HOMOGENEOUS JOB GROUPS, CANDIDATE POOLS, AND STUDY OF EMPLOYMENT DISCRIMINATION

Delores A. Conway and Harry V. Roberts
University of Southern California and University of Chicago

1. Introduction

In studies of employment discrimination, a common objective is to compare income measures for employees in a protected class with those for employees not in the class to detect potential discrimination. Throughout this paper, we shall assume that the protected group refers to white females, although the results apply equally to other protected groups.

Discrimination studies often involve some type of regression analysis, to allow for job qualification and performance measures that are related to income and partially explain income differences. The regression analysis may consist of a simple conditioning on the margins of a table or may entail more formal development of regression models. For example, in salary discrimination studies, the analysis may involve fitting regression models to compare mean salaries of males and females after adjustment for differences in measured job qualifications.

The data base for regression analysis is usually derived from personnel records of the organization. The information about individual employees typically includes a record of job positions and salaries since joining the company, previous work experience, important job qualifications such as education or special training, and job performance appraisals. Some of this information may be computerized and readily available for statistical analysis; some of it may be scattered across documents in personnel folders. Some information, such as seniority within the company, may be relatively easy to quantify. Other information, such as actual job performance or quality of prior work experience, may be difficult to quantify from the personnel records alone. For brevity, we shall refer to all potential information about an employee's relevant qualifications or job performance as "job qualifications."

Personnel data bases vary in accuracy and comprehensiveness from organization to organization. However, virtually none are designed with the aim of making causal inferences about the presence or absence of discrimination. The data used in salary regressions can be called "nonexperimental" to distinguish them from data obtained from experiments designed to facilitate inferences about causation. When salary regressions are used in discrimination studies, the task is to infer whether or not discrimination is present, and this inference is more difficult because of the fact that the data are nonexperimental.

In making causal inferences from nonexperimental data, one encounters problems such as model inadequacy or confounding of the effects of job qualifications explicitly included in the model with those from variables not included (the "omitted variables" problem). These problems are important and often hard to cope with fully. Furthermore, omitted variables and model inadequacy can introduce bias in the estimated regression coefficients used to infer discrimination (See eg., Goldberger, 1984).

A less widely recognized problem arises when salary regressions fail to take account of the job structure of the organization under study. When the data come from specialized, heterogeneous workforces, estimated sex effects can reflect the confounding effects of nondiscriminatory income differences across jobs. If the job structure is ignored, an important structural component of the employment process may be omitted from the model specification. Unlike the usual "omitted variables", which are simply not available or are inadmissible for legal reasons, job information can usually be extracted from personnel records.

The central concern of this paper focusses on potential bias in conclusions about discrimination that may result when job is omitted from the conditioning variables used in regression studies. There are two specific concerns. First, in many discrimination studies, the sample comprises the entire workforce of an organization, or some large component thereof, such as a major department. When the relationship between salaries and job qualifications is studied without explicit consideration of different jobs, the omission of job can distort conclusions about salary discrimination.

Second, placement into a job is also an income measure, and salary and placement constitute a bivariate income variable. A comprehensive regression study would take both salary and placement into account. A salary regression that omits job can tell little about possible discrimination in job placement. A fuller perspective of the data is obtained by studying both salary and placement discrimination.

Because an employer can strongly influence the placement of employees into different jobs, it is often argued that job should not be included in a salary regression. This stems from a concern that the analysis might incorrectly exonerate an employer who discriminates in placement of employees. For example, females could be placed in a job for which they are overqualified relative to males holding the same job. But, their salaries could bear the same relationship to job qualifications as the salaries of males in that job. Within the job, there is no discrimination in salary, but the job placement of females entails a type of discrimination that is often called "shunting". Shunting is related indirectly to salary discrimination by the fact that shunting precludes the opportunity to receive the higher salaries available in jobs for which the females' qualifications are appropriate.

The problem posed by the employer's influence on placement is genuine, but we do not believe that it should be solved by avoiding consideration of different jobs in salary regressions. Rather, the scope of analysis should be enlarged so as to be able to detect either salary discrimination or placement discrimination. In this paper, we suggest two stages of analysis for this purpose.

1. Analysis of Salary Discrimination

The different jobs, or relatively homogeneous groups of jobs, represented in a workforce are used as conditioning variables in regression analyses of salary discrimination.

Thus, for males and females, we compare the relationship of salaries to job qualifications "within homogeneous job groups".

2. Analysis of Placement Discrimination

Possible discrimination in placement of employees into jobs is studied separately from salary discrimination. If the data base includes only information about employees currently holding the jobs, the placement study entails a comparison of mean qualifications of males and females within homogeneous job groups, which can be called a "shunting study". If the data base also includes information about individuals considered for particular jobs but not placed into them, regression analysis of the candidate pool can be used to study placement discrimination.

In brief, the two stages in the study of possible discrimination involve the analysis of salaries given job placement, and a separate analysis of job placement. We now turn to a detailed development of this two-stage approach.

2. Salary Decisions and Homogeneous Job Groups

In most large modern organizations, there are many distinct jobs. Many of these jobs may be unique and only held by a few employees at any time. Different jobs require different levels of accountability, supervisory responsibility, knowledge, problem solving ability, and human relations skills.

Although almost every job has unique features, it is usually possible to classify jobs into a moderate number of relatively homogeneous job groups. One criterion for classification might be based on the operational standards of proof that have evolved in legal cases under the Equal Pay Act of 1963, which requires equal pay for the same work. In a case arising under that act, it might be charged that a female "doing the same work" is paid less than males. The legal issue would be whether or not in fact the males and the female are doing the same work. For example, the females might be in one job and the males in a second. If it is determined that the two jobs are essentially the same, in spite of a difference in their titles or descriptions, then discrepancies in pay must be related to differences in job qualifications. Because equal pay cases have typically involved individuals rather than large groups, statistical questions have not been prominent. The evidence simply considers the similarity of the particular jobs in question.

