Regression Analysis as Statistical Evidence

by

George P. McCabe

Technical Report #84-3

Department of Statistics
Purdue University

February 1984

Regression Analysis as Statistical Evidence

George P. McCabe
Statistics Department
Purdue University

1. Introduction

Does an employer discriminate on the basis of sex in determining salaries? This question and similar ones are often the major focus of class action litigation. Although the fundamental issues are legal, statistical analyses can be used effectively to produce evidence which aids the process of answering such questions.

If an employer has very few employees, say two men and two women, with similar qualifications who perform the same job equally well, then a simple perusal of the salaries should be sufficient to make a determination. When large numbers of employees with different qualifications and different jobs are involved, however, more sophisticated methods are needed.

Although the methods may be sophisticated, anyone with a minimum knowledge of statistical software can easily produce the required calculations. The appropriate use of statistical methodology requires more than running a few computer programs, however. A firm understanding of the details of the method and its limitations is required to construct a meaningful analysis. In this article, some important issues related to such an understanding are discussed.

For convenience of presentation, sex discrimination in salary is used as the canonical example. The comments apply, however, to discrimination against any protected group on any similar measure of reward.

2. Purpose and Use of a Statistical Analysis

Given a collection of employee records containing salary and measures of employee worth, is there a pattern of systematic underpayment of women versus men? This question is a refinement of the question posed in the previous section and is the type of question which is amenable to a statistical investigation. A major function of a statistical analysis is to summarize in a few numbers, i.e. statistics, the information contained in a large collection of data. As with any summary, some information is lost and some is retained. A good summary retains the essential characteristics of the data while discarding as little as possible. If, for example, the men and women employees have substantially different job qualifications, then a summary which ignores this fact is deficient.

Statistical methodology, as described in the above scenario, is often called data analysis. The determinant of quality is the adequacy of the descriptive summary. Data analysis is an important area of statistics which has been much neglected but is currently receiving deserved attention. See, for example, the books by Tukey (1977) Mosteller and Tukey (1977) and Chambers et al. (1983).

Consumers of statistical evidence have rarely been content with data summaries. There appears to be an obsession with "statistical significance" and p-values. Generation of such quantities requires the imposition of probabilistic models and assumptions, the details of which, most consumers are largely unaware. Statistical inference is the branch

of statistics that deals with such matters. In this scenario, a probabilistic model for the salary determination policy is developed. The model contains unknown parameters which are to be estimated by data. Given a well specified model and data conforming to this model, statistical theory provides the procedures for estimating the parameters. Statistical tests can then be constructed to determine the probability that the value of an estimated parameter would be as or more extreme than that observed, if in fact, the true parameter value was some hypothesized number. In the context of the present problem, the parameter of interest is the salary differential between men and women adjusted for qualifications and other variables. Note that the output of a probability or p-value requires the assumption of a probability model as input.

Regression analysis may be viewed as a data analysis technique or as a tool of statistical inference or both. Constructing a good data summary and building a good probabilistic model are similar tasks. However, in using the model for inference, we must be fully aware of the assumptions made and their consequences.

3. Model Building

Procedures used by employers for salary determination are sometimes simple and sometimes quite complex. A probabilistic model should capture the essence of the procedure by accounting for the major salary determinants. A good model is a kind of mirror of the process. It is a mathematical abstraction which attempts to explain how salaries are determined.

In fields such as physics, there exist variables which follow mathematical models quite precisely. Measurement error is the only source of variation. In such circumstances, mathematical models provide a good fit to experimental data and can be used to explain fundamental physical relationships.

It is presumptuous of a statistician, economist or any user of statistics to pretend that the regression analyses typically used as statistical evidence in discrimination capture the essence of the salary determination process. In employment situations, salaries are determined by many factors, some qualitative and some quantitative. Consultation with salary administrators for any large organization will reveal some idea of the complexity. To suppose that a list of variables are compiled, combined in a linear fashion and then added to an independently distributed normal error term is unrealistic.

Should we thus conclude that regression models are useless for studying salary discrimination? The answer is no. We can, in many instances, obtain a reasonable approximation to the salary determination process by using these models. In addition, some of the theoretical models discussed elsewhere in this volume provide a great deal of insight into the use of and limitations associated with these analyses. In litigation, any evidence has its own inherent proper use and limitations. Statistical evidence is no exception.

