ESTIMATING A MICRO WAGE EQUATION: PITFALLS AND SOME PROVISIONAL ESTIMATES

Alan S. Blinder

Econometric Research Program
Research Memorandum No. 131
November 1971

PRINCETON UNIVERSITY
Econometric Research Program
207 Dickinson Hall
Princeton, New Jersey

ESTIMATING A MICRO WAGE EQUATION: PITFALLS AND SOME PROVISIONAL ESTIMATES

Alan S. Blinder*

I. Introduction

Why is the distribution of income so unequal? This is currently, or at least should be, an active topic of economic debate. To answer it, we need quantitative estimates of the many factors that have obvious qualitative impacts on the income distribution. For example, it is clear that racial and sexual discrimination are disequalizing factors; but it is not clear how important they are quantitatively.

Total income of an individual is the product of the wage rate times labor effort, plus the product of net worth times the rate of return earned. Previous research (Blinder 1971, especially Chapter 4) has convinced me that, of these four factors, the dispersion in wage rates is far and away the chief contributor to over-all inequality. It becomes, then, a matter of great importance for public policy to develop some quantitative understanding of the many factors leading to unequal wage rates. This paper is an effort to supply such estimates. Oddly enough, despite its obvious interest, there have been relatively few attempts to estimate such a micro-economic wage equation. The reason for this, apparently, has been the absence of accurate wage data in large cross-sectional surveys. The Bureau of the Census, which has been the data source

*The author owes a considerable debt of gratitude to Robert E. Hall of M.I.T. who furnished the data, assisted greatly with computational problems, and commented astutely on an earlier draft. I would also like to thank Nicholas A. Barr of London School of Economics for counseling me in the use of the Cross Section Processor package of computer programs.

for many studies, does not gather wage data, and a wage figure can be obtained only by dividing money income from labor by a crude estimate of hours of work, obtained by multiplying hours of work in the previous week by number of weeks worked in the previous year. Such a procedure gives rise to errors which may well by systematic. For example, it will under-estimate the wage for any person who worked full-time during the week prior to the survey, but worked part-time for some part of the preceeding year.

Faced with this difficulty, most investigators have opted for some income or earnings measure as the left-hand variable.[1] Using earnings rather than wages as the left-hand variable, of course, confounds two sets of causal factors. The variables accounting for variation in wage rates may be quite different from the variables influencing hours of work. In particular, since tastes -- which generally go unmeasured -- have an important impact on hours, we may expect that regressions to explain earnings will leave more unexplained variance than regressions to explain wage rates. At the same time, however, as long as wages are normally a work incentive, estimated elasticities should be greater in the earnings equation if the same set of variables is used to explain both variables. This is easy to explain. Let x be any independent variable believed to effect the wage rate but to have no

---

[1] Among the many studies which have used annual income or annual labor income as the left-hand variable are: Adams (1958), Hill (1959), Hanoch (1967), Blau and Duncan (1967), Duncan (1968), Weiss (1970) and Bowles (1970). This list is far from complete.

direct effect on hours of work.[2] Let w be the wage and h be hours of work so that $E = wh$ is labor income. Finally, suppose, as is generally the case, that the equations are estimated in logarithmic form. In the equation for log w, the coefficient of x will be:

$$\frac{\partial \log(w)}{\partial x} = \frac{1}{w} \frac{\partial w}{\partial x}$$

while the corresponding coefficient in the earnings regression will be:

$$\frac{\partial \log(E)}{\partial x} = \frac{1}{w} \frac{\partial w}{\partial x} + \left(\frac{w}{h} \frac{\partial h}{\partial w} \right)\left(\frac{1}{w} \frac{\partial w}{\partial x} \right)$$

which is larger as long as $\partial h / \partial w > 0$.

So regression results indicating the positive effect of education on earnings do not give a reliable guide to the effect of education on wages unless labor supply is inelastic.

In specifying a regression model for the present study, I have benefited from the work of many previous investigators.[3] My

---

[2] For example, years of eduation,

[3] In addition to the references cited in footnote 1, I should mention Hall (1970), Morgan, David, Cohen and Brazer (1962), Morgan and David (1963), Lansing and Morgan (1967), Fuchs (1967), David and Schroeder (1970), Taylor (1968), Griliches (1970), Griliches and Mason (1970) and Gintis (1971).

reading of the literature suggests to me several criteria for a methodologically sound study of the dispersion in wage rates (or earnings). Among the most important are:

(1) For the reasons just given, it should use the wage rate, not earnings, as the dependent variable. Alternatively, if labor income is of interest, wages and hours can be considered separately, as in Hall (1970) and Morgan et. al. (1962).

(2) It should specify precisely the structural model of wage determination the author has in mind. This presumably would include structural equations for the level of educational attainment, and for occupational choice. The system should then be solved for the reduced form, which can always be estimated by ordinary least squares. Estimation of the structure presents some knotty problems which are fully discussed in Section IV.

(3) It should include, along with the usual "economic variables" certain "sociological variables" representing the individual's family background. The work of Blau and Duncan (1967), Duncan (1968) and Bowles (1970) suggests that such variables may be of considerable importance.

(4) Variables should be included only if economic <u>theory</u>, or economic <u>reality</u>, can rationalize their inclusion. Data mining, which is very tempting when dealing with rich cross-sections, is

to be studiously avoided. I have tried to adhere to this principle very strictly, and will return to it in discussing the functional specification I have selected.

(5) Among the most important questions such a study should address itself to are:

(a) What is the effect of education on wages?

(b) How much does union membership effect the wage?

(c) Is racial and/or sexual discrimination in wage rates important, and if so where does it enter? Is it the case that blacks have less access to education but, given education, compete on an equal footing with whites in the labor market? Or is there additional direct discrimination in wage rates?

(d) Do the circumstances of one's birth have any lasting influence on economic success and, especially, on the wage rate earned?

In Section II, I present and explain the regression model adopted for this study, and in Section III I discuss the source of the data and the particular empirical variables available. My work differs from most, but not all, previous efforts by taking a simultaneous-equations approach to the estimation problem. This raises entirely new, and very formidable, problems of its own, not all of which have been coped with here. The belief that "learning by doing" effects are very important to the progress of our science prompts me to elaborate these difficulties in some detail in Section IV. Section V presents the regression results, and some conclusions are offered in the final section.

II. A Model of Wages, Education and Occupation

In cross-section studies of the present type there is an inevitable, but uncomfortable, interaction between the model and the data. It is inevitable since cross-sectional surveys are relatively scarce, and one can only test those hypotheses which the limited data permit. It is unfortunate since there is a tendency to let the data dictate the model, i.e., to include those variables measured "accurately" by the survey. In this section, before discussing the data, I formulate a simple model of the wage rate, educational attainment and occupational choice of the individual. I take these subjects up in this order because it was in this order that they were in fact considered. I was fortunate to have access to three national surveys (the Census 1/1000 sample, the Survey of Economic Opportunity and the Survey Research Center's "Panel Study of Income Dynamics" data, and was therefore privileged to formulate my model first and choose the data accordingly. Many other writers have been denied this luxury. Of course, the data was not ideal, so certain modifications in the model were necessary after the data were studied. These will be discussed in the next section.

The model I have in mind is intuitive rather than rigorous. Each individual, at some point in his life-cycle, jointly decides how far he should pursue his education and to what occupational strata he aspires. Simultaneity is quite important here. Given these decisions, and certain other characteristics to be specified shortly, his wage rate is determined.

Let us be more specific. Let w be the wage rate, E be some measure of educational attainment (perhaps a vector), O be some measure of occupation (perhaps a vector) and J be some measure of on-the-job or other vocational training. Let T be a measure (again, perhaps a vector) of tenure on the job, B be a vector of family background variables, U be a dummy variable for union membership and V be a dummy variable for veteran status. Finally, let Z be a vector of characteristics which would have no theoretical rationalein a perfectly functioning labor market, but which obviously effect economic achievement in our imperfectly functioning American economy. For example, I might include sex, race, region of residence, etc. I then propose the following structural model:

$$(1) \quad w = f(E, O, J, T, U, V, Z) + u_1$$

$$(2) \quad E = g(O, V, B, Z) + u_2$$

$$(3) \quad J = h(O, V, U, B, Z) + u_3$$

$$(4) \quad O = k(E, J, V, B, Z) + u_4$$

$$(5) \quad T = l(w, E, O, J, U, V, Z) + u_5$$

$$(6) \quad U = m(E, O, V, B, Z) + u_6$$

$$(7) \quad V = n(E, J, O, B, Z) + u_7$$

Endogenous variables in this model are: w, E, J, O, U, V, and T. Exogenous variables are B and Z.

Equation (1) expresses the notion that the wage rate earned depends directly on the education of the worker, the occupation

followed, the amount (if any) of on-the-job or vocational training the length of time on the job, whether the individual is a union member or not, whether he is a veteran or not, and certain exogenous variables. None of these call for explanation, except the last two. Veteran status may effect the wage for at least two reasons. First, especially at the lower skill levels, the armed services may provide a valuable form of vocational training. Many poorly educated persons come out of the Army skilled in some civilian trade. Secondly, serving in the armed forces results in a loss of experience in the civilian labor force. However, when tenure on the job is already included in the regression, we may surmise that most of this effect is already captured by a lower value of T. Hence, the a priori expectation is that V should have the same sign (positive) and approximately the same magnitude as J. The components of Z which should be included in the wage equation are, inevitably, partially dictated by the data source. I have decided to include:

(a) age -- to get some estimate of the typical age-wage profile, which is the joint result of (1) acquiring experience and (2) depreciation of the human resource over time;

(b) race -- to capture the obvious discrimination against blacks in this country;

(c) sex -- to account for the analogous discrimination against women;

(d) health -- for rather obvious reasons;

(e) residence and local labor market conditions -- to deal with the obvious geographical imperfections in the U.S. labor force. Some comments are in order here. Following my self-imposed dictate to throw out those variables that lack a persuasive reason for inclusion, I have included fewer geographical variables than many previous writers. Other studies have often included the size of the community of residence, or even the precise city.[4] I have included only the broad region of the United States (Northeast, North Central, South or West) and measures of the wage and unemployment level of the local labor market. The reason is that it is these characteristics -- local wage-price levels and labor market tightness -- that should have an impact on the wage rate actually received, not the latitude and longitude on the map.[5]

Finally, observe that I have imposed the identifying restriction that family background variables do not directly influence the wage. They may, as shown in equations (2)-(4),

---

[4] Hall (1970).

