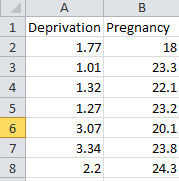
**Worksheet: Scatterplots, correlation and simple linear regression in EXCEL**

Example: Area level deprivation and pregnancy rates per 1,000 young women aged 15-17 years for Local Authorities in East Midlands

**Scatterplots in EXCEL**

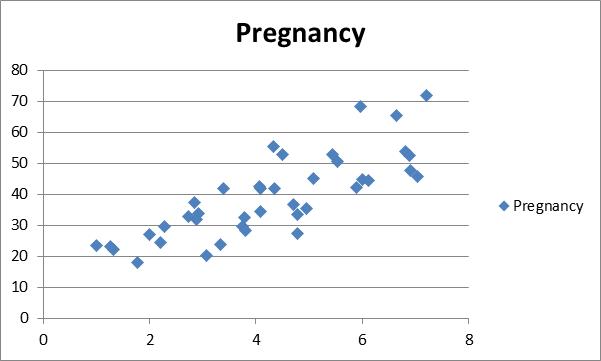
The two variables that you want to plot need to be in adjacent columns with the X (predictor variable) in the first column and the Y (outcome) variable in the second column:



Select the data in the two columns, on the INSERT tab, then Scatter, and the first figure in the list. This will produce a scatterplot. You can change the appearance of the scatterplot by double clicking on the figure and selecting the elements that you want to change:



And here’s the output:



**Correlation in EXCEL**

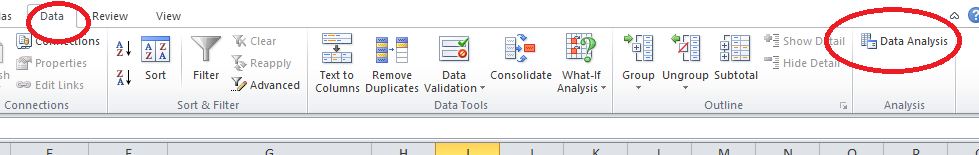
You can either use the Correlation formula if you only want to produce a single correlation between two variables:

*=CORREL(A, B)*

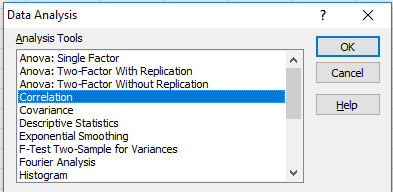
Where A= data range for first variable

B= data range for second variable

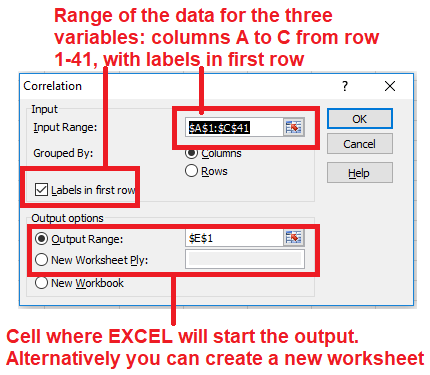
Or you can use the data analysis toolpak if you have several variables that you want to correlate at the same time:



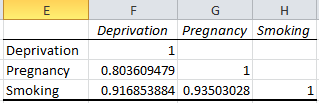
This will open up the Data Analysis dialogue box. Scroll down the list of options until you see Correlation. Select and click OK



And complete the options:



And here’s the output:

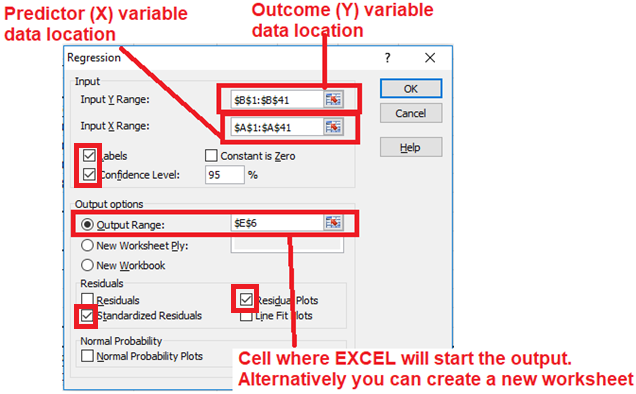


All three variables are strongly correlated: Deprivation and Teenage Pregnancy rate has a correlation of 0.80, Deprivation and Adult smoking rates is 0.92 and between Pregnancy and adult smoking rates 0.94

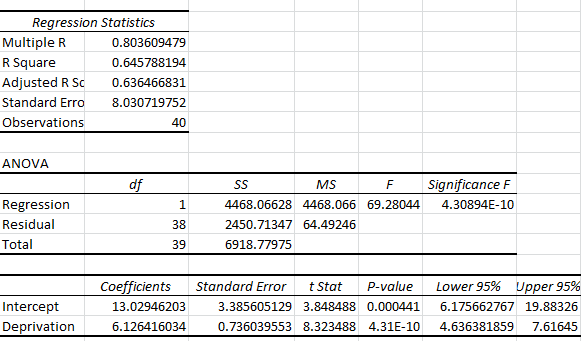
**Simple Linear Regression in EXCEL**

You can do a simple linear regression in EXCEL using the Data Analysis toolpak

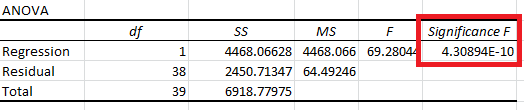




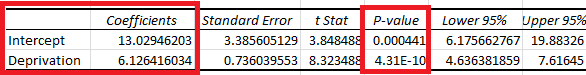
And here’s the output:



And now to make sense of the output:



The ANOVA table enables us to assess whether the overall model is significant. It doesn’t assess individual variables, but allows us to answer the question ‘Is this model better than no model at all?’. The p-value is labelled Significance F. For the current model it is very small (4.30894E-10 i.e. 0.000000000430894) indicating that the model is significantly better than the null model with no variables

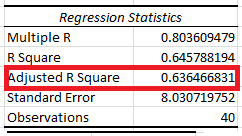


The table of coefficients enables us to assess whether the explanatory variable is significantly related to the outcome. If it is we can use it to construct a regression equation. For the above model, both the intercept and Deprivation coefficients are significant.

Generally we aren’t interested in the value of the Intercept nor whether it is significant, it is almost always retained in the model as it positions the line on the Y axis. Looking at Deprivation we can see that the coefficient is 6.12 and the p-value is very small (< 0.0001). This means that for every increase in Deprivation of one unit, the pregnancy rate increases by 6.12 pregnancies per 1,000 women aged 15-17 years

The regression equation is:

***Teenage pregnancy rate = 13.03 + 6.13 x Deprivation***



The key thing to look at in the regression statistics table is the Adjusted R Squared as this gives us an indication of the amount of variability in the data that is accounted for by the model. It is expressed as a proportion and can be rescaled as a percentage. Here the value is 0.636 or 63.6%.

The final thing to look at is the residuals: do they show any patterns when plotted against the predicted values and are they normally distributed. One of the options in the Regression dialogue box is for residual plots. You can also save the standardised residuals and plot these using a histogram to check that they are approximately normally distributed.