Application Example 19

(Linear Regression)

EL NINO AND FLOW OF THE NILE RIVER

The El Nino phenomenon has significant impacts on the global distribution of weather and on the flow of most tropical rivers. Consider the data in Table 1 on the anomaly of sea surface temperature in the Pacific Ocean in degree Centigrade (an index of El Nino) and the annual flow of the Nile river in cubic kilometers for the period 1972 – 1995.

Year	Temperature	Flow
1972	-1.1	113
1973	-0.8	83
1974	-1.3	121
1975	-1.3	111
1976	0.2	96
1977	1.7	70
1978	-0.5	134
1979	-0.7	124
1980	0.31	97
1981	-0.2	92
1982	-0.7	89
1983	-0.4	105
1984	0.6	87
1985	0.5	93
1986	-1.1	91
1987	-0.3	113
1988	1.3	68
1989	-1.2	92
1990	-0.7	110
1991	-0.2	102
1992	-1.2	124
1993	-0.9	99
1994	-0.5	126
1995	0.7	117

Table 1. Anomaly of sea surface temperature in the Pacific Ocean and Annual Flow of the Nile River

A question of great theoretical and practical interest is whether the El Nino anomaly affects the atmospheric circulation over the African Continent. One way to check this hypothesis is to determine whether there is a statistically significant relationship between the temperature anomaly in the Pacific Ocean, X, and the annual flow of the Nile River, Y.

In what follows, we first fit a simple linear regression between X and Y and then make a formal hypothesis test of the significance of the regression (the regression is said to be "significant" if X affects the mean value of Y, hence if the slope of the linear relationship between X and Y is nonzero). At the end we make some observation on the appropriateness of the assumed linear model.

19.1 Fitting a Simple Linear Regression Model

As a first task, we estimate the linear regression between the two variables (temperature and flow), and find the coefficient of determination R^2.

We assume a linear relation between temperature X and flow Y, of the type

$$Y = \beta_o + \beta_1 X + \varepsilon$$

where ε is assumed to have normal distribution with zero mean and unknown variance σ^2. This means that for each observation (X_i, Y_i), i = 1, …, n, the flow Y_i is assumed to satisfy:

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$$

where the variables ε_i are iid with the distribution of ε.

We look for estimators $(\hat{\beta}_o, \hat{\beta}_1)$ such that the sum of the squared of the residuals $\sum_i (\hat{\beta}_o + \hat{\beta}_1 X_i - Y_i)^2$ is minimum. The solution is:

$$\hat{\beta}_o = \overline{Y} - \hat{\beta}_1 \overline{X} = \overline{Y} - \frac{S_{XY}}{S_{XX}} \overline{X} \, , \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

where

$$\overline{X} = \frac{1}{n} \sum_i X_i \;\; \overline{Y} = \frac{1}{n} \sum_i Y_i \, , \qquad S_{XX} = \sum_i (X_i - \overline{X})^2 \, , \qquad S_{XY} = \sum_i (X_i - \overline{X})(Y_i - \overline{Y})$$

In our case, $\overline{X} =$ -0.32, $\overline{Y} = 102.38$, $S_{XX} = 15.72$, $S_{XY} =$ -176.73

So that $\hat{\beta}_o = 98.73, \quad \hat{\beta}_1 = \text{-}11.24$

Hence, the best fitting line is Y = 98.73 – 11.24X. This line is shown in Figure 1, together with the original data.

