EXERCISE : use the data CLIMATE.MAT

Analyse a multivariate regression model, having all the pollutants as response variables,

$$\underline{Y} = \begin{bmatrix} PM10 \\ O2 \\ N \\ CO \end{bmatrix}$$

and some climatic variable as regressor

$$\underline{X} = \begin{bmatrix} \text{mean } T \\ \text{Humidity} \\ \text{mean Wind speed} \\ \text{Rain. mm} \end{bmatrix}$$

Test the significance of the overall regression and then test if some regressor can be eliminated from the model.

(UNIVARIATE) NONLINEAR REGRESSION [Draper - Smith, Applied Regression Analysis, 1981]

Consider models which are nonlinear in the parameters

Examples:

1) $Y = \exp(\theta_1 + \theta_2 t^2 + \varepsilon)$

2) $Y = \frac{\theta_1}{\theta_1 - \theta_2}\left[e^{-\theta_2 t} - e^{-\theta_1 t}\right] + \varepsilon$

$\underline{\theta} = (\theta_1, \theta_2)$ parameters

t = regressor

Y = response $\in \mathbb{R}^1$

Note that:

1) Is <u>intrinsically linear</u>, since $\log Y = \theta_1 + \theta_2 t^2 + \varepsilon$ is linear

2) Is <u>intrinsically nonlinear</u>, since no transformations may reduce 2) to the linear case

We will deal with models like 2).

LEAST SQUARES

Suppose that the postulated model is of the form

$$Y = f(\xi_1, \xi_2, \dots, \xi_k; \theta_1, \theta_2, \dots, \theta_p) + \varepsilon \qquad (A)$$

with

$$\underline{\xi} = (\xi_1, \xi_2, \dots, \xi_k)^T \quad \text{REGRESSORS (before called } X\text{'s)}$$

$$\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^T \quad \text{PARAMETERS}$$

$\Rightarrow$ (A) is equivalent to

$$Y = f(\underline{\xi}, \underline{\theta}) + \varepsilon \qquad (B1)$$

with the additional assumptions

$$\begin{cases} E(\varepsilon) = 0 \\ \text{Var}(\varepsilon) = \sigma^2 \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \qquad (B2)$$

If we now have a joint sample

$$(y_u, \xi_{1u}, \xi_{2u}, \dots, \xi_{ku}) \qquad u = 1, 2, \dots, n$$

we can write n equations like (B1) - (B2) $\forall u = 1, \dots, n$

$$(B3) \quad \begin{cases} y_u = f(\xi_u, \underline{\theta}) + \varepsilon_u & u = 1, \dots, n \\ \underline{\varepsilon} = N(\underline{0}, \sigma^2 I_n) \end{cases}$$

with $\underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$.

A <u>least squares estimate</u> $\hat{\underline{\theta}}$ of $\underline{\theta}$, will minimise the error sum of squares:

$$S(\underline{\theta}) = \sum_{u=1}^{n} \varepsilon_u^2 = \sum_{u=1}^{n} (y_u - f(\xi_u, \underline{\theta}))^2$$

Also in the nonlinear case the LS estimators of $\underline{\theta}$ are also ML estimators, under the assumption of normality of errors.

In the linear case we found $\hat{\underline{\theta}}$ by using the normal equations, which can be found by differentiating $S(\underline{\theta})$ with respect to θ_i:

$$\frac{\partial S(\underline{\theta})}{\partial \theta_i} = \sum_{\mu=1}^{n} \left(Y_\mu - f(\xi_\mu, \underline{\theta}) \right) \left[\frac{\partial f(\xi_\mu, \underline{\theta})}{\partial \theta_i} \right] = 0 \qquad i = 1, \dots, p \qquad (N)$$

But in the linear case, where f had the form e.g. of the type

$$f(\xi_\mu, \underline{\theta}) = \theta_1 \xi_{1\mu} + \theta_2 \xi_{2\mu} + \dots + \theta_p \xi_{p\mu}$$

we had $\dfrac{\partial f(\xi_\mu, \underline{\theta})}{\partial \theta_i} = \xi_{i\mu}$ <u>independent from</u> $\underline{\theta}$

Now this is not any more true.

