RESEARCH REPORT

SURVO 76 EDITOR

Estimation of regression models

BY
Seppo Mustonen

No. 29 September 1981

DEPARTMENT OF STATISTICS
UNIVERSITY OF HELSINKI
SF 00100 HELSINKI 10 FINLAND

ISBN 951-45-2401-2
ISSN 0357-9778

1. INTRODUCTION

SURVO 76 EDITOR is an extension of the interactive statistical system SURVO 76 and permits various text processing and data handling operations in editorial mode. All the data, text and operations are represented in an edit field a part of which is always visible on the screen when EDITOR is in use. The edit field is like a note book for the user and it is easily controlled by the special function keys (edit keys). This editorial approach to statistical data processing is described in Mustonen 1980,1981.

In this report a new operation ESTIMATE of SURVO 76 EDITOR will be introduced. ESTIMATE can be used for estimating parameters of linear and non-linear regression models and more generally for computing maximum likelihood estimates of user-defined statistical distributions.

ESTIMATE allows the statistical model to be expressed in the edit field in normal notation, variables and parameters having alphanumeric names selected by the user. ESTIMATE is capable of interpreting the model and it also forms analytically the first and second partial derivatives of the model function with respect to the parameters to be estimated. All this information, which is necessary in model identification and in the estimation process, will be transformed by EDITOR into BASIC subroutines working subsequently in connection with the main program. Thus this approach means no loss in the computing efficiency.

The automatic capability of utilizing formal derivatives has important consequences in statistical computing. In this connection, for instance, the program is able to recognize whether the model is linear with respect to the parameters (the second derivatives vanish in this case) and thus it may select an optimal computational approach.

2. ANALYTICAL DERIVATIVES

The procedure of forming analytical derivatives is recursive. Each step in this procedure consists of splitting the function to be analyzed into two parts like sum, difference, product or ratio of two functions and then applying basic derivation rules to the partition obtained. For example to form the derivative of f(x)=(x+a)*log(x↑2) with respect to x, the function f(x) is interpreted as f(x)=r(x)*s(x) where r(x)=x+a and s(x)=log(x↑2). In order to apply the derivation rule for a product, the derivatives of r(x) and s(x) are needed. The derivation algorithm employs itself for evaluating these derivatives. In this case r(x) is interpreted as sum of x and a and its derivative is attained after forming the derivatives of x and a. Similarly the derivative of s(x)=log(x↑2) is obtained according to the formula D(log(g(x)))=1/g(x)*D(g(x)) thus requiring derivation of g(x)=x↑2 etc.

The formal derivation algorithm is automatically employed by the ESTIMATE operation. However, when the user likes to form analytical derivatives in the edit field, this is possible by using another oper-

ation DER. For example, to find the derivative of f(x)=(x+a)*log(x↑2) with respect to x we type

```
DER (x+a)*log(x↑2) x
```

on any empty line in the edit field and after activation of this line by pressing RETURN(EXEC) the result will appear on the next two lines as follows:

```
DER (x+a)*log(x↑2) x
 Derivative of (x+a)*log(x↑2) with respect to x is
    log(x↑2)+(x+a)*2*x/x↑2
```

Observe that the resulting expression is not necessarily in the most simplified form, but usually the difference between the form given by our algorithm and the most reduced one is insignificant in practice.

3. 'ESTIMATE' OPERATION

The use of ESTIMATE in standard applications is best described through some simple examples. In the following display a typical regression model and set of data to be processed are presented in the form required for ESTIMATE.

Disp.1

```
    1  SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 1  *
 2  *DATA COUNTRIES,A,B,C
 3  C              Coffee  Tea  Beer   Wine Spirits
 4  A Finland        12.5 0.15  54.7   7.6  2.7
 5  * Sweden         12.9 0.30  58.3   7.9  2.9
 6  * Danmark        11.8 0.41 113.9  10.4  1.7
 7  * Norway          9.4 0.19  43.5   3.1  1.8
 8  * France          5.2 0.10  44.5 104.3  2.5
 9  * Ireland         0.2 3.73 124.5   3.8  1.9
10  * Italy           3.6 0.06  13.6 106.6  2.0
11  * Holland         9.2 0.58  75.5   9.7  2.7
12  * Portugal        2.2 0.03  27.5  89.3  0.9
13  * Switzerland     9.1 0.25  73.5  44.9  2.1
14  * Spain           2.5 0.03  43.6  73.2  2.7
15  B England         1.8 3.49 113.7   5.1  1.4
16  *
17  *MODEL BEER1
18  *log(Beer)=constant+coeff*log(Tea)
19  *
20  *ESTIMATE COUNTRIES,BEER1,21_
21  *
22  *
23  *
```

When the ESTIMATE operation on line 20 is activated we shall have the results displayed from the line 21 onwards in the form:

```
Disp.2
 20    *ESTIMATE COUNTRIES,BEER1,21
 21    * constant=4.488964       (0.1565749)
 22    * coeff=0.3276288        (0.0752709)
 23    * RSS=1.553654  R↑2=0.6545
```

In this display we have the estimated parameters and their standard errors (in brackets), the residual sum of squares (RSS) and the square of the multiple correlation coefficient (R↑2).

Let us now describe what actually happened after ESTIMATE on line 20 was activated. In this operation we have three parameters: the name of the data set (COUNTRIES), the name of the model (BEER1) and the first line for the results (21).

The data set considered has to be defined by a DATA specification which stands here on line 2. Observe that it is possible to use symbolic labels (a character placed in the control column) for the line numbers referred to in the editing operations. In this case the observation values are located on lines from A to B and the labels of the variables (columns) on line C. Note also that each observation must be preceded by a contiguous alphanumeric string (name of the observation).

