Yale University

EliScholar – A Digital Platform for Scholarly Publishing at Yale

Discussion Papers

Economic Growth Center

12-1-1982

Unbiased Estimates of the Wage Equation When Individuals Choose Among Income-Earning Activities

Wil Vijverberg

ECONOMIC GROWTH CENTER

YALE UNIVERSITY

Box 1987, Yale Station
New Haven, Connecticut

CENTER DISCUSSION PAPER NO. 429

UNBIASED ESTIMATES OF THE WAGE EQUATION WHEN INDIVIDUALS CHOOSE AMONG INCOME-EARNING ACTIVITIES

Wim P.M. Vijverberg

December 1982

Notes: Support from NIH Training Grant 2-T32-HD07146-05 and The William and Flora Hewlett Foundation is gratefully acknowledged. I would like to thank those who made valuable comments during the course of the work that led to this paper: Frank Anscombe, John Beggs, Paul Schultz, Ken Wolpin, and participants of the Labor and Population Workshop.

UNBIASED ESTIMATES OF THE WAGE EQUATION WHEN INDIVIDUALS CHOOSE AMONG INCOME-EARNING ACTIVITIES*

1. Introduction

One of the most stable relationships in economics is the human capital wage determination equation. Numerous studies, using data from the U.S. as well as low income countries, have found that education raises the individual's wage rate, and that experience affects wage rates in a non-linear way, peaking in the age cross-section generally at around thirty years of experience.

The "wage rate" is an aggregate concept, in that it represents the returns per hour of work regardless of the type of work performed. This paper is a first step toward a disaggregation of wages according to income-earning activities. The activities distinguished here are work in the wage labor market, and work in self-employment. The motivation for this paper is threefold. First of all, an issue that has not been well analyzed is whether the wage equation is stable across different types of income-earning activities. Instability in this sense may be due to three reasons. First, wage may be affected differently by exogenous variables, if the labor market does not work sufficiently smoothly. Workers cannot always switch readily between activities in response to the wage discrepancies.[1] Second, wage are only a part of the returns to work. If non-pecuniary benefits are affected by the same set of exogenous variables as pecuniary benefits (or a subset of it), wage equations are expected to differ between activities. In this light one could explain the common finding that the marginal return to education appears to decrease with the level of education. Third, the pecuniary benefits of work in different activities may include different kinds of returns. This is quite obviously the case when, as is the focus of this paper, one distinguishes between labor supplied to the wage labor market and labor engaged in self-employment. By dividing total self-employment earnings by self-employment hours of work, one obtains a so-called self-employment wage rate. One may expect that the impact of wage determinants on market wages differs from their impact on self-employment wages, since, in addition to labor income, self-employment earnings include returns to supervisory and management skills, and returns to assets owned by the individual and used in the production process.

The second motivation for this paper is that little is known about self-employment wages. In most wage studies it is recognized, that self-employed workers are different from market workers, but the most common treatment of individuals reporting self-employment earnings is to omit them from the sample and to estimate wage equations for employees only.[2)] As Table 1 indicates, this is not the optimal solution to the problem. In low income countries up to one half of the labor force is self-employed, and even in industrialized countries 10 to 15 percent adds up to a sizeable group in absolute numbers. Restricting the sample to employees is likely to introduce self-selection bias in the estimates: often the proportion of self-employed in an age group increases with age, due, e.g., to inheritance and asset accumulation. If one widens the analysis to all (market and self-employed) workers, such self-selection bias can be avoided and insights are gained into self-employment wages, providing a more complete picture of the distribution of income. Moreover, knowing or being able to impute self-employment wages is beneficial in studies on a variety of other problems, in which wages are a crucial price variable, e.g., fertility, migration, technological change (labor- versus capital-saving).