[5] Of course, to the extent that we do not have measures of the many relevant features of the local labor markets (e.g., the skill mix in each city), the precise locality may remain a "significant" variable. C.f. Fuchs (1967). However, economic theory will not have explained any of this. I have benefitted from a discussion with Robert E. Hall on this point.

influence the educational and occupational attainment of the individual. But, given his job and his education, the model insists that they have no direct impact on wages. Of course, other writers are free to disagree with this restriction.

Consider next equations (2) and (3) which determine the training of the individual. This is assumed to depend on his occupational choice, whether or not he is a veteran,[6] his family background and certain exogenous variables. The exact list of exogenous variables will be discussed in the next section. Also, for vocational education only, union membership is included as possibly relevant.

Equation (4) expresses the occupation of the individual as a function of his schooling, other training, veteran status (a form of training), family background and other exogenous variables.

Equation (5) states that the length of time on the job depends on the wage rate earned, the amount of education and training, the occupation, union membership, veteran status and certain exogenous variables; but not on family background directly. Here the functional dependence is likely to be quite weak. Put differently, T is likely to be "almost" exogenous. So I have worked with two distinct models, one with T endogenous [equations (1)-(7) above] and one with T exogenous [equations (1)-(4), (6) and (7)].

---

[6]Because being drafted may interfere with education; serving in the Army may be a substitute for vocational education: and the GI bill of rights may make college attendance easier.

Equation (6) takes the somewhat unusual position of endogenizing union membership. The hypothesis here is that one's occupational choice is the principal determinant of whether one joins a union. An auto worker joins the union; a farm worker does not. Formal schooling is also included, though I would not imagine the dependence to be strong. Perhaps education at low levels increases one's eagerness to join unions (to increase one's pay), while higher education diminishes the desire to join. Family background variables are included for similar reasons, and veteran status is included on the hypothesis (testable with the data) that the discipline of the Army makes a man more willing to submit to the enforced uniformity of union membership.[7]

Finally, equation (7) takes the unorthodox step of recognizing what everyone suspects to be true of the U.S. Selective Service System; namely, that it does not operate fairly! Casual empiricism suggests that persons with more education can more easily escape the draft and are less prone to enlist. Occupational

---

[7] It has been suggested to me that endogenizing union membership would be an incorrect procedure if:

(i) unions strictly limit their size to force wages up;

(ii) new entrants are admitted to unions only when a vacancy arises through attrition;

(iii) who is admitted to fill this vacancy is purely a haphazard decision.

Even if (i) and (ii) are true, I question (iii) as an empirical proposition. But, even granting (iii), endogenizing union membership is still correct if we recognize that individuals -- at their own discretion -- may or may not place their names on the queue awaiting admittance to the union. Of course, we would expect equation (6) to have a very large standard error.

deferments for "indispensable" individuals and for all farmers have existed for years (although only the farm deferment is still on the books), so occupation is included. It would appear that men from poor families are both more draftable and more prone to enlist than men from middle- and upper-class backgrounds. Finally, such variables as race and region are included to test how even-handedly the SSS deals with the races and the geographical areas of the country.

The structural model just presented can be solved for its reduced form:

$$\begin{aligned} W &= F(B, I) + v_1 \\ &\vdots \\ V &= N(B, I) + v_7 \end{aligned} \quad .$$

My original intention was to estimate all the equations of both the structure and the reduced form. This would enable us to learn the differences between, for example, the <u>total</u> effect of racial discrimination on wages (from the reduced form) and the part attributable to the job market directly (from the structure). It would also show how serious the simulataneous-equations bias that afflicted previous studies was likely to be. Unfortunately, data limitations forced me to settle for a less ambitious program.

III. The Data and the Empirical Variables

A. Choice of a Data Bank

The sample upon which the present study is based was taken from the Survey Research Center 's (SRC) "Panel Study of Income Dynamics."[8] This is an ambitious project to trace the economic evolution of a panel of households over a five-year period. The study is now in its fourth year, and has made available survey results for the first three years (1968-1970). In panel studies of this sort, there is always a certain amount of bias introduced by families dropping out of the sample in a non-random manner. To minimize this problem, I have used the data collected from the first wave of interviews. The information was gathered by the SRC during 1968 and pertains mainly to calendar year 1967.

To understand how the sample departs from an ideal random sample of American households, it is worth looking back at the genesis of the project.[9] In 1966, the Office of Economic Opportunity (OEO) conducted a national sample of 30,000 dwellings called the "Survey of Economic Opportunity" (SEO). Since the poverty problem was the major focal point, the sample was decidedly non-representative of the U.S. population. In order to over-represent the

---

[8] The magnetic tape containing the data was kindly furnished to me by Robert Hall.

[9] The following information is taken from the Survey Research Center's recent explanatory monograph (1970). For further details on the collection, processing and accuracy of the data, see that source, or Morgan and Smith (1969).

disadvantaged, about one-half of the SEO sample was drawn from specially-selected non-white poverty areas. [cf. Hall (1970), p. 4]. In 1967, in conjunction with the Bureau of the Census, the same dwellings (not families) were interviewed again, and after this it was decided by the various government agencies that the goals of the Survey would be best served by following a panel of individuals over a longer period of time (five years). The Census Bureau then asked the 1967 respondents to sign a release permitting the OEO to have access to the information for possible re-interview purposes. About 70% of the respondents did so, and this is another potential source of bias.

From among this 70%, the SRC conducted re-interviews with about 1,900 families in 1968 and also added almost 3,000 new families from its Master Sampling frame. Of the 4,802 families successfully interviewed in 1968, 4,460 were re-interviewed in 1969,[10] and form the pool from which I have drawn my own sample. Note that the group of 3,000 is a representative sample of U.S. households, but that the group of 1,900 from the original Census survey is definitely not. The additional bias inherent in the 1968-69 attrition is probably quite small since about 89% were successfully re-interviewed.

I have further pared down this sample in a number of ways. First, since very young family heads may still be acquiring

---

[10] This includes some new families which split off from the old.

education, and thus give very misleading wage, education and occupation data, I have dropped from the sample all households with heads younger than 25. This excluded 9.3% of the sample. Secondly, since there is a serious question as to whether members of minority groups other than the Negro race should be grouped with "Whites" or "Blacks," the few (2.5% of the sample) non-white and non-black families were eliminated. Since the main focus of the study is the distribution of wages, families whose head did not work for money in 1967 were also dropped from the sample. About 75% of the sample had working heads. Finally, the survey asked questions which enabled me to eliminate some of the bothersome statistical outlyers which plague cross-sections of this sort; e.g., the professional, college-educated self-employed businessman who earned only 88 cents per hour due to business losses. The survey asked: "Was your family's income a lot higher or lower than usual this past year (1967)?"; and then asked the reason. If the response to the first question was "Yes" and the response to the second question was: "Head's income from work was higher or lower than usual," the family was dropped from the sample. Since 18.7% of the sample was eliminated in this way, it was hoped that a reduction in the error term might be achieved in this way.

All of these truncations left 2,131 eligible families. The sample characteristics for most variables used in the study are given in the Appendix.

B. The Empirical Variables

I now turn to the data available in the survey. As has already been mentioned, actual hourly earnings are available and were used as w. This is one of the great advantages of this data bank.

Education, unfortunately, was not given as a continuous variable, so my choice of a functional specification was severely circumscribed. Educational attainment was grouped instead into one of seven exhaustive classes:[11]

E_1 = failed to complete elementary school (0-5 years)

E_2 = completed elementary school; did not enter high school (6-8 years)

E_3 = started, but did not complete, high school (9-11 years)

E_4 = completed high school (12 years)

E_5 = some college, but no degree (13-15 years)

E_6 = graduated college (16 years)

E_7 = advanced degree, academic or professional (16+ years).

[11] This is not quite true. Within the lowest educational group the SRC made a distinction between those who had difficulty reading and those that did not. I chose to ignore this distinction by merging the two groups. Similarly, among persons who graduated high school but did not attend college, I ignored the SRC's distinction between those who had "non-academic training" and those who did not. Instead, vocational training enters as a separate dummy variable.

In equations where education appears as a right-hand variable, a set of six dummies was used. For each individual, each dummy has the value 1 if he reached the relevant educational class and 0 otherwise. Thus, for example, a person with an advanced degree had all seven dummies turned "on".[12] The coefficients therefore, show the impact on the left-hand variable of advancing to the next educational class.

When education appears as the left-hand variable (equation (2) of the model), however, the qualitative nature of the data is quite troublesome. I was forced to estimate six separate equations for the probability of reaching educational class j ($j=2,\ldots,7$). The precise functional form is discussed in Section IV below.

Occupation, even conceptually, has no natural continuous unit of measurement. So once again a set of eight dummies was used in the regressions where occupation is an independent variable. In place of equation (4), eight equations for the probability of choosing occupation class j ($j=1,\ldots,8$) were estimated.

The variable that I have referred to above as vocational or on-the-job training (J) indicates those individuals who reported "some training outside the regular shcool system" such as apprenticeships, manpower training programs, etc. Union membership (U) and veteran status (V) are simple yes-or-no variables coded 1 for "yes", and 0 for "no".

---

[12] E_1 was just dropped since everyone reached at least this class.

Tenure on the job is another variable which the SRC unfortunately reported only in discrete categories. The same procedure was again used, with six dummy variables for "length of time working for present employer."

For family background variables, I was somewhat constrained by the availability of data, although this data bank has much more information on this count than any other comparable source. The background variables used were as follows:

(a) B_1 = level of education of the father. This was transformed into seven dummy variables corresponding to the educational classes already explained: $B_{11}, B_{12}, \ldots, B_{17}$.

(b) B_2 = economic status of the parents, entered as two dummy variables: B_{21} = 1 if parents were poor;
0 otherwise

B_{22} = 1 if parents were rich;
0 otherwise.