The model to be estimated has to be defined by a MODEL specification. In this case the model BEER1 is defined on lines 17-18, the line 18 containing the model in the form <regressand>=<model function>. The model function is written according to the normal BASIC notation for algebraic expressions, but the variables (regressors) are notated by their labels in the DATA specification and the parameters by any other names. Hence, when ESTIMATE tries to analyze the model, it interpretes as parameters all words which are not recognized as variables. After the interpretation the model function is converted into a standard form, which in this case is

```
A(1)+A(2)*LOG(X(4))
```

and where A(1),A(2),... stand for parameters to be estimated and X(1),X(2),... are variables of the data set in the order they appear in the data matrix.

The regressand, here log(Beer), may also be a function and has to be written according to the same rules as the model function. It is interpreted and converted in a similar way. (Also the regressand may include parameters to be estimated; see 7 and 12.)

After the model has been analyzed its first and second partial derivatives with respect to the parameters will automatically be evaluated and then the model function, the regressand and these derivatives are presented as BASIC subroutines working in connection with the main program.

In the linear case, where the second derivatives vanish, the routine for these derivatives is automatically omitted.

4. COMPUTATIONAL METHODS

The main estimation method of ESTIMATE is the ordinary least squares (OLS) method. (For alternatives see 8.)

The iterative numerical algorithm needed for minimizing the residual sum of squares may be selected by the user by an extra specification METHOD typed in the edit field. At the moment we have three alternatives:

METHOD=N	Newton-Raphson	(see e.g. Walsh, 1975 p.108)
METHOD=D	Davidon-Fletcher-Powell	(see e.g. Walsh, 1975 p.110)
METHOD=H	Hooke-Jeeves	(see e.g. Walsh, 1975 p. 76)

If the METHOD specification is missing (as in our first example), ESTIMATE selects the computational algorithm according to the type of the model.

The Newton-Raphson method finds the optimum for a quadratic objective function (i.e. for a linear model with respect to parameters) in one iteration round which corresponds exactly to the conventional procedure of solving the linear normal equations. Hence, in case of a linear model METHOD=N is always default.

In other cases the default is METHOD=D, since although the Newton-Raphson method is the most efficient, when it works, it is unreliable in more complicated models and especially when the initial values for the parameters are poor. The Davidon-Fletcher-Powell (variable metric) method seems to be one of the best numerical procedures for general unconstrained optimization problems and it may be noticed that in case of a linear model the result is reached after a finite number of iteration steps. In fact the number of iterations required equals to the number of parameters to be estimated.

The simple but ingenious direct search method by Hooke and Jeeves (selected by METHOD=H) is here meant primarly for improving the initial estimates and for very irregular models (e.g. when the model function is not differentiable).

5. INITIAL ESTIMATES

The initial values for the parameters to be estimated are not needed at all when dealing with linear models. In non-linear cases, however, good approximations for the final estimates are always desirable.

The default for each initial value is always 0, but the user can enter his own suggestions simply by typing <parameter name)=<initial value) in the edit field. Since the final results are displayed in the same form (see Disp.2), results from a previous estimation may also be directly employed as initial values for the next one.

Thus when developing a model e.g. from a simple to more a complicated one the user may simply use the present results as starting values for the next attempt. Also the iteration may always be interrupted by pressing '.' (full stop) and when continuing (after changing the model, data or the computational method) the latest values of the parameters serve as initial values, unless stated otherwise by the user.

For example, in Disp.1 we could generalize the model on line 18 to the following non-linear form:

Disp.3
```
 14    * Spain           2.5  0.03  43.6  73.2  2.7
 15    B England         1.8  3.49 113.7   5.1  1.4
 16    *
 17    *MODEL BEER1
 18    *log(Beer)=constant+coeff*log(Tea+C*Coffee)
 19    *
 20    *ESTIMATE COUNTRIES,BEER1,21
 21    * constant=4.488964      (0.1565749)
 22    * coeff=0.3276288        (0.0752709)
 23    * RSS=1.553654  R↑2=0.6545
 24    *
 25    *METHOD=N
```

If then ESTIMATE on line 20 is re-activated, the present values of 'constant' and 'coeff' on lines 21 and 22 will be used as initial estimates and since no initial value for the new parameter (C) is given, C=0 will be used as such. Observe also that we have inserted METHOD=N on line 25 thus requiring that the Newton-Raphson method ought still to be employed, although the model is not linear anymore.

After 9 iterations the following results will finally be displayed:

Disp.4
```
 14    * Spain           2.5  0.03  43.6  73.2  2.7
 15    B England         1.8  3.49 113.7   5.1  1.4
 16    *
 17    *MODEL BEER1
 18    *log(Beer)=constant+coeff*log(Tea+C*Coffee)
 19    *
 20    *ESTIMATE COUNTRIES,BEER1,21_
 21    * constant=4.163718      (0.5073659)
 22    * coeff=0.5183450        (0.2507947)
 23    * C=0.0609008            (0.1059470)
 24    * RSS=1.488258  R↑2=0.6691
 25    *METHOD=N
```

6. CONSTANTS IN THE MODEL

Numerical constants appearing in the model can, of course, be notated normally as numbers. Sometimes, however, it is useful to have symbolic notations. In that case the value of the constant should be entered in the edit field in the form

<name of the constant>=<numerical value>

where <name of the constant> is a string starting with a '#'.

Examples of symbolic constants: #PI=3.14159265 #Mean=370.23333. Observe that in the model the symbolic constants are notated without the prefix '#'.

By using this facility it is easy also to fix any parameter in the model temporarily.

