This paper examines methods of decomposing a difference in levels between groups for a dependent variable such as income. Applied to regression equations, this technique estimates the contribution to the difference from divergent characteristics and divergent rates of converting characteristics into the dependent variable. The consequences of an "interaction" component being present in the decomposition is examined. The paper, using data from the 1960 Census, shows how ignoring the interaction term can influence results.

DECOMPOSITION OF DIFFERENCES

A Cautionary Note

HOWARD M. IAMS
Hope College

ARLAND THORNTON
University of Michigan

In comparative social research investigators frequently search for explanations of social differences. Investigators have studied racial differences (Duncan, 1967, 1969); sexual differences (Cohen, 1971; Malkiel and Malkiel, 1973; Suter and Miller, 1973; Levitan et al., 1971); time differences (Miner, 1960); and welfare group differences (Schiller, 1970). Whether groups were defined by income, race, sex, or time, these investigators attempted to disentangle the factors producing differences between groups in the level of the dependent variable being studied. Furthermore, all of these investigators applied the procedure of demographic standardization to the

AUTHORS' NOTE: *We wish to thank Robert P. Althauser, Duane F. Alwin, O. D. Duncan, Reynolds Farley, Albert I. Hermalin, and James N. Morgan for commentary on various drafts of this paper. Of course, we alone are responsible for any errors contained in the content of the paper. We are grateful to the Population Studies Center of the University of Michigan for support and for making data available for the study.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 3 No. 3, February 1975

results of least-squares regression equations to answer their research problem. This paper cautions researchers about one of the problems associated with this technique.

The studies cited above tried to separate and to estimate the effect of groups having different values on characteristics, or composition, and the effect of the groups having different rates at which they convert their characteristics into values on the dependent variable. For example, it may be that blacks generally have less education and work in lower paying jobs than do whites, and that this poor background of blacks contributes to their lower income. These lower characteristics, the composition of the group, contribute to the lower income of blacks. In addition, it may be that blacks receive less income for doing the same type and amount of work. That is, they may not be able to convert their characteristics into income at the same rate as whites. It would be enlightening to estimate separately these components of the income differences between the races. The ability to separate the effect of lower characteristics from the effect of lower returns to characteristics obviously has widespread application in social science.

Kitagawa (1955) presented methods for decomposing the difference between the values of two social groups on the same dependent variable. Winsborough and Dickinson (1969) and Althauser and Wigler (1972) have shown how this technique can be extended to dependent variables utilizing regression techniques. As Althauser and Wigler indicate, the regression decomposition technique involves estimating a means or composition component by weighting the differences in composition (means) by a set of regression coefficients. Similarly, a slopes or rates component is estimated by weighting the differences in coefficients by a set of means. The researcher can choose weights from just one of the populations, or the weights can be combinations of the means and coefficients of both populations.

PROCEDURES FOR DECOMPOSITION OF DIFFERENCES

We shall utilize the notation of Althauser and Wigler in summarizing the procedures for decomposing a difference between two means. In the usual regression format let:

Y_w = the overall mean of the dependent variable for whites;

Y_n = the overall mean of the dependent variable for blacks;

X_{iw} = the mean of the ith explanatory variable for whites;

X_{in} = the mean of the ith explanatory variable for blacks;

b_{ow} = the regression constant or intercept for whites;

b_{on} = the regression constant or intercept for blacks;

b_{iw} = the partial regression coefficient for the ith explanatory variable for whites;

b_{in} = the partial regression coefficient for the ith explanatory variable for blacks.

We also know that $Y_w = b_{ow} + \Sigma b_{iw} X_{iw}$ and that $Y_n = b_{on} + \Sigma b_{in} X_{in}$.

