SCREENING FOR IMPORTANT INPUTS BY BOOTSTRAPPING

Russell Cheng
School of Mathematics

UNIVERSITY OF Southampton

TALK OVERVIEW

1. **Value of Bootstrapping for Analysing Statistical Distributions, Including those appearing in Discrete Event Simulation**

2. **Amenability to Parallel Computation**

3. **A Detailed Example: Screening for Important Factors in the Linear Statistical (Meta)Model.**

Bootstrapping (BS) can be used:

To handle **Non-standard** problems like

1. Generating critical values in difficult distributions like the null distribution of the Anderson-Darling goodness of fit test.

and

To provide fresh insight into **standard problems** like

2. Screening for important factors in the standard Linear Model.

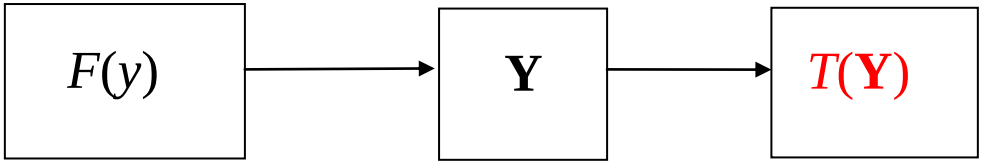
Parallel Implementation

Because the computer intensive aspect of BS involves simply making many independent replications of the same sampling experiment, it is amenable to straightforward parallel computing implementation.

Open applications programming interface (API) software is now available for the latest generation of graphic processing units (GPU) such as GeForce 8 Series NVIDIA GPU, in computer unified device architecture (CUDA) which makes the stream processors of such GPUs accessible for parallel computing applications. (Park and Fishwick, 2011).

A Basic Statistical Problem

Basic Sampling Process



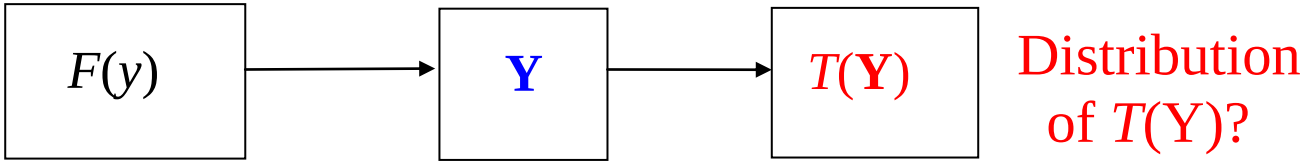
(Null) Distribution Sample (Test) Statistic Of Interest

What is the distribution of $T(\mathbf{Y})$?

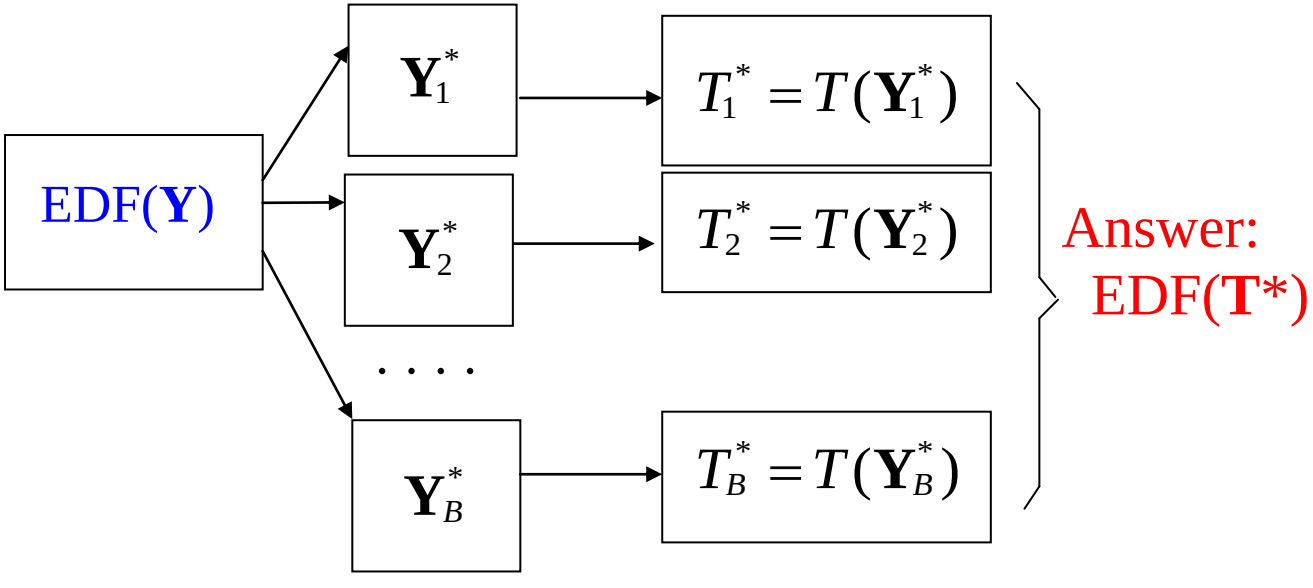
Bootstrapping answers this by applying (twice) the fundamental theorem of sampling

