

4-3-2008

Causal Inference in Civil Rights Litigation

D. James Greiner

Harvard Law School, jgreiner@law.harvard.edu

Follow this and additional works at: http://lsr.nellco.org/harvard_faculty



Part of the [Civil Law Commons](#), [Criminal Law Commons](#), and the [Law and Economics Commons](#)

Recommended Citation

Greiner, D. James, "Causal Inference in Civil Rights Litigation" (2008). *Harvard Law School Faculty Scholarship Series*. Paper 16.
http://lsr.nellco.org/harvard_faculty/16

This Article is brought to you for free and open access by the Harvard Law School at NELCO Legal Scholarship Repository. It has been accepted for inclusion in Harvard Law School Faculty Scholarship Series by an authorized administrator of NELCO Legal Scholarship Repository. For more information, please contact tracy.thompson@nellco.org.

Causal Inference in Civil Rights Litigation

D. James Greiner*

Draft of March 16, 2008

“If they can get you asking the wrong questions, they don't have to worry about the answers.”
Thomas Pynchon

Civil rights litigation often concerns the causal effect of some characteristic on decisions made by a governmental or socioeconomic actor. An analyst may be interested, for example, in the effect of victim race on jury imposition of the death penalty, in the effect of gender on a firm's hiring or promotion decisions, or in the effect of candidate ethnicity on election results. For the past 30 years, such analyses have primarily been accomplished via regression. But as used in civil rights litigation, regression suffers from several shortcomings: it facilitates biased, result-oriented thinking by expert witnesses; it encourages judges and litigators to believe that all questions are equally answerable; and it gives the wrong answer in situations where such might be avoided. These difficulties, and several others, all stem from the fact that regression does not begin with a paradigm for drawing causal inferences. This paper argues for a wholesale change in thinking in this area, from a focus on regression coefficients to an explicit framework of causation called “potential outcomes.” The potential outcomes theory of causal inference, which (for lawyers) may be analogized to but-for causation with a renewed emphasis on time, addresses many of the shortcomings of regression as the latter is currently used in modern civil rights litigation, and it does so within a framework courts, litigators, and juries can understand. This paper unpackages the potential outcomes paradigm and discusses its application in death penalty, employment discrimination, and redistricting settings.

In modern litigation, courts, attorneys, and expert witnesses use statistics in the hope of shedding light on questions of causation. This is particularly true in the civil rights context, where repetition of similar events makes the use of data analysis techniques attractive. The dialogue between law and quantitative methods in the civil rights area has lasted for decades, but few would characterize the relationship as happy. The disquiet is evident on both sides. On the legal end, for example, a 1980 commentator concluded that courts disregarded a substantial portion of statistical analyses they encountered in the employment discrimination context.¹ Twenty-five years later, a survey of Title VII² cases demonstrated that little has changed.³

* Assistant Professor of Law, Harvard Law School; Ph.D., Department of Statistics, Harvard University, 2007; J.D., University of Michigan, 1995. © 2008 by D. James Greiner. Thanks to Sergio Campos, Heather Gerken, Don Greiner, Ellen Greiner, Sam Gross, Sam Hirsch, Dan Ho, Louis Kaplow, Jennifer Lewis, Andrew Martin, Chris Robinson, Matthew Stephenson, Cora True-Frost, and Adrian Vermeule for helpful comments and suggestions. No one listed endorses this paper, and all mistakes are my own.

¹ Michael O. Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases*, 80 COL. L. REV. 737, 737 (1980) (collecting cases).

² 42 U.S.C. § 2000e to 2000e-17.

On the quantitative end, innovation has stagnated. During roughly the decade or so that the Supreme Court was placing its imprimatur on statistics in general and regression in particular as appropriate forms of evidence in Title VII cases,⁴ the academy was responding with scholarly examinations of quantitative issues arising in employment discrimination,⁵ capital punishment,⁶ redistricting,⁷ and other contexts.⁸ Quantitative analysts were convening panels and holding symposia to make recommendations to improve judicial understanding and use of statistical methods in litigation.⁹ Those recommendations were ignored,¹⁰ however, and a perusal of the hornbooks and looseleaves discussing the use of statistics as evidence in civil rights litigation demonstrates that the field seems fixated on methods introduced decades ago, particularly regression,¹¹ despite judicial dissatisfaction.

Judicial reluctance to engage, though not excusable, is perhaps understandable, but the lack of progress on the quantitative end is more puzzling. In the area of quantitative analysis of socioeconomic phenomena (at least, outside the law), the Earth has moved in the past two decades. The development of fast, inexpensive computers, as well as statistical techniques (such as hierarchical and Bayesian modeling) that take advantage of intense computation,¹² has contributed to a revolution in the sophistication and complexity of the analysis of political, economic, and sociological data.

³ Robert L. Nelson & Eric Bennett, *Judicial Treatment of Statistical Evidence in Cases Alleging Racial Discrimination in Employment: Now (2000-2002) and Then (1980-82)*, at 23 (unpublished monograph, available from authors, copy also on file with this author) (Courts “typically are skeptical of efforts by plaintiffs to use statistics to prove discrimination, although they also very frequently chide plaintiffs if they offer no statistical proof to bolster claims of disparate treatment.”). There have been occasional exceptions. *E.g.*, *Vuyanich v. Republic National Bank of Dallas*, 505 F. Supp. 224 (N.D. Tex. 1980), *rev’d*.

⁴ *See, e.g.*, *Bazemore v. Friday*, 478 U.S. 385 (1986); *Hazelwood School District v. United States*, 433 U.S. 299 (1977); *Castaneda v. Partida*, 430 U.S. 482 (1977); *Teamsters v. United States*, 431 U.S. 324 (1977); *see also*, *McCleskey v. Kemp*, 481 U.S. 279 (1987).

⁵ *See, e.g.*, Delores A. Conway and Harry V. Roberts, *Regression Analysis in Employment Discrimination Cases*, in *STATISTICS AND THE LAW* (Morris H. DeGroot ed. 1986) [hereinafter *STATISTICS AND THE LAW*]; Finkelstein, *supra* note 1; Fisher, *Multiple Regression in Legal Proceedings*, 80 *COL. L. REV.* 702 (1980).

⁶ *See, e.g.*, DAVID C. BALDUS ET AL., *EQUAL JUSTICE AND THE DEATH PENALTY: A LEGAL AND EMPIRICAL ANALYSIS* (1990).

⁷ *See, e.g.*, Richard L. Engstrom & Michael D. McDonald, *Quantitative Evidence in Vote Dilution Litigation: Political Participation and Polarized Voting*, 17 *URB. LAW.* 370 (1985); Bernard Grofman et al., *The “Totality of the Circumstances Test” in Section 2 of the 1982 Extension of the Voting Rights Act: A Social Science Perspective*, 7 *LAW & POL’Y* 1999 (1985).

⁸ *See, e.g.*, *STATISTICS AND THE LAW*, *supra* note 5 (antitrust, school finance, environmental regulation); DAVID C. BALDUS & JAMES W. L. COLE, *STATISTICAL PROOF OF DISCRIMINATION* (1980).

⁹ *See, e.g.*, PANEL ON STATISTICAL ASSESSMENTS AS EVIDENCE IN THE COURTS, NATIONAL RESEARCH COUNCIL, *THE EVOLVING ROLE OF STATISTICAL ASSESSMENTS AS EVIDENCE IN THE COURTS* 13-16 (Stephen E. Fienberg ed. 1989); American Statistical Association Committee on Law and Justice Statistics, *Proceedings of the Second Workshop on Law and Justice Statistics* (1983); Royal Statistical Society, *Discussion Meeting on the Role of the Statistician as an Expert Witness*, 145 *J. ROYAL STAT. SOC’Y, SERIES A* 395 (1982).

¹⁰ *See* Nelson & Bennett, *supra* note 3, at 4 (“The courts and the legal profession more broadly have not adopted the suggestions of the Panel on Statistical Assessments as Evidence in the Courts.”).

¹¹ For example, in employment, *see* RAMONA L. PAETZOLD AND STEVE L. WILLBORN, *THE STATISTICS OF DISCRIMINATION*, Chapter 6, and WALTER B. CONNOLLY, JR. ET AL., *USE OF STATISTICS IN EQUAL EMPLOYMENT OPPORTUNITY LITIGATION* § 11.03 & appdxs. D-E (Law Journal Press, New York: NY). In redistricting, *see* the discussion in Appendix A of D. James Greiner, *Ecological Inference in Voting Rights Act Disputes: Where Are We Now, And Where Do We Want To Be?*, 47 *JURIMETRICS J.* 115 (2007).

¹² *See, e.g.*, Joseph B. Kadane & George G. Woodworth, *Hierarchical Models for Employment Decisions*, 22 *J. BUS. & ECON. STAT.* 182 (2004).

Meanwhile, in the area of drawing inferences of causation, a more fundamental and less esoterically technical change has occurred. Previously, research into causation in social science, especially in observational studies, depended on the same framework still used in civil rights litigation today: a near-fetishlike focus on regression coefficients.¹³ In language similar to that currently in statistics-and-discrimination hornbooks and looseleaves,¹⁴ scholarly publications in non-legal quantitative fields would periodically intone that regression coefficients demonstrate correlation only, and that correlation does not equal causation. They would nevertheless proceed to use causal language, speaking in terms of the “effects” of variables after “controlling for” potential confounders; the “effects” were regression coefficients, and one controlled for potential confounders by including them in the right-hand side of a regression equation.¹⁵

More recently, however, much cutting edge causal research outside the law has moved away from regression coefficients toward what has become known as a “potential outcomes” framework.¹⁶ As discussed in greater detail below, the potential outcomes framework begins with a definition of a “causal effect,” not as a regression coefficient, but rather as the difference in outcomes that would occur with versus without some “treatment.” One can think in terms of a pill purporting to reduce blood pressure. Because for any particular “unit” (e.g., a patient), an analyst can only observe one potential outcome (e.g., blood pressure when the patient takes the pill), the challenge in causal inference is to fill in the outcome value that would have occurred had the unit done other than it actually did (e.g., blood pressure if the patient had taken a placebo).

Judges, litigators, and expert witnesses rarely take notice of, much less adopt in some way, paradigmatic shifts in the social sciences, and often with good reason. The near-total isolation of discrimination litigation from a potential outcomes understanding of causation is nevertheless both surprising and unhealthy. It is surprising because the fundamentals of the potential outcomes framework are familiar, perhaps even instinctive, to any survivor of first year torts; as the above blood-pressure pill analogy illustrates, the simplest form of the paradigm may be thought of as but-for causation with a special focus on time. The isolation is unhealthy because the framework can at least begin to address many problems festering in the application of statistical analysis to civil rights issues, and it can do so in a way accessible to adjudicators. For example, a complaint among judges is that the opinions of expert witnesses appear to be for

¹³ Christopher Winship and Michael Sobel, *Causal Inference in Sociological Studies*, in HANDBOOK OF DATA ANALYSIS 481, 481 (Melissa Hardy and Alan Bryman eds., Sage 2004); William A. Darity, Jr. and Patrick L. Mason, *Evidence on Discrimination in Employment: Codes of Color, Codes of Gender*, 12 J. ECON. PERSP. 63, 73-76 (1998) (collecting studies).

¹⁴ PAETZOLD & WILLBORN, *supra* note 11, §§ 6.01, 6.03, at 1, 14-15; CONNOLLY, JR. ET AL., *supra* note 11, § 11.03, at 11-8.1 (Rel. 21).

¹⁵ See *infra* note 26 and accompanying text.

¹⁶ See, e.g., Christopher Winship and Stephen L. Morgan, *The Estimation of Causal Effects from Observational Data*, 25 ANN. REV. SOC. 659, 662 (1999); see also PANEL ON METHODS FOR ASSESSING DISCRIMINATION, NATIONAL RESEARCH COUNCIL, MEASURING RACIAL DISCRIMINATION 78 (National Academies Press 2004).

The potential outcomes paradigm has also been labeled a “counterfactual” definition of causality, *id.*, although this term has been used to describe other frameworks. As is true of most good ideas, ownership over the paradigm is disputed. Compare, e.g., Paul W. Holland, *Statistics and Causal Inference*, 81 J. AM. STAT. ASS’N 945, 946 (1986) (crediting Don Rubin), with David A. Freedman, *Graphical Models for Causation, And the Identification Problem*, 28 EVALUATION REV. 267, 287 (2004) (crediting Jerzey Neyman) with James J. Heckman, *Rejoinder: Response to Sobel*, 35 SOC. METH. 135, 138-39 (2005) (crediting various econometricians) with Michael E. Sobel, *Discussion: “The Scientific Model of Causality*, 35 SOC. METH. 99-101 (2005) (spreading credit among various thinkers). Seeking to avoid this sort of dispute, I stick to “potential outcomes,” the most descriptive term.

sale.¹⁷ In the potential outcomes framework, however, the primary focus is on reproducing an imagined randomized experiment, so most of the hard quantitative work should be accomplished before the analyst knows what answer the study will produce. In other words, the framework allows an expert witness to commit to the most important aspects of a study before he or she knows whether it will favor the plaintiff or the defendant. Judges may respond better to experts who testify (truthfully) that they committed to their analyses before knowing the results.

There are other benefits to using the potential outcomes framework in civil rights litigation, particularly as compared to the current practice of addressing almost any problem with regression. Regression often gives the wrong answer, or contradictory answers, to questions lawyers and judges care about, *e.g.*, it implies that a firm is discriminating against men when in fact it is discriminating against women, or it suggests that the firm is discriminating against both men and women at the same time. In other situations, regression (particularly in more advanced forms) provides answers that are difficult for judges and juries to interpret, for instance, conclusions in terms of logarithms of odds ratios. The potential outcomes paradigm suffers less from these difficulties. Because the framework begins with the inquiry of what would have happened had the perception of some characteristic (*e.g.*, race, gender) been different at a defined time point, courts, juries, litigators, and experts need focus less attention on comprehending statistical minutiae.

Further, regression provides no framework within which to assess whether sharp causal questions, questions linked to information in available data, have been articulated. Before using any statistical technique, a quantitative analyst must articulate a question by translating the applicable legal standards and the background facts into an inquiry focusing on one or more quantities of interest, and must then assess whether available data have any information on these quantities. It is hubris to suppose that our present knowledge of the world is such that we can analyze any situation with currently available quantitative methods, and it is downright foolish to suppose that we can tackle any problem with one set of techniques. There are some situations as to which, at present, we can say nothing useful. In contrast to regression, the potential outcomes understanding of causal inference helps analysts understand when they are unable to translate facts and law into a sharp causal question linked to available data. I provide an example of such a situation in my discussion of in vote dilution suits under Section 2 of the Voting Rights Act.¹⁸

Finally, and most importantly, the potential outcomes framework makes clear that critical choices required to draw causal inferences in civil rights litigation are not mathematical. Instead, they depend on decisions about the law and on an understanding of how the world works. These are matters about which judges, lawyers, and lay people can speak as intelligently as those trained in quantitative methods.

In this paper, I introduce the potential outcomes paradigm for causation and apply it to civil rights litigation.¹⁹ I begin in Part I with the difficulties regression (as currently used in civil

¹⁷ See Samuel R. Gross, *Expert Evidence*, 1991 WIS. L. REV. 1113 (collecting references). Even if witnesses are not for sale, the practice of shopping for an expert with a favorable opinion is well-known. *E.g.*, Paul Meier, *Damned Lies And Expert Witnesses*, 81 J. AM. STAT. ASS'N. 269, 273-74 (1986).

¹⁸ 42 U.S.C. § 1973.

¹⁹ I have only run across a handful of references to the potential outcomes framework in any sort of legal setting. For example, a recent report on by a National Academy of Sciences panel on measuring racial discrimination outlined some of the basics. See PANEL ON METHODS FOR ASSESSING DISCRIMINATION, *supra* note 9, at 77-89. Perhaps because it concerned itself with measuring the effects of discrimination generally, however, the panel struggled with the issue of the timing of treatment assignment, and it dedicated little attention to litigation. For other examples, see Lee Epstein et al., *The Supreme Court During Crisis: How War Affects Only Non-War Cases*, 80

rights litigation) encounters in this area. I show that, at bottom, these problems stem from the fact that regression does not begin with any definition of a causal effect, much less one that would lead to the near-exclusive focus on coefficients characteristic of most modern expert and judicial analysis. In Part II, I unpackage and develop the potential outcomes framework for causal inference. In Part III, I apply the paradigm to the civil rights litigation setting generally as well as to three particular areas in which causal issues arise: imposition of capital punishment, private employment discrimination class actions, and lawsuits alleging vote dilution under Section 2 of the Voting Rights Act. Part IV concludes.

To illustrate some of the issues discussed, I frequently refer to a running example of a fictional employer sued for salary discrimination under Title VII by a class of female employees. I assume that the employer's computer records include information on each employee's gender, education achievement (on some scale), date of hire (which allows calculation of how long each employee has worked at the firm), and current job level (again, on some scale), as well as salary. Where necessary to illustrate principles, I simulate data for this hypothetical.²⁰

I. The Difficulties with Regression

In this Part, I briefly summarize the regression technique as it is typically used in civil rights disputes, then discuss the difficulties it encounters as a method for drawing causal inferences in this context. I demonstrate that all difficulties come from the fact that regression does not begin with a coherent framework for causal inference.

A. What Is Regression?

Regression has been explained many times; I keep this section as brief as possible.²¹ I will use exactly one technical/mathematical symbol in this paper: β , the universal representation of a regression coefficient.

To understand regression, begin with the running example identified above, and focus on the question of whether salary levels at the hypothetical firm reflect gender bias. In the overwhelming majority of employment discrimination cases,²² the analyst would propose the following, "simple" model.

Simple Model

$$\text{Salary} = \beta_0 + \beta_G^*(\text{gender}) + \beta_{JL}^*(\text{job level}) + \beta_{YE}^*(\text{years educ.}) + \beta_{YW}^*(\text{years work}) + \text{error}^{23}$$

N.Y.U. L. REV. 1, 65 (2005) and Daniel E. Ho, *Why Affirmative Action Does Not Cause Black Students to Fail the Bar*, 114 YALE L. J. 1997 (2005) (with response and reply).

²⁰ In this paper, I discuss only issues of intentional discrimination, disparate treatment as opposed to disparate impact. In addition, I limit this paper to a discussion of issues of gender, race, and ethnicity; and I use the term "race" as a shorthand for both of the latter two characteristics.

²¹ The references in note 11, *supra*, include more expansive explanations for lawyers.

²² See, e.g., Arthur S. Goldberger, *Comment*, 3 STAT. SCI. 165, 165 (1988); Thomas J. Campbell, *Regression Analysis in Title VII Cases: Minimum Standards, Comparable Worth, and Other Issues Where Law and Statistics Meet*, 36 STAN. L. REV. 1299, 1312 (1984).

²³ Gender is 0 for men, 1 for women. Note that the "error" here is not a mistake of any kind. Instead, it is the difference between what the equation above predicts and what is actually observed; the term "residual" is also common. Statisticians do not expect predictions from a model to be perfect (in fact, they would not know what to do if the predictions were perfect). The idea is that the "error," the difference between what is predicted and what is observed, is due to random variation.

