Minimum Description Length based model selection in linear regression*

Reijo Sund

Research and Development Centre for Welfare and Health (STAKES),
Outcome and Equity research,
P.O. Box 220, 00531 Helsinki, Finland

University of Helsinki,
Department of Statistics

Email: reijo.sund@stakes.fi

Introduction

The purpose of all scientific studies is - or at least should be - to search answers for the problem which is induced by inquisitiveness, practical need or the purposes of theory development. Keeping in mind that knowledge about our world is, and always must be, partial knowledge, one can state that the aim of statistical model fitting is to "understand" the system behind the studied phenomena via the observed data. Thus, the question of how one should decide among competing explanations of data is at the heart of the scientific enterprise. Over the decades, scientists have used an assortment of statistical tools to select among alternative models of data. However, there has not been an underlying theoretical framework to guide the enterprise and evaluate new developments.

Implicitly the principle of parsimony (or Occam's Razor) has been the soul of model selection. To implement the parsimony principle, one has to quantify "parsimony" of a model relative to the available data. Applying this measure to a number of candidate models, the goal is to find a model, which is a compromise between desirable yet conflicting properties: goodness of fit, generalizability and concision.

* Homework assignment for the course: Lectures on Statistical Modeling Theory, fall 2001, Department of Computer Science, University of Helsinki

Minimum Description Length

If we by "understanding the data" mean the ability to remove redundancies in the data and hence to discover regular statistical features, the ultimate measure of the success of understanding must be the length with which the data can be described. Indeed, if such a shortest description of the data, to be called stochastic complexity, is found in terms of the models of a selected class, there is nothing further anyone can teach us about the data; we know all there is to know. This is the rationale behind the MDL (minimum description length) principle.

The MDL principle represents a drastically different foundation for model selection and, in fact, statistical inference in general. It has a number of distinctive features: There is no need to assume anything about the data generation mechanism. In particular, unlike in traditional statistics, it is not needed that the data form a sample from a population with some probability law. Hence, the objective is not to estimate an assumed but "unknown" distribution, be it inside or outside the proposed class of models, but to find good models for the data. Most essentially, the principle permits comparison of any two models, regardless of their type. It is also important to realise that the MDL principle has nothing to say about how to select the suggested family of model classes. In fact, this is a problem that cannot be adequately formalised. In practise the selection of models is based on human judgement and prior knowledge of the kinds of models that have been used in the past, perhaps by other researches. In addition, the application of the principle requires the calculation of the stochastic complexity, which can sometimes be a difficult task.

The MDL has it's roots in information theory and in the invariance theorem of Kolmogorov Complexity. The Kolmogorov Complexity of a sequence is defined to be the length of the shortest computer program that prints the sequence and then halts (halting makes the code to be a prefix code, i.e. none of the codewords is a prefix of another). Unfortunately the Kolmogorov Complexity is not computable. The idea behind the MDL is to scale things down in a way that it becomes possible to compute the complexity: instead of using a code based on a universal computer language, we should use an arbitrary class of models and do the encoding with the help of this model class. However, there is different forms of description length based on a model, even in that sense that they achieve the universal coding lower bounds.

The most straightforward description length is the one based on two-stage coding scheme. The idea in the first stage is to calculate the code length required to discretize model's parameter space and communicate the estimated parameter. In the second stage the actual data string is coded using the distribution indexed by the communicated parameter. In more complicated situations more than two stages of coding might be required.

In the mixture form of description length we base our description of a data string on a distribution that is obtained by taking a mixture of the members in the family with respect to a probability density function on the parameters. The mixture description length results in integral formula, which has closed form expression only in special cases. An analytical approximation to the mixture can be in certain situations obtained by Laplace's expansion and essentially results in a two-stage description length which is called the stochastic information complexity.

The recent form of description length bases on the normalised maximum likelihood (NML) coding scheme. In general, the NML description of a data string works by restricting the second stage of coding to a data region identified by the parameter estimate. This criterion is not only sensitive to functional form and the number of parameters but also invariant under reparameterisation. From the differential geometric point of view this criteria selects the model that gives the highest value of the maximised likelihood per distinguishable distribution, which may be called the "normalised maximised likelihood". In other words, the model complexity is related to the number of (distinguishable) probability distributions that a model can generate, not to the functional form of a model or its number of parameters.