4. Regression as Data Analysis

Let Y denote salary and $X_1, X_2, \ldots, X_p$ denote variables which are potentially useful in explaining salary. The subscript i denotes the i^{th} employee for $i = 1, \ldots, n$. Consider the following system of equations

$$(1) \qquad Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

For the i^{th} employee, $(Y_i, X_{1i}, \ldots, X_{pi})$ is observed. The β's are unknowns and for the present, ε_i can be viewed as the difference between Y_i and the other terms on the right-hand side of (1). If a particular β, say β_p, is zero then there is no information in X_p that is useful for predicting salary using (1), i.e. given that linear prediction is used and that the other p-1 X's are in the equation.

Since the β's are unknown, a method is needed for computing them. A standard procedure is to use least squares. The idea is as follows. For any set of $\hat{\beta}$'s, the quantities

$$(2) \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_p X_{pi}$$

can be calculated. The least squares $\hat{\beta}$'s are those which minimize the quantity

$$(3) \qquad \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Measures of fit include the mean squared error,

$$(4) \qquad MSE = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

and the squared multiple correlation,

$$(5) \qquad R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} .$$

Notice that if $Y_i = \hat{Y}_i$ for all i, then $R^2 = 1$, whereas if $\beta_1 = \ldots \beta_p = 0$ then $\hat{\beta}_0 = \bar{Y} = \hat{Y}_i$ and $R^2 = 0$. All other cases lie between these extremes.

If

$$(6) \qquad X_{pi} = \begin{cases} 1 & \text{if the } i^{th} \text{ employee is male} \\ 0 & \text{if the } i^{th} \text{ employee is female} \end{cases}$$

then $\hat{\beta}_p$ is the average difference between salaries of men and women adjusted for the other p-1 variables in a linear fashion using the least squares method. Computer programs are widely available for performing these calculations.

The above discussion is entirely concerned with descriptive statistics. The quantity $\hat{\beta}_p$ is a good summary statistic if the least squares method is an efficient way to summarize the data. On this level $\hat{\beta}_p$ represents only itself; it is not an estimator of anything nor are there any hypotheses to test about it or p-values associated with such hypotheses.

5. Regression as a Probabilistic Model

To discuss estimators, hypotheses, etc., we need to impose a probabilistic model upon the problem. The usual model is given by the equations described in (1) with the additional assumption that the ε_i are identically and independently distributed normal variables with mean zero and unknown variance σ^2, i.e.

$$(7) \qquad \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2) \text{ for } i = 1,\ldots,n.$$

In this model the unknown β's are called parameters and the $\hat{\beta}$'s are parameter estimators. The least squares method is the one generally used to calculate the $\hat{\beta}$'s and this method coincides with a theoretically sound procedure called maximum likelihood estimation.

The seemingly innocuous assumption, (7), allows us to pass to the domain of statistical inference. By assuming that the ε's are random variables, it follows that so are the Y's and also the $\hat{\beta}$'s. In addition, for a given set of data, $\hat{\beta}_p$ no longer represents just itself; it is an estimate of the unknown parameter β_p. With this framework we can ask the extent to which the calculated $\hat{\beta}_p$ is compatible with a null hypothesized value of zero and we can calculate the probability that, under such a hypothesis, we would observe a value as or more extreme than the one obtained, i.e. the p-value.

The distribution of $\hat{\beta}_p$ divided by its estimated standard error, $SE(\hat{\beta}_p)$, is a t-distribution with n-p-1 degrees of freedom. The t-statistic,

$\hat{\beta}_p/SE(\hat{\beta}_p)$, can be interpreted as the number of standard deviations by which the adjusted male and female salaries differ. In fact, one could obtain the same result by viewing the problem as an analysis of covariance on salaries with two groups and $X_1,\ldots,X_{p-1}$ as covariates.

Several excellent texts on regression are available. Among these are Draper and Smith (1981) and Neter and Wasserman (1974). Mosteller and Tukey (1977) present the data analytic viewpoint and discuss alternatives to least squares. The survey paper of Hocking (1983) gives a good overview of research in the area and has an extensive bibliography.