One part of the appeal of this model is its simplicity and interpretability. β_0 can be understood as some baseline salary level common to all employees. β_G (“G” for gender) is the addition or subtraction in salary associated with being a woman; by assumption, this amount is constant for all women. β_{JL} (“JL” for job level) is the salary amount associated with each additional unit of job level. β_{YE} is the amount associated with each additional year of education, and β_{YW} the amount associated with each additional year worked. The question of interest under this model is whether β_G is negative and statistically significant, meaning negative in a way that is unlikely to be due to chance. If it is, the appealing thing about regression is that β_G immediately provides a rough measure of how much the defendant firm owes each member of the (female) class, because by assumption β_G represents the constant amount of salary deduction associated with being a woman.²⁴ Most introductory statistics books discuss the math used to implement this type of regression. The math can generate what are called “point estimates” for the β s, which are single numbers representing the best guess for each, and estimated “standard errors” or “standard deviations,” which are measures of uncertainty about the point estimates. Analysts can use these figures to produce, for example, an approximate interval within which the true β_G is likely (in some sense) to be found.²⁵

I clarify some vocabulary. First, the quantity on the left-hand side of the equals sign (Salary, above) is often called the “response” or the “dependent variable,” while the observed quantities on the right-hand side (gender, job level, years education, years worked) are called “explanatory” or “independent” variables. Variables included on the right-hand side of a regression equation are said to be “adjusted for” or “controlled for,” *e.g.*, a litigator might ask an expert witness, “Did your analysis control for years of education?”. As mentioned above, the β s are often referred to as “effects.” Note the causal connotation in the phrases “dependent variable,” “independent variable,” “explanatory variable,” “control for,” “response,” and “effect.”²⁶

A brief word about implementing this model is necessary. Any elementary statistical package or spreadsheet program will produce point estimates and standard errors for the β s. Whenever possible, however, a conscientious analyst checks whether a model under consideration fits the data, *i.e.*, whether the mathematical assumptions necessary to implement this model (which I do not discuss here) are reasonable for the dataset at issue. When a diagnostic shows a lack of fit (*i.e.*, that the mathematical assumptions are unlikely), a conscientious analyst alters the model. In regression, the alteration usually consists of adding combinations or different forms of the independent variables to the right-hand side of the regression equation, such as two variables multiplied together or the same variable squared.

²⁴ Even if such a figure may not serve as a basis for an award to each plaintiff, *see* International Brotherhood of Teamsters v. United States, 431 U.S. 324, 361-62, 371-76 (1977); Franks v. Bowman Transp. Co., 424 U.S. 747, 769 (1976), its availability should greatly increase the likelihood of settlement.

²⁵ Because I am interested in other, more intuitive concepts, I deliberately finesse most technical issues, such as the frequentist definition of confidence intervals, normality assumptions for small samples, the central limit theorem for large samples, etc. Note also that the description above corresponds to a frequentist framework of statistics; for some causal inference questions, it may be advisable to be Bayesian.

²⁶ The following statement is typical: “A common method for assessing statistics in employment discrimination cases is regression analysis, which isolates the impact of certain variables in affecting a particular outcome. Multiple regression analysis . . . is a method of examining the effect of independent or explanatory variables on a dependent variable.” CONNOLLY ET AL., *supra* note 11, § 11.10, 11-4. In the death penalty context, see DAVID C. BALDUS ET AL., *supra* note 6, at 164-65 (1990) (“After adjusting simultaneously for [certain] variables . . . the average race-of-victim effect” was estimated to be a stated quantity.).

Intuitively, these changes represent a belief that a one-unit change in an independent variable is not associated (even approximately) with a constant change in the response.

To illustrate, in the salary example above, after trying the Simple Model and becoming less than satisfied, an analyst might hypothesize that the longer people work, the more they get paid, but only up to a certain point. After a certain number of years, salary might tend to stabilize, or even decrease. If so, the analyst might add the variable (years worked squared) to the above regression equation, producing the following, “revised” model.

Revised Model 1

$$\text{Salary} = \beta_0 + \beta_G * (\text{gender}) + \beta_{JL} * (\text{job level}) + \beta_{YE} * (\text{years educ.}) + \\ \beta_{YW} * (\text{years work}) + \beta_{YW2} * (\text{years work squared}) + \text{error}$$

One of the maddening aspects of statistics in general, and regression in particular, is that when one adds a variable in this way, the results of the revised model may look nothing like those of the simple model. Recall that in the running salary discrimination example, our focus is on β_G . It may well happen that in a regression in which each variable appears plainly (the Simple Model, above), the estimate for β_G might be positive, statistically significant, and thus favorable (in litigation terms) to the defendant firm; but if an analyst adds the term (years worked squared) to the equation (Revised Model 1, above) the estimate for β_G is negative, statistically significant, and thus favorable (in litigation terms) to a putative female class.²⁷ Although the intuition behind this phenomenon is not easy to explain in lay terms, one can consider that adding a variable includes a different kind of information. If it is not duplicative of the information already in the model (from the other variables previously there), then this new information could change the overall results a great deal.

I close this section with two observations on model-checking, the first a general point, the second specific to regression. First, assessing how well any statistical model fits a dataset requires the exercise of judgment. Few hard and fast rules are available because whether the model is adequate for a dataset depends on, among other things, the question being asked, the nature of the lack of fit, and how bad the indicators are. Some commonly used diagnostics consist of simple graphs, with the analyst making a heuristic judgment about whether the shape of the plot is in some sense “good enough.” Second, and most important, when using regression, the analyst cannot examine the fit of the model without first implementing that model, *i.e.*, directing software to produce estimates of the β s that are the quantities of interest (particularly β_G in the above example). Thus, an expert witness in a salary discrimination case almost unavoidably sees the litigation “answer” a particular regression would produce *before* he or she assesses the goodness of fit.

B. Problems with Regression

²⁷ This general phenomenon is called “Simpson’s Paradox.” See E. H. Simpson, *The Interpretation of Interaction in Contingency Tables*, 13 J. ROYAL STAT. SOC’Y, SERIES B 238 (1951). An example of Simpson’s Paradox that may be familiar to baseball fanatics is the following: it is possible for Batter A to have a higher average than Batter B separately in both 2004 and 2005 but for Batter B to have a higher average than Batter A in the overall 2004-2005 period.

In this section, I discuss a variety of problems with the regression technique as it has been used in the civil rights context during the past three decades. A fundamental theme of this paper is that these difficulties all stem at least in part from a common source: the lack of a framework that articulates a question of interest, links it to a set of quantities to be estimated, and provides guidance for the choice of methods to accomplish the estimation. In short, what is missing from the analysis courts typically see in the civil rights context is a framework for causal inference.

1. Bias of the Analyst

“Bias” is often used in its statistical sense, where it has a precise, mathematical definition. Here, in contrast, I speak of good, old-fashioned bias, the human tendency to favor one side or another more than what an impartial assessment of available information warrants.

Over twenty years ago, J. Morgan Kousser, after years of experience testifying in voting rights cases, wrote an article entitled, “Are Expert Witnesses Whores?”²⁸ If some of them are, then regression provides ample opportunity for the fallen to ply their trade. The problem stems from three facts articulated above: first, an analyst fitting a regression sees the litigation answer before he or she assesses goodness of fit; second, deciding whether a model is adequate for the data requires judgment; and third, adding or removing variables from a regression can result in wholesale changes in the results.

The ease of stating this difficulty contrasts with its seriousness. Model-checking is typically a multi-stage process: the analyst implements a first model, assesses fit, becomes less than perfectly satisfied, implements a second model, assesses fit, compares the fit of the first model to that of the second, implements a third, etc.²⁹ A model’s fit is never perfect. At each stage of this process of exploration and assessment, the substantive result, the litigation answer, stares the analyst in the face. Only the superhuman can completely disregard the temptation to lean towards a result favorable to a chosen side, consciously or no.³⁰

2. Ill-posed Questions

Given the stranglehold regression currently enjoys on quantitative proof in civil rights litigation, it is hard to shake the belief that one can measure the causal effect of any variable by including it on the right-hand side of the equals sign in a regression equation. Such overconfidence is unfortunate. There are certain matters about which, at present, we cannot even articulate sharp, answerable quantitative questions about which available data provide information. Legislators and judges might be able to describe processes, including what they suspect to be causal processes, they would like to understand better. But before statistical techniques can be brought to bear, these vagaries must be translated into specific questions asked in terms of quantities of interest, and these questions must be examined to assess whether they are, at the present stage of our knowledge of causation and with available data, answerable.

Little in this translation process depends on an understanding of mathematical obscurities, particularly in the civil rights context. Perhaps for that reason, regression provides little assistance in the translation. One can put virtually anything on the left-hand side of the

²⁸ J. Morgan Kousser, *Are Expert Witnesses Whores? Reflections on Objectivity in Scholarship and Expert Witnessing*, 6 PUB. HISTORIAN 5 (1984).

²⁹ See Daniel E. Ho et al., *Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference*, at 1 (December 4, 2005), available at <http://gking.harvard.edu/projects/cause.shtml> (last visited September 6, 2006).

³⁰ Franklin M. Fisher, *Statisticians, Econometricians, and Adversary Proceedings*, 81 J. AM. STAT. ASS’N 277, 285 (1986); Joseph B. Kadane, *Ethical Issues in Being and Expert Witness*, 4 LAW PROBABILITY & RISK 21, 23 (2005).

equals sign in a regression equation, place virtually anything else on the right-hand side, and obtain a result (including the statistical significance of coefficients) from a computer. But without a well-posed question, this result is of no value. The result is simply a number, or more accurately, a range of numbers; what does this range of numbers mean?³¹ I discuss this issue further in the context of Section 2 of the Voting Rights Act.

3. Nature of the Model

The following difficulties with regression have a common theme, namely, they concern the form of the model used, also known as the “specification.” The issues discussed here are often treated mathematically and with technical terms. That is a mistake. Resolving these issues requires fundamental, and fundamentally legal, choices that can and should be understood intuitively. To make the issues here more accessible, I refer frequently to the running example of a gender discrimination class action alleging unequal salaries.

a. Which variables?

Should analysts include any and all variables (squares, products of variables, etc.) in regression equations? If not, what principles should guide their choices? Unfortunately, courts cannot simply leave to the experts the task of choosing among variables. For example, in salary discrimination cases, courts must often decide whether a variable measuring job level or job type should be included in the right-hand side of a regression equation. It may be that women (say) are concentrated in lower level jobs, which pay less. Typically, defendants present regressions controlling for job level, *i.e.*, including it as an independent variable on the right-hand side of the equation, while plaintiffs present regressions excluding this variable. Plaintiffs argue that the job level measure is tainted in that the defendant is assigning women to lower-level (and thus lower-paying) jobs on the basis of gender, and thus including job level in a regression “controls away” discrimination. After disputing the allegations of taint, defendants respond that including job level increases the explanatory power (the “fit”) of the model. The statistical significance of the gender regression coefficient (β_G) may depend on whether the job level variable is included. If the job level variable is included, the gender coefficient is not significant, while if the job level variable is excluded, the gender coefficient is significant. In other words, the litigation answer of the study depends on whether the job level variable is included. The question of whether to include job level (or other potentially tainted variables, such as performance evaluations or disciplinary action)³² can even be case-dispositive, and the situation has arisen so often in employment discrimination cases that it has become the subject of its own mini-literature.³³

Meanwhile, courts have struggled. Some have stumbled upon what I will demonstrate later is part of the right answer, *i.e.*, that job level (or another similar variable) should be

³¹ Douglas Adams, *THE HITCHHIKER’S GUIDE TO THE GALAXY*, in *THE HITCHHIKER’S TRILOGY*, at 129 (2000) (“‘The Answer to the Great Question . . . Of Life, the Universe, and Everything . . . Is . . . Forty-Two,’ said Deep Thought.”)

³² See, e.g., *Segar v. Smith*, 738 F.3d 1249, 1276 (1984); Connolly, *supra* note 11, at A-89, A-97.

³³ See, e.g., Ramona L. Paetzold, *Multicollinearity and the Use of Regression Analysis in Discrimination Litigation*, 10 BEHAV. STAT. & L. 207 (1992); Srijati Ananda & Kevin Gilmartin, *Inclusion of Potentially Tainted Variables in Regression Analyses for Employment Discrimination Cases*, 13 INDUS. REL. L. J. 121 (1992); Walter Fogel, *Class Pay Discrimination and Multiple Regression Proofs*, 65 NEB. L. REV. 290, 303 (1986); Finkelstein, *supra* note 1, at 750.

excluded if it is tainted by discrimination.³⁴ But they have allocated the burden of proving or disproving taint inconsistently.³⁵ Other courts have focused on a debate between “human capital” and “establishment oriented” theories of labor economics, which is said to inform this question.³⁶ Further, at least one prominent commentator argues that concededly tainted variables should be included in regressions despite the fact that they represent the mechanism for a defendant firm’s discrimination, if they improve fit and if specific, technical circumstances are present.³⁷

The job level issue in the employment discrimination context is part of a larger debate in the legal community about potentially tainted variables, which is itself part of a wider discussion on which variables to include on the right-hand side of a regression equation (or other statistical model), which in turn is a subset of a broader debate on identifying variables in a statistical model generally. An example of a dispute at the first level of generality occurred recently on the subject of whether affirmative action programs at law schools cause lower bar passage rates for African-American students. Here, the allegedly tainted variable was lower GPA, the value of which is determined after the application of an affirmative action program.³⁸

Stepping back from the question of tainted variables, the question of which variables to include on the right-hand side of a regression equation is ubiquitous but is especially problematic in the civil rights litigation context. In voting rights, for example, Justice Thomas³⁹ has suggested that regressions used to assess whether voting patterns in a jurisdiction are racially polarized should include potentially explanatory variables other than race, such as party affiliation or educational differences. Similarly, returning to the employment context, how should analysts treat variables such as gender in a race discrimination case? On the one hand, it is unlawful for an employer to make most employment decisions even partially on the basis of gender; on the other hand, gender may be a powerful explanatory variable (even in the absence of sex discrimination) for the dependent variable of interest, and a lawsuit alleging only race discrimination cannot result in an award of compensation to a class of women.

Finally, variable selection issues are not limited to the right-hand side of a regression equation but extend to the identification of the dependent variable (on the left-hand side, the

³⁴ *E.g.*, *Ottaviani v. State University of New York at New Paltz*, 875 F.2d 365, 374-75 (2d Cir. 1989); *Craik v. Minnesota State University Board*, 731 F.2d 465, 479 (8th Cir. 1984), *rev’d on other grounds*, 483 U.S. 711 (1987); *Sobel v. Clutario*, 566 F. Supp. 1166, 1180-81 (S.D.N.Y. 1983), *rev’d, remanded for reconsideration*, 797 F.2d 1478 (2d Cir. 1986); *Trout v. Hidalgo*, 517 F. Supp. 873, 886 n.47 (D.D.C. 1981); *see also* Paetzold & Willborn, *supra* note 11, § 6.13, at 35-37 & nn. 3, 5 (collecting cases).

³⁵ *Compare, e.g.*, *Trout v. Hidalgo*, 517 F. Supp. 873, 886 n.47 (D.D.C. 1981) (burden on defendant to prove lack of taint) *with* *Coates v. Johnson & Johnson*, 756 F.2d 524 (7th Cir. 1985) (burden on plaintiffs to prove taint); *see also* Ananda & Gilmartin, *supra* note 33, at 143-46 (collecting cases).

³⁶ A nice summary for lawyers of these two theories, together with an application of both to the tainted variables question, appears in Fogel, *supra* note 33, at 303-05; *see also*, *Vuyanich v. Republic National Bank of Dallas*, 505 F. Supp. 224, 266 (N.D. Tex. 1980).

³⁷ Paetzold, *supra* note 33, at 227. The specific technical circumstance is lack of multicollinearity. Speaking non-technically, multicollinearity occurs when two explanatory variables contain the same information; an extreme example is including both time in years and time in weeks as explanatory variables. *See* FRED L. RAMSEY & DANIEL W. SCHAFFER, *THE STATISTICAL SLEUTH: A COURSE IN METHODS OF DATA ANALYSIS* 347 (2002). Multicollinearity can cause estimates of regression coefficients to have a great deal of uncertainty in them.

As I discuss in the text accompanying note 69, *infra*, Paetzold is wrong here. The issue is not multicollinearity but statistical bias, particularly post-treatment adjustment bias.

³⁸ *Compare* Richard H. Sander, *A Systematic Analysis of Affirmative Action in American Law Schools*, 57 *STAN. L. REV.* 367 (2004) *with* Ho, *supra* note 19, at 105.

³⁹ *Holder v. Hall*, 512 U.S. 874, 904 n.13 (1994) (Thomas, J., concurring in the judgment).

“response”). An example of the confusion existing in this area is the debate over “reverse regression” in employment discrimination cases. Instinctively, most analysts take salary, promotion success, hiring success, etc. to be the response in regression equations, and they put observable qualification measures (*e.g.*, years worked, years education, test scores, etc.) on the right-hand side. But a group of academics has proposed another procedure, namely, considering a qualification measure as the response and making salary, for example, an independent variable. In other words, one predicts qualifications from salary instead of predicting salary from qualifications.⁴⁰ Advocates of reverse regression argue that before adjudicating a case, courts must choose between asking (i) whether salaries for men and women of equal qualifications should be the same, or (ii) whether the qualifications of men and women of equal salaries should be the same. And if the focus of causal analysis in civil rights litigation is on regression coefficients, reverse regression has surface appeal because, depending on the assumptions one makes, it might be possible to use either “forward” or “reverse” regression to estimate coefficients. Alas, the choice of whether to predict salary from qualifications or qualifications from salary can be litigation-dispositive because reverse regressions invariably provides less “evidence” of discrimination than do more traditional models.⁴¹

b. Constant Additive Effect, Problems with the “Prohibited” Variable

In the salary regression equation example above, is it reasonable to believe that the effect on salary of being a woman is the same for a plaintiff with a high job level, perhaps an upper level manager, as it is for a plaintiff with a low job level, perhaps a file clerk? To clarify, suppose the former’s current salary is \$300,000, while the latter’s is \$15,000, and the regression equation produces an estimated β_G (found to be statistically significant) of -\$1,500.00. Is the -\$1,500.00 figure plausible for either plaintiff? With respect to the upper level manager, the figure is so small relative to overall salary that one might question whether even a misogynist firm would bother to discriminate at this level. With respect to the file clerk, the figure is so large relative to salary that one might question whether a misogynist but minimally rational firm could hope to avoid detection of discrimination. Yet the assumption of the “Simple Model” above is that the effect on salary of being a woman is the same for both the manager and the file clerk, *i.e.*, that the effect at issue is constant across all individuals.