The third motivation is related in part to the data set used for this study, the Malaysian Family Life Survey. The Malaysian government has recently enacted programs aimed at improving the economic well-being of the Malay population relative to the Chinese. The question asked is to what extent the Chinese - Malay wage gap, observed to be 37 percent in the data set used, is due to an unexplained purely ethnic wage differential. Similar problems in other studies of wages abound: if one could identify choices that have a influence on observed wages, those choices will have implications for the size of any wage gap.

This paper proposes a simple two-stage procedure to compare market and self-employment wages, incorporating the individual's choice between self-employment and market work. In the first stage one estimates the parameters of the underlying structure of the choice facing the individual. This structure is an example of the now familiar index function framework (see the survey by Heckman and MaCurdy (1981)). While in this paper the choice is restricted to two alternatives, the index function framework can in principle handle choices between three or more alternatives. The second stage of the empirical analysis uses the information gained in the first stage in order to obtain consistent estimates of the wage equation in each activity.

The empirical analysis in this paper distinguishes regional, ethnic and human capital variables as well as unmeasured differences in wages. Among workers in Malaysia, the effect of race and human capital variables on market and self-employment wages appear to differ, but evidence that region affects these two sources of wages differently is ambiguous. With respect to the 37 percent Chinese - Malay wage gap, it is found that about one quarter is due to an unexplained purely ethnic wage differential. The rest is explained by, among others, choice of activity and human capital differences.

In section 2 there is a short survey of the literature. The model analyzing the choice between merket and self-employment is described in section 3, and estimated relationships are derived. Data and variable definitions are given in section 4, and the empirical results are reported in section 5. Section 6 summarizes, and examines the new evidence in a broader perspective, considering issues raised in this introduction.

2. Overview of Existing Evidence

To the basis of the human capital wage determination equation lies the work of among others Becker(1975), Ben-Porath (1967) and Mincer (1970,1974). The underlying idea is that individuals incur expenditures and forego earnings in order to accumulate skills (human capital) that increase earnings at a later stage of the lifecycle. Investments in human capital take the form of full-time or part-time education, on-the-job training or any type of skill accumulation that increases future prodictivity.

Since the 1960's a large number of studies have been devoted to a test of the human capital hypothesis. Although sceptics may say that the estimated parameters vary quite a bit across these studies, the hypothesis has received strong support. As the surveys by Blang (1976), Mincer (1970, 1976), Rosen (1977) and Psacharopoules (1973) indicate, the rate of return to education appears to lie in the range of 7 to 12 percent, with some studies reporting rates of return in excess of 15 percent. Rates of return in developing countries tend to be a little higher than those found in the high income countries.

Little attention is paid to the difference in the sources of income. Most studies using data from developing countries, where, as argued in the introduction, the problem may be more severe, ignore self-employed workers and estimate wage equations with a sample of employees.[3)]

One could easily avoid this issue by assuming that self-employed workers are merely workers who happen to work in the entreprise they founded themselves. But from a theoretical viewpoint there are at least three reasons that employers-- to emphasize the distinction -- perform different roles in a firm from employees. First of all, they organize the firm, analyzing the abilities of its employees, matching employees to tasks and to teams, and managing and investing in the firm-specific stock of human capital of the employees. (Prescott and Visscher (1980)). Secondly they bear the risk of production. (Knight (1921), Schultz (1980), Kihlstrom and Laffont (1979)). In a static sense, the environment (e.g. weather) renders the outcome of the firm's production uncertain. In a dynamic sense, production opportunities (technology) change in an uncertain way, to which the entrepreneur can react by committing resources in a timely fashion: there are windfall profits for temporarily being ahead of the competition.[4)] Similarly, entrepreneurs gain by adjusting quickly to changes in input and output prices. Thirdly, self-employment income reflects the contribution of inputs brought into the firm by the employer. Such inputs include both marketable inputs that could command a certain competitive price when sold, as well as fixed endowments.

The organizational and allocative aspects of the productivity of self-employed labor are likely to be enhanced by the individual's stock of human capital. Due to reasons indicated in the introduction as well as the difference between self-employed and market workers, the impact of human capital measures on self-employment and market wages may differ. One may also expect that such differences vary between data sets. Returns to education are, according to the argument above, much lower in a backward agricultural society than in a rapidly developing environment with many dynamic opportunities.

