experts or consultants could use some guidance from Professor Gray in order to avoid being caught up in the adversary nature of the proceedings and pitfalls created by the legal rules of procedure and evidence.

In *Ottaviani*, the plaintiffs' expert first asserted that a data set was too small to analyze. I believe the defendant's expert agreed. Subsequently, the plaintiffs desired to apply a formal statistical test to the data, but the court did not allow them to. Presumably, procedural rules designed to ensure fairness to both parties justify the court's decision. A similar situation arose in another case when at a pre-trial deposition an expert asserted that a 2 × 2 table should be analyzed by the chi-square test. Because of the small sample size, at trial the expert desired to use Fisher's exact test, as the computer output for the chi-square included a warning that the expected cell count was less than five in some cells so the conditions for the validity of the chi-square approximation were not satisfied. Again the court did not allow this testimony as the opposing side could not be prepared for a proper cross-exam. While new computer programs such as STATXACT may alleviate the small sample-size problem, as the data set can readily be analyzed, new approaches often occur to us after we make our first analysis. How can statisticians, especially at pre-trial depositions, appear knowledgeable and yet leave the door open for alternative analyses to be given later at trial? The problem with small samples is their low power to detect meaningful differences. Unfortunately, courts have often failed to appreciate this. With STATXACT and other programs (Goldstein, 1989) hopefully we will be more persuasive in future cases.

The ethical constraints on lawyers differ from those of academia, and experts face a number of unusual problems (Fienberg and Straf, 1991). Should one carry out an analysis that will likely not be in the best interest of the client? Should one do something that the lawyer should not do because it violates their ethical canons? A problem I have faced is the existence of other data sets that the lawyer did not tell me about. When analyses of the new data are submitted by the opposing party we do not have time properly to assess the comparative reliability and relevance to the issue at hand of the two data sets. The lawyer who has put you on the stand desires you to criticize the "new" data set, for example, to point out that some data are missing, some applicants are counted twice and so on. Statistical experts might well wish to avoid commenting without studying the data for a while, and it is tempting to assert that one should avoid any testimony. However, some of the flaws just cited may apply to the new data set. Is it fair to the court not to point them out? Is there a way to obtain a reasonable amount of time to carry out an assessment of the data? Remember, the lawyer who hired you did not tell you about it, so assume it will not help the party that hired you. I am unaware of any way prospective experts can assure that they will be given all the data relevant to the issue they are asked to study before the trial. I hope Professor Gray might offer some suggestions for avoiding these problems.

Comment

Harry V. Roberts

INTRODUCTION: WHAT IS THE RIGHT QUESTION?

When I first looked at the title, "Can Statistics Tell Us What We Do Not Want to Hear?" my reaction was, "Only with great difficulty." Professor Gray almost immediately echoed my reaction by saying, "It often appears that the most, indeed perhaps the only, effective role of statistics is to bolster decisions policymakers were prepared to take on other grounds." She added, "A corollary to the assertion that statistics are believed only when they conform to how one wants the world to look is the theory that the more closely statistics challenge one's own interest, the less likely they are to be relied upon."

The specific testing ground is the area of employment discrimination, with alleged salary discrimination against female faculty members as the principal illustration. Professor Gray provides a lucid overview of the problems in using statistics—regression analysis in particular—to illuminate the legal question of whether or not discrimination against females, minorities, or other protected groups has occurred. Her description of the evolving legal background, groundrules and guidelines for the use of statistics in discrimination cases is most helpful. The difficulties and seemingly erratic variations in the response of courts to statistical argumentation are skillfully and accurately depicted.

Harry V. Roberts is Professor, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, Illinois 60637.

By examining in detail three cases in which salary discrimination against female professors had been alleged, Professor Gray brings to life the conceptual and practical problems faced by any statistical witness who ventures onto legal battlefields. As a survivor of several such battles, I appreciate the opportunity she has given me to reevaluate the experiences I have lived through. In one respect, which I shall explain at the end of this discussion, this reevaluation has led me to change my mind on a fundamental substantive issue.

But in the end the question, "Can statistics tell us what we do not want to hear?" seems to be the wrong question, or at least an uninteresting question: we all know that people are reluctant to believe what they don't want to believe. The more interesting question is, "Can statistics tell us what we are trying to find out about employment discrimination, assuming that we are willing to listen open mindedly?"

I sense that I differ substantially from Professor Gray about the answer to this question. She says, "Confronted at their own institutions with an analysis of faculty salaries showing evidence of discrimination, often the very faculty whose stock-in-trade is persuading others of the efficacy of statistics refuse to believe what is presented to them." The key here is "an analysis . . . showing evidence of discrimination." This seems to mean that after regression analysis, to adjust for the effects of "untainted" explanatory variables, there remains a statistically significant salary shortfall for females or minorities. This is a statistical association. Can we infer causation from this association? As I shall explain, there are good reasons for having doubts.

