Data Mining as an Industry

Frank T. Denton

The Review of Economics and Statistics, Volume 67, Issue 1 (Feb., 1985), 124-127.

NOTES

DATA MINING AS AN INDUSTRY

Frank T. Denton*

Abstract—"Data mining" by an individual investigator can distort the probabilities in conventional significance tests. This paper argues that the same effect can occur when a given data set is used by more than one investigator, *even if no individual investigator engages in data mining*. A problem of publication selection bias is recalled and note is taken of its implications for the interpretation of published test results when there is collective data mining. Some illustrative calculations of probabilities associated with collective data mining are provided.

I. Introduction

A recent article by Lovell (1983) provides helpful insights into the implications of "data mining." Lovell assumes a lone investigator who estimates regression equations with alternative sets of explanatory variables and he demonstrates how misleading conventional significance tests can be when reported only for a single "best" equation. One might assume that all would be well if every investigator were to shun the practice of data mining and follow the "textbook dictum" that the choice of explanatory variables should be made a priori. However, I shall argue that such an assumption would be an example of the "fallacy of composition"—that the use of the same data set by more than one investigator distorts the probabilities associated with reported hypothesis tests, *even if no individual investigator engages in data mining*. Indeed, this problem may be even more serious because it is not as easily recognized. A single investigator can take Lovell's advice to report his data mining but if there are several investigators using the same data, each may be unaware of what the others are doing, and hence of the extent of *collective* data mining. That distortions can result from the collective analysis of a data set inspires the observation that econometric data mining may be viewed as an industry as well as an individual occupation.

II. Some Simplifying Assumptions

We postulate a population of investigators concerned with the estimation of linear regression equations to explain the behavior of some dependent variable y. Each investigator is either a Data Miner or a Classical Statistician. If he is a Data Miner and is given a set of data he will fit as many alternative equations as there are alternative subsets of potential explanatory variables and will choose the "best" equation. If he is a Classical Statistician he will choose a single set of explanatory variables a priori, and will fit only the equation that includes those variables.[1] In both cases the investigator will subject each equation to a conventional hypothesis test.

Assume a data set consisting of m potential explanatory variables x_i $(i = 1, 2, \ldots, m)$, with $T > m$ observations on each. Assume that all variables are mutually orthogonal and that investigators fit only equations with one explanatory variable. There are then m equations of the form $y = \beta_i x_i + \epsilon$ that can be estimated, where ϵ is NID$(0, \sigma^2)$. Assume that σ^2 is known while the β_i are not.[2] Corresponding to the i^{th} equation is the null hypothesis H_0^i: $\beta_i = 0$. If an investigator is a Data Miner he will test all m null hypotheses but report only the test with the "best" result; if he is a Classical Statistician he will test only one. Assume finally that all m of the null hypotheses are true.[3]

Received for publication October 27, 1983. Revision accepted for publication July 5, 1984.

*McMaster University.

*Some very helpful comments on earlier versions of this paper were provided by Michael C. Lovell, Donald N. McCloskey, and anonymous referees.

[1] That applied econometrics seldom conforms to the Classical Statistician model is well known by anyone familiar with the art. (See Leamer, 1978, 1983, for example.) Nevertheless, it is the model on which all standard econometric hypothesis testing theory is based.

[2] The strong assumptions of orthogonality, bivariate regression, and knowledge of σ^2 are chosen to make the analysis as simple as possible. Simulation results reported by Lovell (1983) suggest that the assumptions of orthogonality and knowledge of σ^2 may roughly cancel each other in their effects on significance test probabilities.

[3] An investigator who recognized that competing hypotheses should be treated as a set would not be "data mining," as here defined. We assume m simple non-nested null hypotheses. An appropriate procedure in practice would be to nest them artificially in a multivariate regression equation and test the composite hypothesis $\beta_1 = \beta_2 = \ldots = \beta_m = 0$ using a standard F test. Test procedures for non-nested hypotheses have received considerable attention recently (Davidson and MacKinnon, 1981, 1983; Fisher and McAleer, 1981; Godfrey, 1983; Pesaran and Deaton, 1978; Pesaran, 1982). If it were assumed that one of the β parameters must be nonzero, the Davidson-MacKinnon J test would provide a very simple procedure for testing any particular hypothesis against the set of alternatives. However, the true Data Miner ignores all such possibilities and continues to misapply the classical methods.

III. Alternative Testing Scenarios

It is convenient to consider three scenarios. The first, which may be termed "individual data mining," is of the type which underlies Lovell's analysis, and which we usually have in mind when we refer to "data mining." The second may be termed "coordinated collective data mining" and the third "uncoordinated collective data mining."

Individual Data Mining

In this scenario there is a single Data Miner who fits all m equations and tests each of the estimated coefficients β_i $(i = 1, 2, \ldots, m)$. If the investigator chooses a significance level α the probability that he will reject any single null hypothesis is, or course, α. However, the probability that he will reject at least one of the null hypotheses is $1 - (1 - \alpha)^m$.

Coordinated Collective Data Mining

Assume now that I am an inveterate Data Miner. I have the specified data set but instead of doing the testing myself I engage other investigators to do it for me. The others are all Classical Statisticians and abhor the idea of data mining. I must be careful, therefore, to speak to each one privately and to ask him to test only a single null hypothesis. I engage m investigators and ask each to report to me whether or not he rejects his null hypothesis. Each one believes (correctly) that there is a probability α that he will commit a Type I error. What he does not know is that $m - 1$ other investigators are also testing hypotheses with the same data set and that collectively the investigators are mining the set in the same way as the individual Data Miner. Only I know that, and only I realize that the probability of at least one of the investigators reporting to me a rejection of his null hypothesis is $1 - (1 - \alpha)^m$.

Uncoordinated Collective Data Mining

We do away now with the coordinator and assume n (not necessarily equal to m) investigators using the given data set, each working independently of the others, and each a confirmed Classical Statistician. Each investigator must choose a single null hypothesis. For simplicity, assume that the choice is random and that every null hypothesis is equally likely to be chosen; thus $p_{i,j} = 1/m$, where $p_{i,j}$ is the probability that the i^{th} null hypothesis will be chosen by the j^{th} investigator.[4] Some investigators may test the same null hypothesis and some null hypotheses may not be tested at all. The probability that the i^{th} null hypothesis will be tested and rejected by the j^{th} investigator at significance level α is $q_{i,j} = \alpha/m$.

Consider the probabilities for the full sets of null hypotheses and investigators. All null hypotheses are true but chance variation will place some in the *potential rejection subset* R; that is to say, some will be rejected, if in fact they are tested. The others are in the *potential acceptance subset* A. Let the null hypotheses be ordered from 1 to m. The probability that the first k of them will be in A and the last $m - k$ in R is $(1 - \alpha)^k \alpha^{m-k}$, the probability that every investigator will choose a null hypothesis from A is $(k/m)^n$, and the probability that no investigator will reject a null hypothesis is $(k/m)^n (1 - \alpha)^k \alpha^{m-k}$. If all subsets of given sizes k and $m - k$ are considered, the probability of no rejection is $(k/m)^n \binom{m}{k} (1 - \alpha)^k \alpha^{m-k}$. Finally, if k is allowed to take on all possible values, the probability that *at least one null hypothesis* will be rejected by *at least one investigator* is given by

$$Q(m, n, \alpha) = 1 - \sum_{k=0}^{m} (k/m)^n \binom{m}{k} (1 - \alpha)^k \alpha^{m-k}. \tag{1}$$

As n tends to infinity, Q approaches $1 - (1 - \alpha)^m$, which is the probability of at least one rejection in the individual data mining scenario; uncoordinated data mining is thus equivalent to individual data mining when the number of investigators is large relative to the size of the data set. As m tends to infinity, Q approaches $1 - (1 - \alpha)^n$, which is the probability of at least one rejection of a null hypothesis when all investigators are working with disjoint subsets. If *both* m and n tend to infinity, Q approaches unity.

The assumption of strict independence among researchers yields a polar case. It can be argued that in practice researchers communicate with each other and know about each other's results. However, this would merely tend to accelerate the mining of a given data set by eliminating redundancy: researchers would restrict their choices of null hypotheses to ones that they believed had not yet been tested with the given set and the result would approach that of the coordinated collective data mining scenario. To the extent that researchers do communicate with each other, the key question is whether they allow for this in calculating their significance test probabilities.