Only a few studies address this issue empirically. Among these, Kuznetz (1966, pp. 177-179) focuses on the contribution of entrepreneurial labor and of equity or capital to total income of entrepreneurs. In lack of better data, one could calculate one of the components by some reasonable assignment and treat the other component as a residual. For a small set of industrialized countries, with aggregate income data, Kuznetz prefers to assign to entrepreneurs a per capita labor income equal to per worker compensation of employees.

Later studies work with household level data. Blaug (1974) treats the problem by entering a dummy variable into the wage equation indicating whether the individual is an employee or a self-employed or family worker. In his sample from Thailand employees appear to have 20 to 25 percent higher earnings. Fields and Schultz (1982) separated their Colombian sample into employees and self-employed workers, carrying out the same analysis on both subsamples. In a sample combining rural and urban areas, employers appear to have 7 percent higher earnings than employees, but human capital variables affect earnings of both groups similarly. Disaggregation of the two subsamples according to the rural/urban classification uncovered differences between employers and employees, mainly in rural areas between the different Colombian regions.

Chiswick (1977a) is able to pool self-employed and market workers into one sample by assuming that a fixed proportion of the self-employment earnings is attributable to labor. The estimated wage equation contains an additional exogenous variable measuring the ratio (p) of self-employment earnings to total earnings. The coefficient on this variable indicates that for Thai males labor earnings comprise 57 percent of self-employment earnings. Chiswick's approach is interesting as a direct attempt to analyze self-employed earnings. However, it is questionable whether the ratio p is exogenous. In fact, Chiswick argues: "Self-employment and wage employment are alternative modes competing for the same labor; workers choose one or the other according to their relative earnings potential..". (1977a, p. 78, emphasis added) This choice is exactly the rationale for the approach offered in this paper.

3. Different Income Sources and the Wage Equation

In the most commonly estimated wage equation the logarithm of the wage rate is regressed on a set of explanatory variables including education, experience, race and region. This section demonstrates how the wage equation can be modified to compare wages earned in different activities.

3.1. Endogeneity in the Wage Equation

When the sample contains observations of market workers as well as self-employed workers, the wage equation contains an endogenous element. To see this, let Y_i be income earned in activity i, i=m (market work) or s (self-employment). Let H_i and W_i be hours of work and wage rate in activity i. Then total income Y from employment is

$$\begin{aligned} Y &= Y_m + Y_s \\ &= W_m H_m + W_s H_s \end{aligned} \tag{3.1}$$

Dividing Y by total hours of work yields a wage rate W that is the weighted average of W_m and W_s:

$$W = \frac{Y}{H_m + H_s} = \alpha W_m + (1-\alpha) W_s \tag{3.2}$$

where

$$\alpha = \frac{H_m}{H_m + H_s} \tag{3.3}$$

Clearly α is an endogenous variable, as both H_m and H_s are determined by W_m and W_s. If one estimates the parameters of some assumed functional specification for W_m and W_s without correcting for the endogeneity of α, parameter estimates would be biased. Such bias can be avoided by using an instrumental variable for α.[5)] We proceed by constructing a time allocation model analyzing the choice of market and self-employment hours. This model will indicate which variables play a role in determining α. We then estimate the parameters linking these variables to α and use these estimates in the analysis of the wage equation.

