Rupp, Thomas

Working Paper

Meta analysis of empirical deterrence studies: an explorative contest

Darmstadt Discussion Papers in Economics, No. 174

Provided in Cooperation with:
Darmstadt University of Technology, Department of Law and Economics

Suggested Citation: Rupp, Thomas (2006) : Meta analysis of empirical deterrence studies: an explorative contest, Darmstadt Discussion Papers in Economics, No. 174, Technische Universität Darmstadt, Department of Law and Economics, Darmstadt

This Version is available at:
http://hdl.handle.net/10419/32091

Darmstadt Discussion Papers in Economics

Meta Analysis of Empirical Deterrence Studies an explorative contest

Thomas Rupp

Nr. 174

Arbeitspapiere
des Instituts für Volkswirtschaftslehre
Technische Universität Darmstadt

Applied Research in Economics

Meta Analysis of Empirical Deterrence Studies

an explorative contest

Thomas Rupp*

Keywords: meta analysis, data mining, deterrence, criminometrics

JEL classification: C81, K14, K42

June 27, 2006

Abstract

A sample of 200 studies empirically analyzing deterrence in some way is evaluated. Various methods of data mining (stepwise regressions, Extreme Bounds Analysis, Bayesian Model Averaging, manual and naive selections) are used to explore different influences of various variables on the results of each study.

The preliminary results of these methods are tested against each other in a competition of methodology to evaluate their performance in forecasting and fitting the data and to conclude which methods should be favored in an upcoming extensive meta-analysis. It seems to be the case that restrictive methods (which select fewer variables) are to be preferred when predicting data ex ante, and less parsimonious methods (which select more variables) when data has to be fitted (ex post). In the former case forward stepwise regression or Bayesian Model Selection perform very well, whereas backward stepwise regression and Extreme Bounds Analysis are to be preferred in the latter case.

*Department of Applied Economics and Microeconometrics, Institute of Economics Darmstadt University of Technology
Tel.: +49 (0)6151 16-5512, Fax: +49 (0)6151 16-5652
email: rupp@vwl.tu-darmstadt.de

1 Introduction

Crime and deviant behavior matter to society. This insight has been common knowledge for millennia and theoretical concepts to deter undesired behavior with criminal prosecution were already developed centuries ago (e.g. the rational behavior approach by Beccaria (1819) from 1764 and Bentham (1830) from 1770 or thereabouts)[1]. By comparison, empirical studies verifying the effectiveness of criminal prosecution have emerged only recently. Following several isolated studies such as Michaels (1960), Schuessler and Slatin (1964) or Clarke (1966), numerous studies emerged after Becker's seminal paper (Becker, 1968) founding an economic theory of the behavior of rational offenders and Ehrlich's extension and empirical verification (Ehrlich, 1973). A current summary of the theory of public enforcement can be found in Polinsky and Shavell (2006).

Up to now more than 1000 studies which empirically verify the deterrent effect of the probability of apprehension and sanction (or similar determinants) have been published. This paper is part of an extensive meta-analysis of these studies and deals with the question of which determinants are most influential on the result of such a study, using a random sample of 200 of the acquired studies[2]. The available data base contains several hundred variables which, since there is no overall unifying theoretical underpinning available, could all be important in explaining the huge diversity of results found in the sample.

This study does not focus on the detailed determinants but is instead concerned with the question by which means significant determinants can be found. Several methods (stepwise regressions, Extreme Bounds Analysis, Bayesian Model Averaging, manual and naive selections) for evaluation and multiple ways to compare them are presented and how well the methods perform against each other is analyzed. This is in line with the findings of Fernández et al. (2001), but utilizes a unique data set in a field where interrelations are neither theoretically nor empirically assured and is thus truly explorative in nature. We focus on common methodology in data mining and well-known benchmarks of estimators to identify relevant information.

[1] The years in brackets refer to the cited versions whereas the other years refer to the first Italian publication (Beccaria) and the writing of the manuscripts (Bentham). The latter dates are taken from biographies of both authors.

[2] At the time this paper was written we had acquired approximately 500 studies. We expect to eventually acquire ca. 800 studies from which we will use 700 in the upcoming meta-analysis.

We find that methods which include fewer variables (forward stepwise regression, Bayesian Model Selection) are favored when (out of sample) predicting unknown data and less parsimonious methods (backward stepwise regression, Extreme Bounds Analysis) perform better when existing data has to be fitted (ex post prediction). Although our results may not be applicable for completely different data they may nevertheless be a guide for future meta-analyses in other fields or data sets with many variables.
The paper is organized in five sections. Section 2 describes the data and especially the endogenous variable which is to be estimated. Section 3 introduces and briefly describes the various methods which are used to estimate the data. Section 4 then explains the final weighting scheme used and the various loss-functions implemented to assess the estimation quality of the estimators. Finally section 5 tabulates the results of the tournament of methodology and concludes with an assessment of the findings.

2 The Data

Our data base consists of the information from 200 studies which empirically verify the deterrent effect of general (legal or social) threats to prevent individuals from committing offenses (i.e. general deterrence). These offenses include all kinds of crimes such as the classic types like homicide and petty thefts, as well as driving under the influence, tax evasion, environmental pollution, cheating in class and other offenses. In keeping with Beyleveld (1980), studies dealing primarily with specific deterrence (the effect of actual punishment on recidivists) are not considered[3] as well as studies dealing with antitrust-enforcement, immoral but not illegal behavior, animals or nuclear deterrence.
The acquisition of these studies and the recording of the data was undertaken by two different groups. One group, located in Heidelberg, was responsible for sociological (or similar fields) studies while we, located in Darmstadt, were responsible for economic (or similar fields) studies. The aforementioned 200 studies are a random sample of the approximately 500 studies we had acquired upon commencement of this paper. We expect to eventually retrieve approximately 800-900 out of about 1100 studies found (some, mostly old working papers, cannot be re-

[3]Exceptions are studies with individual data of released prisoners as long as recidivism was not the focus of the study.