As the sample size $n \to \infty$,
Empirical Distn Function $\to$ *Cumulative Distn Function with probability 1.*

Basic Sampling Question



Bootstrap Answer



Linear (Meta)Model:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} \ldots + b_P x_{ip} + \varepsilon_i \quad i = 1,2, \ldots, n$$

where

y is the output variable of interest

$x_1, x_2, \ldots x_P$ are P independent variables (factors) on which y depends

$b_0, b_1, \ldots, b_P$ are $(P+1)$ unknown coefficients

i.e. $\mathrm{y} = \mathrm{Xb} + \varepsilon, \quad \varepsilon \sim \text{Normal}(0, \sigma^2)$

Let $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_P, \hat{\sigma}^2$ be the MLEs of $b_0, b_1, \ldots, b_P, \sigma^2$.

If we are not sure which factors are important, there are 2^{P+1} possible models to choose from (including b_0 as a factor). This is the basic statistical problem of linear model selection. (Screening problem)

We characterize a (sub)model (with p factors) by

$$m = \{j_1, j_2, \ldots, j_p\}$$

where

$$j_1 < j_2 < \ldots < j_p, \quad p \leq P,$$

Are the indices of factors of the full model retained in the model.

Full model denoted by M

Problem has been discussed at length, (e.g Wu and Hamada 2000). Especially

(i) Sequential backward, forward factor selection methods.

Stepwise variants to deal with non-orthogonally designed experiments, to try to avoid selecting a model that does not include all those factors that are important.

(ii) More sophisticated Bayesian strategies, employing Gibbs sampling,gives greater insight.

(Not be considering these here.)

We consider **bootstrap approach**.

Three issues:
(i) Criterion Measure used for deciding which is the 'best' model.

(ii) Need to know if the selected 'best' model is a sufficiently good fit.

(iii) Need to know if best model is a clear winner, or if there are competitive alternative models

Full Details in:

Cheng, R.C.H (2009) Computer Intensive Statistical Model Building. In *Advancing the Frontiers of Simulation. Eds C. Alexopoulos, D. Goldsman and J.R. Wilson.* Springer, 43-63.

Criterion Measure

A number of criteria have been suggested for model selection.

A very interesting criterion is

(i) The **C_p statistic** proposed by Mallows (1973):

$$C_p(m) = [n - p(m)]\hat{\sigma}^2(m)/\hat{\sigma}^2(M) + 2p(m) - n,$$

a combined measure of the bias and variance of the fitted model. Related to

(ii) The **Akaike Information Criterion** (Akaike, 1970).

$$AIC(m) = -2n\log[\hat{\sigma}^2(m)] + 2p(m). \qquad (11)$$

Asymptotically C_p and ***AIC*** have essentially the same distribution (see Nishii, 1984).

Mallows (1973) shows that if the model m (with p factors) is satisfactory in the sense that it has no bias, then the expected value of C_p is close to p:

$$C_p(m) \approx p.$$

So we would expect a model selected as 'best' to satisfy

$$C_p(m) \leq p$$

A simple selection method, if we can examine all models, is therefore:

"Min C_p" Model Selection Method

(i) Consider each of the $2^{P+1} - 1$ possible models and for each model m calculate $C_p(m)$.