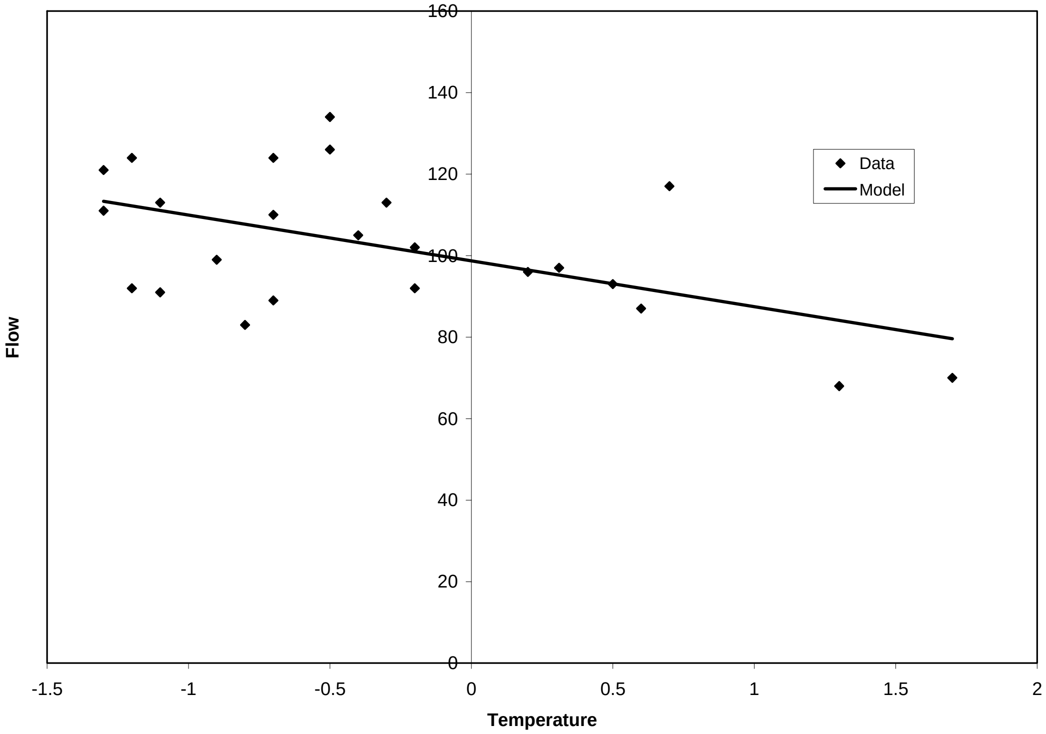

Figure 1. Temperature – flow data and fitted least - squares regression

Recall that the coefficient of determination R^2 is given by:

$$R^2 = \frac{SS_R}{S_{YY}} = 1 - \frac{SS_e}{S_{YY}}$$

where

$$S_{YY} = \sum_i (Y_i - \overline{Y})^2 \text{ = total sum of squares}$$

$$SS_R = \sum_i (\hat{\beta}_o + \hat{\beta}_1 X_i - \overline{Y})^2 \text{ = regression sum of squares}$$

$$SS_e = \sum_i (\hat{\beta}_o + \hat{\beta}_1 X_i - Y_i)^2 \text{ = residual sum of squares}$$

Also, the standard error of the regression, $\hat{\sigma}$, is given by $\hat{\sigma} = \sqrt{SS_e/(n-2)}$

In our case, SS_e = 4966.49, S_{YY} = 6953.63, SS_R = 1987.14 and n = 24. Therefore, $\hat{\sigma}$ =15.02 and $R^2 = 0.29$.

19.2 Testing the Statistical Significance of the Regression

Next we examine the regression significance (i.e. is the slope $\hat{\beta}_1$ significantly different from 0)?

To address this question, we perform the following hypothesis test:

$H_o : \hat{\beta}_1 = 0$
$H_1 : \hat{\beta}_1 \neq 0$

Statistic:

$$T = \frac{\hat{\beta}_1}{\sqrt{MS_e / S_{XX}}}$$

Rule: Accept if where for a test at significance level , c must satisfy

$$P[-c \le T \le c \,|\, H_o] = 1-\alpha$$

Result: $c = t_{\frac{\alpha}{2}, n-2}$

For $\alpha = 5\%$ and n=24, $t_{0.025,22} = 2.41$. Since 2.97 > 2.41, we reject H_o at the 5% significance level and conclude that the regression is significant, i.e. that there is a statistically significant dependence of river flow on temperature.

19.3 Comments on the Appropriateness of the Modeling Assumptions

We have concluded from the previous analysis that the regression of Y against X is statistically significant and therefore that anomalies in the surface temperature of the Pacific Ocean influence the rainfall regime over the Nile River Basin. This conclusion rests on our modeling assumptions for how Y is related to X.

Inspection of Figure 1 suggests that the assumption of a liner relation between Y and X may not be appropriate. Over the range of negative temperature anomalies, the data are dispersed around a nearly constant (or even slightly upwards sloping) mean, whereas for positive X, the annual flow Y seems to decay rather markedly. This different behavior of the data for positive and negative temperature anomalies would suggest a different (perhaps quadratic or piecewise linear) regression model. Such models can be handled through multiple linear regression. If a nonlinear model is used, then one should be forced into looking at the issue of why X influences Y only in “warm” years and does not in “cold” years.

Another observation is that the conclusion that the regression of Y against X is significant relies mostly on the 2 data points with largest positive ocean temperature (and lowest Nile River flow). If these points are removed, the regression is no longer significant. This sensitivity to a few data points should alert one not to firmly conclude that dependence of Y on X is real until these extreme data points are closely examined and other possible explanations are discarded.