The basic statistical problem is well known. When we leave the happy hunting ground of experimental investigation and intervention analysis – where the data more or less speak for themselves – and enter the bleak terrain of observational studies, conclusions about causation depend crucially on prior distributions, that is, upon all information other than that of the data under analysis. Prior distributions vary greatly from one person to another, so the data do not speak for themselves. Moreover, the problem of formulating and checking the underlying statistical model can be very hard. The model may not be believable; it may omit crucial features of reality.

In adversarial contests in which regression is used, the plaintiff's and the defendant's expert statistical analyses typically provide contradictory answers, even when they are working from the same underlying data base. In many instances, I believe, neither analysis succeeds: neither statistical model comes to grips convincingly with reality.

Therefore, the question I shall address is, "Can we believe what the statistical analysis seems to be telling us, regardless of whether or not we like the answer?" At the end I shall address a personal variation of that question: "Can statistics tell me what I was trying to find out in my own statistical studies of alleged discrimination?"

REGRESSION STUDIES OF ALLEGED EMPLOYMENT DISCRIMINATION

In a typical study of alleged salary discrimination, simple salary comparisons between, say, males and females, show very substantial mean female salary shortfalls. When regression adjustments are made for various explanatory variables – job-qualification proxies – the adjusted shortfalls narrow but do not wholly go away. Statistical experts for the plaintiffs and defendants argue heatedly about the validity of explanatory variables, statistical significance of adjusted shortfalls, and details of the statistical analysis. Typically little or nothing is conceded. Experts and attorneys on both sides play hard ball; the attorneys, as is their duty, are not concerned with fairness or politeness in trying to discredit the other side. Judges and juries are rarely in a position to evaluate the merit of the statistical arguments. I vividly recall a judge trying to understand levels of statistical significance: he finally convinced himself that the standard error multiples corresponding to p-values were like numbers on the Richter scale. Two was a mild tremor; five was a real earthquake.

Part of the problem is that many statisticians, like other kinds of expert witnesses, tend (as Mark Twain put it) to stretch the truth a little, or even a lot. For example,

- A plaintiff's expert may try to create the impression that *statistical significance* of an estimated female salary shortfall is tantamount to proof of discrimination.
- A defendant's expert may disaggregate the data into such small fragments that the estimated female shortfalls in the fragments are all statistically insignificant, thus creating the impression that there are no shortfalls at all.

Professor Gray gives other good illustrations of this point; I heartily agree that such tactics are unprofessional.

A STATEMENT OF POTENTIAL CONFLICT OF INTEREST

As is common in adversarial proceedings, statistical experts tend to sort themselves out as specialists on the plaintiff's side or the defendant's side of the case. Professor Gray was on the plaintiff's side in the three cases she discusses. For over a decade, I was on the defendant's side in a number of affirmative action cases. How does this sorting out occur? In my first case, I was approached by the defense. I carried out the kinds of statistical analyses that I would have done in a purely academic study. These analyses suggested reasonable doubt that a causal inference of discrimination could be supported by regression alone. Subsequently I was approached mainly by attorneys for the defense. I was approached by a plaintiff's attorney only once; I was unable to work with him for reasons having nothing to do with the merits of the case.

So my experience was exclusively with defendants. I was always able to do the statistical analysis that I fancied I would have done strictly from an academic perspective. No attorney twisted my arm to depart from my best judgment, though some attorneys eventually became pretty good amateur statisticians and often wanted to polish my prose. To help to keep myself honest, I published or otherwise made available my findings and methodology. Many of these are cited in my discussion of Dempster (1988). Most of the later citations were joint with Delores A. Conway.

In none of the cases of alleged salary discrimination that I worked on were there "smoking guns"–that is, individual events that provided direct and convincing evidence of systemic discrimination. The statistical arguments tended to be at the center of the stage. The nature of the statistical argumentation is well conveyed by Professor Gray's paper, although the three cases she considers place a greater-than-usual strain on the statistical tools, as I shall explain.

In my discussion of Dempster (1988) I reviewed in some detail the various regression studies that I worked on. In order to convey my views about Professor Gray's paper in self-contained fashion, I shall need to review briefly the major developments of those studies. First, however, I need to develop a basic point–central to the question in Professor Gray's title–that is well known to statisticians but easily forgotten in the heat of litigation.

