2. Practice – 2014

Empirical correlation coefficient, method of least squares, coefficient of determination, Wald method

1. Exercise

- Let's investigate if there exists a linear relationship between the alcoholics per 10000 capita and the students in Hungary!
- Using trend line fitting the best fit straight line has to be determined along with the coefficient of determination!

Solution:

We can answer this question with the empirical correlation coefficient (denoted by R):

$$R(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

- $-1 \leq R \leq 1$
- If $R \leq -0.8$ then the relationship is progressive and linear
- If $R \geq 0.8$ then the relationship is degressive and linear
- If $R \approx 0$ there is no linear relationship, but there could be a nonlinear relationship.
- The value of R doesn't say anything about a nonlinear relationship.

First let's plot the data in a scatter diagram:

- Select B6:C21
- go to **Insert/Charts/Scatter/Scatter with only Markers**
- We see that there is a degressive linear relationship between the alcoholics and students in Hungary.

Let's calculate R!

- The averages:
  - $\bar{x}$ = **B22=AVERAGE(B6:B21)** and pull to the right, OR
  - $\bar{y}$ =**C22=AVERAGE(B6:B21)**
- To calculate the sum-s first we need to make 5 new columns:
  - E5: (x-x_Ave)
  - F5: (y-y_Ave)
  - G5: (x-x_Ave)^2
  - H5: (y-y_Ave)^2
  - I5: (x-x_Ave)(y-y_Ave)
- Then calculate the values according to the column names:
  - **E6=B6-B$22** and pull down and to the right
  - **G6=F6^2** and pull down and to the right
  - **I6=F6*G6** and pull down
- For R we need the sum of the last 3 columns:
  - **G22=SUM(G6:G21)** and pull to the right two times
- Now we can calculate the value of R:
  - D25: Correl. Coeff.:

- **E25=I22/SQRT(H21)/SQRT(G21)**
- The excel function for the correlation coefficient:
  - D26: Excel Correl.:
  - **E26=Correl(B6:B21;C6:C21)**
- Note: D25=D26=-0.9654

Let's fit a straight trend line along the points on our diagram!

- Right click on one of the point on the diagram
- **Add trendline..**
- **Linear** trend line
- **Display Equation on chart**
- **Display R-squared value on chart**

2. Exercise

The measured points show the position of a sucktion pump (harmonic motion). The position of the head of the pump is described by the

$$s(t) = a * sin(10{,}68 * t) + b$$

function.
Calculate the value of the *a* and *b* parameters, and decide the quality of the fit using the coefficient of determination.

Solution:
The method of least squares can only be used if the independent data is exactly known, and only the dependent data has some error.

- First plot the points on a scatter diagram to see if there is indeed a sinusoidal relationship.
- For the method of leas squares we want the sum of the square distances to be minimal.
  - The distance between the $s(t)$ curve and the measurement s_i:
    $$e_i = a * sin(10{,}68 * t_i) + b - s_i$$
  - The sum of the square distances:
    $$F(a,b) = \sum_{i=1}^{n} (a * sin(10{,}68 * t_i) + b - s_i)^2$$
  - We are looking for the minimum place of F. It is well known that minimum place can only be where the partial derivatives are zero:
    $$\frac{\partial F}{\partial a} = 2\sum (\mathrm{a} \cdot \sin(10{,}68 \cdot \mathrm{t}_i) + b - s_i)\big(\sin(10{,}68 \cdot \mathrm{t}_i)\big) = 0$$
    $$\frac{\partial F}{\partial b} = 2\sum (\mathrm{a} \cdot \sin(10{,}68 \cdot \mathrm{t}_i) + b - s_i)(1) = 0$$
  - Rearranging the equations we get:
    $$a\sum sin^2(10{,}68 \cdot \mathrm{t}_i) + b\sum sin(10{,}68 \cdot \mathrm{t}_i) = \sum s_i \cdot \sin(10{,}68 \cdot \mathrm{t}_i)$$
    $$a\sum sin(10{,}68 \cdot \mathrm{t}_i) + b \cdot \mathrm{n} = \sum s_i$$
  - This is a linear algebraic equation system with two equations and two unknowns. In order to solve such a system in excel we first need to write it into a matrix form:
    $$\begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$$
    where the coefficient are the following:
    - $c_1 = \sum sin^2(10{,}68 \cdot \mathrm{t}_i)$ and $c_2 = c_3 = \sum sin(10{,}68 \cdot \mathrm{t}_i)$ and $c_4 = n$
    - $d_1 = \sum s_i \cdot \sin(10{,}68 \cdot \mathrm{t}_i)$ and $d_2 = \sum s_i$