So what? The problem is that this assumption of a constant, additive effect for all women can mask discrimination when it is present and can give a false impression of discrimination when none exists. I provide an example of how the former might occur. Imagine a sexist but somewhat economically rational firm assigning salaries to men and women occupying similar positions. Some women are so productive that the firm, despite its sexism, wants to retain their services, so it provides salaries equal to similarly qualified men (*i.e.*, the effect here is \$0). Some men are so unproductive that the firm wants to ease them out, so it provides salaries designed to induce them to look for other jobs; it also, of course, does the same to unproductive women. (*i.e.*,

⁴⁰ See, *e.g.*, Conway & Roberts, *supra* note 5, with comments and reply; Delores A. Conway and Harry V. Roberts, *Reverse Regression, Fairness, and Employment Discrimination*, 1 J. BUS. & ECON. STAT. 75, 78 (1983). Reverse regression, unfortunately, refuses to die. See, *e.g.*, United States v. Delaware, 93 Fair Empl. Prac. Cas. (BNA) 1248, 2004 U.S. Dist. LEXIS 4560, at **67-77 (D. Delaware March 22, 2004). By referencing the reverse regression debate, I do not mean to lend credence to the technique, which suffers from a host of statistical problems, only some of which have been explored in the literature thus far. See, *e.g.*, Arthur S. Goldberger, *Redirecting Reverse Regression*, 2 J. BUS. & ECON. STAT. 114 (1984); John J. Miller, *Some Observations, a Suggestion, and Some Comments on the Conway-Roberts Article*, 2 J. BUS. & ECON. STAT. 123 (1984).

⁴¹ Arthur S. Goldberger, *Comment, in STATISTICS AND THE LAW*, *supra* note 5, at 182.

the effect here is also \$0) In the medium range of productivity, however, the firm feels able to express its sexist nature, providing salaries to members of both genders designed to keep most of them working, but still paying women less than equally productive men (*i.e.*, the effect is \$X). In such a situation, the simple regression model with a constant additive effect will produce one estimate of the effect on salary of being a woman, a single number for all women, that constitutes a complicated kind of average of (a) the \$0 for the highly productive, (b) the \$X for those of middling productivity, and (c) the \$0 for the unproductive. Recalling that there is some uncertainty in all statistical estimation, the averaging of the \$0, \$X, and \$0 amounts might well be statistically indistinguishable from an overall \$0 (*i.e.*, not statistically significant).⁴²

From the point of view of an economist attempting to understand how a firm functions, as opposed to an expert witness attempting to produce an analysis understandable and useful to judges and juries, relaxing this constancy assumption is easy. Returning to the running salary discrimination example, and recalling the Simple Model, above, one potential fix is to create a new variable by multiplying gender and job level together; the resulting model is still readable.

Revised Model 2

$$\text{Salary} = \beta_0 + \beta_G * (\text{gender}) + \beta_{JL} * (\text{job level}) + \beta_{YE} * (\text{years educ.}) \\ + \beta_{YW} * (\text{years work}) + \beta_{GJL} * (\text{gender}) * (\text{job level}) + \text{error}$$

The new variable is the (gender)*(job level) term (and the subscript “GJL” stands for gender-job-level). Because of it, the effect on salary associated with being female now depends on job level. For a woman⁴³ at job level 1 (whatever that means), the effect is $\beta_G + \beta_{GJL} * 1$, while for a woman at job level 2, the effect is $\beta_G + \beta_{GJL} * 2$.

In terms of usefulness in a discrimination case, however, the model has grown complicated. In the original model, we could pose a single, simple question: is the estimate for β_G , the only coefficient associated with gender, negative (and statistically significant)? In the revised model, we now have two coefficients associated with gender, β_G and β_{GJL} ; upon which should the analysis focus? What if one is statistically significant and the other is not? What if one is negative (and thus favorable to the litigation proof of a female class) while the other is positive (and thus adverse to such proof)?⁴⁴ Notice that the problem grows as we add more gender-specific terms to the regression, such as the multiplication of gender and years worked, gender and years of education, etc.

Experimental, experiential, and judicial evidence all suggest that the problems articulated above are, unfortunately, not hypothetical. An example of experimental evidence is a recently published paper in which two authors studied the callback rate for (fictitious) resumes randomized to have either African-American-sounding (such as Lakisha) or Caucasian-sounding (such as Emily) names. The two authors found that an African-American-sounding name lowered the callback rate, but they also found that the size of the negative effect varied according

⁴² See *infra* note 46 and accompanying text.

⁴³ Recall that the gender variable was 1 for women and 0 for men.

⁴⁴ Conway & Roberts, *supra* note 40, at 78 (“If [interaction effects] are not slight, the problem of interpretation becomes severe not only for statisticians but also for judges. Pronounced interaction effects could lead to data patterns in which any concept of simple discrimination or unfairness against females or minorities is lost.”)

to whether the fictional resume was “high” or “low” quality.⁴⁵ In other words, the name-race effect was not constant. On the experiential front, a pioneering expert witness in employment discrimination cases has written, “The simple regression model assuming a constant shift [β] is unrealistic . . . Virtually all highly (low) qualified employees will (not) be promoted. It is in the middle range of qualification levels that discrimination is most likely to occur.”⁴⁶ For evidence of confusion in the judiciary, see *Penk v. Oregon State Board of Higher Education*,⁴⁷ where a court downgraded the evidentiary weight of the plaintiffs’ regression coefficients because of the inclusion of a variable analogous to gender*job level in the example.

c. Problems with Variables Other than the “Prohibited” Variable

The previous section focused on the difficulties arising when the effect associated with gender is not the same at all job levels. But this problem might also arise with other variables as well; for example, the effect associated with an additional year of education might not be the same at all job levels. Or, as outlined above in the discussion of model fitting, it may be that salary tends to increase with years worked only up to a point, and that after that point, additional years worked are associated with a lower salary.

One might think that if we are focused on the effect associated with gender, problems with the other variables would matter less. Alas, such is not the case. To see why, return to the salary discrimination running example, and consider Figure 1, which plots the salaries of 100 employees (50 male, 50 female) on the Y axis against the years worked on the X axis. Because salary is on the Y axis, higher is “good” in the sense of more money. It appears that men, represented by the (blue) squares, do better than women, represented by the (lavender) triangles, for any reasonable range of years worked. Take, for example, the period from five to ten years worked, as marked by the two vertical, dotted (green) lines. In this range, most of the open squares are above most of the triangles, meaning men tend to have higher salaries than women when both have worked between five and ten years.

⁴⁵ Marianne Bertrand and Sendhil Mullainathan, *Are Emily And Greg More Employable than Lakisha And Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991, 992, 1000-01 (2004) (“[H]aving a higher-quality resume has a smaller [positive] effect [on callback] for African-Americans. . . . Most strikingly, African-Americans experience much less of an increase in callback rate for similar improvements in their credentials.”).

⁴⁶ Joseph L. Gastwirth, *Comment*, 3 Statistical Science 175, 176 (1988) (citing *Riorden v. Kempiners*, 44 F.E.P. 1355 (7th Cir. 1987) (Posner, J.)).

⁴⁷ 48 F.E.P. 1724, 1985 U.S. Dist. LEXIS 22624, at ** 147-54. (D. Ore. Feb. 15, 1985).

Male and Female Salaries Versus Years Worked

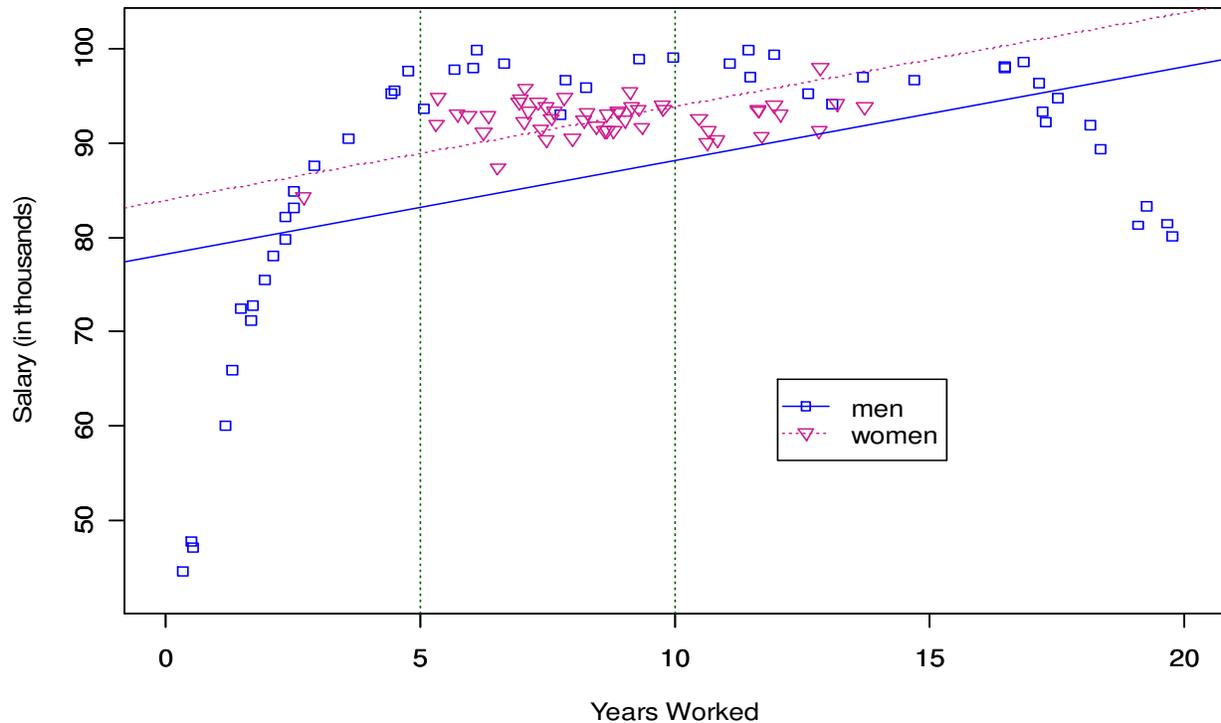


Figure 1: Salaries Versus Years Worked, Male and Female, for a Hypothetical Firm: At any given level of Years Worked, it appears that men generally have higher salaries than women. But in a simple regression approach, represented by the two parallel lines, the line for women is above the line for men, suggesting discrimination against men! Under this framework, the amount of “damages” for each man equals the (constant) distance between the two lines. The explanation for the simple regression model’s incorrect result is (i) a curve in the data (at least for men) that is not reflected in the regression equation, and (ii) the fact that men have a wider range of years worked than women.

What is the problem? The problem is that if analysts use the typical model in such cases, they would conclude that there is statistically significant salary discrimination in favor of women, *i.e.*, that men should be suing the firm. In other words, this is another instance in which the usual regression gives the wrong answer. By the usual regression, I mean the following, reduced version of the Simple Model, above:

$$\text{Salary} = \beta_0 + \beta_G * (\text{gender}) + \beta_{YW} * (\text{years worked}) + \text{error}$$

Geometrically, this model corresponds to two parallel lines, one for men, the other for women. The lines are the “best fitting;” stated loosely, they are the lines that stay as close to as many points as possible. Both lines appear in Figure 1, above, and there, one can see that the dotted line for women is *above* the solid line for men. Up means higher salary, so the fact that the female line is above the male line means that the regression model “thinks” that, given years worked, women are earning more than men. How did that happen? Notice that at the far left and far right parts of Figure 1, there are a fair number of squares but no triangles. Notice also that

most of these squares at the far left and right are below the points in the middle of the graph. The male line has to account for those lower-salaried men near the edges of the graph. In other words, the lower-earning men at the right- and left-hand portions of the plot “pull” the male line down. But there are few if any women who have worked fewer than five or more than 15 years,⁴⁸ so there is nothing to “pull” the female line down (assuming, which one should not, that female salaries would fall off in a pattern similar to the men). In the big picture, the analyst has tried to fit a single, straight line to all the male data, when the male pattern is in fact curved. That resulted in the wrong answer for the question of interest, which was whether the firm discriminates against women (and, incidentally, a statistically significant wrong answer).

Again, one might ask, what is the problem? If analysts can see this sort of thing in a graph, they will adjust the model. There are two responses. The first is that when analysts try to include additional variables such as years of education and job level, they can no longer plot all variables on a flat piece of paper, or even on a hologram. Unfortunately, the world we live in has only three spatial dimensions, not enough to visualize non-trivial data. In many (perhaps most) cases of interest, then, an analyst would not be able to see this problem. The second issue is that even with only one variable (as in the example above), this problem can be difficult to see. I constructed the upside-down U-curve in the above graph to make the problem obvious, but it will not ordinarily be so clear.

It is difficult to know whether the problem articulated here exists in real-world situations, particularly if there is more than one independent variable. Some evidence suggests, however, that the problem does exist. For example, there is some evidence that a regression specification used by economists for years to model earnings suffers from the difficulties explained above.⁴⁹

Finally, notice that in Figure 1 matters might not have gone amiss if the analyst had confined herself to the range of years worked for which there existed an appropriate number of both male and female observations, say between five and 14 years. I return to this thought in Part III.

d. Differing variable types

In the running gender example used thus far, the variable on the left-hand side of the regression equation is salary. Suppose salary takes on values from, say, \$15,000 to \$300,000. Statistically, regression depends on the assumption that salary can be any value at all, but both negative values and unreasonably high figures have such low probability that analysts typically do not bother with this annoyance. Sometimes, however, we have to deal with variables where one cannot ignore the annoyance. For example, suppose that instead of salary, we want to analyze a set of promotions recently awarded by a firm for gender bias. Here, the variable of promotion success may be only two values, 0 (failure) and 1 (success). Treating a 0-1 variable as though it could take on any value at all often goes badly.

Again, the problem is not a shortage of statistical techniques; the terms “logistic regression” or “probit regression” may be familiar. Both are methods that focus on the probability that an applicant receives a promotion, and they often work well for 0-1 response variables like promotion. These models may be conceptualized as follows.

⁴⁸ There are several reasons, some gender-neutral and some not, why men and women might have different patterns of a variable such as years worked. For example, a years-worked measurement might reflect maternity leave time. Educational achievement patterns might reflect early childhood socialization that steered women away from certain degrees or fields. Or it might be just random variation.

⁴⁹ Kevin M. Murphy & Finis Welch, *Empirical Age-Earnings Profiles*, 8 J. LAB. ECON. 202 (1990).

$$\text{Promotion probability} = \text{Fancy Math}(\beta_0 + \beta_G^*(\text{gender}) + \beta_{JL}^*(\text{job level}) \\ + \beta_{YE}^*(\text{years educ.}) + \beta_{YW}^*(\text{years work}))$$

One can again ask the question of whether β_G is non-zero and statistically significant. Luckily, for many models of this type, it is still the case that a negative estimate for β_G is good for the female plaintiffs' lawsuit. But the fancy math⁵⁰ involved in the above model is a substantial drawback in terms of both intuitive understanding of the model and its usefulness in litigation; β_G is no longer the salary addition or subtraction associated with being a woman, as it was with the Simple Model, so there is no intuitively obvious measure of the relevant effect. In logistic regression, for example, β_G would represent the logarithm of an odds ratio, not something that is readily accessible intuitively. Courts and commentators have disparaged or even rejected alternative models of discrimination for precisely these reasons.⁵¹

4. One Regression, or Two?

In an effort to address some of the issues identified in the previous sections, some expert witnesses and commentators have proposed running separate regressions, one for members of the plaintiff class and one for those outside the protected group. The idea, which is absolutely correct, is that a single regression with a constant additive treatment effect is a blunt instrument, a model ordinarily too clumsy to capture the nuances of the data, and that running separate regressions for protected and unprotected groups allows for greater flexibility.⁵² Continuing with the example of a gender discrimination lawsuit focusing on salary, an analyst might state the following two equations:

$$\text{Salary}_W = \beta_{W0} + \beta_{WJL}^*(\text{job level}) + \beta_{WYE}^*(\text{years educ.}) + \beta_{WYW}^*(\text{years work}) + \text{error}_W$$

$$\text{Salary}_M = \beta_{M0} + \beta_{MJL}^*(\text{job level}) + \beta_{MYE}^*(\text{years educ.}) + \beta_{MYW}^*(\text{years work}) + \text{error}_M$$

The symbols W and M refer, of course, to women and men; note that there is no $\beta_G^*(\text{gender})$ term in either equation because, for each equation, there is only one gender.

Continuing with the salary example, an analyst following the two regression approach divides the data into a male dataset and a female dataset, then uses the now-standard regression

⁵⁰ See RAMSEY & SCHAFER, *supra* note 37, at 580-635, for an accessible explanation of the principles involved in logistic and probit regression.

⁵¹ *See, e.g.*, *Penk v. Oregon State Board of Higher Education*, 48 F.E.P. 1724, 1985 U.S. Dist. LEXIS 22624, at **192. (D. Ore. Feb. 15, 1985) (“Because the regression analysis is a logistic regression analysis, the sex coefficients in [an expert report] represent the logarithm of the ratio of the odds of a man being in a higher group to the odds of a woman being in the higher group after all qualifications represented as variables in the equation are accounted for. These sex coefficients are not readily understandable, and plaintiffs did not make them any more so in their rebuttal report.”); *Craik v. Minnesota State University Board*, 731 F.2d 465, 476 n.14 (8th Cir. 1984); *Campbell*, *supra* note 22, at 1313 n.43 (An alternative technique “would not directly yield dollar estimates of disparate treatment, however, so continued use of regression models is preferable.”); *see also* BALDUS ET AL., *supra* note 6, at 71 (“The key measure of the impact of a given variable is the logistic-regression coefficient, which is difficult to interpret in its own right . . .”).

⁵² *See, e.g.*, PAETZOLD & WILLBORN, *supra* note 11, § 6.10; *Ottaviani v. State University of New York*, 875 F.2d 365, 374 (2d Cir. 1989).

techniques twice, once for each dataset. Now what? For an analyst in a quantitative culture such as ours that fixates on regression coefficients, one initial instinct is to compare the coefficients from the two regressions. Suppose the analyst finds that the estimate for β_{WYE} , which corresponds to the amount associated with a one-year increase in education level among women, is lower than the estimated β_{MYE} , which corresponds to the amount associated with a one-year increase in education level for men. Assuming this difference is statistically significant,⁵³ the analyst might take this difference as evidence that the defendant firm is valuing years of education less for women than for men, which might support an inference of gender discrimination.

As courts and commentators have recognized,⁵⁴ however, the problem with this focus on regression coefficients is that it often yields contradictory results. Suppose the analyst finds that β_{WYE} is significantly lower than β_{MYE} , but simultaneously finds that the estimate for β_{WYW} is significantly higher than β_{MYW} . In other words, it appears that the firm values additional years of education for men greater than for women, but simultaneously values additional worked for women greater than for men. Perhaps both men and women should sue the firm. Note that even if the comparison of regression coefficients does not produce contradictory results, the literature apparently still lacks a proposal for how to measure the damages plaintiffs have suffered.

C. The Fundamental Problem: Absence of a Causal Framework

In the gender discrimination context, should experts, litigators, juries, and judges care whether a defendant firm pays men more than equally qualified women or whether men are less qualified than equally salaried women, as the debate over reverse regression might suggest? Should the court be forced to pick between human capital and establishment oriented theories of labor economics, as the debate over including a job level variable in a regression might suggest? The first inquiry is bizarre because it seems as though a defendant is liable in either case, while the second seems removed from the legal task of assessing whether discrimination is present. In the voting rights context, should courts attempting some sort of causal inquiry focusing on the races of candidates, the races of voters, or both? What sort of evidence from statistical models, regression or otherwise, would convince them that something causally connected to race is driving voting patterns?

Few would doubt that litigation often commonly turns on the esoterics of specialized fields, and when legal questions depend on exploring the unfamiliar, courtroom actors learn what they must. An oft-cited danger in such situations is that generalist judges and lay juries will fail to understand what they see and hear, even with material presented by able litigators and articulate, chaste experts. But in my view, an underappreciated danger in such contexts is that of undue immersion such that legal arbiters or triers of fact lose track of what it is that they need to decide, on the questions they need to answer.

