# INFLUENCE CURVES FOR THE CORRELATION COEFFICIENT

SEPPO MUSTONEN

Abstract. An update formula for the correlation coefficient is derived. Using this formula it is shown how influence curves are plotted in the Survo system. This paper is essentially based on notes of the author written in late 1970's.

## 1. Update formula for the correlation coefficient

From a data set of two variables $x, y$ and $n$ observations

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

following statistics have been computed

$$\overline{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y}(n) = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$s_x^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i - \overline{x}(n)]^2, \qquad s_y^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}[y_i - \overline{y}(n)]^2,$$

$$r(n) = \frac{s_{xy}(n)}{s_x(n)s_y(n)}$$

where

$$s_{xy}(n) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i - \overline{x}(n)][y_i - \overline{y}(n)].$$

If a new observation $(x, y) = (x_{n+1}, y_{n+1})$ is obtained then the updated statistics $\overline{x}(n+1), \overline{y}(n+1), s_x^2(n+1), s_y^2(n+1), r(n+1)$ will be derived as follows.
We get immediately updates for means

$$\overline{x}(n+1) = \frac{1}{n+1}\left(\sum_{i=1}^{n} x_i + x\right) = \frac{1}{n+1}[n\overline{x}(n) + x]$$

and

$$\overline{y}(n+1) = \frac{1}{n+1}[n\overline{y}(n) + y].$$

*Date*: May 7, 2005.

The updated variance of $x$ is

$$\begin{aligned}
s_x^2(n+1) &= \frac{1}{n}\left(\sum_{i=1}^{n}[x_i-\overline{x}(n+1)]^2+[x-\overline{x}(n+1)]^2\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}[x_i-\overline{x}(n)+\overline{x}(n)-\overline{x}(n+1)]^2+\frac{1}{n}[x-\overline{x}(n+1)]^2\\
&= \frac{1}{n}\sum_{i=1}^{n}[x_i-\overline{x}(n)]^2+[\overline{x}(n)-\overline{x}(n+1)]^2+\frac{1}{n}[x-\overline{x}(n+1)]^2\\
&= \frac{n-1}{n}s_x^2(n)+\left(\frac{x-\overline{x}(n)}{n+1}\right)^2+n\left(\frac{x-\overline{x}(n)}{n+1}\right)^2\\
&= \frac{n-1}{n}s_x^2(n)+\frac{1}{n+1}[x-\overline{x}(n)]^2.
\end{aligned}$$

Similarly the variance of $y$ is

$$s_y^2(n+1)=\frac{n-1}{n}s_y^2(n)+\frac{1}{n+1}[y-\overline{y}(n)]^2.$$

and the updated covariance is

$$\begin{aligned}
s_{xy}(n+1) &= \frac{n-1}{n}s_{xy}(n)+\frac{1}{n+1}[x-\overline{x}(n)][y-\overline{y}(n)]\\
&= \frac{n-1}{n}r_(n)s_x(n)s_y(n)+\frac{1}{n+1}[x-\overline{x}(n)][y-\overline{y}(n)].
\end{aligned}$$

From these results we get an update formula for $r(n+1)$

$$\begin{aligned}
r(n+1) &= \frac{s_{xy}(n+1)}{s_x(n+1)s_y(n+1)}\\
&= \frac{\frac{n-1}{n}r_(n)s_x(n)s_y(n)+\frac{1}{n+1}[x-\overline{x}(n)][y-\overline{y}(n)]}{\sqrt{\left[\frac{n-1}{n}s_x^2(n)+\frac{1}{n+1}[x-\overline{x}(n)]^2\right]\left[\frac{n-1}{n}s_y^2(n)+\frac{1}{n+1}[y-\overline{y}(n)]^2\right]}}.
\end{aligned}$$

By substitutions

$$u=\sqrt{\frac{n}{n^2-1}}\left(\frac{x-\overline{x}(n)}{s_x(n)}\right),\qquad v=\sqrt{\frac{n}{n^2-1}}\left(\frac{y-\overline{y}(n)}{s_y(n)}\right)\tag{1}$$

$r(n+1)$ is simplified to the form

$$r(n+1)=\frac{r(n)+uv}{\sqrt{(1+u^2)(1+v^2)}}.\tag{2}$$

## 2. Influence curves

Update formulas (1),(2) are applied in an example made in the Survo system. The work scheme below is intended for plotting of scatter diagrams with appropriate contour curves describing robustness of the correlation coefficient. Actually these influence curves will appear as contours of a raster image of the influence function $|r(n+1)-r(n)|$ given as $z(x, y)$ on lines 14-17 thus telling how much the correlation coefficient increases or decreases when a new observation $(x, y)$ is obtained. The final graph is produced by a series of Survo operations (CORR, PLOT scatter diagram, PLOT contours, and PRINT).

