Econometrics in the Courtroom

Daniel L. Rubinfeld

Berkeley Law

Follow this and additional works at: https://scholarship.law.berkeley.edu/facpubs

Part of the Economic Theory Commons, and the Law Commons

Recommended Citation

Econometrics in the Courtroom, 85 Colum. L. Rev. 1048 (1985)

This Article is brought to you for free and open access by Berkeley Law Scholarship Repository. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of Berkeley Law Scholarship Repository. For more information, please contact jcera@law.berkeley.edu.
ECONOMETRICS IN THE COURTROOM

Daniel L. Rubinfeld*

The use of statistical methods for resolving disputes has found increasing acceptance within the adversary system.1 This greater acceptance of statistics has opened the door to law-related econometric studies, particularly in connection with the use of multiple regression models.2 While the most frequent uses of multiple regression have been in cases of sex and race discrimination3 and antitrust violation,4

---

1. According to Professor Fienberg, between January 1960 and September 1979 the words “statistic” or “statistical” appeared in four percent of the reported federal district court opinions. Fienberg, The Increasing Sophistication of Statistical Assessments as Evidence in Discrimination Litigation, 77 Am. Stat. A.J. 784 (1982). Professor Tribe argues against the use of statistical methods in legal proceedings in part because of the appearance of accuracy and precision that such methods impart to the process. Tribe, Trial by Mathematics: Precision and Ritual in the Legal Process, 84 Harv. L. Rev. 1329 (1971). While sympathetic to Professor Tribe’s concerns with the exaggerated precision that can be improperly inferred from the expert, I am convinced that the gains associated with the use of statistical methods far outweigh the costs. I proceed on that assumption. For a critical response to Professor Tribe’s argument, see Lempert, Modeling Relevance, 75 Mich. L. Rev. 1021 (1977).

2. The growth of multiple regression in the courtroom has been more recent than the growth of statistical techniques generally. Fienberg claims to have found only three references before 1972 and a total of only 26 references between 1960 and 1979. Fienberg, supra note 1, at 784. In a recent scan of Lexis, I found 84 cases covering federal decisions from December 1969 to July 1984 that refer to regression analysis. For much of this paper I will treat multiple regression as synonymous with econometrics. “Econometrics,” however, may be more broadly defined as the application of statistical methods to the study of economic concepts. Statistical methods that do not involve multiple regression include: (a) moving average and autoregressive time series methods; (b) spectral analysis; and (c) econometric simulation. I assume that the reader has a working knowledge of the basic fundamentals of multiple regression analysis. For those wishing an elementary review, see Rubinfeld & Steiner, Quantitative Methods in Antitrust Litigation, 46 Law & Contemp. Probs., Autumn 1983, at 69, 88–104. A more advanced and comprehensive review is given in R. Pindyck & D. Rubinfeld, Econometric Models and Economic Forecasts (1981).


---

* Professor of Law, Professor of Economics, University of California, Berkeley. I wish to thank Marianne Constable for her helpful research assistance, and Peter Steiner for stimulating much of my thoughts on the subject. Thomas Campbell, David Kaye, Al Klevorick, Richard Lempert, David Peterson, Susan Rose-Ackerman, Jan Vetter, and Anne Witte made helpful comments on an earlier draft. All errors remain my own.

1048

HeinOnline -- 85 Colum. L. Rev. 1048 1985
other applications have ranged across a wide variety of cases, including those involving census undercounts,\(^5\) voting regulations,\(^6\) and the study of the deterrent effect of the death penalty.\(^7\) This Article raises a number of theoretical questions regarding the rules of evidence, the choice of appropriate burden of persuasion in different kinds of litigation, and the use of experts in the legal process. In addition, the Article recognizes the important practical problem that "standard" statistical procedures are subject to manipulation when applied in a courtroom setting. Thus, a fruitful undertaking for law and economics research is to develop procedures that are both practical and suitable for a statistical litigation context. The suggestions made by the Article are intended to raise provocative questions for both legal professionals and social scientists to debate, rather than to provide definitive policy recommendations.

Section I introduces the most prominent application of econometric methods—the use of significance levels for hypothesis testing. I argue that instead of accepting statistical rules of thumb such as the five percent significance test, courts should use an instrumentalist, efficiency-oriented criterion for determining appropriate standards of proof. In Section II, I discuss several problems associated with hypothesis testing—some that are inherent in the particular method, and others that result from the manipulable nature of statistical technique. I urge that significance tests be tailored both to the hypotheses to be evaluated and to the particular method to be used for obtaining data. Section III focuses explicitly on the problems of analysis and presentation that are specific to multiple regression. After introducing the technique, I examine the sensitivity of regression results to the choice of a model and its theoretical premises.

Section IV applies multiple regression analysis to a case involving the drug ampicillin, thus highlighting the problems and solutions discussed in Sections II and III. Section V moves beyond explicit hypoth-


[6. Regression was used in suits charging that at-large area-wide voting was instituted to neutralize black voting strength. Multiple regression demonstrated that the race of the candidates and of the electorate was a determinant of voting. See Kirksey v. City of Jackson, 461 F. Supp. 1282 (S.D. Miss. 1978); Brown v. Moore, 428 F. Supp. 1123 (S.D. Ala. 1976), aff'd mem., 575 F.2d 298 (5th Cir. 1978); Bolden v. City of Mobile, 423 F. Supp. 584 (S.D. Ala. 1976), aff'd, 571 F.2d 238 (5th Cir. 1978), rev'd, 446 U.S. 55 (1980), vacated, 446 U.S. 236 (1980).]

esis testing to the use of forecasting and simulation—methods that build directly on the outcome of a multiple regression study. Although these techniques can be widely applied, they sometimes generate misleading results that are mistakenly accepted as accurate. In the Conclusion, I raise some questions about the use of statistical evidence and experts in the courtroom, especially in the context of complex litigation. While there are no easy solutions, I propose an expanded role for neutral experts and argue that the process for presenting statistical information to the trier of fact should be seriously reevaluated and reformed.

I. TOWARD AN INSTRUMENTALIST CONCEPTION OF THE STATISTICAL BURDEN OF PROOF

Courts often accept conventional practices of the statistics profession without considering whether such practices are valid in the context of litigation. The most apparent of these practices has been the determination of a statistical level of confidence associated with the burden of persuasion set by a court—preponderance of the evidence, clear and convincing evidence, or proof beyond a reasonable doubt. I have some doubt as to whether a specific level of statistical significance should be attached to a particular burden of persuasion. But I am convinced that if significance levels are to be used, it is inappropriate to set a fixed statistical standard irrespective of the substantive nature of the litigation. The choice of evidentiary standard should take into account the behavior of individuals involved in the process and of others whose behavior might be affected by the outcome of the litigation. More broadly, a completely satisfactory determination must be grounded on an understanding of the legal process that is both concep-

8. The distinction between burden of production (coming forward with the evidence in order to raise an issue) and burden of persuasion (ending up with sufficient evidence) is an important one. I wish to thank David Kaye for helping to make this clear. For useful discussions of the distinction, see McCormick on Evidence § 336 (3d ed. 1984); James, Burdens of Proof, 47 Va. L. Rev. 51 (1961).


10. This point has, of course, been acknowledged by others. See, e.g., Block & Sidak, The Cost of Antitrust Deterrence: Why Not Hang a Price Fixer Now and Then?, 68 Geo. L.J. 1131 (1980).

11. The same argument should be applied to the choice of legal standards of liability, such as the standard of care in a negligence setting. See, e.g., Blume & Rubinfeld, The Dynamics of the Legal Process, 11 J. Legal Stud. 405 (1982); Brown, Toward an Economic Theory of Liability, 2 J. Legal Stud. 323 (1973).
tually sound\textsuperscript{12} and empirically accurate\textsuperscript{13} with respect to issues such as who initiates litigation, how parties negotiate, which cases are litigated, and which are settled.

Among the instruments available to the court to influence litigant behavior are (1) the choice of discovery rules that affect the evidence adduced at trial, (2) the determination of the burden of production of evidence, (3) the determination of the burden of persuasion, and (4) the determination of an award to be paid by the liable party to the injured party. Each of the choices can have important effects on the propensity of litigants to pursue or avoid litigation, and to settle or incur some level of trial expenses.

The appropriate considerations in making use of these instruments include not only the social costs of the activity, but also the legal costs to the private parties—the defensive prelitigation expenditures incurred by prospective defendants, and the pretrial and trial expenditures incurred by actual parties.\textsuperscript{14} The choices of instruments should also aim to minimize the costs to society of fact-finding errors. Statisticians describe these fact finding errors as either Type 1 or Type 2 errors. Type 1 errors involve the cost of concluding that an activity was illegal—for example, that there was discrimination amounting to a violation of Title VII\textsuperscript{15}—when in fact it was not. Type 2 errors involve the cost of wrongly concluding that an activity was not illegal, when in fact it was.\textsuperscript{16}

Courts have implicitly understood the relative costs that are imposed by different choices of burden of persuasion. For example, in In


\textsuperscript{13} For two among many interesting examples of the behavioral perspective, see Danzon & Lillard, Settlement Out of Court: The Disposition of Medical Malpractice Claims, 12 J. Legal Stud. 345 (1983); Priest, Selective Characteristics of Litigation, 9 J. Legal Stud. 399 (1980).

\textsuperscript{14} A proper resolution of this question could yield interesting insights into the reasons why the burden of persuasion should vary with the nature of the offense, suggesting perhaps that there ought to be greater variation than currently exists under the law. For example, if one were to weigh expenditures by guilty defendants in inverse proportion to the seriousness of the criminal offense, the cost-minimizing burden of persuasion would vary by offense. This calculus is sketched out in the following paragraphs.

\textsuperscript{15} For the purposes of this example, I assume that wage discrimination is both a necessary and a sufficient condition for proof of illegality under Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e (1982).

\textsuperscript{16} For a discussion of the costs of convicting innocent parties and the theory of deterrence, see Wittman, Two Views of Procedure, 3 J. Legal Stud. 249 (1974).
Justice Harlan stated:

The standard of proof influences the relative frequency of these two types of erroneous outcomes. If, for example, the standard of proof for a criminal trial were a preponderance of the evidence rather than proof beyond a reasonable doubt, there would be a smaller risk of factual errors that result in freeing guilty persons, but a far greater risk of factual errors that result in convicting the innocent. Because the standard of proof affects the comparative frequency of these two types of erroneous outcomes, the choice of the standard to be applied in a particular kind of litigation should, in a rational world, reflect an assessment of the comparative social disutility of each.

Courts in civil cases ought to acknowledge explicitly that setting standards for statistical proof involves just such an assessment of comparative social costs. The specific details of the proposed analysis will depend in part upon one's appraisal of the relationship between prior information about the liability of the defendant and the information that is presented at trial.18

Consider a negligence case tried before a judge without a jury. The judge has certain prior notions concerning the probability of liability on the part of the defendant, based in part on his experience with previously tried cases and their disposition on appeal, as well as on the experience of others. The litigation process provides the judge with additional information, which the judge combines with the prior estimate to form an updated or posterior estimate of the defendant's likely liability. In this "model" the judge's role is to compare his "subjective"


Justice Harlan said that "[a] preponderance of the evidence standard therefore seems peculiarly appropriate for, as explained most sensibly, it simply requires the trier of fact 'to believe that the existence of a fact is more probable than its nonexistence . . . .'" In re Winship, 397 U.S. 358, 371 (1970) (Harlan, J., concurring) (quoting F. James, Civil Procedure 250–51 (1965)) (footnote omitted). One might be tempted to translate Justice Harlan's statement, as others have, into a rule that the preponderance of evidence translates into a determination that the probability of a fact is greater than .5. Such a simplistic interpretation or application of the standard of proof is likely to be inappropriate, since there is unlikely to be a single level of significance that ought to correspond to the appropriate standard of proof; rather the appropriate burden of persuasion ought to vary with the circumstances involved with cases of a given type.
estimate of the probability that the defendant was negligent to the appropriate burden of persuasion. Assume that the chosen standard is a probability of .95 (which need not be consistent with the preponderence of the evidence standard). If the subjective probability of the defendant being liable is greater than or equal to .95, the plaintiff’s contention should be accepted. If it is less than .95, the judge should rule in favor of the defendant.

In the context of such a model, it is possible to begin an evaluation of how changes in the choice of burden of persuasion will change the costs associated with the legal system. A complete evaluation and a more robust model must await further research. A more detailed look at this example should give the reader a sense of the trade-offs involved, however.

Assume that the court considers lessening both parties’ burden of persuasion by decreasing the standard from ninety-five percent to ninety percent. Initially, this change in “policy” may encourage more plaintiffs to bring cases since, other things equal, the lighter burden of persuasion will increase their expected monetary gain from litigation. How this change will affect the likelihood of settlement before trial is an especially difficult question, since the decision to settle depends in substantial part on the fact that the parties subject to litigation have different subjective estimates of their likelihood of winning the case. But changes in expected outcomes can affect settlement possibilities when there are differences in perceptions about outcomes and differences in attitudes towards risk. If the case goes to trial, the lighter burden of persuasion may make effort by the plaintiff relatively more effective than effort by the defendant, thus leading to increased expenditures by plaintiffs (and by defendants, as well, in response).

Finally, the lighter burden of persuasion might actually deter defendants from behaving negligently, since the likelihood of litigation has increased. Along with this deterrence, however, comes an increase in “defensive” expenditures by defendant, expenditures that might not be socially productive.

This illustration indicates at the very least that, because the choice of policy instruments by the legal system generates individual incentives with certain social and economic consequences, these costs ought to be known or estimated first. More importantly, it suggests that the policy instruments ought to be selected entirely or in part according to the incentives and costs that they generate. Nonetheless, it is too early to reach any definitive conclusions about policy, since a number of philosophical, conceptual, and empirical questions remain: how should

Type 1 and Type 2 errors be weighed? How should the behavior of the
trier of fact be modelled? How substantial are the deterrence effects?