By a homogeneous job group, then, we mean "essentially the same work". It may not be easy to define homogeneous job groups. Different personnel experts could reach somewhat different classifications. However, a serious attempt to make such a classification is essential for study of discrimination.

Individual jobs or homogeneous job groups are often classified into broader groupings designated by "salary grades". The classification is based both on internal job analysis and study of the external job market. Movement of an employee from a job in a lower salary grade to one in a higher salary grade is often regarded as a promotion. Thus, it is natural to talk about lower and higher "job levels".

An important reason for explicit consideration of job groups in discrimination studies concerns the role of job in the employment process. Employment decisions regarding hiring, salaries, and promotion are frequently made with reference to specific jobs. The range of qualifications considered for specific employment decisions is often restricted by consideration of the job. For example, it is unlikely that employment decisions involving executive officers consider the same range of qualifications as those involving mail clerks within the company. The two positions are quite different and require different levels

Figure 1: Direct Regression Fit to a Hypothetical Distribution of Employee Salary and Educational Levels with Allowance for Jobs 1 and 2

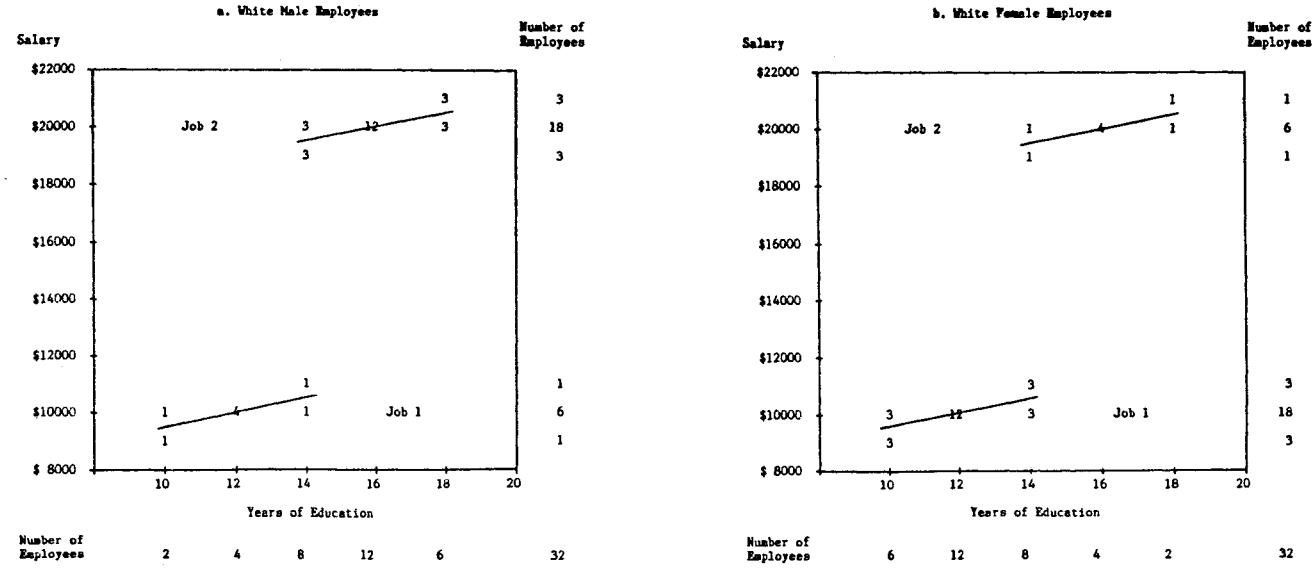


Table 1: Direct Regression Results for Hypothetical Data When Salary is Regressed on Education, Sex, and Job Level

(Salary = 7000 + 0 Sex + 250 Education + 9000 Job)

Variable	b Coefficient	Std. Error(b)	t Statistic
Constant	7000	400	17.48
Sex	0	167	0.00
Educational Level	250	32	7.75
Job Level	9000	167	54.00
	$\text{Adjusted } R^2 = 0.995$	s = 365	

of knowledge, responsibility, and skill. Also, salary levels may vary across different jobs due to external economic factors in the labor market.

Homogeneous job groups help to delineate and reference the relevant range of income and qualification levels actually considered in specific employment decisions. As a result, consideration of homogeneous job groups more closely approximates actual practices used in the employment process.

2.1. A Simple Hypothetical Illustration

Initially we consider a simple hypothetical illustration that shows how salary regressions can give substantially different results depending on whether or not jobs are used as conditioning variables. This hypothetical reflects, in simplified form, characteristic features of data bases used in discrimination studies. For example, Conway and Roberts (1985) consider data from a legal case of employment discrimination with these features.

The hypothetical illustrates a workforce with substantial job heterogeneity. Although job heterogeneity is commonly found in large organizations with specialized workforces, the extent of job heterogeneity in a particular organization can be checked explicitly with the data. When substantial job heterogeneity exists, the hypothetical highlights the need for studying jobs in order to make a comprehensive study of alleged discrimination.

For simplicity, we assume that there are just two homogeneous jobs. Job 1 is in a relatively low salary grade. It carries a target salary of $10,000, and a salary range from $9,000 to $11,000. Job 2 is in a relatively high salary grade. It carries a target salary of $20,000, and a salary range from $19,000 to $21,000.

Suppose further that equal pay discrimination against females is alleged for each of these jobs. We assume that the job qualifications, relevant to salaries, are measured by a single variable "education" that ranges from 10 to 18 years. Ten years reflects some high school education, 12 years reflects a high school diploma, 14 years reflects some college education, 16 years reflects a bachelor's degree, and 18 years reflects an M.B.A. degree. The assumption of a single job qualification is introduced only for expositional simplicity. In practice, additional qualifications would be used.

