Data Mining

Michael C. Lovell

The Review of Economics and Statistics, Vol. 65, No. 1. (Feb., 1983), pp. 1-12.

Stable URL:
http://links.jstor.org/sici?sici=0034-6535%28198302%2965%3A1%3C1%3ADM%3E2.0.CO%3B2-7

The Review of Economics and Statistics is currently published by The MIT Press.

The Review *of* Economics *and* Statistics

VOL. LXV FEBRUARY 1983 NUMBER 1

DATA MINING

Michael C. Lovell*

I. Introduction

THIS paper investigates certain consequences of data mining, a research paradigm that masquerades under a variety of aliases. Applied econometricians describing their own efforts are inclined to speak of "experimentation." "Data grubbing" and "fishing" are terms of opprobrium applied to the work of others. The data miner's research strategy, while usually not defined in the textbooks, is clearly revealed by considering some typical quotations culled from leading professional journals:

> "Because of space limitations, only the best of a variety of alternative models can be presented here."
>
> "The precise variables included in the regression were determined on the basis of extensive experimentation (on the same body of data). . . ."
>
> "The method of step-wise regression provides an economical way of choosing from a large set of variables . . . those which are most statistically significant. . . ."
>
> "Since there are no firmly validated theories of the process . . . we consciously avoided a priori specification of the functions we wished to fit. . . ."
>
> "We let the data specify the model. . . ."

Of course, many authors are not as candid as those quoted above, and the extent of "experimentation" may not become apparent until the investigator is asked why one model was employed rather than an equally plausible alternative.

The efficiency with which data miners go about their work has increased considerably as a result of technological advance. The desk calculator has given way to successive generations of electronic computers. The development of data banks, such as that introduced by Dr. Boschan (1972) at the National Bureau of Economic Research, has increased tremendously the efficiency with which the investigator marshals evidence. The art of fishing over alternative models has been partially automated with stepwise regression programs. While such advances have made it easier to find high $\bar{R}^2$s and "significant" t-coefficients, it is by no means obvious that reductions in the costs of data mining have been matched by a proportional increase in our knowledge of how the economy actually works.

The majority of econometric textbooks discuss estimation and hypothesis testing procedures that are valid only when a priori considerations rather than exploratory data mining determine the set of explanatory variables to be included in the regression.[1] When a data miner uncovers t-statistics that appear significant at the 0.05 level by running a large number of alternative regressions on the same body of data, the probability of a Type I error of rejecting the null hypothesis when it is true is much greater than the claimed 5%. While those textbooks that do discuss the art of data mining usually caution their readers that judgment is required in interpreting the resulting t-statistics, practitioners find it difficult to evaluate what constitutes due caution in the absence of information on the degree to which exaggerated t-statistics are likely to be generated by intensive search over alternative candidate ex-

Received for publication August 7, 1979. Revision accepted for publication May 4, 1982.

* Wesleyan University.

While I am indebted to numerous colleagues and friends for helpful comments on earlier drafts of this paper, I retain full responsibility for remaining errors and heresies. Computer programming was executed by Charles Miller.

[1] An exception is Judge et al. (1980), who devote a chapter to a review of the literature on the problem of selecting the set of regressors. The present paper is more concerned with investigating the implications of data mining as it is often practiced than with an evaluation of the state of the art as exemplified by such recent contributions to econometric theory as those of Leamer (1978), Amemiya (1980) and Fisher-McAleer (1981).

planatory variables. The next section of this paper contributes a Rule of Thumb providing rough guidance in deflating the exaggerated claims of significance generated by data mining activity. Simulations presented later in the paper appraise the degree of success with which a practitioner of the data miner's art is likely to uncover those explanatory variables that actually contribute to the generation of the dependent variable.

II. "Significant" Results with Random Data

Insight into the effects of search is provided by considering a simple form of data mining. Suppose that an investigator has a number of equally plausible candidate explanatory variables, a priori information not prescribing with certainty which variables belong in the regression. Because of a belief in the efficacy of "simple models" or an adherence to Occum's Razor the investigator presumes as part of the maintained hypothesis that at most two of the candidate variables actually play a role in the determination of the dependent variable. Specifically, he presumes

$$Y_t = \beta_0 + \beta_1 X_{it} + \beta_2 X_{jt} + \epsilon_t; \qquad (1)$$

here X_{it} and X_{jt} are to be selected from the set of candidate explanatory variables and ϵ_t is a normally distributed random variable with finite variance σ_ϵ^2. If our investigator uses standard t-statistic procedures in testing at the 5% level the significance of a winning candidate explanatory variable, what will be the true significance level of the test? That is, what is the probability that the t-statistics in the final regression obtained by data mining will appear significant at the 5% level if the null hypothesis is true, there being no causal relationship so that in fact $Y_t = \beta_0 + \epsilon_t$?

Implications of searching for the best two candidate explanatory variables when the null hypothesis is true are reported in table 1. The analysis has been simplified in order to make the Berneulli model applicable by assuming that all the candidate explanatory variables are orthogonal (no collinearity) and that the variance of the ϵ_t is known to the investigator.[2] The first column of data on the table shows for comparison purposes the probabilities of obtaining zero, one, or two significant coefficients when there are only two candidate explanatory variables; this is the classical "textbook" case in which all candidate explanatory variables are included in the reported regression, there is no data mining, and the actual significance level (the probability of committing a Type I error of rejecting the null hypothesis when it is true) is the claimed 5%. The other columns of table 1 show how tests of significance are distorted when the search is conducted over an increasing number of candidate explanatory variables. For example, the second column shows that a researcher testing at a claimed 5% level after picking the best two out of five candidates reduces the probability of correctly reporting no significant coefficients when the null hypothesis is true to 77.4%, which is equivalent to a valid t-test at only the 12% significance level.[3] The third column shows that testing at a nominal 1% level after picking the best two out of five variables is equivalent to a valid test at only the 5% level. A less inhibited investigator considering ten candidates has only a 59.9% probability of accepting the null hypothesis at the 5% level when it is true, more energetic search causing the true significance of the t-tests to deteriorate to 23%.