In the next example it is shown, how to compute a mean of a variable and then use its centered values in another model.

In fact we are continuing the previous example and at first make a border line of consecutive '.'s (line 27 in the next display). Thus we can define independent regression schemes in the same edit field according to the same rules as we have for computation schemes. Observe however, that the data sets (DATA specification) and models

(MODEL specifications) are always global and can be referred to from any subfield. On the contrary all specifications using the connector '=' are local and accessible only from the same subfield limited by the *..... lines. Thus initial values, symbolic constants and extra specifications like METHOD=D are set separately for each subfield.

Now, in order to compute the arithmetic mean (and the standard deviation) of the variable 'Tea', we can enter the following model and ESTIMATE call:

Disp.4

```
      1  SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 26   *
 27   *.....
 28   *
 29   *MODEL A1
 30   *Tea=Tmean
 31   *ESTIMATE COUNTRIES,A1,32_
 32   *
 33   *
 34   *
```

Since SUM (Tea-Tmean)↑2 is minimized with respect to Tmean when Tmean is the arithmetic mean of the variable 'Tea', activation of ESTIMATE on line 31 then leads to

Disp.5

```
      1  SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 26   *
 27   *.....
 28   *
 29   *MODEL A1
 30   *Tea=Tmean
 31   *ESTIMATE COUNTRIES,A1,32_
 32   * Tmean=0.7766667        (0.3851944)
 33   * RSS=19.58547  R↑2=0
 34   *
```

where we have the arithmetic mean of 'Tea' Tmean=0.7766667 and its standard deviation (0.3851944). To compute a quadratic model for 'Beer' with 'Tea' as the sole regressor we may enter another model A2, where 'Tea' appears in the centered form Tea-Tmean. To employ Tmean as a constant in this model we add a '#' in front of Tmean on line 32. Activation of ESTIMATE on line 37 then leads to the result:

Disp.6

```
      1  SURVO 76 EDITOR   (C)1979 S.Mustonen    (100x100)
 26   *
 27   *.....
 28   *
 29   *MODEL A1
 30   *Tea=Tmean
 31   *ESTIMATE COUNTRIES,A1,32
 32   *#Tmean=0.7766667          (0.3851944)
 33   * RSS=19.58547  R↑2=0
 34   *
 35   *MODEL A2
 36   *Beer=a+b*(Tea-Tmean)+c*(Tea-Tmean)↑2
 37   *ESTIMATE COUNTRIES,A2,38_
 38   * a=111.3094               (17.62399)
 39   * b=82.63949               (23.18871)
 40   * c=-28.02650              (10.27111)
 41   * RSS=3195.018  R↑2=0.7721
 42   *
```

7. WEIGHTING OF OBSERVATIONS

The observations can be weighted by using a WEIGHT specification

WEIGHT=<weight function>

where the weight function is a function of any variables appearing in the data set (typically the weight function is simply a variable). If WEIGHT is not given, WEIGHT=1 is used as default. The weight function is expressed according to the same rules as the model function, but no unknown parameters are allowed.

When WEIGHT is in use, it is possible to estimate models of the general type

g(X,A)=f(X,A)+eps/sqr(w(X))

where

X and A are the variables and the parameters, respectively,
g(X,A) is the regressand (function),
f(X,A) is the model function (regressor function),
w(X) is the weight function,
eps is a normal error term with zero mean and unknown constant variance.

To specify this kind of a model for the ESTIMATE operation we have to define MODEL in the form g(X,A)=f(X,A) and WEIGHT=w(X).

If g(X,A) is independent of A, which is the normal case, we obtain maximum likelihood estimates for the parameters A when the standard OLS criterion is used and the observations are independent.

If g(X,A) depends on A, the estimation procedure will not take into account the Jacobian of the g-transformation (see 8). To guarantee that the optimization problem is well-defined, the model is to be formulated so that the regressand will be approximately independent of A.

To demonstrate the use of ESTIMATE in this general situation we make a simulation experiment. In the next display 20 independent observations are generated according to model

Y=a+b*sin(c*t)+sqr(t)*eps

where t=1,2,...,20, a=100, b=10, c=0.1 and eps is N(0,0.3↑2). This is done by activating the COMP operation on line 57:

Disp.7

```
      1  SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 52   *.....................................................................
 53   *
 54   *    Y=a+b*sin(c*t)+sqr(t)*eps
 55   *    a=100, b=10, c=0.1, eps=N.G(0,.09,rnd(1))
 56   *
 57   *COMP 61,80,60,59_
 58   *DATA TEST,X,Y,Z
 59   *          XX        123.123
 60   Z           t           Y
 61   X  1        1        100.681
 62   *  2        2        102.151
 63   *  3        3        103.017
 64   *  4        4        104.069
 65   *  5        5        105.693
 66   *  6        6        106.182
 67   *  7        7        105.315
 68   *  8        8        107.637
 69   *  9        9        107.860
 70   * 10       10        109.704
 71   * 11       11        109.471
 72   * 12       12        109.982
 73   * 13       13        109.817
 74   * 14       14        109.942
 75   * 15       15        111.600
 76   * 16       16        111.273
 77   * 17       17        112.443
 78   * 18       18        110.721
 79   * 19       19        108.345
 80   Y 20       20        106.565
 81   *
```

Using this artificial data set we have tried to estimate the same model first without weighting the observations (lines 84-91 in the next display) and then by employing correct weighting (weight function w(t)=1/t; lines 93-98).

