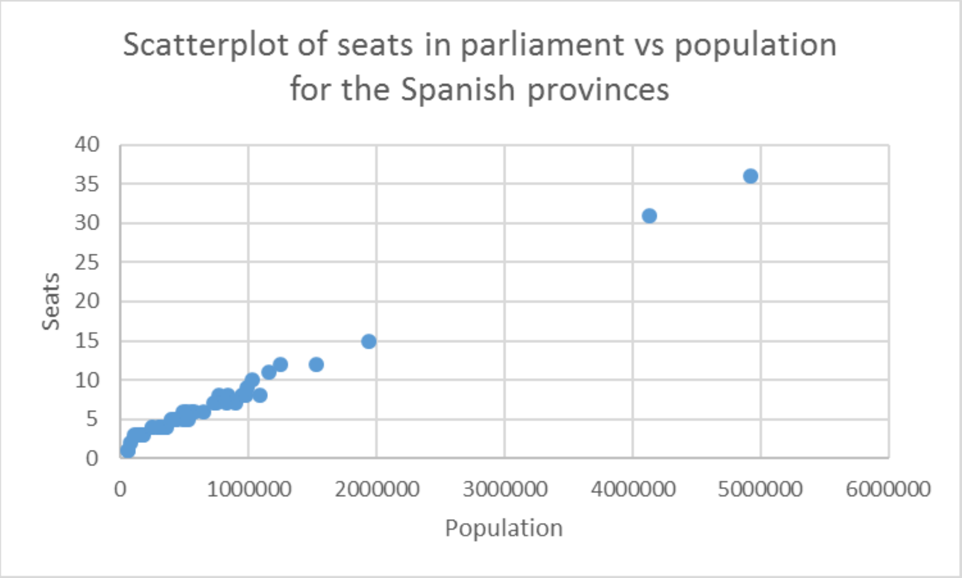
Practical Session 7

Covariance, correlation and regression

In this session, we shall see how to assess and model a linear relationship between two variables.

1. Drawing a scatterplot for two variables.

   In class we looked at data on seats in parliament and population. We can graph these data using the scatterplot Excel graphics option.



2. Calculating the covariance and correlation

   To calculate the covariance, we can use 3 functions in Excel: COVARIANCE.P, (COVARIANCE.P), COVARIANCE.S (COVARIANZA.M) and COVAR which is left over from earlier versions of Excel. As in previous sessions, **COVARIANCE.P** uses a divisor of N as we have done in class and **COVARIANCE.S** uses a divisor of N-1. If our x and y data are in cells **A1:A10** and **B1:B10** respectively, for example, then the syntax is:

   =COVARIANCE.P(A1:A10;B1:B10)

   For the Spanish parliamentary seats data, the value of the covariance is 5198072.04.

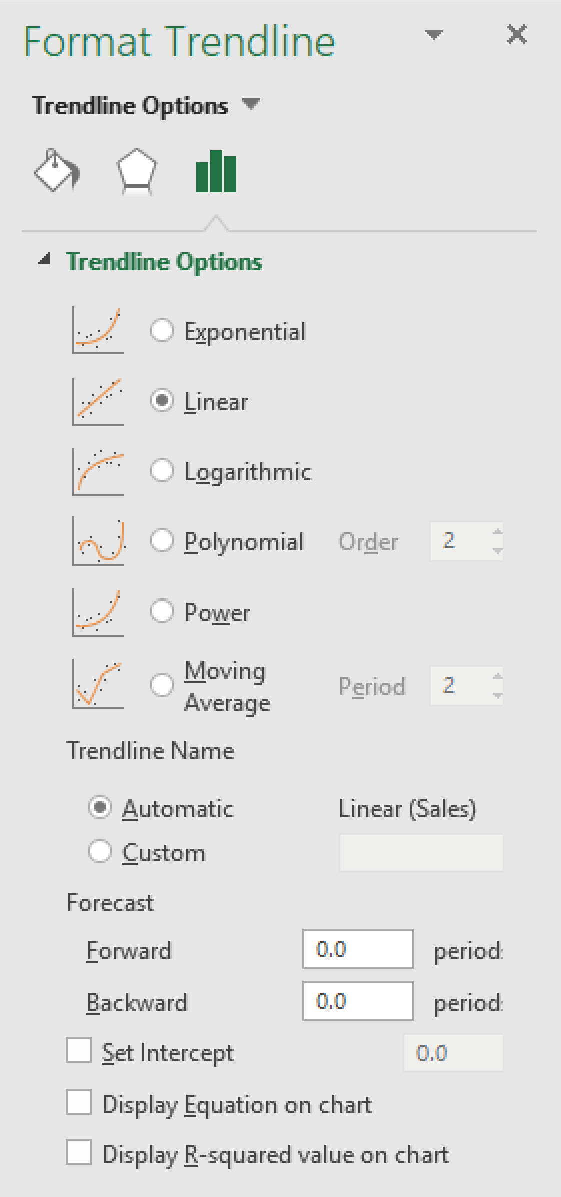
   You can see what happens when you change the scale of the data by dividing the population by a million for example. The covariance will change dramatically.

   To calculate the correlation, we can use the function CORREL (COEF.DE.CORREL). For this data, the correlation is approximately 0.996, very close to 1. You can check that when you rescale the population, the correlation does not change.