In the civil rights context, when experts analyze data with information about repetitive events, such as imposition of capital punishment, employment decisions, and elections, what's the question? The discussion above has demonstrated that all too often, the quantitative question has been the significance of regression coefficients (usually a single regression coefficient). Is that the correct question? The argument in favor of coefficients is that they capture a particular

⁵³ I have not personally run across a proposal to test for significance in this setting, but constructing one does not seem too difficult. One could, for example, examine the posterior distributions of the two coefficients after fitting the regression models with Bayesian techniques. *See, e.g.*, ANDREW GELMAN ET AL., BAYESIAN DATA ANALYSIS 353-85 (2d. ed. 2004).

⁵⁴ *See, e.g.*, Vuyanich, 505 F. Supp. at 278; PAETZOLD & WILLBORN, *supra* note 11, § 6.10, at 27-28 & n.7.

kind of conditional association, and that a legal decision maker can, perhaps in conjunction with other evidence, use the existence of the association to support an inference of causation. But the previous sections have demonstrated that this alone will not suffice. Without a theory to specify the legally relevant association, analysts lack a principled basis to decide, for example, what variables to include in a regression, in what form, whether to use one regression or two, and even what to use as the outcome of interest. The need for theoretical guidance leads to two conflicting impulses: first, (over)immersion in the substantive field in which the civil rights question is operating (punishment theory, labor economics, and electoral politics); and second, a reactionary demand that analysts produce something simple and understandable (*e.g.*, a regression coefficient corresponding to a constant, additive effect on salary associated with gender).

We should tolerate this situation only if we must. In other words, we should continue using regression in the way that we have used it thus far only if we cannot come up with something better.

II. Potential Outcomes

In this Part, I articulate the potential outcomes understanding of causal inference, a general framework that defines a causal effect and provides guidance on how to use data to draw inferences about the effects associated with an identified cause.⁵⁵ A thumbnail sketch of the framework is as follows. An analyst begins by identifying “units” that can be subjected to different “treatments” and by picking an “outcome” variable. It may help to think of hospital patients taking a pill or not taking a pill, where the hospital is measuring blood pressure levels. A causal effect is defined as a comparison between (i) the value a particular unit would have for the outcome variable after the application of a treatment at a given time, and (ii) the value that same unit would have for the same outcome variable after application of some alternative treatment at that same given time. Unfortunately, for any single unit (call it “Unit A”), an analyst can only apply one treatment at a given time, so only the outcome associated with the treatment actually received can be observed. Thus, the analyst must search for a way to fill in Unit A’s missing (counterfactual) outcome value, the one corresponding to the alternative treatment. One way to accomplish this task is to find a unit (call it “Unit B”) which received the alternative treatment; Unit B can donate its value as Unit A’s missing counterfactual outcome. For such a donation to make sense, Unit A and Unit B must be similar to one another; if they are dissimilar, any difference in their potential outcome values could be due to dissimilarity rather than the differing treatments to which they were exposed. The best way to assure similarity over a set of units is via a randomized experiment because randomization assures that all variables other than the treatment are statistically similar; thus, in causal inference, a randomized experiment is the gold standard. In the civil rights context, a randomized experiment is impossible, so analysts should use observational data to recreate the circumstances of a randomized experiment to the extent possible. At bottom, this process of recreating a randomized experiment typically involves searching for units that are similar to one another in all observable ways (as measured before treatment, not after) except treatment, and ignoring the data from units that have no counterparts.

⁵⁵ For technical primers on potential outcomes, see Guido W. Imbens and Donald B. Rubin, *Rubin Causal Model*, in PALGRAVE DICTIONARY OF ECONOMICS 1-17 (2006), Winship & Morgan, *supra* note 11, and Paul W. Holland, Rejoinder, 81 J. AM. STAT. ASS’N 968 (1986), in addition to the sources cited in note 16.

A. Treatment, Units, and the Fundamental Problem

A potential outcomes understanding of causation begins with identification of four fundamental concepts: units, a treatment, the timing of treatment assignment, and an outcome of interest. The units are the things upon which a treatment operates. The nature of the causal inquiry is to discern how the treatment affects the value of a specified outcome. Assume for the moment that there are precisely two forms of treatment, denoted “M” and “F”; if so, for a single unit, then there are two potential outcomes, the first that would occur if a unit receives treatment M at a particular moment, the second that would occur if the unit receives treatment F at the same moment. The causal effect for a single unit involves some comparison of the value of the outcome under treatment M to the value of outcome under treatment F; one typical comparison is the difference between the two quantities (*i.e.*, one minus the other). The role of time in these definitions is critical. The analyst must specify the timing of treatment assignment. The very concept of a “treatment” refers to something applied at a particular time.

For a single unit called “Unit 1” that receives treatment M, the concepts in these definitions can be represented effectively in the following chart, which also introduces some abbreviations (*e.g.*, “T” for “Treatment,” “xx” if something is observed, “??” if something is not observed). The chart will grow horizontally and vertically as additional concepts are defined.

Table 1: Causal Inference Table for a Single Unit, Assuming Exactly One Form of Treatment

Unit # = “#”	Treatment	Outcome(if treatment M received)	Outcome(if treatment F received)
1	M	some observed value = xx	<i>missing/unobserved = ??</i>

Again, the causal effect of the treatment on Unit 1 is some kind of comparison between the outcome under treatment M and the outcome under treatment F (at the same moment in time), often Outcome(F) – Outcome(M). Immediately, what has been called the “fundamental problem of causal inference”⁵⁶ becomes evident: it is not possible to observe directly the causal effect for any single unit. Even in the simplest possible case, a single unit and a treatment that can take on only two values,⁵⁷ one of the potential outcomes is missing. Assumptions and more information are needed for quantitative techniques to become relevant.

To make these concepts more concrete, return to the running example of a salary discrimination case focusing on gender, and define the treatment as being perceived⁵⁸ as male, “M,” versus being perceived as female, “F,” at the same moment in time.⁵⁹ Then Table 1, above, for a unit assigned treatment F looks like the following.

⁵⁶ Holland, *supra* note 16, at 947.

⁵⁷ If the treatment could take on, say, four values, the chart would have a potential outcome column for each treatment value (and thus four potential outcome columns).

⁵⁸ See notes 86-88, *supra*, and accompanying text, for a discussion of why I use “perceived” as opposed to “actual” gender here.

⁵⁹ A critical assumption here is that there is something definable called “M” and something definable called “F” that is the same for all units. There must be only one form of “M” and one form of “F.” This assumption would be violated if, for example, it were nonsensical to speak about a unit’s being perceived as male without discussing how “manly” the unit is. In that case, we would need to have separate columns labeled, say, “Salary under T = Very Manly” and “Salary under T = Moderately Manly,” along with the T = F (or multiple forms of the T = F column).

The assumption that there is only one form of each treatment, together with the requirement (discussed *infra*) of replication across different units, may be seen as the potential outcomes framework’s defense against the theoretical and philosophical attack that all forms of counterfactual reasoning are impossibly non-specific and inherently value-laden. See, *e.g.*, Robert N. Strassfeld, *If . . . : Counterfactuals in the Law*, 60 GEO. WASH. L. REV.

Table 2: Causal Inference Table for a Single Unit, Salary Discrimination, Gender

#	T	Salary(F)	Salary(M)
1	F	some observed value = “xx”	<i>missing/unobserved = ??</i>

At this point, those with minimal training in tort law should be comfortable. Tables 1 and 2 can be thought of as representations of but-for causation, the most basic causation concept in law.⁶⁰ Individual tort cases typically involve a comparison of what actually did occur with what would have occurred had some specific act or omission not taken place. In such a case, one can think of the treatment as the act or omission (say, T = 1) or the absence of the act or omission (say, T = 0). The trier of fact in individual cases uses the evidence presented at trial and its own understanding of how the world works to fill in the missing potential outcome and, subject to other relevant legal principles, decides the case accordingly. The critical concept implicit in but-for causation as applied in ordinary tort cases but often lost in the civil rights context is time.⁶¹ In tort cases, the focus on a particular act or omission typically makes it easy to identify a specific time at which the treatment (the allegedly tortious act or omission) occurred. In the salary discrimination example in Table 2, however, the problem can be more difficult, as discussed below.

B. Additional Units, Non-Interference

As noted above, the fundamental problem of causal inference is that for any individual unit at least one of the potential outcomes is missing, and this problem makes additional information and assumptions necessary before statistical techniques become useful. Ordinarily, the additional information comes in the form of observations of other units, and a standard assumption is that the units do not interfere with one another.⁶² In other words, with a set of multiple observations, the potential outcomes for one unit do not depend on the treatment a different unit receives.⁶³ If so, then an expanded version of Table 2, above, for a dataset with,

339 (1992); *see also* H.L.A. HART & TONY HONORE, CAUSATION IN THE LAW (2d ed. 1985). In the potential outcomes framework, we imagine each unit’s situation exactly as it was except for a single, sharply defined change as of a particular moment in time. Replication assures that puzzles posed by, for example, an unusual *novus actus interveniens* (superseding event), *see id.* at xlv, should not affect overall estimates because such bizarre chains of events typically do not happen repeatedly.

⁶⁰ *E.g.*, Strassfeld, *supra* note 59, at 346.

⁶¹ *See* Gastwirth, *supra* note 46, at 176 (“The failure of some courts to realize the importance of time and the need for exchangeability, in my opinion, has led to far more ‘legal mischief’ than some of the technical issues concerning regression analysis that have dominated the statistical literature.”).

⁶² If the units *do* interfere with each other, the chart for two units, where the first received active (T = 1) and the second received placebo (T = 0), would be as follows (where “O” stands for “outcome”).

#	T	O(T ₁ = 1, T ₂ = 1)	O(T ₁ = 1, T ₂ = 0)	O(T ₁ = 0, T ₂ = 1)	O(T ₁ = 0, T ₂ = 0)
1	1	??	xx	??	??
2	0	??	xx	??	??

⁶³ In some situations, this non-interference assumption is hard to recognize and assess. For example, if one attempts to draw causal inferences about judicial behavior from appellate decisions, and in doing so one treats cases as units, then the non-interference assumption means that precedent has no relevance or force in judicial decision making. *See, e.g.*, Epstein et al., *supra* note 19. Litigators might find such an assumption unsettling.

say, three units receiving treatment F (perceived as female) and five receiving treatment M (perceived as male), appears below.⁶⁴

Table 3: Causal Inference Table for Multiple Units, Salary Discrimination, Gender, Assuming Non-Interference and Exactly One Form of Treatment

Unit #	Treatment	Salary(M)	Salary(F)
1	F	??	91,200
2	F	??	92,400
3	F	??	94,100
4	M	47,100	??
5	M	97,700	??
6	M	98,900	??
7	M	95,300	??
8	M	81,300	??

Again, notice that half of the information we would like to have is missing. Specifically, what is missing is the counterfactual for each unit, *i.e.*, the outcome under the treatment that was not applied to that unit. To proceed, there must be some way to fill in all of the boxes in Table 3 that have the ?? symbol.

C. Donating Values

With the problem defined in this manner, one instinctive way to fill in Unit 1’s missing salary under treatment M is to substitute an observed value from a unit that actually did receive treatment M, *i.e.*, an observed salary from Units 4-8. Lacking a better idea, the analyst might choose among Units 4-8 at random with equal probability (1/5) and “impute” the chosen unit’s salary as Unit 1’s salary under treatment M (then do the same for Units 2-3). For Units 4-8, the analyst does the opposite, that is, picks a unit from 1-3 with probability 1/3 and imputes the corresponding salary. In essence, observations “donate” observed outcomes to one another to fill in each other’s missing potential outcomes; one can also refer to “imputation” of potential outcomes. An example of one iteration of this procedure produces the following table, with imputed values in *bold italic* type.

Table 4: Causal Inference Table for Multiple Units, Salary, Discrimination, Gender, Values Imputed for Missing Potential Outcomes

Unit #	Treatment	Salary(M)	Salary(F)
1	F	<i>81,300</i>	91,200
2	F	<i>47,100</i>	92,400
3	F	<i>95,300</i>	94,100
4	M	47,100	<i>91,200</i>
5	M	98,900	<i>94,100</i>
6	M	97,700	<i>91,200</i>

⁶⁴ I have taken these values, as well as those in subsequent tables in the section, from some of the artificial data graphed in Figure 1.

7	M	95,300	92,400
8	M	81,300	94,100

With the missing potential outcomes filled in, calculating any quantity of interest is straightforward. For example, for an estimate of the average difference in salary caused by the perceived gender treatment, an analyst can calculate the average in the Salary(F) column minus the average in the Salary(M) column.⁶⁵

D. Randomization

From the description above, it should be clear this procedure will produce misleading results without another critical assumption: the units perceived male cannot be systematically different from the units perceived female in some relevant way the analyst does not observe.⁶⁶ If, for example, Units 1 and 8 differ in a salary-relevant way other than the treatment given to each way, then this business of using Unit 8’s observed salary for what Unit 1’s salary would have been had Unit 1 been perceived male is a bad idea. Any causal effect that an analyst attributes to the treatment might really be due instead to the other difference between Units 1 and 8.

The best way to assure⁶⁷ that the units assigned M are not systematically different from the units assigned F is to assign the treatment (M or F) randomly. This is what makes random assignment such a powerful procedure in causal inference. Random assignment assures that, in the absence of bad luck, units who receive one treatment are not systematically different from those who receive the other treatment.

More specifically, randomization assures that, with an important set of exceptions discussed immediately below, any variable that might affect the potential outcomes will look approximately the same for the group assigned treatment M as it does for the group assigned treatment F. By “look approximately the same,” I mean that the distribution of the variable will be the same in the M and the F groups, *i.e.*, the same among men and women. In the salary discrimination running example, if it were possible (see below) to randomize the perceived gender treatment, the randomization would assure that the pattern of years of education values for men would look roughly the same as the pattern of years of education values for women. Because of this similarity, something quantitative analysts call “balance,”⁶⁸ it is unlikely that any systematic difference in salary between men and women is due to disparate lengths of time in school. Note that this balance would occur even if no one measured the years of education variable. That, again, is the power of randomization; it even balances variables that analysts do not see.

⁶⁵ There are various ways to generate uncertainty measures. For example, an analyst can repeat this (random) imputation procedure a great many times, calculating the average of the Salary(F) column minus the average of the Salary(M) column each time, to obtain a distribution of the possible values of average Salary(F) minus the average Salary(M).

⁶⁶ In more precise terms, the assumption is that the potential outcome vectors (the *two* salaries for each unit in the example above) for the persons perceived as male (treatment M) are not systematically different from the potential outcome vectors for the persons perceived as female (treatment F).

⁶⁷ Technically, “assure” is too strong. When there are a sufficient number of units, random assignment makes systematic differences between treated and control units unlikely in a way analysts can quantify and handle.

⁶⁸ See, *e.g.*, Paul R. Rosenbaum & Donald B. Rubin, *Reducing Bias in Observational Studies Using Subclassification on the Propensity Score*, 79 J. Am. Stat. Ass’n. 516, 517-19 (1984).

The previous paragraph explained that randomization balances variables that might have a role in determining the potential outcomes, subject to an important set of exceptions. The exceptions are variables that are themselves affected by the treatment. This is as it should be; that is, analysts do not ordinarily want balance in variables affected by treatment.⁶⁹ An example (standard in the quantitative literature) demonstrates why: suppose an analyst is assessing the effect of smoking on death, and suppose it were possible to randomize some people to smoke and some to avoid smoking. Would the analyst expect randomization to assure that the percentage of people who contract lung cancer be similar among smokers and non-smokers? The answer is “No,” if incidence of lung cancer is itself affected by the treatment received (smoking or non-smoking). Nor should the analyst desire balance on this variable, particularly if lung cancer is a means by which smoking (the treatment) induces death (the outcome of interest). Speaking in commonly-used terms that are dangerously unclear, an analyst assessing the effect of smoking on death should not “control away” the “effect” of lung cancer.⁷⁰

Thus, in a randomized experiment, the gold standard in causal inference, a clear distinction exists between variables that cannot be affected by the treatment and variables that might be so affected. Randomization can, and should, balance the former. Randomization does not, and should not, balance the latter. The importance of time, specifically the time at which treatment is assigned and applied, is again evident. If a variable is measured before the assignment of treatment, then that variable cannot be affected by treatment. If the variable is measured after application of treatment, an analyst must think carefully as to whether the treatment could affect it. To return again to the salary discrimination running example, suppose that the years of education variable is “measured,” in the sense of recorded in the employer’s computer database, after the employer finds out whether the potential employee is male or female, *i.e.*, after treatment. It might be plausible to believe nevertheless that the years of education variable is unaffected by perceived gender. An analyst might assume reasonably that the process of recording the number of years an employee spent in school is a mechanical matter of transferring a value written on a resume or a job application to a computer file, and that there is no bias in this process related to perceived gender. If so, then an analyst should desire and expect that randomization would balance years of education. In contrast, and as explained in greater detail below, it is not clear that a similar assumption would always be plausible with respect to a job level variable; a firm might assign persons to jobs on the basis of perceived gender.

Finally, a word about vocabulary: variables that cannot be affected by treatment are often called “covariates,” while variables that are affected by treatment are called “intermediate outcomes.” Thus, the analyst’s job is to separate covariates from intermediate outcomes and to examine the latter with particular care.

E. Observational Studies

⁶⁹ Paul R. Rosenbaum, *The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment*, 147 J. R. STAT. SOC’Y, SERIES A, 656 (1984); *see also* PAUL R. ROSENBAUM, OBSERVATIONAL STUDIES 73-74 (2d ed. 2002).

⁷⁰ Although harder to visualize, the reverse can (and does) also happen. That is, in the smoking, lung cancer, and death hypothetical, one danger is that incorrectly “controlling for” or balancing on lung cancer (an intermediate outcome, a variable that might itself have been affected by treatment) might make it appear that no causal effect exists when in fact such an effect does exist. In some situations, however, incorrectly “controlling for” or balancing an intermediate outcome can make it appear as though a causal effect exists when in fact there is no such effect. *See, e.g.*, Ho, *supra* note 19.

Experiments that randomize race and gender are not possible.⁷¹ In the civil rights context, courts are typically limited to observational data. In observational studies, the analyst controls neither the treatment each unit receives nor the time at which the unit receives the treatment. This lack of control means that the analyst cannot depend on randomization either to distinguish covariates from intermediate outcomes or to balance the covariates. For these reasons, even though potential outcome tables such as Table 4 can (and should) be constructed in observational studies, it is not immediately clear how to identify units to donate outcome values to one another because it is not clear whether the “donor” unit is sufficiently similar to the “donee” unit to make the transfer of values plausible. Systematic differences of the type discussed in section II.D may exist.

A fundamental premise of the potential outcomes framework as applied to observational studies is that because randomization has neither distinguished covariates from intermediate outcomes nor balanced covariates, analysts must do so themselves. In other words, the goal in an observational study is to recreate to the extent possible a randomized experiment. To accomplish this goal, the analyst must (i) sharply identify units, treatment, the timing of treatment assignment, and the outcome of interest; (ii) assess the plausibility of underlying assumptions, such as that units do not interfere with one another; (iii) separate covariates from intermediate outcomes; (iv) examine the distributions of the covariates in the treatment groups to see whether they are balanced, and if not; (v) balance these distributions. The first four steps are clear enough. The last, balancing the covariate distributions, means looking for subsets of the data in which the two treatment groups have similar patterns of covariate values. Units with covariate values in one treatment group that lack rough analogs in the other treatment group are “discarded” in the sense that they are not used for inference.