6. Regression as an Approximation

The statistical theory underlying least squares and regression analysis has been well known for a long time and the advent of the modern computer has made the calculations cheap and easy to perform. Why then, should there be any difficulty in constructing a good statistical analysis, data analytic or inferential, for use as statistical evidence in discrimination litigation? One reason is that very few salary administrators have ever seen equations like (1) and even if they had, they would be unlikely to use such a procedure for determining salaries.

Therefore, in most cases, the best that can be expected is that (1) provides a reasonable means of generating useful summary statistics and with (7) is a reasonable approximation to the salary process. The word reasonable is difficult to define and perhaps is beyond the scope of the statistical analyst. On the other hand, it is sometimes easy to detect unreasonable analyses. In what follows, we examine in detail some aspects of regression analysis which are important to consider if misleading or unreasonable analyses are to be avoided.

7. Choice and Quantification of the Variables

The quality of a regression analysis is fundamentally dependent upon the quality of the input, i.e. the $Y, X_1, \ldots, X_p$ variables that are analyzed. When we name variables, we often neglect the fact that judgements have been made with regard to how they are quantified, or as John Tukey would say, expressed.

Salary sounds like a fairly unambiguous measure. However, it may include a variety of extra components including bonuses, retirement fund contributions, etc. The reported earnings on W-2 forms is often the easiest data to obtain. The proper definition of salary in a particular circumstance seems to be a legal question and beyond the expertise of the statistician.

Economists who study this field often build theoretical models using the logarithm of salary. In practice, it is reasonable to run analysis using both salary and log (salary). With most data sets, the results will

agree qualitatively. If not, then additional examination of the data is required to find out why and to determine which approach is more reasonable.

A regression analysis on salary is like a snapshot in time. Long term employees were hired and received raises in years before civil rights legislation was in effect. It can be argued that a more appropriate measure of employer performance with regard to salary is the salary change over time. Thus, a regression analysis of salary raises, perhaps as a percentage of salary, may assess more effectively the yearly decisions for which an employer is accountable. Issues related to this viewpoint are discussed by Churchill and Shank (1976).

Variables used to predict salary, the $X_1,\ldots,X_p$ of (1), generally fall into one of three classes: a) characteristics of the employee such as education, prior job experience, standardized tests scores, age, race and sex; b) characteristics of the job such as job grade, administrative responsibilities and quantitative assessments of job value; and c) measures which depend jointly on the employee and the job such as length of service with the company, time in the particular job, absences, productivity and performance in the job.

Some variables, such as employee evaluations, pose particular difficulties. If a variable is measured in objective fashion with no discriminatory component, then its use in a regression is clearly justified. However, suppose that the variable itself is measured in a biased way with women receiving lower evaluations than men because they are women. In this case, one may learn something about the process by analyzing an equation including this variable but the estimated regression coefficient $\hat{\beta}_p$ would not give a true measure of the properly adjusted salary differential. Some ideas from path analysis

concerning direct and indirect effects can be useful here. The safest course is to steer clear of variables which cannot be defended as unbiased measures.

Some variables are direct measures of characteristics which have an obvious relation to salary. Others, on the other hand, provide indirect information about characteristics not easily quantified. For example, age is often used as a proxy for experience, maturity, responsibility and other qualities that generally increase with age.

Although regression as described herein is essentially a linear method, the procedure is sufficiently flexible to provide for a great deal of nonlinearity. Thus, quadratic or other nonlinear functions of the input variables can be used as predictors.

Experience suggests that the relationship between salary and age is not linear over a wide set of ages. There tends to be a leveling off of salary with increasing age. Such a relationship can often be adequately described by using the log of age or by including terms for both age and the square of age in the model. (For numerical reasons it is often better to use the square of the deviation from the average age or some similar central value.) Other nonlinear relationships can be approximated by the inclusion of similar higher-order terms.

Some variables are clearly categorical and should be quantified by using dummy variables. Thus, sex can be coded as in (6). If a variable has more than two possible values, then more than one dummy variable is needed. For example, suppose there are five types of jobs. These can be coded using four dummy variables with X_1 equal to one or zero for job one, X_2 equal to one or zero for job two, etc. In this scheme, each employee will have a one for at most one of the variables X_1, X_2, X_3 and X_4. Employees in job five will have zeros for all four. In general, the number of variables needed is one less than the number of different categories.

