June 1982

Underestimated Threat to Multiple Regression Analyses Used in Job Discrimination Cases, An

Arnold Barnett

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

Recommended Citation

Link to publisher version (DOI)
https://doi.org/10.15779/Z38BD01

This Article is brought to you for free and open access by the Law Journals and Related Materials at Berkeley Law Scholarship Repository. It has been accepted for inclusion in Berkeley Journal of Employment & Labor Law by an authorized administrator of Berkeley Law Scholarship Repository. For more information, please contact jcera@law.berkeley.edu.
An Underestimated Threat to Multiple Regression Analyses Used in Job Discrimination Cases*

Arnold Barnett†

I

INTRODUCTION

For more than a decade, courts in the United States have been admitting statistical evidence into job discrimination proceedings.1 The potential importance of such data was underscored in Teamsters v. United States, decided in 1977. There the Supreme Court held that a prima facie case of job discrimination, in proper circumstances, could be established on the basis of statistical evidence alone.2

The growing emphasis placed on statistical proof of discrimination has led to greater sophistication in the techniques used to prepare evidence presented in court. The statistical tools that have been applied include the “standard deviations” analysis used by the Supreme Court in Hazelwood School District v. United States,3 ratio tests such as the Equal Employment Opportunity Commission’s “four-fifths” rule, correlation coefficients of various kinds,4 and the Chi-square and

---

* I am thankful for the thoughtful comments of Arthur Menard (partner at law, Morgan, Brown, Kearns, and Joy, Boston), Paul Gray (Southern Methodist University), Warren Walker (Rand Corp.), and my MIT colleagues Shlomo Kalish, Daniel J. Kleitman, Richard Larson, Anthony LoFaso, and Jeffrey Meldman. I am grateful above all to David Dolinko (clerk to Judge Harry Pregerson, U.S. Circuit Court of Appeals, Ninth Circuit, Los Angeles) for numerous editorial suggestions and extensive help with the footnotes.

† Associate Professor, Operations Research Center, Sloan School of Management, Massachusetts Institute of Technology.


3. 433 U.S. at 308 n.14, 311 n.17.

MULTIPLE REGRESSION ANALYSIS

This article focuses on the particular statistical technique of multiple regression analysis, perhaps the most sophisticated test to date.

A 1975 Harvard Law Review note argued that especially in connection with charges of salary discrimination, regression analysis furnishes a new opportunity "to facilitate the truth-finding process in any given case." That same year, a discrimination case in which a multiple regression analysis was used reached a federal district court; since then, several others have followed. Surveying the situation recently, Morris wrote that "it seems safe to predict that regression analyses will appear increasingly in future job bias cases."

Because regression analysis considers many possible influences on a variable under study, the statistic can in principle determine whether differences in status across groups reflect invidious discrimination or simply the usual application of reasonable criteria. But, like any statistical method, regression analysis can be abused. Fortunately, the courts have shown an awareness of both the possibilities for and consequences of such abuses. For example, they have occasionally rejected results of multiple regressions that unaccountably ignored major influences on the process under study or employed irrelevant or unreliable data.

The courts have been somewhat less inspiring, however, in their treatment of the issue of "functional form." This phrase refers to the requirement that the regression analyst, having chosen the variables to be included in the model, specify mathematically how each variable exerts its influence. If the specification is wrong, the accuracy of the

---

11. See infra text accompanying notes 14-19.
regression results is compromised to an unknown and possibly dangerous extent.

In some cases courts have been led to believe that details about functional form are technicalities of only secondary importance. Thus it is possible that standard forms simply are being assumed adequate rather than demonstrated to be so. If so, this is an unhappy circumstance, for there are reasons to suspect a priori that the usual linear and loglinear regression models might well be inappropriate for the analysis of employee-salary data. Such suspicions arise not from abstruse mathematical theory but rather from common sense and simple geometry. Furthermore, the problems may "tilt" the supposedly neutral analysis toward a particular conclusion about the existence of discrimination.

This essay argues that functional form errors, always a theoretical danger in regression analysis, may be particularly hazardous in the context of job discrimination. The point is not that regression analysis is inherently flawed, but that inattention to critical components of the model could make it worse than useless.

The first part of the article will briefly review recent developments involving regression analysis and job discrimination law. Then certain regression models of salary patterns will be considered and I will argue that use of any of three familiar functional forms can readily lead to difficulties. Finally, the article will indicate how to identify and cope with functional form problems.