The importance of determining meaningful statistical significance
levels is a concern which permeates nearly every context in which statis-
tical methods are introduced in litigation. The choice of statistical sig-
nificance level is crucial not only in determining such typical legal
issues as liability or nonliability, but also in its influence on the use of
multiple regression techniques, where the manipulative nature of statis-
tical procedures may be closed to scrutiny. In addition, the choice of
significance levels can have important implications for the selection of
null and alternative hypotheses.  

II. HYPOTHESIS TESTING

Tests of statistical legal assertions, known as hypothesis testing, can often indicate whether a violation of the law has occurred in areas
where particular evidence is inaccessible or inconclusive. Cases in
which this is the paramount issue, such as sex and race discrimination
cases, may use hypothesis testing to determine the presence of discrimi-

natory effect and possibly, discriminatory intent. Where the determi-
nation of the magnitude of the effect associated with a particular
violation is the crucial question, as in many antitrust cases, other more
direct statistical methods are appropriate.  

20. Hypothesis testing is the statistical procedure that tests a theory within the con-
text of a specific model. One first specifies a well-defined null hypothesis—for example,
that race has no effect on wages—in a model including all proper determinants of wages.
The null hypothesis is then evaluated against the alternative hypothesis that there is
some effect (again within the confines of the model). If the statistical evidence—for ex-
ample, a sample of wage data distinguished by race—suggests that it is sufficiently likely
that race has an effect on wages, then the null hypothesis is rejected in favor of the
alternative. If the evidence is not sufficiently strong, the null hypothesis is sometimes
presumed to be correct, but a more accurate description would simply say that the evi-
dence was not sufficiently strong to allow for its rejection.  

21. In general, it is possible for a hypothesis test to reject the null hypothesis even
though the likelihood of the null hypothesis being correct is greater than the likelihood
that any single alternative hypothesis is true. This is a problem that is associated with
classical hypothesis testing, since prior information about the likelihood of each hypoth-

esis being correct is not taken into account. For a detailed discussion of this point, see
Kaye, Hypothesis Testing and the Burden of Persuasion (July 1984) (unpublished manu-
script) (on file at the offices of the Columbia Law Review). Kaye argues for a purely
Bayesian approach to the evaluation of statistical evidence. I am inclined, however, to
modify the classical hypothesis testing approach, in part because it is the approach that
has been used by the courts, and because the Bayesian approach is difficult to apply in
complex circumstances.  

22. In general, a hypothesis test cannot in itself determine whether a violation of
the law has occurred, since information other than statistics is pertinent to the case.

23. Throughout most of the discussion of hypothesis testing, I frame the null hy-
pothesis so that its rejection is associated with the defendant being liable. This is natural
in most civil cases since no effect or impact is usually associated with the nonliability of
the defendant. The null hypothesis might also be framed in terms suggesting the culpa-
I begin with the question of the appropriate level of significance of a hypothesis test. Subsection A argues that, in the context of a particular litigation, the choice of the burden of persuasion ought to depend upon the nature of the alternative and null hypotheses24 as well as upon the quality of the data. Subsection B discusses the additional difficulty that statistical samples may not be random. Such nonrandomness is especially problematic with multiple regression analysis, where tests of statistical significance are sensitive to the choice of explanatory variables and to the form of the regression equation (the "specification" of the model).25

A. The Choice of the Alternative Hypothesis

When the only alternative courses of action to sex-neutral treatment are deemed by society to be inappropriate ones, then a test comparing the null hypothesis of equal treatment to the rather vague alternative hypothesis of unequal treatment will be the appropriate statistical approach. When only some forms of unequal treatment are inappropriate, however, then either of two approaches is reasonable. First, one can utilize the same statistical methods, but frame the alternative hypothesis to include only those kinds of unequal treatment that are illegal. Second, one can utilize a multiple regression framework in which additional variables are added to the statistical analysis to control for appropriate or legal reasons for unequal treatment. Since the analysis of the two situations is both important and difficult, I treat the situations separately. First, I treat the case in which all forms of unequal treatment are inappropriate. In this context, I illustrate the principle that the form of the alternative hypothesis can affect the conclusion that one reaches from the statistical analysis. Second, I treat a more complex situation in which some forms of unequal treatment are reasonable, and I show the importance of properly specifying the alternatives to unequal treatment.

Assume that as a part of a lawsuit involving sex discrimination, an

---

bility of the defendant in cases in which the defendant is charged with failing to satisfy an affirmative obligation of some sort.

24. Typically the null hypothesis is the hypothesis that there is no effect, while the alternative hypothesis is that there is some effect. Alternative hypotheses, however, can vary substantially in their specificity, and the choice of the alternative hypothesis will affect the decision to reject the null hypothesis. For further discussion of the decision-theoretic aspects of this issue, see M. DeGroot, Optimal Statistical Decisions (1970).

25. In its simplest form, a multiple regression model relates a dependent variable to a set of independent or explanatory variables and a random error term that accounts for the effect of omitted variables. Five percent significance tests (corresponding to 95% probability standards) are frequently used because economists wish to minimize the chance that they will wrongly reject the null hypothesis when it is in fact true. A higher numerical level of significance, say 10%, is associated with a higher probability that the null hypothesis will be wrongly rejected, and therefore corresponds with "less significant" results.
econometrician is used to analyze the hiring and wage policies of a firm. The econometrician specifies a model that relates wages paid to the determinants of that wage (say, experience and training) and the sex of the employee. A substantial effect of sex on wage rates—in other words, differential wages by sex for equally qualified individuals—will be taken as support for the presence of discrimination, while an absence of effect will support the opposite conclusion.

Assume next that the econometrician evaluates a large sample of data describing wages, sex, and other variables and concludes that the null hypothesis of no effect of sex on wages can be rejected at the five percent significance level, or equivalently, that the probability of wrongly rejecting the hypothesis of no effect is less than .05. Put somewhat differently, the significance test tells the econometrician and the court that it is very unlikely that the pattern of evidence that was found could have arisen if the null hypothesis were true—in other words, if deviations from the average were only random. The court, often encouraged by the expert, is tempted to conclude that there is effect and therefore discrimination.

One problem is that the statistical statement that an outcome is significant will of necessity be dependent upon a correct choice for all variables included in the study. If a variable is incorrectly omitted, the study may falsely suggest that the effect of sex is significant. Additionally, where the sample being studied is large, statistically significant results are likely to be obtained even for small differences in wages. As a consequence, one must require practical significance before differential effect by sex is deemed important.

1. Specifying the Alternatives. — Another problem is that the choice of alternatives to the null hypothesis will affect the statistical significance of the results. Imagine, for instance, a hiring discrimination case in which only statistical evidence is pertinent to plaintiff’s claim. Under a preponderance of the evidence standard, the court should find discrimination if the available data are more consistent with the plaintiff’s claim of discrimination than with the defendant’s no discrimination alternative. Suppose initially that a firm has been hiring individuals for a job that requires no particular skills or experience, that there is a large pool of individuals in the relevant geographic labor market, and that fifty percent of them are women. Ten job vacancies are then filled: three with women and seven with men. All eligible men and women are assumed to be equally interested in obtaining employment and equally able to perform the appropriate tasks.

---

26. Throughout this discussion, I use the term “discrimination” to refer to differential treatment by sex without reference to the illegality of that treatment.

27. The numbers are chosen to simplify the presentation and the calculations. Since my concern lies solely with the issue of whether one alternative is more likely than another, the question of whether results are statistically significant at the five percent level (an unlikely prospect with such a small sample) is not relevant here.
The court in this simple example would want to determine whether the fact that three women were hired is different statistically from the five who would be hired on the average, were hiring independent of sex. The usual procedure is to calculate the standard deviation of the estimated mean number hired under a no discrimination hypothesis to see whether three is more than a sufficient number of standard deviations from five. In this particular example, the usual statistical approach, based on the binomial model, would lead to a standard deviation of the mean of 1.58. It follows that three is approximately 1.26 standard deviations away from five. Whether this difference is statistically significant depends on the choice of significance level. It would not be significant, for example, were a five percent level used, since a minimum of 1.96 standard deviations would be necessary for significance. Thus, a statistician might conclude that there is not sig-

28. To understand the meaning of the statistical concept of standard deviation, it is useful to distinguish the sample mean from the true or population mean. The latter describes the number of women who would be hired (1 of every group of 10) were data to be available for all past, current, and future hires. The larger the sample studied, the smaller the standard deviation of the mean, in other words, the smaller will be the changes in the estimate of the mean when and if a new sample is studied. The smaller the standard deviation, the smaller the likelihood that the observed value of three is not just attributable to chance variation. On the other hand, with a high standard deviation, chance variability may be great enough so that three does not differ in a statistical sense from the mean of five that would arise if hiring were independent of sex.

29. For a discussion of the "Bernoulli process" underlying the binomial distribution, see, e.g., W. Feller, An Introduction to Probability Theory and its Application 146 (1968). There is good reason to question the use of the binomial distribution in employment discrimination cases, since the process by which individuals apply for jobs and are selected may not be random. How good an approximation is given by the binomial distribution would have to be evaluated on a case-by-case basis by asking (a) whether it is reasonable to assume that the probability that a given individual who is hired will be female is the same as the overall percentage of women in the relevant labor market, and (b) whether each of the hires was made independently (in a statistical sense) of the others. The major difficulty arises when there is nonstatistical evidence concerning discrimination. See Braun, Statistics and the Law: Hypothesis Testing and its Application to Title VII Cases, 32 Hastings L.J. 59 (1980); Meier, Sacks & Zabell, What Happened in Hazelwood: Statistics, Employment Discrimination, and the 80% Rule, 1984 Am. B. Found. Research J. 139.

30. Assuming a binomial distribution and assuming that women make up half the eligible labor force, the standard deviation is equal to $\sqrt{(.5 \times .5 \times 10)} = 1.58$.

31. $(5 - 3) / 1.58 = 1.26$.

32. The number 1.96 comes from the normal distribution. With a sufficiently large sample, the binomial distribution approximates the normal. With a normal distribution, approximately 2.5% of the distribution lies in each tail, where the tail is defined to include the area more than 1.96 standard deviations away from the mean. The five percent significance test just described is a "two-tailed" test, which allows for the possibility of discrimination against either men or women. If one uses a "one-tailed" test, which tests only for discrimination against women, one would need a difference of only 1.65 standard deviations to decide that there is a significant difference between three and five. The use of one-tailed and two-tailed tests is discussed in Pennsylvania v. Rizzo, 20 Fair Empl. Pract. Cas. (BNA) 130 (E.D. Pa. 1979).
significant evidence of disparate treatment of female applicants.  

Suppose, however, that there are three possibilities:  

(a) the firm does not consider gender in its hiring (i.e., the null hypothesis);  

(b) the firm hires every qualified male applicant, but refuses to hire any married women who apply for the job (two-fifths of all women applicants are married);  

(c) the firm hires only one-half of all married women, refusing to hire married women with children (one-half of all married women who apply have children).

Suppose also that both the plaintiff and defendant have the requisite information to specify these alternative hypotheses, so that the question of who is best able to obtain information and therefore ought to bear the burden of production of that evidence is not relevant. Finally, suppose for argument’s sake that it is impossible or improper to obtain information about marital and family status for purposes of litigation.

It is now possible for a statistician to analyze the available data and to compare the possibility that each of the three hypotheses is correct.

(1) Assuming that the null hypothesis (a) is true, the probability

---

33. The source of this argument is Castaneda v. Partida, 430 U.S. 482 (1977), in which Mexican-Americans brought suit against the sheriff of Hidalgo County, Texas, for discrimination in grand jury selection. The Court stated: “As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist.” Id. at 497 n.17 (The number of standard deviations was approximately 12.).

Courts have generally followed the Castaneda approach, although ambiguities about the standard remain. At least one court has maintained that a disparity of two standard deviations is insufficient to establish a prima facie equal protection claim. See Boykins v. Maggio, 715 F.2d 995, 996 (5th Cir. 1983). In the much cited employment discrimination case of Hazelwood School Dist. v. United States, 433 U.S. 299 (1977), the Court dropped Castaneda’s qualifying language about social science and noted that under Castaneda’s “precise” methodology, a disparity of slightly less than two standard deviations was not “suspect.” Id. at 309 n.14.

34. If such data were available, the appropriate procedure would be to perform the statistical tests for each of the relevant subgroups, e.g., married women and nonmarried women. In sex discrimination cases, a failure to “stratify” in this way can bias statistical tests.

35. A number of assumptions would go into such an analysis. First is the assumption that a hypothesis test is the appropriate procedure to use. Hypothesis tests of this type are either neutral or restrictive, depending upon one’s point of view, in the sense that the specification does not prior information that the court might have about the probability that each of three hypotheses might be true. Second is the assumption that the sample of job applicants can be represented by the Bernoulli process. See supra note 29. In the most general formulation, this process can allow for male and female applicants to have different likelihoods of being hired, although it does assume that all women are alike. It is not to dissuade the reader about the usefulness of the statistical tools that these assumptions are made explicit. Rather, they are made in the spirit of good statistical analysis, in which the expert is continually thinking about
that the number of females hired will be three is .12.36

(2) Assuming that alternative hypothesis (b) is true, the probability that the number of females hired is three is equal to .20.37

(3) Assuming that alternative hypothesis (c) is true, the probability that the number of females hired is three is equal to .06.38

Under the preponderance of the evidence standard, a hypothesis testing approach ought to involve a comparison of the probability of the evidence given the null hypothesis, to the probability of the evidence given an alternative hypothesis that is adequately specified for the purposes at hand. Suppose, again, solely for purposes of argument, that the only inappropriate form of unequal treatment that is called into question by this suit is discrimination against married women. Then, if the null hypothesis is (a), and the appropriate specification of the alternative hypothesis is (b), the null is less likely than the alternative and the trier of fact should rule for the plaintiff.39 If instead the only form of inappropriate treatment is discrimination against married women with children, then the trier of fact should rule for the defendant,40 since alternative hypothesis (c) would be less likely than the null hypothesis (a).

