RESEARCH REPORT

DIGRESSION ANALYSIS

Fitting alternative regression models to heterogeneous data

by

Seppo Mustonen

No.12 May 1978

DEPARTMENT OF STATISTICS
UNIVERSITY OF HELSINKI
SF 00100 HELSINKI 10 FINLAND

DIGRESSION ANALYSIS

Fitting alternative regression models to heterogeneous data

by

Seppo Mustonen

No.12 May 1978

To appear in the Proceedings of Compstat 1978, the third symposium on computational statistics, Leiden 1978

ISBN 951-45-1412-2

Digression Analysis: Fitting Alternative Regression Models to Heterogeneous Data

S. Mustonen, Helsinki

SUMMARY

A selective least squares method for parameter estimation of alternative regression models in heterogeneous data is introduced. A test for heterogeneity of the sample is described. Computational aspects of the method are discussed and some test results are presented.

KEYWORDS switching regression, cluster analysis, least squares, heterogeneous data.

1. Introduction

Heterogeneous data appearing in a statistical investigation is seldom a pleasant surprise when a neat unimodal sample was expected. Heterogeneity proves that the observations have been affected by some unrecorded factors. If these disturbances cannot be measured afterwards heterogeneity is bound to remain in the data and it may hamper all the statistical analyses.

In this paper the problem of heterogeneity will be considered in connection with linear and nonlinear regression analysis. A method will be proposed which allows dealing with regression models despite heterogeneous data.

As an illustrative simple example, let us imagine an ambiguous situation in the ordinary linear model $y=\alpha x+\beta+\varepsilon$ where α,β are unknown parameters and ε is the error term. Assume that our data to be used for parameter estimation contains a considerable portion of exceptional observations having a different β parameter, say γ. Thus the important trend parameter α is supposed to be the same for all observations. In Fig.1 such a heterogeneous sample with $\alpha=1$, $\beta=-0.6$, $\gamma=1$ and $\varepsilon\sim N(0,0.5^2)$ is displayed. 50 observations of each kind have been generated. All the simulations, computations and graphic presentations in this paper have been carried out with the statistical data processing system SURVO 76 (Mustonen,1977).

It is disastrous to fit an ordinary linear model to this data. Although α is same for the both species of observations the common least squares estimate of this trend parameter is 0.736.

The method to be considered in this paper, digression analysis, will give without any prior information on the origin of each observation the estimates $\alpha=0.981$, $\beta=-0.632$ and $\gamma=0.968$. Simultaneously with the estimation the observations will also be classified into two groups which in this case agree well with the original partition.

Fig.1a A heterogeneous sample

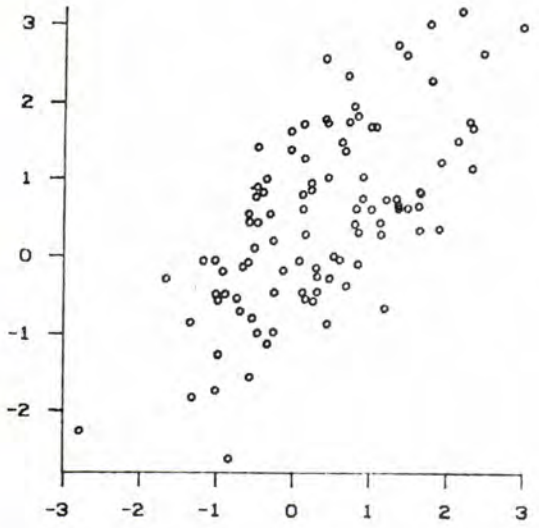
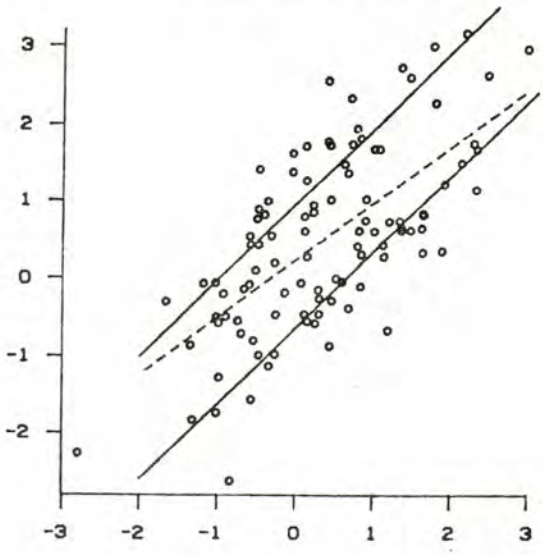