To illustrate this balancing process, I return to the data from Figure 1, above, which depicts salaries versus years worked for men and women and is thus a continuation of the running salary discrimination example. Recall that Figure 1 demonstrates a problem with the regression model standard among employment discrimination experts, as represented by the two parallel lines that stretched across the entirety of the graph. Figure 1 shows that even though a visual inspection of the data depicts men being paid more than women at any level of years worked in which both genders are represented, the standard regression model produces a statistically significant result in women’s favor, a result adverse to a female class’s litigation proof. Graphically, the line for women is above the line for men.

Figure 2 reproduces the Figure 1 data with a critical difference: the outcomes of interest (salaries) have been removed. The search for subsets of the data that might be reasonably analogized to a randomized experiment involves looking for regions of the years-worked variable in which both treatment types are well-represented. In Figure 2, the vertical lines at

⁷¹ In so-called “audit studies,” some entity is visited by two “testers,” one (say) white and one non-white, who resemble each other as much as possible in other ways and who follow pre-written scripts in applying for some good or service. The hope is that a comparison of the offers received by the two testers allows measurement of the effect of race on the entity’s behavior. In the language of the potential outcomes framework, the hope is that the two testers are sufficiently similar so that the white tester can donate his or her observed outcome to be the counterfactual outcome for the non-white tester, *i.e.*, the value that would have been observed for the non-white tester had he or she been white. In this setting, a great many things can be randomized. *See, e.g.*, Ian Ayres & Peter Siegelman, *Race And Gender Discrimination in Bargaining for a New Car*, 85 AM. ECON. REV. 304 (1995) (randomizing, among other things, the order in which the testers approached the target). Unfortunately the race (or gender) of the testers cannot be randomized. This is not to say that audit studies lack probative value, but rather that they are a species of what one usually must resort to in the civil rights context, namely, observational studies.

values of 5, 10, and 14 years worked divide the data into Regions 1-4, as labeled on the graph. The data in Regions 1 and 4 do not look as though they came from a randomized experiment; only one of the 30 or more units in these regions received treatment F (perceived female).⁷² It would be unwise to attempt inferences in these two regions. In Regions 2 and 3, however, the years worked covariate appears reasonably balanced, *i.e.*, there appears to be a sufficient number of units assigned each treatment to allow inference to proceed. Notice that the analyst can (and should) identify these regions without using the outcome (salary) data.

To be perfectly clear: the implication of applying the potential outcomes framework to these data is that estimation will not be attempted for employees with fewer than five or greater than (about) fourteen years worked. The analyst discards data, and discarding data ordinarily carries a cost, namely, loss of precision. In other words, the resultant causal estimates will typically have wider confidence intervals and are thus less likely to be deemed statistically significant. That is the price to be paid for relaxing the implausible assumptions regression makes.⁷³

⁷² A randomized experiment need not have a .5 probability of assigning each of two treatments to produce the favorable properties discussed above. If, however, the probability of one of the two (or more) assignments gets too low, it becomes difficult to draw inferences about the effect caused by that treatment without a lot of data. With extremely low treatment probabilities, there often are not enough observations with a particular treatment to discern patterns. Figure 2 again provides an example. How does salary vary by years worked for units perceived as female in Region 1? With only one observation, it is difficult to say.

⁷³ If an analyst perfectly specified a regression equation over the whole range of the data, discarding units in Regions 1 and 4 (and the corresponding loss of precision) would not be necessary. As discussed *supra*, however, claims of perfect specification are difficult to credit.

To the extent that it counsels discarding uninformative data and thus suffering a loss of precision, the potential outcomes framework might be deemed to favor defendants. In contrast, the implication that intermediate outcomes not be included in balancing (corresponding roughly to the assertion that tainted variables not be included on the right hand side of a regression equation), see *supra* notes 69-70 and accompanying text, is ordinarily deemed to favor plaintiffs.

Male and Female Years Worked Only (Data from Figure 1)

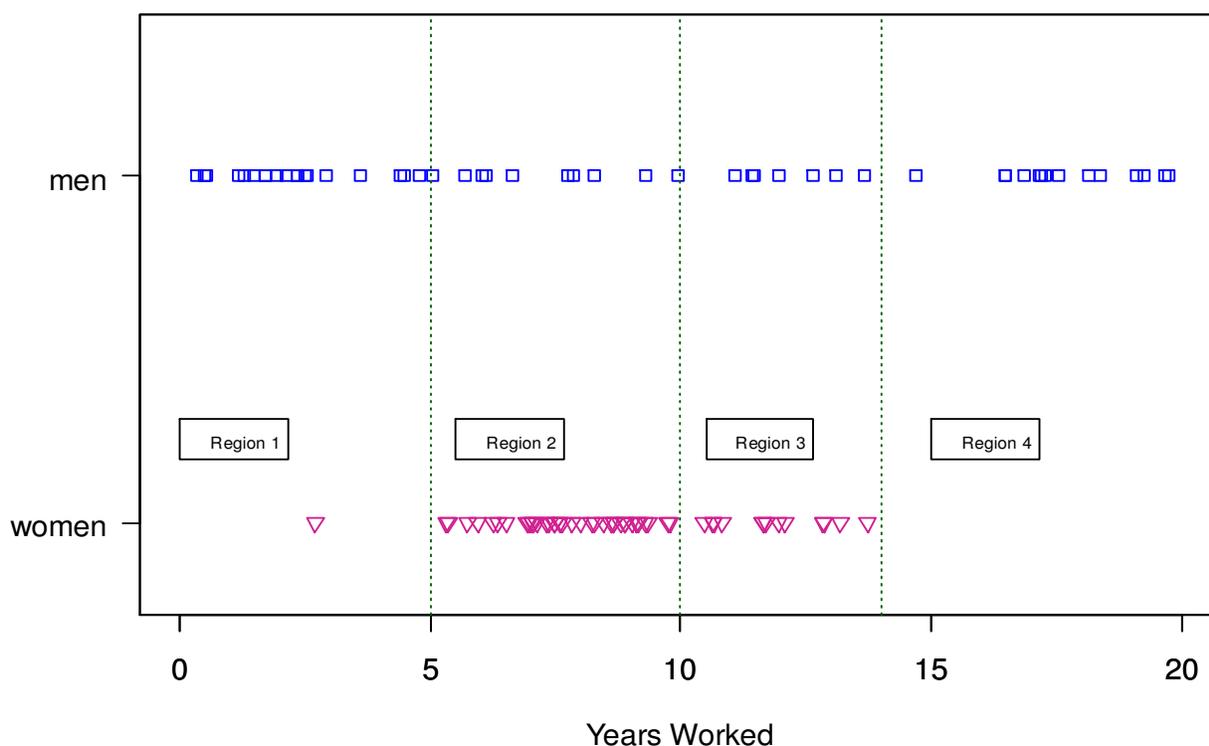


Figure 2: The data from Figure 1 but without the salary variable. The analyst can isolate subsets of the data that can be analogized to a randomized experiment, such as Regions 2 and 3, above. This process can (and should) take place without reference to the outcome (salary) values.

Having identified these subsets of data, one can proceed in a variety of ways; I mention two here. One method is to fit regression lines in the subsets of the data.⁷⁴ Figure 3 shows four such lines, one treatment F and one treatment M line for the data in each of Regions 2 and 3. Notice that the two male (solid) lines remain above their female (dashed) counterparts in both regions, suggesting that being perceived female is causing a lower salary. These four regression lines here are worlds apart from the two regression lines in Figure 1 in critical ways. The analyst here is *not* assuming a constant additive treatment effect even within Region 2 or Region 3, in contrast to the assumption discussed in section I.B.3.b; visually, this is clear from the fact that the lines in Figure 3 are not parallel. The separate, non-parallel lines for units receiving treatments M and F allow for more complex patterns in the data. Moreover, the Figure 3 regression lines are fit over smaller ranges of the data than their Figure 1 counterparts. The technical assumptions of regression⁷⁵ (some of which I have not discussed in this paper) are

⁷⁴ This use of regression lines demonstrates that the potential outcomes framework does not suggest wholesale abandonment of the regression technique but rather judicious use of it in a way that does not purport to interpret the coefficients causally. See Paul W. Holland, *The Causal Interpretation of Regression Coefficients*, in *STOCHASTIC CAUSALITY* (Maria Carla Galavotti et al. eds. 2001).

⁷⁵ See, e.g., Yulia Gel et al., *The Importance of Checking the Assumptions Underlying Statistical Analysis: Graphical Methods for Assessing Normality*, 45 *JURIMETRICS* 3 (2005).

more plausible over the smaller ranges of data represented by the Figure 3 regions than over the whole of data, as in Figure 1.⁷⁶



Figure 3: Blowup of Figure 1 Data Concentrating on Regions 2 and 3. The data are a subset of those pictured in Figure 1. This graph focuses on areas of the data in which an analyst may reasonably attempt causal inference (Regions 2 and 3) and ignores those (Regions 1 and 4) where inference should not be attempted. The solid lines represent the best-fitting lines in each region for men. The dashed lines are the best-fitting in each region for women. Note that here, in contrast to Figure 1, the female lines are below the male lines for almost the entire graph, conforming to the intuition that women (not men) suffer from salary discrimination in this dataset. The figure also shows how to use the lines to predict the salary a woman would have received had she been male. Specifically, a dotted-and-dashed line rises up from the triangle at Years Worked = 10.5, then over to Salary = \$99,700. Thus, the prediction for the woman in Region 3 with the lowest seniority, the left-most triangle in that region, is that she would have earned \$99,700 had she been a man.

Nor does the fact of separate regression lines for M and F units pose a problem for inferences in the potential outcomes framework, as it did with ordinary regression, as discussed in section I.B.4. The potential outcomes framework makes clear that the goal of the analysis is not to estimate regression coefficients, but rather to fill in missing potential outcome values. To see how an analyst might use the separate regression lines to do so, consider the unit assigned treatment F that has the lowest value for years worked in Region 3, *i.e.*, the left-most triangle in Region 3 of Figure 3, with about 10.5 or so years worked. An analyst might fill in the missing potential outcome for that unit, the salary that would have been observed had that unit received treatment M, by putting her finger on the corresponding triangle and traveling straight up until

⁷⁶ The desire to avoid using a single model (or regression line) over a large range of the data is the reason why Regions 2 and 3 are treated separately instead of combined into a single region.

the finger intersects the male line. Then, the analyst would move her finger straight to the left until it intersects with the y (salary) axis. The corresponding number on the y-axis constitutes a predicted value for the counterfactual salary outcome for this unit. Figure 3 shows, with a dotted-and-dashed line, this tracing process. For the unit under consideration, the model predicts that had she been perceived male, she would have earned approximately \$99,700, as opposed to the approximately \$93,000 she did earn. The analyst repeats the process for each unit until the potential outcomes table is full, i.e., all female units (in Regions 2 and 3) have an imputed value in the “Salary(M)” column and all male units (in Regions 2 and 3) have an imputed value in the “Salary(F)” column.⁷⁷

Another way of filling in the missing values of the potential outcomes table is a species of a process called “matching.” To explain this process, I abandon Figure 2 (and its Regions) and reproduce Table 3, above, with an additional column to show each unit’s number of years worked. I also pretend for a moment that only these eight observations exist.

**Table 5: Table 3, Reprinted for Convenience with Colors
To Emphasize the Matching of Units**

Unit #	Years Worked	Treatment	Salary(M)	Salary(F)
1	6.2	F	??	91,200
2	9.0	F	??	92,400
3	13.2	F	??	94,100
4	.5	M	47,100	??
5	9.3	M	98,900	??
6	5.7	M	97,700	??
7	12.6	M	95,300	??
8	19.1	M	81,300	??

Focus for the moment on Unit 1. To proceed with a causal inference, an analyst needs a value for what this unit’s salary would have been had it received treatment M. Of the five units in the table that actually did receive treatment M, the one that looks the most like Unit 1 is Unit 6, because among those five units, Unit 6’s 5.7 years worked is the most similar to Unit 1’s 6.2. Thus, it makes sense for Unit 6 to donate its observed salary value to be Unit 1’s missing, counterfactual value. Units 1 and 6 are a reasonably close “match.” One can proceed to find

⁷⁷ In more technical and precise language, one estimates the regression coefficients associated with the M units, then combines these estimates with the F covariate values with to predict counterfactual potential outcome values for the F units. To predict the counterfactual values for the units that received treatment M, reverse the process. There are several ways to generate uncertainty estimates. The easiest and most intuitive is to specify a prior on the regression coefficients and use Bayesian simulation techniques. *See, e.g., GELMAN ET AL., supra note 53, at 353-85.* In other words, one fills in missing potential outcome values over and over again with a certain amount of random noise included each time. That generates a distribution of the quantity of interest.

The discussion above is intended only to illustrate the two-regression technique (as applied within the potential outcomes framework) conceptually. An analyst actually applying the technique would need to account for at least two sources of uncertainty, the uncertainty in the regression coefficients (by drawing from their posterior) and the uncertainty in the counterfactual salary (by drawing from its posterior conditional on a draw of the regression coefficients). Thus, the predicted counterfactual salaries would never be exactly where the finger-tracing technique above would put them, and in any event the regression lines used to do the tracing would change from simulation iteration to iteration.

reasonable matches for Units 2, 3, 6, and 7 (the pairings 2 & 5, 3 & 7 seem intuitive). Units 4 and 8 present a problem; both received treatment M, and both have values of years worked that are fairly far removed from the values represented in any unit that received treatment F. Under a potential outcomes framework, causal inference is probably not possible with respect to these units. The best procedure is to ignore them, which is the matching equivalent of ignoring the data in Regions 1 and 4 of Figure 2, above. Overall, the intuition with matching is that one has created a mini-randomized experiment in each pair of units, with the pairs identified according to their values on some other important variable.⁷⁸

The result of this matching process is conceptually the same as the result of the regression lines approach, namely, a completed potential outcomes table for all units as to which an analogy to a randomized experiment makes sense. For the subset of units represented in Tables 3 and 5, the result might resemble Table 6.

Table 6: Potential Outcomes Table, Fully Imputed

Unit #	Years Worked	Treatment	Salary(M)	Salary(F)
1	6.2	F	97,700	91,200
2	9.0	F	98,900	92,400
3	13.2	F	95,300	94,100
5	9.3	M	98,900	92,400
6	5.7	M	97,700	91,200
7	12.6	M	95,300	94,100

With the potential outcomes table fully imputed, any quantity of interest can be calculated. One obvious candidate for calculation is the average causal effect for all units.⁷⁹ In a salary discrimination lawsuit, however, the real target of inference is the members of the plaintiff class, so a more interesting quantity might be the average causal effect for women, *i.e.*, for those units who received treatment F. From Table 6, that quantity would be $1/3 * ((91,200 - 97,700) + (92,400 - 98,900) + (94,100 - 95,300)) = -4733$. Thus, in the salary discrimination running

⁷⁸ Years ago, employment discrimination expert witnesses debated the merits of a primitive form of what I label “matching.” These experts called the technique “cohort analysis.” *Compare, e.g.*, Carl C. Hoffman & Dana Quade, *Regression and Discrimination: A Case of Lack of Fit*, 11 SOC. METHODS & RES. 407 (1983), *with, e.g.*, Stephan Michelson, *Comment, in Statistics and the Law*, STATISTICS AND THE LAW, *supra* note 5, 169, 177-78. The Hoffman & Quade article is extraordinary in its lucid discussion of the problem of finding similar units for appropriate comparison, its concomitant emphasis on balancing covariate distributions, and its articulation of the failings of regression, particularly the issue of extrapolating across wide ranges of the covariate space. So far as I am able to tell, the article was ignored, which is a pity.

Other commentators appear to be feeling their way toward something like matching, although they appear to appreciate fully neither the issues at stake nor the power of the technique in drawing inferences of causation. *See, e.g.*, CONNOLLY ET AL., *supra* note 11, § 11.10[1], 11-25. As demonstrated by the cases Michelson, *supra*, collects, courts have thus far largely rejected cohort analysis because (supposedly) (i) it was unsupported in the literature, and (ii) it had low power to detect discrimination. In my view, the former problem stemmed from the fact that the technique was never linked to an explicit definition of and framework for causal inference, while the latter resulted from difficulties encountered when experts attempted to achieve exact or nearly exact matches in the presence of multiple covariates and limited data. Multiple covariates and limited data resulted in few pairs being deemed permissible matches, and thus low power to detect discrimination. As the present article demonstrates, both problems can now be addressed. The explicit causal paradigm is, of course, the potential outcomes framework, while the problem of achieving balance on multiple covariates is discussed in section II.F.

⁷⁹ *See supra* note 65 and accompanying text.

example, after an analysis of the toy data from Table 3, an expert witness could testify as follows: For members of the plaintiff class as to which a causal conclusion could be made, the fact that the defendant firm perceived them to be female caused it to reduce their salaries by approximately \$4700 on average. Notice, moreover, that the expert (with a larger dataset) could also make calculations specific to subsets of the plaintiff class, such as those at the low end or the high end of years worked, by isolating the relevant units in the potential outcomes table and doing similar arithmetic. Such flexibility was not available from the traditional regression model.⁸⁰

The two methods briefly outlined above, separate regression lines and matching, should provide roughly similar completed potential outcomes tables, and thus roughly similar substantive results. The reason is that the analyst has done the hard work of isolating subsets of the data to treat as mini-randomized experiments first, and with data that came from a randomized experiment, the statistical technique used during analysis matters less.⁸¹ If the results these methods differ substantially, the analyst should be wary; it would appear that even after attempting to make the data look as though they came from a randomized experiment, the results are still sensitive to modeling assumptions.

F. Lots of Covariates

The previous sections explained that obtaining balance in covariates is a key part of causal inference, and that randomized experiments balance regardless of whether the covariates are measured by the experimenter or are unobserved. In observational studies, however, the analyst must balance by isolating subsets of data with similar covariate values. These facts lead to an important difficulty with observational studies, which in turn leads to a critical (often the most critical) assumption. The basic problem is that analysts cannot balance variables that are not observed. To draw causal inferences from an observational study, an analyst must assume that he or she has measured all of the really “important” covariates, *i.e.*, the covariates that affect both the probability of receiving a particular treatment and the values of the potential outcomes.⁸²

In the salary employment discrimination example, for instance, an analyst never has direct measurements of variables such as “motivation” or “competence” (assuming such concepts to be well-defined). Their absence could cause difficulties if, for example, (i) these variables are distributed differently among persons perceived male versus persons perceived female, and (ii) the defendant firm bases its decisions on something related to these variables that the analyst does not observe to set salaries. In a courtroom setting, these difficulties should not be overstated; defendants will invariably claim that missing variables explain any observed disparity, and judges who uncritically accept such explanations render discrimination law

⁸⁰ See *supra* section I.B.3.b. Again, I am deliberately suppressing certain technical details in this discussion, such as how to generate estimates of uncertainty. Note that active research is ongoing on this specific subject. Alberto Abadie & Guido W. Imbens, *On the Failure of the Bootstrap for Matching Estimators* (May, 2006), available at <http://elsa.berkeley.edu/~imbens/wp.shtml> (last visited September 6, 2006).