<u>Example</u>: $f(\xi_\mu, \theta) = e^{-\theta \xi_\mu} \Rightarrow \dfrac{\partial f}{\partial \theta} = -\xi_\mu e^{-\theta \xi_\mu}$

i.e. $Y_\mu = e^{-\theta \xi_\mu} + \varepsilon_\mu$

and the normal equations become

$$\sum_{i=1}^{n} \left[y_i - e^{-\theta \xi_i} \right] \left[-\xi_i e^{-\theta \xi_i} \right] = 0$$

which can not easily be solved with respect to θ.

The minimum of $S(\theta)$ can be found via different iterative algorithms.

Here we will introduce some of these methods.

METHOD 1: LOCAL LINEARIZATION (GAUSS-NEWTON METHOD)

Let $\underline{\hat{\theta}}_0$ be an initial guess for $\underline{\hat{\theta}}$, $\underline{\hat{\theta}}_0 = (\theta_{10}, \theta_{20}, \dots, \theta_{p0})^T$

We make a Taylor series expansion of $f(\xi_u, \underline{\theta})$ about $\underline{\theta}_0$ and we approximate

$$f(\xi_u, \underline{\theta}) \simeq f(\xi_u, \underline{\theta}_0) + \sum_{i=1}^{p} \left[\frac{\partial f(\xi_u, \underline{\theta})}{\partial \theta_i} \right]_{\underline{\theta} = \underline{\theta}_0} (\theta_i - \theta_{i0})$$

Now setting $f_u^0 := f(\xi_u, \underline{\theta}_0)$

$$\beta_i^0 := \theta_i - \theta_{i0}$$

$$z_{ui}^0 := \left[\frac{\partial f(\xi_u, \underline{\theta})}{\partial \theta_i} \right]_{\underline{\theta} = \theta_0}$$

the model (B3) is approximated by

$$(y_u - f_u^0) = \sum_{i=1}^{p} \beta_i^0 z_{ui}^0 + \varepsilon_u \qquad u = 1, \dots, n$$

↑ LINEAR IN THE β_i'S !!

Now by defining

$$\underset{\sim}{Z}_0 = \begin{bmatrix} z_{11}^0 & z_{12}^0 & \cdots & z_{1p}^0 \\ \vdots & & & \\ z_{n1}^0 & z_{n2}^0 & \cdots & z_{np}^0 \end{bmatrix} = \{ z_{ui}^0 \} \qquad n \times p$$

$$\underline{\beta}^0 = \begin{bmatrix} \beta_2^0 \\ \vdots \\ \beta_p^0 \end{bmatrix} = \begin{bmatrix} \theta_1 - \theta_{10} \\ \vdots \\ \theta_p - \theta_{p0} \end{bmatrix} \qquad \text{and} \qquad \underline{Y}_0 = \begin{bmatrix} Y_1 - d_1^0 \\ \vdots \\ Y_n - d_n^0 \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

the model becomes, with matrix notation

$$\begin{cases} \underline{Y}_0 = \underset{\sim}{Z}_0 \underline{\beta}^0 + \underline{\varepsilon} \\ \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n) \end{cases} \qquad \text{LINEAR MODEL}$$

and the LS-estimator of $\underline{\beta}^0$ is $\hat{\underline{\beta}}^0 = S_0^{-1} \underset{\sim}{Z}_0 \underline{Y}_0$ with $S_0 = \underset{\sim}{Z}_0^T \underset{\sim}{Z}_0$, assuming it is invertible.

From $\hat{\beta}_0$ we can recover an update of the estimate of $\underline{\theta}$:

$$\underline{\hat{\theta}}_1 = \underline{\theta}_0 + \hat{\beta}_0$$

Then we use $\underline{\hat{\theta}}_1$ as initial guess and iterate the procedure:

$$\Rightarrow \quad \underline{\hat{\theta}}_{j+1} = \underline{\hat{\theta}}_j + \hat{\beta}_j = \underline{\hat{\theta}}_j + S_j^{-1} \underset{\sim}{Z}_j \left(\underline{Y} - \underline{f}^j \right)$$

($\underline{f}^j = \underline{f}(\xi, \underline{\hat{\theta}}_j)$)

($\underset{\sim}{Z}_j$: matrix of first derivatives of f computed in $\underline{\hat{\theta}}_j$)

up to when

$$\left| \left\{ \theta_{i(j+1)} - \theta_{ij} \right\} / \theta_{ij} \right| < \delta \qquad \forall i = 1, \dots, p$$

(δ: fixed tolerance)

At each iteration, $S(\underline{\hat{\theta}}_j)$ can be computed to control if a reduction in its value has been achieved.