Fig.1b The regression line --- and the digression lines



2. Principle of digression analysis

In digression analysis the properties of regression and cluster analysis are combined. Let us have n observations

$$y_j,\ x_{1j},x_{2j},\ldots,x_{mj},\quad j=1,2,\ldots,n$$

on the variables $y,x_1,x_2,\ldots,x_m$ and assume that these observations are divided in an unknown way into two groups (restriction to two is not essential), and in group i (i=1,2) we have

$$E(y)=f_i(x_1,x_2,\ldots,x_m,\alpha_i)$$

where the form of the regression function f_i is known but possibly different for each group, and α_i is the vector of parameters to be estimated.
Our task is to estimate the parameters α_1,α_2 simultaneously without any previous information on the type of any single observation.

A normal procedure in this situation is to classify the observations using some general clustering algorithm and thereafter the parameters will be estimated by means of regression analysis. This is not, however, efficient since classification and estimation are to be carried out independently of each other.

Simultaneous classification and estimation has occurred at least in econometrics when considering discontinuous parameter changes in time series (Goldfeld and Quandt,1976). In these switching regression problems additional information in a form of a time variable or other extraneous variables is usually available, and it naturally simplifies the problem.

In our approach the parameters are estimated by generalizing the least squares (OLS) criterion

$$\sum_{j=1}^{n} (y_j - f(x_{1j},\ldots,x_{mj},\alpha))^2 = \min_{\alpha}$$

to a selective form

$$\sum_{j=1}^{n} \min\{(y_j - f_1(x_{1j},\ldots,x_{mj},\alpha_1))^2, (y_j - f_2(x_{1j},\ldots,x_{mj},\alpha_2))^2\} \qquad (1)$$

and this selective least squares (SLS) criterion is to be minimized with respect to α_1,α_2.
Thus each observation will be attributed to the nearest regression curve and the parameters of each submodel f_1,f_2 are determined only by observation points of its own.
We are in fact applying the nearest-mean classification rule (Fukunaga,1972,p.332) in a generalized form.

The digression model defined above will be notated by

$$E(y) = \begin{cases} f_1(x_1,x_2,\ldots,x_m,\alpha_1) \\ f_2(x_1,x_2,\ldots,x_m,\alpha_2). \end{cases} \qquad (2)$$

This set up and the corresponding SLS criterion can be easily extended to more than two submodels. The introductory example shows that the submodels may have common parameters. In many potential applications it is natural to expect that the submodels have a similar form and they differ from each other only by a few parameters.

At first it may seem rather odd that this simple selective criterion can work in practice when the subgroups are not distinct but partially overlapped. Then many serious misclassifications occur which may disturb parameter estimation. Let us notice, however, that the misclassifications will usually fall on neutral observations located "between" the regression curves and having no major influence on estimation.
Of course, the estimates obtained by the SLS method may be biased particularly if the data is only slightly heterogeneous and the subgroups are strongly overlapped. This digression bias diminishes rapidly when the heterogeneity grows. The magnitude of the bias depends also on the nature of the parameter. For instance, in the digression model

$$E(y) = \begin{cases} \alpha x+\beta \\ \alpha x+\gamma \end{cases}$$

the location parameters β and γ may have strongly biased estimates while the trend parameter α is always almost unbiased.
In some cases the digression bias can be estimated as well and its detrimental effects may be reduced (Mustonen,1976).