(c) B_3 = number of siblings. Unfortunately, the SRC gives this numerically only up to seven siblings and then has an upper category of "eight of more." I entered siblings, therefore, as two variables:

B_{31} = number of siblings if less than eight;
0 otherwise

B_{32} = 1 if eight or more siblings;
0 otherwise.

(d) B_4 = place where the head grew up. Substantial information on this matter was available. Out of this I selected only three dummy variables for Southern, farm and urban origins. I presume that those capture most of the important effects.

It remains only to consider the vector of exogenous characteristics, Z. Since different members of Z enter each equation, I shall first list each Z_i, explaining those that need explanation, and then indicate which variable enters each equation.

(a) Z_1 = sex; 1 for males, 0 for females

(b) Z_2 = race; 1 for blacks, 0 for whites

(c) Z_3 = age in years; and $Z_4 = (\text{age})^2$

(d) Z_5 = region of the United States, entered as a set of dummies.

(e) Z_6 = labor market conditions in the county of residence.

Several aspects of county labor market conditions were avilable, from which I chose the unemployment rate and the wage level for unskilled labor as the most interesting, and defined two non-exhaustive sets of dummies:

Z_{61} = low unemployment rate (3.9% or lower)[13]

Z_{62} = high unemployment rate (6% or higher)

Z_{63} = low wage level county (the average wage for unskilled temporary labor in the county is under \$1.50)

Z_{64} = high wage level county (the average wage for unskilled temporary labor in the county is \$2.50 or more).

---

[13]Making this choice in 1971 instead of 1967, I seleted an unfortunate dividing point for a high-employment year like 1967. By this dichotomization, almost 2/3 of my sample lived in low unemployment counties. Thus Z_{61} does not represent what it purports to represent.

(f) Z_7 = geographical mobility. After some experimentation with a number of possible representations of geographical mobility, I settled upon the following non-exhaustive pair of dummies:

Z_{71} = 1 if head has moved out of a community where he once lived in order to take a job somewhere else:

0 otherwise.

Z_{72} = 1 if head has turned down a job rather than move;

0 otherwise.

(g) Z_8 = health. This is represented by three dummies. The first two would presumably have permanent ill-effects on the individual's economic success, while the last might account for a transitory drop in his wage; but should effect nothing else:

Z_{81} = 1 if the person has an obvious disfigurement;

0 otherwise;

Z_{82} = 1 if person reports some work limitation;

0 otherwise;

Z_{83} = 1 if the person suffered a serious illness within the past 12 months;

0 otherwise.

(h) Z_9 = 1 if the amount of work varies seasonally;

0 otherwise.

The functional forms selected for the structural equations give each coefficient an intuitive economic meaning. For the wage equation, log w is assumed to be a linear function of all the independent variables (including age-squared as a variable). This assumes that impacts on the wage rate enter independently and

multiplicatively, and makes the (approximate) interpretation of the coefficients the percentage changes in w caused by unit increments of the independent variables. For all the binary left-hand variables, the linear probability model[14] was chosen. The coefficients, therefore, indicate the addition to the probability that the left-hand variable equals unity attributable to each independent variable. Table 1 below indicates which variables are included and excluded in each equation; the a priori sign expectation, if any, should be obvious in most cases.

C. The Problem of Missing Data

Before discussing estimation techniques and problems, there is one more difficulty which must be faced by any cross-sectional study of this sort. When a survey collects answers to several hundred questions, as this one does, there is certain to be a great deal of missing data. Few families will be able and/or willing to respond to every query, and for some families quite a few answers may be missing.

There are several ways to cope with this problem. If the number of families with any missing data is very small (as may be the case for a survey which asks few questions), and if there is no reason to think that these families are different from the overall sample, then the best solution is probably to throw out all

[14] Cf. Goldberger, p. 248-250.

TABLE 1

Specification of Variables in Structural Model

Equation for:	Included variables (endogenous in CAPITALS):
log w	EDUCATION, VOCATIONAL TRAINING, UNION MEMBERSHIP, VETERAN STATUS, TENURE, sex, race, age, region of residence, labor market conditions, geographic mobility, health, seasonal dummy.
E_2	sex, race, age, health (excluding Z_{83}), siblings, father's education, parents' economic status, place where grew up.
E_3	Same as above, plus E_2 .
$E_4,\ldots,E_7$	VOCATIONAL TRAINING, OCCUPATION, VETERAN STATUS, sex, race, age, health (excluding Z_{83}), siblings, father's education, parents' economic status, place where grew up, region of residence, geographic mobility, labor market conditions, all prior EDUCATION variables (e.g., the equation for E_6 contains E_2, E_3, E_4, E_5).
J	Same as E_7 except delete: VOCATIONAL TRAINING, and add: UNION MEMBERSHIP.
$O_1,\ldots O_8$	EDUCATION, VOCATIONAL TRAINING, VETERAN STATUS, sex, race, age, region of residence, region where grew up, health (excl. Z_{83}), siblings, father's education, parents' economic status, geographical mobility, labor market conditions.
U	EDUCATION, OCCUPATION, VETERAN STATUS, sex, race, age, region of residence, labor market conditions, health.
V	EDUCATION, OCCUPATION, sex, race, region of residence, place where grew up, labor market conditions, health (excl. Z_{83}), siblings, father's education, parents' economic status.
$T_1,\ldots,T_6$	WAGE (perhaps), EDUCATION, VOCATIONAL TRAINING, UNION MEMBERSHIP, VETERAN STATUS, all prior TENURE classes, sex, race, region of residence, labor market conditions, geographical mobility, health (excl. Z_{83}), seasonal dummy.

families with incomplete information. This is, of course, the approach almost always taken in time-series estimation. Unfortunately, it is hard to know if the pattern of non-response is, in fact, random. And with detailed surveys (like the SRC panel), throwing out every family with any missing data is likely to eliminate much of the sample!

A second possible solution is to make an arbitrary assignment for such missing datum. For example, one might substitute the sample mean. However, when the variables are not all orthogonal (and if they are, the regressions won't be much good!), it is well known that we can do better than this.

Loosely speaking, we can exploit whatever correlations the sample reveals by "predicting" the missing data on the basis of the non-missing data. Specifically, let M_i be any variable which is missing for any family, and let N be the set of variables which are never missing. Then, when M_i is missing, we can replace M_i by its predicted value from the linear regression:

$$M_i = NB_i + e_i$$

which we estimate in some way from the sample. To save on computational costs, I estimated all missing data regressions using only that subset of families which had no missing data.[15] More

[15]There were 1,252 such families out of the sample of 2,131.

information can be extracted from the data, however, by first estimating M_1 in this way, and then adding M_1 to the list of non-missing data for estimating M_2, and so on, expanding the sample at each stage.

Lest there be any misunderstanding, the aforementioned technique corresponds neither to reduced form nor structural estimation. It is strictly an atheoretical "data mining" project, which may entail such apparent absurdities as predicting an exogenous variable like race from endogenous variables like the wage rate, etc! To give the reader some idea as to how much information this extracts from the data, Table 2 below lists the R^2's obtained

TABLE 2

Goodness-of-Fit of Some Missing-Data Regressions

Variable being Predicted	R^2
Seven educational dummies	from .144 to .056
Eight occupational dummies	from .256 to .057
Race	.360
Head grew up in South	.641
Head grew up on a farm	.184
Head grew up in a city	.157
Has work limitation	.030
Union Membership	.117
Veteran status	.267

Note: The variables used as predictors were those variables never missing in the sample; namely, the wage, age, family size and composition, sex, region of residence, local labor market conditions, and a number of interactions among these variables.

in the missing data regressions for certain selected variables. A high R^2 indicates that computing the regression is worthwhile, while an R^2 near zero suggests that we might just as well have replaced the missing data with the sample mean. The ability to predict race, veteran status, and especially whether the individual grew up in the South is quite extraordinary. For many of the truly exogenous characteristics, like health, assigning the sample mean would have done almost as well.

IV. Pitfalls in Estimating a Micro Wage Equation

The estimation problem outlined at the end of Section II seems straightforward enough. Specify the structure, and solve for the reduced form. Then estimate the reduced form by ordinary least squares (OLS) and the structural equations by, say, two stage least squares (2SLS). In practice, this turns out to be an extremely demanding program. The present section explains the difficulties of carrying out such an estimation project -- some of which inhere in my particular data source, but most of which inhere in the problem itself -- and explains the estimation procedure I have had to settle upon.

A. The Identification Problem I

The formal problem of identification in econometrics is well-known. Loosely speaking, if we are to identify an equation of a model of m equations by the omitted-variables method (the

standard technique), we must expurge at least m-1 variables on theoretical grounds. Now imagine yourself in possession of survey data on the socio-economic status of a sample of individuals. Your goal is to estimate a micro-economic wage equation as part of a complete structural system, such as equations (1) - (7) above. What variables can you exclude on a priori grounds? Not very many, I am afraid. I have excluded only the various dimensions of family background, but some economists would probably bridle at even this suggestion. Remember, these excluded variables must be included in one of the other equations of the model. That is, the variables must effect educational achievement, or occupational choice, or some other endogenous variable, but have no direct structural impact on the wage rate. Try thinking of some such variables. Even in principle it is hard to imagine what they might be. Now try perusing the questions asked in a typical cross-section survey to find m-1 such variables. I submit that this task will almost inevitably be impossible unless the structural model is so large that it includes equations for groups of variables quite disconnected with the wage, education and occupation. For example, if there were a block of equations for savings and portfolio behavior, consumption patterns, and labor supply, it might be possible to exclude enough variables to identify the wage equation. Such an ambitious micro estimation project has never been carried out.

B. The Identification Problem II

I have not, so far, discussed any interactions among the list of independent variables affecting the wage. This is because I am very reluctant to play the interactions game. It is not that I do not believe that such interactions exist. It is quite likely that years of education have a different effect on the wage of a 30-year-old white man than on the wage of a 65-year-old black woman. The problem is that once you start adding interaction terms it is difficult to know where to stop. All the combinations of thirty variables taken two at a time gives some 435 possible interaction effects, and a clever economist can probably provide a reasonable argument for each of them! Even with large cross-sections, this can be expensive in terms of degrees of freedom. And the danger of degenerating into data-mining is ever present. For these reasons I have been very hesitant to rely on too many interaction effects. In the estimates reported in the next section, I have considered two models. The first allows for no interactions at all. The second allows for full interactions between race and sex and all other variables by estimating separate regressions for each race-sex group.

