Statistics for a Linear Equation[1]

This analysis applies to statistics for a line with only *one variable* ($y = mx + b$).

The regular plot routine in Excel can generate a value of r^2, but not a confidence interval for the two coefficients (m and b). The easiest method to generate these statistics is to use the regression package in the Analysis ToolPak in Excel 2004, as follows:

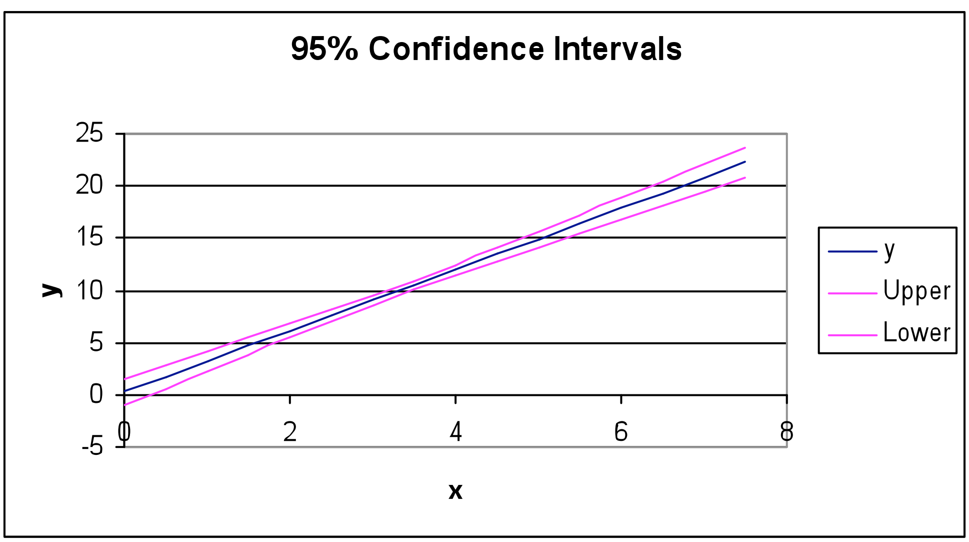
1. Under the Tools pull-down menu, select "Add-Ins."
2. Check the box next to "Analysis ToolPak."
3. Under the Tools pull-down menus, select "Data Analysis."
4. Click on "Regression."
5. Select range of x and y values, select confidence interval, and select some of the output options. A sample page is attached as an Excel file.

This procedure gives you the upper and lower confidence intervals on m and b. You may also use the DeLevie Macro Linest to obtain the same result.

For a line, a formula can be generated to calculate the confidence interval on y for any value of x. The value of $t_{\nu,\alpha/2}$ must be computed from a statistics table using the degrees of freedom ($\nu = n_{points} - n_{variables} - 1$) and the selected confidence interval ($\alpha = 0.05$ for 2-tailed 95% confidence interval). There is also an easy way to calculate $t_{\nu,\alpha/2}$ in Excel (=TINV(α,ν)). For 9 points, $\nu = 7$ and $t_{\nu,\alpha/2} = 2.365$ (i.e., =TINV(0.05,7)). The equation for y is then:

$$y_0 \pm t_{\nu,\alpha/2}\sqrt{\frac{s_{ey}^2}{n} + s_{em}^2 (x_0 - \bar{x})^2}$$

where y_0 is the value computed from $y_0 = mx_0 + b$. You can use this formula to generate lines above and below your linear fit that represent a 95% confidence interval. The value of s_{ey} is the standard error listed in the "Regression Statistics" table, and n is the total number of data points. You must compute the average value of x from your data ($\bar{x}$). The value of s_{em} is in the "Standard Error" for the slope coefficient. Note that this confidence interval line is a quadratic (i.e., non-linear), since you have less confidence as you move away from the mean x value in your data. The plot might look something like this:



[1] Source: http://www.et.byu.edu/groups/uolab/files/lecturenotes/ 2009.

The method for determining confidence intervals on coefficients also works well for *multiple regression* (i.e., $y = m_1x_1 + m_2x_2 + ... +b$), but is a bit more complicated. The following are excel directions:[2]

You will need to calculate the variance of an estimated point on the multiple regression. Suppose X is the "design matrix" (the array that you would pass to LINEST, augmented with a column of ones [unless you are not fitting a constant term]).

If b is the corresponding vector of coefficient estimates (a column, and in reverse order to the LINEST output), then =MMULT(X,b) gives the estimated multiple regression at your data points, i.e. the same output as =TREND(known_y's,known_x's,,const). The predicted value at a given point on the multiple regression would be =MMULT(v,b) where v is the row of X corresponding to the point (if it is in the data set), or is constructed similarly (if it is not in the data set).

You will need D which is calculated as
=MMULT(MMULT(v,MINVERSE(TRANSPOSE(X),X),TRANSPOSE(v))

The variance of an estimated point on the multiple regression is then =D*MSE, where MSE is =sey^2 and sey is one of the quantities output by =LINEST(known_y's,known_x's,const,TRUE). Similarly, the variance of a predicted future point that follows the same multiple regression is =(1+D)*MSE.

A 95% 2-sided confidence interval for a point on the multiple regression line is then
=MMULT(v,b) +/- SQRT(D*MSE)*TINV(0.05,df)

A 95% 2-sided prediction interval for a new point that follows the multiple regression is
=MMULT(v,b) +/- SQRT((1+D)*MSE)*TINV(0.05,df)

Some simplification may be possible given knowledge of the particular regression model that you want. For instance, with simple linear regression:
MSE reduces to STEYX(known_y's,known_x's)
=MMULT(v,b) reduces to =FORECAST(x,known_y's,known_x's)

or equivalently to
=INTERCEPT(known_y's,known_x's) +x*SLOPE(known_y's,known_x's)
and D reduces to
=1/COUNT(known_x's) +(x-AVERAGE(known_x's))^2/DEVSQ(known_x's)

You may compare these equations to the details on page 1.

---

[2] http://www.eggheadcafe.com/software/aspnet/29971722/confidence--prediction-i.aspx