(ii) Select as the best model that m for which $C_p(m)$ is minimum, with the expectation that this model will be satisfactory if $C_p(m) \leq p$.

The following alternative model selection procedure is more parsonious:

"Unbiased Min *p*" Selection Procedure

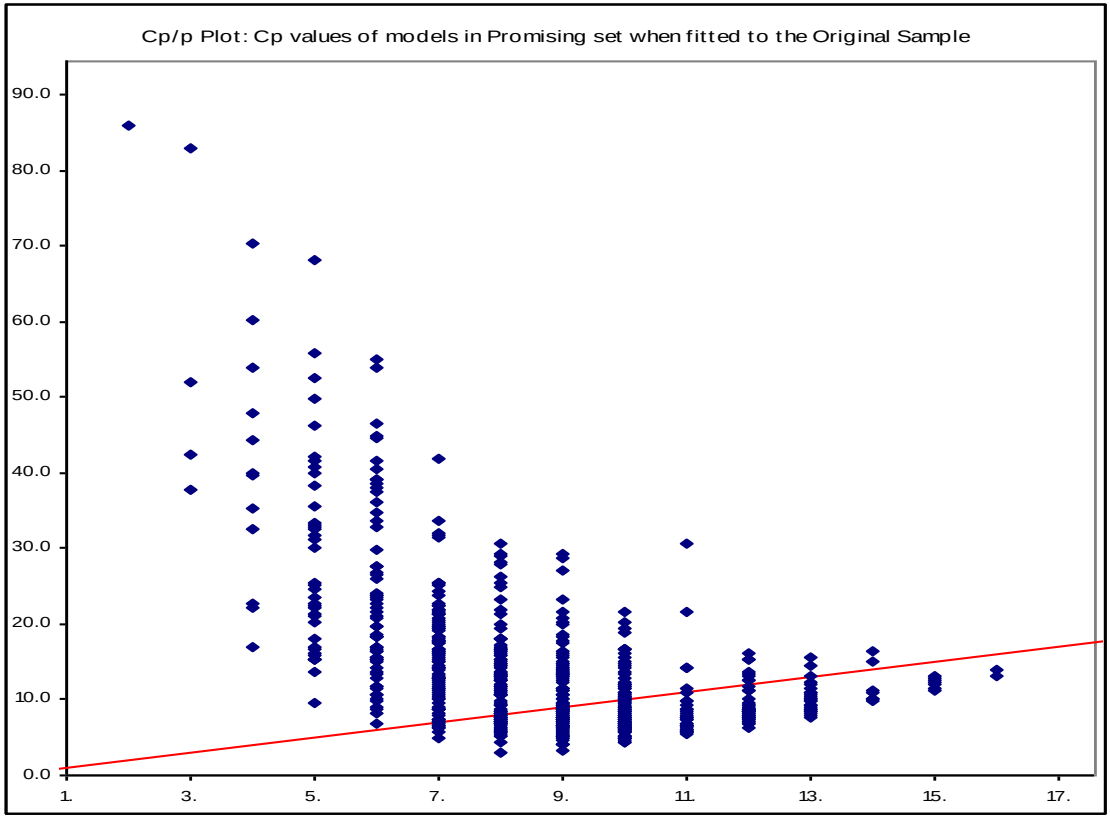
(i) Find the smallest p for which there are models m satisfying $C_p(m) \leq p$:

$$p_0 = \min\{p : C_p(m) \leq p\}.$$

(ii) Amongst all such models m, with $p(m) = p_0$, find the one for which $C_p(m)$ is minimum.

As $C_p \leq p$ for the selected model, this means the model contains no bias whilst having the smallest p possible. For this reason we call it the "Unbiased min p" method.

A typical C_p versus p plot (From a fire & rescue simulation example). Red line is $C_p = p$.



For the orthogonal design case where there are a large number of factors with coefficient values uniformly distributed in the neighbourhood of zero with density λ, Mallows (1995) has shown that the scatterplot has a lower boundary that is the (convex) cubic polynomial in p

$$C_p - P \approx (12\lambda^2)^{-1}(P-p)^3 - 2(P-p)$$

and that this boundary intersects the line $C_p = p$ at

$$P - p = 2\sqrt{3}\lambda$$

Inspection of all models is tractable only when P is small.
e.g. 20 explanatory variables $\Rightarrow$ 1,048,576 models.