- Calculate the coefficients:
  - Create three new columns for the sums:
    - **C10 : sin(10.68 t_i)**
    - **D10 : sin^2(10.68 t_i)**
    - **E10 : s_i* sin(10.68 t_i)**
  - Calculate the values according to the column names, then sum the columns up to have all the values for the matrix C and the vector d.
  - Put matrix C in cells M11:N12 and vector d in cells R11:R12
- To get the desired parameters we need the inverse matrix of C:

$$\begin{pmatrix} a \\ b \end{pmatrix} = C^{-1} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$$

- To calculate the inverse matrix of C
  - Select a 2x2 block for the inverse matrix (e.g. M15:N16)
  - **=MINVERSE(M11:N12)**
  - Instead of ENTER press **CTRL+SHIFT+ENTER** (only this way will be the inverse matrix a 2x2 matrix)
- To get the vector (a,b)
  - Select a 2x1 block (e.g. R15:R16)
  - **=MMULT(M15:N16;R11:R12)**
  - Instead of ENTER press **CTRL+SHIFT+ENTER**
- Now, we have the function that fits best the given points. Let's put this curve into our diagram to see that it is really a good fit.
  - Create to new columns (e.g. in I10:J111)

t	s
0	
0,02	
0,04	
0,06	
…	
2	

    In the first two cells of column t write 0 and 0.02 then pull it down until 2. We want to calculate the fitted curve in these points. In the first cell calculate a*sin(10.68*t) +b, put a dollar sign so the values of a and b would be fixed, then pull down this column.
  - Plot these t-s points in the diagram:
    - **Right click in the diagram**
    - **Select data..**
    - **Add**
    - Give our new data
  - Format the diagram:
    - **Right click on one of the new point in the diagram**
    - **Format data series..**
    - **Marker options -> None**
    - **Line color -> Automatic**
  - We can see that our curve fits quite well the measured data.
- How well our curve fit the data? The numerical answer to the question is the coefficient of determination:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_i (s_i - s(t_i))^2}{\frac{1}{n}\sum_i (s_i - \bar{s})^2}$$

- When our curve passes through the measurement points the numerator is zero, hence$R^2 = 1$, so when it is near to 1, we say that our curve fits well on the data.
- When or curve is the average, then $R^2 = 0$, so our curve is not good.
- Let's calculate the coefficient:
  - Calculate the curve in the given $t_i$times: **F11= R$15*SIN(10,68*A11) +R$16**
  - The square distances from the measurement points: **G11= (B11-F11)^2**
  - The numerator: **G32= AVERAGE(G11:G30)**
  - The denominator: **G33 =STDEV.P(B11:B30)**
  - The coefficient of determination: **G34: =1-G32/G33**
  - We see that the curve fits well.

3. Exercise

In this exercise we will use Wald method to fit a linear to some given data, which is clearly separable along at least one axis. The data shows the australian rabbit and fox populations. Let's check whether there is a linear connection between the two, and if there is, fit a line with Wald method!

Solution:
The Wald method is used when both the independent and dependent data has some error.

- First plot the number of foxes versus the rabbits on a scatter diagram. We see that there is a linear digressive relationship between the two sets of data.
- Also calculate the correlation coefficient to see if there really is a linear relationship: **A35=CORREL(A9:A32;B9:B32)**
- The idea of the Wald method is to separate the data into two disjoint sets, calculate their average point, and fit a line that goes through both of the average points.
- It is clear that the separation should be at around X=200.000. All the point witch have a smaller X coordinate than 200.000 will be in set 1, and the point with larger X coordinates will be in set 2.
  - **A36:Separator**
  - **B36=200000**
- Let's determine the points of the two sets. Create four new columns for X_1; Y_1; X_2; Y_2.
- To separate the points use the function IF:
  - **C9=IF($A9<200000;A9;"")**, and double click, and pull it to the right.
    (If the value in A9 is smaller than 200000, we write that value in C9, if not we write nothing in C9)
  - **E9= IF($A9>200000;A9;"")**, and double click, and pull it to the right.
- Calculate the averages in the bottom of the columns.
- The equation of a line can be written in the form $y = ax + b$. The parameters *a* and *b* can be determined from the averages:
  - The slope *a:* $a = \frac{\overline{Y_2} - \overline{Y_1}}{\overline{X_2} - \overline{X_1}}$
  - The constant b: $b = \overline{Y_1} - a\overline{X_1} = \overline{Y_2} - a\overline{X_2}$
- To place the line in the diagram first calculate the value of the fitted line in the X_i coordinates ($F_i = aX_i + b$.) in a new column. And place the points in the diagram the same way we did in the previous exercise.

- Place a linear trend line on the original points, the equation of the line should be seen. Excel fits the trend line with the linear least method, and it is clear from the equations that the two methods are not the same.