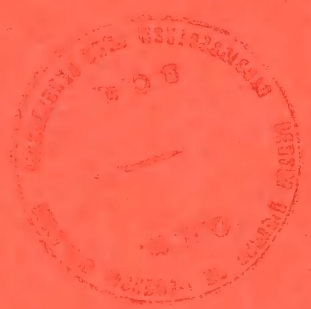
SCK • CEN
0096300

STUDIECENTRUM VOOR KERNENERGIE

CENTRE D'ETUDE DE L'ENERGIE NUCLEAIRE

NON PARAMETRIC FITTING OF A STRAIGHT LINE

J.L. VAN DER PARREN



July 1985

E. Plaskylaan 144, 1040 BRUSSEL
(BELGIE)

BLG 579

144, avenue E. Plasky, 1040 BRUXELLES
(BELGIQUE)

ex. 1

J.L. VAN DER PARREN
BLG (July 1985)-

NON-PARAMETRIC FITTING OF A STRAIGHT LINE

Summary.- This report contains a detailed description of the methods used in CEN-SCK programs LINFIT0, LINFIT1 and LINFIT2 together with historical background and considerations on efficiency.

Keywords : linear regression, distribution-free methods; non-parametric methods.

J.L. VAN DER PARREN
BLG (July 1985)

NON-PARAMETRIC FITTING OF A STRAIGHT LINE

Résumé.- Ce rapport contient une description détaillée des méthodes utilisées dans les programmes CEN-SCK LINFIT0, LINFIT1 et LINFIT2 ainsi qu'un aperçu historique et des considérations sur l'efficience.

J.L. VAN DER PARREN
BLG (July 1985)

NON-PARAMETRIC FITTING OF A STRAIGHT LINE

Samenvatting.- Dit verslag bevat een uitvoerige beschrijving van de methoden die in de SCK-CEN programma's LINFIT0, LINFIT1 en LINFIT2 gebruikt werden, een historisch overzicht, en beschouwingen aangaande hun doelmatigheid.

TABLE OF CONTENTS

Chapter 1.

NON-PARAMETRIC FITTING OF A STRAIGHT LINE WITH x ERROR-FREE

SUMMARY

A method is proposed for the estimation of the parameters γ and β of the straight line $\tilde{y}(x) - \gamma = \beta(x - \bar{x})$ relating the median $\tilde{y}(x)$ of $y(x)$ to the value x at which y is measured. The 'errors' $\varepsilon(x) = y(x) - \tilde{y}(x)$ are supposed identically independently distributed with an unknown continuous distribution. The latter is not supposed symmetrical. Several x_i may assume the same value.

We define $\tilde{\beta}$ as the estimated slope and $\tilde{\gamma}$ as the estimate for the ordinate at $\bar{x}$. $\tilde{\beta}$ is obtained by minimizing the correlation between $\varepsilon(x)$ and x, $\tilde{\gamma}$ by requiring that

$$\text{median } \{y_i - \text{estimated median at } x_i\} = 0.$$

The method may be used instead of least squares in the case the $\varepsilon(x)$ are normally distributed, but presence of outliers is suspected in order to reduce their biasing influence.

1. THE PROBLEM

Measurements y_i are obtained at points x_i. The ε_i's, where $\varepsilon_i = y_i - \tilde{y}_i$, are supposed identically independently distributed with the same unknown continuous distribution that is not necessarily symmetrical (we define $\tilde{y}_i$ as the median of the variable $y(x_i)$).

We want to estimate the parameters γ and β, resp. the ordinate at $\bar{x}$ (the mean of the x_i's) and the slope of the straight line $\tilde{y}(x) = \alpha + \beta x$. It is supposed that more than one measurement can be present for each x-value.

2. PREVIOUS RELATED WORK

The problem, in a somewhat restricted form (the x_i's are supposed different), has already been treated by MARITZ (1979).

The idea applied here for the estimation of the ordinate at origin, once an estimate for the slope has been obtained, is not new but its simplicity does not mean it is used; recent publications propose complicated procedures which were not proved to yield a better estimate. The principle was sketched already in the years fourty by NAIR & SHRIVASTAVA in an artisanal method reported by McNEIL (1977).

A complete procedure inspired by this slide-a-paper method was suggested by TUKEY (1977). The method is as follows :

- 3 groups of equal importance are formed according to x_i-order
- 'medial points' of the outer groups are found : having the co-ordinates : median $\{x_j\}$ and median $\{y_j\}$ in each group, where j is an index taking the values of i included in that group.
- the estimated slope is that of the straight line joining the 'medial points'.

THEIL (1950), SEN (1968) and CIFARELLI (1978) give an estimate for β only.

The methods of these four authors are related to the one used here, as they realize a minimization of the correlation between the x_i's and the deviations : $\Delta y_i = y_i - \hat{\alpha} - \hat{\beta}x_i$ ($\hat{}$ means : estimate of), a feature that THEIL did not point out.

In SEN (on the steps of THEIL) and MARITZ, the absolute value of the Kendall's τ correlation coefficient is minimized whereas in CIFARELLI, that of the GINI's association coefficient is.