Problems of this method:

1) It may converge very slowly

2) $S(\hat{\underline{\theta}}_j)$ may oscillate very widely, even if at some point it stabilizes

3) It may not converge at all, and even diverge $\Rightarrow$
In this case, reparametrize the model or use the Marquardt's compromise

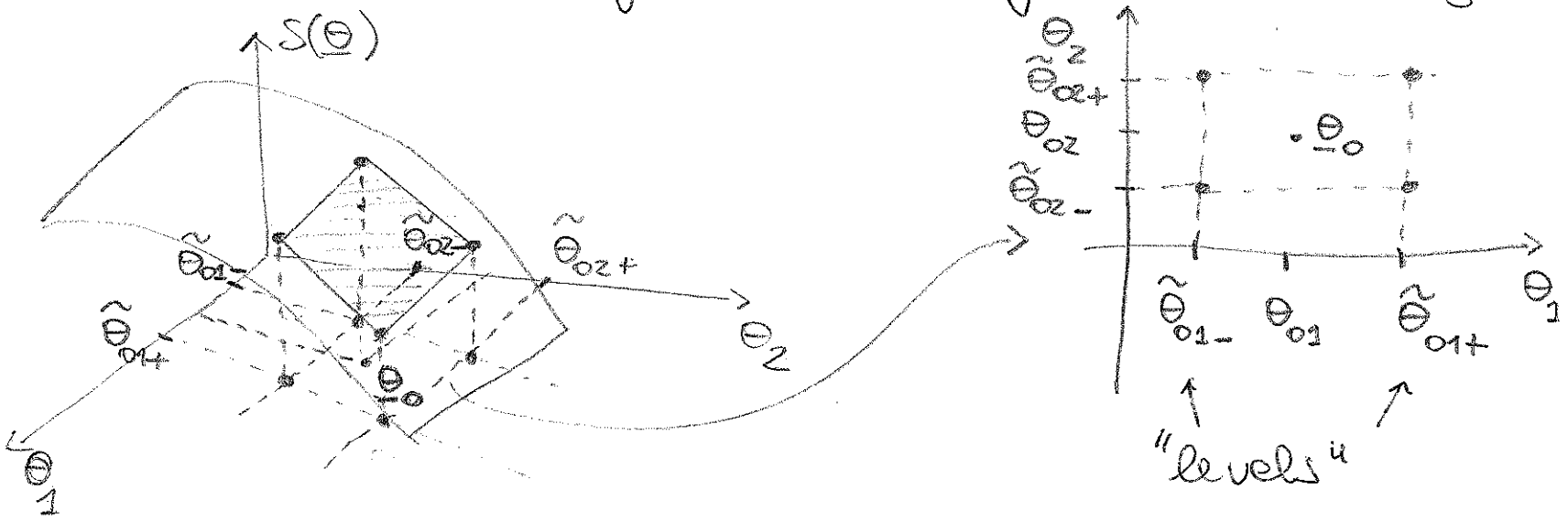
4) It may converge to a local minimum of $S(\underline{\theta})$ or to a maximum $\Rightarrow$ test many and widely spread initial guesses $\underline{\theta}_0$!!

METHOD 2: OF STEEPEST DESCENT

Concentrate on $S(\underline{\theta})$.

Start from an initial point (guess) $\underline{\theta}_0$ then

- fix an experimental factorial design around $\underline{\theta}_0$



- Compute $S(\underline{\theta})$ in the points of the experimental design
- approximate $S(\underline{\theta}_0)$ locally with the hyperplane passing through the points of the design:

$$S(\underline{\theta}) \simeq \beta_0 + \sum_{i=1}^{p} \frac{\beta_i (\theta_i - \bar{\theta}_i)}{\delta_i} + \varepsilon \qquad (*)$$

δ_i ← scaling factor such that

$$\sum_{u=1}^{n} (\theta_{iu} - \bar{\theta}_i)^2 / \delta_i^2 = \text{const.}$$

$\bar{\theta}_i$ = means of the levels θ_{iu}, $u = 1, 2, \dots, n$ of the levels of the design.