II

REGRESSION ANALYSIS AND EMPLOYMENT DISCRIMINATION

The underlying premise of multiple regression analysis is that there exists a mathematical formula which predicts the value of a "dependent variable" from the factors that influence it (e.g., predicting employee salaries from experience, sex, education, etc.).\textsuperscript{12} The technique estimates from available data what this formula is. In essence, a regression analysis considers geometric curves of a given type and selects the one that "best fits" a given set of data points. From the mathematical equation of the curve selected, one can infer how big a role each "influence factor" plays in explaining the variation of the dependent variable.

The geometric curve identified by regression analysis is never a precise representation of the relationship among the variables. The sources of imperfection generally include random measurement error

\textsuperscript{12} See Wonnacott, supra note 6, at 248-324.
in the data and the inevitable need to neglect some minor factors that could conceivably affect the dependent variable. However, like the margin of random error calculated for opinion polls, the uncertainty levels of various regression results are estimated by a well-developed theory. But such “error bounds” are themselves dependent on some strong assumptions. Much as polling results might be useless if respondents are not chosen at random, the implications of regression analysis can be highly misleading unless three issues are resolved sensibly:

1. which variables should be included in the model;
2. which data should be examined;
3. which mathematical form specifies the relationship between the dependent variable and its hypothesized determinants.

The mathematical (or functional) form issue arises because a given regression analysis is restricted to approximating the data with a curve of a particular type. For example, the statistician may choose to summarize the data points with a straight line. The decision about which form of curve to use is made at the outset of regression analysis. The haplessness of searching for the straight line that “best fits” the points on a circle is a vivid reminder that, if an unwise decision is made about functional form, a regression analysis may be incapable of accurately depicting the patterns in the data.

The courts have shown their appreciation for the vulnerabilities of regression models by flatly rejecting some that fell far short of the ideal. In *Presseisen v. Swarthmore College*, a Pennsylvania district court discounted a multiple regression analysis of faculty salaries because academic rank was not among the model’s explanatory variables. In *Dickerson v. U.S. Steel Corp.*, again a Pennsylvania district court considered regression analysis “irrelevant” because much of the data pertained to years beyond the appropriate statute of limitations.

Indeed, courts have shown sensitivity even to subtle problems arising in the preparation of data for regression analysis. In *Presseisen*, a regression model was criticized for its tacit assumption that annual salary increases were the same in all academic departments. The court sustained the objection that because women are not equally concentrated in all areas of study, ignoring cross-departmental variations in pay scales could yield spurious evidence of sex discrimination. In *Agarwal v. Arthur G. McKee & Co.*, the court noted the difficulty of

---

13. See id. at 272-80 and 297-99.
15. “The answer is a simple one, and that is that [the regression analyst] was not properly justified in excluding rank.” Id. at 614.
17. Id. at 79 n.25.
expressing certain variables in the numerical form required for regression analysis.\textsuperscript{19} When, for example, a professor’s research accomplishment is equated to the number of papers produced, it is unclear whether the analysis is an exercise in science or satire.

On the functional form issue, however, the courts largely have been silent. In \textit{Agarwal}, for example, a California district court identified four criteria for assessing the usefulness of regression results:\textsuperscript{20}

\begin{enumerate}
  \item the significance of the influence factors discovered through the analysis;\textsuperscript{21}
  \item the quantity of data used to evaluate a particular variable;
  \item the quality of the data; and
  \item the ease with which independent variables may be converted to numerical equivalents.
\end{enumerate}

Noteworthy is the absence of any reference to the question of functional form.

Perhaps the judiciary has said little about functional form because it has been led to believe that standard regression models (generally linear, as will be shown)\textsuperscript{22} are quite satisfactory for investigating job bias, especially when the analysis is meant less to measure the precise extent of discrimination than to indicate the likelihood that any discrimination exists. In \textit{Pennsylvania v. Local 542, Operating Engineers},\textsuperscript{23} one of the few cases reporting a dispute over functional form, a statistician defended his linear regression model by saying:

\begin{quote}
  It is my belief based on experience and analysis of regression techniques, that a linear model is adequate to calculate a race effect, and that adding non-linear terms and interactions would rarely produce a downward change in the race effect.\textsuperscript{24}
\end{quote}