Sample size does not influence the probability that a data miner will commit a Type I error, *given* the number of candidate explanatory variables and the nominal significance level.[4] While an enlarged sample size increases the probability that high t-coefficients will be uncovered if it encourages the investigator to search over a large number of candidate explanatory variables, the larger sample will decrease the probability of "significant" coefficients if it encourages the investigator to test at a more conservative significance level.[5]

---

[2] Precedence for the orthogonality assumption is provided by the Ando and Kaufman (1966) simulation study of the effect of search on the distribution of $\bar{R}^2$; they did not investigate the hypothesis testing implications of search; see also Christ (1966) and Bacon (1977). In contrast, Ames and Reiter (1961) used historical times series.

[3] Because of orthogonality, the probability of correctly accepting the null hypothesis when two regression coefficients are individually tested at the 12% level is $(1 - .12)^2 = .774$.

[4] Note, however, that the assumption that the candidate explanatory variables are orthogonal implies that the sample size is greater than the number of candidates.

[5] Simulations reported by Ando and Kaufman (1966) show that an increase in sample size is likely to lower the $\bar{R}^2$ that

Can the tendency for data mining to yield exaggerated claims of significance be offset by conservatively restating the claimed significance level, conducting the test with the standard textbook procedure at the 1% level, say, rather than 5%, in order to counter the effect of search? Examination of the fifth column of table 1 reveals that this adjustment is appropriate if the data miner has picked the best 2 variables out of 10 candidates; that is to say, under the null hypothesis the probabilities of obtaining 0, 1, or 2 nominally significant explanatory variables at the 1% level in searching for the best 2 out of $c = 10$ orthogonal candidates is the same as when validly testing at the 5% level when a priori information has appropriately narrowed the number of candidates to $c = 2$ in advance of looking at the data. While this suggests that in evaluating the significance of regression coefficients at the 5% level it may be appropriate to require a critical t-value of three rather than the customary cutoff value of two when the best 2 out of 10 candidates are reported in the final regression, this constitutes an inadequate degree of conservatism when a more intensive search has been conducted. For example, the eighth column of table 1 shows that a valid 5% test when the best 2 out of 20 candidate variables have been selected should be conducted at the 0.5% nominal significance level. Here is a more general *Rule of Thumb*:[6]

> When a search has been conducted for the best k out of c candidate explanatory variables, a regression coefficient that appears to be significant at the level $\hat{\alpha}$ should be regarded as significant at only level
>
> $$\alpha = 1 - (1 - \hat{\alpha})^{c/k}; \tag{2}$$
>
> or as a *short cut guide*, the significance level is approximately
>
> $$\alpha = \frac{c}{k}\hat{\alpha}. \tag{3}$$

To illustrate for the case in which $k = 2$ explanatory variables are to be included in the final regression:

Number of Candidates (c)	Claimed Significance Level ($\hat{\alpha}$)		True Level (α)
	Rule-of Thumb (equation (1))	Short-Cut (equation (2))	
5	.0203	.02	.05
10	.0102	.01	.05
20	.005116	.005	.05
100	.0010253	.001	.05

Examination of the examples as presented in appropriate columns in table 1 suggests that the short-cut guide (equation (3)) is reasonably reliable.

Both the Rule of Thumb and the exaggerated significance levels reported in table 1 were derived with a procedure which abstracts from two important complications that are likely to arise whenever an investigator searches over economic data. First of all, in practical applications the variance of ϵ is usually unknown and must be estimated from the data. This means that whether a particular variable appears to be significant or not depends in part on the other variables also included in the regression. Since the data mining process is likely to yield an underestimate of σ_ϵ^2, there is a consequent exaggeration of t's and a still further overstatement of significance levels. Second, the fact that candidate explanatory variables are likely to be highly intercorrelated means that the trials are not independent, reducing the gains from search.[7] These two complications, probably operating in opposite directions,[8] mean that my Rule of Thumb and table 1 provide only rough indications of the extent to which data mining activities are likely to cause exaggerated claims of significance. Once these two complications are admitted, the likeli-

should be expected in searching over a given number of explanatory variables that are orthogonally distributed with zero mean and unit variance; this tendency would be confounded in the presence of trend. Analytical results related to the Ando-Kaufman study are developed by Robert W. Bacon (1977).

[6] This rule of thumb is obtained by equating the probability of accepting the null hypothesis for all k explanatory variables in the absence of search, the tests being conducted at level α, with the probability of accepting the null hypothesis for all c potential candidates when the test is conducted at claimed significance level $\hat{\alpha}$; i.e., $\hat{\alpha}$ achieves $(1 - \alpha)^k = (1 - \hat{\alpha})^c$. The short-cut guide, suggested by Thomson Whitin, is obtained by using the first term in the binomial expansion of (2).

[7] For example, an investigator who experiments with M1 versus M2 or with real versus nominal GNP explanatory variables might argue that little is gained from search because of the high collinearity of the alternative candidate time series. Readers could be convinced if a comparison of the alternative regressions revealed that the basic conclusions of the analysis were robust with respect to such choices; they should be suspicious if only the final results of search are presented.

[8] Research by Dunn (1959) suggests that intercorrelations among candidate explanatory variables are likely to reduce the apparent gains from search.

TABLE 1.—PROBABILITY WITH RANDOM DATA OF FINDING "SIGNIFICANT" REGRESSION RESULTS WHEN SELECTING THE TWO BEST EXPLANATORY VARIABLES (known variance and orthogonal candidate explanatory variables)

	Number of Candidate Explanatory Variables													
	$c = 2$	$c = 5$		$c = 10$		$c = 20$			$c = 100$			$c = 500$		
Nominal Significance Level ($\hat{\alpha}$):	5%	5%	2%	1%	5%	1%	5%	1%	0.5%	5%	1%	0.1%	1%	0.02%
Number of "Significant" Coefficients Reported:														
None Significant	.903	.774	.904	.951	.599	.904	.358	.818	.905	.006	.366	.905	.007	.905
One Significant	.095	.204	.092	.048	.315	.091	.377	.165	.091	.031	.370	.091	.033	.091
Two Significant	.0025	.023	.0038	.001	.086	.004	.264	.017	.004	.962	.264	.004	.960	.004
True Significance Level (α):	5.0%	12.0%	4.9%	2.5%	22.6%	4.9%	40.1%	9.6%	4.9%	92.3%	39.5%	4.9%	91.9%	4.9%