2. More Complex Alternatives. — This simple example intentionally overstates the point, however. Neither alternative (b) nor alternative (c) need be appropriate in any given case. Perhaps it would make more sense to allow for a more complex alternative containing several possible types of discriminatory behavior.41 For example, one alternative

---

36. The .12 figure (rounded from .1172) is calculated directly from the binomial distribution, with $p$ (the probability of being selected) equal to .5, $n$ (the number in the sample) equal to 10, and $x$ (the number of selections) equal to 3. See, e.g., H. Brunk, An Introduction to Mathematical Statistics 373 (1965) (Table I).

37. .20 (rounded from .2013) is calculated directly from the binomial distribution, with $p=.2$, $n=10$, $x=3$.

38. .06 (rounded from .0574) is calculated directly from the binomial distribution, with $p=.1$, $n=10$, $x=3$.

39. In fact, the ratio of the likelihood that the null is true to the likelihood that the alternative is true is .58. This analysis assumes that in terms of prior information the null and alternative hypotheses are equally likely.

40. The ratio of the likelihood that the null is true to the likelihood that the alternative is true is 2.04. Note also that if one were testing the hypothesis that alternative (c) were true using a five percent significance test, a calculation that three is greater than two standard deviations (2.10) from one (the expected number) would allow one to reject the alternative.

41. One might consider, for example, specifying the rather vague alternative that the firm discriminates against women in any form ranging from not hiring any women to not hiring one out of every ten. When the alternative hypothesis involves such a possibility the calculation of the likelihood of the alternative (to be compared with the likelihood of the null) becomes substantially more complex. The principal difficulty is that one must weigh each possible type of discrimination by one's prior beliefs about how likely each is to be true. This involves a Bayesian approach, which I followed by assum-
might include all possible forms of differential treatment of women in which women are less likely to be hired than men. Because the possible defenses (e.g., business necessity) to a charge of discrimination are likely to change as the qualifications for the job change, it is important to specify the alternative hypotheses carefully.

In the previous example, I assumed that the labor pool is well defined and that it is known with certainty that women make up fifty percent of that pool. In addition, I assumed that the job did not require special skills or substantial educational background. As a result, it may be difficult to suggest reasonable alternative hypotheses that do not involve discrimination. Thus, the rather general alternative that includes all forms of differential treatment would be appropriate.

The direct hypothesis testing approach is not the only way to proceed, however. Perhaps a more sensible variant of this approach would be to allow the plaintiff initially to focus on the broadest alternative allowing for discrimination in any form in presenting statistical evidence. After some evidence of a relationship between sex and hiring is presented, the burden of production could then be shifted to the defendant to attack the plaintiff's hypothesis.

The shifting of the burden of production becomes more important

See, e.g., A. Zellner, An Introduction to Bayesian Inference in Econometrics 40 (1971). I concluded that the likelihood of this particular alternative is .90, substantially greater than the likelihood of the null (.12). If the case can be made for this particular alternative, the conclusion that there is discriminatory hiring would be clear.

42. In fact, the choice of the relevant labor market and the information concerning the eligible applicants for the job are both extremely important factors in a sex or racial hiring discrimination case. Choice of labor market can affect the outcome of the case. See, e.g., Gay v. Waiters' & Dairy Lunchmen's Union, Local No. 30, 489 F. Supp. 282, 301 (N.D. Cal. 1980); see also EEOC v. United Va. Bank/Seaboard Nat'l, 615 F.2d 147, 149 (4th Cir. 1980) ("When special qualifications are required to fill particular jobs, comparisons to the general population (rather than to the smaller group of individuals who possess the necessary qualifications) may have little probative value."). ("When special qualifications are required to fill particular jobs, comparisons to the general population (rather than to the smaller group of individuals who possess the necessary qualifications) may have little probative value."). ("When special qualifications are required to fill particular jobs, comparisons to the general population (rather than to the smaller group of individuals who possess the necessary qualifications) may have little probative value.") (quoting Hazelwood School Dist. v. United States, 433 U.S. 299, 308 n.13 (1977)). Similar issues arise in Mayor of Philadelphia v. Educational Equity League, 415 U.S. 605 (1973) (racial-composition percentage comparisons rejected); Patterson v. American Tobacco Co., 535 F.2d 257 (4th Cir.) (Holding that district court used wrong population to determine ratio of acceptable performance.), cert. denied, 429 U.S. 920 (1976); Hester v. Southern Ry., 497 F.2d 1374 (5th Cir. 1974) (faulty evidence of an unbiased applicant pool). For additional cases and for additional difficulties raised when past discriminatory practices "have chilled" the applicant pool, see Boardman & Vining, The Role of Probative Statistics in Employment Discrimination Cases, 46 Law & Contemp. Probs., Autumn 1983, at 189, 195–201.

43. I am referring, of course, to discrimination in hiring. Other forms of discrimination are possible.

44. In Griggs v. Duke Power Co., 401 U.S. 424, 432 (1971), disparate impact was chosen as the appropriate standard for evaluating employment practices under Title VII. The Court ruled that if plaintiff could show that black applicants had a substantially lower pass rate than whites on a qualifications test, the use of the test would be illegal, absent a business necessity defense.
when the number and forms of alternative hypotheses are more complex. This occurs naturally when defendants argue that differential treatment of men and women is valid. One difficulty relates to the availability of information concerning possible discrimination. If the defendant has the data or other information that are necessary for the alternative hypotheses to be well specified, then it may be appropriate to make it easy for the plaintiff to shift the burden of production to the defendant.

Another difficulty relates to the qualifications that are necessary for the job. Assume, for example, that the job involves substantial educational background as well as previous experience. The appropriate use of hypothesis testing will now differ depending upon how the "eligible" labor pool is defined. Perhaps the most direct and sensible approach is to redefine the relevant labor market to include only those individuals with sufficient skills and education to be qualified for the jobs in question. In this case there might be substantial debate as to the meaning of the term "eligible." Still, once the relevant labor market is chosen, the hypothesis testing could proceed as described.

In practice, however, it is difficult to define the relevant labor market, and it is incorrect to assume that all eligible workers have equivalent skills and education. As a result, the labor pool that is treated as eligible is likely to consist of some individuals who are not qualified and some who are better qualified than others. When the hypothesis testing approach is followed, the set of alternative hypotheses is likely to be substantial, since one can devise a number of hypotheses that may explain why women have lesser job skills or education than men have. In my example, the fact that less than half of the hirers were women need not indicate a finding for the plaintiff, since the probability of the evidence, given a number of acceptable alternative hypotheses such as (b) or (c), may be higher than the probability given the null hypothesis that sex is not a factor in hiring.

An additional layer of complexity is added because one or more of the alternative hypotheses arguably might involve appropriate defendant behavior. In such a case, one could require the defendant, once an initial statistical showing had been made to clearly state and to provide evidence supporting these alternative hypotheses. For example, if the defendant could show that one of every three women applicants had substantially less educational background than others in the job pool and therefore is less qualified for the job, then alternative hypothesis (c) might be consistent with the evidence, but arguably would not involve a legal violation.

A different approach would place upon the plaintiff the burden of showing that if the null hypothesis of no discrimination is assumed correct, then the probability that the evidence would arise is low, even after one controls statistically for variables such as experience and education. The statistical procedure would allow the court to focus ini-
tially on the question of whether male or female hiring rates were significantly different among individuals with essentially the same job qualifications. As the alternative hypotheses become more complex, the use of multiple regression becomes an obvious generalization of the previously described statistical tests. Only when the null hypothesis of no discrimination is rejected in the multiple regression framework, which includes controls made for the effect of experience and education on worker productivity, might the burden be shifted to the defendant to show that there are additional reasonable alternative hypotheses that have not been accounted for by the plaintiff.

It would be presumptuous to argue on the basis of such a brief discussion that the current legal treatment of employee hiring discrimination ought to be changed. Nor am I in a position to urge a shift in the direction of giving a greater benefit of the doubt to defendants. I raise these questions, rather, to stress the fact that research and thought by economists and others has great potential. The value lies not in narrowly treating questions of statistical significance, but in broadly evaluating the question of whether, and if so when, the burden of production ought to be shifted from one party to another, and what the burden of persuasion should be once the evidence has been produced.

3. The Burden of Persuasion and Significance Levels. — The preceding discussion leads naturally to a further question: Why apply a five percent significance test before the burden of production can be shifted from the plaintiff to the defendant, particularly where the only possible alternative hypotheses involve discrimination. A ten percent as opposed to a five percent significance level would make it easier to reject the null hypothesis of no relation between sex and hiring. At this point, the burden of persuasion would be placed on the defendant to show that there was a rational basis for the hiring practices in question. A low significance level (perhaps five percent rather than ten percent) would be appropriate where there were a number of reasonable alternatives to the no discrimination null hypothesis. The low significance level would make it difficult for the plaintiff to provide statistical evidence that would reject the null hypothesis and therefore make it less likely that the burden of production would be shifted to the defendant.45

The selection of a significance level should include consideration

45. In Spurlock v. United Air Lines, Inc., 475 F.2d 216, 219 (10th Cir. 1972), the court stated:
When a job requires a small amount of skill and training and the consequences for hiring an unqualified applicant are insignificant, the courts should examine closely any pre-employment standard or criteria which discriminate against minorities. In such a case, the employer should have a heavy burden to demonstrate to the court's satisfaction that his employment criteria are job-related. On the other hand, when the job clearly requires a high degree of skill and the economic and human risks involved in hiring an unqualified applicant are great,
of broader social issues that go beyond the narrow question of whether an individual defendant should be held liable. For example, the level of significance ought to be lower in situations where there are higher costs to concluding mistakenly that there is discrimination. Thus, one might value the cost of mistakenly holding an employer liable for discrimination against blacks to be higher than for discrimination against women. Accordingly, a lower significance level in racial discrimination cases than in sex discrimination cases might be appropriate.

Finally, one might alter the level of significance in order to change the behavior of attorneys and statistical experts. Assume, for example, that the type of litigation enables an expert to search among a number of alternative models of discrimination by changing the control variables selectively. A low level of significance would make such model searching difficult, and thus might discourage it. This approach is fraught with danger, however. Significance levels are relevant for testing hypotheses within the confines of a particular model. A poorly specified, somewhat incoherent model might yield very significant results, when a well-specified model would not.

B. Statistical Evidence as the Outcome of a Sampling Process

Most statistical evidence about discrimination or price fixing is treated by experts and by the courts as if it were obtained by taking a random sample drawn from either the population or set of all possible statistical evidence. In other words, there is a tendency for all parties to treat the process by which statistical evidence is developed in legal situations as if it were equivalent to the process by which statisticians sample data in controlled experimental situations. In a particular lawsuit such an assumption may closely approximate proper statistical procedure and therefore be warranted. Indiscriminately viewing evidence as the outcome of a random sampling process, however, is questionable statistical practice, since the plaintiff’s role in initiating litigation tends to make the outcome of the process nonrandom. The sample selection of cases to be litigated is skewed initially by the fact that plaintiffs perceive themselves to have a relatively high probability of success (or a high expected payoff). The effect is exacerbated by attorneys who have a good understanding of the relationship between statistical evidence and the likelihood of success of a lawsuit. For example, the statistical evidence in a hiring discrimination suit is not likely to represent a random sample taken from the population of all statistical evidence dealing with hiring situations. This might lead the court to infer

---

46. One might argue, however, that the incorporation of social values into the litigation process ought to be done only with explicit legislative authority.

47. One might argue further that since the employment process is highly nonrandom, the application of statistical methods is dangerous.
discrimination in a higher proportion of cases than would be warranted under a more conservative statistical analysis.

Assume, for example, that there are a large number of firms in an industry, each of which makes its hiring decisions independently of the other firms in the industry. Any hiring discrimination that does occur (in the form of differential treatment of men and women of equal skill)\textsuperscript{48} therefore occurs at the firm level, rather than at the industry level. Assume also that each firm hires the same number of individuals, so that the probability that any one firm hires a substantially smaller number of women, so as to suggest sex discrimination, is equal for all firms. The number of employees that would be sufficient to constitute a significant difference would, of course, vary with the number of hires. Thus, if each firm hired one hundred employees, a five percent significance test would reject the hypothesis of no discrimination if forty-one or fewer women were hired.\textsuperscript{49}

It would not be surprising in this situation to find that a careful statistical sampling of the hiring practices of each firm shows that several were hiring significantly fewer women than men, even at the five percent significance level. One would expect potential employees of these firms to be much more likely to sue for hiring discrimination than potential employees of the other firms. Thus, in an industry with no sex discrimination present, one or more cases may be litigated—yet the litigated cases will not reflect firm hiring practices throughout the industry.\textsuperscript{50} By basing its decision on statistics—and by using a five percent significance level to test that hypothesis—the court would falsely conclude that discrimination had occurred. The possibility of a false conclusion would increase were the test to have a higher significance level.

The potential social cost of such a conclusion is not trivial, if one supposes that firms respond defensively to the prospect of discrimination suits by altering their hiring practices. This can involve substantial cost to the firm, both in monitoring its hiring and in hiring less productive workers.\textsuperscript{51}

---

\textsuperscript{48} I am treating discrimination and differentials in hiring synonymously in this example. A more complete discussion would account for the fact that discrimination can be specific to the individual, so that a finding of discrimination need not depend upon statistical evidence showing differentials in hiring by sex. If differentials were to be examined, they would have to be adjusted for differences in job qualifications.

\textsuperscript{49} I am assuming again that hiring decisions can be described by a Bernouilli process. See supra note 35.

\textsuperscript{50} My analysis greatly oversimplifies the problem, since the selection of cases for litigation also involves the decision to litigate, rather than to settle the case. Cases that are very strong for a plaintiff on statistical grounds could well be settled, rather than litigated. For a general empirical discussion of this phenomenon, see Priest & Klein, The Selection of Disputes for Litigation, 13 J. Legal Stud. I (1984).