The hypothetical is constructed so that the equal pay charge is false. The data for the hypothetical appear in Figure 1. Salary and educational levels are graphed for 64 employees in the two jobs. Notice that within each job, the joint distribution of salaries and qualifications is the same for males and females. Consequently, males and females have the same mean salaries for a given level of qualifications and also the same mean qualifications.

If we omit consideration of job, the 32 males in the hypothetical have a mean salary of $17,500 and a mean education of 15 years. Similarly, the 32 females have a mean salary of $12,500 and a mean education of 13 years. Thus, the females in the workforce have lower mean salaries and lower mean qualifications than the males. This result stems from the different numbers of males and females within each job. Notice that there are 24 females and 8 males in Job 1, whereas there are 8 females and 24 males in Job 2.

To consider the regression of salary on qualifications within a job, we introduce two indicator variables. Sex is an indicator that equals 1 for female employees and 0 for males. Job is an indicator that equals 1 for employees in Job 2 and 0 for employees in Job 1. Table 1 summarizes the regression results when Salary is regressed on Education, Sex, and Job. The fitted regression lines are drawn in Figure 1. Notice that the sex coefficient in Table 1 is 0, reflecting the fact that there is no equal pay discrimination within a job for the hypothetical. The estimated coefficient for Job indicates that there are substantial salary differences between the two jobs, for a given level of education.

2.2. Combined and Separate Analyses of Jobs

It is also possible to compute separate regressions for the 32 employees in each of the two jobs. The separate regression results appear in Table 2. The estimated coefficients are the same as in Table 1 and the fitted salaries are the same. But there are interesting differences in other aspects of the fit.

First, the standard deviation of residuals is virtually the same for the combined and separate regressions. The slight difference is purely technical. It arises because the estimate of the standard deviation of residuals is based on 30 degrees of freedom in each separate analysis, whereas it is based on 61 degrees of freedom in the combined analysis.

Second, the standard errors of the constant and the Education coefficient are somewhat larger for the separate regressions. This is due to the greater variation of educational levels in the combined sample, than in the separate samples. The standard errors of the coefficients will be inversely proportional to this variation.

Third, the adjusted R-squared statistic is much lower in the separate analyses. This occurs because the salary variation to be "explained" by regression (i.e. the denominator of the R-squared statistic) is much lower within each job than it is for the combined group of employees. Although the "unexplained" variation is about the same in the combined or separate regressions, the variation to be "explained" is greater in the combined regression. The separate regressions,

Table 2: Separate Regression Results for Job Groups When Salary is Regressed on Sex and Educational Level

Variable	b Coefficient	Std. Error(b)	t Statistic
	a. Employees in Job Group 1		
	Salary = 7000 + 0 Sex + 250 Education		
Constant	7000	572	12.23
Sex	0	152	0.00
Educational Level	250	46	5.39
	Adjusted R^2 = 0.466	s = 371	
	b. Employees in Job Group 2		
	Salary = 16000 + 0 Sex + 250 Education		
Constant	16000	747	21.43
Sex	0	152	0.00
Salary	250	46	5.39
	Adjusted R^2 = 0.466	s = 371	

in effect, disaggregate the data according to job, a major component of overall salary variation.

2.3. Overlapping Qualifications Across Jobs

An interesting subgroup of employees in the hypothetical concerns those with exactly 14 years of education. Some of these employees are in Job 1, whereas others are in Job 2. Because there are large salary differences between the two jobs, the employees with 14 years of education in Job 1 receive substantially lower salaries.

Notice that 6 of the 8 males with 14 years of education, or 75 percent, are in Job 2. The mean salary for all 8 males at this educational level is $17,250. By contrast, only 2 of the 8 females, or 25 percent, are in Job 2. The mean salary for females with 14 years of education is $12,750. The difference in mean salaries between the two jobs results in higher mean salaries for males and lower mean salaries for females, due to the different proportions of males and females in each job at this education level.

It would have been possible to modify the hypothetical so that the qualifications in Jobs 1 and 2 do not overlap. Then the puzzle would disappear. However, if the hypothetical is modified this way, it would have less claim to realism. In actual applications, there is often some degree of overlap in job qualifications across job groups.

To examine more closely the subgroup of employees with 14 years of education, it is helpful to restrict attention to male employees. The 2 males in Job 1 have a salary of $10,500, whereas the 6 males in Job 2 have a salary of $19,500. Education alone does not explain the difference of $9,000 in salary for males in the two jobs.

Similar reasoning applies to the 8 females with 14 years of education. The 6 females in Job 1 have a mean salary of $10,500, whereas the 2 females in Job 2 have a mean salary of $19,500. The $9,000 disparity in salaries for females in the two jobs is the same as that observed for males at this level of education. Thus, education alone fails to explain the salary differences between the two jobs for either male or female employees with 14 years of education.

The reason that the 8 females do worse as a group than do the 8 males stems from the fact that there are three times as many females as males in Job 1 and three times as many males as females in Job 2. If there is a problem of discrimination, it is to be found in placement, rather than in salary discrimination. We would want to examine how employees were selected for Jobs 1 and 2, and why more males than females with this level of education were selected for Job 2.

This subgroup of employees highlights the nonexperimental nature of the data. The observed data alone cannot explain the salary differences between the two jobs. There may be additional job qualifications, other than education, that do explain the differences. Furthermore, the differential mean salaries highlight the need to consider aspects of job placement in order to obtain a full perspective of employment practices.

3. Confounding of Sex Effects when Job is Omitted

Up to now we have used the hypothetical to illustrate an analysis appropriate to an equal pay study. Consider now the perspective of the Civil Rights Act of 1964 (or of the Executive Order), where discriminatory behavior is not restricted to unequal pay for the same work. Many regression studies often ignore consideration of distinct jobs and simply look at the salary and qualifications relationship for employees in the entire workforce. For example, salary discrimination studies might consider the regression of salary on sex and job qualifications, using data for all employees in the workforce.