⁸¹ See, e.g., Robert J. LaLonde, *Evaluating the Econometric Evaluations of Training Programs with Experimental Data*, 76 AM. ECON. REV. 604 (1986).

⁸² This assumption is also present in regression. In fact, the role of this particular assumption in an observational study is not new or unique to the potential outcomes framework. The problem goes by various names in various fields; “selection bias” and “presence of confounders” are two that are popular in econometrics. One may also analogize this issue to the problem of “omitted variable bias,” although this term is more commonly used in conjunction with a particular model (such as regression), while the concept here is more general.

unenforceable, at least via the class action device. Nevertheless, the fact that analysts cannot balance what they do not see suggests that they should attempt to see as many of the variables the defendant firm used (or claims to have used) in its decision making.⁸³ If measurements of the variables actually used are unavailable, the analyst might resort to related covariates. The hope is that by achieving balance in the related covariates, the analyst achieves balance on the important variables.

This discussion leads to a final principle that must be explained before one can apply the potential outcomes framework to civil rights datasets: how to balance multiple covariates simultaneously. To understand the problem, consider the matching technique outlined in the previous section, and recall the discussion of how in Table 5 the values of the years-worked variable suggest matching Unit 1 to Unit 6. Imagine now that a second variable, such as years of education, was measured, and that with respect to the years-of-education variable, Unit 1's closest potential match was Unit 5. To which of the two should Unit 1 be matched, Unit 6 (its closest fit on years worked) or Unit 5 (its closest fit on years of education)? Moving away from the matching technique specifically, how does one achieve balance on several covariates at once?

The statistical literature growing from the potential outcomes framework has focused on this problem for decades, and a variety of answers now exist.⁸⁴ One popular option is to estimate something called a "propensity score," which is the probability that a particular unit receives a particular treatment. The idea is to use the observed covariates, together with a statistical model or technique, to estimate a probability for each unit that it received one of the two treatments. The idea, again, is to recreate a randomized experiment by imagining that treatment application was in fact random, but that the analyst "lost" the probabilities used in the random assignment. Statisticians have proved mathematically that if lost probabilities can be recovered with reasonable accuracy, and if certain technical assumptions are met, then effective balancing of the propensity score will induce effective balancing on the observed covariates. Note that an analyst can use a variety of statistical methods to check whether such balance has in fact been achieved, and if it has not, the analyst can modify the original propensity score model. Again, the process is somewhat technical, but the point is that by estimating propensity scores, analysts can reduce the problem of matching on many covariates down to a problem of matching on one variable. In this sense, the propensity score acts as a one-dimensional summary of the covariates.⁸⁵ Nothing in this process of balancing multiple covariates, the hard work of causal inference in an observational study, requires the use of the outcome variable (*e.g.*, salary in the running gender discrimination hypothetical).

III. The Civil Rights Litigation Setting

⁸³ Quantitative techniques are available to assess the sensitivity of causal inferences to unmeasured covariates. *See, e.g.*, Paul R. Rosenbaum and Donald B. Rubin, *Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome*, 45 J. ROYAL STAT. SOC., SERIES B 212 (1983).

⁸⁴ In addition to the propensity-score-based methods, other algorithms are discussed in, for instance, Alexis Diamond & Jasjeet S. Sekhon, *Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies* (2005), available at <http://sekhon.berkeley.edu/> (last visited September 6, 2006).

⁸⁵ The literature about, or using, propensity score methods is extensive. The articles collected in *MATCHED SAMPLING FOR CAUSAL EFFECTS* (Donald B. Rubin ed. 2006) are a good place to start.

In this section, I demonstrate how the potential outcomes definition of and framework for causation may be applied in the civil rights context. In doing so, I show how the framework addresses problems discussed in section I.B, *supra*, that regression could not.

A. In General

1. Identifying Primitives

In the civil rights context, as in any other area of causal inference, an analyst's first task is to identify units, the treatment, the timing of treatment assignment, and the outcome variable. In most instances, the units are individual people or cases: criminal defendants or victims in the capital punishment context, employees or applicants in employment discrimination. The outcome of interest is something assigned by an institutional or social actor, such as the imposition of capital punishment by juries, or salary received from a defendant employer. The treatment is ordinarily the institutional or socioeconomic actor's perception of a prohibited personal characteristic (gender, race, national origin), "prohibited" in the sense that the law disallows the actor from basing the value of the outcome variable on the specified characteristic.

Why identify the treatment as perceived gender (or race), as opposed to self-identified gender (or race), or actual gender (or race), assuming that we could all agree on what we mean by actual gender (or race)?⁸⁶ There are a variety of reasons, many of them technical.⁸⁷ For the purposes of this discussion, one is sufficient: time. As explained previously, one must specify the time at which treatment was applied. Furthermore, and as noted above, in the civil rights context we typically wish to assess whether some actor's decisions as to a group of individuals was affected by some characteristic common to the group. The law's focus on the specific actor's decision making requires, indeed compels, the analyst to regard as "given" the characteristics of individuals that were in place prior to the individuals' interaction with the actor. The only way to do that is to define the treatment as taking place at some moment of perception by the actor of the characteristic common to the group.

A hypothetical example may clarify matters. In an employment discrimination disparate treatment case focusing on hiring, a class of African-American applicants claims intentional discrimination, and the plaintiffs demonstrate that a lower percentage of black applicants than white applicants was hired. In response, the employer proves that it hired based on education, and that members of the African-American class had lower educational achievement levels. At rebuttal, plaintiffs demonstrate that their concededly lower education achievement levels were the result of societal and governmental race discrimination, including (to make this hypothetical perfectly clear) a system of *de jure* segregated schools. What result? Assuming no doubt as to these proofs and no other relevant circumstances, the result is a ruling for the employer. In legal terms, Title VII compels the defendant firm to avoid its own discrimination, not to take remedial action based on someone else's. In statistical terms, the employer is entitled to condition on the

⁸⁶ In fact, in the racial context in particular, we might have difficulty agreeing on an "objective" definition. Thinkers in a wide variety of fields believe that race is a social construct, not a biological set of characteristics. See, e.g., Holland, *supra* note 27, at 3; American Anthropological Association, Statement on "Race," (May 17, 1998), available at <http://www.aaanet.org/stmts/racepp.htm> (last visited June 18, 2006); American Association of Physical Anthropologists, *Statement on Biological Aspects of Race*, 101 AM. J. PHYSICAL ANTHROPOLOGY 569 (1996), available at <http://www.physanth.org/positions/race.html> (last visited June 18, 2006); IAN F. HANEY LOPEZ, WHITE BY LAW: THE LEGAL CONSTRUCTION OF RACE xiii (1996); see also *Saint Francis College v. Al-Khazraji*, 481 U.S. 604, 610 & n.4 (1987). But see 18 U.S.C. §§ 1093(2), 1093(6).

⁸⁷ Don Rubin and I examine these more technical concerns in a forthcoming paper, *Potential Outcomes and Causal Effects of Immutable Characteristics* (2007) (draft on file with author).

educational achievements of all applicants, to take them as it finds them without asking what led up to them. And in terms of the potential outcomes framework, the law compels analysts to treat educational achievement as a covariate, something unaffected by treatment. Thus, for timing purposes, what matters is not actual race or actual gender (which is usually conceptualized as being set before birth), even if we could all agree on what those terms mean. Rather, what matters is the state of the world as of some moment of the social or governmental actor's perception.

This discussion demonstrates that an analyst should ordinarily consider treatment assigned as of the moment in which the social or governmental actor first perceives the prohibited characteristic of interest. Ordinarily, the actor's perception of a unit's gender or race will not change over time, that is, the treatment remains stable as of the moment of first interaction. Moreover, it is usually unsafe to assume that if the socioeconomic or governmental actor does discriminate, it delays or limits its discrimination in some way. For example, in the salary discrimination running hypothetical, if an analyst is assessing whether the defendant firm awards employees perceived female lower salaries than employees perceived male, does it make sense for the analyst to assume that the firm evaluates performance without regard to gender?⁸⁸ Such evaluations ordinarily include a heavy dose of subjectivity and are determined after the evaluator perceives employee gender. Thus, the fact that this kind of variable may be affected by the prohibited characteristic under study suggests that it may be an intermediate outcome as defined in the previous section, *i.e.*, a variable whose value is affected by treatment.

The issue of sharply identifying the timing of treatment assignment dovetails with another of critical importance, namely, precisely identifying the institutional or socioeconomic actor whose behavior is to be studied. Much depends on the analyst's exercise of judgment here. For example, in the capital punishment context, should the analyst assess the effect of (say, the victim's) perceived race on the criminal justice system as a whole, on the prosecutor's charging decisions, on jury behavior, or on the decisions of appellate courts? If an analyst attempts to assess the effect of race on the whole system, can he or she safely treat any characteristics of the case as unaffected by the treatment? In less technical terms, would it be safe to assume in such a study that police investigate homicides of African-Americans or Hispanics in the same way as homicides of whites? If, as I believe, such assumptions should be viewed with skepticism,⁸⁹ then variables such as the heinousness of an offense or even the number of victims in a particular criminal "occurrence" may be intermediate outcomes. If the analyst attempts to study the effect of the victim's perceived race on the whole system, there may be few variables that could safely be assumed to be unaffected by treatment.

In this regard, the use of years worked as a balancing variable in this paper's salary discrimination running example is deliberately provocative. Buried inside a variable such as years worked are a host of decisions that might well be characterized as intermediate outcomes. Some employees may have been fired, perhaps discriminatorily, or may have chosen to leave rather than face the unpleasant task of fighting a Title VII lawsuit. Certainly, if the analyst chooses the firm (as opposed to one of its employees) as the institutional actor to study, and

⁸⁸ It is possible that a firm does limit its discrimination to certain areas of activity. Irrational prejudices are, alas, irrational, and they may operate only in certain socioeconomic spaces or at certain times. My point here is that it will ordinarily be unwise to assume that such discrimination is not taking place.

⁸⁹ Samuel R. Gross & Robert Mauro, *Patterns of Death: An Analysis of Racial Disparities in Capital Sentencing and Homicide Victimization*, 37 STAN. L. REV. 27, 44-47 (1984) (discussing empirical evidence that "official descriptions of homicides by prosecutors were affected by racial considerations") (emphasis in original).

conceptualizes the timing of treatment assignment as occurring at the moment of first perception (meaning, probably, at the time of an employment application), years worked is post-treatment and thus presumptively an intermediate outcome as opposed to a covariate.

The discussion above demonstrates the delicate balance an analyst attempting causal inference in the civil rights context must strike among competing concerns such as the importance and nature of the problem to be analyzed and the plausibility of the assumptions necessary to proceed. Much depends on the choice of question. A fundamental aim of this paper is to persuade (i) that these choices are far more important to the answers produced than are disagreements about models, statistical techniques, or computer software; and (ii) that none of these choices is mathematical or technical. No such decision is beyond the ken of the average layperson. In previous paragraphs, I suggested that it might be difficult, without strong and potentially risky assumptions, to assess the effect of the victim's perceived race on the entirety of the system that administers capital punishment. But difficulty is not the only concern. If an analyst can identify primitives (units, treatment, treatment timing, and outcome of interest) and articulate the assumptions that would be required to proceed, perhaps the risky assumptions are worth making, so long as the analyst states them clearly.

Having identified the units, the treatment, the timing of treatment assignment, and the potential outcomes of interest, the analyst should inquire as to the nature of the data available. This is not the same as examining the data. The analyst should do the hard work of (i) classifying variables into covariates versus intermediate outcomes, and (ii) balancing covariates, all without having access to the outcome measurements. In other words, particularly in a litigation setting, an expert witness should specifically request that she not be provided with the outcome data at first. This way of proceeding decreases the danger identified in section I.B.1 as a principal failing of regression, namely, improperly favoring one side or the other by snooping for a preferred result during the model fitting process, because without access to the outcome data the analyst has less idea what the litigation answer will be as he or she works.⁹⁰

This is a situation in which early intervention and ongoing oversight by the trial judge (or, more likely, a magistrate or a special master) can make settlement more likely. For example, the court in a large-scale Title VII class action can begin by prohibiting transmission of data to any testifying expert until, through discovery and any necessary rulings, available variables have been identified and the process of their generation understood. Next, the court could, to the extent consistent with the role of a jury (if any), adjudicate disputes as to what variables are pre- and post-treatment, and as to the post-treatment variables, which ones are unlikely to have been affected by treatment. Then, the court might allow transmission of the data corresponding to pre-treatment variables, post-treatment variables unaffected by treatment, and the treatment itself to the parties' experts. The experts could implement the balancing process and disclose to each other which subsets of data each deems sufficiently similar to a randomized experiment to allow causal inference to proceed. The point is to require experts to commit, in a way difficult to disavow later, to critical parts of their analyses before knowing what the results will be. Finally, upon being satisfied that both sides have committed in this way, the court could order the release of the outcome data. The hope is that because both sides have done the hard work of balancing covariates before knowing the results, their conclusions will be similar to one another, easing settlement. At a minimum, important stages of the analysis will be less susceptible to biasing by experts, some of whom may be fighting a subconscious battle with themselves.

⁹⁰ There are, concededly, some limited ways for the fallen to cheat, but they are both harder to implement and easier to detect than the model-snooping that is so easy to accomplish when checking the fit of regression equations.

If this procedure cannot be implemented, civil rights litigants might nevertheless wish to ask their testifying expert to proceed via potential outcomes as opposed to regression. As noted above, a trier of fact may (should) give greater weight to an expert who truthfully testifies that she committed to an analysis before knowing its substantive results. In this way, one might hope that the incentive structure of litigation could make the use of potential outcomes self-executing.

As suggested above, classifying variables as covariates or intermediate outcomes, *i.e.*, unaffected by treatment versus affected by treatment, requires the analyst to understand the data-generating process thoroughly. This classification process benefits from a sharp definition of the treatment, particularly its timing. That is, the analyst using potential outcomes has a better idea of which variables to include in balancing. The basic rule is that any variable measured prior to the application of treatment is a covariate that should be included in balancing. Anything measured contemporaneously with treatment application or assessed afterward is suspect and should be treated as a covariate only if the analyst may reasonably conclude that the variable's value is unaffected by treatment. In the salary discrimination running example, years of education is a reasonable candidates for inclusion in balancing; absent other evidence, it might be plausible to assume that this variable's values were recorded in an essentially mechanical process from an employment application or resume. Other variables are more suspect, particularly those with inherent subjectivity, such as performance evaluations. Should a dispute arise as to whether to balance a variable whose values are determined after treatment application, the burden of persuasion should be on the party seeking to categorize such a variable as a covariate instead of as an intermediate outcome.

Thus, within the potential outcomes framework, an analyst finds guidance on the issue of which variables to include, while the balancing process (which, again, should take place without reference to measured outcomes) provides information on which observations should be the focus of analysis. Courts find guidance on how to assign burdens of proof. The potential outcomes paradigm addresses problems that regression (as commonly used in modern civil right litigation) could not.

After achieving acceptable covariate balance, the analyst can use a variety of techniques, such as matching or separate regressions for treated and control groups, to fill in the counterfactual outcome for each unit. Regression coefficients themselves (if regressions are used at all) are of little interest. What matters is the outcomes the model predicts, not the estimated coefficients. Thus, expert witnesses, litigators, and courts need no longer attempt to shoehorn data into models with poor fits but with easily interpretable coefficients, such as the constant additive effect model described in section I.B.3.b. Courts and litigators may focus less on grasping or explaining logarithms of odds ratios, per section I.B.3.d. Instead, they can concentrate on figuring out what would have occurred "but-for" the identified treatment, a task litigators, judges, and juries can understand.

2. A Tug-of-War

In drawing inferences of causation in the civil rights context, one issue arises often enough to deserve special consideration. An example may illustrate. In the employment discrimination context, if an analyst chooses the employer as the institutional actor whose behavior is to be assessed (a reasonable choice given that the employer as a firm, and not an individual officer, is the defendant in the lawsuit), then the analyst should ordinarily conceive of treatment (perceived gender, race, ethnicity, etc.) as assigned at the time of employees' job applications. Many application forms include race and gender boxes, and certainly if any sort of

interview is required for the job, the prohibited characteristic of interest will be apparent at that time. The treatment presumptively occurs quite early in the relationship between the units and the actor of interest. Recall that a critical assumption of an observational study is that the analyst has measured a sufficient number of the important covariates (pretreatment variables) so that, after achieving balance in the covariates, members of one treatment group are not systematically different from members of the other. Recall also that analysts should not balance on variables whose values are affected by treatment. The problem in conceiving of treatment as being assigned at the time of a job application is that a great many variables as to which measurements may be available (*e.g.*, tasks or projects accomplished, types of experience gained, nature of training received, not to mention performance evaluations, disciplinary actions, initial job levels, bonuses, etc.) have values that are determined after treatment. In fact, much of the information in a class member's employment file, including even the fact that he or she was hired, is determined after treatment. Thus, few variables may be presumptively considered covariates. With fewer covariates, the assumption that the analyst has measured a sufficient number of the important ones becomes strained.

This concern suggests conceptualizing treatment as occurring later in the employment relationship, if doing so would be plausible. For example, if a Title VII disparate treatment class action focuses on the class members' failure to achieve a particular kind of promotion, and if promotion decisions were made by an official within the defendant firm who had little prior interaction with employees eligible for promotion, the analyst might identify this official as the socioeconomic actor whose behavior is under study, and thus conceptualize treatment as having occurred when this official perceives the gender (or other prohibited characteristic) of the promotion applicants. That choice would allow more variables, such as performance evaluations, job assignments, etc., to be considered covariates and thus be subject to balancing. But this course of action has costs. One of these costs is that any discrimination perpetrated by the defendant firm before the promotion application stage, such as while evaluating performance, would go undetected. Balancing on the performance evaluation variables would mask the discrimination in the same way that balancing on lung cancer could mask the effect of smoking on death.

The problem described here is a general one that occurs whenever one attempts to draw inferences of causation in a civil rights context in which the interaction between potential plaintiffs and the institutional actor of interest extends over a period of time. A tug-of-war exists between (a) the desire to include in balancing as many variables as possible to make plausible the assumption that one has taken care of the important ones, which counsels considering treatment to have been assigned late in the interaction between the units and the institutional actor of interest; and (b) the desire to understand and model properly the fullness of the relationship between the institutional actor and the units, including by assessing the possibility of discrimination early in that relationship, which counsels conceptualizing treatment to have been assigned as early as possible.

One way to address this tug-of-war is by making assumptions, supported by reasonable judgment, about what variables have been affected by the treatment. Here, the analyst identifies variables whose values are (i) determined after treatment, but (ii) less likely to have been affected by treatment, and (iii) likely to be related (strongly related, one hopes) to the values of important unobserved covariates. The analyst includes these post-treatment variables in the balancing process, the hope being that in doing so, he or she balances important pretreatment covariates while not inducing bias (in a statistical sense) in the results. In the promotions

example two paragraphs up, perhaps investigation into the defendant firm's operations reveals that employees receive assignments based on their current workload and the type of tasks that require completion. Or it may be that the training programs a firm offers are ordinarily open to all employees who wish to undertake them. Or it could be that a part of the process of evaluating employee performance is mechanical, *e.g.*, the number of DVD packages processed per minute.⁹¹ In that case, the analyst might attempt to resolve the tug-of-war articulated above by balancing on variables measuring types of job experience, number and nature of training programs, and the mechanically reported aspects of performance. The analyst should, however, be particularly wary of variables that inherently involve a partially subjective judgment, such as overall job evaluation.