Quite a different approach is to think the description length from a predictive coding point of view. Predictive coding means that we model the conditional density for the possible values of the "next" observation using the part of the data which is already "seen", i.e. the joint distribution of a data string can be written as a product of conditional distributions. If each of the conditionals share the same parameter, then the joint distribution based on the particular model class is free of unknown parameters and the cost of encoding a data string can be directly seen from the joint distribution. This form of description length is called predictive description length and it is especially useful in situations where the data is sensible ordered.

Model selection in linear regression

The linear least squares regression problem is a fundamental modelling problem, for which it is possible to derive exact formulas for the different minimal description length criteria. In linear regression the model to be fit is

$$y = Xb+e,$$

where y is a vector corresponding to the dependent variable, X is a matrix of regressors, b is the vector of regression coefficients and e is the vector of error terms.

The idea is to find a linear combination of regressor variables, which explains the systematic variation in the dependent variable. The regression coefficients can be estimated using the ordinary least squares (OLS) technique, in which the error sum of squares is to be minimised. If normality with mean zero and constant variance is assumed for the error terms, the OLS estimates coincide with the maximum likelihood estimates. The main problem is to choose appropriate regressors for the model. In practice the domain knowledge is usually the best criteria to find interesting models, but sometimes a kind of "objective" assistance can be very helpful. As a matter of fact there is a lot of different model selection criteria available.

Since the purpose of this paper is to give a practical view for model selection based on MDL, the exact derivations of the formulas are not explicitely presented; interested reader can find those from the publications mentioned in the list of sources. For generality and comparison purposes also some popular non-MDL-based model selection formulas are given here. In the following formulas various authors have suggested other multipliers which keeps the general form of these criteria. In other words, the numerical values obtained from the formulas are not directly comparable (the "scale" is not same for all formulas). However, in every case the smaller value means the better model.

In the following formulas, y, X, b and e are as in the regression model, SSE is the error sum of squares, σ^2 is the variance estimate from fitting the full model, k is the number of parameters, n is the number of observations, $S = SSE / (n-k)$, $F = (y'y - SSE) / kS$ and R^2 is the coefficient of determination.

Mallow's Criteria (C_p):

$$C_p = \frac{SSE}{s^2} + 2k - n$$

Akaike's Information Criteria (AIC):

$$AIC = n\log\left(\frac{SSE}{n}\right) + 2k$$

Bayesian Information Criteria (BIC):

$$BIC = n\log\left(\frac{SSE}{n}\right) + 2(k+2)\left(\frac{n}{s^2}\right) - 2\left(\frac{ns^2}{SSE}\right)^2$$

Schwarz's Bayesian Information Criteria SBIC (two-stage MDL):

$$SBIC = n\log\left(\frac{SSE}{n}\right) + k\log n$$

Stochastic Information Complexity (SIC):

$$SIC = \frac{n-k-2}{2}\log SSE + \frac{k}{2}\log n + \frac{1}{2}\log\det[X'X]$$

G-prior mixture MDL (gMDL):

$$gMDL = \begin{cases} \frac{n}{2}\log S + \frac{k}{2}\log F + \log n & , R^2 \geq k/n \\ \frac{n}{2}\log\left(\frac{y'y}{n}\right) + \frac{1}{2}\log n & , \text{otherwise} \end{cases}$$

Normalised Maximum Likelihood MDL (nMDL):

$$nMDL = \frac{n}{2}\log S + \frac{k}{2}\log F + \frac{1}{2}\log(n-k) - \frac{3}{2}\log k$$

Exact Normalised Maximum Likelihood MDL (MDL):

$$MDL = (n-k)\log\frac{SSE}{n} + k\log[b'X'Xb] + (n-k-1)\log\left(\frac{n}{n-k}\right) - (k+1)\log k$$

In the following predictive formulas y_i and X_i correspond to the values of the response and regressor variables on ith observation, b_{i-1} and $\hat{s}^2_{i-1}$ are the estimates based on i-1 first observations, m is the first integer so that b_i is uniquely defined and n is the number of observations.