MARITZ estimates the intercept at origin α by means of

$$\text{med } \{(x_i y_j - x_j y_i) / (x_i - x_j)\}, \quad i < j.$$

MOOD & BROWN (1950) find $\hat{\alpha}$ and $\hat{\beta}$ by solving the two equations

$$\text{med } \{y_i - \hat{\alpha} - \hat{\beta}x_i\} = 0 \quad \text{for } x_i \leq \text{med } \{x_i\}$$

$$\text{med } \{y_i - \hat{\alpha} - \hat{\beta}\, x_i\} = 0 \quad \text{for } x_i > \text{med } \{x_i\}$$

by trial and error, which is a heavy procedure.

Furthermore, SEN found that other estimates of β are more efficient.

Different approaches have also been offered by ADICHIE (1967), BROWN (1980) and SIEGEL (1982) for estimating a straight line and HOGG & RANDLES (1975) for estimating the slope.

Recently, BHATTACHARYA, CHERNOFF and YANG (1983) studied the estimation of the slope in the case of a truncated regression, extending the work of THEIL and SEN.

3. GENERAL OUTLINE OF THE METHOD

β is estimated by requiring that no correlation exists between the deviations $\Delta y_i\ (\alpha, \beta) = y_i - \alpha - \beta x_i$ and the x_i's. The correlation measure here is the Spearman correlation coefficient. It is well known (SIEGEL 1956) that the power of the test for zero correlation using Spearman correlation coefficient or Kendall correlation coefficient is the same. So, the use of this coefficient instead of the Kendall τ should be expected to maintain all the advantages of the Kendall τ but as the sum extends to the index i only instead of to the couples (i, j), a computational advantage will be obtained in the cases where n is not very small. We will note that our correlation coefficient will depend only on β as the addition of a constant will not modify the ranks of the Δy_i. It is however advisible to choose an α close to its real value to improve the precision of the computation.

$\rho_s(\beta)$ is a step function (piecewise constant). There may be a step exactly at zero, but this is not generally the case. We will choose as $\tilde{\beta}$ that point where $\rho_s(\beta)$ jumps from a negative to a positive value. In the case a step is zero, at machine precision, β is the centre of the interval.

Once $\tilde{\beta}$ is obtained, $\tilde{\gamma}$ is found by requiring that

$$\text{med}\ \{y_i - \tilde{\gamma} - \tilde{\beta}\ (x - \bar{x})\} = 0$$

i.e. the median of the deviations should be zero.

4. DETAILED ALGORITHM

- Order the couples (x_i, y_i) so that the x_i's are in increasing order.

- Find the centre of gravity $(\bar{x}, \bar{y})$.

- Find the centre of gravity of the r first points : $(\bar{x}_r, \bar{y}_r)$ where $r = [n/3]$.

- The first guess is the straight line through $(\bar{x}, \bar{y})$ and $(\bar{x}_r, \bar{y}_r)$ with slope β_0 and intercept at $\bar{x}$: y_0.
  (A slope $\beta_0 = 0$ at machine precision is not accepted as first guess, it is automatically replaced by $\beta_0 = .01$).

- The Spearman correlation coefficient $\rho_s(\beta_0)$ is computed between the deviations $\{\Delta y_i\}$ and the $\{x_i\}$: $\Delta y_i(\gamma_0, \beta) = y_i - \gamma_0 - \beta(x_i - \bar{x})$

- Trials β_1, β_2 are made on both sides of β_0 with increments $.25\,\beta_0$, till two of the values $\rho_s(\beta_1)$, $\rho_s(\beta_0)$ $\rho_s(\beta_2)$ are of opposite sign ($\beta_0 = 0$ would provide zero increments). In the successive trial fits, β is modified while γ_0 is not but ρ_s depends on β only.

  The jump point is obtained by linear interpolation except if a ρ_s value that is zero at machine precision is met (for β^*).
  Procedure * is then used.
  * In this case, one starts on one side with β^+ and β^*, on the other with β^- and β^* and find the limits β_L, β_U of the interval where β is presumed to be zero by an halving process, $\tilde{\beta}$ is then $0.5\,(\beta_L + \beta_U)$.

- The slope is now estimated by $\tilde{\beta}$.
  In general, med $\{\Delta y_i(\gamma_0, \tilde{\beta})\}$ will not be zero.
  We have to correct the ordinate at $\bar{x}$ and find $\tilde{\gamma}$ such that
  med $\{\Delta y_i(\tilde{\gamma}, \tilde{\beta})\} = 0$.
  $\tilde{\gamma}$ is, of course, equal to $\gamma_0 + \text{med}\{\Delta y_i(\gamma_0, \tilde{\beta})\}$.
  An estimate for the intercept at origin α is then obtained from $\tilde{\gamma}$ and $\tilde{\beta}$, i.e. :

$$\tilde{\alpha} = \tilde{\gamma} - \tilde{\beta}\,\bar{x}$$

5. PROPERTIES OF THE ESTIMATES

The process is invariant for a linear transformation i.e. the transformed fitted line is identical with the straight line fitted to the transformed points.