Our approach is to identify a set of promising models using bootstrap resampling.

The number of models in this set is easily controlled and so can be much smaller than 2^{P+1}.

But it will almost certainly contain many good candidate models.

So it is satisfactory to select a `best' model from this subset.

Bootstrap Analysis

Step (1) To find a set of Promising Models

"One Model per Bootstrap Sample" Method

1.1 Generate *B* parametric BS samples

$$\mathbf{Y}^{(j)*} \; j = 1,2, \ldots, B$$

$$Y_i^{(j)*} = \mathbf{X}_i \hat{\mathbf{b}} + e_i^{(j)*}, \; i = 1,2,\ldots,n \qquad j = 1,2,\ldots,B$$

where

$$e_i^{(j)*} \sim N(0, \hat{\sigma}^2),$$

with $\hat{\mathbf{b}} = (\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_P)$, **and** $\hat{\sigma}^2$ **the estimates from the original sample, so that each BS sample** $\mathbf{Y}^{(j)*}$ **has the same form as the original sample Y.**

1.2 For each BS sample $\mathbf{Y}^{(j)*}$ (and the original sample):

1.2.1 Fit the full model, *M*, to the sample

1.2.2 Calculate the *|t|-value* of each of the fitted coefficients.

1.2.3 Retain just those coefficients with *|t|*-value > *a* (a preselect value)

This is the selected promising model for the given BS sample.

The set of distinct models comprises our set of *promising models*, *S*.

| *t* |-Values and *p*-Values

For a fitted model associated with each factor *j* a t-value

$$t_j = \hat{b}_j / \sqrt{\hat{\sigma}^2 d_j}$$

where d_j is the *j*th entry in the main diagonal of the dispersion matrix

$$d_j = (\mathbf{X}^T \mathbf{X})^{-1}_{jj}.$$

If the true value of b_j is $b_j = 0$ then t_j has Student's *t*-distribution with $n - p$ degrees of freedom.

Magnitude of t_j provides a simple criterion for when deciding to include b_j in the model or not.

In the orthogonal case, the "One Model per BS Sample" selection method with $a = \sqrt{3}$ is equivalent to using the "Unbiased min-*p*" selection method. Equivalent retaining factors with p-value less than 0.083.

The above method produces at most *B* promising models, but can be far fewer, if the same model is repeatedly obtained from different BS Samples.

Once a set of promising models has been obtained:

Step(2)

Use the "unbiased min p" method to select, from the set of promising models, the ‘best’ model for the original data set.

Use the "unbiased min p" method to select, from the set of promising models, the ‘best’ model for each of the *B* bootstrap samples.

Step(3) Display the models of *S*, ranked in order of the proportion of times that they are selected in Step(2) as being the best model in one of the bootstrap samples, displaying these proportions as well.

The computing effort in the above analysis is at worst quadratic in *B*

For either method of generating the set of promising models each possible model has a positive (often small) probability of being chosen. Thus, as $B \rightarrow \infty$, Step 3 will tend to the situation where every model satisfying $p \leq p_0$ is considered for possible selection as the best.

The frequency table of Step(3) estimates $\alpha(m)$, the probability that model m will be selected as the best model.