However, in order to see if interactions might help me overcome the identification problem, I did experiment with some interactions among family background variables (the only set of variables excluded from the wage equation). It is, after all, quite easy to

see, for example, why having a poorly educated father would be much less of a handicap in a rich family than in a poor family. It doesn't appear to have handicapped Lamar Hunt![16]

Allowing a few such interactions easily gets us over the formal criteria for identifiability, i.e., satisfies the rank and order conditions. However, this is merely a formal solution. The essence of the identification problem is that exclusions allow us to distinguish one equation of the model from all the others. That is, if equation 1 excludes enough variables, it cannot be confused with a linear combination of the other m-1 equations of the model. Looking at identification this way makes it clear that merely large numbers of excluded variables will not suffice unless they are important exclusions. That is, excluding a set of relatively trivial variables will not give us the power to distinguish equation 1 from a linear combination of the other equations. Computationally, the X'X matrix of the second stage of 2SLS will not be exactly singular (as it is in the under-identified case), but instead will be nearly singular. This will be reflected in large standard errors for the estimated coefficients. This, in fact, is what I found when I added interactions (which are almost always of minor importance) to the model. Though the rank and order conditions were met, the resulting 2SLS estimates were so unreliable as to be useless.

---

[16]The next paragraph owes much to a discussion with Robert E. Hall.

What then, is the harried econometrician to do? One thing that can certainly be done is to estimate the reduced form instead. Such estimation is rigorously correct, and not without interest. After all, if we are interested in knowing why wages differ in terms of "ultimate causes" (but not the mechanisms through which these causes influence wages), the reduced form gives the information we want. For example, it will tell us how much being black costs. But will not tell us whether this is due to inadequate education or outright discrimination in the labor market.

Still, those of us not enamored of the Chicago methodology,[17] would like to get a look inside the "black box," i.e., would like to have some structural information. In order to provide some provisional structural estimates, I have had to look elsewhere for identifying restrictions. Before proceeding farther, let me make it clear that this is a decidedly second-best estimation procedure. Full-fledged 2SLS estimates would be much preferred were they possible. However, the estimation procedure I am about to outline has been, at least tacitly, adopted by every previous writer in this field. And some capable econometricians have been included in this group. Thus the procedure, unsatisfactory as it may be, does not lack for precedents.

The identifying restriction I have appealed to in order to obtain structural estimates is the absence of correlation among the structural disturbances, u_i. Look back at the model outlined

[17] Friedman, (1953), Part I.

in equations (1) - (7) above. Notice that the wage does not appear as a right-hand variable in any equation other than (5). This means that, *if* u_2, u_3 *and* u_4 *are all uncorrelated with* u_1, the classical regression assumptions that $E(Eu_1)=E(Ou_1)=E(Ju_1)=E(Uu_1)=E(Vu_1)=0$ all hold so that OLS is the best linear unbiased estimator. In a word, equations (1) - (7), along with the restrictions on the disturbances just mentioned, constitute a *block-recursive system*. Of course, it is not a fully recursive system. The simultaneity among the higher educational classes $(E_4,\ldots,E_7)$ and the occupational classes, for example, still remains. But some structural information, in particular the coefficients of the wage equation, can be obtained without using simultaneous equations estimators.

It is easy to prove that if the first endogenous variable (w in my system) never appears as a right-hand variable in any equation of the model (and, if the weak dependence of T on w is ignored, my model satisfies this requirement), then OLS is the best linear unbiased estimator of the first structural equation *provided that the structural disturbances* (u_i) *are mutually uncorrelated*.[18]

---

[18] The "proof" is simply to observe that, in the block recursive specification just outlined, all remaining endogenous variables will be uncorrelated with the first structural disturbance. Thus the Gauss-Markov theorem will apply, assuming none of its other requirements are violated.

Is the assumption that $E(u_i u_j) = 0$ for $i \neq j$ a "reasonable" one? This is a difficult question to answer. To the extent that the disturbances are true "random shocks," the assumption probably holds. After all, the decision as to whether to advance from the eighth grade to the ninth grade is made long before an occupational choice is settled upon, and certainly before the current wage is determined.[19] This temporal argument however, breaks down if the u_i really represent unobserved personality traits that make one man more prone to (a) pursue education, (b) rise on the occupational ladder; and (c) earn a higher wage, than another identically-situated man. Since the disturbance term is only a name we give to our ignorance, there is no way we can resolve this issue. We can only appeal to the general principle of econometrics that small departures from theoretical assumptions [in this case that $E(u_i u_j) = 0$] generally cause only small biases and/or inefficiencies in estimation.

C. Two-Stage Least Squares with Binary Data

It seems clear that simultaneous equations estimates of a microeconomic wage equation -- even with ideal data -- almost inevitably founder on the identification problem. Worse, the qualitative nature of the data forces us into the linear probability

---

[19] I have benefited from a discussion with Nicholas A. Barr on this point.

model,[20] which entails unique problems of its own.

The present study is the only attempt known to me to combine simultaneous equations estimates with the linear probability model. As Murphy's Law[21] suggests, the problems that arise when the two estimation techniques are combined exceeds the sum of the problems inherent in each separately. The additional problem is that the first stage regressions to predict, for example, the probabilities of being in each occupation class, are bound to fit the data very poorly. Predicting a discrete variable with, say, eight possible outcomes is much more difficult than predicting a single continuous variable.[22] If the first-stage (reduced form) equations fit poorly, when we replace, e.g., actual occupation with predicted occupation we introduce a great deal of "noise" into the system. Worse yet, a very poorly-fitting regression is little more than just predicting a constant. To the extent that one or more of our predicted variables is "almost" a constant, we run into the one econometric problem that cross-sectional studies usually manage to avoid -- multi-collinearity among the right-hand variables! This is reflected in near singularity of the $X'X$ matrix and therefore in very unreliable estimates. For this reason,

---

[20] There are alternative models, e.g., the logit and probit models. But these require expensive non-linear estimation techniques, Cf. McFadden (1968) on the logit model, and Goldberger (1964), pp.248-255.

[21] Freely translated: "If anything can go wrong, it will!"

[22] This is illustrated in this study by the much higher R^2 in the reduced form for the wage equation than for any of the binary variables.

where the assumption of uncorrelated disturbances is insufficient to rid us of simultaneity problems (e.g., in the equations for higher educational achievement and occupational choice), it is impossible to estimate the structural equations. Perusal of Table 1 will show the reader that OLS is an unbiased estimator only for the structural equations for log w, E_2, E_3 and U. Fortunately, it is the wage that is of primary interest.

V. Reduced Form and Structural Estimates of the Micro Wage Equation

This section presents the regression results for the wage equation, comparing the reduced form estimates with the structural estimates. The estimating techniques have been extensively discussed, so no further comments on this matter should be necessary.

Rather than present exceedingly long tables with the regression coefficients, I have chosen to present the data in smaller blocks indicating the structural and reduced form impacts of each group of independent variables. These results are prefaced by Table 3 below which gives the summary statistics for each wage equation. Not surprisingly, the structural equations have greater explanatory power than the reduced form: much greater according to R^2, though the standard errors suggest that the differences on this count may not be all that great. About 50-55% of the variance of wages is explained by the stuctural model.

TABLE 3

Summary Statistics for Wage Equations

Sample Group	Sample Size	No. of Ind. Vars.	$\bar{R}^2$	standard error
Reduced Form Regressions				
All	2131	31	.382	.543
White Males	1239	29	.200	.550
Black Males	467	28*	.329	.516
White Females	194	29	.335	.500
Black Females	231	28*	.297	.494
Structural Regressions				
All	2131	40	.550	.464
White Males	1239	38	.498	.437
Black Males	467	38	.467	.467
White Females	194	36**	.449	.465
Black Females	231	34***	.417	.456

*Since no blacks had fathers with an advanced degree, this dummy was dropped from the regressions pertaining to blacks.

**Since no females were farmers, this occupational dummy was dropped from the female regressions. Veteran status was also dropped as a variable for females, although three whites and two blacks were veterans.

***Same as above. Also, no black females were in the "managers, officials and proprietors" job category.

Comparison of the standard errors of the individual race-sex group regressions with the s.e. of the pooled regression suggests that the interactions of race and sex with other variable

are not terribly important. That is, the additional explanatory power of the many interaction terms is not very great. But at the formal statistical level, the null hypothesis of no interactions (against the alternative hypothesis of a full set of interactions) is rejected at the 1% level and beyond.[23]

A. Effects of Race, Sex and Age

Table 4 below contrasts the estimated effects of sex, race and age in both the structural equation and the reduced form for the entire sample. It is interesting to compare the regression coefficients of sex and race with the corresponding unadjusted wage differentials. In this sample, the white-black differential was 44.6% of the white wage (or, 80.5% of the black wage), while the male-female differential was 78.3% of the female wage (43.9% of the male wage). The reduced form coefficients suggest that

---

[23] The test is the F-test on the sum of squared residuals resulting from the pooled regression. In particular, if the number of coefficients being estimated in the pooled regression is K_p, the total number in the four race-sex group regressions is K, the sum of squared residuals in the pooled regressions is SSR_p, and the total sum of squared residuals in the four disaggregated regressions is SSR, then the test statistic is:

$$F(K-K_p,\ 2131 - K) = \frac{SSR_p - SSR}{K - K_p} \Big/ \frac{SSR}{2131 - K} .$$

For the structural wage equations, the computed value of this statistic was 2.47, as compared to the 1% point of the $F(106,1985)$ distribution of 1.29. For the reduced form, the test statistic had the value of 1.90, compared to a 1% point of 1.33.

TABLE 4

Coefficients of Sex, Race and Age of the Logarithm of Wage: Whole Sample

Variable	Reduced Form Coefficient	Structural Coefficient
Sex		
Male	+ .462 (.031)	+ .339 (.031)
Female	-- --	-- --
Race		
Black	- .353 (.032)	- .223 (.028)
White	-- --	-- --
Age	+..0398 (.0075)	+ .0277 (.0068)
$(Age)^2$	- .00046 (.00008)	-.00032 (.00007)

Note: Standard errors are in parentheses.

these unadjusted differentials sharply overestimate the sex differential by not holding "other things equal,"[24] and moderately over-estimate the race differential as well. According to the regressions, being a male tacks 46.2% onto the wage, and being black costs 35.3%.