\textsuperscript{51} On the other hand, a threat of litigation can provide social benefits if a firm decides to implement a nondiscrimination policy.
Two conclusions may be drawn. First, the courts must view the statistical evidence presented in a case as a single sample drawn from the population of relevant evidence, which includes all cases that might have been brought. The determination of that population can be difficult, since the possibilities for discriminatory policies of hiring and promoting are pervasive. One rule of thumb that might serve as a useful starting point is to view the population as all firms in an industry if hiring is at issue, and all divisions or sections of a firm if promotion is the issue. If that population is large, the court must treat seriously the possibility that the case was brought because nondiscriminatory behavior appeared to be discriminatory. This suggests a look at the hiring practices of comparable firms, or if a suit involves discrimination within one branch of a firm, an examination of the practices of other branches of the firm.\footnote{For an application of my argument to the question of promotion within firms, see Follett & Welch, Testing for Discrimination in Employment Practices, 46 Law & Contemp. Probs., Autumn 1983, at 171.}

A second conclusion is that, to the extent that the selection bias is deemed to be serious, the court should consider asking for other evidence or lowering the significance level; these actions would reduce the likelihood of finding an innocent defendant liable and would discourage frivolous lawsuits. Of course, if fewer cases are brought, the probability of deterring discriminatory hiring would decrease. Moreover, even if cases are brought, the “randomness defense” may be improperly applied by the courts. If the courts wish to pursue aggressively improper discriminatory hiring policies, then a higher significance level may be appropriate, despite the potential costs that I have outlined. In the end, statistical values may be different from legal values, but both should be made more explicit.

III. Introduction to Multiple Regression

A. Reporting Multiple Regression Results

Multiple regression is generally well suited to the analysis of data and the testing of theories in which there are several possible explanatory variables. Included in the multiple regression are not only those variables of direct relevance to the legal issue—for example, sex—but also those that must be controlled—for example, education and experience—to avoid a false conclusion.

Multiple regression is valuable in numerous situations in which simple statistics will not suffice, including many employment discrimination and antitrust suits. The reasons can vary from case to case, but they invariably relate to the possibility of spurious correlation. Spurious correlation arises when two or more variables move in the same direction and apparently are closely related to one another, but in fact bear no causal relationship to one another. To avoid spurious correla-
tion, a theory must be developed that relates one or more explanatory or independent variables to the dependent variable, whose movements are caused by the explanatory variables. Applied to the theory of a case, multiple regression allows one to choose among alternative hypotheses and to sort out those correlations that are spurious from those that are not.53

Multiple regression involves a well-specified model relating a dependent variable to a series of independent or explanatory variables and a random error term in a form that is usually taken to be linear. The random error term, which picks up the effects of omitted variables, is often assumed to be normally distributed (looking like a bell-shaped curve) and to be uncorrelated with the independent explanatory variables. Consider, for example, the following wage equation, that might be employed in a case involving alleged pay discrimination against women. $W$ represents the hourly wage rate; $E$ is a measure of the years of experience; $A$ measures the age of the worker in years; $S$ is a variable equal to one if the worker is a female, zero if a male; and $e$ represents a random error term:

$$W = \beta_1 + \beta_2 E + \beta_3 A + \beta_4 S + e$$

The $\beta$'s represent constants (sometimes called parameters) that describe the slope of the many-dimensional line which relates differences in the wage, $W$, to differences in experience, $E$, age, $A$, and sex, $S$. For example, the parameter $\beta_4$ represents the mean difference between wages of men and women, when experience and age are controlled for. The controlling process allows wage differences to be evaluated for men and women of comparable age and experience.54

Assuming that the model is correct, both in including the correct variables and in describing the relationship as linear, it is not difficult to compute the "least-squares" estimates of the parameters of the model.55 These estimates appear in the regression equation:

$$W = 1.50 + .25E + .02A - .75S$$

$$R^2 = .30$$

53. The choice of variables to be included in the model comes not only from theory (e.g. what variables are known by labor market economists to determine a wage rate), but also from concerns of the law (e.g. sex).

54. The regression model must have variables that are well measured, but not all determinants of the wage rate need to be included. Excluded variables are implicitly accounted for by the error term. The essential properties of the regression model will be retained, so long as there is no correlation between the omitted variables in the error term and the explanatory variables of interest. For example, assume that hair color is a determinant of the wage in a particular job, but that both men and women are equally likely to have the same hair color. Then, failure to include hair color explicitly in the regression model will not bias any statistical tests that involve sex discrimination.

55. Least-squares, the most popular multiple regression estimation procedure, selects parameter estimates that minimize the sum of the squared deviations of the predicted values of the model (the right-hand side of the regression equation evaluated with the parameter estimates as weights for each variable) from the actual observations.
The coefficient of .02 on the age variable, $A$, for example, tells us that the average hourly wage of employees in the sample increases by two cents as the age increases by a year. The $-.75$ coefficient on the variable $S$ suggests that the average wage of women is seventy-five cents less than that of men, after one controls for experience and age. The standard errors of the parameters or coefficients, which appear in parentheses below the coefficients, indicate reliability of the estimates.

A central issue in any sex discrimination suit will be whether there is any relationship between the wage paid to employees and the sex of those employees, after other wage determining factors are held constant. It is natural, therefore, to frame the null hypothesis of equal wage payment by sex as the hypothesis that $\beta_3 = 0$. The coefficient of $-.75$ is clearly different from zero, but how different depends upon the standard error, in this case .25. If the error term were normally distributed and the sample size were large, we would expect that the distribution of estimates obtained each time a new sample of employees was analyzed would have a mean of zero and a standard deviation of .25. With a small sample, the $t$-distribution replaces the normal distribution, but the hypothesis testing approach is generally the same.

In this particular case, it is very unlikely that an estimate of $-.75$ would be obtained from such a $t$-distribution. In fact, it is not difficult to calculate that the probability that the estimate is greater than .75 (either positive or negative) is only .003. Thus, if the "model" explaining wages is correct in its description or specification, one could conclude that the hourly wages of women are significantly less than those of men, even after controlling for age and experience.

The $t$-statistic, which is sometimes reported in regression studies, is the ratio of the estimated coefficient to its standard error. With 100 observations, for example, the $t$-statistic on the sex coefficient of $-3.0$ is statistically different from zero at the five percent significance level. While the $t$-statistic of $-3.0$ is a useful shorthand measure for reporting significance, I prefer the use of standard errors. The standard error of .25 makes it easy to calculate a ninety-five percent confidence interval, which suggests that the wage differential between men and women is very likely to range between $0.25$ and $1.25.56

Of course, $t$-statistics are quite sensitive to the size of the sample being studied. With a sufficiently large sample, the likelihood of getting a $t$-value greater than two can get large. This occurs because additional data allows us to estimate the coefficient on $S$ with more accuracy, that is, with a smaller standard error. Intuitively, the larger the data set, the easier it is to decide that the female wage is significantly different from the male wage. Statistical significance and practical significance, however, are two different things. Presumably the court would be, and ought to be, more concerned with a wage differen-

---

56. I have added and substracted two standard errors from the coefficient.
tial of $2.00 per hour that is not significant due to limited data, than with a wage differential of $0.02 that is statistically significant. Evaluation of statistical results cannot be accomplished by the use of any single test statistic. Ideally, findings should allow the trier of fact to make independent inferences based on information relating to both statistical and practical significance.

In Melani v. Board of Higher Education, a Title VII suit was brought against the City University of New York for allegedly discriminating against female instructional staff in the payment of salaries. The court ruled for the plaintiff after treating carefully a number of multiple regression analyses. An analysis of the data showed that among all instructional staff the average annual male-female salary differential was approximately $3,530. This overall difference was not appropriate for statistical analysis, however, because it did not account for differences in characteristics of individuals. Such characteristics were arguably related to worker productivity and therefore to the wages that employees would have been paid in a market without sex discrimination. Much of the debate between the statistical experts concerned which set of variables should have served as controls in the multiple regression analysis.

One approach of the plaintiff's expert was to describe mean salary differentials by individual productivity-related factors, such as highest academic degree, age, and years of service. Statistically significant salary differentials continued to persist in the analysis, although the means then ranged from $835 to $1803 per year in the sample of all instructional staff, and $478 to $1629 in the sample of individuals hired after October 1972. Although better than looking at overall sample means, this procedure was inappropriate, since it did not take into account the joint effect of all measurable productivity-related variables.

The actual multiple regression study presented by plaintiff's expert included as many as ninety-eight independent variables. The coefficients of the 0–1 variable that reflected sex were approximately equal to $1800 when all years were included, and $1600 when only years after 1972 were included. Practically and legally, these are statistically significant results. The court, however, did not clearly indicate whether and to what extent practical and statistical significance were individually important in its ruling for the plaintiff. Initially, the court stated that "[p]laintiffs have produced statistically significant evidence . . . ." Yet in the same sentence the court added that "women hired as CUNY instructional staff since 1972 received substantially lower salaries than sim-

58. Id. at 773.
59. Id. at 775.
60. Id. at 774.
61. Id. at 781 (emphasis added).
ilarly qualified men.” 62 Apparently, in Melani, the court found the evidence to be both statistically and practically significant and thereby avoided a potentially difficult problem. One could imagine evidence from a small sample that was deemed substantial, but not significant, or evidence from a very large sample that was deemed significant, but insubstantial.

Finally, one should note that the court chose not to place great weight on the defendant’s objections to the plaintiff’s multiple regression analysis. While some of the objections seem credible, the court’s omission of much of the defendant’s empirical analyses in the opinion makes an evaluation difficult. One objection went to the choice of sample for analysis. The stratification defense, akin to my argument regarding the earlier illustration that a court might want to distinguish married women from unmarried women, was that faculty and administrative positions ought to be analyzed separately. The plaintiff’s expert’s response—to allow for differences in job categories through control variables such as degree and years of service—is ultimately not convincing to me. An additional objection is that the plaintiff failed to include important control variables such as number of publications, years of teaching experience, quality of teaching, and community service. Whether these omissions were serious is hard to discern without additional information. I will have more to say about the problems of omitted variables in multiple regression analysis in subsection B.

A final objection, which also seems credible, is that the plaintiff failed to allow for differences in academic departments. The defendant claimed that differences in mean salaries could arise simply because in a tighter labor market overall salaries are higher in departments that tend to attract relatively more males. As a matter of arithmetic, for example, it is possible that male-female average salaries could have been equal in each department viewed alone (adjusting for other variables), while the average across all departments was substantially higher for men. 63 Defendants apparently did not drive this point home; in any case, the court was ultimately unconvinced. Whether the defendant’s effort was not more successful because a stratification by department would not matter empirically, or because the court was deterred by the problems of adjusting the multiple regression technique to account for differences in departments, is not clear. I would like to see some calculations as to whether male-female differences were practically significant once departmental distinctions were made, but none were presented in the district court’s opinion.

62. Id. (emphasis added).
63. This could occur if departments with high salaries had many men and few women, and the opposite were true for low salary departments.
B. Evaluating Significance Tests in the Multiple Regression Context

The issue of robustness—whether statistical results are sensitive to changes in assumptions and procedures—is of vital importance for the courts. Decisions should be made with assurance of the reliability of the statistical results upon which the decisions are based. While courts often require experts to utilize "standard statistical practice," such as reporting t-tests with multiple regression models, courts have not yet begun to require new techniques that test the robustness of a regression model with respect to some of its assumptions. Assuming that the standard regression model is correct, t-tests can help determine statistical significance, but when the assumptions of the model are inaccurate, t-tests can substantially overstate the significance of statistical results. Accordingly, courts should require experts to report a number of tests illustrating how robust their results are, even if such reporting requires the use of as yet "nonstandard" statistical techniques. Social scientists and statisticians interested in the litigation process should evaluate the usefulness of old as well as new statistical procedures that may be appropriate to litigation.

To illustrate potential improvements, I describe below some errors and failures of assumptions that are likely to arise in legal proceedings. Each description is followed by a suggestion of how changes in reporting regression results can accommodate the need to consider the robustness of the results.

1. Nonnormality of the Error Term and Sensitivity to Individual Data Points.

   The t-tests of significance are valid only if the distribution of the error term, which picks up the effect of omitted variables, is normal. For sufficiently large samples (larger than one hundred perhaps), the normality assumption is usually a reasonable one, but for small samples, it should be suspect. When the errors are not normal, it is possible that the usual t-test will often overstate (but may understate) the significance of a test of the null hypothesis that there is not differential treatment, falsely encouraging (or discouraging) one to conclude that discrimination is present. Related to nonnormality is the occasional extreme sensitivity of regression results to individual data points (suggesting that other evidence might be pertinent).

   Sensitivity can arise because the error process follows a probability distribution that has numerous observations at the high and low extremes, in which case the nonnormality discussion would be relevant. Or, sensitivity can arise because one or two errors are unusually large, relative to the mean and standard deviation, while the remaining portion of the error distribution is approximately normal. It is this latter situation on which most attention will be focused. It is not unusual, even with a relatively large sample, to find that the coefficient in a mul-

64. Fed. R. Evid. 703 states that materials or methods must be of a type that is "reasonably relied upon by experts in the particular field."
Multiple regression can change substantially, perhaps by several standard deviations, if only one data point is dropped from the sample. A large t-statistic implies statistical significance, but it does not imply robustness, so that one must test separately for data sensitivity.

It is always good procedure to list or to plot the individual residuals from the regression equation. For a normal distribution, one would expect that five percent of the residuals would lie outside the interval containing plus or minus two standard errors of the multiple regression. If substantially more than five percent are outside the interval, nonnormality is a problem.