Note: Each column reveals for the specified number of candidate explanatory variables (c) and claimed significance level ($\hat{\alpha}$) the probability of obtaining zero, one, or two "significant" regression coefficients. Because the Bernoulli model is applicable, the entries in the "none significant" row are simply $(1 - \hat{\alpha})^c$. The entries in the next row are $b(1, c, \hat{\alpha}) = c\hat{\alpha}(1 - \hat{\alpha})^{c-1}$, which is the probability that precisely one variable will be reported as significant at the $\hat{\alpha}$ level out of the c candidates. Two variables will be reported as significant with probability $\sum_{k=2}^{c} b(k, c, \hat{\alpha}) = 1 - (1 - \hat{\alpha})^c - b(1, c, \hat{\alpha})$. The true significance level reported at the bottom of each column is that value of α which yields the correct probability of concluding that none of the candidate explanatory variables are significant; i.e., $(1 - \alpha)^2 = (1 - \hat{\alpha})^c$.

hood that data mining will yield exaggerated claims of significance depends in part on the universe from which the candidate explanatory variables are drawn. Nonetheless, the simulations reported later in this paper based on actual economic time series as candidate explanatory variables suggest that the Rule of Thumb provides a rough and ready guide as to the exaggerated claims of significance that are likely to arise from data mining.

III. Simulated Data Mining

Is data mining activity likely to uncover those candidate explanatory variables actually generating the data under study? Because the answer to this question depends in part upon the universe from which the candidate explanatory variables are selected, actual economic time series are used for this purpose in the simulations reported in this paper. However, it would be inappropriate to use historical data as the dependent variable for two reasons: it is necessary to know the actual structure generating the data, and it is necessary to be able to replicate the experiment in order to estimate the probability that the data miner's strategy will be successful. For the dependent variable the simulations reported in this paper use pseudo-realistic consumption time series generated by explicitly specified stochastic processes.

The 20 candidate explanatory variables used in the simulations are listed in table 2; all but the trend variable were culled from the NBER data tape. These time series were not selected at random. Rather, I chose 20 variables from the more than 3,000 on the NBER tape that I thought a reasonable investigator might conceivably be inclined to consider. Because of an interest in observing whether the search procedures would be likely to pick out the correct variable, I deliberately included two sets of quite closely related time series, the *fiscal variables* X_3, X_4, and X_5, and *monetary variables* X_{10}, X_{11}, X_{12} and X_{13}. The time series are used to explain nine alternative artificial dependent variables generated with the nine models reported in table 3.

The nine different models used in generating dependent variables are listed in table 3. The first three generate artificial dependent variables ran-

TABLE 2.—CANDIDATE EXPLANATORY VARIABLES

Variable Code Number	Title
1	Index, 5 Coincident Indicators
2	Gross National Product Implicit Price Deflator
3	Government Purchases of Goods and Services
4	Federal Government Purchases of Goods and Services
5	Federal Government Receipts, NPA
6	Gross National Product, GNP
7	Potential Level of GNP in 1958$
8	Disposable Personal Income
9	Expected Investment Expenditure
10	Total Member Bank Reserves
11	Monetary Base (St. Louis FED Concept)
12	Money Supply, M1
13	Money Supply, M2
14	Dow-Jones Industrial Stock Prices
15	AAA Corporate Bond Yield (Moody's)
16	Civilian Labor Force (CLF) (16+, Excluding Armed Forces)
17	Unemployment Rate, Total CLF
18	Unfilled Orders, MFG Durable Goods Industries
19	New Orders, All Manufacturing Industries
20	Trend

Note: Observe that candidate variables numbers 3, 4, and 5 constitute a closely related set of fiscal variables while numbers 10, 11, 12, and 13 are a set of closely related monetary candidates. All time series are annual from the NBER data tape of August 1973.

domly rather than with a causal process:[9] the first is simply a normally distributed random variable while the second involves an autocorrelated error term and the third uses a second-order autoregressive process plus trend.[10] An ideal search procedure, if there were one, would lead the investigator to accept the null hypothesis that none of the candidate economic variables plays a causal role. An honest search procedure will lead an investigator to erroneously conclude that a significant relationship is present with a frequency equal to the claimed level of significance.

The remaining six models are causal. Number 4 is a "monetary model," for it uses variable $X_{12,t}$ (money) to generate the dependent variable; number 5 is a "fiscal model," for it uses $X_{3,t}$ (government spending).[11] It is hoped the search procedure will reveal the correct variable and indicate that only one is involved. Model 6 is eclectic, for both $X_{3,t}$ and $X_{12,t}$ participate jointly in the generation of the artificial dependent variable. Hopefully, the search procedure will reveal that this pair of explanatory variables participate in the determination of $Y_{6,t}$. Models 7, 8 and 9 are like the three preceding except that the disturbance is generated by a first-order autoregressive process.

One problem with simulation is that it is possible to test only a finite number of alternative structures. While the choice of models used in generating the artificial dependent variables employed in this study may be considered arbitrary, the parameters are not. The parameters used in generating the artificial dependent variables were obtained from regressions using another time series on the NBER data tape, personal consumption expenditure, as the dependent variable. Thus each model generates pseudo consumption time series. For example, the parameters in model 3 were obtained by regressing the log of consumption spending on lagged values of itself and trend; similarly the parameters of models 4 and 5 were obtained by regressing consumption spending on the money supply (M1) and government spending on goods and services. The variance of the random disturbances used in each simulation was set to that yielded by the regression. The parameters of model 6 were obtained by averaging models 4 and 5. And models 7, 8 and 9 have the same parameters as 4, 5 and 6 but utilize an autocorrelated rather than an independent random disturbance.

