NBER TECHNICAL WORKING PAPER SERIES

RESEARCHER INCENTIVES AND EMPIRICAL METHODS

Edward L. Glaeser

Technical Working Paper 329
http://www.nber.org/papers/t0329

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2006

This essay was written for the NYU Conference on Economic Methods. I am grateful to the Taubman Center for State and Local Government for financial support. My thoughts on this topic are the result of conversations over years with many people, but I am particularly indebted to Robert Barro and Lawrence Katz. Guido Imbens, Lawrence Katz and especially Edward Leamer provided helpful advice. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Researcher Incentives and Empirical Methods
Edward L. Glaeser
NBER Technical Working Paper No. 329
October 2006
JEL No. A11,B4

ABSTRACT

Economists are quick to assume opportunistic behavior in almost every walk of life other than our own. Our empirical methods are based on assumptions of human behavior that would not pass muster in any of our models. The solution to this problem is not to expect a mass renunciation of data mining, selective data cleaning or opportunistic methodology selection, but rather to follow Leamer's lead in designing and using techniques that anticipate the behavior of optimizing researchers. In this essay, I make ten points about a more economic approach to empirical methods and suggest paths for methodological progress.

Edward L. Glaeser
Department of Economics
315A Littauer Center
Harvard University
Cambridge, MA 02138
and NBER
eglaeser@harvard.edu

I. Introduction

A central theme of economics is respect for the power of incentives. Economists explain behavior by pointing towards the financial gains that follow from that behavior, and we predict that a behavior will increase when the returns to that behavior rise. Our discipline has long emphasized the remarkable things – from great inventions to petty crimes – that human beings can do when they face the right incentives.

But while economists assiduously apply incentive theory to the outside world, we use research methods that rely on the assumption that social scientists are saintly automatons. No economist would ever write down a model using the assumptions about individual selflessness that lie behind our statistical methods. A small but strong literature that dates back to Sterling (1959) and Tullock (1959) and that is most strongly associated with Edward Leamer, counters this tendency and examines the impact of researcher initiative on statistical results. But while many of the key points of this literature (especially those associated with data mining and specification searches as in Leamer, 1974 and Lovell, 1983) are both understood and accepted, the only obvious impact of this literature is a healthy skepticism about marginally significant results. Certainly, there is no sense in which standard techniques have been adjusted to respond to researcher incentives.

Indeed, the modest number of econometric papers in this area is probably best understood as the result of the very modest interest of empirical researchers in using techniques that correct for data mining and other forms of researcher initiative. After all, the impact of such techniques is invariably to reduce the significance levels of results. The same incentives that induce researchers to data mine will induce them to avoid techniques that appropriately correct for that data mining.

The importance of these issues increases during periods of rapid change in empirical techniques. Even if economists are not using formal Bayesian methods for dealing with data mining in standard regressions, experience-based rules of thumb proliferate.

Everyone knows to be skeptical about a new growth regression purporting to present a new variable that predicts growth (see Sala-I-Martin, 1997). However, new methods, such as the use of lab experiments or genetic information or instrumental variables techniques, give researchers new opportunities to use their initiative to enhance their results. During these periods of rapid change, it becomes particularly important to think about how to adjust statistical inference for researcher initiative.

In this essay, I make ten points about researcher incentives and statistical work. The first and central point is that we should accept researcher initiative as being the norm, and not the exception. It is wildly unrealistic to treat activity like data mining as being rare malfeasance; it is much more reasonable to assume that researchers will optimize and try to find high correlations. This requires not just a blanket downward adjustment of statistical significance estimates but more targeted statistical techniques that appropriately adjust across data sets and methodologies for the ability of researchers to impact results. Point estimates as well as t-statistics need to be appropriately corrected.

The second point is that the optimal amount of data mining is not zero, and that even if we could produce classical statisticians, we probably would not want to. Just as the incentives facing businessmen produce social value added, the data mining of researchers produces knowledge. The key is to adjust our statistical techniques to realistically react to researcher initiative, not to try and ban this initiative altogether.

The third point is that research occurs in a market where competition and replication matters greatly. Replication has the ability to significantly reduce some of the more extreme forms of researcher initiative (e.g. misrepresenting coefficients in tables), but much less ability to adjust for other activity, like data mining. Moreover, the ability to have competition and replication to correct for researcher initiative differs from setting to setting. For example, data mining on a particular micro data set will be checked by researchers reproducing regressions on independent micro data sets. There is much less ability for replication to correct data mining in macro data sets, especially those that include from the start all of the available data points.

Fourth, changes in technology generally decrease the costs of running tests and increase the availability of potential explanatory variables. As a result, the ability of researchers to influence results must be increasing over time, and economists should respond for regular increases in skepticism. At the same time however, improvements in technology also reduce the cost of competitors checking findings, so the impact of technology on overall bias is unclear.

Fifth, increasing methodology complexity will generally give the researcher more degrees of freedom and therefore increase the scope for researcher activity. Methodological complexity also increases the costs to competitors who would like to reproduce results. This suggests that the skepticism that is often applied to new, more complex technologies may be appropriate.