The CORR DECA,3 command computes the means, standard deviations, and correlation coefficients of active variables in the data set DECA (48 best athletes in

```
--------------------------------------------------------------------
 1 *
 2 *CORR DECA,3 / VARS=Height,Weight
 3 *Means, std.devs and correlations of DECA  N=48
 4 *Variable  Mean        Std.dev.
 5 *Height    186.9583    5.090493
 6 *Weight    85.56250    6.847600
 7 *Correlations:
 8 *            Height  Weight
 9 * Height       1.0000  0.8522
10 * Weight       0.8522  1.0000
11 *........................................................................
12 *r=0.85 mx=186.96 my=85.56 sx=5.09 sy=6.85 n=48
13 *HEADER=Influence_curves_for_the_correlation_coefficient
14 *PLOT z(x,y)=abs(r*(1-w)+u*v)/w
15 *  u=sqrt(n/(n*n-1))*(x-mx)/sx
16 *  v=sqrt(n/(n*n-1))*(y-my)/sy
17 *  w=sqrt((1+u*u)*(1+v*v))
18 *TYPE=CONTOUR  ZSCALING=20,0     (1/0.05=20)
19 *              SCREEN=NEG
20 *XSCALE=150(10)220 YSCALE=40(10)130 SIZE=1164,1164
21 *x=150,220,0.125 y=40,130,0.125
22 *DEVICE=PS,INF.PS
23 *........................................................................
24 *PLOT DECA,Height,Weight
25 *XSCALE=150(10)220 YSCALE=40(10)130 SIZE=1164,1164
26 *HEADER=
27 *DEVICE=PS,DECA.PS
28 *........................................................................
29 *PRINT CUR+1,END TO INF2.PS
30 % 1200
31 - picture INF.PS
32 - picture DECA.PS
33 E
34 */GS-PDF INF2.PS
--------------------------------------------------------------------
```

Decathlon in 1973) and displays the results from the line 3 onwards. Thus, in this case, lines 3-10 are output from the `CORR` operation. Active variables `Height` and `Weight` are now set by the specification `VARS=Height,Weight` (typed by the user on the line 3 as a comment).

The user has copied the basic statistics obtained by `CORR` in a rounded form to the line 12. The `PLOT` scheme needed for making a contour plot of the influence surface is located on lines 12-22. The actual `PLOT` command to be activated (on the line 14) plots a function `z(x,y)` of two variables `x`,`y` as a contour plot (specified by `TYPE=CONTOUR` on the line 18). The expression defining `z(x,y)` depends on the current value of the correlation coefficient `r` and on three auxiliary functions `u`,`v`, and `w` which in turn depend on parameters `n`,`mx`,`my`,`sx`, and `sy`. These functions are defined as specifications on lines 15-17.

According to (1),(2) the function `z(x,y)` gives the change in the value of correlation coefficient `r` when a new observation `x`,`y` is obtained.