Table 3 presents the regression results for the 64 employees in the hypothetical without any allowance for job heterogeneity. The fitted regression lines appear in Figure 2. The regression results are very different from those in Table 1.

The appearance of parity in salaries between males and females when Job is included changes to an appearance of salary discrimination against females when Job is omitted. The estimated sex coefficient in Table 3 is $1,800 and statistically significant. This estimated coefficient would suggest an apparent shortfall of $1,800 in the mean salaries of females at a given educational level. In Table 1, there is no sex effect in the fitted model, and the results indicate parity in salaries within job groups.

A second difference is that the regression model in Table 1 provides a closer fit to the data than the one in Table 3. For example, the standard deviation of residuals is $365 from Table 1, rather than $2,550 from Table 3. Also, the adjusted R-squared statistic is higher in Table 1, 0.99 versus 0.75 in Table 3, reflecting a better fit. It is also important to note that

Table 3: Direct Regression Results for Hypothetical Data When Job Level is Omitted from the Model

(Salary = - 6500 - 1800 Sex + 1600 Education)

Variable	b Coefficient	Std. Error(b)	t Statistic
Constant	- 6500	2186	- 2.97
Sex	- 1800	699	- 2.58
Educational Level	1600	143	11.22

Adjusted R^2 = 0.746 s = 2550

the coefficient of Job in Table 1 is significant and indicates that Job is an important predictor of employee salaries. Finally, there are differences in the estimated coefficient for Education. The estimated coefficient in Table 3 is $1,600, as opposed to $250 in Table 1, and has a much larger standard error.

Two key assumptions underlie the hypothetical and lead to important differences in regression results. First, there are substantial salary differences between the two jobs with much higher salaries in Job 2. Second, over the entire workforce, there are three times as many males as females in Job 2. This leads to a difference in the bivariate distribution of salary and qualifications over the entire workforce for males and females. Overall, males have higher salaries and qualifications than females. It is only within each job that the male and female salary and qualification distributions coincide.

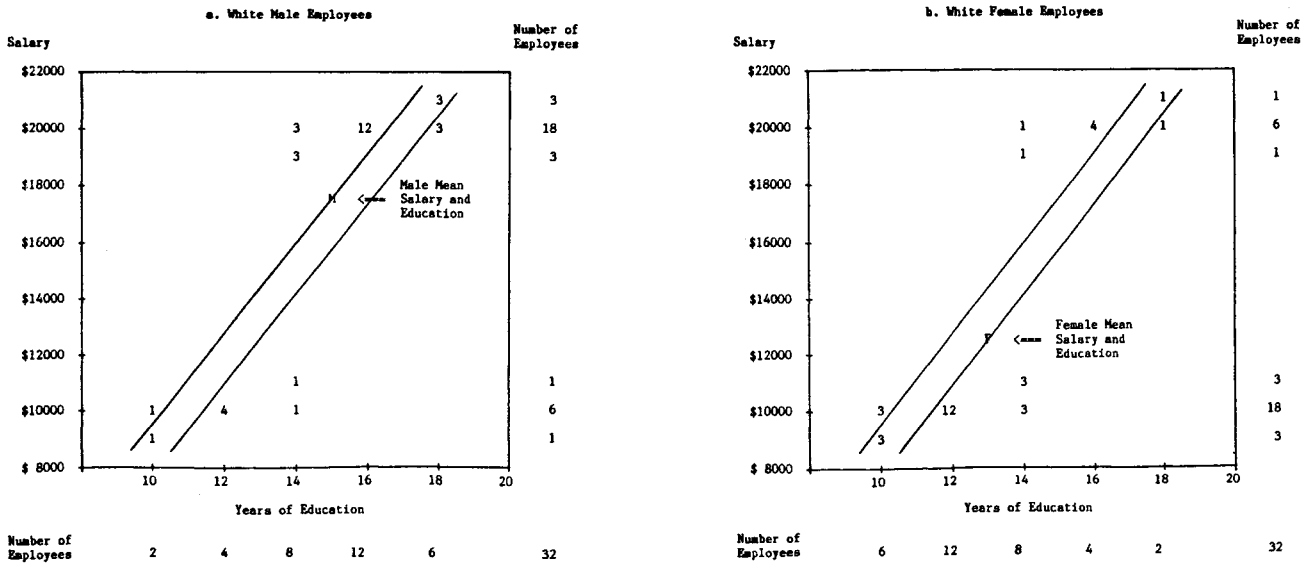
Consequently, an impression of discrimination appears for the hypothetical when job level is omitted from the regression analysis, even though there is no salary discrimination within job groups. This is due to the fact that there are different proportions of females and males within the two jobs. Irrespective of whether we consider males alone or females alone, there is an estimated $9,000 salary difference between Jobs 1 and 2 for a given level of education. The salary difference between jobs combines with the different proportions of males and females within each job to result in the $1800 estimated salary shortfall for females from Table 3.

The hypothetical can be summarized by saying that the effects of Sex and Job are confounded. For any given level of education, Job and Sex are related. Because 75 percent of the females are in Job 1 and 75 percent of the males are in Job 2, the value of the job indicator variable coincides exactly with the values of the indicator variable, 1 - Sex, for 75 percent of the employees. When Job is omitted from the regression, the sex variable proxies for Job and shows a negative relationship with salaries given education. To the extent that the hypothetical reflects relationships found in real data bases, an appearance of overall salary discrimination could result even when there is no salary discrimination within job groups.

We suggest that, as a first stage, it is useful to study salary discrimination from the perspective of the Equal Pay Act. For this purpose, an appropriate tool is the comparison of male and female mean salaries for given job qualifications within homogeneous jobs. Alternatively one can use job indicator variables in a single regression analysis that includes all the individual job groups. Equal pay regressions require careful development of the data base to form relatively homogeneous job groups.

