[redacted]

DIGRESSION ANALYSIS

Digression analysis is a method for clustering of observations and for estimation of regression models in heterogeneous data. Digression analysis can be considered as a generalization of normal regression analysis*; the ordinary least-squares method is replaced by a *selective least-squares* (SLS) *method**. Digression analysis is also closely related to *switching regression**. In digression analysis, however, no extraneous information, as time or other support variables, is used for classification of observations.

Let there be n observations $y_j, x_{1j}, \dots, x_{mj}$, $j = 1, \dots, n$, on the variables $Y, X_1, \dots, X_m$ and assume that these observations belong in an unknown way to two (or more) groups. In group i ($i = 1, 2$) we have $E(Y \mid x_1, \dots, x_m) = f_i(x_1, \dots, x_m, \alpha_i)$, where the form of the regression function f_i is known but possibly different for each group and α_i is the vector of the parameters.

The parameters are estimated by generalizing the least-squares criterion to a selective form

$$S(\alpha_1, \alpha_2) = \sum_{j=1}^{n} \min\left\{ (y_j - f_1(x_{1j}, \dots, x_{mj}, \alpha_1))^2, \right.$$
$$\left. (y_j - f_2(x_{1j}, \dots, x_{mj}, \alpha_2))^2 \right\}$$

and this SLS criterion is minimized with respect to α_1, α_2. Thus each observation will be attributed to the nearest regression curve. This set up and the SLS criterion can be extended to more than two submodels. In applications it is natural to expect that the submodels have a similar form and the parameters may also be partially common.

The estimates obtained by the SLS method may be biased particularly if the subgroups overlap strongly. This digression bias diminishes rapidly when the heterogeneity increases. The magnitude of the bias depends also on the nature of the parameter.

Minimization of the SLS criterion is usually a nonlinear optimization problem that must be solved by iterative methods. It is also difficult to study the SLS principle theoretically.

In digression analysis it is necessary to check that the data are really heterogeneous in the intended manner. One possibility to do this is to fit also an ordinary regression model $E(Y) = f(x_1, \ldots, x_m, \alpha)$ and compare the residual sum of squares S_R of regression analysis and $S_D = \min S(\alpha_1, \alpha_2)$ of digression analysis. If $f = f_1$, we have $S_D \leqslant S_R$, since the digression model is always more flexible than its submodels. Typically, $2S_D < S_R$ even in homogeneous samples.

Testing for heterogeneity may be based on the ratio S_D/S_R, which is asymptotically normal in homogeneous samples when the error terms are normal. In the decomposition of two univariate normal distributions S_D/S_R is asymptotically

$$N(1 - 2/\pi, 8(1 - 3/\pi)/(\pi n))$$

when the population is homogeneous, i.e., $N(\mu, \sigma^2)$. For $n \geqslant 8$ a good approximation for the mean is $1 - 2\pi^{-1} - n^{-1}$.

In Fig. 1, an artificial heterogeneous sample of 200 observations

$$Y = \begin{cases} \sigma_1 x + \beta_1 + \epsilon_1 & \text{with probability } p_1 \\ \alpha_2 x + \beta_2 + \epsilon_2 & \text{with probability } p_2 \end{cases}$$

$[\alpha_1 = 1.2,\ \beta_1 = 1,\ \epsilon_2 \sim N(0, 0.8^2),\ p_1 = 0.4,\ \alpha_2 = 0.5,\ \beta_2 = 2.5,\ \epsilon_2 \sim N(0, 0.4^2),\ p_2 = 0.6]$ is displayed. In Fig. 2 the estimated digression lines $y = 1.344x + 0.507$ and $y = 0.504x + 2.525$ are plotted together with the theoretical digression lines (dashed). By simulation it has been found that in this case the digression estimates have the following properties:

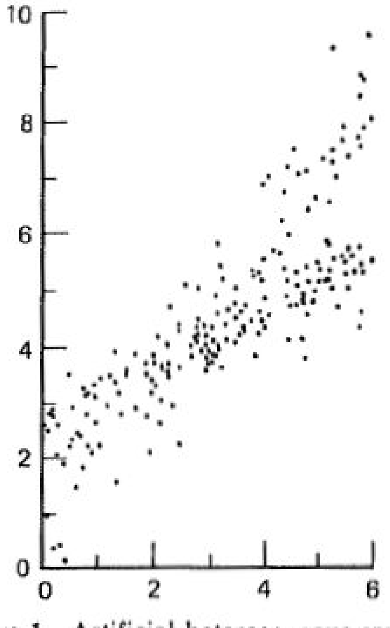

Figure 1 Artificial heterogeneous sample.

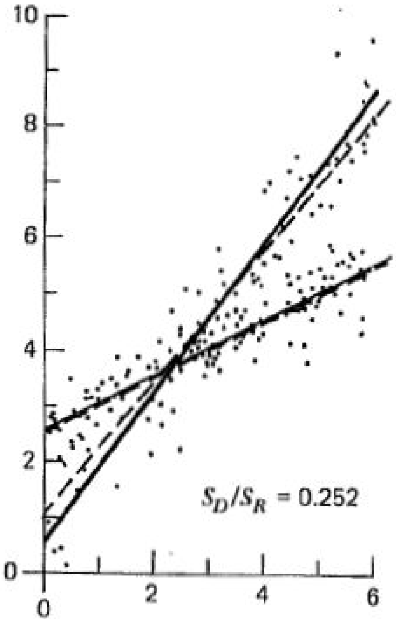

Figure 2 Estimated and theoretical digression lines.

Parameter	Mean of Estimate	Bias	Mean Square Error
$\alpha_1 : 1.2$	1.324	0.124	0.0226
$\beta_1 : 1$	0.645	− 0.355	0.2180
$\alpha_2 : 0.5$	0.497	− 0.003	0.0008
$\beta_2 : 2.5$	2.532	0.032	0.0133

Bibliography

Mustonen, S. In *COMPSTAT 1978, Proc. Comp. Statist.* Physica-Verlag, Vienna, 1978, pp. 95–101.

(HOMOGENEITY, TESTS OF
MIXTURES
REGRESSION ANALYSIS
SELECTIVE LEAST SQUARES
SWITCHING REGRESSION)

SEPPO MUSTONEN

[redacted]