[24] These "other things" are the exogenous variables, not such things as education and occupation which themselves reflect discrimination.

Comparing these figures with the structural estimates shows how much of this differential is due to unequal attainment of the other endogenous variables (education, occupation, union membership, and veteran status) and how much appears to be strictly a labor-market differential. Of the .462 differential in favor of males, only about .123 is attributable to their superior education, occu--pations, etc., while .339 remains even holding these equal. Similarly, .130 of the total disadvantage of blacks appears to come from inferior education and occupation, while .223 is the result of discrimination and disadvantaged background. It would appear that outright discrimination in the labor market is quite substantial.

The age profiles of the two equations also differ, although both suggest a peak-earnings age between 43 and 44 years old. Figure 1 below graphs these profiles. That the structural wage-age profile rises and falls much slower than the reduced form profile, indicates that a substantial part of the age profile is due to the age pattern of education, occupation, etc. But the significant impact of age that remains even when these things are held constant indicates that the combined influences of experience versus depreciation of skills do create a significantly concave wage-age profile, which goes a long way towards explaining the observed age-income profile.

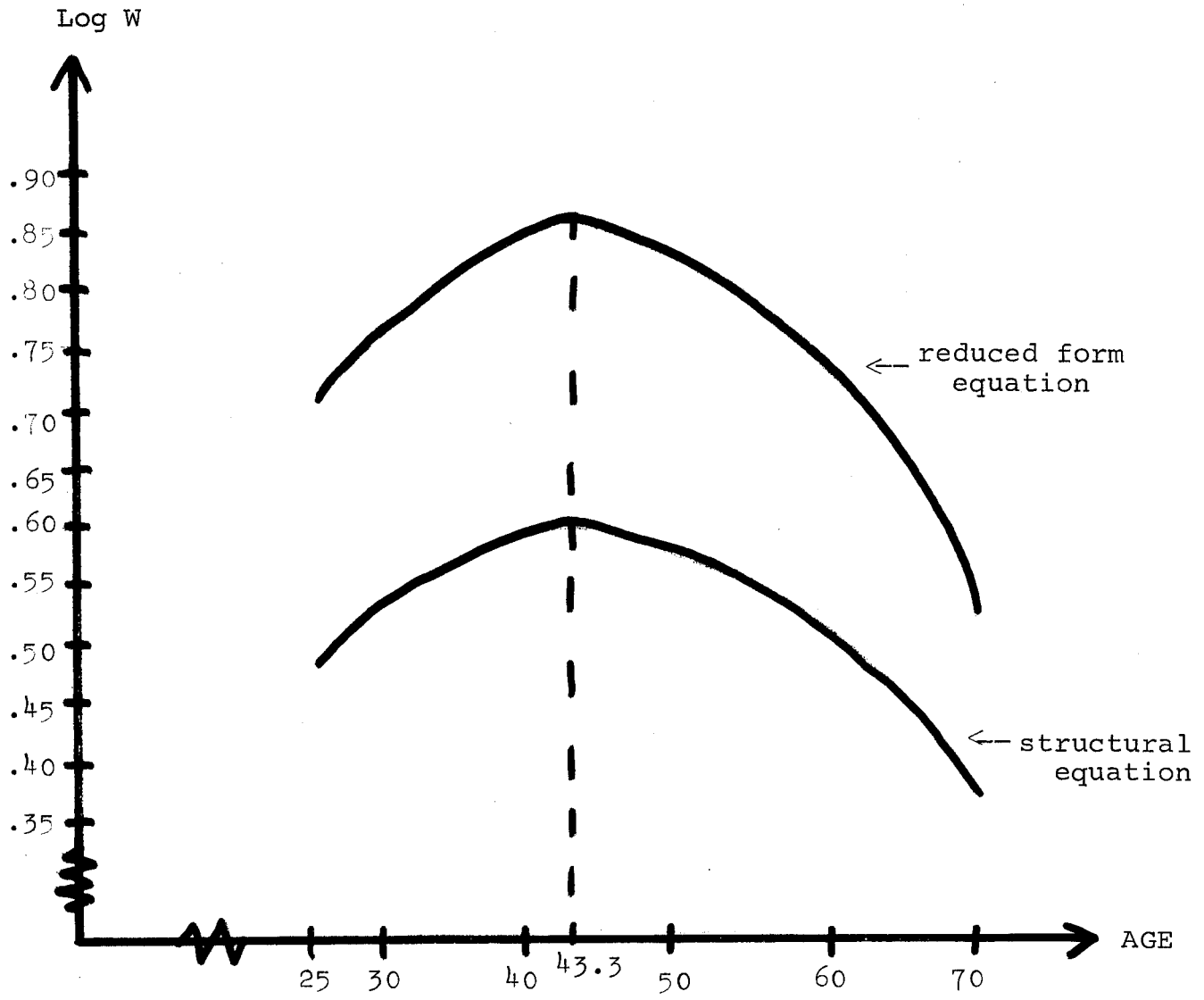

FIGURE 1

Wage-Age Profiles: Whole Sample

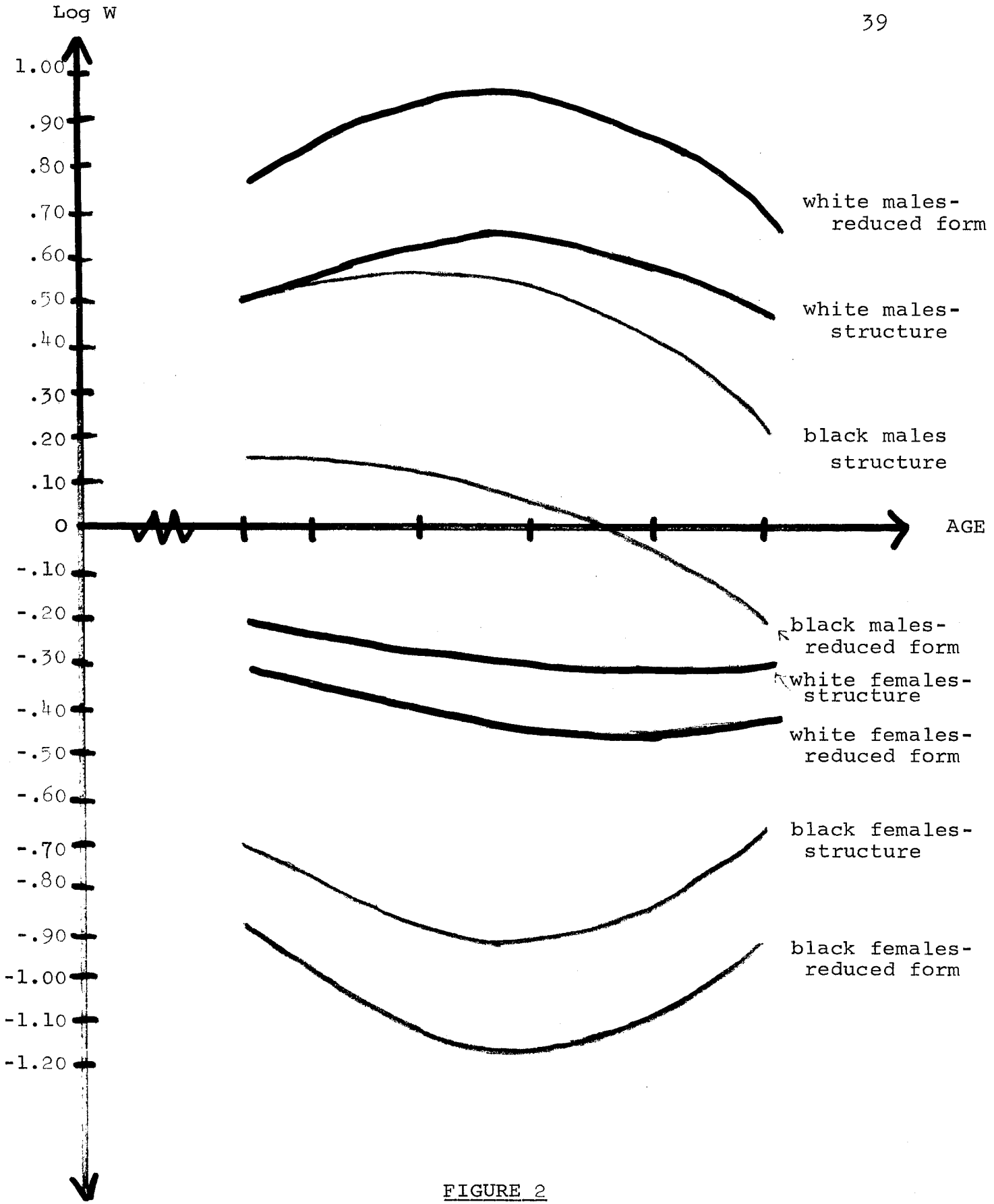

FIGURE 2

Wage-Age Profiles:
Disaggregated Race-Sex Groups

Table 5 below shows that the age-wage profiles are rather different for females, and Figure 2 above portrays them diagrammatically. Instead of the concave profile exhibited by the men, both groups of females show unconventional convex profiles; although

TABLE 5

Coefficients of Age for each Race-Sex Group

Variable	White Males	Black Males	White Females	Black Females
	Structural Coefficients:			
Age	+.0271 (.0078)	+.0296 (.0195)	- .0105 (.0239)	-.0386 (.0270)
$(Age)^2$	-.00029 (.00008)	-.00038 .00022)	+.00009 (.00026)	+.00041 (.00030)
	Reduced Form Coefficients:			
Age	+.0427 (.0093)	+.0110 (.0201)	-.0158 (.0244)	-.0481 (.0275)
$(Age)^2$	-.00047 (.00008)	-.0002 (.0002)	+.00014 (.00027)	+.0005 (.0003)

in the case of white females the profile is close to being gradual monotonic decline with age since the minimum occurs at age 58.3 (in the structure). Although the standard errors for females are rather large, due to the small sample sizes, the repeated sign pattern gives us some confidence that it is true. Note that only for black males, who have the least pronounced age profile, does the structure imply a more intense age pattern than the reduced form.