[9] In a sense these three models relate to the approach of Ames and Reiter (1961), who investigated the distribution of correlation coefficients obtained by pairing up 100 economic time series drawn at random from *Historical Statistics for the United States;* they found that given one randomly selected series it is possible, on the average, to find by random drawing in from two to six trials another series explaining at least 50% of the variance of the first. An essential difference between the approach of this paper and that of Ames and Reiter (1961) is that they looked at the correlations between pairs of historical time series; presumably, the null hypothesis is not always satisfied. In contrast, the present approach uses artificial dependent variables generated by the specified stochastic process.

[10] Granger and Newbold (1974) analyze the spurious regressions that are likely to arise from autocorrelated error terms. Model 3 provides a slight elaboration of a model used by Orcutt (1948) in his fundamental critique of Tinbergen's (1939) econometric model of the U.S. economy. Orcutt cautioned that the set of 52 economic time series used by Tinbergen might have been obtained by drawing from the population of series generated by the model $(Y_t - Y_{t-1}) = 0.3(Y_{t-1} - Y_{t-2}) + \epsilon$. However, Orcutt's process was explosive. A series generated by model 3 has finite variance around its expected value.

[11] My choice of these two alternative models of consumption behavior was prompted by the debate sparking the Friedman-Meisselman (1963) comparison of the relative effectiveness of the money stock and autonomous spending as determinants of aggregate consumption spending.

TABLE 3.—MODELS USED TO GENERATE ALTERNATIVE ARTIFICIAL "CONSUMPTION" DEPENDENT VARIABLES

Random Disturbances

$\epsilon_t: N(0,1)$
$\epsilon^*_t = .75\epsilon^*_{t-1} + \epsilon_t\sqrt{7}/4$

Dependent variables:

1: $Y_{1,t} = 130.0\epsilon_t$
2: $Y_{2,t} = 130.0\epsilon^*_t$
3: $\log Y_{3,t} = .131 + .865 \log Y_{3,t-1} + .121 \log Y_{3,t-2} + .0016t + .0178\epsilon_t$ $\quad \bar{R}^2 = .997$
$\quad (.898) \quad (.224) \quad (.266) \quad (.0097) \quad \bar{S}_e = .018$
4: $Y_{4,t} = -325.0 + 4.44X_{12} + 12.55\epsilon_t \quad \bar{R}^2 = .991$
$\quad (13.8) \quad (0.09) \quad \bar{S}_e = 12.54$
5: $Y_{5,t} = 75.3 + 2.4X_3 + 18.58\epsilon_t \quad \bar{R}^2 = .980$
$\quad (8.9) \quad (0.7) \quad \bar{S}_e = 18.58$
6: $Y_{6,t} = -125.0 + 2.22X_{12} + 1.2X_3 + 15.56\epsilon_t$
7: $Y_{7,t} = -325.0 + 4.44X_{12} + 12.55\epsilon^*_t$
8: $Y_{8,t} = 75.3 + 2.4X_3 + 18.58\epsilon^*_t$
9: $Y_{9,t} = -125 + 2.22X_{12} + 1.2X_3 + 15.56\epsilon^*_t$

Note: Reasonable parameters for models 3, 4, and 5 were obtained by regressing actual personal consumption expenditure on the indicated explanatory variables, using annual data, 1948 through 1970. The parameters of model 6 were obtained by averaging models 4 and 5. Models 7, 8, and 9 have the same parameters as 4, 5, and 6 but utilize an autocorrelated rather than an independent random disturbance.

As a practical matter there exists a rich repertoire of criteria that data miners employ in deciding on the subset of explanatory variables to be reported in their final regression. Three ad hoc strategies are considered in this simulation study:

1. stepwise regression
2. maximum $\bar{R}^2$
3. max-min $|t|$.

The maximum $\bar{R}^2$ criterion can be defended as maximum likelihood. However, finding the pair of variables maximizing $\bar{R}^2$ is computationally demanding; for example, with 20 candidate explanatory variables there are 190 possible regressions to search over. While stepwise regression programs are more economical in terms of computer time, they do not always yield the $\bar{R}^2$ maximizing pair of explanatory variables.[12] The stepwise algorithm used in the simulations reported here proceeds by first introducing the explanatory variable with the highest simple correlation with the dependent variable; then the algorithm selects as a second explanatory variable that time series which maximizes goodness of fit, given the presence of the first selection.[13] The max-min $|t|$ criterion mimics the behavior of an investigator who is willing to sacrifice goodness of fit for high t-coefficients. While all three criteria yield precisely the same results when the candidate explanatory variables are orthogonal, this does not necessarily happen with highly intercorrelated economic time series. Thus the third criterion may lead to a different set of explanatory variables when those selected by the $\bar{R}^2$ maximizer are highly collinear.

IV. Simulation Results

Quite respectable correlation coefficients are likely to be obtained when data mining over 20 candidate time series, judging by table 4, which reports the correlation coefficient obtained with each of the nine models used in generating the artificial dependent variables, averaged over 50 simulation runs.[14] Only for model 1, in which the dependent variable is an independently distributed random series, are the correlation coefficients of disappointing magnitude. The results for model 2 reveal, as expected, that substantially tighter fits are obtained when the random dependent variable is autocorrelated. The

[12] Write-ups of stepwise procedures customarily warn the reader that they will not always maximize $\bar{R}^2$; but they fail to caution the user that the t-coefficients printed out by the computer do *not* have the t-distribution. Recent versions of the Biomedical Computer Package contain an "all possible subset regression" option in addition to step-wise regression.

[13] While more sophisticated stepwise procedures, such as that provided in the SPSS statistical package (cf. Nie et al., 1971), will discard variables and replace them with alternatives when this will lead to a higher R^2, the simple stepwise procedure used in these simulations, borrowed from IBM's SPS Statistical Package, does not incorporate this refinement.

[14] Were 50 replications sufficient? Examination of intermediate output obtained with but 25 runs provided essentially the same results as reported in this paper, suggesting that if anything an excessive number of replications were executed.

tightest fits are obtained with model 3, which generates the dependent variable with a second-order autoregressive process with trend rather than with the aid of any of the 19 candidate economic time series. This effect, while disconcerting, was built into the simulation experiment by the parameter values used in generating the data; indeed, the simulation $\bar{R}^2$s are close to the observed $\bar{R}^2$s reported in table 3.[15]

The correlation coefficients obtained by data grubbing are rather insensitive to the criterion used in selecting the "best" regression. For model 1, in which the dependent variable is purely random, there does exist a substantial gap between the average $\bar{R}^2$ generated by the stepwise algorithm and that obtained by invoking the max $\bar{R}^2$ criterion; and max-min $|t|$ does almost as well as $\bar{R}^2$. The same tendency is to be observed, but to a lesser degree, with model 2. For all the other models the three criteria yield uniformly tight fits, although max-min $|t|$ does lead to a slight sacrifice relative to the stepwise algorithm.

The proportion of t-coefficients that appear significant at the 5%, 1% and 0.1% level with each of the models is reported in table 5. The simulations reveal that max-min $|t|$ is much more productive than the two alternatives at uncovering explanatory variables that appear significant in terms of the customary criteria. Because the substantially higher yield of sizable t-statistics is achieved with only a very minor reduction in $\bar{R}^2$, the evidence of tables 4 and 5 suggests that the max-min $|t|$ criterion is much more likely to yield impressive regression results. It is tempting to conjecture that max-min $|t|$ most closely approximates the behavior of a regression runner under pressure to generate publishable results. This does *not* mean that the max-min $|t|$ dominates in terms of being most informative about the underlying structure generating the data under study.

The anticipated tendency for data mining activity to generate exaggerated claims of significance is clearly revealed by table 5. Although the null hypothesis is valid for model 1, the dependent variable being random rather than generated by any of the 20 candidate explanatory variables, the claimed significance level grossly understates the probability of rejecting the null hypothesis. And the results for model 2, for which the null hypothesis is also true, show that a naive data miner who neglects problems of autocorrelation will almost always find significant results when the dependent variable is generated by a simple autoregressive process. Although the artificial dependent variables for models 4, 5, 7 and 8 are generated with only one rather than two of the candidate explanatory time series, table 5 shows that the max-min $|t|$ data mining strategy always yields two "significant" regression coefficients, which constitutes a Type I error of rejecting the null hypothesis when it is true. While both the stepwise and the max $\bar{R}^2$ data mining criteria are much more cautious, the true probability of a Type I error is substantially in excess of the claimed significance level.

The simple Rule of Thumb advanced earlier in this paper (equation (2)) is reasonably successful at correcting the exaggerated claims of significance reported in table 5. Specifically, the rule suggests that when the best $k = 2$ out of $c = 20$ orthogonal candidate explanatory variables are being selected, significance tests at nominal level $\hat{\alpha} = 5\%$ are more appropriately regarded as at the $\alpha = 40\%$ significance level, which means that the null hypothesis will be correctly accepted for both variables, if independent, 36% of the time. With $\hat{\alpha} = 1\%$ the same Rule of Thumb yields a corrected significance level of $\alpha = 9.6\%$ and an 87% probability of finding no coefficients significant; $\hat{\alpha} = 0.1\%$ yields $\alpha = 1\%$ and a 98% probability of no significant coefficients. These Rule of Thumb adjustments, relevant for model 1, are rather more

[15] Recall that the coefficients of the various models were estimated from U.S. consumption spending data, which apparently conform to Orcutt's suggestion about the stochastic process generating the data used in Tinbergen's (1939) econometric model.

TABLE 4.—AVERAGE VALUE OF $\bar{R}^2$
(50 runs; 2 best out of 20 candidate explanatory variables, 23 observations)

	Criterion		
	Stepwise	Max $\bar{R}^2$	Max-min $\|t\|$
Model 1	.103	.181	.180
Model 2	.374	.520	.507
Model 3	.997	.998	.993
Model 4	.991	.992	.973
Model 5	.981	.982	.959
Model 6	.986	.987	.973
Model 7	.996	.997	.978
Model 8	.991	.992	.983
Model 9	.993	.994	.979

Note: The models are defined in table 3; candidate explanatory variables are listed in table 2.

TABLE 5.—"SIGNIFICANT" t-STATISTICS—BEST PAIR OUT OF 20 CANDIDATE EXPLANATORY VARIABLES

Nominal Significance Level ($\hat{\alpha}$)	None "Significant"			Both "Significant"			Mean $\lvert t \rvert$	
	$\hat{\alpha}$ = 5%	$\hat{\alpha}$ = 1%	$\hat{\alpha}$ = 0.1%	$\hat{\alpha}$ = 5%	$\hat{\alpha}$ = 1%	$\hat{\alpha}$ = 0.1%	Variable 1	Variable 2
Model 1								
Stepwise	64%	88%	100%	20%	4%	0%	1.85	1.38
Maximum $\bar{R}^2$	36	72	96	58	24	2	2.45	2.26
Max-min $\lvert t \rvert$	36	72	96	60	24	2	2.45	2.30
Model 2								
Stepwise	16	38	74	52	18	10	3.44	2.23
Maximum $\bar{R}^2$	0	8	32	92	78	46	4.79	4.11
Max-min $\lvert t \rvert$	0	10	34	94	88	48	4.77	4.25
Model 3								
Stepwise	0	0	0	94	86	64	22.50	4.21
Maximum $\bar{R}^2$	0	0	0	100	100	98	32.63	7.90
Max-min $\lvert t \rvert$	0	0	0	100	100	100	18.23	10.55
Model 4								
Stepwise	2	8	12	38	8	0	13.26	1.87
Maximum $\bar{R}^2$	0	4	8	44	18	2	13.04	2.02
Max-min $\lvert t \rvert$	0	0	0	100	100	100	9.85	7.64
Model 5								
Stepwise	4	10	16	26	2	0	12.98	1.75
Maximum $\bar{R}^2$	2	4	6	48	34	20	12.37	2.78
Max-min $\lvert t \rvert$	0	0	0	100	100	100	8.57	6.70
Model 6								
Stepwise	0	12	36	46	12	2	7.60	2.16
Maximum $\bar{R}^2$	0	8	24	60	36	18	9.81	2.80
Max-min $\lvert t \rvert$	0	0	0	100	100	100	10.12	7.56
Model 7								
Stepwise	0	0	6	86	68	34	21.80	3.63
Maximum $\bar{R}^2$	0	0	2	88	78	44	23.48	3.83
Max-min $\lvert t \rvert$	0	0	0	100	100	100	11.16	8.53
Model 8								
Stepwise	0	0	4	88	62	16	15.98	3.11
Maximum $\bar{R}^2$	0	0	2	90	86	58	14.86	5.78
Max-min $\lvert t \rvert$	0	0	0	100	100	100	10.88	8.89
Model 9								
Stepwise	0	0	12	96	68	28	11.07	3.46
Maximum $\bar{R}^2$	0	0	2	96	84	68	14.75	4.66
Max-min $\lvert t \rvert$	0	0	0	100	100	100	12.97	8.92

Note: The entries in the first six columns show the percentage of times in 50 replications that the observed t-statistics appeared significant at the nominal level $\hat{\alpha}$. The last two columns show the average value of the largest and the smallest t-statistics.

accurate than we had any right to expect, suggesting that it may provide a useful guide even when the simplifying assumptions under which it is derived are violated.[16] Caution is advised, however, for the Rule does underpredict the proportion of times in which the selected pair of explanatory variables both appear significant.[17] And the Rule does not go nearly far enough in deflating the tendency for the max-min $|t|$ criterion to reject with excessive frequency the null hypothesis in models 4, 5, 8 and 9.

Is data mining likely to yield accurate information about the underlying structure generating the dependent variable when the null hypothesis is false? Table 6 shows how well data mining strategies do at uncovering true hypotheses. Each entry on this table reports the number of times in 50 replications the indicated explanatory variable was selected; numbers in italics indicate correct selections. To illustrate, when the dependent variable was generated by model 4 the stepwise regression procedure correctly identified variable 13 (M1) in all 50 simulation runs, a rather remarkable achievement. The max $\bar{R}^2$ procedure did almost as well, selecting M1 correctly in 46 out of 50 regressions of model 4. However, the max-min $|t|$ data mining strategy

[16] As explained in section II, the Rule of Thumb was derived under the assumption that the candidate explanatory variables are orthogonal and that the variance of the dependent variable is known to the investigator.

[17] Under independence the probability of zero significant coefficients at level α is $(1 - \alpha)^2$; the probability of two significant coefficients is α^2. A more sophisticated rule would take departures from orthogonality of the candidate explanatory variables explicitly into account.

never picked M1, the correct variable, although some solice may be derived from the thought that it did select a closely related monetary variable, member bank reserves, in 49 out of 50 trials. Stepwise regression also does spectacularly well with model 5, correctly selecting government spending on goods and services (variable 3) in 49 out of 50 replications. The maximum $\bar{R}^2$ criterion slips slightly, finding the correct explanatory variable in 31 out of 50 simulation runs. But the results obtained with the max-min $|t|$ search criterion are particularly disconcerting. This criterion, which places a premium on orthogonality, habitually picks up total member bank reserves and trend regardless of the actual structure. And this monetarist bias manifests itself even in models 5 and 8, where the artificial dependent variable is generated by a fiscal rather than a monetary process; that is, the max-min $|t|$ criterion is inclined to indicate that monetary rather than fiscal variables play a significant role in generating aggregate consumption spending even in a world in which it is government spending that counts.[18]

[18] How serious a matter this is depends in part on the loss function and in part on the cost of manipulating control variables. Suppose, for example, that we know that either model 4 or model 5 is correct and that there is no cost in manipulating either fiscal or monetary variables. Then misspecification is no problem, for we can use our estimate of the fiscal multiplier in order to adjust government spending to insure the desired level of consumption *if* that model is correct; simultaneously, we adjust the monetary base so as to get on target if in fact it is *M* that matters; either way we will be on target. The specification problem is serious when the eclectic possibility of model 7 is admitted and/or when the costs of manipulating fiscal and monetary control variables are considered.

TABLE 6.—EXPLANATORY VARIABLES SELECTED IN DATA MINING SIMULATIONS
(number of time each variable selected in 50 simulation runs)

Variable	1	2	*3*	*4*	*5*	6	7	8	9	*10*	*11*	*12*	*13*	14	15	16	17	18	19	20
Model 1																				
Stepwise	0	2	0	8	4	0	0	2	5	14	1	1	3	12	9	1	20	11	5	2
Maximum $\bar{R}^2$	6	1	5	2	8	6	3	7	8	4	7	7	9	3	5	5	4	0	8	2
Max-min $\|t\|$	6	0	6	2	8	7	3	10	8	3	7	7	8	3	6	5	3	1	6	1
Model 2																				
Stepwise	1	2	0	3	1	0	0	0	3	16	1	2	7	15	13	1	13	16	1	5
Maximum $\bar{R}^2$	4	9	2	2	4	6	8	2	4	3	7	14	8	1	8	4	0	5	3	6
Max-min $\|t\|$	5	10	3	3	4	10	8	1	3	4	6	13	8	0	6	4	2	4	2	4
Model 3																				
Stepwise	0	3	1	0	2	12	16	8	0	2	0	4	12	1	12	18	0	5	0	*4*
Maximum $\bar{R}^2$	0	2	0	0	0	10	24	2	0	0	0	0	18	5	14	5	1	8	0	*11*
Max-min $\|t\|$	1	2	5	5	0	1	0	0	0	3	1	0	35	7	4	2	1	1	0	*32*
Model 4																				
Stepwise	2	4	3	3	9	1	2	1	5	0	2	*50*	3	1	2	2	3	4	3	0
Maximum $\bar{R}^2$	3	6	3	3	8	0	2	1	5	0	4	*46*	5	1	2	2	3	3	3	0
Max-min $\|t\|$	0	0	0	0	0	0	0	0	0	49	0	*0*	0	0	1	0	0	0	1	49
Model 5																				
Stepwise	1	1	*49*	3	9	0	1	0	5	0	1	4	1	1	6	1	8	5	2	2
Maximum $\bar{R}^2$	1	1	*31*	19	9	1	1	1	4	1	4	4	3	1	4	1	7	3	2	2
Max-min $\|t\|$	1	0	*0*	9	0	0	0	0	0	39	0	0	3	16	4	3	0	2	0	23
Model 6																				
Stepwise	3	1	*28*	12	7	3	1	1	2	2	5	*28*	1	1	1	0	0	3	0	1
Maximum $\bar{R}^2$	4	2	*19*	17	4	1	6	1	1	2	6	*26*	1	1	2	1	0	5	0	1
Max-min $\|t\|$	0	0	*0*	1	0	0	0	0	0	48	0	*0*	0	2	2	0	0	0	1	*46*
Model 7																				
Stepwise	1	3	1	5	1	2	2	2	3	3	8	*47*	6	1	3	1	1	3	1	6
Maximum $\bar{R}^2$	1	5	1	4	1	2	3	2	3	3	7	*45*	5	1	4	2	1	4	0	6
Max-min $\|t\|$	0	1	1	1	0	0	1	0	0	40	1	*2*	1	3	6	0	0	0	6	37
Model 8																				
Stepwise	0	8	*49*	2	2	1	0	1	1	5	3	1	7	2	6	2	0	8	0	2
Maximum $\bar{R}^2$	0	7	*26*	23	2	2	3	2	1	2	4	5	6	2	4	5	0	5	0	1
Max-min $\|t\|$	0	2	*0*	26	0	0	1	3	1	22	3	3	5	2	5	5	0	2	0	20
Model 9																				
Stepwise	1	6	*26*	7	2	2	0	2	3	0	13	*25*	1	0	2	3	0	4	1	2
Maximum $\bar{R}^2$	1	3	*12*	17	1	4	5	1	1	0	14	*23*	1	0	5	6	0	5	0	1
Max-min $\|t\|$	0	0	*1*	6	0	0	1	1	0	43	0	*0*	0	3	4	0	0	1	0	40

