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.

$$y = a_n x_n + \cdots + a_1 x_1 + a_0 \tag{1}$$

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

$$y = mx + b \tag{2}$$

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 $(n + 1)$.

<u>Method 1: Matrix Calculations</u>

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

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

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

$$v = \begin{pmatrix} a_n \\ \vdots \\ a_1 \\ a_0 \end{pmatrix} \tag{4}$$

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

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

{=MMULT(MINVERSE(MMULT(TRANSPOSE(M),M)), MMULT(TRANSPOSE(M),h))}

Data				M			h	v
y	x1	x2						
2.31	0.1	0		0	0.1	1	2.31	-3.02019
1.42	0	0.2		0.2	0	1	1.42	2.90421
1.9	0.3	0.3		0.3	0.3	1	1.9	2.016483
1.68	0.4	0.5		0.5	0.4	1	1.68	
3.14	0.5	0.1		0.1	0.5	1	3.14	
0.81	0.1	0.5		0.5	0.1	1	0.81	
1.1	0.1	0.4		0.4	0.1	1	1.1	
2.59	0.5	0.3		0.3	0.5	1	2.59	
2.92	0.4	0.1		0.1	0.4	1	2.92	

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.

G8

	A	B	C	D	E	F	G	H
1								
2		Data				Linear Regression		
3		y	x1	x2		a2	a1	a0
4		2.31	0.1	0				
5		1.42	0	0.2				
6		1.9	0.3	0.3				
7		1.68	0.4	0.5				
8		3.14	0.5	0.1				
9		0.81	0.1	0.5				
10		1.1	0.1	0.4				
11		2.59	0.5	0.3				
12		2.92	0.4	0.1				

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.

F4 | {=LINEST(B4:B12,C4:D12,TRUE,FALSE)}

	A	B	C	D	E	F	G	H
1								
2		Data				Linear Regression		
3		y	x1	x2		a2	a1	a0
4		2.31	0.1	0		-3.02019	2.90421	2.016483
5		1.42	0	0.2				
6		1.9	0.3	0.3				
7		1.68	0.4	0.5				
8		3.14	0.5	0.1				
9		0.81	0.1	0.5				
10		1.1	0.1	0.4				
11		2.59	0.5	0.3				
12		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.

F4 | {=LINEST(B4:B12,C4:D12,TRUE,TRUE)}

	A	B	C	D	E	F	G	H
1								
2		Data				Linear Regression		
3		y	x1	x2		a2	a1	a0
4		2.31	0.1	0		-3.02019	2.90421	2.016483
5		1.42	0	0.2		0.08202	0.076357	0.033914
6		1.9	0.3	0.3		0.998016	0.041715	#N/A
7		1.68	0.4	0.5		1509.295	6	#N/A
8		3.14	0.5	0.1		5.252781	0.010441	#N/A
9		0.81	0.1	0.5				
10		1.1	0.1	0.4				
11		2.59	0.5	0.3				
12		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	$\cdots$	a_1	a_0
Δa_n	$\cdots$	Δa_1	Δa_0
R^2	σ_{res}		
F	ν_2		
SS_{reg}	SS_{res}		

The first row still contains the coefficients $a_n, \ldots, a_1, a_0$. 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 $k^{\underline{\text{th}}}$ regressed data is denoted by $\hat{y}_k$, and the average of y_k is denoted by $\bar{y}$, i.e.

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

$$\bar{y} = \frac{\sum_{k=1}^{P} y_k}{P} \tag{7}$$

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

Item	Description	Formula	Application
R^2	Coefficient of Determination	$\frac{\sum_{k=1}^{P}(\hat{y}_k - \bar{y})^2}{\sum_{k=1}^{P}(y_k - \bar{y})^2}$	- range: $0 \leq R^2 \leq 1$ - if close to 1, model is good
σ_{res}	Standard deviation of the residuals	$\sqrt{\frac{SS_{\text{reg}}}{\nu_2}}$	- smaller value means model predicted points are close to data points.
F	F-distribution value	$\frac{SS_{\text{reg}}/\nu_1}{SS_{\text{res}}/\nu_2}$ where $\nu_1 = n - 1$	- if greater than $F_{\text{critical}}(\alpha, n-1, \nu_2)$ then regression is justified with an α-confidence level.
ν_2	Degree of Freedom for residuals	$P - n$	- if zero then solution demands an equality instead of approximation
SS_{reg}	Sum of Squares of regressor errors	$\sum_{k=1}^{P}(\hat{y}_k - \bar{y})^2$	- indicates variability of regressed data
SS_{res}	Sum of Squares of residual errors	$\sum_{k=1}^{P}(\hat{y}_k - y_k)^2$	- indicates variability of the residuals

Remarks:

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

$$\Delta a_k = \sigma_{\text{res}}\sqrt{G_{kk}} \tag{8}$$

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

$$t95 = \textbf{TINV}(1 - 0.95, \nu_2) = 2.4469$$
$$\Delta_{95}a_2 = \Delta a_2 \cdot t95 = 0.2007$$
$$\rightarrow \text{ 95\% confidence interval for } a_2 = -3.0202 \pm 0.2007$$

c) To perform the F-test, we can also use the **FINV** function in Excel. For our example, $F95 = \textbf{FINV}(1 - 0.95, \nu_1, \nu_2) = 5.1432$. Since, $F = 1509.3 > 5.1432$, we can justify the linear model proposed with 95% confidence.