Disp.8

```
 82   *.....................................................................
 83   *
 84   *MODEL TRIG
 85   *Y=a+b*sin(c*t)
 86   *
 87   *ESTIMATE TEST,TRIG,88
 88   * a=99.26542              (0.7043820)
 89   * b=11.28469              (0.8837113)
 90   * c=0.1084699             (0.0043882)
 91   * RSS=18.31138  R↑2=0.9129
 92   *.....................................................................
 93   *WEIGHT=1/t
 94   *ESTIMATE TEST,TRIG,95_
 95   * a=99.60848              (0.2828318)
 96   * b=10.85851              (0.4674275)
 97   * c=0.1060417             (0.0048035)
 98   * RSS=7.852696  R↑2=0.9695
 99   *
```

8. ESTIMATION CRITERIA

The normal estimation criterion in ESTIMATE is ordinary least squares (OLS) which in case of the general model presented in the previous chapter (model g(X,A)=f(X,A), WEIGHT=w(x)) implies the minimization of

SUM w(X)*(g(X,A)-f(X,A))↑2

with respect to the parameters A.

By using an extra specification CRITERION=Lp where p is any positive constant the estimates will be obtained by minimizing the generalized criterion

SUM w(X)*ABS(g(X,A)-f(X,A))↑p

CRITERION=L2 is always default and thus corresponds to OLS.

CRITERION=L1 can also be given in the form CRITERION=ABS and it implies the sum of absolute deviations to be used as the object function to be minimized.

The influence of the criterion selected is illustrated in the following display where a simple data set having a "serious outlier" on line 15 is analyzed with the model Y=a*X (true a=1) and by using p=2,1,0.1 and 10 successively.

In the results obtained by Hooke-Jeeves' method we have RSS=minimum value of the object function and N(fnct)=number of function evaluations.

Disp.9

```
      1  SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 1    *
 2    *MODEL YX
 3    *Y=a*X
 4    *
 5    *ESTIMATE KOE,YX,6
 6    * a=1.025974              (0.0328548)
 7    * RSS=3.740260  R↑2=0.9555
 8    *
 9    *DATA KOE,11,20,10
10    *         X  Y
11    *  1      1  1
12    *  2      2  2
13    *  3      3  3
14    *  4      4  4
15    *  5      5  7
16    *  6      6  6
17    *  7      7  7
18    *  8      8  8
19    *  9      9  9
20    * 10     10 10
21    *.....................................................................
22    *CRITERION=L1 METHOD=H
23    *ESTIMATE KOE,YX,24
24    * a=1.000583375
25    * MIN Lp=2.026251875  N(fnct)=37  Final step length=.0009765625
```

```
 26    *.....................................................................................
 27    *CRITERION=L0.1 METHOD=H
 28    *ESTIMATE KOE,YX,29
 29    * a=1
 30    * MIN Lp=1.071773462536  N(fnct)=25  Final step length=.0009765625
 31    *.....................................................................................
 32    *CRITERION=L10 METHOD=H
 33    *ESTIMATE KOE,YX,34
 34    * a=1.123046875
 35    * MIN Lp=37.78600958657  N(fnct)=37  Final step length=.0009765625
 36    *
```

9. RESIDUALS AND PREDICTED VALUES

The residuals g(X,A)-f(X,A) and the predicted values f(X,A) and g(X,A) may be computed jointly with the ESTIMATE operation and displayed as new columns in the data matrix. In this case the ESTIMATE call must include an extra (fourth) parameter which is the number (or label) of an image line. This image line indicates the places and formats of the pertinent new columns so that

-RR.RRR is image for residuals g(X,A)-f(X,A),
-GGG.GG is image for values g(X,A),
-FFF.FF is image for values f(X,A).

Any of these options may, of course, be omitted. Also the order is immaterial, but all columns indicated by these images must be located on the right side of the data set involved.

In the next display our first example (displays 1,2) is repeated by using this extra parameter (images are on line 16) in ESTIMATE.

Disp.10

```
      1  SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
  1   *SAVE DATA
  2   *DATA COUNTRIES,A,B,C
  3   C              Coffee  Tea  Beer   Wine Spirits
  4   A Finland        12.5 0.15  54.7   7.6  2.7      0.134  4.00  3.86
  5   * Sweden         12.9 0.30  58.3   7.9  2.9     -0.028  4.06  4.09
  6   * Danmark        11.8 0.41 113.9  10.4  1.7      0.538  4.73  4.19
  7   * Norway          9.4 0.19  43.5   3.1  1.8     -0.172  3.77  3.94
  8   * France          5.2 0.10  44.5 104.3  2.5      0.060  3.79  3.73
  9   * Ireland         0.2 3.73 124.5   3.8  1.9     -0.095  4.82  4.92
 10   * Italy           3.6 0.06  13.6 106.6  2.0     -0.957  2.61  3.56
 11   * Holland         9.2 0.58  75.5   9.7  2.7      0.013  4.32  4.31
 12   * Portugal        2.2 0.03  27.5  89.3  0.9     -0.025  3.31  3.34
 13   * Switzerland     9.1 0.25  73.5  44.9  2.1      0.262  4.29  4.03
 14   * Spain           2.5 0.03  43.6  73.2  2.7      0.434  3.77  3.34
 15   B England         1.8 3.49 113.7   5.1  1.4     -0.164  4.73  4.89
 16   *                                               -R.RRR -GG.GG -FF.FF
 17   *MODEL BEER1
 18   *log(Beer)=constant+coeff*log(Tea)
 19   *
 20   *ESTIMATE COUNTRIES,BEER1,21,16_
 21   * constant=4.488964       (0.1565749)
 22   * coeff=0.3276288         (0.0752709)
 23   * RSS=1.553654  R↑2=0.6545
```

10. SELECTING OBSERVATIONS

The image line specified by the extra fourth parameter in the ESTIMATE operation (see 9) may also be used to indicate the observations which actually are to be handled. Setting an image I to any position on this image line implies the corresponding column in the data set to be selected as the indicator. If a 'blank','0' or '-' occurs in that column, the corresponding observation will be omitted. All other characters let the observation to be analyzed.