Disaggregation by homogeneous job groups is limited by sample size because there is relatively low power for detecting possible salary discrimination in job groups that have few employees. In legal cases, however, the main focus is usually on the aggregate picture, not on small subgroups of employees. A measure of the aggregate difference between males and females can be defined by combining the differences found in smaller subgroups, as in the method of standardized

Figure 2: Direct Regression Fit to the Hypothetical Distribution of Employees in Figure 1 with No Allowance for Job Level



averages. This aggregate measure can have high precision even when precision is low in individual job groups. A useful treatment of this general strategy is provided by Mosteller and Tukey (1977, Chapter 11).

Of course, if disaggregation is carried to the point that there are as many distinct jobs as there are employees, even standardized averages will not help. On the other hand, sometimes relatively large subgroups, such as all professional employees in one group and all clerical employees in another, may give reasonable homogeneity. In each study, some compromise between large and small subgroups must be reached. In deciding how far disaggregation should be carried, detailed study of company organization, personnel practices, and job descriptions might supplement statistical analyses based on different levels of disaggregation.

4. Two-Stage Analysis of Placement and Salary Discrimination

We have seen that failure to consider job heterogeneity creates serious difficulties when interpreting overall salary regressions. We now turn to the question of additional analysis required to extend an equal pay study to the analysis of discrimination under the Civil Rights Act.

Regression of salary on job qualifications and Sex, without conditioning on job, is one way to extend an equal pay study to a general civil rights study. This approach is based on the concept of nondiscrimination as equal expected pay for given measured job qualifications. However, the analysis for the hypothetical suggests that serious confounding problems may result from this approach.

It is helpful to ask how discrimination might exist, even when there is no evidence of salary discrimination within jobs. One answer is found in the placement of employees into the jobs. Job placement decisions, like salary decisions, are substantially influenced by the employer. Consequently, placement of males and females into different jobs might be discriminatory.

Instead of trying to study the combined effects of both salary and placement discrimination by an overall salary regression that does not condition on jobs, we suggest separate study of the two types of discrimination. First, salary discrimination can be studied through salary regression analyses within job groups. Second, placement discrimination can be studied by reference to the candidate pool of employees considered for the job.

The approach is based on the study of two different employment decisions. First, among the potential candidates for a job, one is selected and hired by the company. This type of employment decision illustrates what we have called "Type-2 Employer Behavior" (eg. see Conway and Roberts, 1985). Second, for the placed candidate, a separate salary decision is made with reference to the employee's qualifications. The second decision illustrates "Type-1 Employer Behavior." When these employment decisions are repeated many times across candidate pools for the job, what emerges is a process whereby the employer selects candidates for placement into a particular job and then prices the qualifications of placed employees to determine salaries. For discrimination studies, we might first evaluate the pricing of qualifications as Stage 1 of the analysis, and then separately consider the placement process as Stage 2 of the analysis.

The rationale can be stated in probability notation by considering the joint distribution of salary and qualifications within a job. Let Job refer to the particular job considered in the employment decisions. Let S and Q represent the respective salary and qualification levels observed for employees placed into the job. Because the exact relationship between salaries and qualifications within a job is not known, we consider the joint distribution of salaries and qualifications to assess potential discrimination. Consequently, the item of interest is $P(S,Q|Job,Sex)$. This joint distribution can be factored as the product of a conditional and marginal distribution in the following way:

$$(1)\quad P(S,Q|Job,Sex) = P(S|Q,Job,Sex)\ P(Q|Job,Sex).$$

Stage 1 of the analysis entails consideration of the first factor on the right of (1), whereas Stage 2 corresponds to the second factor. If Job does not contribute to the fit in Stage 1, it is unnecessary to condition on it. If so, we would conclude that the employer's pricing of qualifications is carried out irrespective of job. Then, $P(S|Q,Job,Sex) = P(S|Q,Sex)$ and

$$(2)\quad P(S,Q|Job,Sex) = P(S|Q,Sex)\ P(Q|Job,Sex).$$

Note, however, that Job is still part of the picture in (2) due to the factor, $P(Q|Job,Sex)$. This factor requires separate statistical study for a complete audit of possible discrimination. Relevant background information about employment practices, obtained from memoranda, manuals, or interviews with personnel managers and employees, may also be pertinent to the study of the placement factor.

5. Placement Decisions and Candidate Pools

Placement of individuals into jobs illustrates an important aspect of the employment process that involves selection decisions. For example, the initial search for potential employees entails selection of candidates who receive serious consideration for hiring. So does actual hiring of some of these potential employees. Hiring may be simultaneous with placement into particular jobs, but sometimes placement occurs only after an initial training program.

Other examples of selection decisions are promotion to higher job levels, transfer, demotion, discharge, and even forced early retirement. If the data base has information about the rejected as well as selected individuals, selection decisions can be studied directly to assess potential discrimination. If only data on those selected are available, a more limited analysis is possible through shunting studies.

We suggest the term "candidate pool" to designate the group of people considered for an employment selection decision. Our use of the term "candidate pool" is always with reference to a particular job or a homogeneous job group. Consequently, members of the candidate pool for a

selection decision will typically have a more narrow range of job qualifications, than those present across all jobs in the firm.

As an illustration, consider the job of insurance adjustor for car accident claims in an insurance company. In the company's classification of jobs, this job might have a specific title and job description. It may also carry a fixed target salary, and a range of salaries above and below the target salary in which management can adjust salaries to accommodate differing qualifications of individual employees.

The requirement to hire one or more adjustors is derived from the company's need to accomplish a specialized function required by market demand. Management searches for eligible candidates, possibly both inside and outside the company. The aim is to find candidates who are neither overqualified or underqualified for the job. Such candidates constitute the candidate pool for the placement decision.