Passing from OLS to SLS means a considerable increase of computational work even in the case of linear submodels f_1,f_2 since minimization of the selective criterion (1) is a nonlinear optimization problem to be solved by iterative methods. It is also difficult to study the SLS principle theoretically, and this may be the reason why it has been ignored.

3. Testing for heterogeneity

In digression analysis it is necessary to make sure that the data in question is really heterogeneous in the intended manner. One possibility to do this in practice is to fit an ordinary regression model

$$E(y)=f(x_1,x_2,\ldots,x_m,\alpha) \tag{3}$$

and compare the residual sums of squares in (2) and (3). Let these sums be S_D and S_R, respectively. If $f=f_1$ we have $S_D \leq S_R$ since the digression model is always more flexible than its submodels. As a natural test criterion for heterogeneity the ratio S_D/S_R may be used. Now the crucial question is whether the value of this ratio is small enough to indicate that the digression hypothesis (2) is to be preferred to a plain regression hypothesis (3).

We shall study the behaviour of the statistic S_D/S_R in the simple digression model

$$y=\begin{cases}\mu_1+\varepsilon_1, & \varepsilon_1 \sim N(0,\sigma_1^2)\\ \mu_2+\varepsilon_2, & \varepsilon_2 \sim N(0,\sigma_2^2)\end{cases} \tag{4}$$

which is equivalent to dissection of a mixture of two normal distributions. Assume that the components of this mixture are presented in our data in the proportions p_1,p_2.

A theoretical counterpart for SLS estimation of this model is to minimize the expected value

$$E(\min\{(y-\lambda_1)^2,(y-\lambda_2)^2\}) \tag{5}$$

with respect to λ_1,λ_2. The minimum of (5) is denoted by σ_D^2.

In general, it is not possible to write λ_1,λ_2 and σ_D^2 in a closed form. In Table 1 these values are listed for some combinations of $\mu_1=-\mu_2=\mu$, p_2/p_1 and σ_2 ($\sigma_1=1$).

Table 1

μ	σ_2	p_2/p_1	λ_1	λ_2	σ_D^2	σ_D^2/σ_R^2
0.0	1.0	1.0	0.7978	-0.7978	0.3634	0.3634
0.5	1.0	1.0	0.8955	-0.8955	0.4491	0.3593
1.0	1.0	1.0	1.1666	-1.1666	0.6389	0.3195
1.5	1.0	1.0	1.5586	-1.5586	0.8207	0.2525
2.0	1.0	1.0	2.0169	-2.0169	0.9317	0.1864
2.5	1.0	1.0	2.5040	-2.5040	0.9799	0.1352
3.0	1.0	1.0	3.0007	-3.0007	0.9954	0.0995
0.0	0.8	0.8	0.7269	-0.7269	0.3115	0.3709
0.0	0.8	0.6	0.7380	-0.7380	0.3202	0.3703
0.0	0.6	0.8	0.6560	-0.6560	0.2851	0.3985
0.0	0.6	0.6	0.6782	-0.6782	0.3000	0.3948
1.0	0.8	0.8	1.2974	-0.9331	0.5532	0.3027
1.0	0.8	0.6	1.3402	-0.8941	0.5551	0.3080
1.0	0.6	0.8	1.3723	-0.8709	0.4645	0.2728
1.0	0.6	0.6	1.3968	-0.8144	0.4766	0.2808
2.0	0.8	0.8	2.0531	-1.9602	0.7946	0.1659
2.0	0.8	0.6	2.0566	-1.9410	0.8150	0.1766
2.0	0.6	0.8	2.0615	-1.9478	0.6747	0.1446
2.0	0.6	0.6	2.0629	-1.9289	0.7136	0.1582