My sixth point is the data collection and cleaning offers particularly easy opportunities for improving statistical significance. One approach to this problem is to separate the tasks of data collection and analysis more completely. However, this has the detrimental effect of reducing the incentives for data collection which may outweigh the benefits of specialization. At the least, we should be more skeptical of results produced by analysts who have created and cleaned their own data.

A seventh point is that experimental methods both restrict and enlarge the opportunities for researcher action and consequent researcher initiative bias. Experiments have the great virtue of forcing experimenters to specify hypotheses before running tests. However, they also give researchers tremendous influence over experimental design, and this influence increases the ability of researchers to impact results. .

An eighth point is that the recent emphasis on causal inferences seems to have led to the adoption of instrumental variables estimators which can particularly augment researcher flexibility and increase researcher initiative bias. Since the universe of potential instruments is enormous, the opportunity to select instruments creates great possibilities for data mining. This problem is compounded when there are weak instruments, since the distribution of weak instrument t-statistics can have very fat tails. The ability to influence significance by choosing the estimator with the best fit increases as the weight in the extremes of the distribution of estimators increases.

A ninth point is that researcher initiative complements other statistical errors in creating significance. This both means that spurious significance rises spectacularly when there are even modest overestimates in statistical significance that are combined with researcher initiative. This complements also creates particularly strong incentives to fail to use more stringent statistical techniques.

My tenth and final point is that model driven empirical work has an ambiguous impact on researcher initiative bias. One of the greatest values of specialization in theory and empirics is that empiricists end up being constrained to test theories proposed by others. This is obviously most valuable when theorists produce sharp predictions about empirical relationships. On the other hand, if empirical researchers become wedded to a particular theory, they will have an incentive to push their results to support that theory.

A Simple Framework

Consider a researcher who is trying to produce famous results that appear to change the way that see the world. The researchers objective function is $\theta \bullet S(f(E)+R)-E$ where θ is the exogenous importance of the question, S(.) reflects the extent that the public attributes importance to the result, R is the real significance of the research which is a random variable not known by the public characterized by distribution g(R), E is the non-negative effort of the researcher in enhancing its significance and f(.) is a monotonically increasing function mapping effort into more impressive results.

The function S(.) is meant to reflect Bayesian inference on behalf of the pubic about the real value of R. For example if f(E)=1, for all E, then S(f(E)R)=f(E)R=R, and there would be no incentive to put in any effort. More generally, the researcher's first order condition is $\theta \bullet S'(f(E)+R))f'(E)=1$ which determines an optimal choice of effort E*. This problem is solved if S(f(E)+R)=f(E)+R-f(E*). In this case, the researcher sets $\theta f'(E^*)=1$. Here there is no misinformation, the public corrects perfectly for the researcher effort. The social loss is just the wasted researcher effort. This model predicts that there will be more researcher effort on more important problems and that the public will correctly be more skeptical about new results in important areas. I believe that this situation roughly characterizes situations with established researchers and established methodologies.

I now assume instead that the researcher has access to a new research technology, that is less perfectly understood by the public. In this case, $f(E)=\alpha\phi(E)$ where $\phi(E)$ is a known function and α is productivity parameter. If α is known, then in equilibrium $S(\alpha\phi(E)+R))=\alpha\phi(E)+R-\alpha\phi(E^*)$ and E* is defined by $\theta\alpha\phi'(E^*)=1$. Again, there is no misinformation, but the amount of researcher effort and associated cynicism is increasing with the value of α. New research technologies that increase the ability of worker effort to produce significant results will lead to more effort, more results that overrepresent significance and, in equilibrium more discounting of worker results.

The more complicated and interesting case occurs when the parameter α is unknown to the general public and known to the researcher. I assume that there is a distribution $h(\alpha)$, that is known. If R is unknown to the researcher when the choice of E is made, then the equilibrium is characterized by the optimal effort equation $\theta\alpha\phi'(E^*)\int S'(\alpha\phi(E^*)+R))g(R)dR=1$ and the Bayesian inference equation $S(x)=\dfrac{\int(x-\alpha\phi(E^*(\alpha)))g(x-\alpha\phi(E^*(\alpha)))h(\alpha)d\alpha}{\int g(x-\alpha\phi(E^*(\alpha)))h(\alpha)d\alpha}$. In this case, the uncertainty about the new research technology loses information because the public doesn't know how much the new technology changes the ability of researchers to impact significance.

This small framework has made four points. First, the degree of skepticism about results should be a function of the returns for providing significant results. Second, new technologies that increase the ability of researchers to influence results should increase skepticism. Third, new technologies, until they are fully understood, will lead to confusion because it is hard to figure out whether high levels of significance reflect reality or an increased ability of researcher effort to skew findings. I now move to a series of points about the impact of researcher effort. The first point follows directly from the model.

Point # 1: Researchers will always respond to incentives and will be more skeptical than standard statistical techniques suggest.