Econometricians do not usually test for and resolve the nonnormality problem. Testing for robustness to individual data points, however, is an important technique and is rapidly becoming standard econometric practice. Such tests ought to become part of the presentation of statistical evidence, especially when small samples are involved. Belsely, Kuh, and Welsch suggest a series of statistics that are relatively easy to calculate and that tell the expert (and hopefully the court) how sensitive the multiple regression results are to individual observations. Among those statistics, the two types discussed in the following paragraphs are perhaps the easiest for a nonspecialist to grasp.

a. **The Difference in the Betas.** — This is a measure that is provided for each observation and each coefficient or “beta.” It is the change in the coefficient of the variable that occurs when a particular observation is omitted from the multiple regression. To make the measure interpretable irrespective of the units of measurement, the difference is scaled or normalized (by the standard deviation of the estimated beta). A difference in the betas that is greater than 2.0 suggests that a particular observation is extremely influential, since it can easily change a significant result to an insignificant one or vice versa. When such observations are found with respect to a coefficient such as the coefficient of the sex variable, the results should be reported with and without the influential observation or observations. An application of the use of this statistic is given in Section IV.

b. **The Difference in the Predicted Values.** — This is a measure which is provided for each observation. It is the scaled or normalized (by the standard error of the regression) change in the predicted or fitted value of the regression when a particular observation is omitted from the regression. The predicted value provides a measure of the best forecast

---

65. The studentized residuals (the residuals divided by their individual standard errors) would be more appropriate.

66. The treatment of this issue is beyond the scope of this Article. See, e.g., Belsley, Kuh & Welsch, Regression Diagnostics (1980).

67. Id.

68. Specifically, it is the change in the beta divided by the standard error of the coefficient calculated when the observation in question has been dropped from the regression.
that one would make for the dependent variable, given the values of each of the independent variables for the observation in question. A large change in a predicted value, for example greater than 1.0, indicates that the omitted observation has a substantial influence on the prediction made by the multiple regression model. This would be important, for example, if the regression model were to be used to estimate damages in a wage discrimination case. Once again, if individual observations are influential, results both with and without the observations should be reported.

Observations that are influential with respect to the betas are often influential with respect to the predicted values, but they need not be, since the former focuses on the effect of an individual observation on a single coefficient, while the latter measures the effect of the observation on the net effect of all of the coefficients.

2. Sensitivity to Specification Error—Omitted Variables. — All multiple regression studies are built on a theory that describes the form that the multiple regression study ought to take. Failure to develop the proper theory, failure to use variables that are appropriate for the multiple regression study, and failure to use alternative procedures when they are more appropriate can substantially bias the multiple regression results. In a study of the relationship between wage rates and sex, for example, one would want to investigate and be prepared to rebut other possible causes of the observed wage differentials. The fact that the regression results are sensitive to errors in the specification of the regression model dictates a careful scrutiny of the model, and in certain cases may lead to the rejection of the regression approach altogether.69

In a wage study, as in most econometric applications in the courtroom, the choice of explanatory variables is to some extent both arbitrary (in that the theory does not tell the researcher exactly what variables ought to be placed in the equation) and limited (in that some variables cannot be accurately measured). The usual procedure followed by experts is to include in the wage equation any relevant measurable variable (to the extent that these can be identified—and the measurements obtained) that controls for nonsex-related productivity differences between men and women. In addition to age and years of

experience, one might include a variable measuring total hours worked, a variable to control for differences in labor union or bargaining status, and possibly a series of variables to control for employee scores on interviews when hired, or in performance and promotion reviews. A good theory need not have a large number of explanatory variables. The list above could usually be pared down after discussions with the employer or personnel director.

The large number of possible explanatory or control variables available tempts statisticians to run a large number of regressions. It can be dangerous, however, if the expert uses this approach to search for the best coefficient that he or she can find. "Best" might mean a high t-value on the sex variable if the expert is working for the plaintiff, and a low value if she is working for the defendant. After a search of this kind, reported t-values are no longer good indicators, since they are likely to misstate the likelihood of discrimination.

One relatively simple solution to this problem is to require experts to report, as a standard practice, both the nature of the experimentation carried out and the sensitivity of the results to that experimentation. A somewhat more formal procedure would require two steps. First, the expert would divide the control variables into two groups: those that "should" enter the equation because most experts would include them and because they are measurable; and those that are of questionable explanatory value or especially difficult to measure accurately. Next, the expert would run a series of multiple regressions in which all agreed-upon variables were included or excluded in all possible combinations. The expert would then report the range of coefficient estimates obtained for the variable, for example, the sex variable in a discrimination suit. Requiring or expecting the latter procedure would force the expert to seriously evaluate the nature of the variables to be included in the study and would reduce the incentive to search among a large number of possible models before reporting regression results.

Omission of a relevant variable from a regression study should cause serious concern. Defendant's ability to find an omitted variable,


71. For a more formal version of this approach that accounts systematically for all model searching options, see Leamer & Leonard, Reporting the Fragility of Regression Estimates, 65 Rev. Econ. & Statistics 306 (1983); accord E. Leamer, Specification Searches (1978); Leamer, Sets of Posterior Means with Bounded Variance Priors, 50 Econometrica 725 (1982).
however, should not be dispositive. The omitted variable must be sufficiently highly correlated with the variable involved in the hypothesis test—for example, sex or race—so as to show that the test offered by the plaintiff is substantially biased.72

3. Sensitivity to Specification Error: Functional Form. — Significance tests assume a correct specification. Even if the proper set of variables is included in the multiple regression equation, however, the expert must be careful to allow for both non-linearities and interactions. In wage cases it has become standard practice to include both experience and experience squared as explanatory variables, since the effect of experience on the dependent variable is often non-linear; for example, an employee’s wages may rise rapidly at first and then level off after a number of years on the job. Failure to account for this type of non-linearity can lead to either overstatement or understatement of both the size and significance of the sex variable coefficient.73 It is generally good practice to include non-linearities when they are thought to be relevant. When in doubt, however, the expert should report the robustness of the results.

An interaction variable is the product of two other variables that are included in the multiple regression equation. It allows the expert to take into account the possibility that the effect of a change in one variable on the variable to be explained may not be constant; rather it may change as the level of another explanatory variable changes. The application can best be seen if we consider the following wage equation.

\[ W = \beta_1 + \beta_2 S + \beta_3 E + \beta_4 A + \beta_5 AS + e \]

where \( A \) is equal to 1 for those workers over 40 and 0 for workers 40 and under, \( S \) is equal to 1 for females and 0 for males, and \( e \) is a random error term.

The coefficient \( \beta_2 \) measures the difference in average wage rates between women under the age of forty and men under the age of forty. The coefficient sum \( \beta_2 + \beta_5 \) measures the average difference in wages between men and women over the age of forty. The inclusion of the interaction term allows the expert to test for sex discrimination by age (or age discrimination by sex). A significant negative coefficient on the \( S \) variable suggests that younger women are discriminated against.

---

72. For example, in Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224, 306-07, 310 (N.D. Tex. 1980), the court did not accept the argument that blacks and women possessed lower potential productivity because of certain unmeasurable productivity influences such as career motivation and supervisory capability.

while a significant sum of the coefficients on the $S$ variable and the interaction variable suggests that older men are discriminated against. Either type of discrimination can occur (or both can occur), but what is especially important for the multiple regression approach is that failure to account for the interactive nature of the discrimination may lead to a false conclusion that there is no discrimination.\textsuperscript{74}

Since there are often a number of possible interactions, it is once again important for the researcher to make every effort to develop a theory that suggests the correct specification of the multiple regression model. Discussions with the employer during the discovery process might suggest a specific form in which the alleged discrimination might occur. Absent this theory, however, the expert should report a number of alternative model specifications along with the robustness of the regression results. The procedures outlined previously would be useful in this regard.\textsuperscript{75}

\textsuperscript{74} Some individuals might argue that the limited nature of the discrimination theory makes it difficult for the proper interactions to be specified in a multiple regression context, or more generally, for the correct functional relationship to be chosen. As a result, labor economists and econometricians have begun to seriously evaluate alternatives to the approach that involves a multiple regression model of wage on productivity variables and a sex variable. David Peterson, for example, advocates the use of cohort analysis, in which mean wage differentials are calculated for individuals of similar skills and experience. These mean differentials are then averaged over the entire sample to obtain a grand mean, which can be compared statistically to a mean of zero to see whether there is significant support for the hypothesis that discrimination is present. See Measuring Employment Opportunity Equality: Cohort Analysis v. Regression, Personnel Research Report, Jan. 1983. In cohort analysis, results are often similar to the multiple regression results, but do not require such strong assumptions about the exact form of the interactions present.

Other related approaches that have received some attention recently are urn models and reverse regression. See Conway & Roberts, Reverse Regression, Fairness, and Employment Discrimination, 1 J. Bus. & Econ. Statistics 75 (1983); Levin & Robbins, Urn Models for Regression Analysis, with Applications to Employment Discrimination Studies, 46 Law & Contemp. Probs., Autumn 1983, at 247.


The courts have been concerned with the fact that productivity variables may themselves be affected by discriminatory behavior. In James v. Stockham Valves & Fittings Co., 559 F.2d 310, 332 (5th Cir. 1977), cert. denied, 434 U.S. 1034 (1978), the court found that skill level and merit ratings incorporated discriminatory behavior, whereas in Mecklenburg v. Montana State Bd. of Regents, 13 Empl. Prac. Dec. (CCH) 11,438 (D. Mont. Feb. 17, 1976), promotion and tenure variables were held to be based on discriminatory practices. For additional discussion, see Boardman & Vining, supra note 42, at 209.

\textsuperscript{75} As an alternative to the inclusion of interactions, provided that sufficient data are available, it would be appropriate to run a separate regression for each relevant group, e.g., young workers and older workers. This procedure allows for all possible interactions between age and other variables in the equation. For further details, see Fisher, supra note 4, at 724; R. Pindyck & D. Rubinfeld, supra note 2, ch. 5. Some
4. Nonexogeneity of the Explanatory Variables. — In the multiple regression model each of the explanatory variables is assumed to be independent of the dependent variable in the sense that a change in the explanatory variable would affect the dependent variable, but a change in the dependent variable would not affect the explanatory variable. Statisticians have a number of ways of naming this assumption, including exogeneity of the explanatory variables, lack of feedback, and absence of simultaneity. Whatever it is called, the assumption allows one to conclude that any correlation between an explanatory variable and the dependent variable is due to the effect of the former on the latter and not vice versa. Were this not the case, spurious correlation might cause the expert and the trier of fact to reach the wrong conclusion.

The exogeneity issue arises occasionally in multiple regression litigation situations because it is possible for the defendant (if responsible for price fixing or discrimination) to affect the values of the explanatory variables and thus to bias the usual t-tests that are used in the multiple regression approach. An example should help to make the point clear.

Consider an employment situation in which one of the explanatory variables is the score attained by applicants during the interviewing process. If interview scores accurately measure ability to perform on the job, then the multiple regression procedure for testing for wage discrimination is appropriate. If the employer scores women lower than men for inappropriate reasons, however, then the interview score variable will become a good predictor of the wage, and the sex variable will become less insignificant.76

When exogeneity is in question, three approaches are possible, each of which has its limitations. First, the expert can drop the questionable variable from the regression to see whether the variable's exclusion makes a difference. If it does not, then the exogeneity issue becomes moot. If there is a difference, then dropping the variable (in the wage example) biases the case towards the plaintiff, since the interview grades may be gender and race neutral.

Second, the expert can expand the multiple regression model by adding one or more equations that explain the feedback relationship between the explanatory variable in question and the dependent varia-

---

76. If women who score low are not hired, then an analysis of only those hired may be biased. Selection bias of this type is a special example of the simultaneity problem that I have discussed.
ble. For example, one could develop information that illustrates how the employer adjusted test scores by penalizing women. This information could potentially be translated into an equation which allowed for a statistical test of whether test scores are a function of the worker's wage and sex.

Third, one might conclude that the biases and risks associated with the multiple regression approach were simply too great. In this case, statistical or other alternatives would have to be tried. Econometric techniques cannot inform all problems, and experts should be willing to forego the econometric approach when it is not promising.

5. Specification Error: The Inclusion of Irrelevant or Inappropriate Variables. — In multiple regression analysis the inclusion of irrelevant variables, variables that have no effect on the dependent variable, does not bias statistical tests involving appropriate variables in the regression.\textsuperscript{77} Thus, the trier of fact should not reject out of hand a multiple regression study that contains a number of explanatory variables whose relevance is questionable. The major effect of the inclusion of irrelevant variables is to reduce the statistical significance of the regression results;\textsuperscript{78} in other words, the inclusion lowers $t$-values. If the variable that is included, however, is inappropriate for theoretical reasons but does have an effect on the dependent variable, then the problem can be more serious. As will measures of goodness-of-fit, $t$-values may rise, and the trier of fact may reach inappropriate conclusions from the statistical analysis. The courts should be wary of the practice of including a large number of variables solely to overfit the data.

The trier of fact should be concerned, however, when an inappropriate variable is included in the regression. An inappropriate variable is one which (a) has an effect on the dependent variable, and (b) is correlated with the variable whose effect on the dependent variable is of concern to the court.\textsuperscript{79} Building on the previous discussion of exogeneity, consider the use of a skill rating in a wage equation, a rating that is determined in part by the current job classification of the worker.\textsuperscript{80} If the form of the alleged discrimination is to exclude women from the job classification, then inclusion of job classification as a control variable will make it appear that there is no discrimination when the usual significance tests are applied.\textsuperscript{81}

\textsuperscript{77} See, e.g., R. Pindyck & D. Rubinfeld, supra note 2, ch. 6.

\textsuperscript{78} Of course, a strategy of including irrelevant variables might be of advantage to a defendant seeking to disprove discrimination.

\textsuperscript{79} The term "inappropriate variables" has recently been applied in the employment discrimination area. See Finkelstein, The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases, 80 Colum. L. Rev. 737, 738 (1980). The discussion that follows relies heavily on Finkelstein's article.