In the case of a symmetrical distribution of errors, the estimate $\tilde{\beta}$ of the slope as obtained following 3 and 4 is an unbiased estimate of β and $\tilde{\gamma}$ is also an unbiased estimate of $y(\bar{x})$. In that case, $\tilde{\beta}$ and $\tilde{\gamma}$ are independently distributed. $\tilde{\beta}$ is a very efficient estimate of β.

For equally spaced x_i and one measurement for each x_i, SEN has shown that the estimator using Kendall correlation coefficient minimization has an asymptotic relative efficiency (A.R.E.) versus least squares estimate never less than .864.
This estimator does not differ fundamentally from that used by SEN and its performance will be the same.

Confidence limits on $\tilde{\beta}$ can be obtained in the same way as when Kendall coefficient is used. Natural limits for β are those values of β for which $\rho_s(\beta)$ becomes significantly different from zero. In fact, for small n, the relative efficiency versus least squares seems to be sensibly better than the A.R.E.

The estimation of $\tilde{\gamma}$ has not the same efficiency at all. It behaves as a median versus a mean (A.R.E. compared with normal estimation : .637).

For small n, it is, of course somewhat better. A Monte-Carlo experiment was done with 10000 samples of 8 measurements with normal errors at equally spaced different x_i's.

The experimental relative efficiency versus least squares was .90 for $\tilde{\beta}$ and .71 for $\tilde{\gamma}$.

6. ALTERNATIVE APPROACH

From computational point of view, the use of the GINI's association coefficient would be even lighter.

CIFARELLI shows that it provides a slightly smaller efficiency than the estimates obtained with the Kendall τ in the normal hypothesis, but is better with a Laplace or Cauchy distribution. In these two cases, least squares are not an acceptable choice.

7. APPLICATIONS

The program was used with known "pathological" data i.e. FISCHLER & BOLLES's data (see Table 1) cited by REY and HOGG & RANDLES's data (see Table 2).

The results for the first set give a fit that in the author's opinion is more satisfactory than that proposed by ROUSSEUW (1984) in that it takes good account of the six "acceptable" points whereas the alternative method ignores the sixth point passing nearly exactly through the five points that are in line.

The parameters found by our method are $\tilde{\beta} = .6667$, $\tilde{\alpha} = .3333$.

For the second set, the result is a bit different from that proposed by HOGG & RANDLES but looks quite satisfactory. We find $\tilde{\beta} = .06$, whereas the above cited authors come to an estimate .08 for β and the least squares estimation yields .05.

Our intercept at origin is 3.42 compared to 3.35 for the least squares method (this close agreement is not surprising as a large part of the information corresponds to $x = 0$).

TABLE 1 : FISCHLER & BOLLES DATA

x_i	y_i
0.	0.
1.	1.
2.	2.
3.	2.
3.	3.
4.	4.
10.	2.

TABLE 2 : HOGG & RANDLES's DATA : ACT SCORE and first year GPA

x	y
1	4.00
2	1.93
1	3.47
0	3.00
0	3.27
0	4.00
3	3.62
2	3.89
0	3.87
2	4.00
1	3.00
2	3.73
4	4.00
0	3.56
0	3.36
1	3.55
3	3.20
0	3.30
2	3.00
3	2.88
1	3.06
1	3.00
0	3.47
0	3.27
1	3.75

x	y
0	3.62
0	3.25
0	3.18
0	2.33
0	3.75
0	3.14
1	3.06
1	3.33
1	3.92
2	3.60
0	3.00
0	3.43
0	2.40
0	4.00
0	2.50
0	4.00
1	3.77
1	4.00
1	3.50
1	3.00
2	3.06
2	4.00
0	3.27
0	3.50
1	3.76

8. CONCLUSIONS

The method is conceptually simple.
Its efficiency is as good as presently used methods.
It is computationally lighter, when n is not small.
It allows the fitting to data with more than one measurement for each x-value.
It is invariant for a linear transformation.

BIBLIOGRAPHY

- ADICHIE, J.N. (1967)
  Estimates of regression parameters based on rank tests
  Ann. of Math. Statist. 38, 894-904

- BHATTACHARYA, P.K.; CHERNOFF, H. and YANG, S.S. (1983)
  Non-parametric estimate of the slope of a truncated regression
  The Annals of Statistics 11, no. 2, 505-514

- BROWN, B.M. (1980)
  Median estimates in simple linear regression
  Austral. J. Statist. 22 (2), 154-165

- CIFARELLI, D.M. (1978)
  La stima der coefficiente di regressione mediante l'indice di cograduazione di GINI
  Rivista Matimatica Sci. Econ. Social. 1, no. 1, 7-38

- GINI, C. (1915-1916) Sul criterio di concordanza tra due caratteri
  Atti del Reale Instituto Veneto di Science, Lettere ed Arti
  Anno Acc 1915-1916, 309-331

- HOGG, R.V. and RANDLES, R.H. (1975)
  Adaptative distribution-free regression methods and their applications
  Technometrics 17, no. 4, 399-407

- LEROY, A and ROUSSEEUW, P. (1984)
  A portable FORTRAN program for the median of squares regression
  Rev. Belge de Statistique, d'Informatique et de Recherche Opérationnelle 24, no. 2, 28-38

- MARITZ, J.S. (1979)
  On Theil's method in distribution-free regression
  Australian J. Statist. 21 (1), 30-35

- Mc NEIL, D.R. (1977)
  Interactive Data Analysis, Chapter 3, p. 48 sqq
  Wiley
  (reference gratefully obtained from an anonymous referee)

- MOOD, A.M. (1950)
Introduction to the theory of statistics
Mc Graw Hill- N.Y.