Empirical researchers are not automata. They make decisions about what puzzles to study, what variables to investigate, what specifications to estimate, what data to use and how to collect that data. Every decision permits flexibility and can potentially increase extent to which results are exciting and seem significant. The decisions that are made greatly impact the way that we should interpret results and create what I will refer to as "researcher initiative bias."

The most well-studied example of researcher initiative bias is data-mining (Leamer, 1978, Lovell, 1983). In this case, consider a researcher presenting a univariate regression explaining an outcome, which might be workers' wages or national income growth. That researcher has a data set with k additional variables other than the dependent variable. The researcher is selecting an independent variable to maximize the correlation coefficient (r) or r-squared or the t-statistic of the independent variable (which equals $\frac{r}{\sqrt{1-r^2}}$), all of which are identical objective functions. I will assume that each one of these explanatory variables is independently and identically distributed. Finally, I will assume that the researcher is constrained to present a univariate relationship that will then be judged on its ability to explain the data.

If the true correlation between every one of the k independent variables and the dependent variable is zero, and if all of the independent variables are independent, then the probability that the researcher will able to produce a variable which is significant at the ninety-five percent level is $1-.95^{k}$. If the researcher has ten potential independent variables, then the probability that he can come up with a false positive is 40 percent. If the researcher has 100 independent variables then the probability with which he can come up with a false positive is more than 99 percent. Even 99 percent confidence levels yields a 64 percent of a false positive if the researcher has 100 variables over which he can search.

Lovell (1983) provides a rule of thumb "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 the level $\alpha = 1-(1-\hat{\alpha})^{c/k}$." This provides us with a convenient way of adjusting significance levels for data mining. Alternatively, the formula can be inverted to determine the appropriate t-statistic needed to guarantee a significance level recognizing that the researcher chooses the best c out of k instruments.

Using this approach, to get a significance level of α , one needs to insist on a t-statistic that delivers a conventional statistical significance level of $1-(1-\alpha)^{k/c}$. For example in the case where the researcher is choosing the best out of ten candidate explanatory variables, and we want results to be statistically significant at the 95 percent level, we should save our interest for coefficients that are significant at the 99.5 percent level which would require a t-statistic of 2.6. If the researcher had 100 potential explanatory variables to draw from, and we wanted a 95 percent confidence level, then we would insist on a t-statistic of around 3.3.

This approach assumes no other forms of flexibility and assumes that the independent variables are uncorrelated. High degrees of correlation among the potential explanatory variables will make the bias less severe. Other forms of flexibility, such as the ability to choose functional form or selectively clean data will increase the bias, and as I will discuss later, often these types of flexibility will be complements in the production of bias. More generally, there will be a great deal of heterogeneity across settings in the ability to engage in both data mining and other forms of researcher initiative. In the language of the framework, the parameters α and θ. Ideally, if these parameters are known this would lead to a straightforward adjustment of significance levels. The problem is in fact for more difficult, since the reader knows quite little about the actual ability of the researcher to influence significance.

Denton (1985) follows Sterling (1959) and makes it clear that exactly the same statistical problems occur with data mining and publication bias. For example, if ten different statisticians work independently testing only one variable in the classic way, and if the only work that is published is that which is statistically significant at the 95 percent level, then this is equivalent to having a single researcher choose the best of ten potential variables. This publication bias is best seen as the result of journal editors using their initiative to obtain their own objectives. Journal editors also face incentives to present exciting significant work and this causes them to publish only papers that have significant results and this will create exactly the same bias that occurs when the selection occurs at the researcher level.

The fact that both journal editors and authors have an incentive to collaborate in producing exciting results that seem statistically significant may perhaps explain why the econometric field of addressing research initiative bias has grown so slowly. There is no quick fix for this problem, and addressing it requires the vast majority of researchers to admit that they routinely violate the admittedly unrealistic norms suggested by classical statistics. Econometrics, like every other field, responds to demand and the demand for techniques to correct statistical significance downward has historically been limited.

One example of the relative modesty of this field is that while these papers have convinced at least some readers to hold out for t-statistics greater than two, there has been little attempt to develop the appropriate ways of correcting point estimates as well as significance levels. Yet the researcher initiative bias compromises both significance tests and point estimates. At the extreme, a regression coefficient should be seen as the biggest coefficient among a distribution of coefficients rather than an average coefficient among that distribution. If the researcher is just trying to produce significant results, this will bias coefficients away from zero. If the researcher has an ideological bias that pushes towards one type of result, this will create bias in a single direction.

For example, assume again that the true relationship between each of k candidate variables and the outcome variable is zero and the standard deviation of the coefficient estimates is σ, then expected value of the maximum coefficient in this group is approximately $\sigma\left[\sqrt{2\ln k}-\frac{\ln(\ln k)+\ln(4\pi)-1.544}{2\sqrt{2\ln k}}\right]$. The term in parentheses equals 1.72 when k is equal to ten. As such, the expected value of best regressor drawn from a sample of ten uncorrelated regressors is 1.72 standard deviations. If the regressors had an average coefficient of "b" then the expected largest regressor would be 1.72 standard deviations higher than b. This gives us a hint of how one might correct for data mining in our coefficient estimates as well as our significance tests. A more thorough approach would require complete specification of the reader's assumptions about what the researcher can do, and then use Bayesian inference.