B. Specific Contexts

1. Capital Punishment

Drawing causal inferences about the role of race in the administration of the death penalty poses several challenges, some familiar, some unique to the capital setting. A logical place to begin is with the identification of the treatment. The literature in this area suggests that the treatment should be defined in terms of perceived race of both the victim and the defendant. For example, in a jurisdiction made up almost exclusively of Caucasians and African-Americans, the analyst could divide cases into four treatment categories: white defendant, black victim; white defendant, white victim; black defendant, black victim; black defendant, white victim.⁹² Each unit would then have four potential outcomes, only one of which is observed. Units are cases. Although in a real study an analyst would probably want to have four treatment groups, for the sake of illustrating principles I suppose for the moment that the analyst is interested in race-of-the-victim effects only, so there are two treatments, "B" for victim perceived black and "W" for victim perceived white.

Next, the analyst must choose the institutional actor whose behavior is to be studied. If the analyst follows the presumption that treatment is administered as of the moment the institutional actor first perceives the victim's race, then much depends on this choice. Suppose the analyst identifies the actor as the criminal justice system. From case files, the available variables will be those the police, the prosecutor, and other official sources have uncovered and recorded after they perceive the victim's race, *i.e.*, after treatment application. These post-treatment variables must be assessed carefully to see whether it is reasonable to assume that the treatment does not affect their values. Consider something as simple as whether the murder was for hire. The value the analyst sees for this variable is not whether, in some abstract sense of "truth," the murder was for hire, but rather the value determined by a police investigation. It seems odd, if one is assessing the race effects in the entire criminal justice system, to assume that police investigate homicides of whites with equal vigor as homicides of blacks. Thus, the analyst must think carefully about which variables, if any, may be deemed unaffected by treatment if the institutional actor of interest is the criminal justice system. On the other hand, if

⁹¹ See Susan Sheehan, *Tear, Slap, Clack*, *New Yorker*, Aug. 28, 2006, at 26 (describing a Netflix facility). As Sam Gross has reminded me, discrimination, such as perpetration of a hostile working environment, can affect an employee's performance even as measured by the most neutral of criteria. This is a problem, and it haunts any analysis framework (potential outcomes or otherwise). One would expect that the actions taken to communicate a message of inferiority or stigma to certain employees might leave a non-quantitative evidentiary trail.

⁹² If there are a small number of cases fitting one or more of these treatment categories, the analyst can either remove units in this treatment category from the analysis or collapse it together with another category (thus redefining the treatments). To reiterate a point made several times in this paper, this choice is not mathematical.

the analyst chooses the jury as the actor whose behavior is to be assessed, then the facts as discovered by the police (and the defense investigation) may be taken as given on the principle that any discrimination by the police is not chargeable to the jury. But among other things, any racial bias in police investigations (for example) will go undetected. The tug-of-war identified in section II.A.2 is fully apparent here.

For the sake of illustrating other questions that arise, I suppose for the remainder of this section that the analyst chooses the jury as the actor of interest. The remaining “primitives” are the timing of treatment application and the outcome. The first should be easy to do; as articulated above, the presumption should be that treatment is applied when the jury first perceives the victim’s race. To be safe, an analyst might conceptualize the perception as having occurred at the earliest possible moment, perhaps during jury selection.⁹³ Identifying an outcome to study, can be tricky. After *Furman v. Georgia*,⁹⁴ most capital systems include two phases, guilt and penalty. A great deal of scholarly attention has been focused on the latter, perhaps because the guilt phase of a capital trial is similar to trial of lesser offenses.⁹⁵ The penalty phase of capital trials, in contrast, includes elaborate and unusual procedural safeguards designed, in theory at least, to prevent arbitrary administration of the law’s severest punishment. Thus, assume for the moment that the question of interest is the effect of the perceived race of the victim on the probability that a jury will sentence a defendant convicted of a death-eligible crime to die.

With these sharp definitions of the primitives, much is clarified, including the fact that at least one critical issue remains: the relationship of when the jury perceives the race of the victim to the jury’s decisions on guilt and punishment. The analyst may be interested in the penalty phase alone, but per above, treatment is administered earlier, before the jury renders a verdict on guilt. Is it safe to assume that the treatment had no effect on the jury’s guilt or innocence decision? Probably not. In other words, if we believe it worthwhile to study whether the victim’s perceived race has an effect on a jury’s punishment decision, it seems odd to assume that the victim’s perceived race had no effect on the same jury’s prior verdict of guilt or innocence (of a death-eligible offense).

To clarify, suppose, in a particular case, the treatment assigned is “W,” that is, that the jury perceives the race of the victim to be white; that in the first stage of proceedings, the jury convicts the defendant of capital murder; and that in the second stage of the proceedings, the jury sentences the defendant to die. Does it make any sense to talk about what the jury’s decision in that same case at the second stage of the trial would have been under treatment “B” if, had the jury perceived the victim to be black, it would have acquitted (or found the defendant guilty of only a non-capital offense)?⁹⁶ Here, a focus on a variable occurring after treatment, the phase one verdict, is necessary because unless that variable takes a certain value (conviction of a death-

⁹³ See *Turner v. Murray*, 476 U.S. 28, 37 (1986) (holding that a defendant in a cross-racial capital trial has a constitutional right to inform the venire of the race of the victim and question it about racial prejudices).

⁹⁴ 408 U.S. 238 (1972).

⁹⁵ See *Turner*, 476 U.S. at 34-39 (distinguishing the discretion exercised by a capital jury at the penalty phase from both (i) that exercised in any part of a non-capital case, and (ii) the guilt phase of a criminal trial).

⁹⁶ The fundamental issue here is that the characteristics of one phase of a capital trial might induce outcome-determinative alterations in a jury’s behavior at the other phase. See *Woodson v. North Carolina*, 428 U.S. 280, 302-03 (1976) (plurality opinion of Stewart, J.) (holding North Carolina’s mandatory death penalty statute unconstitutional in part because juries would exercise unbridled discretion in refusing to convict defendants of death-eligible offenses to avoid the mandatory capital punishment) and *Roberts v. Louisiana*, 428 U.S. 325, 334-35 (1976).

eligible crime), the outcome of interest (the choice between a sentence of death or life imprisonment) has no defined value. Recall that under the potential outcomes framework, a causal effect is defined in terms of a comparison of potential outcomes of the variable of interest under alternative forms of treatment. To answer the question of interest, *i.e.*, the effect of the victim’s perceived race on the jury’s penalty phase decision making, an analyst must isolate those units (cases) in which a jury would convict of a death-eligible offense under both treatment B and treatment W. Only then does it make sense to compare the outcomes of the penalty phase under the two treatment assignments.

In table form, the problem can be illustrated as follows:

Table 7: Potential Outcomes Table with a Crucial Intermediate Outcome

Unit #	Treatment (Victim’s Perceived Race)	Conviction if Treatment = Black	Conviction if Treatment = White	Death if Treatment = Black	Death if Treatment = White
1	White	??	No	??	undefined
2	White	??	Yes	??	Yes
3	White	??	Yes	??	Yes
4	Black	No	??	undefined	??
5	Black	Yes	??	No	??
6	Black	Yes	??	No	??

In Table 7, it is clear that Units 1 and 4 do not belong in the study. Neither was convicted of a death-eligible offense, so the analyst knows that because of the potential outcome under one treatment (the one actually received), a comparison of punishment under treatment “B” to punishment under treatment “W” makes no sense. But the analyst cannot stop there. Suppose it were the case that missing ?? values in Table 7 were in truth filled in as per Table 8. These values would never be directly observed, so Table 8 is for illustration.

Table 8: Potential Outcomes Table with a Crucial Intermediate Outcome, All Values Filled In

Unit #	Treatment (Victim’s Perceived Race)	Conviction if Treatment = Black	Conviction if Treatment = White	Death if Treatment = Black	Death if Treatment = White
1	White	<i>No</i>	No	<i>undefined</i>	undefined
2	White	<i>No</i>	Yes	<i>undefined</i>	Yes
3	White	<i>Yes</i>	Yes	<i>No</i>	Yes
4	Black	No	<i>Yes</i>	undefined	<i>Yes</i>
5	Black	Yes	<i>No</i>	No	<i>undefined</i>
6	Black	Yes	<i>Yes</i>	No	<i>Yes</i>

Now, it is clear that Units 2 and 5 also have no place in the study, for the same reason that Units 1 and 4 do not belong. For each of these units, the outcome of interest is undefined under one of the two treatments. The only units for which a causal effect of perceived victim race on penalty is defined are Units 3 and 6. Because the analyst observes Table 7 with its missing values, not the fully-filled-in Table 8, his or her task is to use statistical techniques to fill in the missing values in the two “Conviction if . . .” columns in Table 7, isolate the units for which both “Conviction if . . .” columns have the value “Yes,” fill in the missing values in the two “Death if . . .” columns, then (say) calculate an average death penalty rate using the now-completed

potential outcomes columns. There are various methods the analyst might use to accomplish the first and third steps, and a choice among them might require some familiarity with sophisticated mathematical principles. But nothing in understanding the problem requires anything technical.⁹⁷

2. Employment Discrimination

I have used employment discrimination examples throughout this paper to illustrate basic concepts. Moreover, the number and kind of employment decisions subject to challenge under anti-discrimination laws render complete coverage in this area in a few pages impossible. I dedicate this section to a few general principles, as well as to identification of issues that require further research.

First, a problem similar to the issue identified above in the death penalty context is present in employment discrimination, but here the issue is on a larger scale. Technically, it does not make sense to discuss whether an employee actually perceived male would have received a promotion had that employee been perceived female if, had the employee been perceived female, he/she would not have been hired in the first place. If an analyst assessing a promotion discrimination claim examines only those units currently eligible to receive the promotion, he or she is using a dataset that is contaminated by units as to which a counterfactual comparison makes no perfect theoretical sense. A second issue that stands out as troublesome in the employment discrimination context is the tug-of-war identified above, *i.e.*, the problem of what variables to include in a balancing process, given that gender and race are perceived (*i.e.*, treatment is applied) so early in the employee-defendant relationship. Much here may depend on how the law chooses to characterize events as separate or related. In the capital punishment context, the Supreme Court has recognized a strong connection between the guilt and penalty phases of a capital trial,⁹⁸ triggering the problem illustrated by Tables 7-8, above. Perhaps in the employment discrimination context, hiring and promotion constitute legally separate transactions or occurrences.⁹⁹

This discussion demonstrates that difficult, fundamental issues may arise over and over in categories of employment discrimination cases. Regression as used in modern civil rights litigation does not avoid these issues; it masks them. Thus, expert witnesses, litigators, and courts will be required to make delicate judgments. It will not do to render the employment discrimination laws wholly inapplicable to entire classes of lawsuits because we currently lack a

⁹⁷ The capital punishment problem outlined above is structurally identical to a set of statistical issues in the biomedical context that are referred to as, alas, “censoring due to death” or “truncation due to death.” *See, e.g.*, Junni L. Zhang & Donald B. Rubin, *Estimation of Causal Effects Via Principal Stratification When Some Outcomes Are Truncated by “Death,”* 28 J. EDUC. & BEHAVIORAL STAT. 353 (2003); *see also* Constantine E. Frangakis & Donald B. Rubin, *Principal Stratification in Causal Inference*, 58 BIOMETRICS 21 (2002).

To my knowledge, no study has done what is required to make valid causal inferences about jury decisions at the penalty phase, *i.e.*, remove units from the study as to which a treatment effect is undefined. To the contrary, the issue articulated above has usually either gone unrecognized or been presumed to be amenable to solution via modeling, such as with a properly specified regression. In fact, regression cannot solve this problem because the issue is one of data collection and question identification, not modeling technique. Note that the Baldus study, BALDUS ET AL., *supra* note 6, suffered from this problem, along with the difficulty of including intermediate outcomes in the right hand side of its regression equations. *See* Greiner and Rubin, *supra* note 87.

⁹⁸ *See supra* note 96.

⁹⁹ *See* General Telephone Co. of the Southwest v. Falcon, 457 U.S. 147 (1982) (holding that Fed. R. Civ. P.’s representativeness requirement prevented a plaintiff alleging that he personally suffered race discrimination in promotion from being the named plaintiff for a class alleging race discrimination in hiring).

statistical framework that will satisfy all skeptics.¹⁰⁰ Nor will it do to allow defendants to prevail in any cases in which they point out that the ideal data are not available; the ideal data are never available. Assumptions that might make the quantitative analyst blush¹⁰¹ outside the litigation setting, particularly with respect to the characteristics of persons a firm does not hire (as to which covariate information may not be available or recoverable), must be seriously assessed and explored. Obviously, a prerequisite to such a process of exploration and assessment is that the assumptions be stated clearly. Moreover, experts, litigators, and courts need not be satisfied in all instances with the information available in a defendant's computer files. In some cases, paper records can be searched and coded; in others, surveys may be administered to reconstruct lost or previously unavailable information.

The key point here is that in most employment discrimination contexts, we can identify the units, the treatment, a timing of treatment application, the outcome of interest, and the variables that might or might not be subject to balancing. We can assess critical assumptions, such as noninterference among units. Some uncomfortable, additional assumptions may be necessary to proceed in actual cases, but at least we can ask coherent questions linked to available data. As the next section demonstrates, such is not always the case in civil rights litigation.

3. Causation and Section 2 of the Voting Rights Act

Since *Thornburg v. Gingles*,¹⁰² it has been clear that Section 2 of the Voting Rights Act prohibits dilution of minority voting strength, and that an essential part of a plaintiff's case in a Section 2 lawsuit is proof of racial bloc voting,¹⁰³ which until recently was attempted primarily

¹⁰⁰ Consider the following passage:

Scientifically, the strength of the case against smoking rests not so much on the p-values, but more on the size of the effect, on its coherence and on extensive replication both with the original research design and with many other designs. . . . The results of the studies on smoking are generally coherent in the following ways: (i) There is a dose-response relationship: persons who smoke more heavily have greater risks of disease than those who smoke less. (ii) The risk from smoking increases with the duration of exposure. (iii) Among those who quit smoking, excess risk decreases with time after exposure stopped.

Freedman, *supra* note 88, at 253. Requiring this level of proof before reaching a conclusion a respectable viewpoint for a scientist, one who is interested in investigating and establishing universal truths. It will not do for a court, which must either (i) work towards finding the best possible answer in a particular dispute under a tight (by social science standards) timeline, or (ii) always rule against the party with the burden of proof. See Joseph L. Gastwirth, *Some Issues Arising in the Presentation of Statistical Testimony*, 4 L. PROBABILITY & RISK 5, 5-7 (2005); Steven L. Willborn, *A Lawyer's View of the Statistical Expert*, 4 L. PROBABILITY & RISK 25, 26 (2005).

¹⁰¹ "In countless areas of the law weighty legal conclusions frequently rest on methodologies that would make scientists blush. The use of such blunt instruments in examining complex phenomena and corresponding reliance on inference owes not so much to a lack of technical sophistication among judges, although this is often true, but to an awareness that greater certitude frequently may be purchased only at the expense of other values." *LULAC v. Clements*, 999 F.2d 831, 860 (5th Cir. 1993) (en banc) (Higginbotham, J.).

¹⁰² 478 U.S. 30 (1986); see also *Holder v. Hall*, 512 U.S. 874, 885 (1994) (O'Connor, J., concurring in the judgment); *id.* at 961-963 (Stevens, J., dissenting).

¹⁰³ I assume knowledge here of basic Section 2 law, including the three *Gingles* prerequisites, the totality-of-circumstances test, and the so-called "Senate Report" factors. See 42 U.S.C. § 1973(b); S.Rep. No. 97-417, 97th Cong., 2d Sess., at 28-29 (1982), Pub. L. No. 97-205, § 3, 96 Stat. 131, 134; *Clark v. Calhoun County*, 21 F.3d 92, 97 (5th Cir. 1994); *Colleton County Council v. McConnell*, 201 F. Supp. 2d 618, 632 n.13 (D.S.C. 2002), *clarified*, 201 F. Supp. 2d 618 (D.S.C. 2002).

via regression.¹⁰⁴ Among the innumerable unresolved issues in this area is whether a plaintiff's prima facie case includes a causal element of some kind, either as part of the proof of racial bloc voting or later at a "totality of circumstances stage."¹⁰⁵ Courts adjudicating this issue describe the causal element in various ways. Examples include (i) a focus on "the reasons why white voters reject[] black candidates;"¹⁰⁶ (ii) "an inquiry into the cause for the correlation" between voter race and voter preferences "to determine, for example, whether [the correlation] might be the product of similar socioeconomic interests rather than some other factor related to race;"¹⁰⁷ (iii) a need to eliminate party affiliation or party preferences as a cause of voting patterns;¹⁰⁸ and (iv) whether any vote dilution observed is "being caused by the interaction of racial bias in the voting community and the challenged scheme."¹⁰⁹

The judiciary has spilled a fair amount of ink on this subject, with the prevailing view being that Section 2 does contemplate some sort of causal inquiry.¹¹⁰ Much of the debate has

¹⁰⁴ See Greiner, *supra* note 11, appdx. A; see also *LULAC v. Perry*, 548 U.S. 399, ___ (2006) (Roberts, C.J., concurring in part and dissenting in part) ("At trials, assumptions and assertions give way to facts. In voting rights cases, that is typically done through regression analyses of past voting records.").

¹⁰⁵ See *supra* note 103.

¹⁰⁶ *Gingles*, 478 U.S. at 100 (O'Connor, J., concurring).

¹⁰⁷ *Holder v. Hall*, 512 U.S. 874, 904 (1994) (Thomas, J., concurring). Notice how quickly doctrinal confusion in this area emerges. Usually, for socioeconomic circumstances to provide an alternative explanation for voting patterns that are related with race, these circumstances would need to be themselves related to race. But the relation of socioeconomic circumstances and race is a Senate Report factor tending to support a finding of vote dilution, at least when (as is usually the case in this country) minority race is related to less advantageous socioeconomic conditions.

¹⁰⁸ See, e.g., *LULAC v. Clements*, 999 F.3d 831, 859 (1993) (en banc); *Baird v. Consolidated City of Indianapolis*, 976 F.2d 357, 361 (7th Cir. 1992).

¹⁰⁹ *Nipper v. Smith*, 39 F.3d 1494, 1497 (11th Cir. 1994) (en banc).

¹¹⁰ The debate has concerned at least three broad issues. First, there is the threshold question of whether Section 2 vote dilution plaintiffs must prove any sort of causation. This issue was one (among several) about which the Eleventh Circuit, sitting en banc, divided equally in *Solomon v. Liberty Count*, 899 F.2d 1012 (11th Cir. 1990), and later resolved for circuit purposes in *Nipper v. Smith*, 39 F.3d 1494 (11th Cir. 1994) (en banc). The en banc Fifth Circuit also split, although a majority supported the view that Section 2 requires a causal proof. *LULAC v. Clements*, 999 F.2d 831, 856 (5th Cir. 1993) (en banc) (majority); *id.* at 900, 901-12 (King, J., dissenting). On the question of whether Section 2 includes a causal element, I note in passing that the Supreme Court recently found a Section 2 violation without mentioning causation. *LULAC v. Perry*, 548 U.S. ___ (2006).