\textsuperscript{81} See, e.g., Stastny v. Southern Bell Tel. & Tel. Co., 458 F. Supp. 314, 324 n.3 (W.D.N.C. 1978) (number of years of schooling was deemed to be a male-biased varia-
Clearly inappropriate variables ought not be included in the regression study. The appropriateness of including a variable for marital status, for example, might be at issue in certain cases. It could be true that married workers, especially women, have more erratic work patterns and therefore lower productivity. Better measures of productivity should be used if available; otherwise, marital status may be an adequate "proxy" for unmeasurable variables. On the other hand, marital status at a particular job level could be highly correlated with sex, or at least interactive with sex, and could be a means by which the employer discriminates against women, even those whose productivity is not low. The reporting of econometric analyses to the courts is sufficiently new that it would be presumptuous of me to prescribe specific procedures that should be followed. It is my hope, however, that presentations and analyses will be done in the spirit of the suggestions that I have made.

IV. An Application: The Ampicillin Litigation

The econometric issues surrounding the use of hypothesis testing are both numerous and complex. The following discussion of one particular case involving the drug ampicillin should put some of the major points in context and thus provide some additional explanations and applications.

A. Introduction to the Ampicillin Litigation

The plaintiffs in the case were a substantial number of cities, counties, and states (CCS) that purchased ampicillin, a semisynthetic penicillin developed and patented by Beecham, a British drug firm. For a discussion of certain pretrial issues, see In re Ampicillin Antitrust Litig., 88 F.R.D. 174 (D.D.C. 1980). After several years of pretrial activities, trial of the case began in May 1981. Near the end of plaintiff's case, a settlement was reached between the parties. See In re Ampicillin Antitrust Litig., 526 F. Supp. 494 (D.D.C. 1981) (discussing court's earlier approval of settlement).
The defendant, Bristol-Myer (Bristol), received from Beecham in 1959 an exclusive license for the manufacture and sale of ampicillin in the United States. In addition, Bristol received the exclusive right to sublicense in the United States, and Beecham agreed not to market in the United States bulk powder that would be available for finished production. Plaintiffs claimed that these restrictions and licensing practices had caused bulk to be unavailable to "generic houses" and others (wholesalers) who they alleged would have entered the market and sold ampicillin at prices lower than those charged by Bristol, Beecham, Beecham's industrial customers, and Bristol's industrial customers. Plaintiff alleged that exclusion of the competition with generic houses caused the plaintiffs to pay higher prices for ampicillin than they otherwise would have paid. Such conduct was said to be illegal under the Sherman Act.

The CCS entities all purchased ampicillin through a sealed-bid competitive bidding process. Over the relevant time period, the winning bid prices tended to fall, a pattern consistent with both an increasing number of firms bidding and a steady decline in the cost of producing ampicillin. The central issue in the case was whether, but for the license agreement between Bristol and Beecham, the prices paid by plaintiffs for ampicillin would have fallen faster and thus been lower than they actually were.

B. The Plaintiff's Case

The plaintiff's expert testified that the price of ampicillin sold in the CCS market from 1965 to 1972 was "too high." There are several assumptions underlying this argument that can be tested empirically, but I will concentrate on one issue. Does the introduction of new firms into the CCS ampicillin market cause the price of ampicillin to fall? The answer could well be yes if: (1) new firms win bids, thus having a direct effect on price; or (2) new entrants win no bids, but manufacturers currently in the market lower their bids to fight off the competition from the new firms, thus lowering prices.

Of course, the mere possibility of such a causal link is not itself evidence. If no new firms win bids and manufacturers do not respond

---

86. Bristol did license some others to produce the finished powder, but refused to allow its licensees to sublicense.

87. It is useful to think of the sellers of ampicillin in five groups: (1) drug manufacturers who both make and sell ampicillin (this group includes Bristol and Beecham); (2) major drug manufacturers who make other drugs, but choose to buy their ampicillin in bulk quantities from the manufacturer; (3) foreign manufacturers whose ampicillin production may find its way into the United States despite the fact that they do not hold patent licenses under United States patent law; (4) generic houses that buy ampicillin either in bulk or in bulk quantities of finished products, and typically seek to resell it in unbranded form to the retail/wholesale market for filling generic prescriptions; and (5) wholesalers who may take over the distribution from any seller of finished products by buying in large quantities and reselling to drug stores in smaller quantities.
to the new firms in choosing their bidding strategies, then any correlation between number of firms and price may be spurious, due to a failure to correctly account for important explanatory variables that were omitted from the study.

C. The Use of Hypothesis Testing in the Ampicillin Case

The plaintiff's hypothesis was that the presence of generic houses in the market results in a lower price; the defendant's hypothesis was that the presence of generic houses had no effect on price. The two hypotheses are mutually exclusive. The hypothesis of no effect is a clear, well defined hypothesis, suggesting no violation of the law and, of course, no impact or damages.88 If there is an effect, there may or may not be cognizable impact on any particular plaintiff or class thereof. The alternative hypothesis that there is some effect leaves open the questions of whether a particular plaintiff was injured and of the correct measure of damages.

In analyzing the facts of the ampicillin case, a multiple regression approach to hypothesis testing is appropriate. The first step is to use one's underlying understanding of economics and of the institutions of the ampicillin market to model the factors expected to determine the price of ampicillin in the CCS market, in other words, to specify the dependent and independent variables in the regression model. The scope of explanatory variables could be categorized into four types:

1. **Cost Variables (COST).** Cost and price are expected to be positively correlated, that is, higher costs are expected to lead to higher prices in the market.
   a. \( C \) = a cost index based on the average cost of production of ampicillin (by Bristol-Myers).
   b. \( QD \) = a variable to reflect the higher costs associated with selling ampicillin in low volume, or quantity discount. It will be equal to one if the bid involves less than one hundred bottles, zero otherwise.

2. **Measure of Competition Variables (COMP).** This reflects the competition for sales by other firms in the market.89
   a. \( NMAN \) = number of manufacturing firms that bid.
   b. \( NBRCUS \) = number of Bristol-Myers customers that bid.90
   c. \( NBECUS \) = number of Beecham customers that bid.

88. Violation is obviously a matter of law, not economics or econometrics. For the purposes of this application, "no effect" will be equivalent to "no violation," while some effect is at least consistent with a cognizable violation.
89. There were no close substitutes for ampicillin during this time period. If there were, a variable reflecting the availability of these substitutes would have to be included in the analysis.
90. These companies purchased bulk ampicillin from Bristol and resold the final product to the CCS entities. None of these companies were generic houses.
d. \( NW \) = number of wholesalers (firms that bought bulk ampicillin from nonmanufacturers).

(3) Time Adjustment Variables (TIME). Each of these time variables is a variable which allows for differences in price that occur over the years listed.

a. \( T67T068 = 1 \) in 1967 or 1968; 0 otherwise.
b. \( T69T070 = 1 \) in 1969 or 1970; 0 otherwise.
c. \( T71T072 = 1 \) in 1971 or 1972; 0 otherwise.
d. \( T73T077 = 1 \) from 1973 to 1977; 0 otherwise.

These variables are introduced to account for cost differences that were not reflected in the crude cost index cited previously and for other structural changes in the ampicillin market that might have occurred.

(4) Generic House Competition Variable.

\( N \) = number of generic houses that bid.

The theory that generates the multiple regression model is the economist's theory of supply and demand. The cost variables are all determinants of supply, while the measures of competition, including the generic house variable, are all determinants of demand. The time variables can account for either supply or demand determinants that are not otherwise specified. Any attempt to distinguish supply and demand characteristics separately would be difficult.\(^9\) All that is important for my purposes, however, is to account for the fact that the price of ampicillin will be a function of all determinants of both supply and demand. The model, in its multiple regression format, is summarized below:

\[
P = \beta_1 + \beta_2 N + \beta_3 \text{COST} + \beta_4 \text{COMP} + \beta_5 \text{TIME} + \epsilon,
\]

where \( P \) is the price of ampicillin and \( \epsilon \) is a random disturbance that accounts for the effects of omitted variables.

The null hypothesis of no effect with respect to generic house competition is the hypothesis that \( \beta_2 = 0 \), while one alternative hypothesis might be that \( \beta_2 \) is not equal to zero. Because the plaintiffs' hypothesis was that more firms bidding would have caused the price to fall, a more accurate formulation of the alternative hypothesis is that \( \beta_2 \) is less than zero. Even this alternative is rather vague. I will reconsider the form of the alternative hypothesis after the multiple regression results have been described.

To test the null hypothesis, data for nine states and eighty bidding situations, from 1967 to 1977, are used to estimate the multiple regression model described above. The statistical results are presented in Table 1.

A good first step is to examine the regression results to see whether they appear sensible. First, the cost index coefficient is posi-
Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Part A</th>
<th>Part B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Standard Error</td>
</tr>
<tr>
<td>$N$</td>
<td>0.06</td>
<td>0.09</td>
</tr>
<tr>
<td>$C$</td>
<td>3.45</td>
<td>0.25</td>
</tr>
<tr>
<td>$C^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$QD$</td>
<td>1.46</td>
<td>0.40</td>
</tr>
<tr>
<td>$NMAN$</td>
<td>-0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>$NBRCUS$</td>
<td>-0.27</td>
<td>0.14</td>
</tr>
<tr>
<td>$NBECUS$</td>
<td>0.23</td>
<td>0.29</td>
</tr>
<tr>
<td>$NW$</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>$T69T070$</td>
<td>-0.79</td>
<td>0.41</td>
</tr>
<tr>
<td>$T71T072$</td>
<td>-0.22</td>
<td>0.51</td>
</tr>
<tr>
<td>$T73T077$</td>
<td>-0.12</td>
<td>0.54</td>
</tr>
<tr>
<td>$CONSTANT$</td>
<td>-9.73</td>
<td>1.61</td>
</tr>
</tbody>
</table>

$R^2 = 0.96$ \hspace{1cm} $SE = 0.76$ \hspace{1cm} $R^2 = 0.97$ \hspace{1cm} $SE = 0.64$

tive and significant, indicating that increases in cost are reflected in higher prices. The fact that the cost coefficient has a very low standard error relative to its size suggests that cost is the most important predictor of price.

Similarly, the coefficient of the quantity discount $QD$ variable is positive and significant at the five percent level, indicating that bids on very small volume sales are generally higher than average.

Of the competition variables other than the number of generic houses, $NMAN$ and $NBRCUS$ are negative, as expected; however, neither is statistically significant. The insignificance here is partly due to the multicollinearity which arises because $NMAN$ and $NBRCUS$ are relatively highly correlated. If $NMAN$ were dropped from the sample, for example, $NBRCUS$ would become significant. Thus, to the extent that competition determines price, the competition comes from manufacturers and from customers of the defendant in the case, Bristol-Myers.

Now consider the null hypothesis that generic houses had no effect

92. Multicollinearity arises when two or more explanatory variables are highly correlated. It becomes difficult to distinguish the effect of one variable on the dependent variable from the effect of the other, even though each may be an important predictor of that dependent variable. For this reason, experts and the courts must be careful in evaluating regression results to realize that (a) insignificant coefficients do not always imply that variables are unimportant, and (b) a wrong sign for a coefficient may not indicate that the model is poor. If two explanatory variables that have a positive effect on price are highly correlated, the net effect of both variables will be positive, but the individually measured effects could be either two positive, but small coefficients, or one large positive coefficient and one small negative coefficient (assuming that the units of measurement are comparable). For further analysis of the multicollinearity issue, see R. Pindyck & D. Rubinfeld, supra note 2, ch. 4.

93. The simple correlation between $NMAN$ and $NBRCUS$ is 0.54, high enough to lower the reliability of individual coefficient estimates somewhat, but low enough to avoid any serious problems in computation.
on price. The coefficient on $N$ is .06, suggesting that, other things being equal, an additional generic house bidding adds six cents to the average winning bid price.\textsuperscript{94} The six cent coefficient, however, is highly insignificant, that is, not statistically different from zero (the $t$-statistic is only .67). In statistical terms, an appropriate way to look at the ampicillin results is to state that the null hypothesis that generic houses have no effect on price cannot be rejected. Moreover, given the positive sign of $N$, these data provide no support for the view that generic houses have a negative impact on price in the CCS market.

In addition to examining the signs and the significance of the coefficients of the regression results in Part A of Table 1, one should also examine the magnitudes of the coefficients. In particular, one should ask whether the magnitudes make sense in light of the problem being studied. As an example, consider the coefficient on the cost variable, 3.45. Taken literally, this coefficient suggests that an additional $1.00 of cost will lead to an increase in price of $3.45. On its face this is clearly a surprising result, and it suggests that there may be a problem with the regression results.

A sense of what the problem might be can be gleaned by examining the coefficients of each of the time dummy variables. The value $- .79$ on the dummy $T69T070$ implies that from 1967–68 to 1969–70, price fell on average by seventy-nine cents more than the change accounted for by other variables. The smaller negative coefficient on $T71T072$, however, tells us that, other things being equal, price rose from 1969–70 to 1971–72, and again from 1971–72 to 1972–73. Costs of production were generally falling throughout the period of analysis, and, in particular, costs were higher in 1967–68 than at any later period of time. Thus, the pattern of the coefficients on the time variables may be picking up some of the effect of the decline in costs over time, but other effects appear to be present as well. If costs had been declining continuously at the same rate each year and there were no other factors, the cost variable in the equation would have a negative coefficient and the time variables would show no effect. Therefore, since a cost term is already included in the regression, the relationship between cost and price appears to be more complex, with the effect of cost on price varying with the level of cost, rather than remaining constant.

Of course, one cannot be sure that cost is the only misspecified variable.\textsuperscript{95} It seems likely that other variables related to the product cycle of the drug are also important and yet omitted from the model. A

\textsuperscript{94} The correlation between $N$ and the other computation variables is never greater than .55. Dropping any one of the other variables does not change the sign of the $N$ coefficient.