This hint is only that and to address this issue squarely we need to develop better methods for dealing with researcher initiative bias. These methods are needed partially to be realistic: all of the most well-meaning attempts to get empiricists to act exactly like classical statisticians are doomed to failure because of the amount of discretion that is unavoidable. But these methods are also necessary, because as I will argue in the next section, it is not obvious that we would want empiricists to follow the classic rules of statistical inference even if we could get them to do so.

Point # 2: The optimal amount of data mining is not zero.

Classical statistics suggests that every hypothesis that is ever tested should be written up and published. After all, only in that case will we be sure that we are seeing an unbiased sample of the true set of estimates.   However, economic models have an uncountable number of potential parameters, non-linearities, temporal relationship and probability distributions. It would utterly impossible to test everything. Selective testing, and selective presentation of testing, is a natural response to vast number of possible tests. Moreover, in cases where economic questions have reached iconic status (the impact of education on wages for example), the utopian classical scenario may actually come close to being true. After all, when the question becomes sufficiently important there is an interest in both positive and negative results. Of course, even in this case, estimates that lack sufficient precision to inform the debate often die without publication, but my suspicion is that most sensible tests of fully established questions are generally published. In these cases, as well, it is presumably true that that the real correlation between the variables of interest isn't zero, which also acts to reduce the degree of publication bias.

The situation is far different when we are in the more primitive situation of trying to come up with new explanations for an outcome of interest. Consider the problem facing a social planner who has a set of researchers whose activities can be completely controlled. Assume further that the planer's only ambition is to inform the world about the true causes of variation in the outcome. I assume further that there is a single (incredibly large) data set with a fixed number of variables, and we are interested in the ability of these variables to explain a single dependent variable. Would this dictator instruct his researchers to act like classical statisticians or would he favor an approach that looks like data mining?

To fill in the model, assume that the planner's total allocation of researcher time equals "T" and that it takes "z" units of time for each researcher to run a test and understand its meaning and "q" units of time for the researcher to write up each results in a coherent manner and present it to the public. I am essentially assuming that each new test requires its own paper. I abstract from the costs of editor and referee time and the hard costs of publication. If the planner instructs the researcher to act like classical statisticians, they will write T/(q+z) papers informing the public of T/(q+z) correlations.

If each independent variable has an expected r-squared of ρ^2, and if all of the x variables are completely independent, then the expected amount of total explanatory power created by the T/(q+z) variables introduced to the public is $T\rho^2/(q+z)$. I am assuming that the number of tests involved is sufficiently small that this number is comfortably greater than one. In this case, the amount of knowledge that the public thinks that it has is the same as the amount of knowledge that it actually does have because the reported r-squared are unbiased estimates of the true r-squared.

Consider alternatively, the situation in which the planner instructs the researchers to run "m" tests and report the "n" most significant of those tests where mq+nz equals T. I assume that the decision about how many tests to run is determined before learning the results of any of these tests, so the value of n is also predetermined and equals (T-mq)/z. There will be a cutoff point, denoted $\hat{\rho}^2$ that is determined by the n most significant variables. The basic intuition behind the social benefit of data mining is that there is very little benefit in reporting a variable which literally has an r-squared of zero to a social planner trying to reveal information about the causes of the outcome. If a regression found that wearing blue shirts had no correlation with cancer, then would it really make sense to publish that result?

Assume somewhat implausibly a binary distribution that each one of the variables either had an r-squared of zero or $\overline{\rho}^2$ each with equal probability, and that a regression also yielded binary results so that variables appeared to have power or not and there was a probability of error in each variable of "e." We can then compare the strategy of reporting all of the variables or reporting only those variables that seem to be significant. If we report all of the results, the expected number of variables that the public will think matter equals $T/2(q+z)$. There will be some confusion over which variables are significant and which ones are not.

Conversely, if the researcher only reports the variables that appear significant the public will have learned of $T/(2q+z)$ variables that are significant. The public will not have learned about all the negatives, some of which are false. But if there are actions that can be taken in response to observed correlations then certainly the second strategy is better. In either case, the public must adjust its beliefs about the positive results downward to account for the possibility of a false positive, and in this case, the adjustment factor doesn't change between the two strategies. It is just that with data mining more information gets to the public.

Obviously, the advantages of data mining in this little example increase with the size of z. When writing up results is costless relative to running the experiments, then there is little advantage to be gained from selectively presenting results. However, when z is larger then the gains from data mining are higher. This may explain why data mining appears to be more prevalent in economics than in many of the laboratory sciences. Our presentational requirements are much harder and our experiments—at least those with observational data—are much cheaper. As a result, no social planner would insist that every insignificant result be presented to the public.