The decomposition of the difference $Y_w - Y_n$ is usually handled in one of three basic ways. The first (shown in equation 1) utilizes weights from just one of the populations (in this case blacks). (This is equivalent algebraically to equation 20 in Althauser and Wigler, 1972.) This procedure results in four components. The first component is the intercepts component reflecting the difference in the intercepts of the equations for the two groups. The second component is the rates or coefficients component reflecting the differences of the slopes. The third term is the composition component. This component shows the part of the overall difference produced by differences in the means of the independent variables. The fourth term is usually called the interaction component. Technically, it is not interaction in a statistical sense, but reflects the covariation or collinearity between the means and the coefficients of the two populations. This component can be interpreted, following

Winsborough and Dickinson (1969), as the effect of changing both means and regression coefficients together over the effects of changing them one at a time.

$$Y_w - Y_n = (b_{ow} - b_{on}) + \Sigma X_{in}(b_{iw} - b_{in}) + \Sigma b_{in}(X_{iw} - X_{in}) + \Sigma(X_{iw} - X_{in})(b_{iw} - b_{in}) \qquad [1]$$

The second procedure involves utilizing coefficient weights from one population and composition weights from the other. In equation 2 (taken from equation 18 of Althauser and Wigler) we decompose the difference using the black means to weight the differences in regression coefficients and use the white regression coefficients to weight the differences in means. Alternatively, in equation 3 (equivalent to equation 21 of Althauser and Wigler) we decompose the difference using the white means and black coefficients as weights.[1] It should be noted, however, that both of these decompositions are equivalent to equation 1 except that they include the interaction component with one of the other components. Equation 2 includes the interaction component with the composition component, and equation 3 includes it with the rates component.

$$Y_w - Y_n = (b_{ow} - b_{on}) + \Sigma X_{in}(b_{iw} - b_{in}) + \Sigma b_{iw}(X_{iw} - X_{in}) \qquad [2]$$

$$Y_w - Y_n = (b_{ow} - b_{on}) + \Sigma X_{iw}(b_{iw} - b_{in}) + \Sigma b_{in}(X_{iw} - X_{in}) \qquad [3]$$

The third decomposition procedure is shown in equation 4 (equivalent to equation 23 of Althauser and Wigler). This procedure utilizes an average of the two populations as weights. The resulting decomposition is the same as that shown in equation 1 except that half of the interaction component is, in effect, added to the rates component and half to the composition component (Winsborough and Dickinson, 1969).

$$Y_w - Y_n = (b_{ow} - b_{on}) + \Sigma \frac{(X_{iw} + X_{in})}{2} (b_{iw} - b_{in}) + \Sigma \frac{(b_{iw} + b_{in})}{2} (X_{iw} - X_{in}) \qquad [4]$$

The choice of the decomposition utilized depends on the issue being investigated and the questions being asked. Sometimes the researcher is interested in only one component. The question may be whether or not composition differences by themselves could account for the income differences. Similarly, but from a policy framework, one might ask if the income difference could be eliminated by "giving" blacks the same characteristics as whites. In those cases one would be interested only in the composition component, and one could focus on the third component of either equation 1 or 3 (both components being identical). Similarly, if one were only interested in rates differences, a focus on component 2 of equation 1 or 2 would be appropriate.

The purpose of the research is often to compare the magnitudes of the various components. We may want to know whether "composition" produces more income differences than do "rates." Similarly, from a policy standpoint, the issue may be whether a change in composition will produce more changes in income than will a change in rates. For these cases equation 1 seems to provide the best choice. It allows both the rates component and the composition component to be weighted by values from the same population rather than from some mixture of two populations. As we mentioned earlier, this decomposition produces an interaction term which may be interpreted as the effect of composition and rates beyond their individual effects. Some researchers may choose to simplify this decomposition by distributing the interaction term equally to the rates and composition components. In those instances equation 4 will be the appropriate formula.