trieved although they are certainly relevant). As of now our "data base" always refers to the information provided by those 200 studies.
Since the data acquisition is part of an extensive meta-analysis of empirical deterrence studies we wanted to record a large amount of information. Therefore we developed a set of 490 variables to catch all kind of characteristics of each observation[4] provided by each study. These variables can be partitioned into two separate parts: the first part, about 130 variables, covers all general information about the study itself (characteristics of the publication, the author, the kind of study, the utilized data, quality aspects of the study, etc.) while the second part captures characteristics of each observation (characteristics of the independent and dependent variables, the used explanatory variables, aspects of the model, detailed information of the results, etc.). To evaluate the data we merged both parts by duplicating the study variables for each observation. A study which provides n recorded results is thus represented by n rows in the date base - the first part of each row is exactly the same, the other part may be more or less different (depending on the results) as depicted in table 1.
To perform our meta-analysis we decided to resort to four different variables which evaluate the deterrence hypothesis: the t-values, the p-values (5 categories: $\leq 0.001, \leq 0.01, \leq 0.05, \leq 0.1, > 0.1$ and all other significance-measures (F-values, χ^2-values, z-values, etc.). Additionally, we always have the sign, i.e. whether the result supports (right sign) or rejects (wrong sign) the deterrence-hypothesis and the general opinion of the author for six different subsets of crime and deterrence (violent, property and other crimes - the probability and severity of sanctions). In this study we resort only to the t-values for reasons given in subsection 2.1.

2.1 Imputing t-Values

To include as many observations as possible in our analysis, we have to convert all significance-measures, when feasible, to a common scale avoiding loss of information. Since almost all can be converted to t-values and they are usually used in modern meta-analysis (Stanley, 2001) (since they are very well suited for regression methods), these are our favored candidates. The t-values are imputed in the following way:

[4]By observation we mean a result a study reports. For example, if a study tests the deterrence hypotheses for each of the seven index I crimes in the United States with one regression, it contributes seven observations to our data base.

- when a coefficient, its standard deviation and degrees of freedom is given, it is calculated exactly
- if not given, the degrees of freedom are approximated by the number of observations minus the number of explanatory variables
- p-values are transformed by the inverse t-distribution, choosing the p-value uniformly from its respective interval[5]
- all other values are transformed likewise by their exact p-value
- if the number of observations is missing, it is assumed to be 300

These transformations introduce further measurement errors but the advantages should outweigh arising disadvantages. The major problems are:

- The p-values are given in categories (e.g. $1\%, 5\%$), not in exact numbers. The underlying p-value can lie in any subset; thus the transformed absolute t-value will be underestimated[6].
- Almost no study gives the degrees of freedom but rather (if at all) the number of observations. The error arising from the (overestimated) degrees of freedom should be small. Nonetheless, this introduces a slight underestimation of the transformed t-values.
- Not all authors state explicitly whether the tests they use are one- or two-sided tests. We always treat those (rare) cases as two-sided tests, resulting in a slight overestimation of the transformed t-values.
- Some given t-values are only asymptotically t-distributed, but this should also pose only a minor problem[7].

In the remaining paper we will call all originally given t-values the *true t-values* and call the rest *imputed t-values*.

[5] A result which is significant on a 5% level is treated as being uniformly distributed in the interval $0.01 < p \leq 0.05$.

[6] For various reasons authors restrict themselves to a specific set of levels: therefore a p-value of 0.000234 might still be recorded as a significance on a 5% level.

[7] There is always the potential problem that any value is not distributed as stated when certain assumptions are violated but ignored or not tested.

2.2 Adjustment of Variables

Our data contains many variables with missing entries. There are two main reasons for this: either the information was not available from a study (e.g. whether the used data is representative, some characteristics of surveyed people, the year the data was gathered, etc.) or the information was not-applicable for the kind of study (e.g. survey characteristics for time-series studies, name of journal for books, etc.). Since we want to exclude as little data as possible, we treat missing or not applicable information as zero values. Excluding these observations would either result in zero observations (there are always variables not applicable to a study) or restricting the analysis to very narrow subsets (with rarely more than a few dozens observations). Imputing variables is only reasonable for specific subsets and would be very difficult, even in these subsets, for various reasons:

1. There are variables which could be imputed but every imputation method would be questionable (e.g. the nationality of the author or whether the used data is representative).

2. Variables are not independent. For example the used data set will be correlated with the nationality and the field of the authors.

3. It is not easy to identify the correct neighbors to generate the imputed values. To calculate the imputed values (e.g. the mean or correctly distributed random value) requires to identify all applicable observations. For example to impute various characteristics of surveyed people we would have to identify all similar survey-studies beforehand.

Thus we do not exclude observations with missing values at all but treat missing information as unique values[8].

2.2.1 Weighting

As mentioned before, the data recording was done by two groups. One concentrated on sociological studies, and we worked mostly with economic studies. Since many economic studies report several results to show that their results are robust to changes in the model, to contrast petty models or other reasons,

[8]There are some rare cases when a variable can take the value zero (e.g. the percentage of males in a sample) but these are negligible.

time restraints made it necessary to restrict ourselves to one[9] result per crime per study (otherwise we would have to record about 3.5 times as much). Rupp (2005a) showed that choosing one random result should be better than using the mean or median of all values.
This makes it necessary to weigh the observations in our data base in some way. In principle there are three different approaches from which we chose the last one:

1. Leave everything unchanged: use the unweighted observations (these may heavily bias the result if "our" studies differ systematically from the rest, which seems be evident.

2. Treat each observation equally: weight each observation in such a way that the sum of all weighted observations of each study is equal to the total amount of results it contains. This would be an approximation of the case in which we recorded all results and would bias our estimates in favor of those studies with many results.

3. Treat each study equally: weight every observation by the inverse number of the observations. If a study not recorded by us provides n observations it is weighted by $1/n$. A study recorded by us of which m out of n observations are in our data base, each is weighted by $1/m$. Therefore the sum of all weights of each study amounts to one.

Since "our" studies seem to differ significantly from the other (which can be readily appreciated by examining table 2) and the number of results per study varies substantially (from one to several hundred), we decided to use the latter weighting scheme.

2.3 Model Selection

In principal there are three different approaches to this problem:

1. dropping seemingly unimportant and keeping important variables,

2. use all variables and weight them according to their impact,

[9] When appropriate, we additionally recorded the favorite results of the author (if not already chosen randomly); but these are not studied here.

3. choose those variables which should be important on a subjective basis.

Before or after the application of such a method we can try to condense certain variables, e.g. using factor analysis.
Since we have no prior knowledge of the quality of any methods, we must assess the quality by statistical means. Therefore, we will test all methods in a tournament (consisting of out of sample prediction and fitting) and then evaluate which methods seem to be the most useful in our scenario.