The court quoted this rebuttal approvingly.\textsuperscript{25}

The clearest statement to date about the functional form issue appeared in the 1975 \textit{Harvard Law Review} note advocating greater use of regression analysis.\textsuperscript{26} Its author acknowledged that, under economic theory, one would not expect a worker’s salary to increase in a linear

\begin{thebibliography}{26}
\bibitem{19} Fair Empl. Prac. Cas. (BNA) 503 (N.D. Cal. 1977). “Plaintiff’s regression analyses also contained a number of defects based upon the difficulty of assigning numerical values to independent variables having complex characteristics.” \textit{Id.} at 506.
\bibitem{20} \textit{Id.}
\bibitem{21} The issue referred to is whether the uncertainty levels assigned to particular results by regression theory are large enough to cast doubt that these results reflect genuine patterns.
\bibitem{22} \textit{See infra} text accompanying notes 31-33.
\bibitem{23} 469 F. Supp. 329 (E.D. Pa. 1978). Another case in which functional form was disputed is Vuyanich v. Republic National Bank of Dallas, 505 F. Supp. 224 (N.D. Tex. 1980). The confusion that surrounded the issue in that case is suggested by the court’s observation that “to say the least, the Bank is unclear as to whether Dr. Madden used natural logarithms where she should use actual value, or vice versa.” \textit{Id.} at 341-42.
\bibitem{24} 469 F. Supp. at 377.
\bibitem{25} \textit{Id.} at 377-78.
\bibitem{26} Harvard Note, \textit{supra} note 7.
\end{thebibliography}
fashion, i.e., by the same amount year after year. Nevertheless, the au-
thor argued that:

[In the typical case where an employer will claim reliance on a set of
factors, estimations using the data in the form in which they are ob-
served should be adequate to prove (or rebut) a prima facie case. As an
approximation to an actual employment decision, factors can be ex-
pected to be considered by an employer just as they are observed,
rather than in more esoteric mathematical expressions.27

This means, for example, that employers who measure seniority by
"number of years with the company" presumably arrange for salaries
to grow in direct proportion with this quantity. From this, the author
proceeded to introduce regression models based on the simple linearity
assumption despite its divergence from economic principles.28

The author did qualify the endorsement of the use of simple func-
tional forms:

In the unusual case where no information is known about the em-
ployer's decision, a more complicated expression suggested by business
and economic theory might have to be used to assess observed out-
comes of that decision.29

In other words, the key factor in determining whether one should use a
simple or complex mathematical formula is the recalcitrance of the em-
ployer in identifying the decision variables. If it is conceded that expe-
rience is relevant to salary levels, the dependence relationship is
assumed straight-forward. Absent this concession, a more labyrinthine
pattern is suspected.

The author of the note deserves respect for the forthright articula-
tion of views that generally remain inchoate. Indeed, the note itself
may have engendered confidence that functional form is a straightfor-
ward issue. But I will argue that:

(1) The distinction made between "typical" cases in which
simple models are adequate and the "unusual" ones in
which they are not is untenable.

(2) Similarly, the dichotomy between simple models and those
arising from "business and economic theory" is problem-
atic, because neither kind may be appropriate in a given
case.30

27. Id. at 395-96 (emphasis added).
28. See J. Mincer, Schooling, Experience, and Earnings 76-79 (1974) for an
econometric analysis on this point.
29. Harvard Note, supra note 7, at 396 (footnote omitted) (emphasis added).
30. See infra Sections III-V.
III

LINEAR REGRESSIONS AND SALARY PATTERNS

For ease of exposition, I will discuss the functional form issue for regression analyses in which employee salary is the dependent variable and years with the organization and the employee's sex are the only explanatory variables. In practice, of course, several more explanatory variables would generally be necessary. Ignoring these extra factors, however, is not a gross oversimplification because they generally would not mitigate any problems of functional form. Indeed, such problems could well be compounded by the presence of additional explanatory factors in the model.

Consider first the simple linear model that relates salary to sex and experience by an equation of the general form:

\[ S = A + B y + C z + \epsilon \]

where:
- \( S \) = salary in dollars
- \( y \) = years with the organization
- \( z \) = 1 for women, 0 for men
- \( \epsilon \) = a "random-error" term that, on the average, is zero

A, B, and C are numerical constants (parameters) to be estimated through the analysis of relevant data; they are allowed to assume negative values as well as those of zero or greater.

Given the definition of \( z \), the parameter C is interpreted as the average difference in salary between men and women of identical seniority. Were C to take on a significant negative value, this would be construed as evidence of systematic discrimination against women.

This linear model is similar to many routinely used in practice and discussed and defended in the Harvard Law Review note. They were also considered with apparent approval in Morris' book on the use and misuse of statistical evidence in job discrimination cases. And while judicial opinions do not always clearly describe regression analyses introduced as evidence, it does appear that several linear models have reached the federal courts.