In the symmetric case $\sigma_1=\sigma_2=\sigma$, $p_1=p_2$, $\mu_1=-\mu_2=\mu$ we have

$$\lambda_1=-\lambda_2=\lambda=(2\Phi(\mu/\sigma)-1)\mu+2\phi(\mu/\sigma)\sigma$$

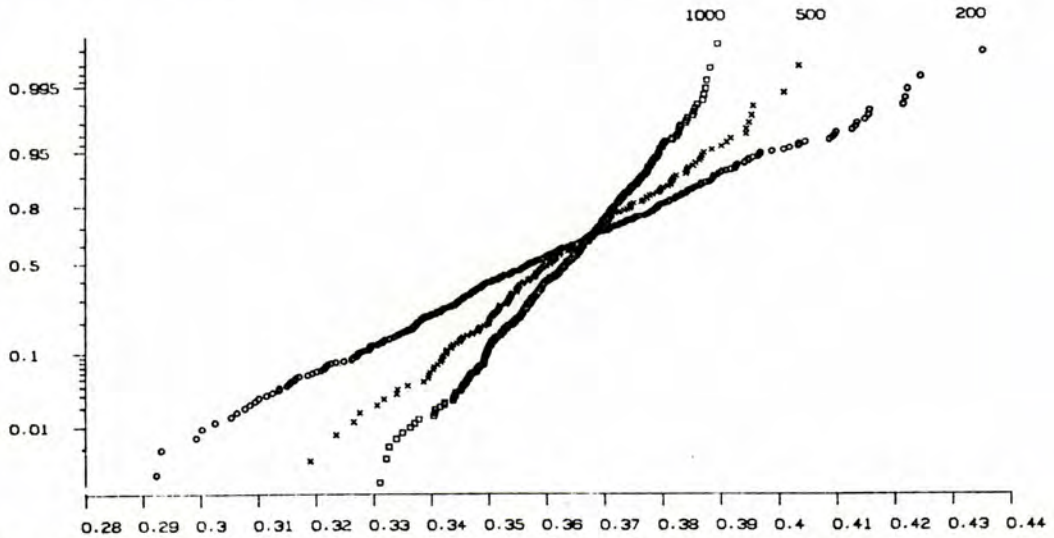
and $\sigma_D^2=\sigma^2+\mu^2-\lambda^2$, $\sigma_R^2=\sigma^2+\mu^2$. In particular, if the distribution is homogeneous then $\lambda=\sqrt{2/\pi}\sigma$ and $\sigma_D^2/\sigma_R^2=\kappa=1-2/\pi=0.36338$. The ratio σ_D^2/σ_R^2 is clearly a counterpart for S_D/S_R in the model (4) and S_D/S_R tends to κ as the sample size grows if the sample is from a homogeneous normal distribution. It is also obvious that for a homogeneous normal sample the asymptotic distribution of S_D/S_R is normal. To evaluate the mean and variance of this asymptotic distribution, a series of simulation experiments have been carried out. The main results are given in Table 2.

Table 2

number of replicates	n	mean of S_D/S_R	std.dev. of S_D/S_R	skewness	kurtosis
200	50	0.3452	0.0456	-0.07	3.03
1320	100	0.3543	0.0327	0.18	3.19
385	200	0.3587	0.0251	0.12	2.97
210	500	0.3620	0.0156	0.13	2.88
578	1000	0.3633	0.0105	-0.15	2.83

The results indicate that $E(S_D/S_R)$ tends to κ and $Var(S_D/S_R)$ is approximately $0.115/n$.
In fact S_R and S_D/S_R seem to be asymptotically independent variables. In this case it can be shown that the asymptotic variance of S_D/S_R is $8(1-3/\pi)/(\pi n)$. Several tests have been performed for various n showing no significant departure from normality.
In Fig.2 sample distributions of S_D/S_R have been plotted on normal probability paper for n=200,500 and 1000.