Predictive MDL (PMDL):

$$PMDL = \sum_{i=m+1}^{n} \left[\log \hat{s}^2_{i-1} + \frac{(y_i - X_i b_{i-1})^2}{\hat{s}^2_{i-1}} \right]$$

Predictive Least Squares (PLS):

$$PLS = \sum_{i=m+1}^{n} (y_i - X_i b_{i-1})^2$$

Practical example

In the appendix 1 is a listing of a data set consisting of 100 observations for the variables x and y. The aim is to build two linear regression models, where y is a dependent variable (both models) and ordinary (model 1) and hermite (model 2) polynomial transformations of x are regressors, and choose the "best" model. The possible models are assumed to be restricted in such a way that the minimum and maximum degrees of polynomials in a model are correspondingly zero and six and adding a higher level regressor to a model is possible only if all the lower degree polynomials are also included.

As can be seen from the figure 1, there is some kind of functional dependency between the variables. Moreover, fitting the full models (all variables included) and checking the distributions of residuals it is found that the error terms are not normally distributed (figure 2). Usually these kind of findings in exploratory analyses stand for serious problems in actual modelling.

However, proceeding with the analyses using straightforward computations using statistical package SURVO (see appendix 1) we end up with the following tables of model selection results. The same results are also presented in "standardised" form in figure 3. The predictive criteria base on the original order of observations and to avoid the initialisation problem the sequential prediction is not started until the 11^{th} datapoint.