Other variables have a somewhat mixed structure being partly categorical and partly continuous. Education is a prime example. Years of education can be a useful quantification. However, degrees obtained are often more informative. Suppose that the highest educational achievement for a set of employees can be categorized as either some high school, a high school degree or equivalent, some college, a college degree, or some post college training. This education information can be coded as follows:

X_1 = years of high school

X_2 = 1 if high school degree, 0 otherwise

X_3 = years of college

X_4 = 1 if college degree, 0 otherwise

X_5 = 1 if some post-college training, 0 otherwise

In this way, the value of both some training and degrees obtained are quantified. The data can be examined to determine the extent to which, this or any other coding scheme quantifies the relationship between salary and education. In some cases, the type of technical training or degree may be important. Furthermore, interaction terms can be constructed. Suppose, for example, that length of service is particularly relevant for a certain job. A variable which is the product of length of service and the dummy variable for the job can be computed. Similarly, by multiplying variables, interactions between other pairs of predictors can be obtained. This approach can be used for categorical as well as continuous variables.

Interactions of variables with sex pose an interesting problem. A regression with all sex interactions can be run and the entire collection of sex-related coefficients tested. This approach is

equivalent to running separate regressions for the two sexes and testing the equality of regression surfaces. The loss of power associated with this method can be substantial and it is therefore not generally recommended. On the other hand, some interactions of this type can be used to pinpoint areas of discrimination or mechanisms related to a discriminatory practice. For example, interactions with job dummy variables can be used to isolate jobs where large sex differentials are present. Similarly, a substantial interaction with length of service could indicate that this factor is given more weight for males than females.

For any given situation, variables which quantify the relevant information can be constructed, examined and refined. This process is often more complex than it appears at a first glance.

In a large data set missing values are often encountered. Ideally, such cases should be identified and the correct values obtained. If this is not feasible, care must be taken in the analysis. Blind use of options provided by software packages can lead to a faulty analysis. First, the extent of the problem must be investigated and described. If only a few cases have missing values, then either reasonable values can be assigned or the cases with missing values can be set aside from the main analysis and examined separately. However, if many cases have missing values on one or more variables then some statistical remedies are necessary. One procedure is as follows. Suppose the variable in question is age. Let X_1 be one or zero depending upon whether age is missing or not. Let X_2 be age. Then by including X_1 and X_2 in the regression, the effect of age will be fit for those employees with age present while a constant term will be fit for the others.

A more subtle problem concerns extreme values and their influence on the regression. Two good sources are Belsley et al. (1980) and Cook and Weisberg (1982). The bottom line is that extreme values can have an enormous effect on the results of least squares procedures.

The first step in dealing with extreme values is to identify them. This can be done with descriptive routines which give minimum and maximum values for each variable, histograms and bivariate plots. The diagnostic plots and calculations available in some regression routines are very helpful in this regard.

Extreme values should be checked for accuracy. If data errors are present, they should be corrected or declared missing. Sometimes a group of employees may have similar extreme values. This situation can lead to the construction of new variables which will improve the explanatory power of the regression.

The problems of missing and extreme values are the subject of active statistical research. Methods have been proposed for replacing missing values by estimates and robust alternatives to least squares have been developed. These procedures may provide good descriptive measures but the effects on hypothesis tests and p-values are not clear. For the latter reason, their use in litigation is somewhat limited.

8. Building the Regression

Planning the details of the analysis before examining the data is an important task. Discussions with salary administrators and others familiar with the process is needed to provide valuable information for building the regression. However, it is a serious mistake to suppose that all potential problems can be forseen in this way.

To produce a good analysis, careful examination of the data at every step is essential. The model building process is iterative with results from previous runs being used to improve the results at the next step.

Difficulties with bad and missing values have been discussed above. Despite assurances to the contrary, such difficulties should always be expected.

Several automatic algorithms are available in software packages for adding and deleting variables in a regression equation. The forward, backward and stepwise procedures can provide some insight but have limited value in the present context. Routines which examine all possible subsets are somewhat better. Although it is not necessary that each variable in the equation be statistically significant, excessive redundancy of information which can cause multicollinearity problems should be avoided. Furthermore, if the number of employees is small relative to the number of variables then undesirable overfitting may result.

The regression framework assumes a homogeneous group of employees in the sense that the effects of the X variables are the same for all. Minor deviations from this assumption can be handled by the inclusion of interaction terms. However, if there are subgroups of employees with vastly different characteristics then separate analyses are usually indicated.