3. Adding a regression line to a scatterplot.

Excel has specific statistical functions for calculating the formula for a regression line as we will see below. However, if we only want to add the line to a scatterplot we already have, this is straightforward.

In the scatterplot, just right click on any data point and use the command **Add Trendline** (**Añadir linea de tendencia**) n the command box that opens up. This will give you various options as below:



Format Trendline

Trendline Options

Trendline Options

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Linear (Sales)

○ Custom

Forecast

Forward 0.0 period

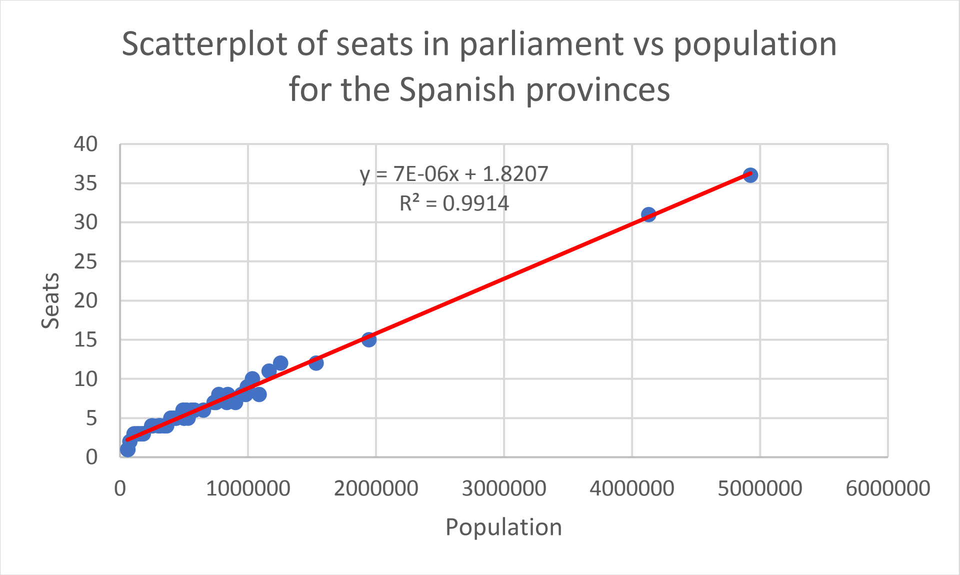
Backward 0.0 period

□ Set Intercept 0.0

□ Display Equation on chart

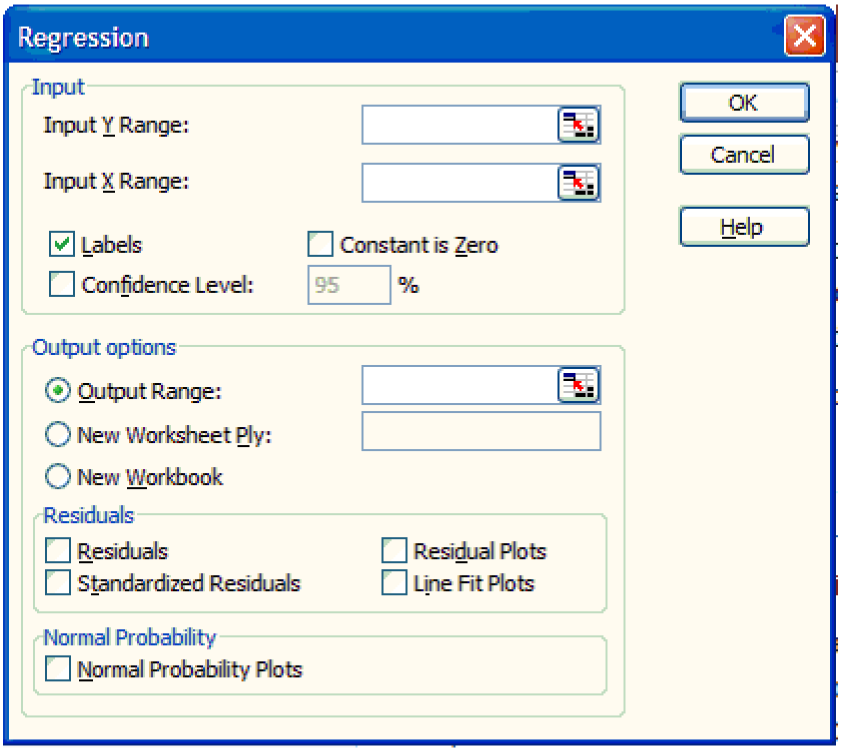
□ Display R-squared value on chart

The option we want to use is the default. Note that if you want to show the regression equation and the R^2 or coefficient of determination value on the plot, you can also tick these options as well.


Scatterplot of seats in parliament vs population for the Spanish provinces

4. Residual analysis

If you want to graph the residuals as well, then you need to do a full regression analysis in Excel. This can be done by clicking the **Data Analysis** option in the **Data** tab in Excel. Remember that in the lab computers, you will need to reinstall this via **File**, **Options**, **Add-Ins**. You can then scroll down to the Regression option.


Regression

Input

Input Y Range:

Input X Range:

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range:

○ New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Here you can insert the x and y ranges, mark **Labels** if you include the column names, mark **Constant is zero** if you want a regression through the origin (we did not talk about this in class, but it makes sense for our parliamentary seats example: zero population should be zero seats. If we want to see a plot of the residuals, we can also mark the **Residual Plots** option. The remaining available options are not so relevant for our course.