Equations 2 and 3 also can be utilized in making comparisons between the magnitudes of the components, but special care must be taken in the interpretation. For policy purposes one might want to know which component is largest under the conditions that one first changes the intercept and regression coefficients and then changes the composition (means). For that comparison equation 2 would provide the proper decomposition. On the other hand, one might be interested in comparing component magnitudes under the conditions that one first changes composition and then changes intercept and coefficients. Then equation 3 would be correct. However, neither of these two decompositions allows us to compare the magnitudes of the coefficients under similar conditions. If our interest is whether a change first in rates would produce greater income changes than a change first in composition, then equation 1 again becomes the appropriate decomposition. In many cases this last question would seem to be the most interesting.

It is the interaction component and its placement (explicit or implicit) in the equation that produces the differences among the four equations. If the interaction component were zero (or nearly zero), the four decompositions would be identical. However, when the interaction component is large, the differences among the four decompositions become substantial.

Failure to recognize the possible importance of the interaction component can result in improper utilization of equation 1. A researcher, interested in only the first three components, might compute any two of these three and then assume that the remaining difference was due to the component of interest not computed. This would implicitly add the interaction component to that uncomputed component. This problem can be illustrated by the work of Levitan et al. (1971) and Miner (1960). In their study of the income difference between men and women Levitan et al. (1971) calculated that $3,458 of the total income difference between men and women was due to the combination of the intercept and rates components and then inferred that the remaining $914 was the effect of achievement or composition factors. It appears that they included the interaction component with the composition component. On the other hand, Miner (1960) appears to have included the interaction component with the intercept component. This is implicit in his statement that "if, then, neither changes in the overall effects of the independent variables nor changes in their mean values account for the upward shift in the proportion of debtors there remains only the constant term as the explanatory factor."

In considering this issue it is important to realize that the size of the interaction component can be nontrivial. In addition, its influence on interpretation may not be consistent for comparisons. This can be illustrated from census data on earnings of white men and black women. The example considers black women as the base group in the decompositions.[2]

ILLUSTRATION

We will be examining differences in hourly wages between white men and black women in this example. The differences between white men and black women for the age groups 20 to 34 and 35 to 44 will be examined separately. Focusing first on the age group 20 to 34, we see from Table 1 that white men earned $2.39 per hour while black women earned $1.10. That is, black women earned $1.29 per hour less than white men or only 46% of the wages of white men.

TABLE 1
Mean Levels of Age Adjusted for Tenure, Occupational Prestige, and Hourly Wages for Black Women and White Men by Age, 1960

	Age 20 to 34		Age 35 to 44	
Variables (1)	Black Women (2)	White Men (3)	Black Women (4)	White Men (5)
Age adjusted tenure (years)	5.3	10.7	7.6	15.0
Occupational prestige	30.5	39.5	26.5	41.1
Education (years)	10.7	11.9	9.9	11.5
Hourly wages (dollars)	1.10	2.39	1.16	2.99

NOTE: The sample consists of full-time wage and salary workers in the civilian nonagricultural labor force. Other tables refer to this sample.
SOURCE: One-in-one thousand sample of the Census of the Population: 1960.

TABLE 2
Partial Regression Coefficients from the Regression of Hourly Wages on Age Adjusted for Tenure, Occupational Prestige, and Education for Black Women and White Men by Age, 1960

Independent Variables (1)	Age 20 to 34 Black Women (2)	Age 20 to 34 White Men (3)	Age 35 to 44 Black Women (4)	Age 35 to 44 White Men (5)
Age adjusted tenure	.113	.210	(.092)	.048
Occupational prestige	.024	.014	.033	.029
Education	.023	.082	.024	.130
Intercept	−.49	−1.38	−.65	−.41
Adjusted coefficient of determination[a]	.34	.20	.45	.15

SOURCE: One-in-one thousand sample of the Census of the Population: 1960.
a. See Rao and Miller (1971: 21).

By using the decomposition techniques we can try to see how much of the $1.29 per hour difference is due to poorer occupational characteristics of black women and how much is due to black women converting their characteristics or skills into wages at a lower rate. The characteristics related to wages that will be examined here include education, occupation, and age adjusted for tenure. In Table 1 we show the mean levels of these variables in the two populations. We see that white men score better on all of these variables than do black women.