The residuals and predicted values set by the same image line will, however, be computed also for observations which are omitted in the estimation procedure.

In the preceding example 'Italy' seems to be an exceptional observation. Treating 'Italy' as an outlier we may repeat the same analysis by using the indicator specified on the image line 16. Thus by reactivating ESTIMATE on line 20 the following results will be obtained.

Disp.11

```
      1  SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 1    *SAVE DATA
 2    *DATA COUNTRIES,A,B,C
 3    C              Coffee  Tea  Beer   Wine Spirits
 4    A Finland        12.5 0.15  54.7   7.6  2.7      0.013  4.00  3.98 1
 5    * Sweden         12.9 0.30  58.3   7.9  2.9     -0.110  4.06  4.17 1
 6    * Danmark        11.8 0.41 113.9  10.4  1.7      0.474  4.73  4.26 1
 7    * Norway          9.4 0.19  43.5   3.1  1.8     -0.279  3.77  4.05 1
 8    * France          5.2 0.10  44.5 104.3  2.5     -0.083  3.79  3.87 1
 9    * Ireland         0.2 3.73 124.5   3.8  1.9     -0.033  4.82  4.85 1
10    * Italy           3.6 0.06  13.6 106.6  2.0     -1.130  2.61  3.74 0
11    * Holland         9.2 0.58  75.5   9.7  2.7     -0.030  4.32  4.35 1
12    * Portugal        2.2 0.03  27.5  89.3  0.9     -0.238  3.31  3.55 1
13    * Switzerland     9.1 0.25  73.5  44.9  2.1      0.170  4.29  4.12 1
14    * Spain           2.5 0.03  43.6  73.2  2.7      0.222  3.77  3.55 1
15    B England         1.8 3.49 113.7   5.1  1.4     -0.106  4.73  4.83 1
16    *                                               -R.RRR -GG.GG -FF.FF I
17    *MODEL BEER1
18    *log(Beer)=constant+coeff*log(Tea)
19    *
20    *ESTIMATE COUNTRIES,BEER1,21,16
21    * constant=4.501612      (0.0909884)
22    * coeff=0.2705604        (0.0454895)
23    * RSS=0.4717560  R↑2=0.7972
```

11. MAXIMUM LIKELIHOOD ESTIMATES FOR UNIVARIATE DISTRIBUTIONS

The ESTIMATE operation also enables the computation of maximum likelihood estimates for a user-defined univariate distribution. In this case the MODEL specification has to be written in the form

*MODEL <name of the model>
*LOGDENSITY=<logarithm of the density function>

Thus the logarithm of the density function of a single observation has to be given and it is assumed that the data set defined by a DATA specification is a random sample of the distribution in question.

Otherwise the ESTIMATE operation is used in the same way as in regression models and some extra specifications and options (like METHOD,

#constants, initial values) are still valid.

As an example we try again to estimate the model appearing in display 1, log(Beer)=constant+coeff*log(Tea) where it is hitherto tacitly assumed that the model has an additive normal error term with zero mean and unknown constant variance (notated by 'var' in sequel).

The same problem may now be handled by entering the logdensity of the normal distribution for 'log(Beer)' with mean 'constant+coeff*log(Tea)' and variance 'var'. This is expressed as the model NORMAL (on lines 18-19 in the next display).

Since 'var' is a "nuisance" parameter for computational reasons, too, it is best to start the estimation by keeping 'var' constant by setting #var=.1 (on line 21).

After ESTIMATE now on line 22 has been activated we shall have the following display where the estimates obtained for 'constant' and 'coeff' are final (due to the form of the normal density)' but their standard errors are not.

Disp.12

1 SURVO 76 EDITOR (C)1979 S.Mustonen (100x100)

```
 1  *
 2  *DATA COUNTRIES,A,B,C
 3  C              Coffee  Tea  Beer   Wine Spirits
 4  A Finland        12.5 0.15  54.7   7.6  2.7
 5  * Sweden         12.9 0.30  58.3   7.9  2.9
 6  * Danmark        11.8 0.41 113.9  10.4  1.7
 7  * Norway          9.4 0.19  43.5   3.1  1.8
 8  * France          5.2 0.10  44.5 104.3  2.5
 9  * Ireland         0.2 3.73 124.5   3.8  1.9
10  * Italy           3.6 0.06  13.6 106.6  2.0
11  * Holland         9.2 0.58  75.5   9.7  2.7
12  * Portugal        2.2 0.03  27.5  89.3  0.9
13  * Switzerland     9.1 0.25  73.5  44.9  2.1
14  * Spain           2.5 0.03  43.6  73.2  2.7
15  B England         1.8 3.49 113.7   5.1  1.4
16  *
17  *
18  *MODEL NORMAL
19  *LOGDENSITY=-0.5*((log(Beer)-constant-coeff*log(Tea))↑2/var+log(var))
20  *METHOD=N
21  *#var=.1
22  *ESTIMATE COUNTRIES,NORMAL,23_
23  * constant=4.488964      (0.1256160)
24  * coeff=0.3276288        (0.0603879)
```

Now to obtain the ML estimate for 'var', we fix 'constant' and 'coeff' by setting a '#' in front of them on lines 23-24 and on the other hand delete '#' from line 21 thus letting 'var' be the only parameter to be estimated.