A related issue concerns the construction of separate equations for men and women. Gray and Scott (1980) among others recommend using only the male data to construct the equation. Female salaries are predicted using this equation and compared to actual salaries. This procedure is more suited to the problem of estimating the amount of the discriminatory effect given that one exists, rather than addressing the question of whether or not there is a sex differential. In addition, if the X's for the two sexes differ appreciably there is some danger of extrapolation present with this technique. An example, due to De Groot and described by McCabe (1980) illustrates the difficulties that can arise.

In this example, the women look underpaid using the men's equation while the men look underpaid using the women's equation.

Some insight can be obtained by running the equation with and without X_p. Recall that an R^2 value can be interpreted as the proportion of variation explained by the predictors. The difference in R^2 values thus represents the variation explained by the sex variable over and above that explained by the other variables. The test of the null hypothesis that there is no change in R^2 is identical to the test of $\beta_p = 0$. Explicitly,

$$(8) \qquad t = \left(\frac{(n-p-1)(R_p^2 - R_{p-1}^2)}{1-R_p^2} \right)^{\frac{1}{2}},$$

$$= \hat{\beta}_p / SE(\hat{\beta}_p)$$

where R_p^2 and R_{p-1}^2 are the R^2 values for the full and reduced models, respectively. Thus, $\hat{\beta}_p$ and the R^2 change are two different ways of quantifying the effect of sex on salary. Examination of the R^2 values calls attention to the variation unexplained by any of the available predictors.

A great deal can be learned by examining the residuals from a regression fit. These are defined as

$$(8) \qquad \hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

where $\hat{Y}_i$ is defined in (2). The residuals measure the deviation between the actual salary and that predicted by the regression equation. The least squares method insures that the residuals are uncorrelated with each of the X's and with $\hat{Y}$. Plots of residuals versus these quantities can reveal the need for additional terms in the equation. Such plots are also used to examine the tenability of assumption (7). Roughly, the residuals should look normally distributed about zero with constant

spread. A systematic variation in spread is an indicator that the homogeneity of variance assumption (constant σ^2) is not valid.

Looking at residuals is an art. Detection of extreme values, evidence of nonlinearity and departures from assumptions can all be detected by this technique. It is always a good idea to pick out a few of the most extreme observations and to examine them carefully.

9. Consequences of Using an Inadequate Equation

As stated above, a search for the perfect model is likely to be fruitless. The best we can hope for is to give a reasonably good fit to the data and to avoid misleading summary statistics. In what follows, possible consequences of some model inadequacies are examined.

Extreme values. The least squares method gives these too much weight. Results obtained may be due entirely to the extreme points rather than the bulk of the data. Inflation of the mean squared error results in a lessened chance of detecting statistically significant results.

Neglect of nonlinearities. If salary depends, for example, on length of service in a nonlinear fashion with a leveling off of salary for long term employees and if women have typically less service than men, then an analysis which ignores the nonlinearity can wrongly lead to a conclusion that a sex differential exists. The point is discussed in McCabe (1980).

Multicollinearity. Overlap of information in the predictor variables means that the effect of the common information can be carried by one or more of the variables. This situation produces instability in the regression coefficients. Small changes in the data can produce large changes in the $\hat{\beta}$'s. Furthermore, the affected $\hat{\beta}$'s will have large standard errors. Multicollinearity involving predictors other than the sex

variable is generally not much of a problem and can be alleviated by eliminating one or more variables. If the multicollinearity involves the sex variable, as is often the case, then the resultant instability in $\hat{\beta}_p$ must be recognized when interpreting this statistic.

Lack of constant variance. If the σ^2 in (7) is not a constant, the estimators of the β's are still unbiased but they may be somewhat inefficient. The net effect is some loss of power in the analysis.

Lack of independent errors. The effects of failure of the independence assumption are difficult to ascertain since the assumption can be invalid in so many ways. If the distribution of the errors depends upon sex then the analysis may not give a true picture of the effect of sex on salary.

10. Missing variables

The most serious threat to a valid regression analysis is the existence of important variables which are not included in the regression. Roughly speaking, an important variable is one which, when added to the equation would have a non-zero $\hat{\beta}$ and would also change the value of $\hat{\beta}_p$. Such a variable would contain information, useful for predicting salary, which is not present in the other variables.