---

31. See supra text accompanying notes 26-29.
32. Morris, supra note 10, at 76-77. While Morris discusses various potential problems in multiple linear regression analysis, the linearity assumption itself is not criticized. Id. Linear regressions were considered in a recent fact-finding hearing among Vermont State Colleges, its Board of Trustees, and the Faculty Federation of the Vermont State Colleges, held before chairman John P. McCrory at Castleton State College, Castleton, Vermont, on November 19, 1980. I served as statistical consultant to the defendant. See discussion at text accompanying notes 45-48 infra.
33. Of the cases mentioned in supra notes 8-9, verbal descriptions of the regression analyses used suggest the employment of linear models in, at a minimum, James v. Stockham Valves & Fittings Co., Greenspan v. Automobile Club of Michigan, Presseisen v. Swarthmore College and Mecklenburg v. Montana State Board of Regents.
Even assuming that seniority and sex are the only possible determinants of salary, is the linear model in Equation (1) adequate for establishing a prima facie case of sex discrimination? In considering this question, three general observations about salary and hiring patterns in a typical organization are useful:

(i) Annual increases in salary tend to be calculated as a percentage of current salary. Thus, the higher one's current salary, the greater the increase in dollar terms.  

(ii) Because of the recent upsurge in the fraction of women who work, one would generally expect that, even in the absence of discrimination, female workers would tend to have lower seniority than males.  

(iii) Given the birth patterns in this country, one would expect a disproportionate part of the work force to be at low experience levels.

While these statements do not apply to all organizations, their relevance is sufficiently general to merit exploring their implications.

With respect to Equation (1), statement (i) suggests that the greater the value of \( y \), the greater the salary increase when \( y \) goes up by one. (i.e., \( S \) is a convex function of \( y \)). Under a linear model like Equation (1), however, the salary increase associated with a unit raise in \( y \) is assumed completely independent of its current value. Indeed, the model implies that as one gains seniority, the corresponding raises in pay are continuously decreasing in percentage terms.

It appears, therefore, that a linear formulation is rarely a perfect description of the way salaries are actually determined. But given the purpose of the regression analysis, such imperfections are relevant only if they might plausibly bias the calculated estimate of \( C \). A fairly simple argument suggests that such a bias is likely.

Suppose that, in reality, there is no sex discrimination in the organization being studied and thus that \( C \) is actually equal to zero. If statements (i), (ii), and (iii) apply, what might one expect about an estimate of \( C \) obtained through linear regression on employee salary data?

Suppose that, initially, \( C \) is assumed to be 0 (correctly, it turns out) and an analysis is performed to estimate \( A \) and \( B \). Geometrically, the

---

34. Indeed, when California Governor Jerry Brown ordered equal dollar pay increases for all state employees ($780 per year) in 1976-1977, the anomaly attracted national attention. N.Y. Times, Jan. 8, 1976, at 21, col. 1.

35. Between 1960 and 1977, the number of female workers in the United States rose by 73% (from 23.2 to 40.1 million); the corresponding rise for males was only 22% (48.9 to 59.5 million). U.S. DEP’T OF COMMERCE, STATISTICAL ABSTRACT OF THE UNITED STATES, Table 644 at 398 (1978).

36. In 1977, the average number of American workers at each age from 20 to 24 was 3.04 million. The analogous figure for each age from 45 to 54 was 1.69 million, only about half as large. Id.
situation is depicted in Figure 1, with each employee represented by a point whose abscissa is current salary and whose ordinate is years of experience. (Points marked with W's are women.) \( \hat{A} \) and \( \hat{B} \) represent the intercept and slope, respectively, of the straight line that best describes the data points under the least-squares "goodness-of-fit" criterion routinely used in regression analysis.\(^{37}\)

Consistent with (i), the points tend to "climb faster" as \( y \) goes up. Because of (iii), there are more data points at lower than at higher \( y \)-values; because of (ii), employees with the greatest seniority are disproportionately male (i.e., the W's get sparse as \( y \) gets large). Note that since the firm really does not discriminate by sex, the data points reflecting women's salaries are not abnormally low given their \( y \)-values.

Because there are so many data points in the low-\( y \) region, providing a "good fit" there dominates the calculation of \( \hat{A} \) and \( \hat{B} \).\(^{38}\) This situation leads, however, to the tendency apparent in Figure 1 for the line to underpredict the salaries of the most experienced workers (i.e., those beyond point P). Since the sum of all the regression line's prediction errors (positive and negative) must be zero,\(^{39}\) if the line underpredicts beyond P, it necessarily tends to overpredict in the region

---

37. In assessing possible regression curves, one observes the distance by which each curve "misses" each point in the data set. The curve for which the sum of the squares of these distances is minimized is chosen to describe the data under the (aptly named) "least squares" criterion. See Wonacott, supra note 6, at 320-26.