5.1. Shunting Studies for Placed Candidates

For a complete study of the placement process, it is desirable to have information on all candidates considered for the job, namely those rejected as well as those selected. Information about the rejected candidates is often absent from data bases used in employment discrimination studies. This may reflect the limited business necessity to retain for long periods of time detailed records of job applicants not hired.

If only information about the placed candidates is available, a more limited analysis of placement discrimination is still possible. Specifically, we can compare the mean qualifications for males and females within the same job to detect possible "shunting" of females into jobs for which they are overqualified. Such a comparison is called a shunting study. Whenever it is reasonable to assume that the qualifications of placed employees within the job are similar to those of candidates rejected for the job, the shunting comparison provides an assessment of placement discrimination.

For the hypothetical data in Section 2, the mean qualifications of males and females are the same within each of the two jobs. The mean educational level is 12 years for Job 1 and 16 years for Job 2. If males and females rejected for each job have similar qualifications to those observed for the placed employees, the shunting comparison suggests the absence of placement discrimination.

The hypothetical uses a single job qualification for simplicity of illustration. One can also use multiple qualifications to perform a shunting study. The pricing of qualifications within homogeneous job groups is modeled by the regression of salary on qualifications and Sex for each homogeneous job group. This regression provides a natural qualification index that is independent of Sex. The fitted values from the regression, minus any sex effect, provides an estimate of how the employer prices job qualifications without regard to sex.

We can compare the mean qualification index for males and females within a job group to assess shunting. This comparison is equivalent to regressing the qualification index on the single variable, Sex. The shunting comparison is called a simple reverse regression, because the qualification index appears as the dependent variable rather than as an independent variable. Hence, simple reverse regression provides a method to check for shunting.

Analysis of shunting may not yield a conclusive verdict about possible placement discrimination. For example, although males and females actually placed into each job may have the same mean qualifications, it is possible that females considered but not selected could have higher mean qualifications than those of employees actually selected. This would suggest discrimination against the non-selected females. The fact that many data bases, like that of the hypothetical, do not have information on rejected job candidates raises a problem of omitted people.

When information about the omitted candidates is not available, the check for shunting carries the investigation of possible placement discrimination as far as the data permit. Sometimes, there will be good reason to believe that a shunting study provides a useful evaluation of placement discrimination. This will be true when it is realistic to assume that the males and females considered for a particular job, whether or not actually placed into the job, have similar distributions of job qualifications. For example, candidates for a particular job may have a relatively narrow range of qualifications. Within this narrow range, the conditional distributions of males and females are likely to be similar.

Even when one is unwilling to assume equal average qualifications of male and female candidates for a job, shunting studies may still provide some information about placement discrimination. For example, suppose that for a given job, an employer hires only the most outstanding female candidates, but hires all males. Then, females actually hired would have higher mean qualifications than males, and the shunting study would correctly suggest placement discrimination. Salary regression alone, even if conditioned on job, would fail to reveal any problem, so long as the employer paid the females as well as equally qualified males.

5.2. Full Information about the Candidate Pool

If full information is available on all members of the candidate pool for the job, we can study an employer's placement practices in a similar way as salary practices. Let H be an indicator variable that equals 1 if the candidate was selected for the job and 0 otherwise. Among applicants in the candidate pool, the bivariate distribution of hiring decisions and candidate qualifications can be factored as,

$$(3)\quad P(H,Q|Sex,Job) = P(H|Q,Sex,Job)\ P(Q|Sex)$$

The first term on the right of (3) might be studied through a logistic regression analysis of hiring decisions given applicant qualifications. Let $p = P(H|Q,Sex)$ represent the conditional probability of being hired given a candidate's qualifications and sex. Then, the logistic regression model has the form,

$$(4)\qquad E[p/(1-p)] = a + b'Q + c\ Sex.$$

The coefficient of Sex in (4) can be used to assess placement discrimination in the same way that the corresponding salary regression coefficient assesses salary discrimination.

In one sense, we are simply pushing the idea of direct salary regression back one stage and applying it to placement. Direct salary regressions consider only those hired into the job. By contrast, the logistic regression considers all applicants in the candidate pool for the job.

We could also push back the analysis one additional stage to audit the search process in the formation of the candidate pool from some larger group of potential candidates. This entails analysis of the second factor on the right side of (3). If complete information on members of the larger group is available, a second logistic regression could assess the search process. The same rationale for the placement logistic regression applies.

If information about the larger group of potential candidates for selection is not available, a shunting analysis can be performed. We can construct the placement qualification index, $a + b'Q$, using the estimated coefficients from the placement logistic regression in (4). Comparison of the mean values of this index for males and females results in a shunting study. Higher mean qualifications for females would provide some reason to suspect shunting in the search process. For example, females might be encouraged to apply for less desirable jobs in the company.

To illustrate further the problems posed by the search process, consider a simple hypothetical example. Suppose that in a study of an airline in the early 1970s, it was found that no females had been placed into the pilot job, so no shunting study was possible. The natural next step would be study of the formation of the candidate pool for pilots to determine whether or not there were qualified female pilots who had been overlooked by the airline during the formation of the candidate pool or passed over in selection from the pool.

Thus we have extended the strategy to three stages, each corresponding to a possible type of discrimination:

1. The search process that results in the formation of candidate pools for placement.

2. The placement process that results in the selection of employees from job candidate pools.

3. The salary process that results in the pricing of employee qualifications within specific jobs.

5.3. Detecting Discrimination in Job Definitions

We have now suggested several stages and methods for studying potential discrimination in the employment process. There may well be other stages and methods of analysis needed to highlight different manifestations of potential discrimination. For example, the proposed analysis conditions throughout on homogeneous job groups and the job structure of the firm. It is also necessary to consider whether the job structure itself might conceal possible discrimination.