Of course, this is just one admittedly extreme example of how data mining can be good. In other cases, it can be an unqualified bad. For example, if the social planner is directing his researchers to all do independent tests on the same variable, then suppressing data is generally socially bad (unless the suppressed data is a random sample of the revealed data). Moreover, if the ability of people to correct the errors in observed estimates declined when data was selectively published, that would also represent a problem. If, for example, individuals didn't know the ex ante probability of getting significant or insignificant results, then providing the insignificant results would have social value.

Point # 3: Competition and replication can help, but its impact will differ significantly from setting to setting.

The role that competition among scientists will have on researcher initiative bias was discussed by Tullock (1959) who argued that competition would counteract the version of publication bias that occurs when 20 researchers each use different data sets to run the same experiment but when only the one significant result gets published. Tullock argued that in this case the other 19 researchers would come forward and discuss their insignificant results. The conclusion Tullock drew from this is that publication bias is more likely to occur in a situation where there is a single data and 20 possible explanatory variables. In that case, there is no obvious refutation that could be published over the false positive. The best that can be done is to publish articles emphasizing the number of potential explanatory variables in the data set (as in Sala-I-Martin, 1997) or the fragility of the results to alternative specifications (as in Levine and Renelt, 1992).

Tullock is surely right that there are cases where competition among peers is likely to either deter researcher initiative bias or at the least limit its impact ex post and there are other cases where this competition is unlikely to have any impact. The de-biasing power of competition lies in the fact that once an article is published showing a fact, then it becomes interesting to publish an article refuting that fact, even in cases where the negative result would not have been interesting on its own. This competition will be stronger in large fields than in small ones and there will be more effort put into debunking famous facts than ones that are less important. This seems at least reasonably efficient.

However, the ability of replication and competition to reduce researcher initiative bias will differ greatly across the types of initiative and across the nature of the data. For example, the cheapest form of researcher initiative bias is just faking the numbers in a regression table. This method provides an easy way of achieving any statistical significance that the researcher might want to achieve. However, as long as the data is public, there is little chance that an important paper written with completely faked tables will survive any scrutiny. I believe that there is almost no wholesale fabrication of results for this reason.

A second way of producing misleadingly significant results is to choose an extreme specification that yields significant, yet highly fragile coefficients. This method of misleading is considerably safer than wholesale fabrication because if caught the researcher is not actually guilty of lying. Moreover, in the subsequent debate that will surely surround claims and counterclaims, it will be difficult for outsiders to sort out who is right. Even a knife-edge specification can usually be given some sort of coherent defense.

Classic data mining is even harder to refute. After all, the correlation is actually there in the data. As the Denton-Lovell dialogue makes it clear, a misleadingly significant result can come from one data mining econometrician or twenty classical statisticians, so there is certainly no reasonable public censure that accompanies claims of such behavior (although one could imagine such censure if norms about specifying hypotheses shifted enough). If the first data set is the only one that exists, the only response is to mention that there were many variables and this is the one that was most significant and to hope that readers downgrade their assessments of the research accordingly.

Competition can check data mining primarily when new and independent data sets exist. For example, once a fact is established on U.S. wages, then a test of that fact using British wages serves as an unbiased test of the first hypothesis, and once the first fact is established as interesting the second fact will generally be considered interesting enough to be published. The existence of multiple sets therefore helps us to counter publication bias and this suggests one reason why micro results that can be reproduced in other data sets seem more reliable than macro data sets. The evolution, for example, of industrial organization from broad cross-industry regressions to within-industry studies surely increases the possibility that replication across industries can serve as a check on researcher initiative bias. Of course, the new data set must be both independent and similar enough to the existing data set for results to be seen as relevant tests of the first finding.

This reasoning also helps us to consider what sort of macroeconomic facts are more or less likely to have researcher initiative bias corrected by replication. For example, cross-country income regressions are particularly prone to this bias because there is little change in this distribution year to year and no chance of duplicating these regressions with some other sample. Cross-country growth regressions are somewhat less vulnerable because new data comes out at regular intervals. A fact established on income growth between 1960 and 1990 can then be tested using income growth between 1990 and 2000. The new growth rates are not truly independent of early growth, but they are more likely to be independent than income levels which are spectacularly correlated across decades. Within country time series is deeply problematic as well because of the high levels of year-to-year correlation of many variables. Replication has some ability to correct researcher initiative bias if time series from other countries can be used or if the variable changes at high enough frequencies so that new information is produced rapidly over time.

In general, there can be little doubt that competition among researchers can act as a check on researcher initiative bias and steps taken to ease replication are surely beneficial. For example, making data widely available seems like a particularly useful step in promoting replication. As I will discuss later, keeping replication costs low is also a case for simple methodologies.

As beneficial as competition among researchers generally may be, it should be emphasized that the competing researchers are also maximizing their own objectives which may or may not be the perfect presentation of truth. Once a famous result has been established, there are strong incentives for debunking that result and researchers will respond to those incentives and take their own initiative. As a result, it surely makes sense to have some skepticism towards the skeptics.