We estimate the rates at which each group converts its characteristics into wages by using multiple regression. For each group we estimate income as a linear additive function of education, occupation, and age adjusted for tenure. Equation 5 shows the model and estimated coefficients for white men while equation 6 does the same for black women. The estimated coefficients and the coefficient of determination are also shown in Table 2.

$$Y_w = -1.38 + .210\, X_{1w} + .014\, X_{2w} + .082\, X_{3w} \qquad [5]$$

$$Y_n = -.44 + .113\, X_{1n} + .024\, X_{2n} + .023\, X_{3n} \qquad [6]$$

where Y is mean income per hour, X_1 is mean age adjusted for tenure, X_2 is mean occupational prestige, and X_3 is mean educational achievement.

We can now decompose the $1.29 mean difference in income between black women and white men using equation 1. The decomposition is shown in row one of Table 3. The composition component of this difference is $.86 per hour. The intercept component was calculated to be –$.90 per hour and the rates component was calculated as $.83. The interaction component was calculated to be $.50 per hour.

Our results would have been altered substantially if we had ignored the interaction component. If we had calculated the rates component correctly as $.83 and the intercept component as –$.90 and then assumed that the remaining difference ($1.29 – [$.83 – $.90]) was the composition component, we would have obtained a $1.36 estimate for the composition component rather than the actual $.86 figure. This procedure would be equivalent to estimating the composition component using the sum of the interaction component and the actual composition component. Because the interaction component is positive, this procedure results in an *overestimate* of the actual component.

We shall now examine the income differences for the age group 35 to 44. From Table 1 we see that black women earn, on the average, $1.16 per hour while white men average $2.99 per hour. Table 1 also shows that white men score higher on all

TABLE 3
Decomposition of Hourly Wages of Black Women and White Men with a Model of Hourly Wages Regressed on Age Adjusted for Tenure, Occupational Prestige, and Education by Age Group, 1960

	Components			
Age Group	Composition	Interaction	Regression Coefficients	Intercepts
24 to 34	$ .86	$.50	$.83	–$.90
35 to 44	$1.20	–$.21	$.60	$.24

NOTE: See text for definitions.
SOURCE: One-in-one thousand sample of the Census of the Population: 1960.

of the important characteristics related to income. Utilizing the same regression model as for the younger age group, we estimated the rate at which the two groups convert their characteristics into income. The estimated coefficients are shown in Table 2.

We can now decompose the $1.83 difference in mean wages into its components. That decomposition is shown in row two of Table 3. There we observe a composition component of $1.20 per hour, a rates component of $.60, an intercept component of $.24, and an interaction component of –$.21. If we had calculated only the rates and intercept components for this age group and then assumed that the remaining difference was the composition component, we would have *underestimated* the actual composition component. This occurs because the interaction term is negative in this instance and would in effect be added to the actual composition component. Thus, ignoring the interaction component would result in an estimate of $1.00 for the composition component rather than the actual component $1.20.

As we have seen, failure to include the interaction component explicitly would influence the results for both age groups. Furthermore, failure to include the interaction component affects the interpretation in a different direction at ages 35 to 44 than it did at ages 20 to 34. This example illustrates the effect of interpreting the gap remaining after adjustment for the rates and intercept components. However, inferences based on the remainder after obtaining the composition and intercept components can be equally misleading.

In many situations the decomposition used should handle the interaction component explicitly. For other questions it may be appropriate not to directly consider the interaction. However, even in these cases, it may be a useful strategy to utilize equation 1 in the initial decomposition and then combine the interaction component with one of the other components as appropriate.