- NAIR, K.R. & SHRIVASTAVA, M.P. (1942-1944)
On a simple method of curve fitting
SANKHYA, Vol 6, 121-132

- REY, J.J. (1983)
Introduction to robust and quasi-robust statistical methods
Springer Verlag

- ROUSSEEUW (1982)
Least median squares regression
Technical Report - Vrije Univ. Brussel
Centre for Statistics and Operations Research CSOOTW/178

- SEN, P.K. (1968)
Estimates of the regression coefficient based on Kendall's τ.
J. of the Amer. Stat. Assoc. (Dec. 1968) 1379-1389

- SIEGEL, A.F. (1982)
Robust regression using repeated medians
Biometrika 69, 1, 242-244

- SIEGEL, S. (1956)
Non-parametric Statistics for the Behavioural Sciences
Mc. Graw Hill - N.Y.

- THEIL, H. (1950)
A rank invariant method of linear and polynomial regression analysis I, II and III
Nederl. Akad. Wetensch. Proc. 53, 386-392, 521-525 and 1397-1412

- TUKEY, John (1977)
Exploratory Data Analysis
Addison-Wesley

APPENDIX to Chapter 1

NON-PARAMETRIC LINEAR FITTING

50 MEASUREMENTS Y(I) AT X(I)

I	X(I)	Y(I)
1	1.000000	4.000000
2	2.000000	1.930000
3	1.000000	3.470000
4	0.000000	3.000000
5	0.000000	3.270000
6	0.000000	4.000000
7	3.000000	3.620000
8	2.000000	3.890000
9	0.000000	3.870000
10	2.000000	4.000000
11	1.000000	3.000000
12	2.000000	3.730000
13	4.000000	4.000000
14	0.000000	3.560000
15	0.000000	3.360000
16	1.000000	3.550000
17	3.000000	3.200000
18	0.000000	3.300000
19	2.000000	3.000000
20	3.000000	2.880000
21	1.000000	3.060000
22	1.000000	3.000000
23	0.000000	3.470000
24	0.000000	3.270000
25	1.000000	3.750000
26	0.000000	3.620000
27	0.000000	3.250000
28	0.000000	3.180000
29	0.000000	2.330000
30	0.000000	3.750000
31	0.000000	3.140000
32	1.000000	3.060000
33	1.000000	3.330000
34	1.000000	3.920000
35	2.000000	3.600000
36	0.000000	3.000000
37	0.000000	3.430000
38	0.000000	2.400000
39	0.000000	4.000000
40	0.000000	2.500000
41	0.000000	4.000000
42	1.000000	3.770000
43	1.000000	4.000000
44	1.000000	3.500000
45	1.000000	3.000000
46	2.000000	3.060000
47	2.000000	4.000000
48	0.000000	3.270000
49	0.000000	3.500000
50	1.000000	3.760000

I	X (I)	Y (I)
4	0.000000	3.000000
5	0.000000	3.270000
6	0.000000	4.000000
9	0.000000	3.870000
14	0.000000	3.560000
15	0.000000	3.360000
18	0.000000	3.300000
23	0.000000	3.470000
24	0.000000	3.270000
26	0.000000	3.620000
27	0.000000	3.250000
28	0.000000	3.180000
29	0.000000	2.330000
30	0.000000	3.750000
31	0.000000	3.140000
36	0.000000	3.000000
37	0.000000	3.430000
38	0.000000	2.400000
39	0.000000	4.000000
40	0.000000	2.500000
41	0.000000	4.000000
48	0.000000	3.270000
49	0.000000	3.500000
1	1.000000	4.000000
3	1.000000	3.470000
11	1.000000	3.000000
16	1.000000	3.550000
21	1.000000	3.060000
22	1.000000	3.000000
25	1.000000	3.750000
32	1.000000	3.060000
33	1.000000	3.330000
34	1.000000	3.920000
42	1.000000	3.770000
43	1.000000	4.000000
44	1.000000	3.500000
45	1.000000	3.000000
50	1.000000	3.760000
2	2.000000	1.930000
8	2.000000	3.890000
10	2.000000	4.000000
12	2.000000	3.730000
19	2.000000	3.000000
35	2.000000	3.600000
46	2.000000	3.060000
47	2.000000	4.000000
7	3.000000	3.620000
17	3.000000	3.200000
20	3.000000	2.880000
13	4.000000	4.000000