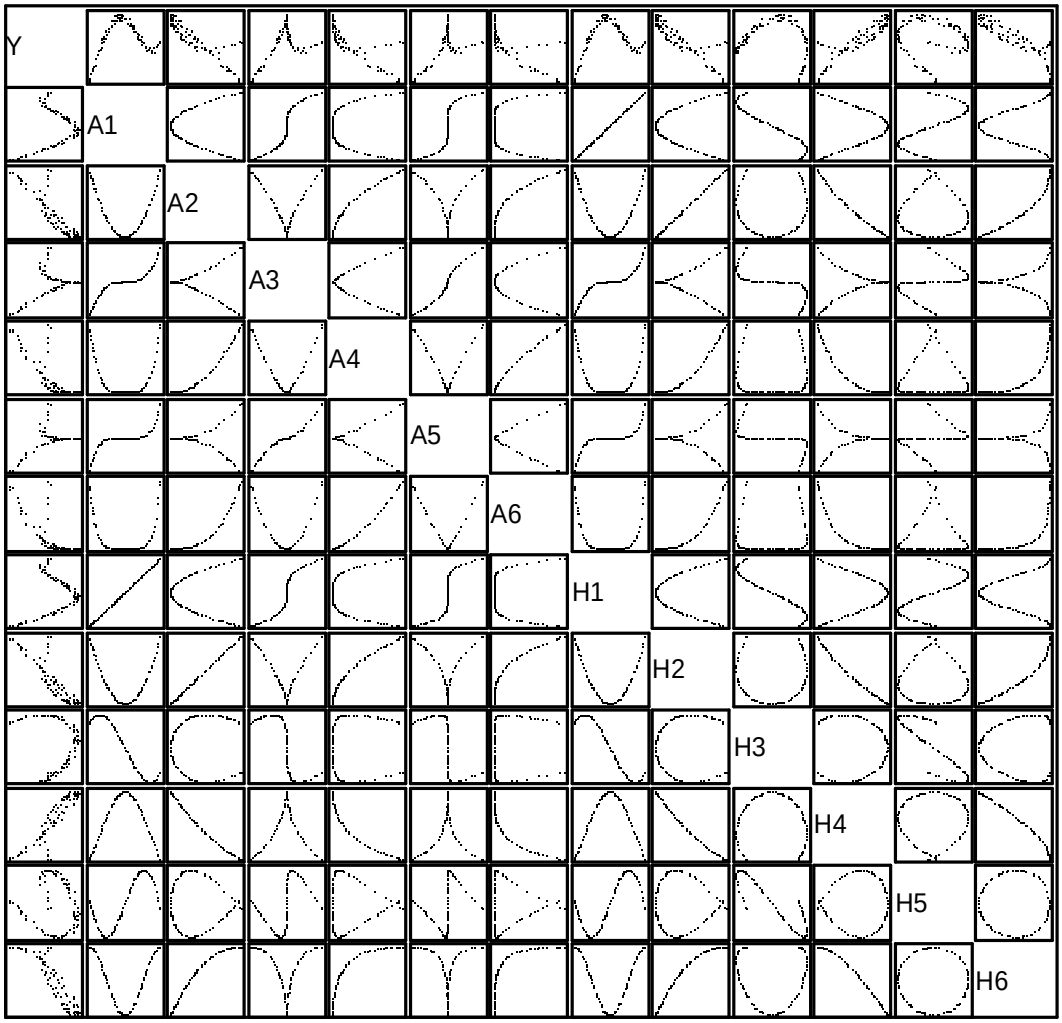

Figure 1. Draftman's display. Variables A1-A6 correspond to the normal polynomial transformations of x and variables H1-H6 to the hermite polynomial transformations. Constants (A0 and H0) are omitted from the plot.

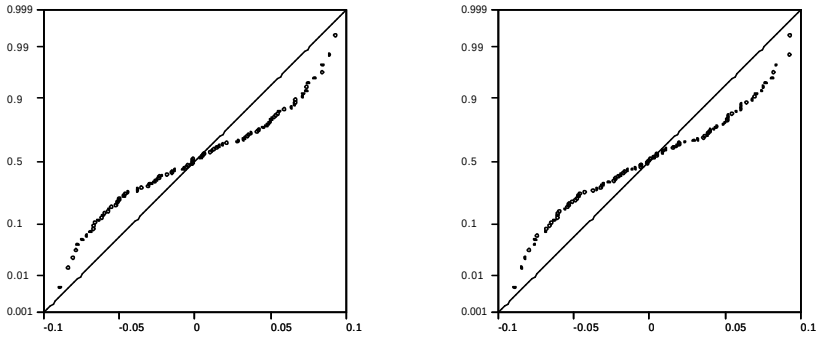

Figure 2. Q-Q-plots for the full models 1 and 2. (Normally distributed = straight line)

Ordinary Polynomials

k	*Cp*	*AIC*	*BIC*	*SBIC*	*SIC*	*gMDL*	*nMDL*	*MDL*	*PMDL*	*PLS*
1	6790	-175.5	-177.4	-172.9	141.9	-1.2	-83.0	-170.4	-42.7	17.0
2	6094	-184.2	-188.1	-179.0	139.3	-83.2	-86.5	-175.8	-34.9	16.6
3	1158	-342.0	-347.2	-334.2	65.5	-157.6	-161.6	-325.5	-37.4	3.82
4	233	-474.9	-479.4	-464.5	4.2	-219.0	-223.4	-448.5	-121.6	1.30
5	**13.7**	**-587.1**	**-585.5**	**-574.1**	**-46.7**	**-269.3**	**-274.0**	**-549.4**	**-310.1**	**0.39**
6	15.7	-585.1	-583.6	-569.5	-44.8	-264.9	-270.0	-540.9	-295.6	0.41
7	14.0	-586.7	-584.7	-568.5	-45.1	-262.3	-267.6	-535.8	16466	26.0

Hermite Polynomials

k	*Cp*	*AIC*	*BIC*	*SBIC*	*SIC*	*gMDL*	*nMDL*	*MDL*	*PMDL*	*PLS*
1	6790	-175.5	-177.4	-172.9	141.9	-1.2	-83.0	-170.4	-42.7	17.0
2	6094	-184.2	-188.1	-179.0	140.0	-83.2	-86.5	-175.8	-34.9	16.6
3	1158	-342.0	-347.2	-334.2	67.5	-157.6	-161.6	-325.5	-37.4	3.82
4	233	-474.9	-479.4	-464.5	8.4	-219.0	-223.4	-448.5	-121.6	1.30
5	**13.7**	**-587.1**	**-585.5**	**-574.1**	**-39.8**	**-269.3**	**-274.0**	**-549.4**	**-310.1**	**0.39**
6	15.7	-585.1	-583.6	-569.5	-34.4	-264.9	-270.0	-540.9	-295.6	0.41
7	14.0	-586.7	-584.7	-568.5	-30.6	-262.3	-267.6	-535.8	16461	26.0