Numerical Example

Statlib Body Fat Example: Obsns of 252 individuals

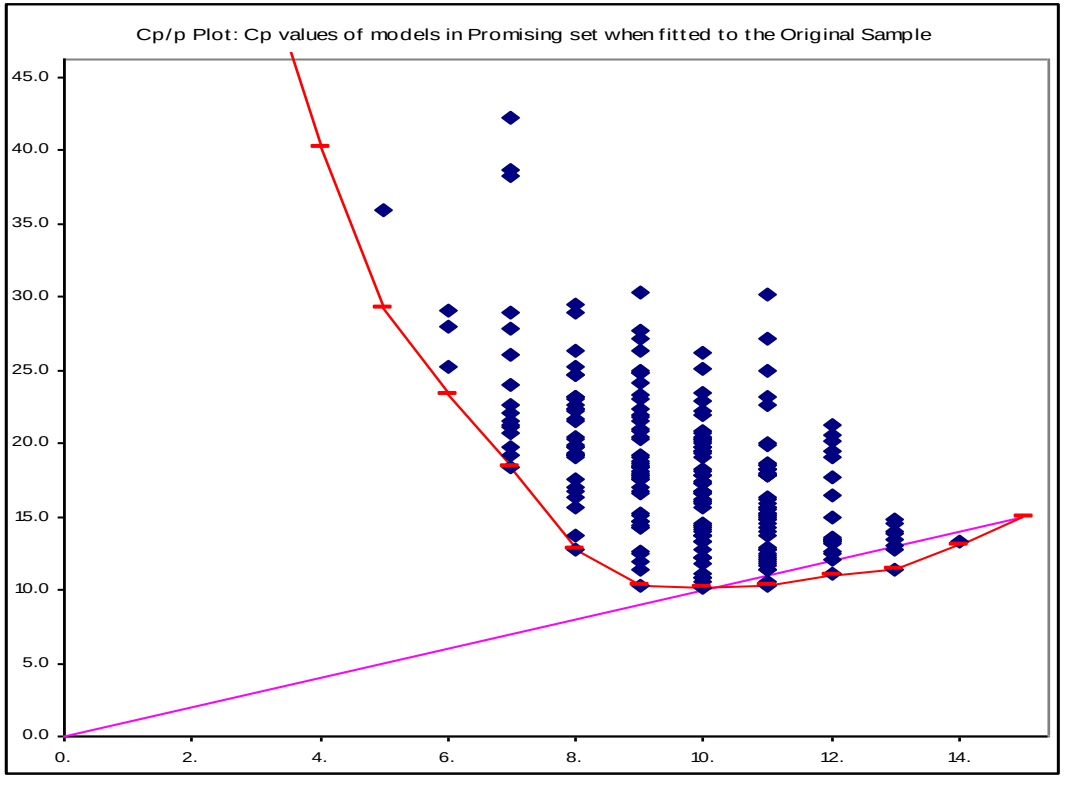
Percent body fat (Y)

1. Body Density determined from underwater weighing
2. Age (years)
3. Weight (lbs)
4. Height (inches)
5. Neck circumference (cm)
6. Chest circumference (cm)
7. Abdomen 2 circumference (cm)
8. Hip circumference (cm)
9. Thigh circumference (cm)
10. Knee circumference (cm)
11. Ankle circumference (cm)
12. Biceps (extended) circumference (cm)
13. Forearm circumference (cm)
14. Wrist circumference (cm)

Body fat data, non-orthog. Y = % Body Fat. 14+1 factors. 252 individuals. 14+1 factors.

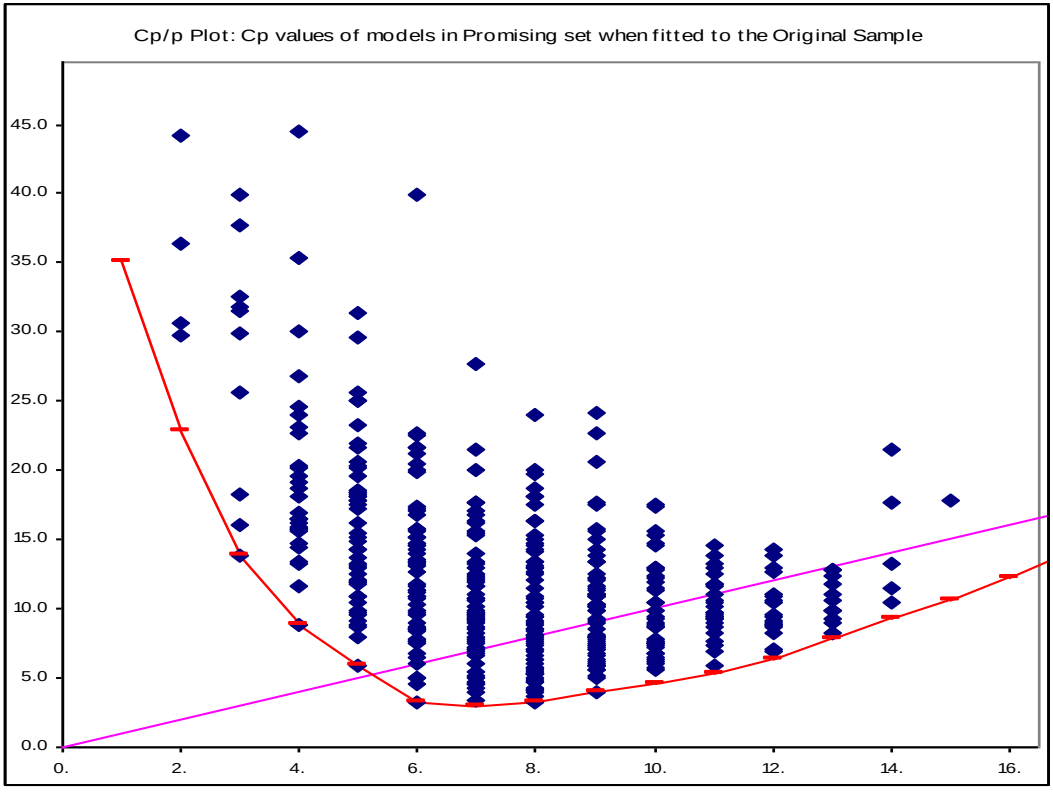
$C_p(m)$ against p: 236 Promising Models from 500 BS samples.