Fig.2



The test, based on the asymptotic distribution of S_D/S_R, can also be used in the more general situation of our introductory example where the digression model can be rewritten in the form

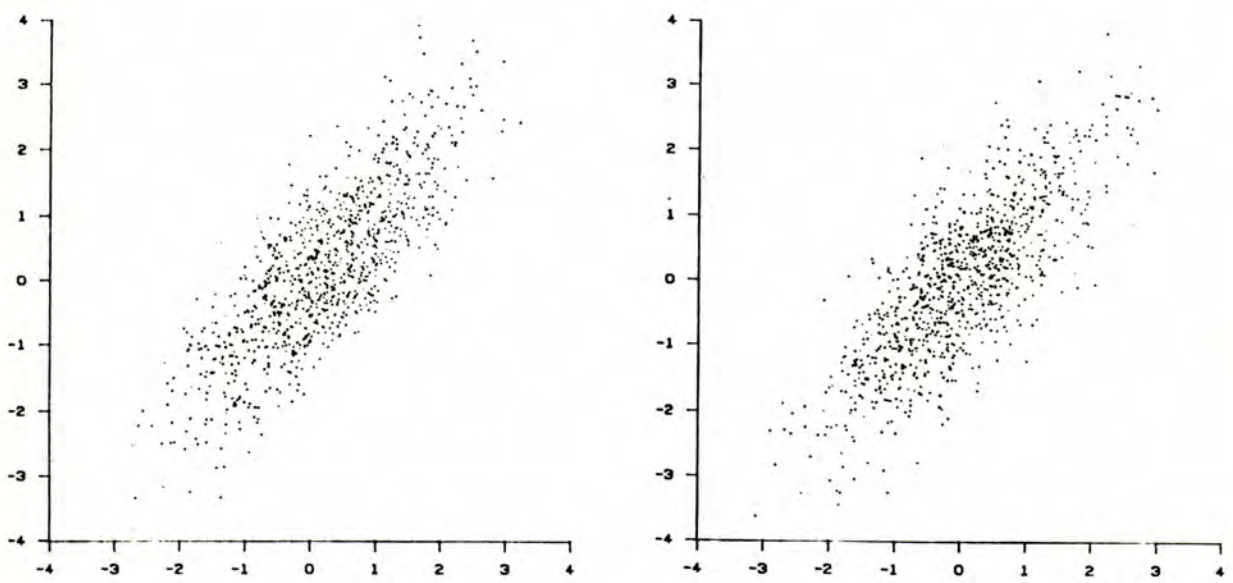
$$y-\alpha x=\begin{cases}\beta+\varepsilon_1\\ \gamma+\varepsilon_2\end{cases}$$

which is same as (4) if α is known. Thus for large n, S_D/S_R has the same asymptotic properties for both models. Also simulations with homogeneous samples for variables x,y support this assertion.

In the introductory example $S_D/S_R=0.248$ and if the data were homogeneous S_D/S_R would be approximately $N(0.35,0.034^2)$. Thus our test indicates that the sample in Fig.1 is very heterogeneous.

In Fig.3 two samples of 1000 observations of the same type are displayed. The first is heterogeneous with $\alpha=1$, $\beta=-\gamma=0.5$ and $\varepsilon_1,\varepsilon_2\sim N(0,0.5^2)$ but the second sample is homogeneous with $\alpha=0.9$, $\beta=\gamma=0$ and $\varepsilon\sim N(0,0.7^2)$. In this case it is rather difficult to detect the heterogeneity by eye. Here, however, S_D/S_R acquires the value 0.328 for the first sample and 0.366 for the second. In this case S_D/S_R should be $N(0.363,0.0107^2)$ for a homogeneous sample. Thus the homogeneity of the first sample would be rejected.