Measures of productivity and job performance are the most frequently omitted variables. These measures are often very difficult to quantify with adequate levels of psychometric reliability and validity. Nonetheless, if such considerations are important in salary determination, then (excluding the mathematically possible but practically irrelevant situation in which the information is perfectly correlated with variables used in the equation), the analysis will be biased if there is a sex differential on these measures. This is not to suggest that all salary differentials would disappear if only more variables were available. However,

an analysis with important missing variables must be understood in the context of the limitations imposed by this phenomenon.

Some aspects of the reverse regression dilemma (see the article by Ash in this volume) can be viewed as a missing variable problem. In Birnbaum's (1979) model, he assumes that both salary (Y) and merit (M) are linearly related to quality (Q) with errors from the two equations being uncorrelated with each other and with sex. Sex differences in quality then lead to a non zero sex coefficient in an equation using merit and sex to predict salary. The missing variable is Q minus the conditional expectation of Q given M. It contains information present in quality which is not present in merit but is useful for predicting salary. Birnbaum's model implies that there is a sex differential in this missing variable. A similar view results from an errors-in-variables approach.

11. Drawing Conclusions

The estimated salary differential adjusted for the X's, i.e. $\hat{\beta}_p$, is the primary output from a regression analysis. The extent to which this quantity provides an answer to the question posed in the first sentence of the Introduction depends upon several considerations.

Clearly the data available must be analyzed in a reasonable fashion. Many suggestions for avoiding unreasonable analyses are described above.

Judgments regarding the existence and potential effects of missing variables are needed. Such judgements come primarily from an understanding of the salary determination process under review. Low R^2 values and high MSE's are indicators of missing variables.

The estimated standard error of $\hat{\beta}_p$ quantifies the variability of the estimated salary differential. A very large value relative to $\hat{\beta}_p$, i.e. a small t-statistic, indicates that the salary differential is small relative to the background variation in salary.

With any hypothesis test, the dependence of the results upon the sample size is an extremely important consideration. The estimated standard error of $\hat{\beta}$ goes down roughly as $1/\sqrt{n}$ and hence the t-statistic increases as $\sqrt{n}$. In other words, if one had four times as many similar employees in an analysis, the t-value would be twice as large. As a consequence, the p-value would decrease. Thus, an employer with 400 employees who discriminates might fail a "two-standard deviation" or 0.05 test while a similar employer with 100 employees using the same discriminatory salary practices could pass the test. There is no easy way out of this dilemma. The situation is a direct consequence of the limitations of classical statistical testing methodologies.

In conclusion, regression analysis is a powerful descriptive and modelling tool. Properly used and interpreted, it can provide valuable information for improving legal judgments.

Belsley, David A., Kuh, Edwin and Welsch, Roy E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Birnbaum, Michael H. (1979). "Procedures for the detection and correction of salary inequities," in Salary Equity, eds. Thomas R. Pezzulo and Barbara E. Brittingham, D. C. Heath, Lexington, Massachusetts.

Chambers, John M., Cleveland, William S., Kleiner, Beat, and Tukey, Paul A. (1983). Graphical Methods for Data Analysis. Duxbury Press, Boston, Massachusetts.

Churchill, Neil C. and Shank, John K. (1976). "Affirmative action and guilt-edged goals," Harvard Business Review 54: 111-116.

Cook, R. Dennis and Weisberg, Sanford (1982). Residuals and Influence in Regression. Chapman Hall, New York.

Draper, Norman and Smith, Jr., Harry (1981). Applied Regression Analysis, 2nd ed. Wiley, New York.

Gray, Mary W. and Scott, Elizabeth L. (1980). "A 'statistical' remedy for statistically identified discrimination." Academe , 174-181.

Hocking, R. R. (1983). "Developments in linear regression methodology: 1959-1982," Technometrics 25: 219-230.

McCabe, George P. (1980). "The interpretation of regression analysis results in sex and race discrimination problems," The American Statistician 34: 212-215.

Mosteller, Frederick and Tukey, John W. (1977). Data Analysis and Regression. Addison-Wesley, Reading, Massachusetts.

Neter, John and Wasserman, William (1974). Applied Linear Statistical Models, Irwin, Homewood, Illinois.

Tukey, John W. (1977). Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts.