

This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Higher Education and Earnings: College as an Investment and Screening Device

Volume Author/Editor: Paul J. Taubman, Terence Wales

Volume Publisher: NBER

Volume ISBN: 0-07-010121-3

Volume URL: <http://www.nber.org/books/taub74-1>

Publication Date: 1974

Chapter Title: Appendix D: Grouping

Chapter Author: Paul J. Taubman, Terence Wales

Chapter URL: <http://www.nber.org/chapters/c3659>

Chapter pages in book: (p. 195 - 196)

## Appendix D: Grouping

The information contained in the cross tabulations in Appendix B can be used to estimate the effects of education and mental ability on income in the following manner. Suppose the equation to be estimated is

$$Y = XB + u \quad (D-1)$$

where  $Y$  = a vector of wage and salary income

$X$  = a matrix of independent variables, one of which is mental ability  $A$

$u$  = a random-disturbance term

Assume that observations are on individuals and that the data are ordered by the value of  $A$ , divided into groups such as tenths, and averaged. These averaged data may then be expressed in terms of a grouping matrix  $G$  applied to Eq. (D-1) after ordering by  $A$ .<sup>1</sup> That is,

$$GY = GXB + Gu \quad (D-2)$$

The elements in  $GY$ ,  $GX$ , and  $Gu$  are the mean values in each of the cells. As is well known, efficient estimation of such grouped data requires weighted regressions, with weights equal to the

<sup>1</sup> For example, with three groups involving three observations in the first group, and two observations in the others,

$$G = \begin{bmatrix} 1/n_1 & 1/n_1 & 1/n_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/n_2 & 1/n_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/n_3 & 1/n_3 \end{bmatrix}$$

square root of the number of observations in each group.<sup>2</sup> We can therefore obtain unbiased estimates of the coefficients in Eq. (D-1) by collapsing the information on the distribution of income (and the other variables) in each ability class into average values for each ability class.

As noted earlier, the data are available cross-classified by ability *and* occupation. The analysis of grouping just discussed also applies to this more detailed breakdown. We have disaggregated our data into occupation-IQ groups for several reasons. First, this disaggregation should increase the efficiency of our estimators by providing more dispersion in the independent variables. Second, this method substantially increases the number of observations to be used in the regressions—from 10 to about 70 for Minnesota males. Because for some purposes we wish to use five education variables and five or more ability variables, the larger sample size is very valuable. Finally, and perhaps most importantly, we wish to examine the effects of education within occupations. By using the data cross-classified by occupation and ability, we can explore some of the questions raised in earlier chapters. Note, however, that if occupation is correlated with the other independent variables as well as income, then dummy variables for occupations must be included as independent variables in order to obtain unbiased estimates of the parameters.

<sup>2</sup>If the grouping variable is correlated with the other  $X$ 's, then unbiased estimates of the coefficients require the grouping variable to be one of the independent variables.