

# Linear Regression using Excel

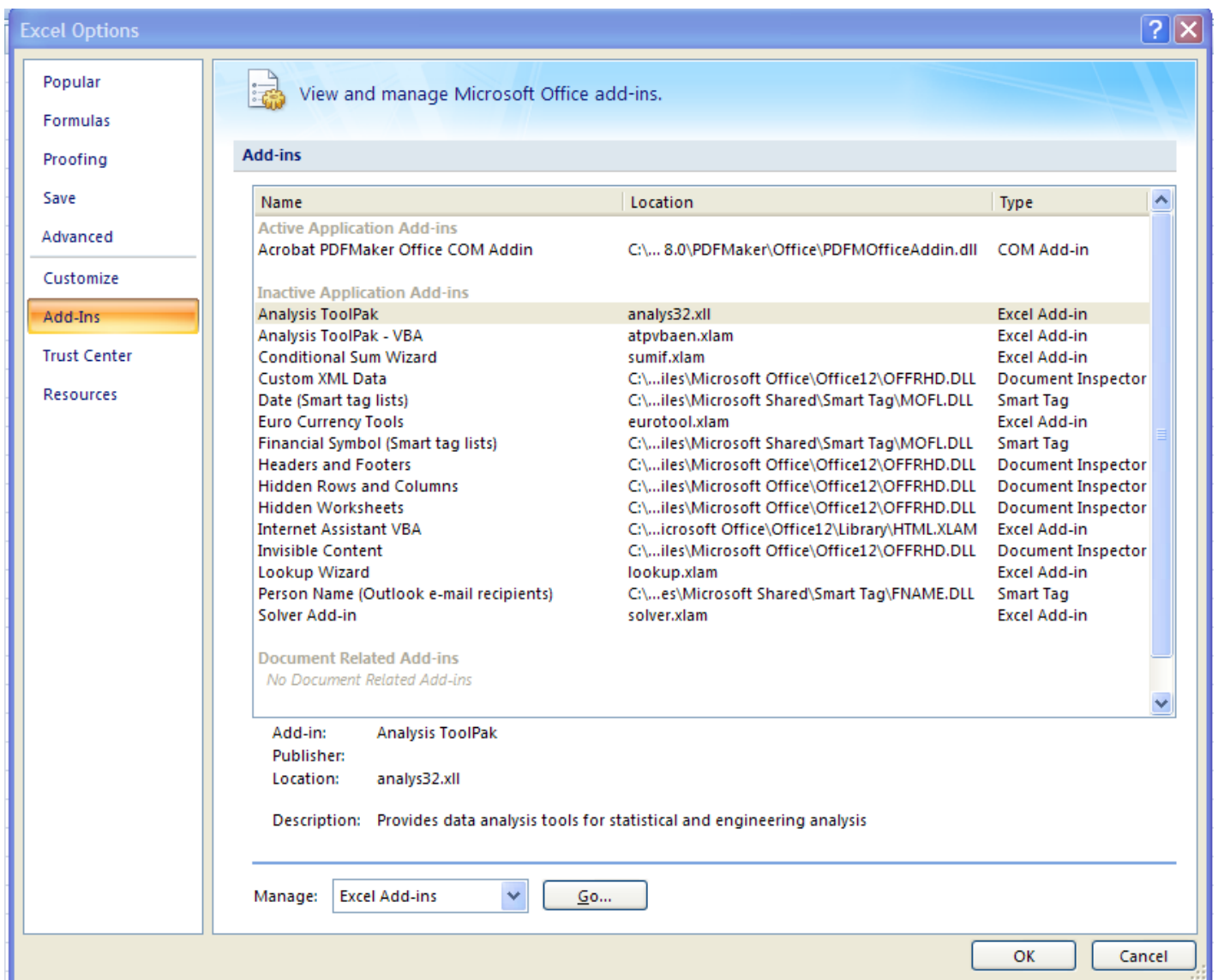
*Subject: Linear Regression using Excel*

*Application: Microsoft Excel 2007*

*Task: I want to find a linear equation that best describes a data set*

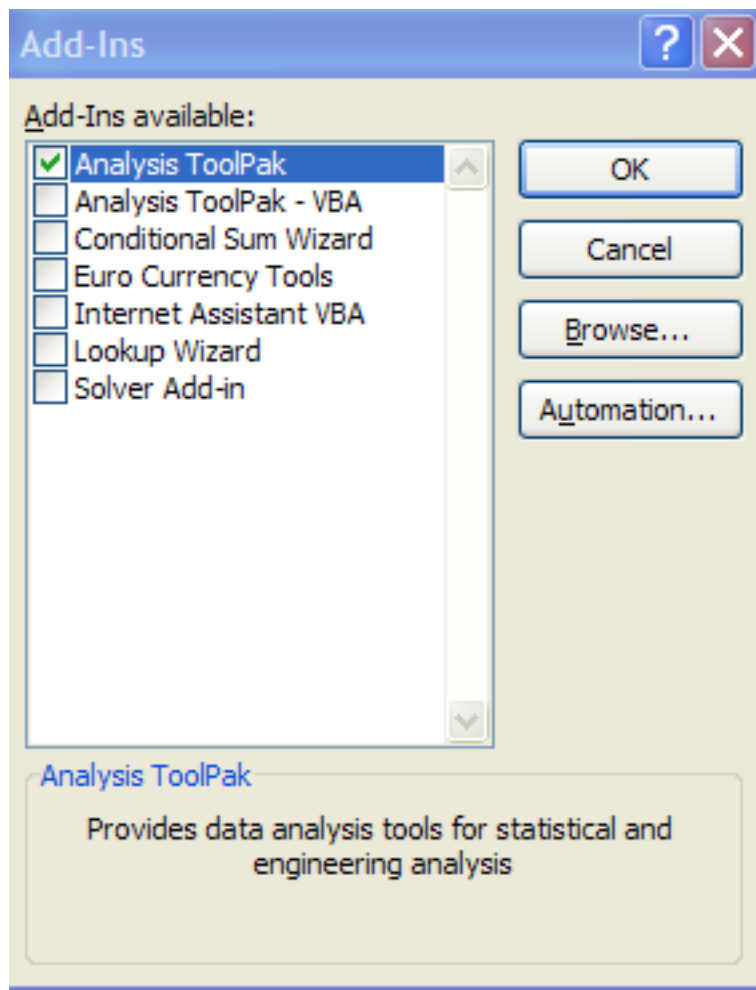
*Tutorial Date: 17th February, 2010 by Nathan Smith*