B. Effects of Place of Residence

Table 6 below shows the effect of place of residence on wage rates. Recall that our a priori expectation was that being in the Northeast, in a low unemployment county and a high wage county should have positive effects, and being in the South, a high unemployment county and a low-wage county should have negative effects.

The regional differences among areas of the United States, even after local labor market conditions are adjusted for, appear to be substantial, except that the West does not appear to differ much from the North Central region (the control group). Living in the Northeast always has a positive effect on the wage (with one insignificant exception), especially for black males. Living in the South, with one exception, has a substantially negative impact on the wage. Surprisingly, the exceptional group is black males! For the other groups, residence in the South appears to cost about 20% - 22% in the reduced form and about 9% - 12% in the structure (after the South's inferior educational and occupational structure is adjusted for). However, after adjusting for education, occupation, etc. (i.e., in the structure), living in the South does not appear to hurt black males; and in the reduced form it lowers their wage only about 10%! Of course, the coefficients have rather large standard errors, so we can't hold this conclusion with any certainty. Black males also are the only group to show substantial gains from living in the West.

TABLE 6

Coefficients of Place-of-Residence Variables

Variable	Whole Sample	White Males	Black Males	White Female	Black Female
		Structural Coefficients:			
Region:					
Northeast	.066	.061	.144	.109	.119
	(.033)	(.038)	(.107)	(.108)	(.139)
South	-.116	-.107	-.013	-.090	-.122
	(.031)	(.038)	(.088)	(.106)	(.101)
West	.009	-.011	.189	-.006	-.057
	(.036)	(.092)	(.117)	(.113)	(.143)
North Central	--	--	--	--	--
County Labor Market:					
Low Unemployment	.002	-.024	.014	-.245	.006
	(.025)	(.031)	(.054)	(.098)	(.085)
High Unemployment	-.077	-.095	-.168	-.255	.017
	(.042)	(.051)	(.097)	(.151)	(.206)
Low Wage Level	-.044	+.010	-.207	+.231	-.248
	(.033)	(.041)	(.068)	(.122)	(.118)
High Wage Level	.108	.073	.315	.203	.190
	(.039)	(.046)	(.109)	(.136)	(.150)
		Reduced Form Coefficients:			
Region:					
Northeast	.079	.080	.130	-.007	.094
	(.039)	(.048)	(.119)	(.117)	(.148)
South	-.207	-.205	-.099	-.235	-.225
	(.044)	(.065)	(.101)	(.156)	(.116)
West	.002	-.024	.198	-.104	-.094
	(.042)	(.053)	(.124)	(.120)	(.154)
North Central	--	--	--	--	--
County Labor Market:					
Low Unemployment	.036	.015	.150	-.249	-.089
	(.029)	(.039)	(.060)	(.102)	(.094)
High Unemployment	-.092	-.087	-.082	-.319	-.188
	(.049)	(.062)	(.104)	(.159)	(.225)
Low Wage Level	-.037	.018	-.226	.369	-.274
	(.038)	(.051)	(.077)	(.132)	(.126)
High Wage Level	.181	.159	.400	.165	-.004
	(.045)	(.057)	(.118)	(.141)	(.159)

As we noted earlier, the definition chosen for "low unemployment counties" was unfortunate, so this variable behaves very erratically. Its coefficient has the wrong sign in four of the ten equations, and is larger than its standard error in only four cases. High unemployment counties show the expected effects. In the reduced form all groups lose, with females showing the greater sensitivity. In the structure (i.e., when education, occupation, etc. are adjusted for), black females appear oddly insensitive (but the s.e. is large) while black males and white females are hurt very seriously. This is in accord with observations of the labor market which show blacks and females to be the last hired and the first fired.

The wage level dummies show a very diverse pattern across race-sex groups.[25] White males appear insensitive on the down side to the wage level for unskilled labor, though they share in the gains on the up side. Black males appear to be the most sensitive to the unskilled wage level. The coefficients in every case (a) have the correct sign; (b) are extremely large (i.e., 21-40%) and (c) are more than three times their standard errors. White females exhibit an anamolous pattern. Wages of black females appear to respond strongly to the unskilled wage level; the approximately zero coefficient in the reduced form is revealed to be an illusion which disappears when we adjust for educational and occupational structure.

---

[25] Recall that the dummy is based on the wage for <u>unskilled temporary</u> labor.

C. Effects of Geographical Mobility, Health and Seasonal Employment

Tables 7 and 8 below present the coefficients for the six remaining exogenous variables which appear both in the reduced form and in the structure: geographical mobility, health and the seasonal dummy.

The coefficients for the mobility variables are quite startling, and lead us to re-think the interpretation of these variables. Recall that each dummy indicates that the individual had once been offered a job in another locality. The first indicates that he accepted the offer; and the second that he turned it down. Now, only a small elite segment of the occupational spectrum generally gets offers of jobs in other localities: doctors, lawyers, professors, scientists, middle- and top-level executives, etc. Thus it is incorrect to take Z_{71} as indicating mobility and Z_{72} as indicating immobility. Instead, both variables serve to pick out the highest rungs on the job ladder (which earn higher wages) and the difference between the coefficients indicates the return to geographical mobility. So the a priori expectation is that both variables have positive coefficients, with Z_{71} getting the larger one if the return to geographical mobility is positive.[26]

---

[26] I have benefited from a discussion with R.E. Hall on this point.

TABLE 7

Coefficients for Geographical Mobility, and Seasonality

Variable	Whole Sample	Males White	Black Males	White Females	Black Females
			Structural Coefficients:		
Geog. Mobility: Moved (Z_{71})	.005 (.026)	.025 (.030)	-.056 (.070)	.032 (.100)	-.066 (.141)
Refused Move (Z_{72})	.034 (.037)	.016 (.043)	.088 (.083)	.230 (.162)	.070 (.220)
Neither	--	--	--	--	--
Job is seasonal	.002 (.029)	-.050 (.038)	.050 (.057)	-.055 (.107)	.140 (.111)
Job is not seasonal	--	--	--	--	--
			Reduced Form Coefficients:		
Geog. Mobility: Moved (Z_{71})	.096 (.030)	.144 (.036)	-.109 (.074)	.034 (.105)	.102 (.145)
Refused Move (Z_{72})	.112 (.043)	.134 (.053)	.059 (.088)	.218 (.167)	.162 (.234)
Neither	--	--	--	--	--
Job is seasonal	-.065 (.033)	-.136 (.046)	.032 (.063)	-.177 (.110)	.112 (.120)
Job is not seasonal	--	--	--	--	--

In fact, the coefficients suggest that the pure return to geographical mobility may be negative. In both the structure and the reduced form, only white males exhibit a positive return to mobility -- and only 1% at that. All other groups, apparently, are better off staying put. What is the reason for this counter-

intuitive result? A recent study of geographical mobility by David and Schroeder (1970) suggests that the measured return to mobility may be negative because movers earn lower salaries in the short-run (within a year after the move), reaping the gains only after a lag. The reason, one surmises, is that they lose seniority in the move. Comparison of the reduced form and structural coefficients -- since the latter adjust for job tenure -- reveals that this certainly is the case. White male movers, for example, gain 14.4% according to the reduced form, but only 2.5% after their loss of job seniority is accounted for. Even this, however, does not fully explain the results. It would appear that accurate estimation of the returns to mobility requires that the data discriminate between habitual movers who "can't hold a job" (this may explain the negative sign for black males), and infrequent movers who relocate only to improve their status.

The seasonal dummy has the wrong sign in half of the equations, suggesting misspecification of the structure and simultaneous equations bias. That is, whether or not ones job is seasonal appears to be endogenous rather than exogenous.[27]

The health handicap variables, which appear quite reasonable in the aggregate sample, exhibit a remarkably diverse pattern across race-sex groups. An obvious disfigurement appears to be

---

[27] If the reader scrutinizes Table 1, he will see that the model makes the theoretical reduced form and structural coefficients of seasonality identical.

TABLE 8

Coefficients of Health Variables

Variables	Whole Sample	White Males	Black Males	White Females	Black Females
		Structural coefficients:			
Obvious Disfigurement	-.092 (.044)	.119 (.053)	-.004 (.103)	-.183 (.131)	+.105 (.201)
Some Work Limitations	-.077 (.032)	-.081 (.081)	-.227 (.082)	-.189 (.135)	+.037 (.095)
Serious Illness Recently	.007 (.037)	-.013 (.050)	-.049 (.074)	-.030 (.140)	+.065 (.108)
Healthy	--	--	--	--	--
		Reduced Form Coefficients:			
Obvious Disfigurement	-.162 (.051)	-.201 (.066)	-.018 (.111)	-.189 (.131)	+.240 (.213)
Some Work Limitations	-.148 (.037)	-.176 (.048)	-.216 (.090	-.231 (.128)	-.011 (.101)
Serious Illness Recently	-.029 (.043)	-.028 (.062)	-.018 (.081)	-.053 (.149)	+.102 (.116)
Healthy	--	--	--	--	--

a greater handicap than "some work limitation". Recent illness appears to have an exceedingly minor effect; presumably its impact is felt on hours of work instead. White males conform to this aggregate pattern, and white females come close to doing so. But the impact of these health disadvantages on blacks appears to be quite different. The data seem to say that "blackness" itself serves as an "obvious disfigurement" so that further disfigurement for blacks is harmless (but note the large standard errors on

these coefficients). Limits on their ability to work, however[28] do exert a negative effect on wages of black males (but not on black females).

D. Effects of Family Background

The preceding exogenous variables appear in both the structural and the reduced form wage equation. The remaining exogenous variables, variables pertaining to the head's family background, were hypothesized to effect the wage only indirectly, i.e., through the reduced form. Table 9 below shows the ultimate effects of the father's education on the wage rate earned by the son or daughter. Racial patterns again are quite different.

TABLE 9

Coefficients of Father's Education on Wage Reduced Form

Level of Father's Education (years)	Whole Sample	White Males	Black Males	White Females	Black Females
0 - 5	--	--	--	--	--
6 - 8	.073 (.036)	.090 (.055)	-.015 (.061)	.166 (.153)	.160 (.084)
9 - 11	.119 (.053)	.163 (.067)	-.055 (.125)	.161 (.160)	.398 (.187)
12	.058 (.064)	.015 (.080)	.109 (.171)	-.012 (.186)	-.174 (.225)
13 - 15	-.087 (.081)	-.143 (.095)	.303 (.392)	.178 (.244)	.159 (.272)
16	.221 (.103)	.228 (.125)	.016 (.410)	.282 (.276)	-.575 (.565)
16+	-.004 (.151)	.082 (.176)	-- --	-.874 (.356)	--- ---

NOTE: Coefficients indicate marginal effect of the father advancing to the next educational level.

[28] Some disfigurements may have no effect on productivity.

Having a father who completed elementary school appears quite helpful to the careers of females, and moderately helpful to white males. However, it seems to accomplish nothing for black males. Entering but not completing high school exhibits a similar pattern, with daughters of blacks gaining enormously (but the s.e. is large). Whether or not the father graduated from high school seems to have little impact on whites, while helping black males and hurting black females (but again s.e.'s are large). The small number of observations on fathers with education above high shool[29] make it hazardous to guess the effects for groups other than white males. For them, having a father enter college seems to be a substantial advantage only if the father earns a degree. Failure to earn the degree actually appears to hurt the son! An advanced degree seems to buy only moderate benefits beyond those achieved by the first degree.

Another potentially relevant background variable is the wealth or income class of the parents. This is crudely captured by our pair of dummies for "poor" and "rich" families. The results are tabulated in Table 10 below. Only white females from poor families show an incorrect (and statistically significant!) sign. White males, like the aggregate sample, indicate only a minor disadvantage from a poor upbringing and a moderate advantage from being brought up in a well-to-do household. Blacks seem to show

[29]Only 11 black males, 19 white females and 6 black females had fathers in the upper three educational classes.

TABLE 10

Coefficients for Family Income Status: Reduced Form

Family of Origin was:	Whole Sample	White Males	Black Males	White Females	Black Females
Poor	-.038 (.029)	-.012 (.036)	-.075 (.075)	.136 (.092)	-.154 (.084)
Middle Class	--	--	--	--	--
Rich	.072 (.042)	.060 (.054)	.033 (.107)	.041 (.111)	.103 (.144)

far greater sensitivity to the economic status of their parents, with the exception that black males do not appear to capitalize on having wealthy parents. It should be noted that the respondents' interpretation of what constitutes poverty vs. affluence may differ sharply among the race-sex groups.

Sociologists generally believe the size of the nuclear family to have an effect on the child's subsequent achievement. However, they do not agree of the direction of this effect. It is clear that having a large number of brothers and sisters tends to rob a child of of both the parents' attention and financial resources. However, it can also help his economic development if older siblings contribute to the family's income and aid in his socialization (Blau and Duncan, pp. 296-7).

The actual coefficients, as presented in Table 11 below, are generally quite small -- suggesting that the coutervailing tendencies tend to cancel one other out. Interestingly for blacks, where

TABLE 11

Coefficients for Number of Siblings: Reduced Form

Variable	Whole Sample	White Males	Black Males	White Females	Black Females
Number of Siblings (if ≤ 7)	.004 (.007)	-.003 (.009)	.013 (.013)	-.012 (.023)	.004 (.017)
Eight or more	-.015 (.040)	-.071 (.059)	.050 (.074)	-.153 (.157)	.093 (.107)

(perhaps) older brothers and sisters may be breadwinners, large family size tends to help. By contrast, for whites, where (perhaps) large family size only serves to spread the parents' attention and financial resources, having many siblings is a (minor) handicap.[30] Note that the coefficients for the sibling variables are "insignificant" because of their small size, not because of large standard errors.

The final family background variable, one of great potential importance, is the place where the head of household grew up. In particular, whether it was in the South, on a farm, or in a city. Table 12 below presents the coefficients for these three variables. Once again, the pattern differs widely across the race-sex groups.

[30]Robert Hall has pointed out an alternative explanation for the different sign patterns by race. For whites, but not for blacks, large family size is an ethnic indicator -- picking out the Catholic (especially, Irish-American and Italian-American) minority. If, for whatever reasons, they tend to have lower wages than Protestants and Jews, the sibling variables might be picking this up.

The insignificant positive coefficient for the South in the whole sample conceals a strongly negative effect on black males, and a smaller negative effect on black females. Men appear to be severely handicapped by farm origins; white females, surprisingly, are aided. (Oh, that farmer's daughter!) An urban upbringing contributes positively to the future success of whites, but does not appear to benefit blacks. In view of the quality of life in big-city ghettos, this is hardly surprising.

TABLE 12

Coefficients for Place Where Head Grew Up: Reduced Form

Head Grew Up:	Whole Sample	White Males	Black Males	White Females	Black Females
In the South	.019 (.041)	.045 (.059)	-.177 (.090)	.070 (.155)	-.074 (.125)
On a farm	-.160 (.030)	-.214 (.040)	-.112 (.064)	.063 (.107)	-.019 (.087)
In a city	.094 (.031)	.099 (.040)	-.017 (.069)	.222 (.093)	.025 (.091)

E. The Structural Impact of Education

The remaining variables, including some of the most important ones for explaining the wage rate, are endogenous and therefore appear only in the structural equations. I first consider the effects of educational attainment on hourly earnings.

The coefficients of each education class dummy (representing the incremental effect over the previous attainment level) are presented in Table 13 below. For the aggregate sample and for white males, the difference between 6-8 years of education and 0-5 years does not appear very great (perhaps 7%). Returns to entering and completing high school are more substantial.[31] White males achieve a considerable increment in wages by attending college, and gain a

TABLE 13

Coefficient of Education on Wages: Structure

Years of Formal Schooling:	Whole Sample	White Males	Black Males	White Females	Black Females
0 - 5	--	--	--	--	--
6 - 8	.065 (.041)	.075 (.065)	-.015 (.065)	-.179 (.227)	.037 (.124)
9 - 11	.135 (.031)	.133 (.043)	.175 (.063)	.247 (.128)	.014 (.082)
12	.074 (.031)	.084 (.040)	.110 (.075)	.022 (.109)	.016 (.098)
13 - 15	.159 (.037)	.155 (.042)	.012 (.115)	.107 (.119)	.132 (.192)
college degree	.169 (.055)	.212 (.059)	.136 (.259)	-.109 (.180)	.411 (.352)
advanced degree	.180 (.067)	.164 (.070)	.248 (.323)	.241 (.218)	-- --
Vocational Training or OJT	.044 (.026)	.031 (.032)	-.069 (.061)	+.147 (.085)	.153 (.092)

[31] Note that the coefficient for the third educational class indicates an 8% increment in wages from what may be a year or less of additional schooling.

further 21% by earning a college degree. Returns from higher degrees also appear quite substantial.

For black men, the picture is very different. Advancing to the 6th-8th grade level accomplishes nothing, but having at least some high school is quite important, as is obtaining a high school diploma. Returns to college education short of the degree are nil, and the gain from earning a baccalaureate appears to be substantially less than for the whites (though still notable). Higher degrees, though the standard error is large due to few observations, pay off very handsomely. Presumably these are black doctors and lawyers.

The impact of education on the wages of white female heads of households turns out to be extremely hard to estimate; witness the large standard errors. Perhaps this is because some females work at jobs which utilize their education, while others do not (e.g., college-educated secretaries), For what it is worth, the data seem to say that the large jumps in earnings occur when a white female enters high school, enters college and achieves an advanced degree. High school graduation is almost worthless, and completing grade school and college have anamolous negative impacts.

The pattern for black females is quite revealing. Returns from education up to and including a high school diploma (and only 13 of the 231 made it higher than this) appear to be negligible. Victims of both racial and sexual discrimination, these individuals seem unable to overcome their disadvantages through education.

Returns to higher education appear quite substantial, but the s.e.'s are too high to hold this conclusion with any conviction.

Voational education and on-the-job training are measured very imperfectly since there is no discrimination between people having a great deal of OJT and people with only a few month's training. Still the coefficients are interesting, suggesting that women stand to gain a lot more from vocational education than men.[32]

F. The Structural Impact of Occupation Class

The eight occupational categories do not arrange themselves in any natural order. I have used "craftsmen, foremen and kindred workers" as the control group, and one would guess that the first two job categories certainly earn higher wages while the last three surely earn less. It is not clear whether membership in "self-employed businessmen," a very heterogeneous group, or "clerical and sales workers" should receive positive or negative signs a priori. In fact, the data show strongly negative coefficients for self-employed businessmen, and a mixed pattern for clerical and sales help. Apparently, the latter are "good" jobs for black females, but "bad" jobs for white males. The coefficients, presented in Table 14 below, contain only a few surprises. For white males, being a professional or technical worker (this includes teachers) is only 5%

[32]The unexpected negative sign for black males may, perhaps, indicate that vocational education was obtaining by sacrificing formal schooling within an educational class.

TABLE 14

Coefficients of Occupational Groups on Wages: Structure

Occupation	Whole Sample	White Males	Black Males	White Females	Black Females
Professional, and technical	.101 (.049)	.049 (.053)	.159 (.174)	.130 (.332)	.170 (.474)
Managers, officials, proprietors	.177 (.048)	.124 (.049)	.567 (.344)	.133 (.354)	-- --
self-employed businessmen	-.308 (.065)	-.395 (.069)	-.019 (.214)	-.290 (.412)	-.569 (.532)
clerical and sales workers	-.040 (.040)	-.085 (.049)	+.011 (.115)	-.024 (.305)	.114 (.358)
craftsmen, foremen, and kindred	0 --	0 --	0 --	0 --	0 --
operatives and kindred	-.146 (.034)	-.162 (.040)	-.193 (.067)	.014 (.309)	.122 (.346)
laborers, service workers, farm hands	-.233 (.036)	-.333 (.053)	-.123 (.067)	-.263 (.305)	-.126 (.341)
farmers and farm managers	-.690 (.073)	-.719 (.076)	-1.051 (.213)	-- --	-- --

more renumerative than being a skilled craftsman. Since this pattern is not repeated in the other groups, I suspect it is a reflection of the inflated wages in the lily-white and all-male construction trades.[33] Managers and officials are the highest paid occupational

[33]Recent work by Orley Ashenfelter (1971a) provides some support for this notion. For the construction industry, he found (Table 4) the following union-nonunion differentials for white males: 33% for craftsmen, 36% for operatives and 39% for laborers. By contrast the corresponding union-nonunion differentials in these three occupation groups for all industries except construction were 3%, 14% and 18%.

category in every group. Black males appear to gain dramatically, by reaching this level of the job market, but the huge coefficient can not be trusted. The very low effective wages for self-employed businessmen reflect what most economists know, but President Nixon (and other "black capitalism" advocates) apparently do not know. Namely, that most small businesses in this country are failures, and earn for their proprietors less than the returns they could earn elsewhere in the economy.[34] Operatives and kindred workers earn 16-19% less than craftsmen. The positive coefficients for females simply reflect the fact that there are hardly any female craftsmen in the sample. Laborers earn 17-27% less than operatives, except for black males who actually earn more. Finally, the money wages of farmers are extraordinarily low.[35] This is not surprising in view of the lingering overpopulation of farm areas and the fact that much farm income is in kind rather than in money.