After altering the last parameter of ESTIMATE from 23 to 25 and by reactivating line 22 a new result will appear on line 25:

Disp.13
```
 17    *
 18    *MODEL NORMAL
 19    *LOGDENSITY=-0.5*((log(Beer)-constant-coeff*log(Tea))↑2/var+log(var))
 20    *METHOD=N
 21    * var=.1
 22    *ESTIMATE COUNTRIES,NORMAL,25_
 23    *#constant=4.488964       (0.1256160)
 24    *#coeff=0.3276288         (0.0603879)
 25    * var=0.1294712           (0.0528564)
 26    *
```

To obtain correct values for the standard errors and to check the results in general it is best to do the same job with all 3 parameters simultaneusly still once. Thus after erasing line 21 and the '#'s from lines 23-24 and by activating ESTIMATE we finally have

Disp.14
```
 17    *
 18    *MODEL NORMAL
 19    *LOGDENSITY=-0.5*((log(Beer)-constant-coeff*log(Tea))↑2/var+log(var))
 20    *METHOD=N
 21    *
 22    *ESTIMATE COUNTRIES,NORMAL,23_
 23    * constant=4.488964       (0.1429327)
 24    * coeff=0.3276288         (0.0687126)
 25    * var=0.1294712           (0.0528564)
 26    *
```

12. SPECIAL APPLICATIONS

It has been stated previously (see 7) that also the regressand in the model defined by a MODEL specification may include parameters to be estimated, but the estimates obtained in this case using the (weighted) OLS criterion are not ML estimates.

As a first simple example of this general type we consider a model of the form (X-a)↑2=b where X is a variable and a,b are parameters to be estimated. It is natural to expect that a is near the mean of X and b is near the variance of X.

We apply this model to COUNTRIES by using 'Beer' as X. Thus we activate ESTIMATE on line 31 in the next display.

Disp.15
```
 27    *
 28    *MODEL ab
 29    *(Beer-a)↑2=b
 30    *
 31    *ESTIMATE COUNTRIES,ab,32
 32    * a=72.79598              (4.935703)
 33    * b=1220.717              (344.8947)
 34    * RSS=13663151  R↑2=0.0000
 35    *
```

To compare the results obtained with the true mean and variance of 'Beer' we may compute these statistics directly either by estimating

model 'Beer=constant' (see Disp.4) or by using the STAT operation as follows:

Disp.16

```
  2    *DATA COUNTRIES,A,B,C
  3    C              Coffee  Tea  Beer   Wine Spirits
  4    A Finland        12.5  0.15  54.7   7.6  2.7
  5    * Sweden         12.9  0.30  58.3   7.9  2.9
.....
 14    * Spain           2.5  0.03  43.6  73.2  2.7
 15    B England         1.8  3.49 113.7   5.1  1.4
 16    *                         XXXXX
 17    *
.....
 31    *ESTIMATE COUNTRIES,ab,32
 32    * a=72.79598              (4.935703)
 33    * b=1220.717               (344.8947)
 34    * RSS=13663151  R↑2=0.0000
 35    *
 36    *STAT COUNTRIES,16,37_
 37    * Beer:      N=12  MEAN=65.5667  STD.DEV.=35.7026
 38    *           SKEWNESS=0.40497  EXCESS=-1.15486
 39    *           MIN=13.6000  MAX=124.500
 40    *
 41    *     STD.DEV.↑2=1274.67564676
 42    *
 43    *     MEAN+0.5*SKEWNESS*STD.DEV.=72.795940961
 44    *
```

After the STAT operation on line 36 has been activated, the basic statistics for 'Beer' (indicated by X's on the image line 16) will be displayed from line 37 (=last parameter in STAT) onwards.

It is seen immediately that a and b do not match exactly with MEAN and STD.DEV.↑2 (which is computed afterwards on line 41).

In fact, it can be shown that the OLS principle in this case leads to an estimate a=MEAN+0.5*SKEWNESS*STD.DEV and this result is demonstrated on line 43.

As another example we shall study the effect of the Box-Cox power transformation in a certain special case where it is assumed that the model (Y↑C-1)/C=a*X+b+eps is valid for some unknown value of C. An artificial data set of 40 observations with X=1,2,...,40, a=-0.2, b=3, eps=N(0,0.1) and C=0 (i.e. log(Y)=a*X+b+eps) is generated by a COMP operation:

