

Linear Regression

(by Dr. Tomas Co 5/8/2008)

Definition:

Linear regression refer to the process of fitting a linear model of the form given by equation (1) to a set of given data.

$$y = a_n x_n + \dots + a_1 x_1 + a_0 \quad (1)$$

where x_1, x_2, \dots, x_n are the n independent variables, y is the dependent variable, and a_0, a_1, \dots, a_n are the coefficients of the model.

The simplest case is the equation of a line, when $n = 1$, often written as

$$y = mx + b \quad (2)$$

where $m = a_1$ is the slope and $b = a_0$ is the intercept.

Approaches:

Let the k^{th} data point be denoted by $[y_k, (x_1)_k, \dots, (x_n)_k]$, $k = 1, \dots, P$, where P is the number of data points which must be larger than $(n + 1)$.

Method 1: Matrix Calculations

1. Set up matrix \mathbf{M} and vector \mathbf{h} . (Note: the size of \mathbf{M} is P rows by $(n + 1)$ columns while the length of vector \mathbf{h} is P .)

$$M = \begin{pmatrix} (x_1) & \dots & (x_n)_P & 1 \\ \vdots & \ddots & \vdots & \vdots \\ (x_1)_P & \dots & (x_n)_P & 1 \end{pmatrix} \quad h = \begin{pmatrix} y_1 \\ \vdots \\ y_P \end{pmatrix} \quad (3)$$

2. Set up solution vector \mathbf{v} . (Note: the length of column vector \mathbf{v} is $(n + 1)$).

$$v = \begin{pmatrix} a_n \\ \vdots \\ a_1 \\ a_0 \end{pmatrix} \quad (4)$$

3. Solve for \mathbf{v} using the least squares formula (aka the *normal equation*),

$$Mv \simeq h \quad \rightarrow \quad v = (M^T M)^{-1} M^T h \quad (5)$$

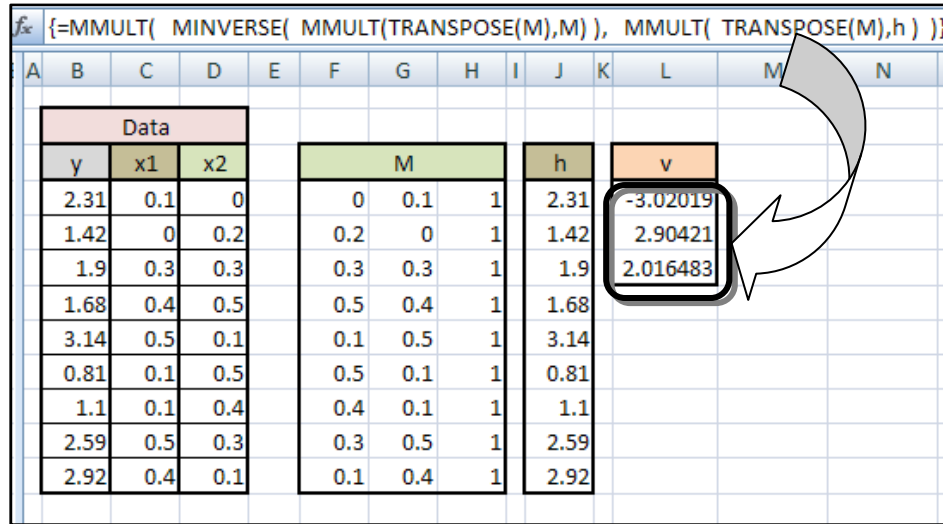


Figure 1. Linear regression via matrix calculations.
 (Note: this is an array formula, i.e. select range for vector **v** and then use [CTRL Shift ENTER]).

Method 2: Using the **LINEST** Function.

1. Set up the data and the solution region.

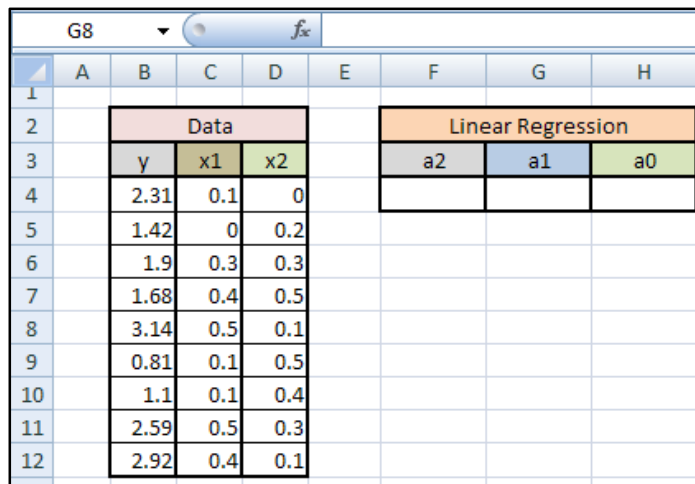


Figure 2. Set up for data and solution region.

2. Select the solution region and then implement the **LINEST** function (Note: use [**CTRL Shift ENTER**] to invoke the array function) as shown in Figure 3.