\textsuperscript{95} One of the dangers of using time variables is that the expert cannot be sure exactly which of the effects of omitted variables is being picked up by the time variables. It is almost always desirable to measure variables directly, rather than to substitute time variables as proxies.

\[1083\]
reexamination of the complex nature of the effect of cost on price can help to clarify the situation. I reestimate the regression model, including an additional variable, $C^2$, the square of the cost variable, as shown in Part B of Table 1. In this formulation, the marginal effect of an increase in cost on price is not constant. Since the coefficient of $C^2$ is positive, the effect of cost on price increases at higher costs. Note also that the time variables are negative, increasing, and significant, suggesting that, other things accounted for, price falls continuously from the first to the last time period. The omitted relationship between prices in the CCS market and other related markets, including the acceptability of generic products over time, helps to explain why prices declined at a greater rate than costs. This regression also makes the interpretation of the effect of cost on price more reasonable. For example, in the 1967–68 period, when costs were relatively high, a $1.00 increase in cost had an effect on price that was just under $1.00. Subsequently, the effect of cost on price decreased.

Finally, notice that the equation used in Part B of Table 1 alters the magnitudes of the other coefficients in the regression model. It is interesting to note that the coefficient of the generic house variable is now .01 and once again insignificant. In terms of the question of violation, the conclusion that the generic houses do not affect price is unchanged and, if anything, strengthened.

Before any conclusions can be reached, however, it is important to evaluate the robustness of the empirical results with respect to both the specification of the model and the choice of sample. In the ampicillin case analysis, the number of measurable explanatory variables is quite limited, and the steps necessary to evaluate robustness are reasonable. The coefficient and $t$-statistic on the variable of concern, $N$, is quite insensitive to specification. A negative coefficient or a significant $t$-statistic is not obtained whatever the reasonable choice of variables, and whatever the functional form. Were contrary results obtained in some cases, it would be advisable to report these other results and to argue, if appropriate, why the specification chosen is preferable to those that give conflicting results.

Because the number of possible explanatory variables is small and the results quite robust, a formal model sensitivity procedure is not used. If the model were substantially more complex, however, a more explicit approach to allow for selection of alternative model specifications would be advisable.

To illustrate the kinds of procedures that are available, I have pursued the question of the robustness of the sample more thoroughly. First, I printed a list of the individual residuals of the regression re-

96. When $P$ is regressed on $N$ and $C$, with all other variables omitted, a negative coefficient on $N$ is obtained. This specification, however, is clearly inappropriate since it omits the competition from the manufacturing firms.

97. For further details and discussion, see supra note 79 and accompanying text.

98. I do not believe that the problem of bias in choice of the case to be litigated is a
ported in Part B of Table 1, and compared those residuals to the standard error or the regression, .76.\textsuperscript{99} I found that only two of those residuals were greater than twice the standard error, a result that is consistent with the assumption of normality of the error term.\textsuperscript{100} Then I listed the values of the change in the beta or coefficient of the variable \textit{N} as each of the observations were omitted in turn. In the usual robustness test these values are scaled to the appropriate standard error,\textsuperscript{101} so that observations with values greater than two in absolute value are of particular concern. I found no such troublesome values in my analysis,\textsuperscript{102} and concluded, therefore, that the study was quite robust with respect to individual data points in the sample.\textsuperscript{103}

D. Specifying the Alternative Hypothesis

The ampicillin case is a valuable example with which to consider the question of how the alternative hypothesis should be specified. The outcome of a test of the null hypothesis would be different depending on whether the alternative hypothesis was that an additional generic house would lower the price by fifty cents or more, or that an additional generic house would have an impact equal to or greater than one dollar.

To begin with, it is certainly possible that if a violation did occur and if price had actually been trivially lower, for example one cent, the sample might have shown a coefficient of \(+.06\). At greater effects involving a lowering of price by five cents, ten cents, twenty cents, or forty cents, however, the likelihood of obtaining a positive coefficient of problem here, since the case focused broadly on the bidding for ampicillin throughout the CCS market over an extended period of time.

\textsuperscript{99} The standard error of the regression provides a measure of the mean magnitude of the regression residuals (deviations of actual from predicted values). The standard error of .76 is approximately 13\% of the mean price of \$5.87, suggesting a reasonable fitting equation. When I examined the studentized residuals, see supra note 65, I found that the same two observations had residuals which were significantly different from zero, and that a third observation (with a residual of 1.52) also was different from zero. Once again, this result is consistent with a normally distributed error term.

\textsuperscript{100} With a normal distribution, five percent, or about four observations, would be greater than twice the standard error.

\textsuperscript{101} Specifically, the scaled change in the beta is equal to the change in beta, divided by the standard error of the coefficient. The standard error is calculated when the observation in question has been omitted from the regression.

\textsuperscript{102} I did find one observation with a change in beta equal to 1.27. This suggests that removal of this particular observation could alter the coefficient on the generic house variable by somewhat greater than one standard error. It would not alter the conclusions of my analysis, however. Note in addition, that the observation which is influential in terms of the sensitivity of the coefficient on the \textit{N} variable is not one of the observations that generated a large residual. This suggests why a failure to find large residuals does not guarantee robust regression results.

\textsuperscript{103} Had I been interested in prediction or forecasting, I would have also calculated the change in the predicted values of price when each observation is omitted from the regression of the predicting equation.
.06 falls. Specific calculations can show how the probability of mistakenly failing to find an effect depends upon the specification of what such an effect might involve.

How often might a coefficient of +.06 or larger occur if the "true" effect were negative? The answer depends on how negative an effect we are concerned with. I calculated that if the price effect of an additional generic house is assumed to be negative five cents, then the probability of getting a coefficient on \( N \) of +.06 or more is .12. This probability falls to .04 when the price effect is assumed to be negative ten cents, and falls to .01 when the price effect is negative fifteen cents. Clearly the larger the true negative effect, the less likely a positive .06 result. Thus, while the small positive coefficient shows that a negative price effect is not likely, it cannot rule out the possibility of a small effect. The .06 coefficient does, however, make even a fifteen cent negative effect extremely improbable.

Where does all this lead? In a suit where the burden of persuasion is on the plaintiff, rejection of the null hypothesis embodying the defense's position of no effect provides strong support for the plaintiff's case. Even then, a conclusion about the case is still uncertain because, as we have seen, rejection of a null hypothesis depends on the level of significance chosen, and on the realism and legal relevance of the regression model, as well as on the quantity and quality of data available.

Absent other considerations, failure to reject the null hypothesis provides support for the defense. Failure to reject the null hypothesis, however, does not completely rule out the possibility that plaintiff's claim is correct. Nothing is ever certain, but some things are highly improbable and statistical analysis can indicate those probabilities.

It is difficult to suggest specific alternative hypotheses without reference to individual cases, but one approach is to select a de minimis impact as the standard for determining violation. A showing of such an effect or of a more substantial one will support a plaintiff's case. In the ampicillin example, one could decide that a negative impact of five percent of price, roughly a lowering of price by twenty-five cents, is de minimis. The plaintiffs then would be required to show such impact in order to prove their case. If a five percent impact were considered de minimis, one could conclude definitively that there was no violation in the ampicillin case. A second approach would be to calculate the level of impact, \( X \), above which one could be quite sure (say with a probability of ninety-five percent) that the conclusion of no violation was correct. \( X \) turns out to be eleven cents in the ampicillin equation. Both approaches are closely related since with each approach plaintiffs

104. An employer in a sex discrimination case, for example, may well concede that the null hypothesis is incorrect and argue instead that wage disparity by sex can be accounted for by other factors, such as high performance on a test or a good attendance record.
ECONOMETRICS

and defendants would devote substantial effort to arguing the question of what is a de minimis effect.

V. FORECASTING AND SIMULATION

Multiple regression methods are particularly well suited to answering "but for" questions because they can be utilized for both forecasting and simulation purposes. This section briefly reviews some of the possible applications of multiple regression methods, while highlighting some of the conceptual and practical problems. I rely on a plywood industry case and on the ampicillin example to develop the forecasting and simulation arguments.

A. Econometric Forecasting: Testing for Effect and Measuring Damages

The forecast obtained from a single equation multiple regression model, among all forecasts from single equation models, generally will provide the best, or most accurate forecast of damages. In a price-fixing case, for example, a regression model that explains price in a period of nonviolation can be used to predict what the price would have been during the period of violation. Unfortunately, it is not unusual for a regression model that allows for the rejection of the null hypothesis with substantial confidence to generate forecasts of effects that are not very reliable. An important reason is that the standard error of forecast is calculated on the assumption that the model is correctly specified. Thus, if any of the specification problems suggested in Section III arise, the true forecast error may be even larger than it appears. The importance of reporting information about forecast reliability is illustrated in the following example.

1. The Plywood Litigation. — One of the primary issues in the plywood antitrust litigation was whether the growth and development of the southern plywood industry, in coordination with an allegedly conspiratorial method of quoting delivered prices, served to raise ply-

---

105. The standard error of forecast is equal to the standard error of the regression multiplied by a factor that is equal to one when the forecast is made at the mean of each of the explanatory variables, and that otherwise is greater than one. See R. Pindyck & D. Rubinfeld, supra note 2, ch. 6, for details. It is not unusual for a model to have a relatively low $R^2$, and therefore a low standard error of the regression, while also having very significant $t$-statistics. This could happen if, for example, the sample is quite large. The courts have often been too willing to accept $R^2$ as a satisfactory measure of goodness of fit. See, e.g., Wilkins v. University of Houston, 654 F.2d 388 (5th Cir. 1981); Lewis v. Bloomsburg Mills, Inc., 30 Fair Empl. Prac. Cas. (BNA) 1715 (D.S.C. 1982); Valentino v. United States Postal Serv., 511 F. Supp. 917, 944 (D.D.C. 1981), aff’d, 674 F.2d 56 (D.C. Cir. 1982).

106. This includes the correct choice of explanatory variables and the functional form of the model.

wood prices in the South higher than they would have been had the market been competitive. Using a forecasting technique helps settle this issue. One begins by constructing an economic model to explain the movement of plywood prices during the period of the alleged conspiracy. The model is then used to predict what the conspiratorial prices would have been in a period in which no conspiracy was alleged, or when all parties agreed that prices were competitive. If predicted conspiratorial prices, after controlling for relevant variables during the nonconspiratorial period, are substantially higher than actual competitive prices, the analysis supports the theory that the conspiracy raised prices, and suggests a measure of damages as well.\(^{108}\) On the other hand, if there is little or no systematic excess of predicted prices over actual prices, the analysis supports the defendant's position that there is no harm.\(^ {109}\)

In the regression study for the plywood case, the price of one-half inch standard exterior pine plywood was chosen as the dependent variable. The explanatory variables (in the "reduced form" regression) included those related to demand (income and housing starts in the region) and supply (stumpage price, forest sales, a production capacity measure, log costs, drying costs, and glue costs).\(^ {110}\) There is no reason to believe that a cartel of plywood producers, if it existed, could have controlled the explanatory variables that have been listed. Housing starts and income are clearly outside the control of such a cartel, and though the possibility of inflated costs should be treated seriously, no suggestion was made in this case that input prices and costs of production were controlled by the cartel.

Least-squares analysis of these regression results suggests that the model fits the data quite well. In addition, many of the individual variables are statistically different from zero at the five percent level.\(^ {111}\) A further way to examine goodness-of-fit in this context is to graph the relationship between the actual plywood prices and the prices predicted

\(^{108}\) Questions of violation are inherently legal and cannot be answered solely on the basis of econometric evidence.

\(^{109}\) There is, of course, the question of how many years such an experiment must be carried out. A lengthy experimental period is necessary, since historical factors may serve to maintain higher price levels for a while.

\(^{110}\) In specifying any model, one must consider whether simultaneity is a problem. In a simultaneous model, one would account for the fact that price and quantity are jointly determined, and that neither has an independent effect on the other variable. The equation used here was a "reduced form" equation, rather than a "structural equation," in that quantity did not appear on the right-hand side as an explanatory variable. For most purposes, reduced form equations are preferable for forecasting, although strictly speaking, the estimates obtained for the parameters may not be the most efficient (i.e., accurate). Structural equations can be used if properly estimated, but the use of simultaneous equation techniques complicates the analysis and will be avoided here. For further details on this issue, see Fisher, supra note 4, at 726.

\(^{111}\) The \(R^2\) was slightly higher than .80. Since my concern is with forecasting, I have omitted the discussion of the individual variables.
by the multiple regression model. In Figure 1, the actual prices are represented by a solid line, and the predicted prices by a dotted line. With the exception of the period running from late 1972 to early 1973,

![Figure 1](image)

the predicted prices are quite close to the actual data. The ability to predict or "track" fluctuations in data—turning points—is another useful test of a regression model. Models that predict only growth paths for variables without explaining turning points are poor forecasting models.\textsuperscript{112} The regression model is only modestly successful in predicting turning points.

2. "Forecasting" with the Plywood Model. — Can the results of the regression model be used to test for an antitrust violation? Not directly. The regression model can be used to explain the price behavior of a cartel just as well as it can test that of a competitive firm or firms. A cartel of price-setting firms could also vary price depending upon conditions of supply and demand. Just as an increase in demand for plywood caused by an increase in housing starts will lead to higher prices in a competitive market, it will also allow cartel members to raise price. Of course, the cartel's price will be higher than the competitive price,

---

\textsuperscript{112} Why does the model fare so poorly in the 1972-73 period? The probable answer is that during this period the government controlled prices, thereby restricting the ability of the market to equate demand and supply. The fact that the supply/demand model works badly in a period when supply and demand were constrained is encouraging. Surprisingly, predicted prices fell below actual prices during most of the period of price control. This suggests that the government erred by setting the price of southern plywood too high.
but this observation involves a question of price level, not price variation.