THE PITFALLS OF OBSERVATIONAL STUDIES

This point is that regression studies to investigate the question of discrimination were observational studies. I believe that Professor Gray is too sanguine about what can be learned about causation from observational studies, *even if we are open-minded and eager to hear the answer, whatever it may be.* She refers, for example, to Sir Ronald Fisher's dismissal of evidence on the bad health effects of smoking as an "embarrassing culmination of a distinguished career." But at the time when Fisher was attacking the claims that smoking was implicated in disease, the observational evidence was by no means unambiguous. Other competent statisticians–Joseph Berkson and K. A. Brownlee, for example–were on Fisher's side, and they were not easy to refute (I tried informally, without much success). They postulated a "constitutional hypothesis": the same underlying factors that predispose people to disease also predispose them to smoking. The evidence during Fisher's time did not clearly rule out the constitutional hypothesis; I am not even sure that today's evidence does so, although I am free to admit that I am happy that I never took up smoking.

There are many other similar questions–such as the harmfulness of sodium, asbestos or radon, or the controversies surrounding global warming–on which observational evidence is far from conclusive. The same kinds of interpretive problems cloud the apparent conclusions of regression studies of discrimination.

On the use of observational studies to study discrimination, however, my position is not entirely negative. I have not seen regression studies purporting to demonstrate discrimination that rule out serious doubt about the nature of causality. But the studies that I have worked on, and many that I know about, suggest that discrimination, if present, is not a pervasive pall hanging uniformly over all aspects of the employment relationship. Moreover, these studies suggest promising directions for further exploration to learn if, where, when and how discrimination occurs and to gain insight into what can be done to mitigate it. I shall develop these issues below.

SPECIAL PROBLEMS WITH STUDIES OF DISCRIMINATION AGAINST FEMALE PROFESSORS

Before I develop my central argument, I need to point out that the three discrimination cases discussed by Professor Gray posed unusually difficult challenges for regression methodology. The key problem is that college faculties are incredibly specialized, so that in economic terms there are many, relatively small, noncompeting job groups. Many of the questions she discusses of excluding certain individuals from the regression analyses or of choosing "untainted" explanatory variables reflect this fact.

To illustrate, suppose that in a Graduate School of Business, there are two male professors aged 69. One is a professor of finance and a Nobel Laureate; the other is a run-of-the-mill professor of statistics. The salary of the professor of finance is much higher than that of the professor of statistics. Here we have two observations and two "special causes"–finance specialization and Nobel Prize. In studies like those of Professor Gray's three cases, there are many legitimate special causes and few faculty who are directly comparable.

It may offend one's sense of fairness or even of good salary administration in a university that such large salary differentials as those between the two professors should exist. As a practical matter we have a "market factor" that cannot be ignored. If both professors decided to move to another school, there is no doubt as to which would get the higher salary.

In general, the faculty salary regressions reported by Professor Gray tend to lead to endless disputes about inclusion or exclusion of particular individuals or particular explanatory variables, especially if all reasonable indicator variables are considered. From my own work, reported in broad summary below, I have come to feel that if the results are not robust to such detailed variations of the regression models, causal interpretation of the resulting analyses is on weak ground.

MY OWN EXPERIENCES: BACKGROUND

In my own experiences I encountered much less formidable obstacles. I worked on salary regressions based on large samples of employees from several large companies. To illustrate the nature of the results, I will summarize what I think I found out for a large Chicago bank [see Gray's reference, Roberts (1979) for some details of my early work at that bank]. My studies took place over several years, 1977–1985. I learned some interesting new things as the studies went on and the regression methodology evolved. What I learned was similar to what I learned in other studies; I suspect it would be similar at many large business organizations that are in the public eye. I was not alone in doing the kinds of analyses reported below. My suspicion is that many other people followed similar approaches and obtained similar statistical results, although there was no general consensus on the proper causal interpretation of these results.

I will simplify a little, but the following account captures the salient features. Initially, we disaggregated the entire sample by "entering cohorts," that is, subgroups of employees hired within narrow time periods. Within each entering cohort, we did "standard" salary regressions, regressing (log) salary on such noncontroversial explanatory variables as age, years of schooling, years of prior experience, and simple transforms thereof. The regression-adjusted female shortfalls were less than the unadjusted shortfalls, but both were statistically and substantively significant. The shortfalls were smaller but were still statistically significant for the more recent entering cohorts.

Then within each entering cohort, we disaggregated in a different way. We studied (log) salaries at hire and the subsequent rate of advancement of (log) salary. We called the first "placement regression" and the second, "advancement regression." The placement regressions gave results similar to those described in the preceding paragraph with significant and substantial female salary shortfalls. The advancement regressions showed that male and female salary advancement was roughly the same, even without regression adjustment.