Two obvious examples of how refutation can be biased are again data mining and specification search. One technique of refuting a statistically significant finding is to search and find the right combination of variables which make the variable of interest appear statistically insignificant. These regressions will be informative, but since the researcher specifically sought to find the variables that would drive the result, standard significance does not apply. The new variables will generally appear to be more significant than they actually are and the original variable may as a result appear less significant than it is in reality.

Specification changes also offer a large number of degrees of freedom to potential debunkers and as such should also be treated skeptically. It is particularly hard for an outsider to infer the meaning of a result that appears significant in some specifications and insignificant in others. Again, there is need for both more econometric research in this area and more use of the econometric techniques that are available.

Point # 4: Improvements in technology will increase both researcher initiative bias and the ability of competition to check that bias.

Lovell's (1983) original work was motivated by the observation that improvements in computing technology and expanded data sets make data mining far easier. He is surely correct. In the framework those changes can be understood as reflecting an increase in the α parameter. Technological change has reduced the costs of searching across different potential explanatory variables and specifications. Since the time of that article, the costs of computing have only continued to fall and the ability to search across a wide range of specifications and variables has increased. One reasonable response to this is to be increasingly skeptical of results produced over time. According to this view, empirical results of the 1950s are more likely reflect classical conditions than results produced over the last five years.

A countervailing force is that the ease of running regressions has also reduced the costs of competition, replication and sensitivity analysis. In the 1950s, it would have been prohibitively costly to have students reproduce famous empirical results. Today such student replication is routine and seems like an ideal way to check for certain types of researcher initiative. The rise of empirics outside the U.S. also offers greater possibility for replicating facts using independent non-U.S. data sets which is also hopeful.

Point # 5: Increasing methodological complexity will generally increase researcher initiative bias.

While technological progress will have mixed effects on the amount of researcher initiative bias, methodological progress—unless it is aimed directly at the problem of researcher initiative bias—is likely to just increase the potential for researcher initiative. There are two effects that come from new and improved methodologies. First, they clearly increase the degrees of freedom available to the researcher. Unless the methodology has truly become *de rigeur*, the researcher now has the freedom to use either the old or the new techniques. If the researcher can choose to present only one of these types of results, then the researcher's action space has increased and so has the expected amount of researcher initiative bias.

A second reason for increased bias is introduced by the existence of particularly complex empirical methodologies. Methodological complexity increases the costs of competitors refuting or confirming results. The simpler the technique, the cheaper it will be for someone to try and duplicate the results either with the same or some other data set. At the extreme, an unvarnished figure allows the reader to reproduce an ocular data test in real time. However, as methods become more complex it becomes harder to reproduce results or even figure out the amount of discretion the researcher.

The appropriate response to the freedom and replication costs created by new methods is far from trivial. In some cases, insisting the older, simpler techniques be used as well can help, but the partisans of newer methods can sometimes justly claim that when new methods produce different results, this represents progress. The real solution must lie in the econometrics profession taking researcher initiative seriously and designing corrections that are based on the degrees of freedom enjoyed by an enterprising researcher.

Point # 6: The creating and cleaning of data increases the expected amount of researcher initiative bias.

A particularly potent opportunity for researcher initiative lies in the collection and cleaning of data. Leamer's (1983) essay on taking the "con out of econometrics" particularly dwells on the dangers of producing data. Anytime the researcher is closely involved in the design of the data gathering process, or even worse its implementation, there will be abundant scope for taking actions to increase expected r-squared and thereby bias results. A researcher gathering cross-national data on a subset of countries could easily decide to avoid certain countries which are unlikely to vindicate his ideas on cost related grounds. This biasing activity can even be subconscious, so a clear conscience is not even enough to ensure that some biasing didn't take place.

Introducing bias through the selective cleaning of data may be even easier. In many cases, there are abundant excuses to eliminate or alter particular data points on seemingly legitimate grounds.[1] After all, most data is flawed one way or another. Selectively cleaning data points, even when truth is replacing falsehood, introduces clear biases into estimated results. Even choosing to clean all of the data points, if this decision is made after results are observed introduces a bias since presumably the decision to clean is more likely when the results don't match the researcher's objectives.

The one reason why skewed data cleaning may end up being less problematic than skewed data production is greater ease of replication. Since the cost of data cleaning is so much lower, once the basic data set is available, it becomes easier for researchers to apply their own cleaning mechanism or use the raw data. Indeed, there are many instances of debates about selective cleaning of data and far fewer debates about selective production of data.

Specialization offers one fix for the skewed data production problem. When data analysis is separated from data production, researchers lose the ability to manipulate. There is still a legitimate fear that data producers will skew results so that data will be more exciting to future researchers, but this seems somewhat less likely. However, enhancing the division

[1] Robert Barro has been a particularly forceful advocate of restricting researcher discretion in data cleaning.

of labor also substantially reduces the incentives for data creation. After all, the reason that so many researchers put in time and energy to produce data is in the hope that it will create new information. Certainly, if we are faced with the choice between no information and biased information, the latter option is preferred.

Point # 7: Experimental methods restrict hypothesis shopping but increase researcher discretion in experimental design and data collection.