Due to these results, all the selection criteria suggest the models of fourth degree polynomials. Moreover, the ordinary and hermite polynomials give the same description length for all criteria except SIC, i.e. according to the geometric interpretation of the exact MDL criteria those are essentially just different parameterisations of the same model.

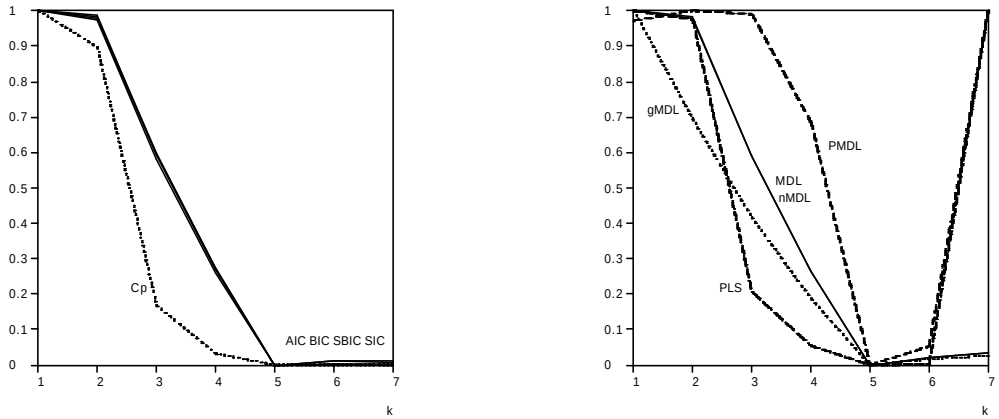

Figure 3."Standardised" results, ordinary polynomials

List of Sources

Hansen, Mark H. & Yu, Bin (2001): Model Selection and the Principle of Minimum Description Length. Journal of the American Statistical Association (to appear).

Mustonen, Seppo (1992): SURVO - An Integrated Environment for Statistical Computing and Related Areas. Survo Systems. Helsinki.

Myung, In Jae & Zhang, Shaobo & Pitt, Mark A. (1999): Geometric Complexity and Minimum Description Length Principle. (Draft).

Puranen, Juha (1997): Data-analyysi (Data analysis). Lecture notes. Department of Statistics. University of Helsinki.

Rissanen, Jorma (1986): Stochastic Complexity and Modeling. Annals of Statistics 14:3, 1080-1100.

Rissanen, Jorma (2000): MDL Denoising. IEEE Transactions on Information Theory, IT-46:7.

Rissanen, Jorma (2001): Lectures on Statistical Modeling Theory. Lecture notes. Department of Computer Science. University of Helsinki.

SAS institute (1999): Model Fit and Diagnostic Statistics. The REG Procedure. SAS OnlineDoc v8. SAS Institute.