Some light on the puzzle of causation seems to have been shed: if there is a problem of discrimination, it seems to be mainly localized to placement salaries. Curiously, even though the plaintiff was a government agency (OFCCP) charged with enforcement of the presidential order requiring nondiscrimination, this finding seems to have had no effect on the development of their case.

REVERSE REGRESSION

There was one interesting aspect of the placement regressions that was also present in the usual aggregate regressions based on current salary. Suppose that we look at *individual fitted values* from these regressions, modified by omitting the contribution of the female indicator variable. These modified fitted values can be interpreted as an *index of job qualifications*, because they are an index of explanatory variables ("job qualifications") with weights determined by the salary regressions.

It turned out that for any given salary, the means of these qualification indices for males and females were about the same. In other words, it appeared that male and female employees earning similar salaries were about equally qualified on average. In statistical language, it was natural to say that the *reverse regression* of the qualification index on (log) salary and the female indicator gave near zero values for the coefficient of the female indicator. Thus there was a kind of parity between males and females. This parity did not suggest that there was no discrimination, any more than did female salary shortfalls on the usual, or *direct*, regression establish that discrimination existed. But the statistical picture was not as straightforward as it had appeared initially.

In the litigation arena, the plaintiffs initially dismissed reverse regression as a devious trick. Serious discussion took place in the academic arena with sophisticated econometric debates; a convenient reference, with quite complete footnotes, is given by Dempster (1988).

HOMOGENEOUS JOB GROUPS

As time passed and litigation dragged on, we disaggregated in an additional way: we divided entering cohorts into *relatively homogeneous* job groups. In the context of a bank, one can think for simplicity of just two job groups: clerical employees and professional employees. Within homogeneous job groups, we did *direct* salary regressions: placement regressions, advancement regressions, and current salary regressions. Generally, these regressions all showed males and females essentially at parity: there were no systematic female salary shortfalls.

We also compared male and female indices of job qualifications within the homogeneous job groups, which is a special case of reverse regression in which the only independent variable is the female indicator variable. There were no systematic differences between males and females. On average, male and female hires within homogeneous job groups appeared to have about the same qualifications.

In Conway and Roberts (1986), there is a simple numerical example that captures the essence of these paradoxical findings. Essentially, the resolution of the paradox is that females were, on balance, less qualified than males; and the percentage of females in the clerical group was much higher than in the professional group. When a standard regression is done without disaggregation by job groups or use of indicator variables for job groups, the coefficient of the female indicator variable reflects the effects of the omitted variable, job group.

None of these findings establishes nondiscrimination, but they serve to localize where discrimination, if it exists, is to be found; *attention is focused on the initial hiring process*. For further study it is useful to think of *candidate pools* or *applicant pools* from which employees are initially hired. None of our salary regressions ruled out discrimination, even gross discrimination, in *hiring from these pools*. For example, there could have been many *rejected* females who were better qualified than any of the males (or females) who were actually hired. For example, highly qualified female professional job applicants could have been steered or shunted into clerical jobs.

Nor did the salary regressions tell anything at all about the ways in which candidate pools were formed; here again there could have been discrimination. Hence the salary regressions, while showing little or no evidence of discrimination, led the search for discrimination back to earlier stages of the employment process, including the possibility of what can be called "societal discrimination," as opposed to "employer discrimination."

Unfortunately, for the data bases used in our salary regressions (and most other data bases that I know of), information on rejected applicants was not available. (Information on rejected applicants is sometimes available for specific job groups, such as airline mechanics, for short periods of time. But the sheer cost and logistical burden of routinely collecting and retaining such information is so large that it is not done.)

Discrimination in formation of, and hiring from, job candidate pools has been approached in a different way. Percentages of females and minorities in an organization, or in specific job groups within an organization, have been compared with estimated percentages in the job market from which the organization hires, however that market may be defined. Salary regression studies, however, appear not typically to have been linked with parallel studies of this nature. Professor Gray does not discuss them.

FURTHER REFLECTIONS ON HOMOGENEOUS JOB GROUPS

In recent years my interests have focused on total quality management, where conventional notions of compensation policy and human resources management have been challenged by many, including Deming (1986), Grayson and O'Dell (1988) and Schonberger (1990). In world-class companies, there has been a tendency to reduce sharply the number of job classifications. In some plants there may now be, say, only two job classifications, when previously there had been one hundred. (The original proliferation of job classifications in unionized companies is often blamed on unions, but management has also been responsible.)