Data				Linear Regression		
y	x1	x2		a2	a1	a0
2.31	0.1	0		-3.02019	2.90421	2.016483
1.42	0	0.2				
1.9	0.3	0.3				
1.68	0.4	0.5				
3.14	0.5	0.1				
0.81	0.1	0.5				
1.1	0.1	0.4				
2.59	0.5	0.3				
2.92	0.4	0.1				

Figure 3. Implement the **LINEST** function.

The range **B4:B12** is the set of y values while the range **C4:D12** is for the x_1 and x_2 values. The third argument is set to **TRUE** to mean that we need the value of $a_0 \neq 0$, otherwise we set it to **FALSE** if we want to force $a_0 = 0$. The fourth argument is set to **FALSE** to mean that we are not requesting for the calculation of statistical parameters. (See section below for the case when the fourth argument is set to **TRUE**).

Statistics from **LINEST**:

1. Some statistic are available when the fourth argument of **LINEST** is set to **TRUE**. To access these, one needs to expand the result region to include four more rows as shown in Figure 4.

Data				Linear Regression		
y	x1	x2		a2	a1	a0
2.31	0.1	0		-3.02019	2.90421	2.016483
1.42	0	0.2		0.08202	0.076357	0.033914
1.9	0.3	0.3		0.998016	0.041715	#N/A
1.68	0.4	0.5		1509.295	6	#N/A
3.14	0.5	0.1		5.252781	0.010441	#N/A
0.81	0.1	0.5				
1.1	0.1	0.4				
2.59	0.5	0.3				
2.92	0.4	0.1				

2. The result region has $(n + 1)$ columns and 5 rows described in Table 1,

Table 1. Regression and statistics resulting from LINEST.

a_n	...	a_1	a_0
Δa_n	...	Δa_1	Δa_0
R^2	σ_{res}		
F	ν_2		
SS_{reg}	SS_{res}		

The first row still contains the coefficients a_n, \dots, a_1, a_0 . The second row contains the standard error values for each corresponding coefficient. For example, the coefficient a_2 has a standard error uncertainty of 0.0802 given in cell F5. The other results are summarized in Table 2, where k^{th} regressed data is denoted by \hat{y}_k , and the average of y_k is denoted by \bar{y} , i.e.

$$\hat{y}_k = a_n(x_n)_k + \dots + a_1(x_1)_k + a_0 \quad (6)$$

$$\bar{y} = \frac{\sum_{k=1}^P y_k}{P} \quad (7)$$

Table 2. Description of the **LINEST** Statics.

Item	Description	Formula	Application
R^2	Coefficient of Determination	$\frac{\sum_{k=1}^p (\hat{y}_k - \bar{y})^2}{\sum_{k=1}^p (y_k - \bar{y})^2}$	<ul style="list-style-type: none"> - range: $0 \leq R^2 \leq 1$ - if close to 1, model is good
σ_{res}	Standard deviation of the residuals	$\sqrt{\frac{SS_{\text{reg}}}{v_2}}$	<ul style="list-style-type: none"> - smaller value means model predicted points are close to data points.
F	F -distribution value	$\frac{SS_{\text{reg}} / v_1}{SS_{\text{res}} / v_2}$ <p>where $v_1 = n - 1$</p>	<ul style="list-style-type: none"> - if greater than $F_{\text{critical}}(\alpha, n - 1, v_2)$ then regression is justified with an α-confidence level.
v_2	Degree of Freedom for residuals	$P - n$	<ul style="list-style-type: none"> - if zero then solution demands an equality instead of approximation
SS_{reg}	Sum of Squares of regressor errors	$\sum_{k=1}^p (\hat{y}_k - \bar{y})^2$	<ul style="list-style-type: none"> - indicates variability of regressed data
SS_{res}	Sum of Squares of residual errors	$\sum_{k=1}^p (\hat{y}_k - y_k)^2$	<ul style="list-style-type: none"> - indicates variability of the residuals

Remarks:

- a) The formula for the standard errors of the coefficients is given by

$$\Delta a_k = \sigma_{\text{res}} \sqrt{G_{kk}} \quad (8)$$

where G_{kk} is the k^{th} diagonal entry of $G = (M^T M)^{-1}$, with matrix M as defined in equation (3).

- b) To obtain a 95% confidence interval for each parameter, we need to multiply the standard error by the t -distribution value corresponding to desired confidence level. In Excel, this can be found using the **TINV** function. For example, for a 95% confidence interval of a_2 in the example above, we have

$$t_{95} = \mathbf{TINV}(1 - 0.95, \nu_2) = 2.4469$$

$$\Delta_{95} a_2 = \Delta a_2 \cdot t_{95} = 0.2007$$

$$\rightarrow 95\% \text{ confidence interval for } a_2 = -3.0202 \pm 0.2007$$

- c) To perform the F -test, we can also use the **FINV** function in Excel. For our example, $F_{95} = \mathbf{FINV}(1 - 0.95, \nu_1, \nu_2) = 5.1432$. Since, $F = 1509.3 > 5.1432$, we can justify the linear model proposed with 95% confidence.