Note: Entries in the body of the table reveal the number of times each candidate variable was selected in 50 simulation runs by the indicated selection criterion when the dependent variable was generated by the specified model. Candidate explanatory variables are identified in table 2; models are defined in table 3. Numbers in italics indicate correct selections. Note that variables 3, 4 and 5 constitute a closely related set of fiscal variables while 10, 11, 12 and 13 are closely related monetary variables.

It will be observed that at times the introduction of autocorrelated error terms into the fiscal model improves the situation; thus, in model 8, the max-min $|t|$ criterion selects $X_{4,t}$, a fiscal variable closely related to the true explanatory variable $X_{3,t}$, in 26 out of 50 simulations; while this is better than with model 5, it is a far from impressive performance.

V. Summary and Conclusions

This paper has examined the likely consequences of using standard regression procedures when the investigator's choice of explanatory variables is not inhibited by well-defined a priori considerations. The simulations reveal that of the three alternative selection criteria considered, max-min $|t|$ is likely to uncover explanatory variables yielding the most impressive regression results, a substantially higher yield of "significant" regression coefficients being obtained with only a modest sacrifice in goodness of fit. Appearances aside, is the data miner likely to uncover those candidate explanatory variables that actually contribute to the generation of the phenomenon under study? The top two rows of table 7 summarize the performance of the three alternative data mining strategies in terms of their success at picking out the correct pair of explanatory variables from twenty candidates.[19] The simple stepwise regression algorithm was remarkably successful, an impressive 70% of the selected variables that appeared significant actually participating in the generation of the dependent variable. The max-min $\bar{R}^2$ criterion was almost as successful. Unfortunately, the max-min $|t|$ criterion was a disaster, never succeeding in picking out the correct variable. And the third row of table 7 demonstrates that the max-min $|t|$ data mining strategy is particularly prone to commit Type I errors, rejecting the null hypothesis when true 81% of the time at a claimed $\hat{\alpha} = .05$ significance level. It is ironic that the data mining procedure that is most likely to produce regression results that appear impressive in terms of the customary criteria is also likely to be the most misleading in terms of what it asserts about the underlying process generating the data under study.

TABLE 7.—DATA MINING PERFORMANCE SUMMARIZED (models 1, 4, 5 and 6; $\hat{\alpha} = .05$)

	Stepwise	Max $\bar{R}^2$	Max-min $\lvert t \rvert$
Correct Variable Selected	70%	52%	0%
Correct or Related Variable Selected	82	75	36
Type I Error (true significance level)	30	53	81
Type II Error	15	8	0

Note: This table summarizes the performance of the three alternative selection criteria for the 4 models satisfying the classical regression assumptions. The first row shows the frequency with which the correct explanatory variables were selected at the $\hat{\alpha} = .05$ significance level. The second row shows the frequency with which a member of the correct monetary or fiscal policy set was selected. The third reports the incidence of Type I errors (rejecting the null hypothesis when it was true). The fourth row reports the incidence of Type II errors (accepting the null hypothesis when false).

While the analysis of this paper confirms that impressive $\bar{R}^2$ and substantial t-statistics are almost inevitably produced when standard regression procedures are used in searching over a modest number of candidates for the best explanatory variable, few data miners are willing to abandon the pretense of estimating structure and hypothesis testing by conceding that their t-coefficients are to be regarded as no more than sample descriptive. One alternative is to interpret the product of exploratory data mining activities conservatively. The Rule of Thumb advanced in part II of this paper provides guidance for deflating exaggerated claims of significance. For example, the Rule suggests that a data miner testing at a nominal $\hat{\alpha} = 5\%$ level after picking the two best explanatory variables out of twenty candidates should claim only 40% significance; that is the probability of committing a Type I error if the candidate explanatory variables are orthogonal. Although the candidate explanatory variables used in the simulations were highly collinear rather than orthogonal, the simulation evidence summarized on the third row of table 7 suggests that the Rule of Thumb offers a reasonable guideline if the data miner is attempting to maximize goodness of fit, either with a stepwise regression program or an exhaustive search for the highest $\bar{R}^2$; the Rule does not go far enough,

[19] In appraising this mixed record, it should be kept in mind that only 23 annual observations were provided to the simulated data miner confronted with the difficult task of selecting the right explanatory variables from 20 highly collinear candidate series. While it is reasonable to presume that longer time series would yield more accurate results, the consumption models generating the artificial dependent variables used in the simulations were very tight; with looser fits data mining performance deteriorated substantially. Additional tables reporting the results of doubling the variance of the disturbances of the models specified in table 3 are available from the author on request.

providing only an upper bound, in deflating the exaggerated claims of significance generated by the max-min $|t|$ data mining criterion.