NOTES

1. To obtain equation 2, $0 = \Sigma X_{in}(b_{iw} - b_{in}) - \Sigma X_{in}(b_{iw} - b_{in})$ is added to the right side of equation 1. To obtain equation 3, $0 = \Sigma b_{in}(X_{iw} - X_{in}) - \Sigma b_{in}(X_{iw} - X_{in})$ is added to the right side of equation 1. The quantities are regrouped to form the respective equations.

2. In order to illustrate the problem, a sample was selected from the one-in-one thousand sample of the *Census of the Population: 1960.* The sample consisted of black women and white men in the civilian nonagricultural labor force who worked at least 35 hours in the census week and 50 weeks in 1959. The total wage and salary earnings in 1959 of each worker was divided by the estimated hours worked in 1959 by the worker (see Fuchs, 1968). The detailed census occupation of each worker was coded for its 1964-1965 NORC occupational prestige score (Siegel, 1971). Educational attainment was coded in years of school completed. The age of white men was multiplied by .381 and the age of black women by .191. This adjustment to age creates an estimate of the amount of continuous employment or job tenure a worker had for each year of age. The values of .381 and .191 were the partial regression coefficients from the equation regressing job tenure on age for white men and for black women (O'Boyle, 1969). The mean characteristics of white men were substantially higher than those of black women (see Table 1). Hourly wages were then regressed on age adjusted for tenure, occupational prestige, and education (see Table 2). The equations were calculated separately for white men and for black women by age cohort.

REFERENCES

ALTHAUSER, R. P. and M. WIGLER (1972) "Standardization and component analysis." Soc. Methods and Research 1, 1 (August): 97-135.

COHEN, M. S. (1971) "Sex differences in compensation." J. of Human Resources 6, 4 (Fall): 434-447.

DUNCAN, O. D. (1969) "Inheritance of poverty or poverty or inheritance of race?" pp. 85-110 in D. P. Moynihan (ed.) On Understanding Poverty. New York: Basic Books.

--- (1967) "Discrimination against negroes." Annals of the American Academy 371, 1 (May): 85-103.

FUCHS, V. (1968) The Service Economy. New York: Columbia Univ. Press.

KITAGAWA, E. M. (1955) "Components of a difference between two rates." J. of the Amer. Stat. Assn. 30 (December): 1168-1194.

LEVITAN, T., R. P. QUINN, and G. L. STAINES (1971) "Sex discrimination against American working women." Amer. Behavioral Sci. 15, 2 (November/December): 237-255.

MALKIEL, B. G. and J. A. MALKIEL (1973) "Male-female pay differentials in professional employment." Amer. Economics Rev. 63 (September): 693-705.

MINER, J. (1960) "Consumer personal debt: an intertemporal cross-section analysis," pp. 400-461 in I. Friend and R. Jones (eds.) Proceedings of a Conference on Consumption and Saving. Vol. II. Univ. of Pennsylvania Press.

MORGAN, J. N. (1968) "Analysis and interpretation of cross-national surveys," in B. Karger (ed.) Interdisciplinary Topics in Gerontology. Vol. II. New York: Basic Books.

--- and J. B. LANSING (1971) Economic Survey Methods. Ann Arbor: Institute of Social Research.

O'BOYLE, E. J. (1969) "Job tenure: how it relates to race and age." Monthly Labor Rev. 92, 9 (September): 16-23.

RAO, P. M. and R. L. MILLER (1971) Applied Econometrics. Belmont, Calif.: Wadsworth.

SCHILLER, B. R. (1970) "Stratified opportunities: the essence of the 'vicious circle'." Amer. J. of Sociology 76, 3 (November): 426-442.

SIEGEL, P. M. (1971) "Prestige in the American occupational structure." Ph.D. dissertation. University of Chicago.

SUTER, L. E. and H. P. MILLER (1973) "Income differences between men and career women." Amer. J. of Sociology 78, 4 (January): 962-974.

WINSBOROUGH, H. H. and P. DICKINSON (1969) "Components of negro-white income difference." University of Wisconsin Center for Demography and Ecology, Madison. Ecology, Madison. (mimeo)