The estimated coefficients of (*)

$$\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$$

indicate the direction of the <u>steepest ascent</u>, thus

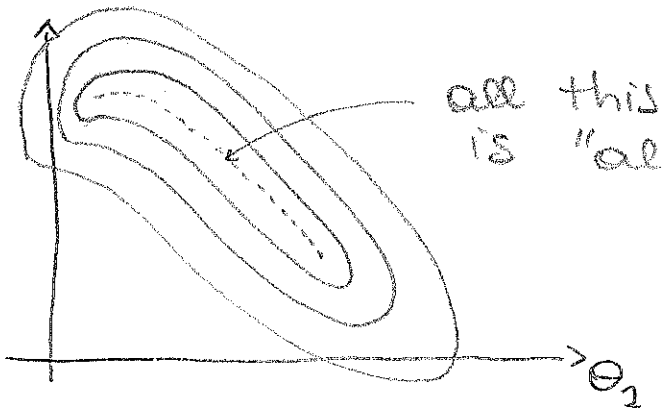
$$-\hat{\beta}_1, -\hat{\beta}_2, \dots, -\hat{\beta}_p$$

indicate the direction of the <u>steepest descent</u> => we will move on $S(\underline{\theta})$ to a $\underline{\theta}_1$ following this direction, then iterate.

Problems of this method:

1) It may converge very slowly, in particular where the contour levels of $S(\underline{\theta})$ are "banana shaped", leading to "zigzagging":

$\underline{\theta} = (\theta_1, \theta_2)$



2) The method is not scale invariant, i.e. is sensitive to the choice of δ_i, if they are not changed by the same factor

An advantage of this method is that it works well when $\underline{\theta}_0$ is far from the point of global minimum $\hat{\underline{\theta}}$, which is often the case.

METHOD 3: MARQUARDT'S COMPROMISE

It is a compromise between linearization (Gauss-Newton) and the steepest descent, and works better of both of them in most practical cases.

IDEA: Start from an initial guess $\underline{\theta}_0$.

- Compute the direction $\underline{\delta}_g$ where to move on $S(\underline{\theta})$ to find $\underline{\theta}_1$, using the steepest descent method;
- Compute the direction $\underline{\delta}$ where to move on $S(\underline{\theta})$ to find $\underline{\theta}_1$, using the linearization method;
- Interpolate suitably $\underline{\delta}_g$ and $\underline{\delta}$ for obtaining a new "best direction" and a suitable step size

Remark: on many examples Marquardt found that the angle ϕ between $\underline{\delta}_g$ and $\underline{\delta}$ is $80° < \phi < 90°$!!

CONFIDENCE REGIONS

(1) Use an ellipsoidal confidence region by assuming that in a neighbourhood of $\underline{\hat{\theta}}_{LS}$, the model can be approximated by its linearized form. The ellipsoid is

$$(\underline{\theta} - \underline{\hat{\theta}}_{LS})^T \underset{\sim}{\hat{Z}} \underset{\sim}{\hat{Z}}^T (\underline{\theta} - \underline{\hat{\theta}}_{LS}) \leq p\, s^2 F_{1-\alpha;\, p,\, n-p} \qquad (**)$$

where $\underset{\sim}{\hat{Z}}$ is the matrix of first derivatives of f, computed in $\underline{\hat{\theta}}_{LS}$ and

$$s^2 = \frac{S(\underline{\hat{\theta}}_{LS})}{n-p} \qquad p = \begin{matrix} n - \text{of parameters} \\ \text{to be estimated} \end{matrix}$$

Note that (**) is only an approximate $1-\alpha$ confidence region.

(2) An exact confidence contour is defined by taking

$$S(\underline{\theta}) = \text{Const.} = c$$

but since we don't know the distribution of $S(\underline{\theta})$, we can't determine c such that the confidence level of the region is exactly $1-\alpha$.
We can anyway use the contour such that

$$S(\underline{\theta}) = S(\hat{\underline{\theta}}_{LS}) \left\{ 1 + \frac{p}{n-p} F_{1-\alpha;\, p,\, n-p} \right\}$$

which is exactly of level $1-\alpha$ if the model is linear, and is <u>approximately</u> of level $1-\alpha$ in the non linear case.

Note that this <u>is</u> a confidence contour, is its confidence level which is approximate !

Confidence intervals for the single parameters can be obtained by sectioning the confidence regions.

Exercise

Consider the climate-pollutants data.
Find a suitable (non linear) model to relate O3 and meantemperature (meanT)

$$O3 = f(\beta, meanT) + \varepsilon$$

f ?