COORD OF CTR OF GRAVITY : 0.680000 3.390995

```
A1=    0.062926 AP1=   -0.011680 A2=    0.047195 AP2=    0.027064
A1=    0.062926 AP1=   -0.011680 A2=    0.058184 AP2=    0.011216
A1=    0.060507 AP1=   -0.007769 A2=    0.058184 AP2=    0.011216
A1=    0.060507 AP1=   -0.007769 A2=    0.059556 AP2=    0.011216
A1=    0.060118 AP1=   -0.007769 A2=    0.059556 AP2=    0.011216
A1=    0.060118 AP1=   -0.007769 A2=    0.059888 AP2=    0.011216
A1=    0.060024 AP1=   -0.007769 A2=    0.059888 AP2=    0.011216
A1=    0.060024 AP1=   -0.007769 A2=    0.059968 AP2=    0.011216
A1=    0.060001 AP1=   -0.003860 A2=    0.059968 AP2=    0.011216
A1=    0.060001 AP1=   -0.003860 A2=    0.059993 AP2=    0.011216
A1=    0.060001 AP1=   -0.003860 A2=    0.059999 AP2=    0.011216
A1=    0.060001 AP1=   -0.002084 A2=    0.059999 AP2=    0.011216
A1=    0.060000 AP1=   -0.002084 A2=    0.059999 AP2=    0.011216
BETA1 =   0.06000030  SP1 = -.0020845;  BETA =   0.06000007   SP = 0.0047607;  BETA2 =   0.06000007
```

SP2 = 0.0047607

DEV. ABOUT STRAIGHT LINE THROUGH CTR OF GRAVITY

-0.338195
-0.068194
0.661805
0.531805
0.221806
0.021805
-0.038195
0.131805
-0.068194
0.281805
-0.088195
-0.158195
-1.008195
0.411805
-0.198195
-0.338195
0.091805
-0.938195
0.661805
-0.838195
0.661805
-0.068194
0.161805
0.601806
0.071806
-0.398194
0.151806
-0.338194
-0.398194
0.351806
-0.338194
-0.068194
0.521806
0.371806
0.601806
0.101806
-0.398194
0.361806
-1.528194
0.431806
0.541805
0.271805
-0.458195
0.141806
-0.398194
0.541805
0.101805
-0.318195
-0.638195
0.421805

MED. DEV. ABOUT STRAIGHT LINE THROUGH CTR OF GRAVITY : 0.081806

CRD. AT MEAN X : 3.472800

DEVIATIONS ABOUT FITTED LINE

-0.420000076
-0.145999618
 0.579999923
 0.449999809
 0.140000343
-0.600004196E-01
-0.119999885
 0.500001907E-01
-0.149999618
 0.159999809
-0.170000076
-0.239999771
 -1.09000015
 0.329999923
-0.279999733
-0.420000076
 0.100002288E-01
 -1.02000045
 0.579999923
-0.920000076
 0.579999923
-0.149999618
 0.799999237E-01
 0.520000457
-0.999927520E-02
-0.479999542
 0.700006485E-01
-0.419999122
-0.479999542
 0.270000457
-0.419999122
-0.149999618
 0.440000534
 0.290000915
 0.520000457
 0.200004577E-01
-0.479999542
 0.280000686
 -1.60999965
 0.350000381
 0.460000038
 0.189999580
-0.539999961
 0.600004196E-01
-0.479999542
 0.460000038
 0.199995040E-01
-0.400000572
-0.720000267
 0.340000152

GRD. AT ORIGIN = 3.420000

Chapter 2.

NON-PARAMETRIC FITTING OF A STRAIGHT LINE WITH x ERROR-FREE THROUGH ORIGIN

SUMMARY

There are n measurements y_i at points x_i. The errors are supposed i.i.d. with a symmetrical distribution. The slope of the straight line $\bar{y}(x) = \beta x$ is estimated by requiring that no correlation exists between the Δ_j's and the x_j's where

$$\Delta_j(\beta) = y_j - \beta x_j$$

The set of couples $\{(x_j, y_j)\}$ is obtained by adding to the original set of points their symmetricals with respect to the origin.

1. GENERAL SCOPE

Measurements y_i are obtained at points x_i, $i = 1, 2, \ldots, n$. More than one measurement at each point is allowed. The errors $\varepsilon(x)$ about the means $\bar{y}(x)$ are supposed to be independent of each other and to follow a same symmetrical distribution.

The set of data is extended by adjoining the points $(-x_i, -y_i)$ to the original set. The estimation of the slope is that β that makes the correlation between $\{\Delta_j = y_j - \beta x_j\}$ and $\{x_j\}$, $j = 1, 2, \ldots, 2n$, as small as possible. More precisely, the absolute value of the Spearman correlation coefficient (see for instance SIEGEL (1956)) is minimized. The method can be useful when the ε's are normally distributed but the occurence of outliers is possible.

2. DETAIL OF THE PROCEDURE

The method is quite similar to that used in Chapter 1 for the fitting of an unconstrained straight line.

- Take the straight line through the origin and the centre of gravity as first approximation : $\beta_0 = \bar{y}/\bar{x}$ ($\bar{x}$, $\bar{y}$: co-ordinates of the centre of gravity of the data points).
  If $|\beta_0| < .01$, take $\beta_0 = .01$ as starting value.
  If $\bar{x}$ is very small compared to the range of x_j's, take instead the straight line through the origin and (x_r, y_r). We define the latter point as the centre of gravity of the $r = [2n/3]$ first points of the set $\{(x_j, y_j)\}$ ordered according to the x-values.

