-0.17388\cdot t\,.$$

From this we can determine the time constant & the time delay:

$$\tau=-\frac{1}{a_1}=\frac{1}{0.17388}=5.75$$
$$\theta=-\frac{a_0}{a_1}=\frac{0.759653}{0.17388}=4.37\,.$$

> Note that these parameters match those of the Taylor series approximation quite well.

Non-Linear Regression

With non-linear regression one can work with the model equation directly and no modifications need be done. The image on the right shows a section of an Excel spreadsheet with the given data, corresponding values from the model equation, & the error in the model predictions.

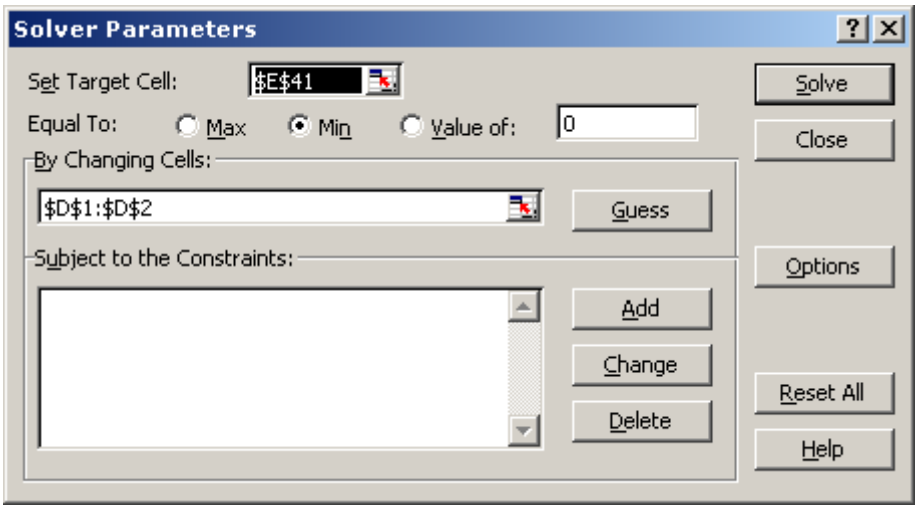
The objective function, the sum of the errors squared, is calculated in cell E41. In this particular spreadsheet, only those error values that we want to consider for the objective function will be calculated. Just as before we'll restrict ourselves to the values associated with $10 \le t \le 45$.

Also note that we must specify initial estimates for all parameters.

	A	B	C	D	E
1			τ =	1	
2			θ =	0	
3			y'_∞ =	10	
4					
5	t	y	y'	Model y'	Δy'
6	0	50.00	0.00	0	
7	0.1	50.00	0.00	0.951626	
8	0.2	50.00	0.00	1.812692	
9	0.3	50.00	0.00	2.591818	
10	0.4	50.01	0.01	3.2968	
11	0.5	50.02	0.02	3.934693	
12	0.6	50.03	0.03	4.511884	
13	0.7	50.04	0.04	5.034147	
14	0.8	50.06	0.06	5.50671	
15	0.9	50.08	0.08	5.934303	
16	1	50.11	0.11	6.321206	
17	1.5	50.28	0.28	7.768698	
18	2	50.53	0.53	8.646647	
19	2.5	50.83	0.83	9.17915	
20	3	51.17	1.17	9.502129	
21	3.5	51.54	1.54	9.698026	
22	4	51.93	1.93	9.816844	
23	4.5	52.33	2.33	9.88891	
24	5	52.74	2.74	9.932621	
25	6	53.55	3.55	9.975212	
26	7	54.32	4.32	9.990881	
27	8	55.04	5.04	9.996645	
28	9	55.70	5.70	9.998766	
29	10	56.29	6.29	9.999546	3.71
30	15	58.34	8.34	9.999997	1.66
31	20	59.30	9.30	10	0.70
32	25	59.72	9.72	10	0.28
33	30	59.89	9.89	10	0.11
34	35	59.96	9.96	10	0.04
35	40	59.98	9.98	10	0.02
36	45	59.99	9.99	10	0.01
37	50	60.00	10.00	10	
38	55	60.00	10.00	10	
39	60	60.00	10.00	10	
40					
41				*Sum Errors Squared =*	*17.0989*

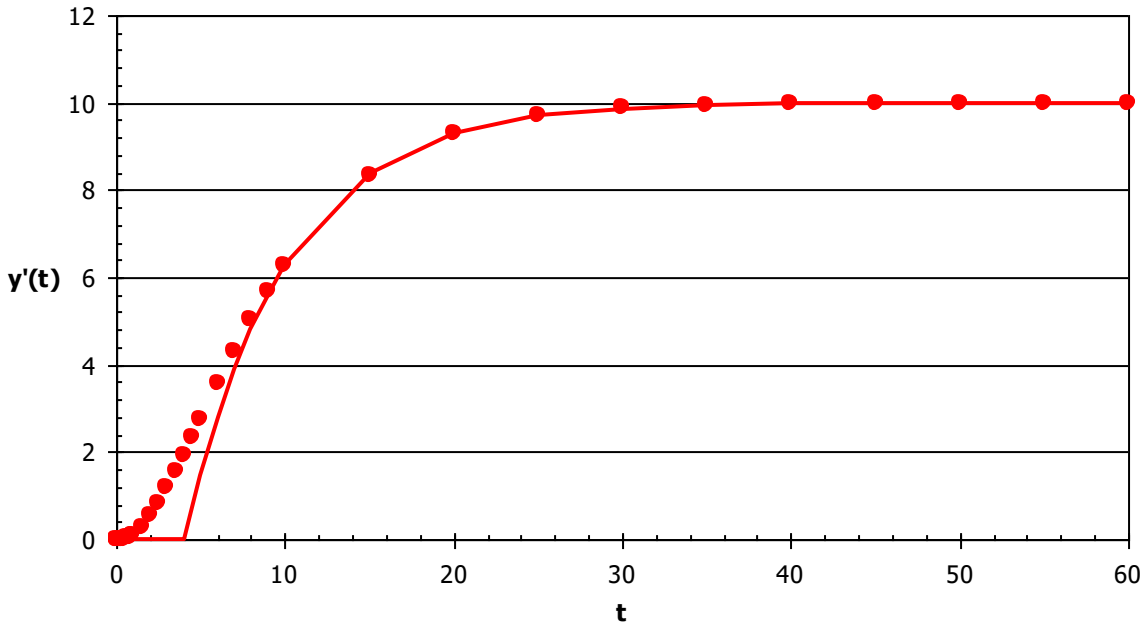
We will use the Solver tool to do the non-linear regression. Start the regression process with the command Tools, Solver … The important items to fill out are:

- What cell represents the objective function? Specify this cell in the Set Target Cell box & that is to be minimized by clicking Min.
- What cells are to be modified? The cells for τ & θ are specified in the By Changing Cells box.



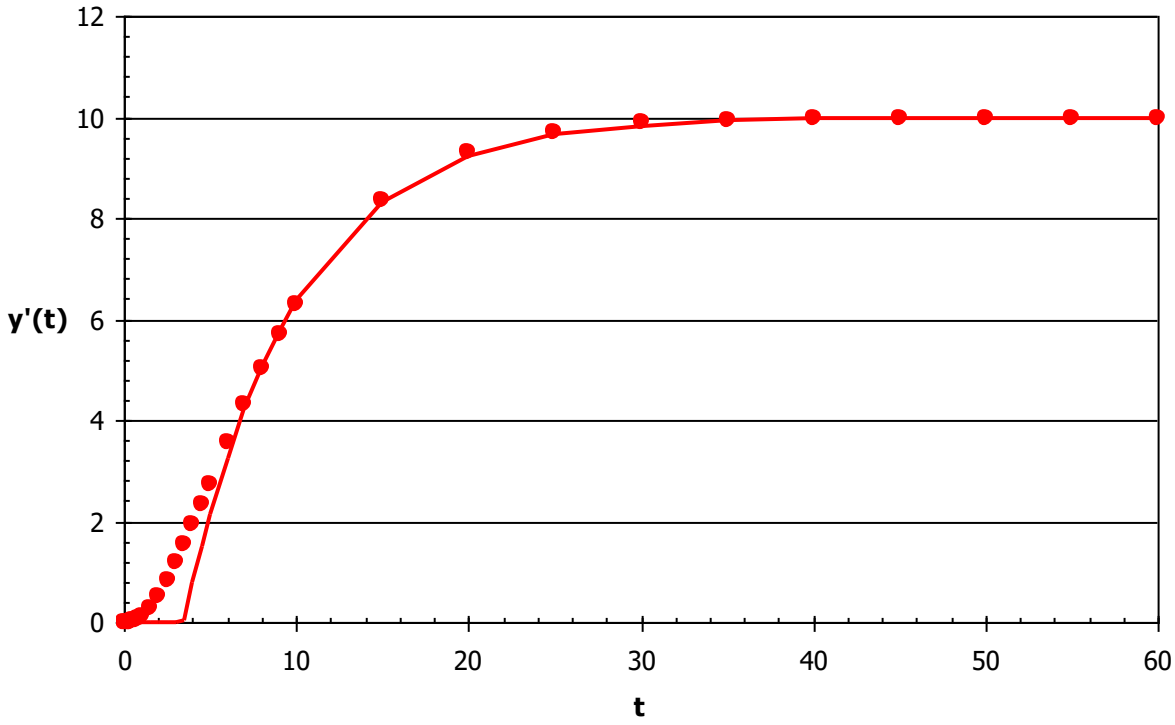
After clicking Solve, we get the following result. The fit is very good for times $t \geq 10$ (the data used for the regression) & the associated parameters are also shown. The figure suggests that we can add in more of the shorter time values, too, & still get a good fit to the large time values.

τ =	6.029618
θ =	4.038509
y'_∞ =	10



The following set of results makes use of the data for $t \geq 7$. The fit still looks very good for these values used for the regression. The associated parameters have changed slightly. With some experimentation one finds that including more short time data makes the fit near the maximum (for times around 20) worse. This is probably the best FOPDT fit to the data.

τ =	6.442579
θ =	3.463511
y'_∞ =	10


12
10
8
6
4
2
0
y'(t)
0
10
20
30
40
50
60
t