One of the most exciting trends in empirical methods over the last 40 years has been the introduction of true experiments into economics. Many of the first economic experiments were like psychological experiments conduced in the "lab" using students as subjects. More modern experiments take place in the fields and sometimes these experiments are small in scale (sending out randomized vitas to potential employers) and sometimes they are extremely large (giving out housing vouchers to a random sample of the population). The introduction of experimental methods both reduces and increases the scope for researcher initiative bias.

The positive impact of experimental methods on researcher initiative bias comes through limiting the scope for data mining. Experiments are always designed with a clear hypothesis which is then embedded in the experimental manipulation. The clear statement of a hypothesis before any data gathering occurs makes data mining much more difficult. When experiments are sufficiently expensive so that each experiment is written up, then data mining essentially disappears. The primary source of selection bias will then occur at the journal stage when insignificant results may have difficulty finding a home in print.

The reverse side of experimental economics is that experimenters have enormous degrees of freedom in experimental design. A whole host of experiments have shown how easy it is to manipulate student behavior in a lab setting. Field experiments also give great scope for experimenter initiative. Moreover, once the data is collected, the experimenter can readily clean the data with a variety of irrefutable justifications. This discretion is also present in field experiments, although the large number of people involved in some of them means that the ability of any one researcher to skew results may be limited.

As such, the basic tradeoff with experimental methods is that greater discretion in data gathering in experimental design and data collection is traded off against reduced discretion in shopping hypotheses. There are a number of nuances to this basic view. First, after testing the basic hypothesis many researchers launch into a search for interactions. Since there may be a large number of variables that can potentially be interacted with the experimental treatment, estimates of interaction effects are much more likely to suffer from research initiative bias.[2] This is a clear area where better econometric methods are needed for appraising the statistical significance of estimated interactions.

A second nuance is the impact of experimental cost on researcher initiative bias. Lower costs mean that it becomes easier for the experimenter to run several experiments and report only one. Lower costs also make it easier for other experimenters to reproduce important results in their own lab. My guess is that the positive effect of reduced costs that works through replication is more important than the negative effect that comes from greater experimenter discretion.

Point # 8: The search for causal inference may have increased the scope for researcher initiative bias.

Perhaps the greatest change in empirical methods over the past 20 years has been the increased emphasis on causal inference. While researchers in the 1970s and 1980s were often comfortable documenting correlations among clearly endogenous variables, today such work is much rarer and often confined to less prestigious research outlets. There is no question that increased attention to causal inference represents progress, but in many cases, the empirical approach to endogeneity has greatly expanded the degrees of

[2] Many researchers are very well aware of this issue, but my own thinking on it has been particularly shaped by Lawrence Katz.

freedom available to the researcher and consequently the scope for researcher initiative bias, especially in the use of instruments, especially those collected and cleaned by analysts themselves.

One impact of the rise of causal inference has been a renewed focus on older hypotheses, like the connection between schooling and education. The tendency to address established questions limits hypothesis shopping. A second impact has been the rise of true experiments, like Moving-to-Opportunity. Although there are worries about the freedom allowed in experimental design, as I discussed above, this trend also forces researchers to specify hypotheses ex ante.

The most worrisome aspect of causal inference is the prevalence of instrumental variables drawn from observational data selected, or even collected by the researcher. In that case, the scope for researcher effort is quite large, although some of this freedom is restricted when the researcher focuses on a clean public policy change that is directly targeted at a particular outcome. The problem will be most severe when the variables are general characteristics that may have only a weak correlation with the endogenous regressor.

At the very least, the choice of instruments creates the standard problem of the ability to find the instrument (or set of instruments) that produces results. In this case, instrumental variables have essentially re-introduced the data mining problem into a the debate on an established hypothesis where freedom of choice in dependent and independent variable had been purged from research. By allowing the researcher to choose any number out of a wide range of possible instruments, instrumental variables methods create great scope for selecting a set of instruments that produce the desired results. It would be shocking (to economists at least) if that freedom doesn't occasionally produce striking results.

This basic problem with instrumental variables is compounded by the other statistical problems that compound instrumental variables, especially the weak instruments problem (Staiger and Stock, 1997). One of the consequences of weak instruments is that the distribution of t-statistics becomes non-normal, and in some cases develops especially fat tails. The problem of researcher initiative bias concerns selecting those coefficients at the upper end of a distribution. Extra weight in the right hand side of a distribution makes it much easier to select a particularly large value. As a result, the problem of weak instrument bias, makes it particularly easy for researchers to find estimators the have particularly large t-statistics which further biases results.

Another way of thinking about this problem is that impact of instruments bias is increased by covariance between the instrument and the independent variable is divided by covariance between instrument and the depended variable. The univariate instrumental variables estimator is Cov(Y, Z)/Cov (X, Z) where Y is the dependent variable, X is the endogenous regressor and Z is the instrument. If there is some reason for spurious correlation between Z and Y, this will be divided by the covariance between X and Z. Another way of writing this formula is $\frac{Cov(Y,Z)/Var(Z)}{Cov(X,Z)/Var(Z)}$ which means that the regression coefficient when Y is regressed on Z is divided by the regression coefficient when X is regressed on Z. When the denominator is small, any biases in the numerator get blown up. As such, either selection of the instrument Z to have a high correlation with Y, or selective cleaning of that instrument can have a much larger impact when the correlation between Z and X is modest, even if the instruments pass conventional weak instrument tests.