- Compute $\rho_s(\beta_0)$, the Spearman correlation coefficient between

  $\{\Delta_j(\beta_0)\}$ and $\{x_j\}$.

- Modify β_0 on both sides by increments $.25\,\beta_0$ till two ρ_s of opposite sign are found.

- Find the β at which $\rho_s(\beta)$ jumps from a negative to a positive value.

- If $\rho_s(\beta)$ is zero at machine precision on an interval, find the endpoints β_U and β_L of the interval by using an halving process and take $.5(\beta_U + \beta_L)$ as the estimation of the slope.

3. PROPERTIES OF THE ESTIMATE

The estimate of the slope is unbiased.
Its efficiency when the measurements are normal seems quite good.
A simulation with 10000 samples of size 5 was done with

$$\{ x_i \} = \{ 1, 2, 3, 4, 5 \},$$

a unit slope and a standard deviation .2.
The efficiency obtained was about .90.

4. EXAMPLE

We used data from SIEGEL (1956) as given in table 1.
A slope $\beta = .6484$ was obtained. The corresponding straight line is nearly identical with that joining the origin and the centre of gravity of the observed points.

TABLE 1

Social status strivings versus authoritarianism.
(authoritarianism is here arbitrarily supposed error-free).

AUTH.	S.S.S.
82	42
98	46
87	39
40	37
116	65
113	88
111	86
83	56
85	62
126	92
106	54
117	81

BIBLIOGRAPHY

- SIEGEL, S. (1956)
  Non-parametric statistics for the behavioural sciences
  Mc-Graw Hill

- KILDEA (1981)
  Brown-Mood type median estimation for simple regression models
  Ann. Statist. 9 n° 2, 438-442, § 6

12 MEASUREMENTS Y(I) AT X(I)

I	X(I)	Y(I)
1	40.000000	37.000000
2	83.000000	56.000000
3	82.000000	42.000000
4	85.000000	62.000000
5	87.000000	38.000000
6	98.000000	46.000000
7	126.000000	92.000000
8	106.000000	54.000000
9	111.000000	86.000000
10	113.000000	58.000000
11	116.000000	65.000000
12	117.000000	81.000000

COORD OF CTR OF GRAVITY : 97.000000 62.333328

```
A1=     0.642612 AP1=    0.065217 A2=    0.803264 AP2=   -0.741738
A1=     0.642612 AP1=    0.065217 A2=    0.655595 AP2=   -0.065217
A1=     0.642612 AP1=    0.065217 A2=    0.649103 AP2=   -0.047826
A1=     0.646356 AP1=    0.018261 A2=    0.649103 AP2=   -0.047826
A1=     0.647115 AP1=    0.018261 A2=    0.649103 AP2=   -0.047826
A1=     0.647665 AP1=    0.014783 A2=    0.649103 AP2=   -0.047826
A1=     0.648004 AP1=    0.014783 A2=    0.649103 AP2=   -0.047826
A1=     0.648264 AP1=    0.014783 A2=    0.649103 AP2=   -0.047826
A1=     0.648264 AP1=    0.014783 A2=    0.648462 AP2=   -0.011304
A1=     0.648376 AP1=    0.014783 A2=    0.648462 AP2=   -0.011304
A1=     0.648376 AP1=    0.014783 A2=    0.648424 AP2=   -0.011304
A1=     0.648376 AP1=    0.014783 A2=    0.648403 AP2=   -0.011304
A1=     0.648391 AP1=    0.014783 A2=    0.648403 AP2=   -0.011304
A1=     0.648398 AP1=    0.014783 A2=    0.648403 AP2=   -0.011304
A1=     0.648400 AP1=    0.014783 A2=    0.648403 AP2=   -0.011304
BETA1 =   0.64840043  SP1 = 0.0147827;  BETA =   0.64840168   SP = 0.0147827;  BETA2 =   0.64840288   SP2 = -.0113039
```

DEV. ABOUT STRAIGHT LINE

```
-10.301392
 -5.137069
 10.214584
-14.730621
-14.027420
 14.730576
 17.543350
 17.410934
 -6.685864
 -2.182663
 11.168530
-11.063534
```