38. Actually, far-off data points (e.g., the highest \( y \)-values) have larger-than-usual influence on \( B \); if they are rare, however, their total impact will still generally be limited.

MULTIPLE REGRESSION ANALYSIS

below P where female employees are concentrated.\(^\text{40}\)

Now, suppose C is allowed to move from zero with \(\hat{A}\) and \(\hat{B}\) held fixed. Since \(z = 0\) for men, the value given C has no effect under Equation (1) on the predictions of male salaries. But if C were assigned an appropriate negative value, the overprediction of female salaries just discussed would be largely counteracted. Thus, the overall predictive accuracy of the model would be improved.

We have just outlined the first stage of a dynamic that could lead to a negative estimate of C.\(^\text{41}\) But this negative value, we must remember, reflects not the presence of sex discrimination but rather an attempt by the regression model to “recover” from its false assumption that S depends linearly on y. While this discussion is hypothetical, it suggests that if the relationship between S and y is nonlinear, the results pertinent to discrimination arising from the linear regression Equation (1) are “untainted” only if the distribution of seniority in the protected group is essentially the same as that for all others. One suspects that this last situation is quite rare; its absence in a particular case would be fairly easy to demonstrate.

Before discussing further the implications of these comments, we turn to two other functional forms that are attractive in regression analyses of salary data.

IV

SOME OTHER SALARY REGRESSIONS

A. Loglinear

An alternative to the linear Equation (1) under which salaries grow by a fixed percentage each year is:
\[
S = \gamma(1 + \alpha)yz^\beta \epsilon
\]
[Equation (2)]
where S and y are defined before,
\(z = 2\) for women, \(1\) for men
\(\epsilon = \) is a “random-error” factor,
and \(\gamma, \alpha,\) and \(\beta\) are constants to be estimated through regression analysis.

Equation (2) is of the loglinear form used frequently in economics.\(^\text{42}\)

Under the definition of z, \(2^\beta\) is the ratio of salaries of a woman and

---

\(^{40}\) While no W’s appear beyond P in Figure 1, the argument only requires that the proportion of employees beyond P who are female be smaller than the corresponding proportion for those below P. We also tacitly assume that females are not jammed at the lowest y-values.

\(^{41}\) Computers do not literally use this recursive approach in estimating regression parameters, but the soundness of this heuristic argument nevertheless remains unaffected.

\(^{42}\) See H. THEIL, INTRODUCTION TO ECONOMETRICS 2 (1978), where the loglinear regression model first appeared on page 2. The familiar economic concept of “elasticity” is tied to the loglinear model.
man of equal seniority; hence, a significant negative value of $\beta$ would be interpreted as evidence of bias against females. The factor $(1+\alpha)^y$ reflects the assumption that salary increases are proportional to current salary so that a phenomenon akin to compound interest is present; $\alpha$ represents the annual growth rate in this process. Note that the disharmony between Equation (1) and observation (i) in Section III is alleviated by the use of Equation (2).

The only gnawing concern about Equation (2) is that, even if all salaries increase by roughly the same percentage in a given year, this percentage itself may vary over time. This would clearly seem true given the recent upsurge in inflation, but even if one corrects for inflation by expressing $S$ in “real dollars” the issue does not disappear. Gyrations in the company’s financial health and the militancy of its labor unions, as well as wage restraints suggested or imposed by government, all contribute to the sense that $\alpha$ probably varies over time. This circumstance raises the question whether the failure of Equation (2) to consider the time-dependence in $\alpha$ could seriously threaten the accuracy of its estimate of $\beta$.

I discuss this issue in the Appendix, where I indicate that if $\alpha$ has tended to drop in recent years (i.e., if real wages grew faster in the past than now), then the estimate of $\beta$ is subject to a “false negative” distortion similar to that suffered by $C$ in Section III. The reason for the problem is essentially that depicted in Figure 1: the logarithm of salary is assumed under Equation (2) to grows linearly with $y$ while the actual logarithm grows faster as $y$ increases. This finding is rather unfortunate, for the circumstances that trigger the problem seem to resemble recent economic trends in the United States. Data from the U.S. Bureau of Labor Statistics indicate that between 1960 and 1967 the average real income for nonagricultural workers increased at a rate of 1.63% per year. Between 1967 and 1973, this annual growth rate slackened to 1.17%. From 1973, the year of the first major oil price hikes, through 1979, real income actually dropped at a rate of 1.29% per year. These data refer to average income for all workers rather than annual changes in salary for any single one and, like all aggregate statistical indices, they suffer imperfections. However, the data are certainly consistent with the possibility that, looking back from 1980, $\alpha$ has been tending downward for at least two decades.