Disp.17

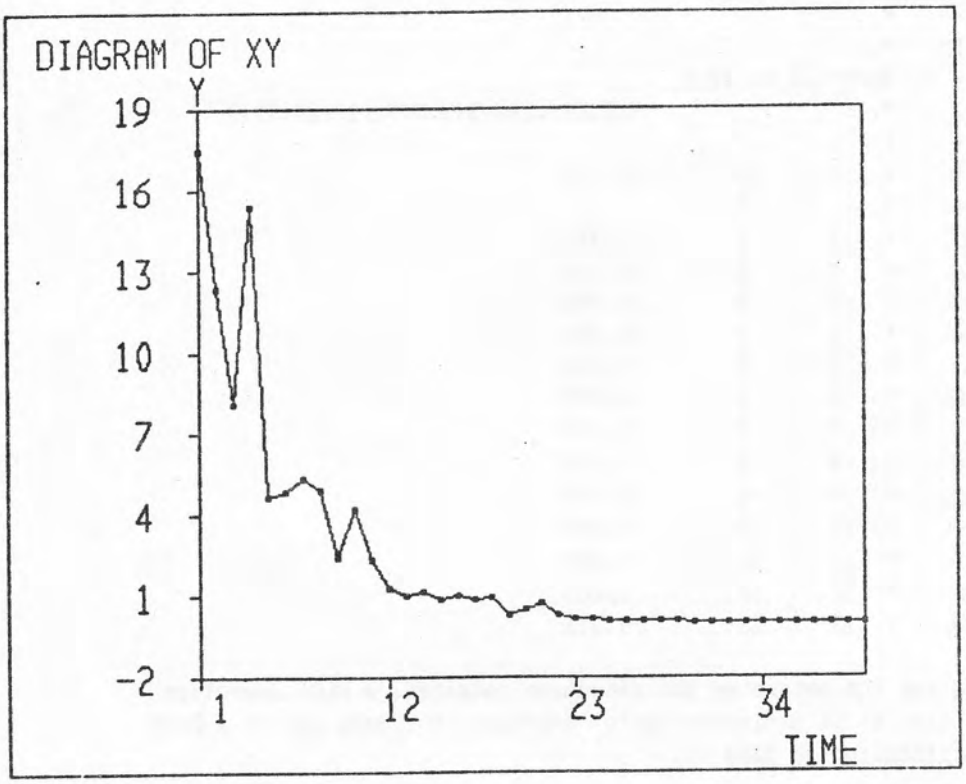
```
      1   SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 1    *
 2    *
 3    *
 4    *
 5    *
 6    *COMP 11,50,10,9_
 7    *                  Y=EXP(-.2*X+3+N.G(0,.1,rnd(1)))
 8    *
 9    *         XX       12.123
10    *          X         Y
11    *   1      1       17.483
12    *   2      2       12.316
13    *   3      3        8.050
14    *   4      4       15.391
15    *   5      5        4.628
16    *   6      6        4.872
17    *   7      7        5.358
18    *   8      8        4.952
19    *   9      9        2.430
20    *  10     10        4.208
21    *  11     11        2.329
22    *  12     12        1.269
23    *  13     13        1.011
```

To see the nature of the situation obtained, a PLOT operation (on line 4) is activated (after labelling the data set by a DATA specification on line 8):

Disp.18

```
      1   SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 1    *
 2    *
 3    *SIZE=500,500
 4    *PLOT XY,5_
 5    *         TT       YYYYYY
 6    *COMP 11,50,10,9
 7    *                  Y=EXP(-.2*X+3+N.G(0,.1,rnd(1)))
 8    *DATA XY,A,B,C
 9    *         XX       12.123
10    C          X         Y
11    A   1      1       17.483
12    *   2      2       12.316
13    *   3      3        8.050
```

and the following plot will appear on the graphic screen:



Finally, an ESTIMATE operation supported by a MODEL specification is activated using C=1 as an initial estimate and the following results are obtained:

```
Disp.19
      1   SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 48   *  38     38        0.010     -0.032
 49   *  39     39        0.005     -0.431
 50   B  40     40        0.007      0.040
 51   *                              -R.RRR
 52   *MODEL TEST
 53   *(Y↑C-1)/C=a*X+b
 54   * C=1
 55   *ESTIMATE XY,TEST,56,51_
 56   * C=0.0332740              (0.0206612)
 57   * a=-0.1886771             (0.0058234)
 58   * b=2.932239               (0.1051248)
 59   * RSS=3.831644  R↑2=0.9802
 60   *
```

Since the residuals were also computed (due to -R.RRR on the image

line 51), they can immediately be plotted on probability paper, too. At first we sort the observations with a SORT operation (on line 61) by using the residual column as a sort key

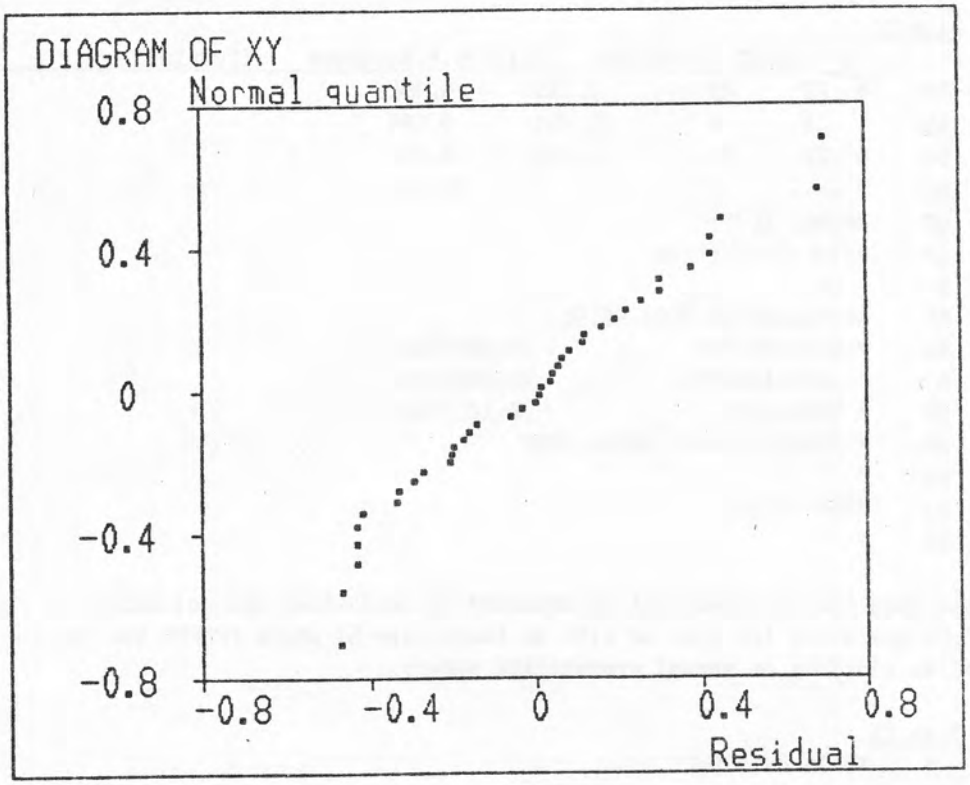
Disp.20

```
      1   SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 48   *  29     29        0.115      0.452
 49   *   4      4       15.391      0.684
 50   B  21     21        0.719      0.701
 51   *                             111111
 52   *MODEL TEST
 53   *(Y↑C-1)/C=a*X+b
 54   * C=1
 55   *ESTIMATE XY,TEST,56,51
 56   * C=0.0332740            (0.0206612)
 57   * a=-0.1886771           (0.0058234)
 58   * b=2.932239             (0.1051248)
 59   * RSS=3.831644  R↑2=0.9802
 60   *
 61   *SORT XY,51_
 62   *
```