3 Employed Methods

The literature on data mining provides us with many possibilities to retrieve any assumed information resulting from the data. We have chosen to implement a set of methods with methodologically different approaches. They can be divided into two categories:

1. Selection of variables. The method selects specific variables which are then used in a standard regression analysis.

2. Selection of the estimator. The methods produce an equation by which the endogenous variable can be estimated.

3.1 Naive Approach

The naive approaches we use here as the lowest benchmark are the mean t-value (i.e. the empty set of variables) and the full set of variables (except those which are dropped by the statistic package due to singularity problems).

3.2 Extreme Bounds Analysis (EBA)

The principle is to regress all possible combination of k (out of N) exogenous variables on the variable to be explained and to track the distribution of the associated t-values - see Leamer (1983), Leamer (1985) or Levine and Renelt (1992). Results are then derived from analyzing the distribution of these t-values. We have implemented three kinds of EBA:

- the strong (weak) sign test: the influence of a variable is important if all (a α quantile) of its t-values are of the same sign;

- the strong CDF-Test: a $(1-\alpha)$ confidence interval of the mean lies left or right from a critical α' confidence interval around zero.
- the extreme CDF-Test: the maximum (minimum) t-value is smaller (bigger) than a critical value (eg. ± 1.96).

Although EBA has been used in many cases ((McAleer and Veall, 1989),(Bartley et al., 1998),(Levine and Renelt, 1992) or Sala-I-Martin (1997)) it has several disadvantages:

- Computing the statistics of all $\binom{N}{k}$ combinations is computationally impossible for even a small amount of variables for large k. Our own ad-hoc implementation in STATA (StataCorp LP, 2006) requires one gigabytes of data per five million regressions and has a runtime of $O(N^k)$. Therefore, the largest k possible is 3, resulting in approximately 15 million regressions generating about 3GB of data ($k = 4$ would result in 1671 million regressions, taking 113x more time to run and would generate ca. 337GB of data; even a supposed optimization of the algorithm would make this not feasible).
- Errors of the second type increase with the number of observations (in this case combinations). Further information on the vote-counting problem can be found in Hedges and Olkin (1985).
- It is an unresolved problem whether the calculated t-values should be weighted (e.g. with the ML of the respective model, refer to Sala-I-Martin (1997)) – this can improve the conclusions (minimizing the influence of obviously improper models) or dampen them (since all models, whether high ML or not, will suffer from severe omitted variable bias).

Results from EBA could be further analyzed by applying a Response Surface Analysis (RSA) as sketched by (Florax, 2001) but this was not done here due to computational constraints (in combination with an EBA the algorithm has a runtime of $O(N^{k+1})$).

3.3 Stepwise Regression

We use the algorithm implemented in STATA (StataCorp LP, 2006). Basically it starts either with all (or no) variables and drops insignificant and includes sig-

nificant variables with subsequent re-estimation until there are no more variables to be in- or excluded.
We set the level of inclusion p_1 to 0.001 to account for selection bias (rule of thumb according to Lovell (1983)) and the exclusion level p_2 to 0.1 since there are 448 possible candidates. A first regression declared 37 to be significant on a ten percent level, so we calculated the adjusted significance levels to be $37/448 \cdot 0.05 \approx 0.004$ (inclusion) and $200/448 \cdot 0.2 \approx 0.1$ (exclusion; the desired maximum number of 200 was chosen arbitrarily). 59 variables were dropped due to singularity problems.
Hendry and Krolzig (2000) showed that these adjustment can be avoided when the data mining algorithm accounts for the selection bias as is the case with PcGets (Hendry and Krolzig, 2006) but since this accumulation of statistical tools didn't achieve any better results[10] we resorted to the well studied stepwise regression methods.

3.3.1 Backward Stepwise Regression

The basic procedure is rather simple:

1. start with the full model,
2. exclude the least significant included variable if its p-value is above p_2,
3. include the most significant excluded variable if its p-value is below p_1,
4. exclude the least significant included variable if its p-value is above p_2,
5. reestimate and repeat steps 3 to 4 until neither is possible.

3.3.2 Forward Stepwise Regression

The forward procedure will typically find less variables than the backward procedure but is methodologically almost identical:

1. start with the empty model,
2. include the most significant excluded variable if its p-value is below p_1,

[10] One reason may be that it does fit time series models better than this data set of very heterogenous cross-section.

3. exclude the least significant included variable if its p-value is above p_2,
4. include the most significant excluded variable if its p-value is below p_1,
5. repeat steps 3 to 4 until neither is possible.

3.4 Bayesian Model Selection (BMS)

The basic idea of calculating the probabilities of all models using Bayes Theorem and choosing (one of the) the most probable is quite simple and intuitive (see Raftery et al. (1997) and Hoeting et al. (1999)).

1. calculate for model Δ the probabilities over all possible models M_k, given the data D (δ_k is the vector of model parameters of Model M_k)

$$\begin{aligned} P(\Delta|D) &= \sum_{k=1}^{K} P(\Delta|M_k, D)P(M_k|D), \\ P(M_k|D) &= (P(D|M_k)P(M)) \,/\, \left(\sum_{l=1}^{K} P(D|M_l)P(M_l)\right), \\ P(D|M_k) &= \int P(D|\delta_k, M_k)P(\delta_k|M_k)d\delta_k, \end{aligned}$$

2. chose the models with the highest posterior probability.

Nonetheless its implementation poses several difficulties (also see Koop and Potter (2003) or Chipman et al. (2001)). We used the implementation in R (R Development Core Team, 2006).

1. The quality of the results hinges on the selection of the hyper-parameters necessary for the calculations. They can be chosen manually, be calculated from the data or simply set on trivial values. Since we do not possess any usable information about the priors, we chose to take uninformative priors.
2. In some cases the conditional probability of the data $P(D|M_k)$ cannot be calculated and has to be approximated.
3. The runtime depends crucially on the number of possible models. Without prior informations and even restricted to the simple in/exclusion of N variables 2^N models have to be taken into account.

3.5 Bayesian Model Averaging (BMA)

BMA is in principle the same as BMS with the exception of using the information of all models with the weighted (with their posterior model probability) average of each coefficient instead of the coefficient of the most probable model.
Again there are several disadvantages to cope with: the previously mentioned hyper-parameters can affect the quality of the predictions (without prior knowledge uninformative priors have to be used or they are optimized according to some criteria), the huge model space incorporates all 2^N models and therefore optimization algorithms and monte carlo methods have to be employed in our case (depending on the individual case, normally 40-50 variables can be used without monte carlo algorithms).