Fig.3



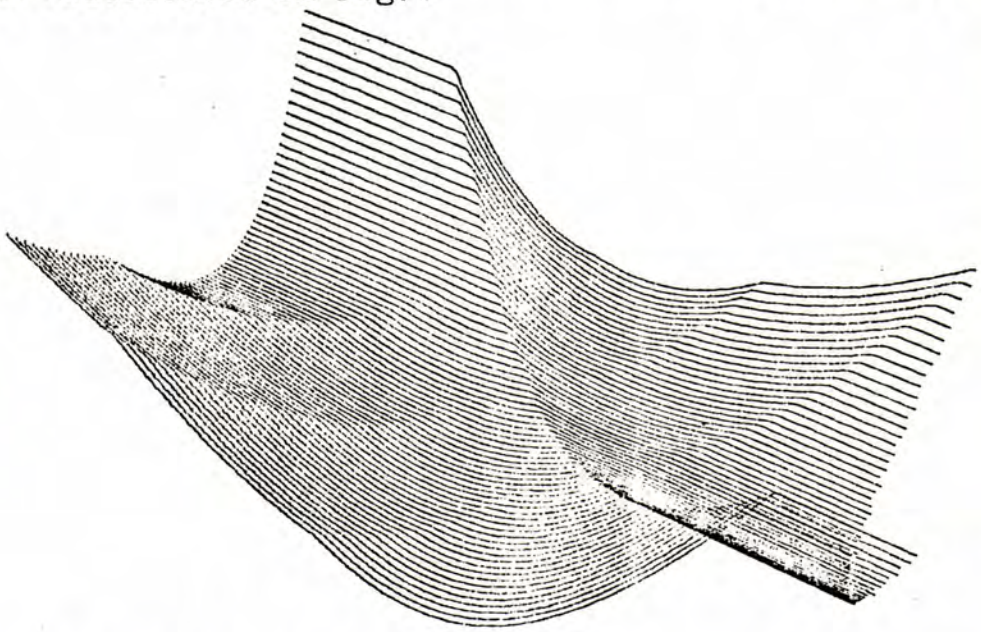
4. Computational aspects

In estimation of digression models we have to minimize the SLS criterion (1). The nature of this problem is illustrated by the following example.

Consider the simple digression model (4). In this case application of the SLS principle means partitioning of observations $y_j, j=1,\ldots,n$ into disjoint subsets $\{y_j, j\in J_1\},\{y_j, j\in J_2\}$ and selecting the values of μ_1,μ_2 so that

$$\sum_{j\in J_1}(y_j-\mu_1)^2+\sum_{j\in J_2}(y_j-\mu_2)^2$$

is minimized. For each fixed partition J_1, J_2 the minimum is attained when μ_i is the arithmetic mean of y values of J_i, i=1,2. Thus by comparing the minima obtained in various partitions the solution will be found. By sorting the observations only a few comparisons are necessary in practice.
This simple procedure can also be applied to some other digression models, but in general, the solution of the estimation problem must be based on some general iterative algorithm. The task is not easy because the objective function (1) may have unpleasant cusps as illustrated in Fig.4



It is important that the algorithm can detect the broad outlines of the function in spite of the cusps. In practice we have experimented with the variable metric method with numerical derivatives and the method of Hooke and Jeeves (1961). The latter method, which is a direct search algorithm, has proved to be reliable in various linear and nonlinear digression models.

References

Fukunaga K., Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972

Goldfeld S.M. and Quandt R.E., Studies in Nonlinear Estimation, Ballinger, New York, 1976

Hooke R. and Jeeves T.A., "Direct search" solution of numerical and statistical problems, J.of the Assn.for Computing Machinery,8,212-229,1961

Mustonen S., Digression analysis, Research Report No.2, Department of Statistics, University of Helsinki, 1976(in Finnish)

Mustonen S., SURVO 76: A statistical data processing system (for Wang 2200), Research Report No.6, Department of Statistics, University of Helsinki, 1977