and then the required plot is obtained by activating the following PLOT operation (on line 62 with an image line 51 where PPPPPP now implies plotting on normal probability paper).

Disp.21

```
  8   *DATA XY,A,B,C
  9   *         XX       12.123
 10   C          X          Y       Residual
 11   A  23     23        0.144     -0.469
 12   *  13     13        1.011     -0.468
 13   *  19     19        0.331     -0.432
.....
 48   *  29     29        0.115      0.452
 49   *   4      4       15.391      0.684
 50   B  21     21        0.719      0.701
 51   *                             PPPPPP
 52   *MODEL TEST
 53   *(Y↑C-1)/C=a*X+b
 54   * C=1
 55   *ESTIMATE XY,TEST,56,51
 56   * C=0.0332740            (0.0206612)
 57   * a=-0.1886771           (0.0058234)
 58   * b=2.932239             (0.1051248)
 59   * RSS=3.831644  R↑2=0.9802
 60   *
 61   *SORT XY,51
 62   *PLOT XY,51_
 63   *
```



ESTIMATE can also be used for solving nonlinear equations in the following way. To find a real root for the equation sin(X↑2+1)=sqr(X)-1 we enter it as a model EQUATION (on lines 49-50 in the next display) and activate an ESTIMATE (on line 53) with a 'dummy' data set (COUNTRIES) and X=1 (on line 51) as the initial value. By the Hooke-Jeeves' method (METHOD=H) the following result will be obtained:

Disp.22

```
      1  SURVO 76 EDITOR    (C)1979 S.Mustonen    (100x100)
 47   *.....................................................................
 48   *
 49   *MODEL EQUATION
 50   *sin(X↑2+1)=sqr(X)-1
 51   *X=1
 52   *METHOD=H     STEP=.1,.00001
 53   *ESTIMATE COUNTRIES,EQUATION,54
 54   * X=1.399169921876
 55   * RSS=.00000000306064587  N(fnct)=50  Final step length=.000006103515625
 56   *
```

REFERENCES

Mustonen, S. (1980), Interactive analysis in SURVO 76, Proceedings in Computational Statistics, ed.by M.M.Barritt and D.Wishart, 253-259, Physica-Verlag, Wien.

Mustonen, S. (1980), SURVO 76 EDITOR, a new tool for interactive statistical computing, text and data management, Research Report No. 19, Dept.of Statistics, University of Helsinki.

Mustonen, S. (1981), SURVO 76 EDITOR, a new tool for interactive statistical computing, text and data management, (RELEASE 2), Research Report No.24, Dept.of Statistics, University of Helsinki.

Walsh, G.R. (1975), Methods of Optimization, Wiley, New York.

RESEARCH REPORTS

Department of Statistics
University of Helsinki

No. 18 Nykyri, Erkki. The estimation of the mixture of two normal distributions. 6 pp. May 1980.
ISBN 951-45-2002-5.

No. 19 Mustonen, Seppo. SURVO 76 EDITOR, a new tool for interactive statistical computing, text and data management. 50 pp. August 1980.
ISBN 951-45-2048-3

No. 20 Tuomikoski, Jaakko. Two approaches to the digression problem: a comparison by simulation experiments. 41 pp. September 1980.
ISBN 951-45-2063-7.

No. 21 Vartia, Yrjö O. Summan varianssin ja hajonnan jakamisesta summan osatekijöille. 24 s. Lokakuu 1980.
ISBN 951-45-2123-4

No. 22 Vartia, Yrjö O. Interpolation of cumulative frequency curves by cubic splines. 27 pp. November 1980.
ISBN 951-45-2139-0

No. 23 Teräsvirta, Timo. Overestimation of mean square error matrix in misspecified linear models. 8 s. December 1980.
ISBN 951-45-2155-2

No. 24 Mustonen, Seppo. SURVO 76 EDITOR, a new tool for interactive statistical computing, text and data management. (Release 2). 65 pp. February 1981.
ISBN 951-45-2209-5

No. 25 Saikkonen, Pentti. Estimates for spectral averages. 16 pp. June 1981.
ISBN 951-45-2343-1

No. 26 Saikkonen, Pentti. Asymptotic moments for an estimator of the inverse autocorrelation function. 9 pp. June 1981.
ISBN 951-45-2343-1

No. 27 Soininvaara, Osmo. Alkoholi ja liikennekuolema. Alkoholitapaukset liikennevahinkojen tutkijalautakuntien tutkimissa kuolemaan johtaneissa liikenneonnettomuuksissa vuosina 1970-1977. 17 pp. August 1981.
ISBN 951-45-2379-2

No. 28 Saikkonen, Pentti. Asymptotic properties of some tests for autocorrelation. 14 pp. August 1981.
ISBN 951-45-2386-5