Red Line: Min C_p Line for all $2^{15} - 1 = 32767$ models

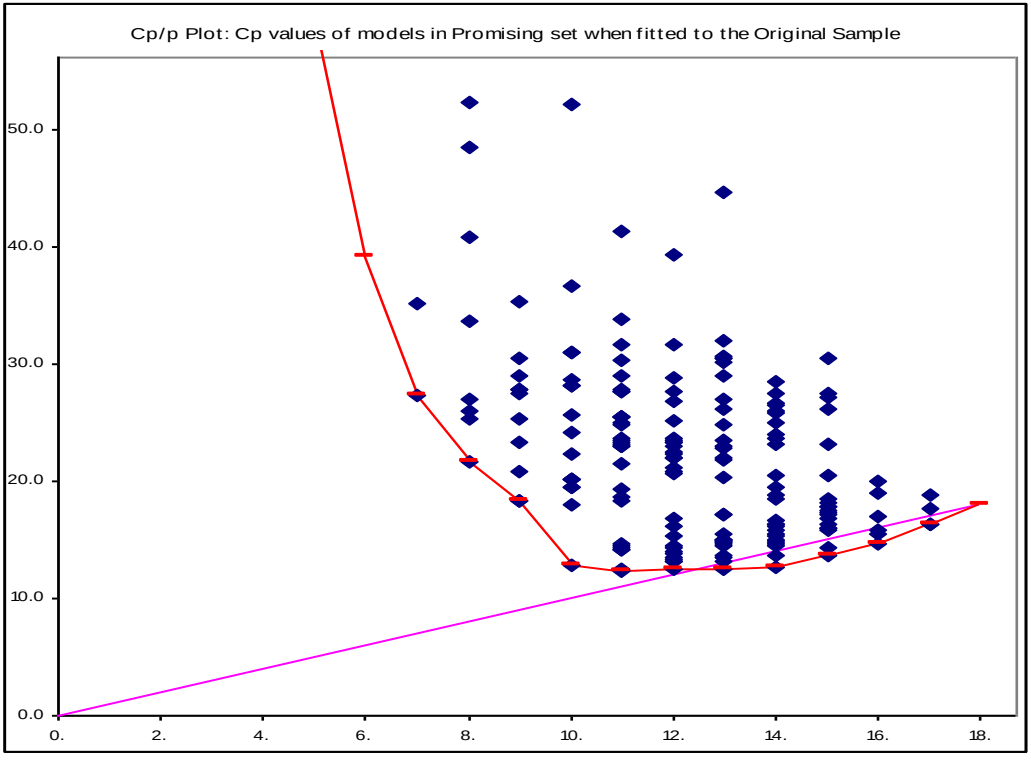


Fire & Rescue Orthog Simulation. Y = Performance cost measure. 19+1 factors. 32 runs. $C_p(m)$ against p for 445 Promising Models from 500 BS samples.