Table 15 below gives the estimated proportional impacts on wages of being a union member and being a veteran of the armed forces (estimated for men only). The gains from union membership (union due dues are <u>not</u> deducted from wages) are quite dramatic for all groups, especially black males. The greater gains for blacks may indicate that inclusion under uniform union wage agreements prevents employees

---

[34]Of course, there may be objectives for the "black capitalism" program other than raising black incomes.

[35]For blacks, farmers are estimated to earn 105% less than craftsmen!

TABLE 15

Coefficients of Union Membership and Veteran Status: Structure

Variable	Whole Sample	White Males	Black Males	White Females	Black Females
Union Member	.292 (.026)	.228 (.032)	.393 (.056)	.276 (.122)	.337 (.101)
Veteran	.038) (.025)	.029 (.028)	.099 (.053)	-- --	-- --

from discriminating against blacks in rates of pay (though discrimination in hiring remains possible). The estimated impacts of union membership -- I do not say "unionization" since unions may selectively organize the higher paid workers -- are much higher than the rather low coefficients that troubled Hall (1970, p. 29). His estimates of the union differential ranged from 7% to 17%; mine range from 23% to 39%. Of course, my estimates are higher than most others. For example, Ashenfelter's estimates (1971b, page 24) range from 7% to 21% and R. Oaxaca's (1971, p. 13) range is essentially the same. The reason for such widely diverging results remains obscure[36] but I, for one, find it hard to believe that a black female adds only 7% to her depressed wage by attaining union membership.

For whites, our a priori speculations about the effect of service in the armed forces (i.e., that it should have about the same effect as vocational training) is strikingly confirmed. Compare the coefficient for vocational education in Table 13 above. However,

[36]Ronald Oaxaca has suggested to me that, since I do not include city size in my regressions, the union variable may be picking up some of the effect of city size. Why city size should matter, given that I have included two measures of local labor market conditions, is a mystery to me. But it might!

blacks appear to achieve a 10% increment in wages through military service.

G. The Impact of Tenure-on-the-Job

The only remaining variable is job tenure, i.e., "length of time working for the present employer." In the theoretical model, I have listed this as an endogenous variable as, indeed, it should be. However, the ability to predict $T_1,\ldots,T_6$ from the reduced form proved to be so minimal that, after experimenting with T_i as endogenous and as exogenous, I decided that the estimates with job tenure exogenous, though sligtly biased, were more reliable. Table 16 below reports these results. Since the control group is workers for whom "tenure" is not applicable (e.g., the self-employed), it is hard to place a priori sign restrictions on the coefficients. However, since the coefficients are for marginal impacts, they certainly should be positive after the first. Unfortunately, except for white males, this does not prove to be the case. The general impression left by this table is that the effects of longevity on the job vary greatly both over time and across race-sex groups. Only very long tenure with the same employer (20 years or more) earns workers a reliable 8-12% gain in wages.

VI. Some Conclusions

The distribution of wages is probably the principal determinant of the distribution of both income and economic welfare, however measured. But, if the ultimate interest is in public policy

TABLE 16

Coefficients of Job Tenure: Structure

Time on the Job:	Whole Sample	White Males	Black Males	White Females	Black Females
Less than $\frac{1}{2}$ year	-.040	-.085	.393	-.027	-.132
	(.054)	(.065)	(.139)	(.237)	(.256)
About 1 year	.018	.025	-.103	.047	.087
	(.048)	(.069)	(.101)	(.146)	(.122)
2-3 years	.065	.020	.122	-.075	.250
	(.042)	(.060)	(.089)	(.129)	(.111)
4-9 years	.043	.056	.020	.189	-.068
	(.034)	(.045)	(.073)	(.119)	(.096)
10-19 years	.040	.078	-.001	.024	.010
	(.032)	(.041)	(.068)	(.110)	(.096)
20 years or more	.091	.081	.082	.125	.099
	(.037)	(.044)	(.081)	(.151)	(.201)

to equalize incomes, knowing this is less important than knowing why wage rates are so disperse. Among the obvious reasons are (a) differences in education; (b) differences in ability; (c) racial and sexual discrimination; and (d) trade unions.

The present study, especially Table 13 above, suggests that educational differences are very important determinants of wage differentials, and that the benefits from education differ radically across race-sex groups. Part of this, no doubt, can be attributed to discrimination in labor markets and part can be attributed to differences in ability and tastes. No sensible measures of ability were available for this study; but recent work suggests that this may not be a serious omission [Gintis (1971), Griliches and Mason (1970)].

Discrimination appears to be a pervasive phenomenon throughout the economy, and, quantitatively, an important source of wage dispersion. Differential attainments of education, occupation, etc. can account for only about one third of the 35% wage differential in favor of whites. The remainder can only be attributed to outright discrimination against blacks in rates of pay.[37] Similar remarks hold for females. Relative to males, they are at a 46% disadvantage in wages, only one-quarter of which can be accounted for by inferior education and occupation. In fact, on the whole, females do not have inferior education.

Members of unions earn dramatically more than otherwise identically-situated non-members. Whether this is evidence of union power or of a tendency for unions to organize the more prosperous workers is an open question. Such large union differentials tend to be disequalizing, unless unions specialize in organizing those persons who would otherwise be disadvantaged. The results from predicting union membership as an endogenous variable (not reported here) do not suggest that this is the case. However, neither do they suggest the opposite --i.e., that being born into an underprivileged family makes it harder to gain admittance to a union. Unions, therefore, may not have a very great impact on the over-all income distribution.

---

[37] Of course, if we seek to measure the total discrimination that takes place in the labor market, we should include discrimination in occupational status as well.

Appendix: The Sample Characteristics

The following table exhibits the major characteristics for the aggregate sample of 2131 household heads, as well as for the separate race-sex groups. Since most variables are dummies, the figures indicate the number of persons falling in each category. The only exception is the "number of siblings if ≤ 7" variable, where the average is given. The reader can see for himself in what ways the sample is un-representative of the U.S. population. Attention is called particularly to the over-representation of blacks, Southerners, and persons who grew up in poor, Southern farm families.

TABLE 17

Sample Characteristics

Characteristic	Whole Sample	White Males	Black Males	White Females	Black Females
Race:					
White	1433	1239	--	194	--
Black	698	--	467	--	231
Sex:					
Male	1706	1239	467	--	--
Female	425	--	--	194	231
Education Level:					
0-5	201	61	115	6	19
6-8	464	244	124	27	69
9-11	457	21	123	35	89
12	554	370	67	76	41
13-15	244	178	28	28	10
college degree	124	103	4	14	3
advanced degree	87	73	6	8	0

-continued on next page-

Duncan, Otis D., "Ability and Achievement," Eugenics Quarterly, 1968, 1-11.

Friedman, Milton, Essays in Positive Economics, Chicago, 1953.

Fuchs, Victor R., Differentials in Hourly Earnings by Region and City Size, NBER Occasional Paper No. 101, New York, 1967.

Gintis, Herbert, "Education, Technology and the Characteristics of Worker Productivity," American Economic Review, May 1971, 266-279.

Goldberger, Arthur S., Econometric Theory, New York, 1964.

Griliches, Zvi, "Notes on the Role of Education in Production Functions and Growth Accounting," in W. Lee Hansen (ed.) Education, Income and Human Capital, NBER, New York, 1970.

_____, and W. Mason, "Education, Income and Ability," mimeo, Harvard University, 1970.

Hall, Robert E. "Wages, Income and Hours of Work in the U.S. Labor Force," M.I.T. Working Paper No. 62, August 1970.

Hanoch, Giora, "An Economic Analysis of Earnings and Schooling," Journal of Human Resources, 1967, 310-329.

Hill, T.P., "An Analysis of the Distribution of Wages and Salaries in Great Britain," Econometrica, 1959, 355-381.

Lansing, John B. and James N. Morgan, "The Effect of Geographical Mobility on Income," Journal of Human Resources, 1967, 449-460.

McFadden, Daniel, "The Revealed Preferences of a Government Bureaucracy," Technical Report No. 17, Project for the Evaluation and Optimization of Economic Growth, Institute for International Studies, Berkeley, November 1968.

Morgan, James N., Martin H. David, Wilber J. Cohen and Harvey Brazer, Income and Welfare in the United States, New York, 1962.

_____ and _____, "Education and Income," Quarterly Journal of Economics, 1963, 423-437.

_____, and James D. Smith, A Panel Study of Income Dynamics: Study Design, Procedures, and Forms; 1969 Interviewing Year (Wave II), Survey Research Center, Ann Arbor, 1969.

Oaxaca, Ronald, "Sex Discrimination in Wages," paper read at the Princeton University Conference on "Discrimination in Labor Markets," October 1971.

Survey Research Center, A Panel Study of Income Dynamics: Study Design, Procedures, Available Data; 1968-1970 Interviewing Years (Waves I-III), Ann Arbor, 1970.

Taylor, David P., "Discrimination and Occupational Wage Differences in the Market for Unskilled Labor," Industrial and Labor Relations Review, 1968, 375-390.

Weiss, Randall D., "The Effect of Education on the Earnings of Blacks and Whites," Review of Economics and Statistics, 1970, 150-159.