Applied researchers are usually quite modest in describing how industrious a search was undertaken in generating reported results; and the criterion by which variables have been selected is usually left unspecified. In order to facilitate an informed interpretation of their results, empirical investigators should be expected to reveal the range of candidate variables and models considered and to report the degree of sensitivity of their primary conclusions to the choice of nuisance variables included for control purposes in their regressions.[20] It is the duty of editors and their referees to interrogate authors in order to insure that articles report, in footnotes of manageable size, details about the extent of search activity that has been undertaken.

This paper has not attempted to prescribe how an investigator should proceed in empirical work in the absence of a tightly structured theory. One ad hoc remedy, more frequently recommended than applied, is to reserve a portion of the data for post-sample prediction.[21] In addition to post-sample verification, the replication of empirical studies on new bodies of data should be encouraged, both by research foundations and journal editors.[22]

While this paper has focused on applied econometrics as often practiced, it is important to note that substantial progress in the state of the art has been made in recent years, including the Bayesian strategies reviewed by Gaver and Geisel (1974), the Pesaran and Deaton (1978) procedures for testing among non-nested non-linear hypotheses, and the work of Leamer (1978).[23] Unfortunately, inspection of the *Social Science Citation Index* indicates that applied researchers are slow to adopt improved procedures; in any event, such techniques are undoubtedly susceptible to misinterpretation when utilized by researchers who do not accurately report the extent of search activity they have engaged in.

[20] Cooley and LeRoy (1981), building on the contributions of Leamer (1978), have demonstrated that much can be learned from the ex post review of the data mining activities undertaken by empirical investigators. Although Cooley and LeRoy focused on studies of the demand for money, other areas of empirical research would doubtless be clarified by similar retrospective reviews.

[21] Thomas Mayer's review (1975) of a number of published studies reporting post-sample predictions reveals that the model yielding the tightest fit in the sample period usually does not perform best in the post-sample period. Theil (1971) suggests that when sufficient observations are available a researcher might partition the data into three parts, the first for choosing the specification, the second for parameter estimation, and the third for verification. With a sample of limited size it is not obvious how many observations should be reserved for either specification selection or post-sample verification.

[22] Prompted in part by the suggestion of Edgar Feige (1975), the editors of the *Journal of Political Economy* announced several years ago that their standard submission fee would be waived and publication expedited for notes submitted to a newly established Confirmation and Refutation section; in spite of this encouragement, very few articles have as yet appeared in this section.

[23] Much of the literature is concisely reviewed in chapter 11 of George Judge et al. (1980).

REFERENCES

Amemiya, Takeshi, "Selection of Regressors," *International Economic Review* 21 (June 1980), 331–354.

Ames, Edward, and Stanley Reiter, "Distributions of Correlation Coefficients in Economic Time Series," *Journal of the American Statistical Association* 56 (Sept. 1961), 637–656.

Ando, Albert, and G. M. Kaufman, "Evaluation of an Ad Hoc Procedure for Estimating Parameters of Some Linear Models," this REVIEW 48 (Aug. 1966), 334–340.

Bacon, Robert W., "Some Evidence on the Largest Squared Correlation Coefficient from Several Samples," *Econometrica* 45 (Nov. 1977), 1997–2001.

Boschan, Charlotte, "The NBER Time Series Data Bank," *Annals of Economic and Social Measurement* (Apr. 1972), 193–216.

Christ, Carl, *Econometric Models and Methods* (New York: John Wiley, 1966).

Cooley, T. F., and S. F. LeRoy, "Identification and Estimation of Money Demand," *American Economic Review* 71 (Dec. 1981), 825–844.

Dunn, Olive Jean, "Confidence Intervals for the Means of Dependent Normally Distributed Variables," *Journal of the American Statistical Association* 54 (Sept. 1959), 613–621.

Feige, Edgar L., "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy* 83 (Dec. 1975), 1291–1296.

Fisher, Gordon R., and Michael McAleer, "Alternative Procedures and Associated Tests of Significance for Non-nested Hypotheses," *Journal of Econometrics* 16 (May 1981), 103–119.

Friedman, Milton, and David Meiselman, "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897–1958," in E. C. Brown et al., *Stabilization Policies*, A Series of Research Studies Prepared for the Commission on Money and Credit (Englewood Cliffs: Prentice Hall, 1963), 165–268.

Gaver, Kenneth M., and Martin S. Geisel, "Discriminating among Alternative Models: Bayesian and Non-Bayesian Methods," in Paul Zarembka (ed.), *Economic Theory and Mathematical Economics* (New York: Academic Press, 1974), 49–80.

Granger, C. W. J., and P. Newbold, "Spurious Regressions

in Econometrics," *Journal of Econometrics* 2 (July 1974), 111–120.

Judge, George G., William E. Griffiths, R. Carter Hill, and Tsoung-Chao Lee, *The Theory and Practice of Econometrics* (New York: John Wiley & Sons, 1980).

Leamer, Edward E., *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (New York: John Wiley & Sons, 1978).

Mayer, Thomas, "Selecting Economic Hypotheses by Goodness of Fit," *The Economic Journal* 85 (Dec. 1975), 877–883.

Nie, Norman, Dale H. Bent, and C. Hadlai Hull, *Statistical Package for the Social Sciences* (New York: McGraw-Hill, 1970).

Orcutt, Guy H., "A Study of the Autoregressive Nature of the Time Series Used for Tinbergen's Model of the Economic System of the United States, 1919–1932," *Journal of the Royal Statistical Society* 10 (1948).

Pesaran, M. H., and A. S. Deaton, "Testing Non-nested Non-linear Regression Models," *Econometrica* 46 (May 1978), 677–694.

Theil, Henri, *Principles of Econometrics* (New York: John Wiley, 1971).

Tinbergen, Jan, *Statistical Testing of Business-Cycle Theories: Business Cycle in the United States of America, 1919–1932* (Geneva: League of Nations, 1939).