Red Line: Min C_p Line for all $2^{20} - 1 = 1{,}048{,}575$ models

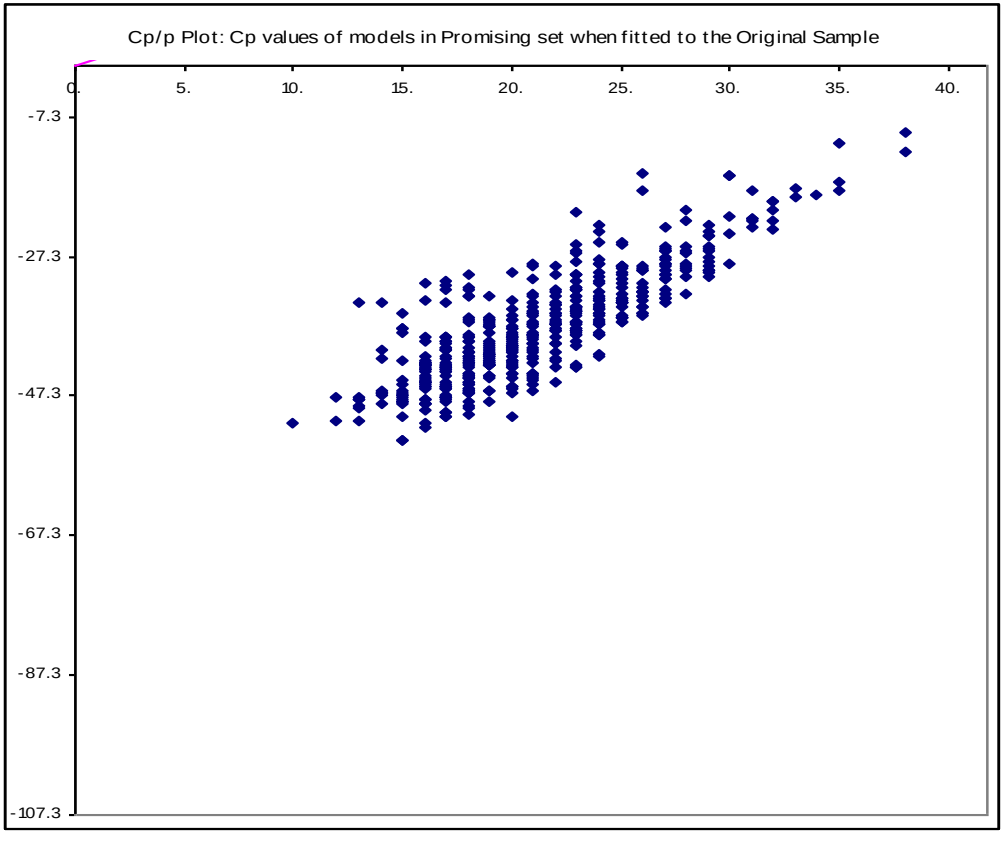


Bank Data, Non-orthog: Y = DEOM (first difference between end of month balances at a mutual bank).
17+1 factors. 59 obsns. $C_p(m)$ against p for 198 Promising Models, from 500 BS samples.
Red Line: Min C_p Line for all $2^{20} - 1 = 262{,}143$ models



Ericsson Supply Chain Orthog Simulation. Y = Operating Cost. 92+1 factors. 128 obsns.

$C_p(m)$ against p for 500 Promising Models from 500 BS samples.



Final Comments

The Excel workbook is available from the author.

Further Work: Port onto a parallel platform.

REFERENCES

Akaike, H. 1970. Statistical predictor identification. *Ann. Inst. Statist. Math.* 22 203-217.

Cheng, R.C.H (2009) Computer Intensive Statistical Model Building. In *Advancing the Frontiers of Simulation. Eds C. Alexopoulos, D. Goldsman and J.R. Wilson.* Springer, 43-63.

Mallows, C. L. 1973. Some comments on C_p. *Technometrics*. 15 661-675.

Mallows, C. L. 1995. More comments on C_p. *Technometrics* 37 362-372.

Nishii, R. 1984. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12, 758-765.

Park, H. and Fishwick, P.A. An analysis of queuing network simulation using GPU-based hardware acceleration. *TOMACS*, 21, Article 18(22 pages)

Wu, C. F. J., and M. Hamada. 2000. *Experiments Planning, Analysis, and Parameter Design Optimization.* New York: Wiley.