How should researchers adjust to the presence of researcher initiative bias in instrumental variables? On one level, the existence of researcher initiative just presents a caution against this technique, especially when the instruments are not clearly experimental treatments. Certainly, it is especially worrying when instrumental variables estimators diverge substantially from ordinary least squares estimators, but it is in those cases that we find instrumental variable estimates most interesting. Again, we need new statistical tools, especially those that combine the robust literature on weak instruments with more realistic assumptions about researcher behavior. A more complete Bayesian approach that recognizes researcher initiative may offer some hope.

Point # 9: Researcher initiative bias is compounded in the presence of other statistical errors.

The complementarity between researcher initiative bias and weak instruments highlights a more general issue: in many cases, there is a complementarity between researcher initiative and other statistical errors in producing falsely significant results. Return to Lovell's basic data mining example and assume now that the researcher is erroneously underestimating standard errors (say by failing to cluster for correlation at the local level). In the absence of researcher initiative this failure to correct standard errors makes it x percent more likely that a variable will be statistically significant at the 95 percent level. The probability of a false positive is therefore .05+x.

If the researcher then has access to k independent variables, the probability that a false positive will be found on at least one of them is then $1-(..95-x)^k$. If x=.01 and k=10, then the probability of a false positive increase with this one percent error from .4 to .46. The one percent increase in false positives is magnified six times through data mining. As long as the error occurs in each one of variables tested or each one of the specifications tried, then the ability to try many variables or many specifications will compound the impact of the bias.

There are two obvious consequences of this complementarity. First, this complementariy makes it all the more important to eliminate other sources of false statistical significance Second, the existence of researcher initiative may make researchers less willing to take steps that would tighten up their standard errors and make significance less likely. Since these errors can have a really spectacular impact on the ability to generate positive results, they will continue to be attractive. Moreover, we should expect the perpetuation of such errors particularly in settings where data mining is easy and replication difficult.

Point # 10: The impact on researcher initiative bias of increasing the connection between theory and empirics is ambiguous.

One of the central divides in modern empirical work is the appropriate degree of connection between economic theory and empirical work. A closer connection between theory and empirics can have both positive and negative effects on the amount of researcher initiative bias. On the negative side, theoretically motivated researchers may be more committed to finding a particular type of result. Atheoretical, empirical researchers presumably care less about the sign of their effects and this may decrease the incentives for researcher initiative. This effect suggests an increased connection between theory on bias.

On the other hand, models, especially when they are written by researchers other than the empiricists themselves, specify particular hypotheses and even particular functional forms. This leads to a reduction in researcher discretion and a consequent reduction is researcher initiative bias. The division of labor is also important, as models produced after results are generated, offer no such hope for reducing bias.

Together these two points suggest that theory-driven researchers are pushed towards finding a particular result in line with the theory, empiricists without theory are prone to running too many specifications. Both forms of researcher initiative need to be considered and there certainly is no clear implication that we need either more or less theory in empirics.

Conclusion

The tradition in statistical work of assuming that researchers all follow the rules of classical statistics is at odds with usual assumptions about individual maximization and reality. Researchers choose variables and methods to maximize significance and it is foolish to act as if this is either a great sin or to hope that this will somewhat magically disappear. It would be far more sensible to design our empirical methods the way that we design our models: embracing human initiative in response to incentives. In most cases, this initiative is good for research, and even things like data mining have substantial upsides. After all, none of us would like to read through paper after paper of insignificant results.

There is the beginning of an empirical approach along these lines in the work of Edward Leamer who is clearly the pioneer in this field. However, even this work has not been used to the extent that it deserves. We need a series of both tests and techniques that respond to researcher initiative bias. Only in that way will the many benefits of researcher creativity be tied to less biased estimation.

References

Denton, F. T. (1985) "Data Mining as an Industry" *Review of Economics and Statistics* 67(1): 124-127.

Leamer, E. (1974) "False Models and Post-Data Model Construction" *Journal of the American Statistical Association* 69(345): 122-131.

Leamer, E. (1983) "Let's Take the Con Out of Econometrics" *American Economic Review* 73(1): 31-43.

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

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

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

Staiger, D. and J. Stock (1997) "Instrumental Variables with Weak Instruments" *Econometrica* 65(3): 557-586.

Sterling, T. (1959) "Publication Decisions and Their Possible Effects on Inference Drawn from Tests of Significance—Or Vice Versa" *Journal of the American Statistical Association* 54(285): 30-34.

Tullock, G. (1959) "Publication Decisions and Tests of Significance—A Comment" *Journal of the American Statistical Association* 54(287): 593.