NON-PARAMETRIC FITTING - STRAIGHT LINE THROUGH ORIGIN

12 MEASUREMENTS Y(I) AT X(I)

I	X(I)	Y(I)
1	-57.000000	-25.000000
2	-14.000000	-6.000000
3	-15.000000	-20.000000
4	-12.000000	0.000000
5	-10.000000	-23.000000
6	1.000000	-16.000000
7	29.000000	30.000000
8	9.000000	-8.000000
9	14.000000	24.000000
10	16.000000	26.000000
11	19.000000	3.000000
12	20.000000	19.000000

```
COORD OF CTR OF GRAVITY :     0.000000    0.333333

A1=    0.846014 AP1=   -0.179208 A2=    0.634511 AP2=    0.232275
A1=    0.753900 AP1=   -0.093954 A2=    0.634511 AP2=    0.232275
A1=    0.753900 AP1=   -0.093954 A2=    0.719516 AP2=    0.007829
A1=    0.722161 AP1=   -0.007829 A2=    0.719516 AP2=    0.007829
A1=    0.720838 AP1=   -0.007829 A2=    0.719516 AP2=    0.007829
A1=    0.720177 AP1=   -0.007829 A2=    0.719516 AP2=    0.007829
A1=    0.720177 AP1=   -0.007829 A2=    0.719846 AP2=    0.007829
A1=    0.720011 AP1=   -0.007829 A2=    0.719846 AP2=    0.007829
A1=    0.720011 AP1=   -0.007829 A2=    0.719928 AP2=    0.007829
A1=    0.720011 AP1=   -0.007829 A2=    0.719965 AP2=    0.007829
A1=    0.720011 AP1=   -0.007829 A2=    0.719990 AP2=    0.007829
BETA1 =   0.72001100  SP1 = -.0078295;  BETA =   0.72000003   SP = 0.0000000;  BETA2 =   0.71998984   SP2 = 0.0078295

A1 =   0.72001100 SPA1 = -.0078295; BETA1 =   0.72000551 SP = -.0078295; B1 =   0.72000003 SPB1 = 0.0000000

A2 =   0.71999884 SPA2 = 0.0078295; BETA2 =   0.71999931 SP = 0.0078295; B2 =   0.72000003 SPB2 = 0.0000000

  SLOPE = 0.7200022

DEV. ABOUT STRAIGHT LINE

  16.039948
  -9.120026
  -4.600014
  10.679987
 -14.480011
  -9.200010
   4.079990
 -13.920010
   8.639992
 -15.800007
  14.479994
  16.719986
```

Chapter 3.

NON-PARAMETRIC FITTING OF A STRAIGHT LINE WITH POSITIVE SLOPE AND WITH ERROR ON BOTH x AND y

SUMMARY

Let be given n couples (x_i, y_i) of measurements with

$$x_i = \bar{x}_i + \varepsilon_i$$

$$y_i = \bar{y}_i + \delta_i$$

The distribution of the variables ε_i and δ_i are supposed independent of i and symmetrical. These errors are uncorrelated.

The slope of the straight line $\bar{y} = \alpha + \beta\bar{x}$ is estimated by requiring that β yields a zero Spearman correlation coefficient between {rank (i)} and $\{\Delta y_i\}$ where rank (i) ~ .5 [rank (x_i) + rank (y_i)]
and $\Delta y_i = y_i - \hat{\alpha} - \hat{\beta}x_i$.

The intercept at origin α is obtained by requiring that median $\{\Delta y_i\} = 0$.

The process is used only after testing that there exists a significant and positive correlation between $\{y_i\}$ and $\{x_i\}$

1. PROBLEM AND PROPOSED METHOD

As KENDALL & STUART (1961) explain in the chapter entitled "Functional and structural relationship", once the x-values are also subject to error the problem becomes a sophisticated one.

Different approaches are described by these authors and in CHAN, LAI & MAK, TAK K. (1981).

The idea is to extend a method used in Chapter 1 for fitting a function measured (with error) at given points x_i i.e. : minimizing the absolute value of the Spearman coefficient between the $\{x_i\}$ or $\{\text{rank } (x_i)\}$ and the $\{\Delta y_i\}$, the deviations from the fitted line.

The problem here is that it is not sure at all that the $\{\text{rank } (\bar{x}_i)\}$ coincide with $\{\text{rank } (x_i)\}$ and $\{\text{rank } (y_i)\}$. However, if a significant correlation exists between $\{x_i\}$ and $\{y_i\}$, it is hoped that it is possible to find a variable conveying enough information to be useful. In the ignorance of the relative merits of $\{\text{rank } (x_i)\}$ or $\{\text{rank } (y_i)\}$ to represent closely $\{\text{rank } (\bar{x}_i)\} = \{\text{rank } (\bar{y}_i)\}$, it is suggested to take their mean.

The Spearman correlation coefficient between the so obtained $\{\text{rank } (i)\}$ and $\{\Delta y_i\}$ does not depend on the additive constant α but on β only.

The multiplicative constant .5 in rank (i) is also irrelevant.

α is obtained independently by requiring that the median of $\{\Delta y_i\}$ to be zero.

As the horizontal deviations from the fitted line $\{\Delta x_i\}$ are functionally related to the vertical discrepancies $\{\Delta y_i\}$, the median of the $\{\Delta x_i\}$ is then also zero. Roughly speaking this means that half of the points are on each side of the fitted line.

2. PROCEDURE

The detailed procedure has been given in Chapter 1. It takes into account the fact that ρ_s (β) is a step function.

If ρ_s (β) is not zero on an interval, β is chosen as that value where ρ_s jumps from a negative to a positive value.

If ρ_s (β) is, at machine precision, zero on an interval β is estimated by the mean of the endpoints of the interval (obtained by an halving process).

3. SPECIAL CASE

The constrained case where the straight line passes through a fixed point can be treated by adjoining to the original set of points their symmetrical with respect to this point and then minimizing the absolute value of the Spearman correlation coefficient between the {rank (j)} and the $\{\Delta y_j\}$, j= 1, 2, ..., 2n. The index j refers to the extended set.

4. EXAMPLE

An example (Psychological data in Table 1 extracted from SIEGEL) has been treated by the present method.

For this example, the method used by THEIL is not indicated at all, as there are discrepancies between the ranks of x_i and y_i. It leads to non-sense if notwithstanding these circumstances, it is used anyway.

By ordering the x-values, one obtains an estimation β = 1.23 with the THEIL's method cited by KENDALL & STUART, whereas our method yields β = .64 which looks quite satisfactory.

The ordinate at the origin obtained by the present method is 4.55.
If the straight line is constrained to pass through origin, as if the x_i were error-free, the slope obtained is .62. (The assumption of a straight line through the origin is quite natural with the data at hands).
The straight line joining the origin to the centre of gravity has a slope .64.

TABLE 1

x_i	y_i	rank (x_i)	rank (y_i)	rank (i)
40	37	1	1	1
82	42	2	3	2.5
83	56	3	6	4.5
85	62	4	7	5.5
87	39	5	2	3.5
98	46	6	4	5
106	54	7	5	6
111	86	8	10	9
113	88	9	11	10
116	65	10	8	9
117	81	11	9	10
126	92	12	12	12

REFERENCES

- KENDALL, M.G. and STUART, A. (1961)
The advanced theory of statistics
Vol. 2 Chapter 29
Charles Griffin and Cy

- CHAN, LAI K. and MAK, TAK K. (1981)
Two adaptative methods for the estimation of a linear structural relationship
Scand. J. Statist. 9, 223-228

APPENDIX to Chapter 3

NON-PARAMETRIC LINEAR FITTING

10 MEASUREMENTS

I	X(I)	Y(I)
1	75.000000	80.000000
2	70.000000	87.000000
3	60.000000	91.000000
4	55.000000	44.000000
5	50.000000	22.000000
6	40.000000	58.000000
7	25.000000	52.000000
8	20.000000	10.000000
9	15.000000	38.000000
10	10.000000	18.000000

I	X(I)	Y(I)
10	10.000000	18.000000
9	15.000000	38.000000
8	20.000000	10.000000
7	25.000000	52.000000
6	40.000000	58.000000
5	50.000000	22.000000
4	55.000000	44.000000
3	60.000000	91.000000
2	70.000000	87.000000
1	75.000000	80.000000

```
COORD OF CTR OF GRAVITY :     42.000000   50.000000

A1=     1.037037 AP1=     0.090909 A2=     0.777778 AP2=    -0.267478
A1=     0.971273 AP1=     0.042424 A2=     0.777778 AP2=    -0.267478
A1=     0.971273 AP1=     0.042424 A2=     0.944764 AP2=    -0.115151
A1=     0.971273 AP1=     0.042424 A2=     0.964141 AP2=    -0.006061
A1=     0.971273 AP1=     0.042424 A2=     0.965032 AP2=    -0.006061
A1=     0.971273 AP1=     0.042424 A2=     0.965812 AP2=    -0.006061
A1=     0.971273 AP1=     0.042424 A2=     0.966455 AP2=    -0.006061
A1=     0.967092 AP1=     0.042424 A2=     0.966455 AP2=    -0.006061
A1=     0.967092 AP1=     0.042424 A2=     0.966569 AP2=    -0.006061
A1=     0.967092 AP1=     0.042424 A2=     0.966635 AP2=    -0.006061
A1=     0.966692 AP1=     0.042424 A2=     0.966635 AP2=    -0.006061
A1=     0.966692 AP1=     0.042424 A2=     0.966642 AP2=    -0.006061
A1=     0.966692 AP1=     0.042424 A2=     0.966648 AP2=    -0.006061
A1=     0.966692 AP1=     0.042424 A2=     0.966653 AP2=    -0.006061
A1=     0.966692 AP1=     0.042424 A2=     0.966658 AP2=    -0.006061
A1=     0.966692 AP1=     0.042424 A2=     0.966662 AP2=    -0.006061
A1=     0.966692 AP1=     0.042424 A2=     0.966666 AP2=    -0.006061
BETA1 =    0.96666873   SP1 = 0.0424243;  BETA =    0.96666873   SP = 0.0424243;  BETA2 =    0.96666557   SP2 = -.0060606

DEV. ABOUT STRAIGHT LINE THROUGH CTR OF GRAVITY

  -1.066605
  14.100052
 -18.733292
  18.433365
   9.933350
 -35.733337
 -18.566681
  23.599976
   9.933289
  -1.900055


MED. DEV. ABOUT STRAIGHT LINE THROUGH CTR OF GRAVITY :     4.433342

   ORD. AT MEAN X :     54.433334

Y-DEVIATIONS ABOUT FITTED LINE

 -5.49993896
  9.66671753
 -23.1666260
  14.0000305
  5.50001526
 -40.1666718
 -23.0000153
  19.1666412
  5.49995422
 -6.33338928


   ORD. AT ORIGIN =     13.833252
```

20 AUG. 1985