Wei, C. Z. (1992): On Predictive Least Squares Principles. Annals of Statistics 20:1, 1-42.

```
*SAVE MDL
*
*The "original data set"
*
*DATA D1,A,B,C
CX                          Y
A 1.655833631224617e-001   1.166637632464838e+000
*-1.530074862978985e-001   1.344150104924718e+000
* 3.102350428152567e-002   1.318746329459874e+000
*-3.320970400564828e-001   1.294663517507757e+000
*-1.341868077865418e-001   1.331434142079297e+000
*-5.481002637111070e-001   8.680243914868706e-001
* 1.596137464991980e-001   1.199878575625089e+000
* 5.207300196086881e-001   7.419873961241118e-001
* 5.964623343213216e-002   1.215798760932922e+000
* 2.810529979796706e-001   1.093665447052138e+000
*-5.818611911231599e-001   9.054452606231694e-001
*-2.403632592984178e-001   1.292319565920448e+000
* 5.666572997354253e-001   6.543719916543290e-001
* 3.616915027944594e-001   9.932847751927388e-001
*-7.780974668660223e-002   1.410614697565672e+000
* 1.356574248576685e-001   1.317101630319699e+000
* 5.884213027452294e-001   6.679970164785789e-001
*-8.816348130578577e-001   2.097720073531828e-001
* 2.057381713339848e-001   1.150228385663520e+000
*-8.994623925062541e-001   1.462964189444808e-001
*-1.692502791135546e-001   1.415267250558598e+000
*-3.900026459930162e-001   1.133089413070555e+000
* 7.487343431752524e-001   5.559102587227264e-001
*-9.699810026467674e-001  -8.494062033243803e-002
* 5.359007800222828e-001   7.670477738618251e-001
* 9.416898785112468e-001   7.245315873083948e-001
* 9.801651852261252e-001   7.820511571058888e-001
* 5.777233844675052e-001   6.741003044898320e-001
*-1.226829324581760e-001   1.294647464121411e+000
*-3.377393103039084e-003   1.327622027005822e+000
*-5.720733368076675e-001   8.641207474635655e-001
* 2.869845757706999e-001   1.024022225317884e+000
*-3.599288450670062e-001   1.286918421083040e+000
* 9.201972007380210e-001   7.226492770888950e-001
* 4.532635332837960e-001   9.104704048135290e-001
*-1.760935836627104e-001   1.367089119431872e+000
* 4.891315662123108e-001   7.036375423113067e-001
*-4.641054985812515e-001   1.099802301041796e+000
*-1.201513808693169e-001   1.419469888718829e+000
* 8.667602163791794e-001   6.970070580377314e-001
* 3.666646486769694e-001   9.031587427486700e-001
*-5.748802713225267e-001   9.162179023500520e-001
* 6.784764806723904e-001   6.037427416993005e-001
* 2.575692000481476e-001   1.169799476647005e+000
*-7.324545030531473e-001   5.602924425583777e-001
*-5.857345407172900e-001   8.125045840101361e-001
* 2.143978890790608e-001   1.139847198324631e+000
* 2.597756976846162e-001   1.036722843901726e+000
*-2.590463478962074e-001   1.398364527906660e+000
* 1.502955580949370e-001   1.194504101364189e+000
*-9.715034647504662e-002   1.287271289685580e+000
*-9.122093493057120e-001   1.045815865489734e-001
*-9.456297540066508e-001   4.166452244870178e-002
*-3.746299038397094e-001   1.172748241108614e+000
*-9.742748506540050e-001  -1.420524012104264e-001
*-2.320654230113966e-001   1.408477983178817e+000
* 3.662319356091996e-001   8.881768792565522e-001
*-8.143150765181604e-001   3.918092487384336e-001
*-9.293233520616888e-001   5.086030909929048e-002
* 2.247909627460449e-001   1.195637952388555e+000
* 2.170807224479832e-001   1.095504291603473e+000
*-9.684803641604986e-001  -9.235599174599896e-002
*-9.672901329000070e-001  -1.533735990257834e-001
*-6.198508218405485e-001   8.467338093556774e-001
* 1.738369437693468e-001   1.205072501046552e+000
*-8.848378202434229e-001   1.014481720976662e-001
*-2.648639223473116e-001   1.292676306956284e+000
* 2.629023294888864e-001   1.151890958576170e+000
* 4.352688429313946e-001   8.355512378948255e-001
* 3.853387894355756e-001   8.522451666195930e-001
*-8.318418784991107e-001   2.194335752862969e-001
*-9.128970048889518e-002   1.286762697472016e+000
*-1.163434061873154e-001   1.376769966306303e+000
*-2.934990899986188e-001   1.227486268523524e+000
*-6.927872749530165e-001   6.237581326885469e-001
* 3.512892992668244e-001   9.061752681120840e-001
* 3.984266554825248e-001   9.536117754647240e-001
* 4.550182584358613e-001   8.097180787565774e-001
*-4.323123808668538e-002   1.408823469725868e+000
* 1.096839726833538e-001   1.180548619821508e+000
*-7.579057739271714e-001   4.771188179918713e-001
*-9.849211804122682e-002   1.441546051921794e+000
* 4.317658963459490e-001   8.259800306321049e-001
* 7.856832162915064e-001   6.408518819412552e-001
*-4.537950595497152e-001   1.031393498628919e+000
*-4.904614088754334e-001   9.651962406093136e-001
* 7.312069554895038e-001   5.320131726356866e-001
*-5.352992587449350e-001   1.033093233574328e+000
* 6.097434882314121e-001   5.772864833238529e-001
* 8.167950868970366e-001   5.425701011288548e-001
*-5.362113637753497e-001   1.042748409967318e+000
*-5.213748710620297e-001   9.354977465706280e-001
*-9.004910318575006e-001   9.118004054129114e-002
*-8.432318504599099e-001   2.052947600685934e-001
* 2.816308197400344e-001   1.094405454969295e+000
*-6.182268592048795e-001   8.685726090005242e-001
* 6.877389977487169e-001   6.404542522765659e-001
*-6.521995030764315e-001   7.429719814874598e-001
*-6.584143725166405e-001   6.371169720977338e-001
B 9.885909810278394e-001   7.387227662550414e-001
```

```
*
*.....................................................................................................
*Create the SURVO data file
*
*FILE CREATE MDL1,160,20
*FIELDS:
*   1 NA_   8 X         (##.########)
*   2 NA_   8 Y         (##.########)
*END
*.....................................................................................................
*Copy the data from the edit field to the file
*
*FILE COPY D1 MDL1
*.....................................................................................................
*Calculate polynomial transformations
*
*VAR A0,A1,A2,A3,A4,A5,A6,H0,H1,H2,H3,H4,H5,H6,P,R TO MDL1
*A0=1      H0=1
*A1=X      H1=2*X
*A2=X^2    H2=-2+4*X^2
*A3=X^3    H3=-12*X+8*X^3
*A4=X^4    H4=12-48*X^2+16*X^4
*A5=X^5    H5=120*X-160*X^3+32*X^5
*A6=X^6    H6=-120+720*X^2-480*X^4+64*X^6
*P=MISSING R=MISSING
*.....................................................................................................
*Copy essential parts of the data to matrices
*
*MAT SAVE DATA MDL1 TO Y / VARS=Y
*MAT SAVE DATA MDL1 TO A / VARS=A0,A1,A2,A3,A4,A5,A6
*MAT SAVE DATA MDL1 TO H / VARS=H0,H1,H2,H3,H4,H5,H6
*.....................................................................................................
*
*Some "short-cuts" to matrix cells
*
*SSE=MAT_SSE(1,1)
*SST=MAT_SST(1,1)
*nrh=MAT_NRH(1,1)
*yy=MAT_YY(1,1)
*
*Some definitions
*
*dxx=det
*fs=(FSSE/N)
*S=SSE/(N-k)
*F=(nrh-SSE)/(k*S)
*
*Definitions of the model selection criteria
*
*   rs=(1-(SSE/SST))
*  ars=1-((N-1)*(1-rs)/(N-k))
*   Cp=SSE/fs+2*k-N
*  AIC=N*log(SSE/N)+2*k
* SBIC=N*log(SSE/N)+k*log(N)
*  BIC=N*log(SSE/N)+2*(k+2)*(N*fs/SSE)-2*(N*fs/SSE)^2
*  SIC=(N-k-2)/2*log(SSE)+(k/2)*log(N)+1/2*log(dxx)
*  MDL=(N-k)*log(SSE/N)+k*LOG(nrh)+(N-k-1)*log(N/(N-k))-(k+1)*log(k)
* nMDL=(N/2)*log(S)+k/2*log(F)+1/2*log(N-k)-3/2*log(k)
* gMDL=if(rs>=k/N)then(N/2*log(S)+k/2*log(F)+log(N))else(gMDL2)
*gMDL2=N/2*LOG(yy/N)+1/2*log(N)
*
*
*Computation of the criteria  (inverted parts are "interactive")
*
*N=100                                          / Number of observations
*k=7                                            / Number of parameters
*MAT X=A(1:N,1:k)                               / Create submatrix from the full regressor matrix
*MAT GRAM-SCHMIDT DECOMPOSITION OF X TO U,R      / Normal equation solving by
*MAT C=U'*Y                                     /   orthogonalization
*MAT SOLVE B FROM R*B=C                         /   (numerically stable solution)
*MAT E=Y-X*B                                    / Calculation of residuals
*MAT SSE=SUM(E,2)                               / Error sum of squares
*MAT SST=SUM(CENTER(Y),2)                       / Total sum of squares
*MAT NRH=B'*X'*X*B                              / Some more matrix calculations
*MAT YY=Y'*Y                                    /   ..
*MAT XX=X'*X                                    /   ..
*MAT DXX=INV(XX,det)          / *det=164.689 7*7       determinant obtained via inverting
*
*FSSE=0.24608197709001                          / Error sum of squares for the full model
*
*
*Criteria values
*
*   rs=0.98548252249392
*  ars=0.98454591104192
*   Cp=14.000000000001
*  AIC=-586.72607443467
* SBIC=-568.48988313276
*  BIC=-584.72607443467
*  SIC=-45.124995330934
*  MDL=-535.83959862175
* nMDL=-267.5724001371
* gMDL=-262.3146644741
*
```

```
*.....................................................................................................
*Simple loop-macro for the predictive criteria
*
*
*TUTSAVE LUUPPI
*{init}{tempo -1}{R}
+ luuppi:
*{line start}{erase}i={print W1}{R}
*{pre}{act}{u12}
*{W1=W1+1}
- if W1 <= W2 then goto luuppi
*{d12}
*{end}
*
*.....................................................................................................
*Computation of predictive criteria
*
*
*sigma=MAT_S(1,1)
*
*MAT X=A                                    / Normal or hermite polynomials
*
*k=4                                        / Number of parameters
*MAT XP=X(1:10,1:k)                         / Submatrices for initialization
*MAT YP=Y(1:10,1:1)                         /   ..
*MAT PNS=CON(1,1,0)                         / Zeros for criteria values (1x1 constant matrices)
*MAT PMDL=CON(1,1,0)                        /   ..
*MAT GRAM-SCHMIDT DECOMPOSITION OF XP TO U,R / Least squares solution
*MAT C=U'*YP                                /   ..
*MAT SOLVE B FROM R*B=C                     /   ..
*MAT E=YP-XP*B                              / Residuals
*MAT S=SUM(E,2)/10                          / Error sum of squares divided by the number of init-obs
*MAT LS=S
*MAT TRANSFORM LS BY log(X#)                / Logarithm of S
*
*
*/LUUPPI 11,100                            / Sequential loop
*i=100                                      / Obs number
*MAT XP=X(1:i,1:k)                          / Submatrices
*MAT YP=Y(1:i,1:1)                          /   ..
*MAT E=YP-XP*B                              / Residuals using estimators based on i-1 obs
*MAT PNS=PNS+(E(i:i,1:1)^2)                 / Update criteria values
*MAT PMDL=PMDL+LS+(E(i:i,1:1)^2)/sigma      /   ..
*MAT GRAM-SCHMIDT DECOMPOSITION OF XP TO U,R / New estimators
*MAT C=U'*YP                                /   ..
*MAT SOLVE B FROM R*B=C                     /   ..
*MAT S=SUM(E,2)/i                           /   ..
*MAT LS=S                                   /   ..
*MAT TRANSFORM LS BY log(X#)                /   ..
*
*MAT LOAD PNS,D                             / Load the criteria values to the edit field
*MAT LOAD PMDL,E                            /   ..
*
DMATRIX PNS
*...
*///             Y
*PNS      1.296731
*
EMATRIX PMDL
*...
*///             Y
*PMDL     -121.581
```