IV. The Publication Filter

The effect of publication criteria on the dissemination of information about the results of significance tests was considered by Sterling (1959) and Tullock (1959). Let

[4] In practice some hypotheses may be so incredible as to be assigned zero probabilites. However, we may assume for present purposes that the m variables in the given data set represent only admissible null hypotheses, the criterion of admissibility being theoretical plausibility. The randomness of choice then applies only within a theoretically admissible set.

TABLE 1.—PROBABILITIES OF REJECTING ONE OR MORE TRUE NULL HYPOTHESES WITH (UNCOORDINATED) COLLECTIVE DATA MINING

	$n = 1$	$n = 2$	$n = 5$	$n = 10$	$n = 25$	$n = 100$	$n = \infty$
$\alpha = .10$							
$m = 1$	.100	.100	.100	.100	.100	.100	.100
$m = 2$	.100	.145	.184	.190	.190	.190	.190
$m = 5$	.100	.172	.296	.374	.408	.410	.410
$m = 10$	.100	.181	.349	.495	.623	.651	.651
$m = 25$	.100	.186	.384	.584	.812	.925	.928
$m = 100$	.100	.189	.403	.634	.903	.999	1.000
$m = \infty$	.100	.190	.410	.651	.928	1.000	1.000
$\alpha = .05$							
$m = 1$	.050	.050	.050	.050	.050	.050	.050
$m = 2$	.050	.074	.095	.097	.097	.097	.098
$m = 5$	.050	.088	.158	.204	.225	.226	.226
$m = 10$	.050	.093	.189	.283	.378	.401	.401
$m = 25$	.050	.096	.210	.348	.558	.716	.723
$m = 100$	.050	.097	.222	.387	.679	.961	.994
$m = \infty$	.050	.098	.226	.401	.723	.994	1.000
$\alpha = .01$							
$m = 1$	.010	.010	.010	.010	.010	.010	.010
$m = 2$	.010	.015	.019	.020	.020	.020	.020
$m = 5$	.010	.018	.033	.044	.049	.049	.049
$m = 10$	.010	.019	.040	.063	.089	.096	.096
$m = 25$	.010	.020	.045	.081	.148	.219	.222
$m = 100$	.010	.020	.048	.092	.200	.471	.634
$m = \infty$	.010	.020	.049	.096	.222	.634	1.000

Note: The figures in the table are the probabilites that n Classical Statisticians working independently of each other will reject at least one out of m independent null hypotheses at significance level α when in fact all of the null hypotheses are true.

there be n investigators, each testing a different null hypothesis with independent data. If all of the null hypotheses are true the expected number of rejections is $n\alpha$. Publication is costly, and while there are certainly problems for which either statistical significance or non-significance would be of general interest, there are many for which the results of tests are much more likely to get published if they indicate significance. Hence there is a danger that readers of journals will be more likely to see results that support rejection rather than acceptance, whether or not the null hypotheses are true.[5] This process, which intervenes between investigators and readers, may be termed the "publication filter."

In the case of coordinated collective data mining, the coordinator was fully informed and able to allow for the effects of data mining on the probabilities when interpreting the tests. However, in the case of uncoordinated collective data mining there is no such fully informed person. A reader of a journal article may see the results of testing some null hypothesis but typically does not know how many others may have been tested without significant results. Even if the author convinces the reader that he himself has not engaged in data mining the reader is unable to know how much collective mining there may have been, and what therefore are the correct probabilities to apply in interpreting the published results.

V. Illustrative Calculations

Table 1 presents the probabilities that (uncoordinated) collective data mining will reject one or more true null hypotheses for alternative values of m, n, and α, based on equation (1). The table provides support for arguing that very high levels of significance should be used. For example, with 10 hypotheses and 10 investigators, and α at 0.10, there is a probability of about 0.50 that one "significant" result will be found; with α at 0.05, the probability is about 0.28; and with α at 0.01 it is about 0.06. The 0.06 is still six times the assumed level of 0.01 but it may be considered small enough to provide some confidence in the result. With $m = 10$ and n approaching infinity, the probability of rejection is about 0.65 for $\alpha = 0.10$, 0.40 for $\alpha = 0.05$, and 0.10 for

[5] It is not necessary that other papers be rejected explicitly; they may simply never be written if researchers believe that "negative" results are of little general interest and are unlikely to be accepted for publication. In laboratory contexts in which experiments can be repeated the publication of a chance rejection may elicit other papers which contradict and eventually nullify the initial (false) published results. Developments in the provision of economic data, especially microdata, have made it possible in some cases to attack a given problem with different sets of observations. However, the possibilities for doing this in applied econometrics are still severely limited.

$\alpha = 0.01$. Thus even with exhaustive collective data mining a "nominal" significance level of 0.01 can be interpreted as an effective level of 0.10, which is still within the "conventional" range for significance testing.

VI. The Number of Investigators as a Function of Time

We have assumed a given number of (Classical Statistician) investigators working with a given data set but n should probably be interpreted as the number who have *ever* tested a null hypothesis with the set, and hence as a (nondecreasing) function of time. If an investigator were restricted to testing only one null hypothesis with any data set he could hardly be expected to devote his life to that one test; he would more likely move on to another data set, while someone else would discover and start to work with the first one. This argues for assuming n to be large, as a safeguard, and again for choosing a very high (nominal) significance level. As a practical measure, it may be desirable to assume n infinite for any assumed value of m and to calculate the probability of rejection of a null hypothesis as $1 - (1 - \alpha)^m$. But this is equivalent to Lovell's "rule of thumb" for a single data miner.[6] The argument then is that the "rule of thumb" should be applied in interpreting the published results of a conventional significance test, *even if the author of the results has not himself engaged in data mining.*

VII. Concluding Observations

It is well recognized that data mining by individual researchers invalidates conventional significance tests. One might assume that if every researcher behaved according to the rules of "classical" or "textbook" hypothesis testing theory the validity of the tests would be restored. The purpose of this paper has been to argue that such an assumption would be an example of the "fallacy of composition." Data mining is an implicit characteristic of the collective process of analysis by a population of researchers, regardless of whether it is practiced explicitly by the individual members.

It has *not* been the purpose of this paper to argue that the investigation of alternative hypotheses with a given data set is "sinful." Attempting to describe a situation in which every researcher behaves according to the classical rules produces a scenario which (a) is unrealistic (if not downright silly) and (b) merely substitutes collective for individual data mining, so that the end result is the same, even if longer in coming. What should be avoided rather is the misleading practice of interpreting classical test results as if the theoretical assumptions on which they are based were satisfied. One can recommend the use of much higher nominal significance levels than the conventional ones.[7] Also, I have suggested that Lovell's "rule of thumb" might guide the adjustment of probabilities to allow for (invisible) collective data mining as well as (visible) individual data mining. If nothing else, an awareness of the problem should lead to healthy skepticism and higher standards in the interpretation of reported test results.

REFERENCES

Davidson, Russell, and James G. MacKinnon, "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica* 49 (May 1981), 781–793.

_____, "Testing the Specification of Multivariate Models in the Presence of Alternative Hypotheses," *Journal of Econometrics* 23 (Dec. 1983), 301–313.

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.

Godfrey, L. G., "Testing Non-Nested Models After Estimation by Instrumental Variables or Least Squares," *Econometrica* 51 (Mar. 1983), 355–365.

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

_____, "Let's Take the Con out of Econometrics," *American Economic Review* 73 (Mar. 1983), 31–43.

Lovell, Michael C., "Data Mining," this REVIEW 65 (Feb. 1983), 1–12.

Pesaran, M. H., "On the Comprehensive Method of Testing Non-Nested Regression Models," *Journal of Econometrics* 18 (Feb. 1982), 263–274.

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

Sterling, Theodore D., "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa," *Journal of the American Statistical Association* 54 (Mar. 1959), 30–34.

Tullock, Gordon, "Publication Decisions and Tests of Significance: A Comment," *Journal of the American Statistical Association* 54 (Sept. 1959), 593.

[6] In notation consistent with present usage, Lovell's rule is to calculate the probability of at least one rejection of a null hypothesis as $\gamma = 1 - (1 - \alpha)^{m/k}$, where k is the number of candidate variables in the final (best) equation, and γ the true probability of committing at least one Type I error with exhaustive search over all subsets of candidate variables of size k.

[7] In the same spirit, one can recommend to journal editors and referees a policy of requiring authors to demonstrate that their estimates and significance tests are not highly sensitive to minor changes in model specification. Of particular concern is whether results are robust with regard to the inclusion or exclusion of theoretically inconsequential variables intended to represent minor "nuisance" effects in a regression model.