Regression models on the whole are not particularly well suited to explaining levels of variables; rather, they explain variation about the mean. Imagine, for example, two price series requiring explanation, both having the same pattern of variation, but one reflecting monopoly pricing that is always fifty percent higher than nonmonopoly pricing. A regression model would explain both price variables equally well, with the same $R^2$ and the same prediction of turning points. The only difference would be reflected in the intercepts, due to the level differences. Regression lines always fit the data on average—the intercept term calculation assures the fit.

The forecasting technique can overcome this limitation of the regression model. The southern plywood model was estimated through the month of March 1977. March 1977 represents the time when the “price conspiracy,” if it occurred, would have broken up in any event, because at that time the defendant plywood companies altered their form of pricing to quote southern freight rates to those purchasing plywood from southern plants. The estimated coefficients were then used to forecast what the prices would have been from March 1977 through early 1978 had the pricing behavior prior to March 1977 remained in effect. The results of comparing the forecasted prices to actual prices are shown in Figure 2.

Had there been a conspiracy prior to March 1977, predicted prices would be substantially higher than actual prices. Had there been no effective conspiracy, there would have been no systematic relationship between the levels of predicted and actual prices. The results suggest no conspiracy. While predicted prices were higher than actual prices some of the time, they were lower at other times. There is no clear pattern and certainly no distinct decline in price during the period following March 1977.

There is a related but somewhat different way of distinguishing cartel prices from competitive prices econometrically. One would develop a model for the entire time period for which data exist. This estimating model would include an additional variable, a variable equal to one for time periods when the cartel allegedly was in effect and zero during other time periods. The coefficient on the cartel variable then would indicate whether the average price during the cartel’s existence was significantly different from, and in particular, higher than, the average price during the competitive period. Thus, the forecasting ap-

113. The adjustment period to the new pricing method could have been gradual, so it is useful to test for robustness with respect to choice of forecast period.

114. The forecasts used in actual litigation included corrections for the presence of serial correlation and thus are somewhat more sophisticated than implied here. For a discussion of the serial correlation issue, see R. Pindyck & D. Rubinfeld, supra note 2, ch. 7.
approach becomes a hypothesis testing approach. An important assumption is made here, however. One assumes that the overall supply and demand behavior built into the regression can be modeled in precisely the same way during both the conspiratorial and nonconspiratorial periods. In some cases, this may be reasonable, but in others it may not be. The forecasting approach described above, involving predictions beyond the period of estimation, requires no such assumption and therefore is generally more appropriate than the hypothesis testing approach.

In the plywood case, the forecasting test is not entirely satisfactory, once forecast reliability is taken into account. In the regression model, the standard error of forecast was approximately 12.0.\textsuperscript{115} When two standard errors of forecast are drawn on either side of the predicted values in Figure 2, the ninety-five percent confidence band is so large that almost anything is possible. Thus, while the forecasting test adds no support to the conspiracy hypothesis, one cannot say with certainty on the basis of this test that a conspiracy could not have occurred. A

\textsuperscript{115}. This standard error of forecast is actually somewhat higher than the standard error of the regression, but it is not substantially higher.
better test would be provided by a better model or a forecast for a longer period of time.\textsuperscript{116}

B. Using Simulation Techniques to Estimate “But For” Prices: Ampicillin

Occasionally, forecasting methods do not yield “but for” prices, or if they do, the reliability of the results cannot be evaluated. In such cases, simulation can be valuable, as illustrated in the following example.

In the ampicillin case, the impact/damages analysis requires an analysis of three questions, all of which must be answered yes if the court is to decide that damages should be awarded. The questions are:

1. If there had been no contract limitation, would generic houses have obtained bulk ampicillin?
2. If the generic houses had obtained bulk ampicillin, would they have bid in the CCS market?
3. If the generic houses had bid in the CCS market, would their bids have lowered the purchase price of ampicillin?

Whatever the answer to the first two questions, the multiple regression technique used in a simulation context can be used to answer the third question.

Assume arguendo that generic houses would have bid earlier in the CCS market. Would those bids have had any impact on price, and second, what would have been the size of the impact? To decide these questions, one must first recall that there are two ways in which the presence of generic houses could affect price. Generic houses can affect price by actually winning sealed bids at prices lower than the prices submitted by nongeneric houses. Even if generics never win a bid, though, the bids of others in the market might be lower in response to the presence of generics. In either case, one could argue that some impact occurred.

To apply a simulation approach, one must first assume that the behavior of firms in the ampicillin market after 1972 can be used to shed light on what the behavior of those firms would have been before 1972. Costs and prices need not be the same, but the bidding strategies of firms of each type must be the same. This assumption involves two parts: (a) the assumption that the bidding strategy of generic houses would have been the same before 1972 as after it; and (b) the assumption that the strategy of existing firms in response to the presence of generic houses after 1972 would have been the same before 1972, had generic houses been bidding.

Multiple regression analysis can be used to determine whether individual firms, such as Bristol-Myers, select their sealed bids in part in

\textsuperscript{116}. With a long forecast period, one can estimate a new model using only the “postconspiracy period” to see whether it is significantly different from the conspiracy model.
response to the presence of generic houses. The hypothesis of no effect was tested by comparing the coefficient on the variable representing the number of generic houses \((N)\) to its standard error in a regression in which the dependent variable was Bristol's bid price.

The coefficient on the \(N\) variable turns out to be .05, which is not significantly different from zero. This result suggests literally that the addition of a generic house may lead to slightly higher bid prices by Bristol, but a more realistic interpretation is that the addition of one or more generic house bidders has no impact on Bristol's bidding, or on the bidding of other manufacturers or industrial customers.\(^{117}\)

Since Bristol's bids and the bids of other manufacturers were unaffected by the presence of generic houses, the only impact that generics could have had on price was a direct one—by winning a bid. Generic houses actually won no bids after 1972. Nevertheless, to answer the question whether they would have won bids before 1972, had they been more active in the bidding process, a simulation technique can be used.

Roughly speaking, the simulation process is utilized to mimic what the CCS ampicillin market would have been like had generic houses been present to bid. Detailed information describing the bids of all types of firms, manufacturers, industrial customers, wholesalers, and generic houses were taken into account. The bid descriptions were divided into each of four time periods: 1965–68, 1969–70, 1971–72, and 1973–77. Within each time period, bids were adjusted for cost to put them all on a comparable basis. Thus, for each type of firm, I allowed a set of possible bids based on actual bidding experiences.

To simulate what the effect of generic houses would have been in 1965–68 (when in fact none existed), the following procedure was used. First, I randomly chose a typical bidding situation (by sampling from the set of all possible situations), for example, one in which five manufacturers, three industrial customers, and two wholesalers were bidding. For each potential bidder, I then selected (again by random selection) a characteristic bid for each of the firms in the market. The bid was chosen from the set of all bids which firms of that type had made throughout the period of study.\(^{118}\) The mean winning bid in each period over one thousand simulations was then calculated, as were the number of wins by each type of firm.

Would generic houses have had an impact on price? To complete the evaluation of this question, a simulation experiment just like the previous one was run, but with the addition of generic houses to the bidding process.\(^{119}\) This simulation experiment was repeated an additional one thousand times to see both whether there was any change in

---

117. Rubinfeld & Steiner, supra note 2, at 1460–61.
118. The bids were adjusted for differences over time due to changes in cost and price levels. If bids were affected by the presence of other firms, then the simulation process would have been more complex.
119. Since it has been shown that generic houses do not affect the bids of others,
the mean winning price, and whether generic houses ever won a bid. During the relevant time period 1965–68, generic houses never bid in practice and never won a bid in the simulation.

Another issue in the case was the effect of the entry of more wholesalers into the market. Wholesalers were difficult to distinguish from generic houses in the analysis and appeared to be in relatively similar cost situations. After I added two wholesalers to the simulation, in one thousand tries, one wholesaler won a bid. This was surprising, given the relatively high costs associated with wholesaling and the high bids that wholesalers typically make. It turned out that I had randomly selected a period in which no other firms bid, an implausible situation. In any case, the average impact of the resulting increase of a few cents in price, occurring only once in one thousand times, is effectively zero.\(^{120}\)

Once again, the potential value of the econometric approach outlined above can be substantial. Courts, however, must carefully evaluate the results presented to them by experts. A number of assumptions were necessary to estimate Bristol-Myers' bidding strategy, and also to simulate the bidding process. Courts ought to require experts to make their assumptions explicit, to make available the data upon which the analysis is performed, and to report some measures of the accuracy of the technique being used. In the simulation exercise there is a particular danger. Because it is inexpensive to perform a large number of simulations, the usual statistics measuring accuracy, such as the standard deviation, are likely to suggest that the results are extremely significant statistically. Statistical significance is only meaningful, however, given a correct choice of model. In the simulation experiments, an alternative to reporting statistical significance is for the expert to report the outcomes of experiments in which some important, but questionable, assumptions are relaxed. In the ampicillin simulation, this could mean varying the choice of bidding situations, or allowing each firm's bids to depend upon the number of other firms against which it is bidding.

VI. Conclusion

The growth in the use of econometric techniques in the courtroom has been rapid, and there is little reason to believe that the trend will not continue. This growth is a beneficial one in my view, since econometric techniques can be valuable in cases involving complex, essentially empirical issues. I have shown, however, that the expanded use of multiple regression techniques is accompanied by the possibility

---

120. This suggests an additional reason why econometric approaches are difficult to apply to questions of violation in the antitrust area. Regression forecasts involve predictions of average effects. If violations occur, but only in a small percentage of the cases, then the average effect will be small and may not be picked up as significant by the regression approach.
of their misuse, and that substantial normative questions about their use remain unanswered.

The potential problems are of two distinct types. First is a tendency to interpret regression results as being more accurate than they are. Regression results may appear deceptively accurate because the underlying assumptions are seldom made explicit. A second problem is the experts' tendency to analyze data and report in a manner which biases the results for the particular side that they represent. This tendency could reflect an intent to mislead or deceive, but it need not. More likely, it reflects economists' standard practice of searching among alternative models, trying numerous functional forms and variables to find the "best fit," and reporting only the final outcome of the search process.

These problems associated with econometric practice suggest a two-pronged solution. First, experts who testify in the courtroom ought to be encouraged to do what most econometricians now consider to be proper practice—to report the sensitivity of their results to (a) choice of data set; (b) individual observations within the chosen data set; (c) choice of variables that appear in the regression model; (d) choice of functional form; and (e) choice of estimation technique—i.e., is exogeneity correct?

Courts ought to consider a number of alternatives to encourage experts properly to report econometric results. This will make it easier for triers of fact to reach an appropriate conclusion as to the import of the econometric evidence. The first set of alternatives involves the presentation of the statistical study itself. The court should rule, to the extent possible, on the admissibility of data prior to the trial, perhaps while the discovery period is still open.¹²¹ This has the advantage of letting the lawyers and experts know what kind of testimony will be of value during the trial, and will give the opposing side sufficient time to evaluate the data as well.

This important first stage in the analysis could be aided if the court were to appoint a neutral expert or a panel of experts to advise it on the professional nature of the technical material. The role of experts becomes more important, however, when we focus on the trial in which the statistical study is to be presented. The court ought to consider the possibility of using a master or masters to evaluate the technical arguments concerning the econometric issues.¹²² The master need not reach a final conclusion or give specific recommendations. He or she need simply evaluate the strengths and weaknesses of the technical ar-

---


¹²² Fed. R. Evid. 706, which allows for court-appointed nonparty experts, has been little used. See Saks & Van Duizend, The Use of Scientific Evidence in Litigation, National Center for State Courts (1983).
arguments made by the opposing sides. Indeed, the master need not serve as a surrogate for the jury or the judge, whoever is the trier of fact. The master need simply report to the court, and could be subject to cross-examination by both sides if that were deemed desirable. 123

The advantage of using a master or a “neutral” expert is to avoid the bias that can arise when lawyers have the option of searching among potential experts before hiring one whose testimony is likely to be compatible with their approach to a case. If the court can appoint well-qualified experts who are open-minded, some of the technical statistical issues that arise might be settled fairly. The use of an appointed expert might allow the court to focus its attention on basic conceptual questions, rather than on more narrowly defined technical issues.

There are a number of difficult issues surrounding the choice of neutral experts that would need to be considered before a final evaluation was made. But the analysis does not have to start from scratch: the model of private arbitration provides a useful analogy. 124 One possibility would be for both parties to propose a set of possible experts and in consultation to reach an agreement about an acceptable choice. The neutral expert chosen in this manner would have an incentive to carefully evaluate the arguments in the case and to reach a fair and reasonable resolution. If there were a bias, it might result from the expert’s desire to seek a “balanced” solution, thereby making herself more attractive to future parties in need of an expert.

Some people fear that the court would give greater weight to the expert testimony of its “non-partisan” appointee than would be desirable. Certainly such testimony is likely to be given greater weight than that given by partisan experts. If both sides are given the opportunity to cross-examine the court-appointed expert, however, the court will have a chance to hear any pertinent criticisms. Of course, judges still may give too much weight to expert testimony if they do not have the technical competence to make an independent evaluation of the material. Nevertheless, I would prefer to put more reliance on the ability of


124. The incentives that arbitrators face in “interest” arbitration have been discussed in the social science literature. One theme is that the effect of a particular arbitration scheme on arbitrated outcomes depends crucially on the individual parties’ knowledge about the arbitrators’ preferences. Where there is uncertainty about preferences the structure of the arbitration process can substantially affect outcomes. Presumably the same arguments could be applied to the use of neutral experts. See, e.g., Ashenfelter & Bloom, Models of Arbitrator Behavior: Theory and Evidence, 74 Am. Econ. Rev. 111 (1984); Crawford, On Compulsory-Arbitration Schemes, 87 J. Pol. Econ. 131 (1979); Farber, Splitting-the-Difference in Interest Arbitration, 35 Indus. & Lab. Rel. Rev. 70 (1981).
judges to evaluate the worth of experts' testimony, while at the same time encouraging judges to acquire the necessary skills and background to make reasonable judgments about technical issues.