## Data gap filling through linear regression

One way of data completion is through linear regression.

With linear regressions a mathematical relation is defined between data of a base station and other stations of the form:

***Y***  ***C***  ***C***1 ***X***1  ***C***2 ***X***2  ***C***3 ***X***3  ***etc***. (2.1)

where

*Y* a series of values of the base station (dependent variable)

*Xi* a series of values of neighbouring station *i* (independent variable)

*C* the equation's constant

*Ci* the equation's coefficients

Multiple regression means that more than one neighbouring station (independent variable) is regarded. In case of a base station and one neighbouring station, the equation reduces to

***Y***  ***C***  ***C***1 ***X***1 (2.2)

The method is based on fitting the 'best' straight line through observations. In general there always will be a difference between a calculated value, *Y'* according to this line and a measured value. In linear regression this difference *(Y-Y')* is minimized using the method of 'least squares'.

The quality of the fit obtained can be investigated by calculating the coefficient of determination *ρ2*. The coefficient of determination is a number between *0* and *1*, where *1* represents perfect fit. A minimum value of *ρ2* should be applied giving the limit of a realistic

regression (minimum of *ρ2* should be around *0.5*). This is also a criterion for the selection of stations to be included in the regression analysis.

The square root of the coefficient of determination is the correlation coefficient *ρ* which value varies between *-1* and *+1.* The correlation coefficient *ρ* is related to the differences between measured and calculated values of the dependent variable *(Y-Y')* and defined as (see also appendix A5)

  ***(Y***  ***- Y***

***=***

******   ***(Y - Y***



***2*** ***0.5***



***)***

***)2*** 



Among the statistical applications of the Excel spreadsheet program is the option of multiple linear regression. This provides directly the coefficient of determination *ρ2* (in the spreadsheet known as *R Square*), as well as the coefficients and the *Intercept* of the regression equation.

##### Assignment

Retrieve the file ‘Data for gapfilling’ with the annual data of four precipitation stations. The (multiple) regression analysis will be performed between P425 (dependent variable) and P119, P5 and P6 as neighbouring stations (independent variables). A worked out example as a reference for the exercise is provided in table 1. The correlation analysis will be carried out for only those years in which all 4 stations have data.

Steps:

1. Open the file and go to the ‘Data’ sheet and select ‘Data Analysis’[[1]](#footnote-1)
2. Select ‘Regression’
3. Select the input Y range (the station you want to predict) and the input X range (can be multiple stations)
4. Select OK to get the results

**Table 1 Example of (multiple) linear regression between 2 and 4 stations**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| January 58/59 | P425 175.3 | P119 123.7 | P5 162.9 | P6 79.5 | **SUMMARY OUTPUT P425 - P6** |
| 59/60 | 79.5 | 63.5 | 76.4 | 84.3 | *Regression Statistics* |
| 60/61 | 56.0 | 49.1 | 110.0 | 84.5 | Multiple R 0.8564737 |
| 61/62 | 142.4 | 118.1 | 93.1 | 188.6 | R Square 0.7335473 |
| 62/63 | 95.7 | 115.8 | 111.8 | 84.7 | Adjusted R Square 0.7202246 |
| 63/64 | 249.0 | 173.2 | 210.3 | 215.5 | Standard Error 60.297187 |
| 64/65 | 12.3 | 56.2 | 14.3 | 40.4 | Observations 22 |
| 65/66 | 546.8 | 672.2 | 625.1 | 587.6 |  |
| 66/67 | 76.6 | 190.1 | 48.5 | 162.1 | *Coefficients* |
| 67/68 | 121.5 | 113.7 | 92.0 | 71.4 | Intercept 10.380164 |
| 68/69 | 157.7 | 188.7 | 125.5 | 111.4 | X Variable 1 0.7697746 |
| 69/70 | 4.7 | 13.0 | 98.4 | 9.6 |  |
| 70/71 | 51.8 | 98.6 | 87.9 | 53.1 | **SUMMARY OUTPUT** |
| 71/72 | 218.0 | 320.6 | 156.8 | 210.4 | **P425 and P119, P5, P6** |
| 72/73 | 56.7 | 57.2 | 79.1 | 63.3 | *Regression Statistics* |
| 73/74 | 108.6 | 209.6 | 151.7 | 299.5 | Multiple R 0.9044524 |
| 74/75 | 81.3 | 183.8 | 138.7 | 232.3 | R Square 0.8180341 |
| 75/76 | 210.2 | 416.7 | 311.4 | 275.0 | Adjusted R Square 0.7877065 |
| 76/77 | 44.3 | 150.9 | 77.0 | 115.5 | Standard Error 52.524354 |
| 77/78 | 122.0 | 354.0 | 305.6 | 202.9 | Observations 22 |
| 78/79 | 122.5 | 171.2 | 60.5 | 129.0 |  |
| 79/80 | 62.2 | 157.7 | 52.9 | 33.8 | *Coefficients* |
|  |  |  |  |  | Intercept 3.821143 |
|  |  |  |  |  | X Variable 1 0.0918169X Variable 2 0.4495039X Variable 3 0.2729471 |
|  |  |  |  |  |  |

1. If ‘Data analysis’ is not available, then go to ‘File’, ‘Options’, ‘Add-ins’, manage Excel add-ins, select ‘Go’, click box for ‘Analysis ToolPak- VBA’ [↑](#footnote-ref-1)