The pertinence to homogeneous job groups is direct: homogeneous job groups—job classifications—may be inherently discriminatory, not necessarily against females or minorities, but against anyone who wishes to improve his or her economic position. Schonberger (1990) puts it this way:

> Job classification systems have been a miserable failure. They have pushed people into narrow-niche jobs and kept them there, letting them rust. I recall my own frustration at a time early in my career. I was an industrial engineer.... Computers came along . . . and I learned some programming. I was eager for a career change—to become a computer systems analyst, where I thought the future lay. No way. The job classification says NO to lateral career shifts. . . .
>
> Here is a better idea: Press for comparable worth, but with pay for knowledge as the basis. It works like this: Mary X, a trainer in HRD [human resources department], starts at a modest base pay; it is the same as the base for many other jobs in the company. Mary can earn step increases, but gets a good raise—after three years or more—only by making and mastering a career shift: perhaps to purchasing, sales, computer programming or operations; it's a lesser raise if the shift is from training to recruiting. The person is paid, based on knowledge—not the position.
>
> To be sure, this system won't help certain acute shortages, such as in nursing. The comparable-worth law must be written to allow market-based pay in occupations for which there is a proven chronic shortage; for nursing that has been the case for decades—in many countries.
>
> On the surface, comparable worth is a pay issue. Just below the surface, there is a riptide surging in another direction: It is the seething anger of women—plenty of men, too—who are stuck in a dead-end job. These are bright, ambitious people, many with a college degree or at least some college.

> . . . They took, say, a clerical job, and can't get out. (pp. 191–192)

Hence it appears that the introduction of homogeneous job classifications into salary regression studies could have pointed the way to better understanding of the phenomenon of discrimination, and to search for ways of reducing discrimination if it occurs. Instead, for the most part, salary regressions retained their original narrow focus.

STATISTICAL LESSONS

The general statistical lessons, I think, are these:

- Observational studies should be based on models aimed at capturing essential things that are going on in the real world; the models are not ends in themselves.
- Statistical models must be modified in light of preliminary analysis and background knowledge about the subject matter, such as the system of job classification.
- One should not take any particular statistical model too seriously; believability is greater when the salient findings are robust to a wide range of detailed variations of the models.
- In the enthusiasm for statistical methodology, the careful examination of case histories should not be overlooked.

The implication for possible discrimination on university faculties is interesting. There are *many* different homogeneous job groups on university faculties, and the increasing specialization of research has tended to create more rather than fewer. Even more than for multiple job classifications in a factory, lateral transfer is generally very difficult, but it is not because the organization prohibits them. Rather, university specializations are largely noncompeting because of the enormous time costs of developing new specializations. I have a faculty colleague in the business school at Chicago who made such a shift – from management science to accounting – about a quarter century ago. I have no doubt that the shift was rewarding to him, but the shift was not easy, and examples like his are rare.

It is therefore clear that salary differences between professors will have a great deal to do with initial choices of specialization and their experiences in acquiring the skills for their chosen specializations. If there is a problem of discrimination against female professors, these choices deserve careful study as a possible explanation: formation of and hiring from applicant pools need study. This leads back to the broader questions of societal discrimination.

"CAN STATISTICS TELL ME WHAT I WAS TRYING TO FIND OUT?"

I return now to my personal reformulation of Professor Gray's question. Work on salary regressions has told me a lot about what we started out to find, although it was not what I expected originally. My personal conclusion from my own studies – which concerned large, public companies in the business sector – is that there was little evidence of *salary discrimination* against females. The process of statistical investigation, however, led to the study of organizational structure and of events that occur at the time of initial hire, where data are usually meager. There we need to know much more.

The really clear-cut examples of discrimination concern precisely these events: for example, the exclusion of blacks from professional sports or certain craft unions. Of course, I am not claiming that this redirection of focus would have shown that the large companies that I studied had been guilty of discrimination. But the resources expended on salary regressions might have been better used if they had been redirected, even if this entailed greater reliance on selected case histories and lesser reliance on statistics.

Indeed, one curious aspect of the discrimination cases I worked on was the lack of enthusiasm of attorneys for case histories – their traditional kind of evidence – once statistical regressions came into the picture. Lawyers still collected, and disputed, case histories – individual employees who claimed discrimination – but they did so in a rather half-hearted way, feeling that the battle of the statistical experts was likely to determine the outcome of the hearing or trial.

Finally, the introduction of homogeneous groups into the regressions led eventually (after a lag of several years and with the aid of Schonberger's insight) to my realization that there is a common source of potential discrimination that had not been raised in any of the cases that I had worked on: the limitation on opportunity for career development created by job classification systems. In retrospect, direct investigation of these systems might have contributed more to the cause of reducing discrimination than have all the standard salary regression studies.