Comment: Causal Mechanism or Causal Effect: Which Is Best for Statistical Science?

Paul W. Holland

My current hobbyhorse is to promote the view that statistical science does more good in the world when it concentrates on the careful *measurement* of the *effects* of causes than when it attempts to *explicate* the *causes* of effects. Well-founded measurements of causal effects are the building blocks of the successful identification of causes. Causal effects come first, not last, in the difficult process of causal inference (Holland, 1986b). In this admirable contribution to the statistics of employment discrimination Dempster seems to be riding in the opposite direction. He gives, in his words, "an explicit view of the basic mechanism of reward determination which is at best left implicit in traditional econometric models." This is a causal mechanism, i.e., an explication of the causes of the salaries that employees receive. Does Dempster's paper convince me that I should turn my hobbyhorse around and ride off with him, identifying causes at every opportunity? I don't think so, and I shall try to show how a significant portion of what Dempster accomplishes can be articulated within the structure of what I call Rubin's model (Holland, 1986a; Rubin, 1978) and does not really require explicit causal mechanisms.

A notion of employment discrimination can be developed along the lines of Rubin's model that illustrates how difficult it is to justify much of what passes as statistical or econometric analysis of this problem. The idea is quite simple—the effect of discrimination on a person's salary is the difference between their salary and what their salary would be if there were no discrimination. Such a position assumes (a) that the person's current salary is obtained under conditions of some relevant amount of discrimination, and (b) that a "control" condition of "no discrimination" can be conceived of in which the person would get a possibly different salary.

To develop some notation, let U be a population of employees and let $u \in U$ denote a particular employee. Then we have

$Y_d(u)$ = u's salary under the current, possibly discriminatory, system = d, and

$Y_c(u)$ = u's salary if there were no discrimination

(c is for "control" in the sense used by Dempster). We also let

$$\mathrm{G}(u) = \begin{cases} 1 & \text{if } u \text{ is male,} \\ 0 & \text{if } u \text{ is female.} \end{cases}$$

Dempster's Y is my Y_d, his Y^* is my Y_c and we both use G to denote gender. The primary difference between Dempster's approach and mine is our attitude toward Y_c. For me, $Y_c(u)$ is a number that is typically not observed. For Dempster, $Y_c(u)$ is an employer's posterior mean of another unobserved quantity, $Y^{**}(u)$, which is u's "true worth" to the employer. For both of us, $Y_d(u)$ is u's salary, and is a known value.

In terms of Rubin's model the causal effect of discrimination is the difference between $Y_d(u)$ and $Y_c(u)$, i.e.,

$$\text{(A)} \qquad \mathrm{D}(u) = Y_d(u) - Y_c(u).$$

Thus, $\mathrm{D}(u)$ is the difference between u's current salary and what u's salary would be if there were no discrimination. I can think of no clearer definition of the effect of discrimination on u's salary.

In the structure of Rubin's model, *causal theories* are specifications or partial specifications of the values of the responses, Y_d and Y_c. Dempster's equation (4) is a very simple causal theory; it is

$$\text{(B)} \qquad Y_d(u) = Y_c(u) + \alpha' \mathrm{G}(u) \quad \text{for } u \in U.$$

Dempster's causal model (B) yields these causal effects of discrimination:

$$\text{(C)} \qquad \mathrm{D}(u) = \begin{cases} \alpha' & \text{if } u \text{ is male,} \\ 0 & \text{if } u \text{ is female.} \end{cases}$$

Hence, due to the way Dempster has parameterized the problem, there is no causal effect of discrimination for females, whereas males have a constant discriminatory increment, α', added to their control salaries, Y_c, to produce their current salaries. It is not my purpose here to criticize this simple model but merely to show what Dempster's equation (4) means in terms of causal effects.

The question then arises as to what can the data say about α'? To begin, what are the data? We can certainly measure $Y_d(u)$ and $\mathrm{G}(u)$. Unfortunately, $Y_c(u)$ is not directly observed in typical employment discrimination cases. Dempster also includes a vector, $X(u)$, of other measured variables thought to be

Paul W. Holland is Distinguished Research Scholar and Director of the Research Statistics Group, Educational Testing Service, Princeton, New Jersey 08541.

relevant to the salary decision, and available to the analyst. At this point I think his notation is incomplete because $X(u)$, being measured in the current, possibly discriminatory, system, d, might be different if it were measured in the control, nondiscriminatory system, c. Hence, in applying Rubin's model it is proper to subscript X by a d just as Y_d is, i.e., X_d. This distinguishes X_d from $X_c = X$ measured in the nondiscriminatory control system, c, just as Y_d is distinguished from Y_c. Of course, $X_c(u)$ is typically not measured for the same reason that $Y_c(u)$ isn't.

What about Dempster's various α's? I have used α' in exactly the same way that he has, i.e., as a discriminatory increment given to males and not to females. I believe that the α in Dempster's equation (1) is the same as the α in the regression function

$$\text{(D)} \qquad E(Y_d \mid G, X_d) = k + \alpha G + X_d\beta,$$

where $E(\ \mid\ )$ denotes a conditional average computed over U. I assume (D) is linear as Dempster does, for simplicity.

Quite clearly, it is wrong, in general, to assume that the α in (D) is the same as the α' in (B) and (C). The former is an empirically determined regression coefficient while the latter is part of a causal theory that involves data not directly observed, i.e., Y_c. In Holland and Rubin (1983), we discuss a similar situation that is called "Lord's Paradox" in the psychometric literature. The result there, as it is here, is that there are assumptions that equate α and α' and there are other assumptions that do not and rarely are there data available to the analyst to distinguish between these sets of assumptions. I think Dempster's analysis also leads to this conclusion.

ASSUMPTIONS THAT EQUATE α AND α'

If we assume Dempster's causal model we have $Y_d - Y_c = \alpha' G$ so that

$$E(Y_d - Y_c \mid G, X_d) = E(\alpha' G \mid G, X_d) = \alpha' G.$$

Hence,

$$\text{(E)} \qquad E(Y_d \mid G, X_d) = \alpha' G + E(Y_c \mid G, X_d).$$

Therefore, the equality of α and α' depends on the regression function, $E(Y_c \mid G, X_d)$. This is a "dataless" regression function because the values of Y_c are not usually observed. Hence, there is usually no way to verify any assumptions we make about $E(Y_c \mid G, X_d)$. Suppose, therefore (with no justification), that it is linear and additive, i.e.,

$$\text{(F)} \qquad E(Y_c \mid G, X_d) = k_c + \alpha_c G + X_d\beta_c.$$

Then, (E) becomes

$$\text{(G)} \quad E(Y_d \mid G, X_d) = k_c + (\alpha' + \alpha_c)G + X_d\beta_c.$$

Observe that (G) is of the same form as (D) so that we may identify k_c with k, β_c with β and $\alpha' + \alpha_c$ with α. Thus, like Dempster, I am also led to a bias. The regression coefficient, α, in Dempster's equation (1) and my (D) equals $\alpha' + \alpha_c$ rather than α'. But the bias, α_c, has a different interpretation than Dempster's bias, α''. In (F), α_c is the dependence of $E(Y_c \mid G, X_d)$ on G which may not be zero for various reasons. For example, α_c might not be zero if X_d is not a complete enough set of variables describing employees (i.e., as Dempster suggests, the analyst should be using X_d^*, but it is not available) or α_c might not be zero if X_d differs from X_c, as it might if the condition of "no discrimination" involves profound changes that affect the education levels and relevant experience and training of employees. I would hold that the proper (linear) version of the assumption that gender has no effect on salary in a system without discrimination is that

$$\text{(H)} \qquad E(Y_c \mid G, X_c^*) = k^* + X_c^*\beta^*,$$

where X_c^* denotes the value of what Dempster called X^* but measured in a system without discrimination. (H) is an assumption and, I believe, Dempster would argue that it might also be false due to what he calls "judgmental discrimination." It is certainly a great, untested, and mostly *untestable* leap from (H) to the assumption that, in equation (F), $\alpha_c = 0$. If we make this leap, then there is no bias and $\alpha = \alpha'$ so that an estimate of the regression function (D) does lead to measurement of the causal effect, α', in (B) and (C). However, I do not see how one can justify the assumption that $\alpha_c = 0$ without data on the values of Y_c, which are generally unavailable. This, too, agrees with Dempster, I believe.

REVERSE REGRESSION

One difference between Dempster's approach and the one sketched above is that "reverse regression" appears to be odd and irrelevant to me, while Dempster seems to be able to incorporate it as potentially useful. I do not wish to give the impression that I wouldn't run a regression in reverse if I were trying to analyze the usual sorts of data that arise in employment discrimination cases. After all, when one is grasping at straws it is nice to have a few straws out there to grasp! However, due to my concentration on measuring the effect of discrimination on *salaries*, I find it odd to think about the reverse regression function, i.e.,

$$E(X_d \mid Y_d, G),$$

because it seems irrelevant to the measurement of the effect of discrimination *on salaries*. It would possibly be relevant to the measurement of the effect of discrimination on employee qualifications, but that is a different problem.

IS GENDER A CAUSE?

I have argued elsewhere (Holland, 1986a) that gender is not usefully thought of as a cause in many social science applications, and I would like to point out that I (and, I believe, Dempster) have remained true to this position in the present discussion. The "causes" involved here are discriminatory practices in salary administration, not the genders of the people involved. It is true that gender plays a role in the causal theory (B), but only in the sense that the causal effect of discrimination varies with the gender of the employee (which is, after all, what *discrimination* means). This distinction is blurred in the regression function, $E(Y_d \mid G, X_d) = k + \alpha G + X_d\beta$, where one is apt to call α the "effect" of G on Y_d. This is unfortunate usage and is often a source of confusion in the casual causal talk that often accompaniės regression analyses. Dempster is to be admired for avoiding such a casual approach to causation.

CONCLUSIONS

I hope I have sketched enough to show that the use of Rubin's model, with its focus on the measurement of causal effects, can be used to produce a crisp analysis of the employment discrimination problem that is very similar to much of that given by Dempster but without his need to interpret Y_c as the result of an optimal decision rule used by a thoughtful employer who invokes posterior means, loss functions and prior distributions. $Y_c(u)$ is a crucial number that we usually do not observe and which, because of this, can easily be swept under the rug and forgotten. Who really knows how Y_c should be determined? Is it possible to make serious efforts to actually *measure* some Y_c values rather than to continue to make them up? Perhaps there are some firms or parts of firms that do not discriminate in their administration of salaries; could their data be used to study Y_c directly in some specialized situations? On the other hand, because of the difficulty (and, often, the impossibility) of measuring Y_c, it should be clear that the analysis of employment discrimination differs significantly from the standard observational study in which the responses of *both* treated and control cases are always obtained. A regression analysis done either forward or backward cannot solve this fundamental problem with the analysis of employment discrimination.

I believe that the problem of employment discrimination is both serious and complex. It surely deserves a better effort than a parade of tired, old regression "paradoxes" by well intentioned men and women through countless courtrooms; if such a parade is the best that statistical science can do, perhaps it is doing more harm than good.

ADDITIONAL REFERENCES

HOLLAND, P. W. (1986b). Which comes first, cause or effect? *The New York Statistician* **38** 1–6.

HOLLAND, P. W. and RUBIN, D. B. (1983). On Lord's paradox. In *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord* (H. Wainer and S. Messick, eds.). Lawrence Erlbaum Associates, Hillsdale, N. J.

RUBIN, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6** 34–58.

Comment: Statistical Science and Economic Science

John Geweke

Professor Dempster has argued in favor of constructing models that explicitly specify stochastic components, and against the alternative of using models that introduce convenient but *ad hoc* chance mechanisms. There is increasing recognition among academic econometricians that this explicit specification is necessary for a model to be causal, that is, for a model to evaluate counterfactuals reliably and therefore to be employed for the purpose of policy evaluation. Explicitly specified stochastic components often arise from economic agents having information sets broader than analysts' information sets, as in Dempster's approach. A very successful application of this strategy is the development of asset pricing

John Geweke is William R. Kenan, Jr., Professor of Economics and Professor of Statistics and Decision Sciences, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27706.