## Install the Analysis ToolPak



1. In the Microsoft Office button, go to excel options to click Add-ins
2. In the Add-Ins box, select Analysis ToolPak and click Go...

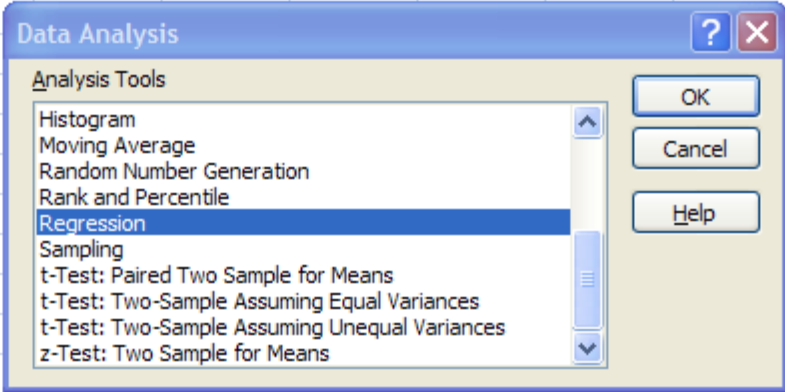
## Install the Analysis ToolPak (continue)



3. In the Add-Ins available box, check the Analysis ToolPak and then OK
4. If Analysis ToolPak is not listed in the Add-Ins available box, click Browse to locate it.

## Linear Regression using the Data Analysis Add-In

x	y
1.5	2.2
3.6	3.1
5.6	8.6
4.5	5.9
6.7	4.4



Suppose we want to determine a whether Y is a function of X

$$(Y)_i = a + b(X)_i + (\text{error})_i$$

where:

$(Y)_i$  = value of Y for observation i

a = mean value of Y when X is zero (intercept coefficient)

b = average change in Y given a one unit change in X, i.e (slope of X)

$(X)_i$  = value of X for observation i

1. In the data analysis tool select the regression and then click Ok.

## Select the Input Data Set for Y and X values

The screenshot shows the Excel interface with a data table and the Regression dialog box open. The data table has columns D and E, with headers 'x' and 'y' respectively. The data points are:

x	y
1.5	2.2
3.6	3.1
5.6	8.6
4.5	5.9
6.7	4.4

The Regression dialog box is titled 'Regression' and has the following settings:

- Input**
  - Input Y Range: \$E\$5:\$E\$10
  - Input X Range: \$D\$5:\$D\$10
  - Labels
  - Constant is Zero
  - Confidence Level: 95 %
- Output options**
  - Output Range: \$G\$5:\$O\$24
  - New Worksheet Ply:
  - New Workbook
- Residuals**
  - Residuals
  - Residual Plots
  - Standardized Residuals
  - Line Fit Plots
- Normal Probability**
  - Normal Probability Plots

Buttons: OK, Cancel, Help

2. After clicking the regression, a regression box appears.
3. Select the Inputs for Y range and X range.
4. Select the place where you want your output
5. Check the Labels and then click Ok

## Output

x	y	SUMMARY OUTPUT								
1.5	2.2									
3.6	3.1	<i>Regression Statistics</i>								
5.6	8.6	Multiple R	0.631253							
4.5	5.9	R Square	0.39848							
6.7	4.4	Adjusted R Square	0.197973							
		Standard Error	2.259047							
		Observations	5							
		<i>ANOVA</i>								
			<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
		Regression	1	10.14212	10.14212	1.987366	0.253402			
		Residual	3	15.30988	5.103295					
		Total	4	25.452						
			<i>Coefficient</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
		Intercept	1.329453	2.687341	0.494709	0.654767	-7.22287	9.881772	-7.22287	9.881772
		X	0.801495	0.568541	1.40974	0.253402	-1.00786	2.610846	-1.00786	2.610846

The output is given in the coefficients column in the last set of output

1.  $a = 1.329$  (Intercept Coefficient)

2.  $b = 0.801$  (Coefficient of X i.e. slope)

So our regression equation is  $Y = 1.329 + 0.801(X)$

Also in the regression statistics output gives the goodness of fit measure

3. Adjusted R square = 0.3984 which measures the fit

This means 39.84% of Y is determined by X