In consequence, a loglinear model like Equation (2), while avoiding one of the apparent deficiencies of the linear model, may itself be subject to a subtler variant of the same deficiency. One could, of course, try to "weed out" the time-dependence of the model’s parameters by adjusting for trends in workers’ real purchasing power. But

these trends themselves vary from region to region, industry to industry, and ultimately, firm to firm. This does not mean that the problem is intractable; it does mean that the hope of a "quick fix" may well be illusory.

B. Quadratic

A regression model of the general type that appeared in Agarwal v. Arthur G. McKee & Co. takes the form:

\[ S = E + Fy + Gy^2 + Hz + \epsilon \]  

[Equation (3)]

where E, F, G, and H are the parameters to be estimated. Here H is the indicator coefficient for discrimination, analogous to C in Equation (1).

Equation (3) differs from both Equations (1) and (2). The earlier formulas are mathematical elaborations of two simple compensation rules: salary rising alternatively by a fixed dollar amount or a fixed percentage each year. In contrast, it is hard to imagine that employers actually use an unintuitive formula like Equation (3) to compute pay raises. Evidently the plaintiff's statisticians thought that Equation (3), although imperfect, was a good approximation of the relevant salary pattern. The reason for this belief, however, was not stated in the opinion.

In all probability, when a model like Equation (3) is calibrated from actual data with H fixed at 0, it is likely to overpredict in some ranges of y-values and to underpredict in others. Any tendency at the relatively low y-values at which women and minorities are concentrated would lead H, when allowed to move freely, to gravitate away from 0 even in the absence of discrimination. Furthermore, the economic trends that threatened the validity of Equation (2) do not disappear when one uses Equation (3). There is, in sum, no self-evident reason to trust the results of Equation (3).

V

Discussion

The preceding sections suggest that, given the social and economic realities that surround the present distribution of wages, analyzing salary data with standard regression models can yield inaccurate estimates of the extent of discrimination. But are we discussing a problem of appreciable magnitude or a miniscule effect? No universal answer is possible, but a recent proceeding might provide some clues. An investigation of salary patterns at a college in Vermont included a linear

---

44. 19 Fair Empl. Prac. Cas. (BNA) at 506.
45. See supra note 32.
regression analysis with salary as the dependent variable and seniority, rank, and sex as explanatory factors. The results implied that under the college's policies a female faculty member could expect $443 per year less on the average than a male with equal rank and experience. Furthermore, the deficit was found "statistically significant." It turned out, however, that nearly half the female faculty members had male counterparts with exactly the same rank and seniority. When comparing salaries within such pairings, there was obviously no need to correct for differences along those two dimensions. In sixty percent of the pairings, the woman had the higher salary, and, over all the pairings, the average salary for women was $35 higher than that for men. Barring the unlikely event that the school penalized the remaining females almost $1000 apiece, this matched comparison suggests that the estimated women's loss of $443 was at least inflated and perhaps entirely a spurious consequence of problems in the regression model. The most obvious of these problems is its linear functional form.

There is another reason, related to the concept of statistical significance, why a functional form problem could be important under present standards. Virtually any observed pattern in data, however strong, could conceivably result solely from chance fluctuations. Even if birth-days are distributed uniformly over the year, fifty people picked at random might all turn out to be born in January. But as it becomes less and less probable that chance alone could have generated a given pattern, it becomes more and more indefensible to dismiss that pattern as fortuitous.

Several courts have noted that the evidentiary value of a statistical finding depends on the probability that fluctuations alone could equal or exceed the size of the observed effect. The most common rule is that evidence is compelling if and only if the probability the pattern obtained would have arisen by chance alone does not exceed five percent. (If this circumstance obtains, the result is said to be "statistically significant" at the five percent level.) This rule implies that the courts will tolerate a risk of up to five percent that an apparently strong pattern in the data is a "statistical accident."

46. Rank was quantified as follows: 1 = Instructor, 2 = Assistant Professor, 3 = Associate Professor, 4 = Professor.
47. This phrase is defined at text accompanying infra notes 49-50.
48. Of course, there are other possible problems with this regression model (e.g., the absence of other explanatory variables). The point is that even if one accepts the judgments of the regression analysts that such exclusions were justified, one should still be skeptical of their conclusions.
50. Such a rule was used, for example, to evaluate statistical correlations in Albemarle Paper Co. v. Moody, 422 U.S. at 340.
Sometimes a functional form problem will be "neutralized" because even though the indicator coefficient for discrimination is incorrectly estimated as nonzero, it will not achieve statistical significance. The coupling of a functional form bias with the usual effects of chance, however, could well push the probability of a "statistical accident" above the permissible five percent to two, three or four times that. In other words, errors in functional form could distort not only the estimates of key parameters but also the statistical significance attributed to these estimates.

Under these circumstances, all parties in a job discrimination case should pay close heed to the functional form assumptions in a multiple regression analysis. The Supreme Court has noted that statistics "are not irrefutable . . . [and] like all other kinds of evidence, they may be rebutted." If, for example, the defendant in a job bias case presents evidence that the plaintiff's assumption of a linear relationship between salary and seniority is clearly inconsistent with reality, the validity of the plaintiff's analysis is impeached as much as if some of the data had been shown to be erroneous or immaterial.

Therefore, those who present regression results should be prepared for challenges on the issue of functional form. Before suggesting how such challenges might be met I will briefly discuss two possible methods that would not be adequate: $R^2$-statistics and the invocation of "business and economic theory." One article on regression models in administrative proceedings suggests the potential popularity of such methods:

The choice between forms of regression equation is determined by considerations of economic theory and by the respective precision of each of the regression estimates. The basic measure of the precision of the equation as a whole . . . is denoted $R^2$.

A. $R^2$ Statistic

This statistic is a number between zero and one, commonly used to describe how well a regression curve fits the data points that generated it. More specifically, it compares "goodness of fit" of the regression model with that of a very crude model that assumes none of the explanatory variables had any effect on the dependent variable. For two separate reasons, $R^2$ statistics cannot establish that particular regression results are worthy of confidence.

As indicated above, $R^2$ is exclusively a relative measure of accu-
racy. Even if $R^2$ is high, there is always the danger summarized by the proverb, "In the valley of the blind, the one-eyed man is king." Extending this logic, it is not especially compelling that the regression analysis presented by one party in court achieved a higher $R^2$ than that presented by the other.

Another problem with $R^2$ is that it concerns the predictive accuracy of the entire model rather than that of its individual components. Sometimes a regression equation appears to be doing well because two serious errors in its construction have "cancelled" one another. In the situation depicted in Figure 1, for example, introducing a negative value for coefficient $C$ might largely dispel the observable consequences of the false linearity assumption.

**B. Economic Theory**

According to both Finkelstein and the Harvard Law Review note, functional forms arising from economic theory have natural relevance to regression analyses on economic subjects. This is not an unreasonable position. But the mere fact that a functional form enjoys "economic sanction" does not ensure its appropriateness in a given situation.

The explanation for this caveat is very simple: economics is not an exact science. Labor economics and the theory of the firm may be very powerful, but it is unlikely they can reliably describe the idiosyncrasies and vicissitudes of any organization over even a small portion of its existence. Thus, independent evidence that general principles apply to a firm in a given job discrimination case is important.

In contrast to the approaches just mentioned, certain model-validation techniques can increase the plausibility of using a given functional form. Many of these validation procedures are based on examining the residuals, the errors that arise when the regression model is used to make predictions in the very data set used to estimate its parameters. The assumptions about "random error" that animated the analysis have sharp implications on such matters as the distribution of sizes of these residuals, the correlation (or lack thereof) between the magnitudes of residuals and the values of explanatory variables, and

---

54. I have argued that even regression models with $R^2$-values exceeding .95 are insufficiently sensitive to determine whether capital punishment could deter killings. Barnett, The Deterrent Effect of Capital Punishment: A Test of Some Recent Studies, 29 OPERATIONS RESEARCH 346 (1981).

55. See supra notes 7 and 52.

56. I will not discuss these techniques in detail here, but instead I refer the reader to texts that do. See, e.g., C. Daniel & F. Wood, FITTING EQUATIONS TO DATA 27-49 (1980); N. Draper & H. Smith, APPLIED REGRESSION ANALYSIS 86-104 (1966); F. Mosteller & J. Tukey, DATA ANALYSIS AND REGRESSION 407-49 (1977).
MULTIPLE REGRESSION ANALYSIS

the comparative signs of the prediction errors for adjacent data points (e.g., workers with very similar seniority). Test results indicating that the residuals do not follow the expected patterns are clear warning flags that something is wrong with the regression analysis. (Such flags would almost surely arise in a situation like that in Figure 1.) On the other hand, behavior in conformity with the anticipated patterns tends to support the choice of functional form.

Another procedure of considerable value is to allow the regression model to assume a somewhat more general functional form than that originally proposed. For example, there exist classes of regression models of which the linear and loglinear forms are but special cases. If, within the larger class of models, the one that fits the available data best does not significantly outperform the specific model initially chosen (i.e., if the added complexity “buys” little increase in accuracy), the original choice seems sound.

Other validation methods exist and, as the discussion implies, modern regression analysis allows the user to move far beyond simple functional forms if that proves necessary. Various transformations of the data can be explored with surprising ease, and procedures also exist to ensure that a few unusual (and possibly erroneous) data points do not dominate the regression results. Furthermore, some sophisticated models outside the context of regression modeling have apparently met successful use in the analysis of salary patterns. These possibilities make it all the more indefensible to use prefabricated functional forms without giving serious thought to their potential defects.

Of course, the choice of a proper functional form can never be reduced to a mechanical process. In salary regressions, for example, the statistician must ask what measures are appropriate to eliminate distortions caused by inflation. And one must always be alert to special details of the situation that should directly affect construction of the underlying mathematical model. For example, the Boston Edison Company changed its new employee pay scales in 1973 after determining that it systematically paid higher wages than private industry. More recently, the utility was charged with pay discrimination against women and minorities. In modeling the company’s compensation policies, the analyst should incorporate the 1973 discontinuity at the

59. Gray, A Faculty Model for Policy Planning, 10 Interfaces 91 (1980).
outset, rather than hope that the model's parameters will somehow depict what happened.

VI
Final Comment

In employment discrimination litigation, mathematical analyses are often essential for a coherent summary of vast quantities of discordant yet vital information. It is unrealistic to hold such analyses to standards of perfection that are not only unattainable but higher than those applied to other forms of evidence. But these realizations, however compelling, cannot justify ignoring the possibility that a mathematical model can distort the phenomenon it is meant to describe.

More specifically, it is naive to believe that simple mathematical analyses yield good approximate results and that more elaborate models merely "fine tune" the accuracy. If the process by which a model is simplified is incorrect, it is perfectly possible that its implications will be false. Those choosing functional forms for regression models about job discrimination should be especially mindful of these points.
MULTIPLE REGRESSION ANALYSIS

APPENDIX

A POTENTIAL BIAS IN LOGLINEAR SALARY REGRESSION CAUSED BY RECENT ECONOMIC TRENDS

Consider the simple situation in which the starting salary $S_0$ for each worker in a firm remains constant over time in real dollars. Suppose the organization increases all salaries by the same percentage in any given year, and let $\alpha_i$ be the across-the-board proportional raise $i$ minus 1 years ago. Thus $\alpha_{25}$ is the percentage increase granted 24 years ago; $\alpha_1$ equals the present increase. The present real salary $S_y$ of someone who has been with the firm for $y$ years would follow:

$$S_y = S_0 \cdot (1 + \alpha_y) \cdots (1 + \alpha_1)$$  \hspace{1cm} \text{[Equation (4)]}

Equation (4) implies that:

$$S_y = S_{y-1}(1 + \alpha_y)$$  \hspace{1cm} \text{[Equation (5)]}

If logarithms are taken, then Equation (5) becomes:

$$\ln S_y - \ln S_{y-1} = \ln(1 + \alpha_y)$$  \hspace{1cm} \text{[Equation (5A)]}

Now if $\alpha_y$ increases as $y$ does (i.e., real salary gains increase as one moves further back in time), then Equation (5A) shows that the larger the value of $y$, the larger the growth in $\ln S$ as $y$ goes up by one. In other words, $\ln S$ is a convex function of $y$.

Suppose $\beta$ is set equal to zero in Equation 2 (corresponding to the absence of sex discrimination), and logarithms are taken. The resulting equation takes the general form:

$$\ln S = A + B y + \epsilon$$  \hspace{1cm} \text{[Equation (2A)]}

where $\epsilon = \ln \epsilon$, $B = \ln(1 + \alpha)$, etc.

The interesting thing about Equation (2A) is that, as before, it represents a linear regression analysis in which the dependent variable ($\ln S$) is a convex function of the explanatory variable ($y$). Thus, under the general assumptions (ii) and (iii), the very problems depicted in Figure 1 would again lead to a tendency to overpredict the salaries of most female workers. It follows that the same forces that could yield a false negative estimate of $C$ in linear Equation (1) could push its loglinear analogue $\beta$ in a similar manner.

As noted earlier, real wage gains have apparently not kept up with inflation in recent years, so the assumption made about $\alpha_y$ seems realistic. While real starting salary $S_0$ is probably not time-invariant as assumed, the evolution of this quantity over time probably does more to complicate the problems cited in this Appendix than to dispel them.

---

61. I assume for simplicity that $y$ is an integer and that $\alpha_1$, the raise this year, has already been granted.

62. See supra text accompanying notes 35-36.