Suppose that job definitions are gerrymandered so that one job is made into two, one high-paying and the other low-paying. Specifically, mean qualifications of males and females would be equal within and between these pseudo "jobs", but females would predominate in the lower-paying "job". An even more subtle example would be two identical jobs with equal current pay for given qualifications but with unequal opportunities for advancement.

To uncover such deception, careful study of the formation of candidate pools is important. Moreover, it is essential to study the details of job evaluation, description, and classification. A central argument of this paper, with its emphasis on the analysis of homogeneous job groups, is that such details are important in deciding whether discrimination has occurred. Investigation might extend beyond any information normally available in personnel records. For example, a memorandum, or personnel manual might reveal the employer's intent to gerrymander job definitions.

Legal considerations, as well as limitations of data, shape the statistical analyses that are possible. Any given case is relatively specific as to the types of alleged discrimination. For example, if the protected class is defined to include only current employees, or only current and past employees, rejected members of applicant pools will not be brought into the picture. Placement discrimination can then be examined only by shunting studies, regardless of availability of data.

5.4. Extensions to Other Stages in the Employment Process

Consideration of the search process in the formation of candidate pools is suggestive. One can define many other stages of the employment process at which discrimination might occur. In addition to search, placement, and salary determination, there are salary increases, promotions, demotions, terminations, forced retirement, and so on. The two-stage process of Section 4 can be expanded to a larger number of stages for analysis. If we arrange these stages in temporal sequence, a comprehensive analysis would consist of a series of conditional analyses, one for each stage.

How far such analyses can be implemented in any application depends in large part on the availability of data. But, the statistical blueprint is clear. With ingenuity in application, we have some hope of ferreting out even very subtle types of discrimination and of understanding more fully the complexities of employment practices.

6. The Dual Role of Reverse Regression

In the previous section, reverse regression provides a method for analysis of placement decisions through shunting studies. This is one important role for reverse regression, especially in studies where information may only be available about placed employees. A second role for reverse regression concerns its use when important factors in the employment process may be omitted from the analysis. Then, reverse regression provides a different method of conditioning the bivariate distribution of income and qualifications than the

direct regression approach. Different conclusions about potential discrimination may result from the two regressions and lead to consideration of other critical factors that may be omitted from the analysis.

6.1. Reverse Regression and Heterogeneous Workforces

From the probability relationship in (2), direct regression provides a method for analyzing the salary factor P(S|Q,Sex,Job), whereas, reverse regression provides a method for analyzing the placement factor, P(Q|Sex,Job). When the analysis is directed to homogeneous job groups, both direct and reverse regression have a function to perform. Now consider what happens when we consider direct and reverse regression for a heterogeneous workforce when job information is omitted from the analysis. Much of the controversy about the merits of direct and reverse regression has taken place in this context. (See e.g., Ferber and Green et. al., 1984).

For a heterogeneous workforce, the direct salary regression of S on Q and Sex, without allowance for Job, considers all employees as a single candidate pool. The direct salary regression models the employer's salary decisions as an overall pricing of qualifications, regardless of job. However, if in fact the employer prices qualifications only within homogeneous job groups, this use of direct regression is beset by the problem of confounding with the omitted job variable, as illustrated in Section 2.

The reverse regression of Q on Sex, without allowance for Job, provides the basis for a shunting study. But, if the employer does price qualifications only within homogeneous job groups, reverse regression loses its simple interpretation as a check for shunting.

However, reverse regression can still provide useful information about shunting whenever salary and job levels are closely related. If salary variations within homogeneous job groups are relatively small by comparison with salary variations between jobs, salary can proxy for the omitted job variable. Reverse regression applied to the heterogeneous group then provides an approximate comparison of the qualifications of males and females within homogeneous jobs, hence an approximate check on the possibility of shunting. Also, if salary variations within jobs are relatively small compared to salary deviations between jobs, shunting looms large in the overall picture.

The hypothetical data is used to illustrate this application of reverse regression. Recall that the hypothetical was constructed so that the bivariate distribution of S and Q is the same for males and females within each job. Consequently, P(S|Q,Sex,Job) = P(S|Q,Job) and P(Q|Sex,Job) = P(Q|Job). The results for the reverse regression of Education on Salary, Sex, and Job appear in Table 4a and show no sex effect.

When the job variable is omitted from the direct regression of S on Q and Sex, the Sex coefficient in Table 3 shows a female salary shortfall of $1800 for given qualifications. By contrast, the reverse regression results in Table 4b show virtual parity of education at given salaries. For a given salary, females have only slightly more education, 0.11 years. The fitted reverse regression lines, expressed in terms of salary, appear in Figure 3 and nearly coincide for males and females.

Hence, to the extent that the hypothetical depicts essential features of a heterogeneous workforce, reverse regression can serve as a rough approximation to what would be found in a fuller analysis that takes job into account. Intuitively, the approximation seems to work for the following reasons. First, the variance component for salaries between jobs may be relatively large by comparison to the variance component for salaries within jobs. Hence, salary may be a good proxy for job. Second, the qualification index based on the entire workforce reflects an averaging of the weights that would be obtained from within job analyses. The overall index may be positively and highly correlated with indices obtainable from within job analyses.