Second, courts have disagreed on whether the causal element of proof, whatever it means, is (i) part of the third *Gingles* prerequisite (white bloc voting usually sufficient to defeat the minority's preferred candidate), or (ii) part of the totality of circumstances, or (iii) synonymous with a Section 2 vote dilution claim, when joined with a potentially dilutive electoral structure such as an at-large districting scheme or a gerrymander. The Fifth Circuit appears to favor the first choice. *LULAC v. Clements*, 999 F.2d at 856-58. The First, Second, Fourth, and Seventh Circuits appear to prefer the second. *Vecinos de Barrio Uno v. City of Holyoke*, 72 F.3d 973, 983 n.4 (1st Cir. 1995); *Goosby v. Town Board of the Town of Hempstead*, 180 F.3d 476, 493 (2^d Cir. 1999); *United States v. Charleston County*, 365 F.3d 341, 347-48 (4th Cir. 2004); *NAACP v. Thompson*, 116 F.3d 1194, 1199-1200 (7th Cir. 1997). Note that the NAACP opinion is not perfectly clear on this point. Finally, the Eleventh Circuit appears to espouse the third choice. *Nipper*, 39 F.3d at 1514-15. Unfortunately, this matter is not one of mere doctrinal nicety, because most courts give at least lip service to the proposition that "[i]t will be only the very unusual case in which the plaintiffs can establish the existence of the three *Gingles* factors but still have failed to establish a violation of § 2 under the totality of circumstances." *Clark v. Calhoun County*, 21 F.3d 92, 97 (5th Cir. 1994) (quoting *Jenkins v. Red Clay Consol. Sch. Dist. Bd. of Educ.*, 4 F.3d 1103, 1135 (3^d Cir. 1993)); see also *NAACP v. City of Niagara Falls*, 65 F.3d 1002, 1019 n.21 (2^d Cir. 1995).

Third, the courts have divided on which party has the burden of proof on causation. The Fifth Circuit places the burden on the plaintiffs, where it remains at all times. *LULAC v. Clements*, 999 F.2d at 856-58. Other courts, while purporting to eschew formal burden-shifting, have stated that if plaintiffs prove the three *Gingles*

been doctrinal and is thus beyond the scope of this paper, but one can learn important things about the nature of causal inference by examining how courts actually apply the concept of causation in the Section 2 context. That examination should begin with two preliminary observations. First the causal inquiry, however defined, appears to matter little in actual cases unless the factual record demonstrates that candidates of minority race have enjoyed some measure of electoral success.¹¹¹ Second, the only additional evidence occasionally brought to the table to address the causation issue, apart from that which addresses the laundry list of factors courts would otherwise consult at the Section 2 totality of circumstances stage, is whether white voter support for a particular party's candidate is generally lower if he or she is the candidate of choice of minority voters than if he or she is not so supported.¹¹² These two facts suggest that, to the extent courts really do focus on causation when adjudicating Section 2 cases (as opposed to when articulating Section 2 doctrine), they are attempting to draw an inference about the effect of the race of the candidate, or (worse yet) the race of voters, on election results.

A potential outcomes understanding of causation pays dividends in this context by demonstrating that neither inquiry can be coherently linked to available data. Consider an attempt to assess the effect of the race of the candidate on election results. One can identify the treatment as (perceived) candidate race and, at least initially, the outcome of interest as election

prerequisites and the defendant is mute on the issue, plaintiffs are deemed to have prevailed on this question. *Nipper*, 39 F.3d at 1524 & nn. 61, 64; *Vecinos*, 72 F.3d at 983 n.4. Other judges would adopt formal burden-shifting. *Goosby*, 180 F.3d at 503 (Leval, J., concurring.).

¹¹¹ At least, the above statement is true after the 1982 amendments to Section 2. *Cf. Whitcomb v. Chavis*, 403 U.S. 124 (1971).

As I did not attempt a comprehensive examination of all Section 2 cases over a particular time period, my support for the statement above comes from a review of several circuit cases grappling with the causation issue, as well as some subsequent rulings. I offer in particular the following three illustrations. First in *LULAC v. Clements*, the case in which the en banc Fifth Circuit first adopted a causation element in the Section 2 context, the court relied exclusively on the success of Republican candidates of African-American race to hold that party, not race, best "explained" observed voting patterns. Later, however, in *Clark v. Calhoun County*, the Fifth Circuit reversed a district court holding in favor of a Section 2 defendant without mentioning causation, although the court did examine success of candidates of minority race, as 42 U.S.C. § 1973(b) compelled it to do. 88 F.3d 1393, 1400 (5th Cir. 1996). Both opinions were authored by Judge Higginbotham. *See also Jones v. City of Lubbock*, 730 F.2d 233, 235 (5th Cir. 1984) (Higginbotham, J., concurring in denial of petition for rehearing en banc). Second, in *Nipper v. Smith*, the Eleventh Circuit spent more than 15 pages justifying its pronouncements that Section 2 included a causal element. 39 F.3d 1494, 1509-27 (11th Cir. 1994). In strongly implying that the requisite causation had been proved, however, the en banc court simply looked to the Gingles factors, weighed the totality of circumstances, and examined special circumstances of the case (there, the unusual considerations attendant to a vote dilution claim when the offices at issue are judicial). 39 F.3d at 1537-43. All of these steps are ones the court should have taken in a Section 2 discussion lacking any focus on causation. Third, after the First Circuit held that Section 2 includes some causal element in *Vecinos de Barrio Uno v. City of Holyoke*, 72 F.3d 973 (1st Cir. 1995), a three-judge court sitting in Massachusetts found a Section 2 violation in a districting scheme for the state legislature. *Black Political Task Force v. Galvin*, 300 F. Supp. 2d 291 (D. Mass. 2004). The three-judge court's entire discussion of causation consisted of the following phrase: "We have also inquired into causation where appropriate . . ." 300 F. Supp. 2d at 300. Judge Selya authored both *Vecinos* and *Galvin*.

I note again that identifying the important role the success of candidates of minority race apparently plays in the causation inquiry does little to clear up the doctrinal confusion in this area. Courts would examine the success of minority candidates as a factor in the Section 2 context apart from any causal focus. 42 U.S.C. § 1973(b).

¹¹² *See, e.g., Old Person v. Cooney*, 230 F.3d 1113, 1128 (9th Cir. 2000); *United States v. Charleston County*, 365 F.3d 341 (4th Cir. 2004) (noting that voting polarization increased when a black candidate ran against a white candidate); *see also Goosby v. Town Board of the Town of Hempstead*, 180 F.3d 476, 495-98 (2d Cir. 1999) (rejecting a party-not-race argument primarily because of the defendant jurisdiction's candidate slating process, which is itself the fourth Senate Report factor).

wins and losses or perhaps vote tallies. If vote tallies or election results are the outcomes of interest, then the institutional or socioeconomic actor whose behavior we are interested in assessing is the set of persons eligible to vote in each election. So far so good. Identifying the units presents more difficulties; if we consider units to be electoral contests, the fact that candidates usually run more than once renders the ordinary assumption that units do not interfere with one another hogwash. In other words, to maintain the non-interference assumption, we would have to assume away things like the incumbency effect and the coattails phenomenon.¹¹³ If we consider candidates to be units, it becomes hard to define the outcome of interest; which of several potential contests in which a single candidate runs should one examine, or should one combine them in some way over time?

The problem of sharply defining units, however, pales in comparison to the issue of identifying the timing of the treatment administration. The reasoning courts employ in their discussions of causation demonstrates their desire to conceptualize the effect of perceived candidate race as of election day. This is evident from the fact that courts contemplate separating the effect of candidate race from the effect of funding levels, or of party loyalty, or of incumbency, or of use of the media, or of a laundry list of other factors,¹¹⁴ some of which continue to change up until voters enter polling booths. Courts (and some commentators) anticipate accomplishing this task by using “multivariate mathematical inquiry,”¹¹⁵ or by adding variables to regression equations,¹¹⁶ or by employing “properly specified models,”¹¹⁷ which are all complicated ways of saying that certain variables should be included on the right hand side of a regression equation.

We are in trouble here. Recall the presumption that treatment should be conceptualized as having been administered as of the moment of first interaction between the unit (here, the candidate, or the candidate-in-a-given-election) and the socioeconomic or institutional actor of interest (here, the members of the public eligible to vote). The public perceives a candidate’s race long before election day.

Is there something special about the electoral process that would allow treatment assignment to be conceptualized later than the moment potential voters first observe a candidate’s race, perhaps on the day of the election? In a word, no. To the contrary, what we think we know about elections suggests that departing from the ordinary presumption about the timing of treatment application is more dangerous here than elsewhere. I will not rehash

¹¹³ This is another example demonstrating that the non-interference among units assumption can be difficult to recognize and, when recognized, fairly startling. *See supra* note 63.

¹¹⁴ *Black Political Task Force v. Galvin*, 300 F. Supp. 2d 291, 313-15 (D. Mass. 2004) (incumbency); *Vecinos de Barrio Uno v. City of Holyoke*, 72 F.3d 973, 983 n.4 (1st Cir. 1995) (“lack of funds, want of campaign experience, the unattractiveness of particular candidates, or the universal popularity of an opponent”); *Jones v. City of Lubbock*, 730 F.2d 233, 235 (5th Cir. 1984) (Higginbotham, J., concurring in denial of petition for rehearing en banc) (“campaign expenditures, party identification, income, media use by cost, religion, name identification, or distance that a candidate lived from any particular precinct”).

¹¹⁵ *Jones*, 730 F.2d at 234; *see also LULAC v. Clements*, 999 F.2d 831, 860 (5th Cir. 1993) (“detailed multivariate analyses”).

¹¹⁶ *Holder v. Hall*, 512 U.S. 874, 904 n.13 (1994) (Thomas, J., concurring in the judgment); *Rodriguez v. Pataki*, 308 F. Supp. 2d 346, 434 (S.D.N.Y. 2004).

¹¹⁷ Charles S. Bullock, III, *Misinformation And Misperceptions: A Little Knowledge Can Be Dangerous*, 72 SOC. SCI. Q. 834, 838 (1991); *see also* Richard L. Engstrom & Michael D. McDonald, *Definitions, Measurements, And Statistics: Weeding Wildgen’s Thicket*, 20 Urban Lawyer 175, 177 (1988) (referring to “multivariate causal analysis”). *But see* Bernard Grofman, *Multivariate Methods & the Analysis of Racially Polarized Voting: Pitfalls in the Use of Social Science by the Courts*, 72 SOC. SCI. Q. 826, 828 (1991).

arguments already made to the effect that perceived race affects campaign contributions, party loyalty, incumbency, access to the media, and most if not all of the laundry list of other factors courts point to as potential non-racial causes for election results. I do pause to point out that one of the dangers here is the same as the danger in the smoking, lung cancer, and death example articulated above. By analogy to the courts' Section 2 reasoning, if a statistician looks for an effect of smoking on death rates after balancing on lung cancer, diabetes, heart disease, throat ailments, etc., he or she might well end up concluding that smoking is a healthy habit.

The real point here is the courts are the ones making strong and implausible assumptions, not Section 2 litigants who argue that race affects election results by way of affecting the alternative variables upon which courts would focus. That is, by conceptualizing treatment application as occurring on election day, courts are assuming that a candidate's race has no effect on his or her ability to raise funds, to command party discipline and loyalty, to campaign, to gain access to the media, etc., all this while deeming it worthwhile to study whether candidate race has an effect on votes cast. It is one thing to say that candidate race may have no effect on these other variables; it is another to assume that no such effects exist, while simultaneously deeming it useful to study whether candidate race affects voter behavior. One would not, without justification based in evidence, assume that smoking has no effect on risk of lung cancer, diabetes, and heart disease, and nevertheless deem it worthwhile to study whether smoking has an effect on risk of death.

Judges, or at least, some judges, understand some of these principles.¹¹⁸ What they do not appear to understand, and what is the fundamental lesson of this paper, is that no multivariate mathematical inquiry, no regression equation, no properly specified model, no amount of common sense and intuitive assessment, and no (to borrow a favorite judicial phrase) examination of all the facts and circumstances, can solve this problem. The difficulty stems from a failure to identify a plausible time for treatment assignment, and thus a failure to articulate a coherent question linked to available data.

Is it possible to articulate a coherent causal question in the Section 2 context by conceptualizing treatment as administered at a time earlier than election day? I have been unable to do so. One might try to focus on the moment a potential candidate announces his or her candidacy. But commentators¹¹⁹ and courts¹²⁰ know that potential candidates choose the offices they run for only after serious thought as to their prospects for winning, and such a calculation must include the racial preferences of voters, party officials, and potential contributors, all of which potential candidates probably know better than do Article III judges.¹²¹ It would be unwise to assume, in a study designed to assess the effects of candidate race, that no such strategic thinking occurs. Worse, race may affect candidate behavior years before he or she contemplates running for office, perhaps at the candidate's initial choice of party affiliation. Again, one need not be certain that such an effect in fact exists. One must, however, ask whether

¹¹⁸ See in particular the cogent discussion in *United States v. Charleston County*, 316 F. Supp. 2d 266, 300-03 (D.S.C. 2002), and the more economic-based examination in *NAACP v. Thompson*, 116 F.3d 1194, 1198-00 (7th Cir. 1997) (Easterbrook, J.).

¹¹⁹ See, e.g., GARY C. JACOBSON & SAMUEL KERNELL, *STRATEGY AND CHOICE IN CONGRESSIONAL ELECTIONS* 19, 35 (2d ed. 1983); V.O. KEY, JR., *THE RESPONSIBLE ELECTORATE* 7 (1966).

¹²⁰ Both the *Charleston County* and *NAACP* courts explicitly recognized this fact; in the former case, there was testimony to this effect. See 116 F.3d at 1198; 316 F. Supp. 2d at 279 n.13.

¹²¹ Putting aside questions of institutional competence, witnesses hoping to run for office in the future have powerful incentives not to testify truthfully and completely on such subjects.

it is plausible (or worthwhile) to assume that it does not exist and simultaneously study whether race affects voting behavior on election day.

At a most fundamental level, the difficulties here are the result of three overarching factors. The first is the tug-of-war articulated in section III.A.2. The long (often decades-long) time period between the moment of first candidate-electorate interaction and vote tallies renders the tug-of-war tension particularly evident.

The second overarching factor is the nature of the institutional or socioeconomic actor whose behavior we are attempting to assess. “The electorate” (more accurately, the potential electorate) is a complicated mass of individuals whose choices and motivations change over time and whose actions even at a particular time point are difficult to understand. The employment discrimination context was hard enough, but there, we had the advantage of not being concerned about the potential for someone else’s discrimination; only the employer’s behavior mattered. In the Section 2 context, by contrast, judges must assess voter “opportunity to participate in the political process and elect candidates of . . . choice,” and applicable law directs them to consider the “totality of circumstances.” The entire political system is fair game.¹²²

The third overarching factor is strategic thinking on the part of persons contemplating a run for political office. In terms of the vocabulary used in this paper, treatment assignment has occurred before treatment application, *i.e.*, the potential candidate knows what the electorate will perceive his or her race to be before he or she enters the electoral contest. The previous portions of this paper have used the terms “treatment assignment” and “treatment application” interchangeably because in many situations the two occur (or can be conceptualized as occurring) simultaneously. That will not do here. Potential political candidates are hyper-strategic actors.¹²³

I conclude by articulating the doctrinal implication of this discussion: until resolved, the problems here are a reason not to interpret Section 2 as including a causal element. If courts feel that other doctrinal considerations, considerations beyond the scope of this paper, compel some kind of causal inquiry in the Section 2 context, they should at least define sharply (i) the units, (ii) the treatment, (iii) the timing of treatment administration, (iv) the outcomes of interest, and (v) the relationships the units have to one another (*i.e.*, how to handle interference among units caused by, say, the incumbency and coattail effects). They should then articulate how available data have any information on the questions identified. Perhaps courts will succeed in doing what I have been unable to do. Until they do so, they are interpreting the statute to include a nonsensical question.

IV. Conclusion

I conclude by clarifying that the potential outcomes framework is not so much a solution to problems of causal inference in the civil rights context as much as it is a coherent structure within which to attempt to solve problems. Its adoption would constitute progress, not final answers. The framework provides a way to reduce expert witness bias, to assess whether

¹²² The above discussion should make clear matters are far worse if one shifts the treatment definition from the race of the candidate to the races of the voters. Which voters? When is treatment applied?

¹²³ Recently, scholars have begun to examine the challenges for causal inference posed by units’ taking anticipatory action between treatment assignment and treatment application. *See, e.g.*, Gerard J. van den Berg, *An Economic Analysis of Exclusion Restrictions for Instrumental Variable Estimation*, available at <http://my.harvard.edu/icb/icb.do?course=fas-gov3009&pageid=tk.page.gov3009.dir.8ff922bbcd8ad41cdfc48d3c5163b2ab> (2006 Spring Presentation Schedule) (copy on file with author).

questions in litigation are at the present time answerable, and to identify unrealistic assumptions, all within a definition of causation accessible to judges and juries. The paradigm helps clarify what data that must be collected to implement studies with believable conclusions. And does a better job of forcing quantitative analysts to articulate the assumptions required to use data to provide answers to questions of interest.

Adoption of the potential outcomes framework would make analyses better, not easier. One of the attractive, but dangerous, characteristics of regression as used in modern civil rights litigation is that it is so simple to implement. At least, the initial steps of regression are easy to implement; the process of assessing fit, attempting different equations, etc. requires more effort. Comparatively speaking, however, much of the process of implementing a regression as it is used in modern civil rights litigation does not require as much effort. Implementing a potential outcomes causal analysis in the civil rights context requires work before the analyst sits down in front of the computer, work in identifying units, treatment, timing of treatment assignment, and outcomes of interest, along with work in maximizing the plausibility of initial assumptions. If these initial stages go well, more hard work is required to identify and balance covariates. Only then, after all of the “design” phase of the study is over, does the “analysis” phase become simpler. Reasoned judgment is required at every step.

Finally, much remains to be done in this area. I have focused this paper on gender, race, and national origin discrimination and shown that many hurdles still exist. Age claims are another matter. Here, an initial challenge is to identify the treatment, *i.e.*, the counterfactual. For example, imagine a 60-year-old employee with 30 years of experience in a given field and an expectation of being able work for another eight years. Should this employee be compared for age discrimination purposes to (i) a 40-year-old worker with ten years of experience and the expectation of eight more years of work, (ii) a 40-year-old worker with ten years of experience and 28 more expected work years, (iii) a 40-year-old with 30 years of experience and eight more expected work years, (iv) a 60-year-old worker with 30 years experience and 28 more expected work years, or (v) something else? Finding actual employees with some of these profiles to donate outcome values could prove difficult. But the benefits of the potential outcomes framework are evident in this example as well: experts, litigators, and courts are forced to articulate their questions clearly, which requires close considerations of statutory objectives and policy goals. The results should be not just more accurate quantitative answers, but better law.