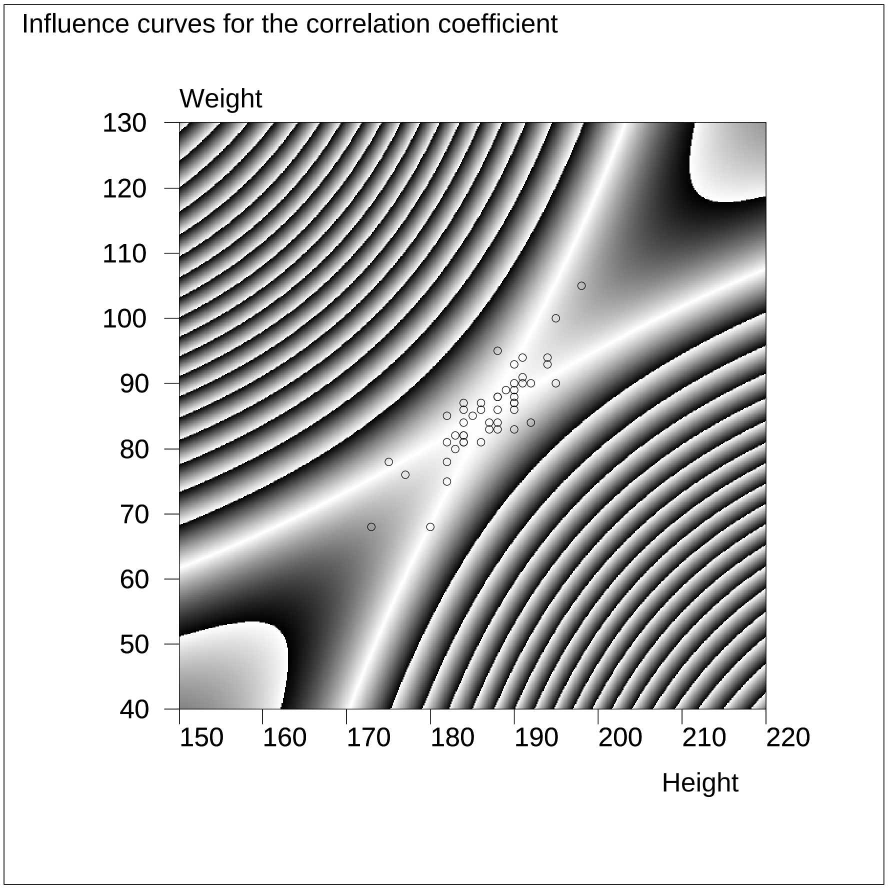

FIGURE 1. For example, a new observation $x = 190, y = 40$ would decrease the original $r$ from 0.85 by 6 × 0.05 to 0.55.

When making the raster image, the values of the function `z` are mapped continuously to various shades of gray in such a way that 0 corresponds to 'black' and 1 corresponds to 'white'.

If the function values exceed 1 the shading is selected 'modulo' 1. In this case, the original function values are multiplied by 20 (by `SCALING=20,0`) which gives a complete cycle of shadings when the function value changes by $1/20 = 0.05$. Thus, the final graph depicts contours of r with increments of 0.05. The `SCREEN=NEG` specification (on the line 19) simply reverses shadings.

Some plotting parameters, regulating ranges of variables and the size of the graph are given on lines 20-21. The last parameters (0.125) of the range specifications on the line 21 tell sizes of the plotting steps i.e. the accuracy of the raster image.

`DEVICE=PS,INF.PS` (on the line 22) implies the graph to be produced as a PostScript picture and saved in file `INF.PS`.

The simpler `PLOT` scheme on lines 24-27 makes a scatter plot of variables `Height` and `Weight` in the data set `DECA` using the same plotting specifications and saves the picture as a PostScript file `DECA.PS`.

Finally, the `PRINT` operation processes lines 30-32 and produces a combined PostScript file `INF2.PS` which is converted into PDF format and displayed by `/GS-PDF INF2.PS`.

## 3. Calculations

```
__________________________________________________________________
36 *ms()=MAT_MSN.M() / abbreviated notation            ACCURACY=3
37 *mx=ms(Height,mean) sx=ms(Height,stddev)
38 *my=ms(Weight,mean) sy=ms(Weight,stddev)
39 *n=ms(Height,N) / n appears in the 3rd column (N) of matrix MSN.M
40 *r=MAT_CORR.M(Height,Weight)
41 *
42 *  r(X,Y):=(r+u(X)*v(Y))/sqrt((1+u(X)*u(X))*(1+v(Y)*v(Y)))
43 *  u(X):=sqrt(n/(n*n-1))*(X-mx)/sx
44 *  v(Y):=sqrt(n/(n*n-1))*(Y-my)/sy
45 *
46 *r(mx,my)=0.852   r=0.852
47 *r(190,40)=0.553
48 *r(210,120)=0.899
49 *r(1000,1000)=1  r(-1000,1000)=-0.997
50 *r(10000,my)=0.003 r(mx,10000)=0.004
__________________________________________________________________
```

Since the `CORR` operation saves correlations in a Survo matrix file `CORR.M` and means, standard deviations, and the sample size in `MSN.M`, formulas (1),(2) can also be readily employed in editorial computing of Survo as shown above.
In these computations SURVO MM ver.2.26+ is required.

The current version of this paper can be downloded from
`http://www.survo.fi/papers/corrcurves.pdf`

Department of Mathematics and Statistics, University of Helsinki
*E-mail address*: `seppo.mustonen@helsinki.fi`