Figure 3: Reverse Regression Fit to the Hypothetical Distribution of Employees in Figure 1 with No Allowance for Job Level

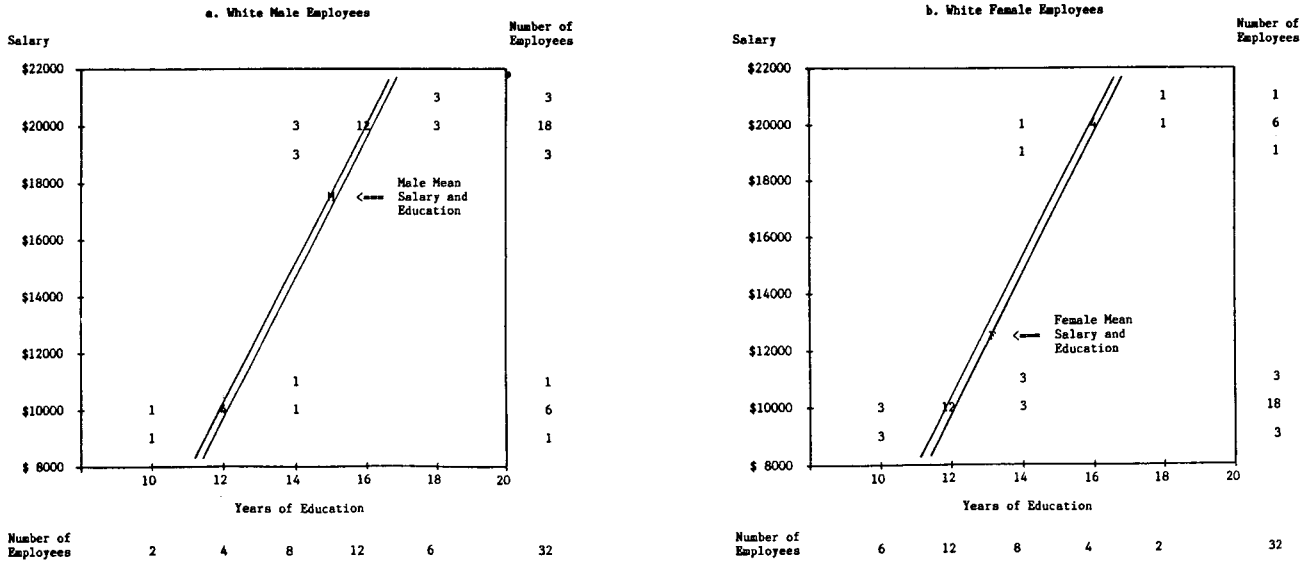


Table 4: Reverse Regression Results for Hypothetical Data When Educational Level is Regressed on Salary and Sex

Variable	b Coefficient	Std. Error(b)	t Statistic
	a. Inclusion of Job Level in Model		
	Education = - 8 + 0 Sex + 0.002 Salary - 16 Job		
Constant	- 8.000	2.598	- 3.08
Sex	0	0	0.00
Salary	0.002	0.000258	7.75
Job Level	- 16.000	2.599	- 6.16
	Adjusted R^2 = 0.825	s = 1.033	
	b. Exclusion of Job Level from Model		
	Education = 7.63 + 0.105 Sex + 0.0004 Salary		
Constant	7.632	0.696	10.96
Sex	0.105	0.377	0.28
Salary	0.000421	0.0000375	11.22
	Adjusted R^2 = 0.719	s = 1.308	

6.2. Reverse Regression and Omitted Variables

In the study of employment practices, reverse regression provides a different way to factor the bivariate distribution of qualifications and income. For example, in the analysis of salary decisions, an alternate way of factoring the bivariate distribution of salaries and qualifications within a homogeneous job group is given by,

$$(5)\quad P(S,Q|Sex,Job) = P(Q|S,Sex,Job)\ P(S|Sex,Job).$$

The factor, P(Q|S,Sex,Job), is a "reverse regression" of qualifications on Salary and Sex within a job group.

The data base available for discrimination studies may contain only a subset of job qualifications used by the employer in actual employment decisions. This stems from fact that the data are observational and not experimental. Assessments of possible discrimination from the direct regression, P(S|Q,Sex,Job), may be biased when there are problems of measurement error and omitted variables in the observed qualifications. Reverse regression provides an alternative assessment of possible discrimination for this case.

The assessments of discrimination from direct and reverse regression coincide whenever the mean qualifications between males and females coincide or income and qualifications are perfectly related. Within homogeneous job groups, we would expect a more narrow range of qualifications than across the entire workforce. Consequently, potential conflicts in the assessments of direct and reverse regression are reduced by conditioning on homogeneous job groups.

However, even within homogeneous job groups, it is instructive to consider both the direct and reverse regression perspectives to evaluate fairness of employment practices. If the conclusions coincide, then either perspective arrives at the same conclusion. However, if the conclusions about possible discrimination do not coincide, this may suggest an important factor has been omitted from consideration. For example, the job groups may be heterogeneous, rather than homogeneous, and external information might validate this possibility. Further disaggregation of the data into homogeneous job groups may be required.

An important consideration for discrimination studies is that there are many factors that enter into employment decisions. Furthermore, our understanding of the employment process is still evolving. When used in tandem, direct and reverse regression help to enlarge our perspective of the employment process and highlight potential biases that may arise from the nonexperimental nature of the observed data.

7. References

Conway, Delores A. and Roberts, Harry V. (1983), "Reverse Regression, Fairness and Employment Discrimination," Journal of Business and Economic Statistics, 1, 75-85.

_____ (1984), "Rejoinder to Comments on 'Reverse Regression, Fairness, and Employment Discrimination'", Journal of Business and Economic Statistics, 2, 126-139.

_____ (1985). "Regression Analyses in Employment Discrimination Cases", in Statistics and the Law, ed. DeGroot, M., Fienberg, S. and Kadane, J.. New York: John Wiley and Sons. (To appear)

Ferber, Marianne A. and Green, Carole A., et. al. (1984), "Discussion of the Statistical Analysis of Fairness and Employment Discrimination", Journal of Business and Economic Statistics, 2, 111-125.

Goldberger, Arthur (1984), "Redirecting Reverse Regression", Journal of Business and Economic Statistics, 2, 114-116.

Mosteller, Frederick and Tukey, John W. (1977), Data Analysis and Regression. Reading, Mass.: Addison-Wesley Publishing Company.

Roberts, Harry V. (1979), "Harris Trust and Savings Bank: An Analysis of Employee Compensation", Report 7946, CMSBE, Graduate School of Business, University of Chicago.