3.6 Manual Selection

We further chose a set of variables based on our own assessment. Since theoretically derived variables are scarce in the field of meta-analysis, most of the chosen variables stem from an ad-hoc selection which seems reasonable.

3.7 Other Methods

The methods above are essentially all linear regressions. Other, non-linear methods should be useful for verifying any conclusions. The methods below were tested only on a smaller data set and we did not include them here. On the one hand, it would be difficult to compare their performance[11] and additionally it would have been very time consuming to prepare the data and perform the calculations.

3.7.1 RSDA

The methods described above are all some kind of regression. It would surely be helpful if other methods, completely different from the principle of minimizing noise, would lend support to any results produced by the various regressions methods.

[11] Although they can be used as selection algorithms this would not fully exhaust their potential.

We have shown in Rupp (2005b) that Rough Set Data Analysis (RSDA) is applicable in principle. The method - refer for example to Pawlak (1984), Pawlak (1991) or Düntsch and Gediga (2000) - relies on finding repeated patterns (rough sets) across the complete data matrix in the nominal data space (thus all ordering information is not considered). During this process the method produces information of the importance of the variables thus indicating which variables should be further studied and which can be neglected.
Since the runtime depends crucially on the number and structure of the variables they should be recoded to fit the method. Since this process and the adjacent analysis are very time consuming, the application was postponed and will be undertaken with the final data set (which will include the data of additional 500 studies).

3.7.2 Decision Trees

Decision trees (Quinlan, 2003) have also been evaluated but the results indicated that there were not enough data to draw any reasonable conclusions. Therefore the implementation has also been postponed to the final data set, which will include approximately four times as many observations.

4 Assessing Prediction Quality

To study the quality of the various methods we randomly partition the data into a training and test-set of equal size. We use (true and imputed) t-values as endogenous variable; we can use 1961 observations from the 200 available studies. We chose the size of 50% for the test-set instead of the usual size of about 10% because otherwise the test-set would only contain observations from a few different studies. Each method described in section 3 is then used to estimate the data and resulting values of various loss-functions (see section 4.1) are recorded.
This procedure is repeated nine times; thus we have ten observations for each loss-function of each method. Since the Bayesian Model Averaging methods required 12-36 hours for each run (with a maximum of 50 variables in each submodel) we did not consider it for additional runs. All estimators (except the bayesian) were computed using weighted OLS and were computed with STATA (StataCorp LP, 2006). All bayesian methods were computed with R (R Development Core Team,

2006).

4.1 Loss-Functions

We employed a wide variety of loss functions to distinguish various characteristics of the different methods of analysis. Each loss function has its own justification. Most of them are symmetrical (punishing deviations in both directions likewise) while asymmetrical loss functions are also possible (e.g. punishing upwards deviations less than those downwards or negative more than positive) but are, at least to some extent, arbitrary.
We calculated the 95% confidence intervals of the mean loss function values to see whether any method is significantly superior to the naive estimators.

- RMSE: the root mean squared error: $\sqrt{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$.

  The most commonly used loss function. The lower its value, the better the estimates are.

- cor.: the Pearson correlation between y and $\hat{y}$. There shouldn't be any negative values; the closer to one the better the estimates are.

- U: Theil's (new) inequality coefficient $\sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N} y_i^2}}$ and its decomposition

  - U.bias $= \frac{(\bar{y}-\bar{\hat{y}})^2}{\frac{1}{N}\sum_{i=1}^{N}(y-\hat{y})^2}$,
  - U.var $= \frac{(s_y-s_{\hat{y}})^2}{\frac{1}{N}\sum_{i=1}^{N}(y-\hat{y})^2}$,
  - U.cov $= \frac{2(1-\rho)s_y s_{\hat{y}}}{\frac{1}{N}\sum_{i=1}^{N}(y-\hat{y})^2}$.

  U should be zero in the case of perfect, and one in the naive, estimation (any values above one indicate that the estimator is worse than the naive estimator). The estimation errors are divided into U.bias (systematic error in the mean value), U.var (systematic error in the variance) and U.cov (unsystematic random error). These should add up to one (except for rounding errors). The perfect estimator has a U.cov of one.

- RMSPE: the root mean squared proportional error: $\sqrt{\sum_{i=1}^{N}(\frac{y_i-\hat{y}_i}{y_i})^2}$.

  Similar to RMSE but measures the error relative to the true values.

- CI.hit: the fraction of predicted values in a $c \cdot s_y$ confidence interval of y. In our case we set c to 0.5.

- Signed: $\sum_{i=1}^{N} 1_{[y>\hat{y}]}$ should be Binomial$(N, 0.5)$ distributed when the errors are distributed randomly around zero.

  The values should not be close to any significant value - in principle the larger the better although any value above 0.2 should be sufficient to reject any systematic unbalance.

- neg2pos4: a loss function which punishes large deviations in the case of positive values much harder: $\sqrt{\sum_{i=1}^{N} \left(1_{[y<0]}(y-\hat{y})^2 + 1_{[y>0]}(y-\hat{y})^4\right)}$

  We implemented this loss function since we have an abundance of negative but only relatively few positive values. With this function we can see if an estimator fares better with positive values.

- fsRMSE: false sign root mean squared error: $\sqrt{\sum_{i=1}^{N} 1_{[y_i \cdot \hat{y}_i<0]}(y_i-\hat{y}_i)^2}$.

  This function is very similar to the RMSE-function but only punishes those estimations which carry the wrong sign. We implemented this because the sign of an estimation is, to some degree, even more important than the extent of a deviation.

- Min. dev. and Max. dev.: the maximum $\max(y, \hat{y})$ and minimum $\min(y, \hat{y})$ deviation.

- Mean pos.: the mean positive deviation $\overline{1_{[\hat{y}_1>y]} \cdot |y_i - \hat{y}_i|}$,

- Mean neg.: the mean negative deviation $\overline{1_{[\hat{y}_1<y]} \cdot |y_i - \hat{y}_i|}$,

- Mean abs.: the mean absolute deviation $\frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$.

Additionally we performed a general encompassing test (Clements and Harvey, 2004) by regressing $y = \sum_{i=1}^{\#E} \beta_i \hat{y}^{(1)} + \varepsilon$ and analyzing the calculated coefficients cEncomp. and the respective p-values pEncomp. ($\#E$ is the number of competing estimators). Good estimators should have coefficients near one and low p-values since they should contain most of the required information.

5 The Tournament

Our final data set consists of 448 variables which are either continues or indicator variables with 1961 observations. The variation of each variable can be very different (from almost none to the uniqueness of each observation). Performing a factor analysis beforehand did not bring forth any usable results and was therefore neglected.

5.1 The Competitors

We gathered all estimators (see section 3)into different groups:

naive The naive approaches which select either none or all variables.

- **SET0** The naive estimator - the mean weighted t-value of all observations (i.e. a regression with the constant only).
- **SET1** All 448 variables are used (except those dropped because of singularity problems).

selective We select sets of variables following simple rules.

- **SET2** All remaining variables after removing the collinear variables and those with a p-value ≥ 0.2; 133 variables remained.
- **SET3** A set of 192 manually selected variables (which are considered important by the authors).

stepwise Backward and forward regressions. With regard to (bib) the inclusion criteria was a p-value below 0.004 and the exclusion criteria was a p-value above 0.1.

- **SET4** Backward stepwise regression (starting with all variables) to determine the sets of variables; 282 variables selected.
- **SET5** Forward stepwise regression (starting with no variables) to determine the sets of variables; 71 variables selected.
- **SET11** Full forward stepwise regression in every run (thus not only the variables but also the coefficients are determined by stepwise regressions; this is implemented only in the case in which less than the full

data set was used as the training set). 44 to 62 variables were selected (depending on the actual run).

EBA Extreme Bounds Analysis with different inclusion criteria. All variables with a gini-coefficient which do not lie within a 95% CI of the mean gini-coefficient are excluded beforehand[12].

SET6 weak Extreme Bounds Analysis: the 1%- and 99% quartile of the t-value-distribution have the same sign. 282 variables were selected with this selection criteria.

SET7 strong Extreme Bounds Analysis: all t-values share the same sign and the 1%- and 99% quartiles are smaller (larger) than -1.96 ($+1.96$). 86 variables were selected with these criteria.

SET8 extreme Extreme Bounds Analysis: all t-values are smaller (larger) than -1.96 ($+1.96$). 31 variables were selected with this criteria.

Other selection criteria were tested but these were chosen in order to achieve the desired amount of variables (one large, medium and small set of variables).

BMA Bayesian Model Selection and Averaging. Due to computational limitations only 50 variables were allowed to be included in any submodel at any time. The best 15 models were selected.

SET9 All variables are selected with a posterior probability ≥ 0.95 after a Bayesian Model Averaging procedure. 40 variables were selected.

SET10 Full Bayesian Model Averaging in every run. 49 variables were selected (the constant is the 50th).

5.2 The Contest

We use three different methods to compare the quality of the various estimators. In table 3 we see how well the estimators fare to predict the data. We randomly selected 50% of the data to calculate the regression coefficients (the estimators)

[12]The gini-coefficient was calculated as the quotient of mean and median t-value. All variables which lie outside an interval of twice its standard deviation around the mean gini-coefficient are excluded since these results seem to be unreliable.

and estimated the remaining data (the estimation). We chose to use 50% (instead of $80 - 90\%$ which is more common) because observations from the same study already share many characteristics. This procedure was repeated ten times - the mean values are shown in the table. The bold cells emphasize the best value in a row while the light (dark) grey cells mark those entries with means which are significantly (5%) better (worse) than the best of the naive approaches.
It is obvious that the forward stepwise regression performs very well compared to all other methods. It performs best in most criteria and especially has the highest correlation, best hit ratio and is uniquely favored by the encompassing test. Only the sign-test indicates that the estimation error is systematically biased (although U.bias is not conspicuous). The BMS method seems to come second; although its explanatory power (correlation, R^2, U.cov) seems to be weaker it has only small mean deviations (especially the mean absolute deviation and the root mean squared percentage error). It is interesting to note that performing stepwise regressions on the individual partitions is worse than doing OLS on the stepwise-selected variables to such an extent. The same applies to the difference between BMS and BMA; it should be mentioned that the means given for BMA inhibit a very large standard deviation - in some runs it performed very well, in others extraordinarily badly.

In table 4 we test how well the estimators are in reproducing the data. The whole data set is used to establish the estimator, and all t-values are then re-estimated. Bold cells are the best entries in each row.
SET6, the variables selected by the weak EBA, seems to perform best. It is somewhat surprising that the encompassing test favors the backward stepwise regression method although its remaining benchmarks seem to be quite poor (the only exceptions are the mean deviations and the root mean squared error of the predictions with false signs). This might be an artifact since the encompassing test includes three methods all with a large number of variables which "play" in opposite directions. The following table will clarify to what extent the methods outperform the naive estimators.

Analogously to table 4 we use the full data to calculate the estimators but estimate only 50% chosen randomly from the data. This is repeated ten times. The mean values are shown in table 5. The same notation as in table 3 applies. This was done to study whether any method performs significantly better than the naive methods.

Naturally the values are very similar to those in table 4. The same arguments apply here. The variables selected by the weak EBA perform significantly better than the naive approaches in most categories. The backward stepwise regression method is favored by the encompassing test but can outperform the naive approaches only in fsRMSE test.

5.3 And the Winner...

depends on the aim of the researcher. Shall the estimator fit (ex post prediction) the existing data as well as possible? Or should the estimator predict (ex ante) unknown[13] data?

One general conclusion seems to be that selecting fewer variables is better for predicting but worse for fitting the data. Bayesian Model Averaging seems to be particularly unsuited to predicting but performs adequately in fitting the data. This may come from utilizing too much detailed information from some studies which are rather specialized and not suited to be used for other studies because Bayesian Model Selection, which is inherently similar to BMA, performs quite good in predicting the data and only slightly worse in fitting them. In our scenario BMA does not perform as well as in Fernández et al. (2001) who compared the naive estimator, one EBA version and BMA in the case of a moderately sized set of variables in the country growth context. Since they concluded that BMA was superior to EBA, this suggests that there is no rule of thumb when to use which method.

Stepwise regressions are suited for both (forward to predict, backward to fit). The same argument as for the Bayesian approaches may be applied here. Not only does EBA perform moderately well in fitting the data, but the least parsimonious EBA performs very well in fitting the data and is many respects better than the stepwise regressions. BMS, as mentioned before, performs very well in predicting the data but only moderately in fitting them. The contrary applies to the manual selection which actually performs only moderately in fitting and is quite poor in predicting the data - we must mention that we did not test a reduced set of manually selected variables which probably would have performed better in predicting the data.

[13]Since many predicted observations come from a study which is also used to calculate the estimator, the unknown data is not really completely unknown.

Since the out of sample prediction is the commonly preferred way to judge a forecasting method, we come to the following conclusions:

our data To judge the importance of the variables we prefer forward stepwise regression and BMS for selecting those variables which should be studied further. A subsequent analysis could then prove if these are really suited for building a model, drawing robust conclusions from it and for judging the influence of these determinants on the results of a deterrence study.

general data When dealing with data sets with many variables in an exploratory context it seems advisable to employ methods which resort to s small selection of really important variables. Including too many variables seems to dilute any findings. It may not be a good idea to rely on "expert opinions" in selecting the variables when they are not derived from an overall unifying theoretical underpinning (and even then important variables may be missed).

The set up presented here is not hard to implement (except for the EBA method which has still not found its way to standard statistical packages) but may be very time consuming when many variables are involved. This makes it even more important to decide before an exploratory analysis how to determine which variables are to be taken into account or how the estimator is to be selected.

References

Bartley, W. A., M. A. Cohen, and L. Froeb (1998). "The Effect Of Concealed Weapons Laws: An Extreme Bound Analysis." *Economic Inquiry.* 36 (2), 258–365.

Beccaria, C. B. M. (1819). *Of Crimes and Punishment*, Philip H. Nicklin, Philadelphia.

Becker, G. S. (1968). "Crime And Punishment: An Economic Approach." *Journal Of Political Economy.* 76 (2), 169–217.

Bentham, J. (1830). *The Rationale of Punishment*, Robert Heward, London.

Beyleveld, D. (1980). *A Bibliography on General Deterrence*, Saxon House, Teakfield Limited.

Chipman, H., E. I. George, and R. McCulloch (2001). "The Practical Implementation of Bayesian Model Selection." *IMS Lecture Notes.* 38, 65–134.

Clarke, R. V. G. (1966). "Approved School Boy Absconders And Corporal Punishment." *British Journal of Criminology.* 6 (4), 364–375.

Clements, M. P. and D. I. Harvey (2004). "Forecast Encompassing Tests and Probability Forecasts." Discussion Paper, School of Economics, University of Nottingham, Nottingham.

Düntsch, I. and G. Gediga (2000). "Rough Set Data Analysis." *Encyclopedia of Computer Science and Technology.* 43, 281–301.

Ehrlich, I. (1973). "Participation In Illegitimate Activities: A Theoretical And Empirical Investigation." *Journal Of Political Economy.* 81 (3), 521–565.

Fernández, C., E. Ley, and M. F. J. Steel (2001). "Model Uncertainty in Cross-Country Growth Regressions." *Journal of Applied Econometrics.* 16 (5), 563–576.

Florax, R. J. (2001). "Methodological Pitfalls in Meta-analysis: Publication Bias." Research Memoranda 0028, Faculty of Economics, Business Administration and Econometrics, Free University Amsterdam, Amsterdam.

Hedges, L. V. and I. Olkin (1985). *Statistical Methods for Meta-Analysis*, Academic Press, San Diego.

Hendry, D. F. and H. M. Krolzig (2000). "Computer Automation of General-to-Specific Model Selection Procedures." Working Paper, Institute of Economics and Statistics and Nuffield College, Oxford.

——— and ——— (2006). *PcGets: General-to-Specific model selection*, Timberlake Consultants Ltd.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). "Bayesian Model Averaging: A Tutorial." *Statistical Science.* 14 (4), 382–417.

Koop, G. and S. Potter (2003). "Forecasting in Large Macroeconomic Panels Using Bayesian Model Averaging." Staff Report 163, Federal Reserve Bank of New York, New York.

Leamer, E. E. (1983). "Let's Take The Con Out Of Econometrics." *American Economic Review.* 73 (1), 31–43.

——— (1985). "Sensitivity Analyses Would Help." *American Economic Review.* 75 (3), 308–313.

Levine, R. and D. Renelt (1992). "A Sensitivity Analysis of Cross-Country Growth Regressions." *American Economic Review.* 82 (4), 942–963.

Lovell, M. C. (1983). "Data Mining." *Review of Economics and Statistics.* 65 (1), 1–12.

McAleer, M. and M. R. Veall (1989). "How Fragile Are Fragile Inferences? A Re-Evaluation Of The Deterrent Effect Of Capital Punishment." *The Review Of Economics And Statistics.* 71 (1), 99–106.

Michaels, R. M. (1960). "The Effects Of Enforcement On Traffic Behaviour." *Public Roads.* 31 (5), 109–113.

Pawlak, Z. (1984). "Rough Classification." *International Journal of Man-Machine Studies.* 20 (5), 469–483.

——— (1991). *Rough Sets - Theoretical Aspects of reasoning about data*, Kluwer Academic Publisher.

Polinsky, M. A. and S. M. Shavell (2006). "Public Enforcement of Law." Working Paper 322, John M. Olin Program in Law and Economics.

Quinlan, J. R. (2003). "Induction of Decision Trees." *Machine Learning.* 1 (1), 81–106.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association.* 92 (437), 179–191.

Rupp, T. (2005a). "Multiple Measures in a meta analysis - guidelines in special cases." unpublished Working Paper, Department of Economics, Darmstadt.

——— (2005b). "Rough Set Methodology in Meta-Analysis. A Comparative and Exploratory Analysis." Darmstadt Discussion Papers in Economics 157, Department of Economics, Darmstadt.

Sala-I-Martin, X. X. (1997). "I Just Ran two Million Regressions." *American Economic Review.* 87 (2), 178–183.

Schuessler, K. F. and G. Slatin (1964). "Sources Of Variation In U.S. City Crimes, 1950 and 1960." *Journal Of Research In Crime And Delinquency.* 1, 127–148.

Stanley, T. D. (2001). "From Wheat to Chaff: Meta-Analysis as Quantitative Literature Review." *Journal of Economic Perspectives.* 15 (3), 131–150.

StataCorp LP (2006). *STATA, Release 9.*

Table 1: Excerpt of the data base

variables	observations				
publication type	journal	journal	journal	...	journal
author	Steven D. Levitt	Steven D. Levitt	Steven D. Levitt	...	Steven D. Levitt
journal	Economic Inquiry	Economic Inquiry	Economic Inquiry	...	Economic Inquiry
publication year	1998	1998	1998	...	1998
author country	USA	USA	USA	...	USA
⋮	⋮	⋮	⋮	⋱	⋮
time span of data (months)	264	264	264	...	264
data base	UCR	UCR	UCR	...	UCR
quality tests	no	no	no	...	no
major problems detected (by reader)	no	no	no	...	no
number of results	84	84	84	...	84
independent variable	arrest rate	arrest rate	arrest rate	...	arrest rate
dependent variable	crimes/pop.	crimes/pop.	crimes/pop.	...	crimes/pop.
crime	murder, homicide	larceny	assault	...	auto theft
number of explanatory variables	8	8	9	...	9
error correction in model	yes	yes	no	...	no
method used	OLS	OLS	OLS	...	OLS
⋮	⋮	⋮	⋮	⋱	⋮
coefficient	-0.03	-0.154	-0.365	...	-0.457
sd of coefficient	0.033	0.021	0.122	...	0.225
t-value				...	
number of observations	819	819	963	...	963

The first column represents an (arbitrarily chosen) sample of the available variables in our data set. The other columns represent the observations from our data (thus, one column in the table stands for one row in our data base) and were drawn randomly from one study. The first block contains the general information about the study and the second block covers information on the individual results detailed in the study.

Table 2: Weighting t-values and source of data

source	Obs.	Mean	sd.	Min	Max
Darmstadt, unweighted	584	−1.498709	2.517713	−14.59	7.83
Heidelberg, unweighted	1377	−1.072549	1.789821	−13.18	11.7
both, unweighted	1961	−1.199463	2.042657	−14.59	11.7
Darmstadt, weighted	584	−1.826989	2.773109	−14.59	7.83
Heidelberg, weighted	1377	−.8748881	2.518862	−13.18	11.7
both, weighted	1961	−1.356271	2.691893	−14.59	11.7

The rows of the unweighted data refer to the first weighting scheme: leave everything unchanged. The rows of the weighted data refer to the third weighting scheme which weights each study equally. Naturally the number of observations and the extreme values are not affected by weighting.

Table 3: How well the models predict the data

Method	$SET0$	$SET1$	$SET2$	$SET3$	$SET4$	$SET5$	$SET6$	$SET7$	$SET8$	$SET9$	$SET10$	$SET11$
# variables	0	448	133	192	282	71	282	86	31	40	49	$44-62$
RMSE	2.1597	645.326	2.51356	2.66033	380.793	**1.95323**	3.90513	2.12179	2.18379	2.11638	30.7854	2.35542
cor.	$3.8 \cdot 10^{-16}$	.015592	.313287	.306828	.015631	**.500925**	.209183	.350409	.268085	.338889	.172424	.318585
U	.85261	253.688	.992844	1.05217	148.28	**.7713**	1.55137	.837739	.862324	.835619	12.2477	.93045
U.bias	.00686	.002708	.002719	.002255	.001898	.003585	.002232	**.00149**	.002009	.002431	.002523	.003829
U.var.	.994243	.985152	.031387	**.019047**	.972338	.07025	.150404	.128107	.178376	.165979	.346649	.031025
U.cov.	0	.013247	.967001	**.979805**	.026872	.927271	.848472	.871511	.820723	.832697	.651936	.966252
rmspe	462.587	917.151	460.415	544.888	843.584	289.424	702.431	816.514	953.844	**231.248**	255.762	326.997
CI.hit	.433569	.399063	.432941	.419064	.452002	**.480113**	.437235	.42977	.42035	.473008	.459976	.42092
Signed	.518773	.550458	.227121	.41579	.49936	.033165	.446462	**.701708**	.420886	.549979	.440751	.439486
neg2pos4	**7.52687**	113244	11.533	13.0819	73684.5	7.53168	42.1987	8.23439	7.63642	7.92112	36.0835	8.98274
fsRMSE	**7.6841**	$3.2 \cdot 10^{6}$	11.5883	12.6364	725374	8.17034	32.4847	8.00736	8.60977	8.58787	3408.39	9.82621
Min. Dev.	−11.9697	−4499.61	−24.8017	−26.8252	−4237.35	**-11.1217**	−40.2951	−12.3844	−11.3944	−12.2562	−617.945	−18.8087
Max. Dev.	12.6373	13000.5	15.5927	13.1828	5792.8	**11.601**	38.2468	12.2886	11.6505	15.5094	449.749	12.8751
mean pos.	1.7022	75.9276	1.61637	1.78615	41.0449	**1.38239**	1.98678	1.54906	1.65816	1.49506	2.61987	1.66908
mean neg.	**-1.43378**	−54.7914	−1.6853	−1.77028	−37.6699	−1.44359	−1.90764	−1.58134	−1.54768	−1.47992	−4.10008	−1.63594
mean abs.	1.57015	66.1303	1.65008	1.77965	39.5358	**1.41072**	1.94926	1.56572	1.60399	1.48825	3.33958	1.6548
cEncomp.	0	.000441	.037557	.057533	−.000052	**.524227**	.009281	.075762	−.02504	.225156	−.025984	−.058645
pEncomp.	.	.327562	.313994	.207189	.618509	$7.0 \cdot 10^{-11}$	.393009	.251078	.441696	.123736	.247644	.302339

10 runs, random 50% of the data used to fit model, remaining 50% is estimated

SET0: no variables, SET1: all variables, SET2: all variables with < 0.2 sign. after removing collinear, SET3: manually selected, SET4: stepwise, backward, SET5: stepwise, forward, SET6: weak EBA, SET7: strong EBA, SET8: extreme EBA, SET9: BMS50 - all with post-prob. > 0.95, SET10: BMA50, SET11: stepwise forward, individual runs.

Light/dark grey cells: mean is 5%-significantly better/worse than the best of SET0 or SET1. Bold cells are the best of each row.

Table 4: How well the models fit the data

Method	$SET0$	$SET1$	$SET2$	$SET3$	$SET4$	$SET5$	$SET6$	$SET7$	$SET8$	$SET9$	$SET10$
# variables	0	448	133	192	282	71	282	86	31	40	49
RMSE	2.12379	9.37319	1.8667	1.90284	5.32825	1.81386	**1.6975**	1.90655	2.0988	1.91472	1.87256
cor.	$8.8 \cdot 10^{-16}$	.198046	.519792	.519887	.284281	.55737	**.616994**	.472324	.325875	.45016	.482237
U	.84821	3.74352	.745534	.759967	2.12803	.72443	**.677958**	.76145	.838231	.764711	.747873
U.bias	.000196	**.000032**	.000096	.000351	.000034	.000475	.000072	.000065	.000257	.0015	.000734
U.var.	1.00035	.629709	.09671	**.055002**	.408353	.080916	.092438	.158316	.152992	.227431	.227068
U.cov.	0	.370807	.903742	**.945196**	.592162	.919157	.908039	.842168	.847299	.771617	.772747
rmspe	509.943	592.371	237.27	319.239	485.022	317.617	529.783	699.437	994.416	258.253	**224.927**
CI.hit	.438837	**.595173**	.4904	.484915	.584202	.502468	.551838	.460779	.4339	.46407	.493143
Signed	**.944391**	.51868	.010846	.021979	.646066	.003784	.574311	.353933	.574311	.696809	.303191
neg2pos4	7.62725	1415.44	6.8035	6.8786	447.548	6.55477	**5.43282**	6.27048	6.89062	6.65784	6.56063
fsRMSE	8.49818	5.74615	6.93651	7.31405	**4.99875**	7.22932	5.72572	6.61779	8.18255	6.44313	6.38583
Min. Dev.	−12.9963	−300.298	−10.6294	**-10.4613**	−162.723	−11.3218	−10.7544	−11.3153	−12.2426	−10.8011	−10.7838
Max. Dev.	13.2337	245.832	12.7149	**11.1167**	138.228	11.3782	11.1909	12.0875	11.9196	12.6964	12.7962
mean pos.	1.58636	1.41132	1.30871	1.36405	1.2835	1.27154	**1.2305**	1.41717	1.5881	1.35885	1.30193
mean neg.	−1.53128	−1.51608	−1.42346	−1.42916	−1.32429	−1.35834	**-1.19229**	−1.4159	−1.50891	−1.47451	−1.43794
mean abs.	1.5578	1.46373	1.36291	1.39502	1.30406	1.3122	**1.21132**	1.41654	1.54835	1.41735	1.36908
cEncomp.	0	−.497241	.035817	.092822	**.929551**	.042404	.251942	.077197	−.046844	−.024083	.093802
pEncomp.	.	$9.7 \cdot 10^{-29}$	.388512	.012472	$3.3 \cdot 10^{-32}$	.363834	$5.6 \cdot 10^{-7}$	.181996	.323632	.794452	.333591

100% of the data used to fit model, 100% is estimated

SET0: no variables, SET1: all variables, SET2: all variables with < 0.2 sign. after removing collinear, SET3: manually selected, SET4: stepwise, backward, SET5: stepwise, forward, SET6: weak EBA, SET7: strong EBA, SET8: extreme EBA, SET9: BMS50 - all with post-prob. > 0.95, SET10: BMA50.

Bold cells are the best of each row.

Table 5: How well the models fit the data

Method	*SET*0	*SET*1	*SET*2	*SET*3	*SET*4	*SET*5	*SET*6	*SET*7	*SET*8	*SET*9	*SET*10
# variables	0	448	133	192	282	71	282	86	31	40	49
RMSE	2.15281	7.07542	1.88113	1.91885	4.22137	1.8198	**1.70719**	1.9318	2.12044	1.93119	1.89042
cor.	$4.1 \cdot 10^{-16}$	.325319	.523688	.522374	.40569	.565711	**.622271**	.469792	.325518	.455287	.486128
U	.849902	2.77664	.742957	.757775	1.6584	.718691	.674062	.762842	.837272	.762507	.746375
U.bias	.00053	.000626	.00109	.001262	.000568	.001481	.00072	**.000472**	.000984	.001943	.001249
U.var.	1.00058	.410092	.104539	**.060694**	.258544	.086906	.100741	.168926	.165053	.241193	.240162
U.cov.	$8.2 \cdot 10^{-17}$	.590391	.89548	**.939152**	.741997	.912721	.899648	.831712	.835072	.757971	.759698
rmspe	518.59	597.179	242.861	324.773	487.706	322.803	534.897	707.721	1002.79	264.412	**230.62**
CI.hit	.434168	**.592404**	.490119	.478825	.584274	.503357	.551704	.456408	.433651	.459622	.491738
Signed	**.856363**	.45928	.046916	.064297	.53218	.02264	.573695	.419755	.498562	.714348	.477502
neg2pos4	7.85069	616.526	6.6941	6.82259	197.222	6.32728	**5.31369**	6.25211	6.95269	6.64895	6.577
fsRMSE	8.47182	5.89845	6.99919	7.29771	**5.04938**	7.26179	5.694	6.65935	8.20988	6.35465	6.37436
Min. Dev.	−12.1293	−145.284	−9.81253	−9.90602	−80.2717	−9.71133	**-9.55075**	−10.1628	−10.84	−10.3719	−10.4169
Max. Dev.	12.4777	81.4712	11.7138	**10.4892**	49.1276	10.9443	10.6293	11.3734	11.3387	11.7103	11.7569
mean pos.	1.61878	1.31646	1.32882	1.38361	**1.23089**	1.27582	1.24533	1.43593	1.59944	1.3796	1.32077
mean neg.	−1.52952	−1.49248	−1.41253	−1.44204	−1.31656	−1.35983	**-1.18917**	−1.42612	−1.51924	−1.47357	−1.43091
mean abs.	1.57235	1.40192	1.36815	1.41078	1.27304	1.31504	**1.21715**	1.43103	1.55971	1.42765	1.37578
cEncomp.	0	−.493387	.032839	.093846	**.927143**	.074551	.24677	.049544	−.043902	−.01659	.08829
pEncomp.	.	$3.7 \cdot 10^{-7}$	.450232	.141248	$3.9 \cdot 10^{-12}$	.318881	.001646	.441695	.505569	.564335	.378019

10 runs, 100% of the data used to fit model, 50% is estimated

SET0: no variables, SET1: all variables, SET2: all variables with < 0.2 sign. after removing collinear, SET3: manually selected, SET4: stepwise, backward, SET5: stepwise, forward, SET6: weak EBA, SET7: strong EBA, SET8: extreme EBA, SET9: BMS50 - all with post-prob. > 0.95, SET10: BMA50.

Light/dark grey cells: mean is 5%-significantly better/worse than the best of SET0 or SET1. Bold cells are the best of each row.

ISSN: 1438-2733