3.2. Choice of Activity

We assume that the individual maximizes his utility U, which is a function of leisure L, and commodities C, purchased in the market at price P_c.

$$U = U(L, C) \tag{3.4}$$

The individual divides his time T between leisure, market work H_m and self-employment work H_s. L, H_m and H_s are non-negative.

$$T = L + H_m + H_s \tag{3.5}$$

Labor used in self-employment is expressed in efficiency units. Per hour of self-employment work the individual contributes s efficiency units of labor. So if H_s^* is the individual's labor input in self-employment measured in efficiency units, we have:

$$H_s^* = s\, H_s \tag{3.6}$$

In addition to H_s^*, self-employment production uses as inputs hired labor H_h^* (measured in efficiency units, hired at price W_h per efficiency unit), other market inputs R, purchased at price P_r, and a set of exogenously given endowments or quasi-fixed factors Z.

$$Q = Q(H_s^*, H_h^*, R, Z) \tag{3.7}$$

Note that in our formulation hired labor is not perfectly substitutable for own labor.[6)]

Sources of income are sales of production Q at price P_q, hours of work spent in the wage market earning a wage rate W_m, and an exogenous flow of income V. Outlays are the purchase of commodities C, the hiring of labor H_h^*, and the purchase of market inputs R. The budget equation indicates that all income is spent:[7)]

$$P_q\ Q + W_m\ H_m + V - P_c\ C - W_h\ H_h^* - P_r\ R = 0 \tag{3.8}$$

The time allocation model is completed by the assumption that U is a concave function, and that Q is concave in the inputs H_s^*, H_h^* and R.

Before tieing this model in with the wage equation, note that, in the case that self-employment and market productivities are identical (i.e., $W_m \equiv W_h\ s$), the model constructed above is quite similar to the work of Huffman (1980) analyzing the allocation of labor to farm and off-farm activities.[8)] It differs from models developed by Rosenzweig (1980) and Barnum and Squire (1979) in the substitutability assumption between own and hired labor in self-employment production. Perfect substitutability implies that an individual will never be part-time self-employed and part-time market worker, while at the same time hiring market labor for self-employment production.[9)] If self-employment and market productivities are different, the perfect substitutability assumption leads to some rather non-sensical results.[10)] Compared to Huffman's model, our distinction between W_m and s leads more directly to an estimable wage equation.

The interior solution of the model specifies demand functions for hours of leisure and self-employment work; market hours are a residual:

$$L = L(W_m, s, V, P_c, P_q, P_r, W_h, Z) \tag{3.9}$$

$$H_s = H(W_m, s, P_q, P_r, W_h, Z) \tag{3.10}$$

$$H_m = T - L - H_s \tag{3.11}$$

Casual observation of most labor markets shows that most individuals spend their time either on market work or on self-employment work. Indeed only few divide their time between both activities. So the corner solutions of H_s=0 and H_m=0 are important constraints in our model.[11)] Let us define α^* as the desired ratio of hours of market work to hours of all work when H_s and H_m could take on negative values. The relation between α^* and α (defined in equation (3.3)) is:[12)]

$$\begin{aligned} \alpha &= 1 \quad \text{if} \quad \alpha^* > 1 \quad \text{(full-time market work)} \\ \alpha &= \alpha^* \quad \text{if} \quad 0 \leq \alpha^* \leq 1 \quad \text{(interior solution)} \\ \alpha &= 0 \quad \text{if} \quad \alpha^* < 0 \quad \text{(full-time self-employed)} \end{aligned} \tag{3.12}$$

where, through (3.9) to (3.11), α^* can be written as:

$$\begin{aligned} \alpha^* &= \frac{T - L - H_s}{T - L} \\ &= \alpha^*(W_m, s, V, P_c, P_q, P_r, W_h, Z) \end{aligned} \tag{3.13}$$

Under the assumption that prices are constant across all individuals and that W_m and s are affected by a set of exogenous variables X, we can linearize α^* as (adding a normally distributed $N(0,\sigma_{uu})$ random error u):

$$\alpha^* = X\,\delta_x + V\,\delta_v + Z\,\delta_z + u \qquad (3.13)'$$
$$= X_\alpha\,\delta + u$$

where X_α is a vector containing X, V, Z_2 and Z_3.

Equations (3.12) and (3.13)' describe an empirical model of how a set of exogenous variables (X, V and Z) affects the choice of activity represented by α. The dependent variable α has both a lower and an upper limit. Thus maximum likelihood is the appropriate estimation technique.[13)]

Let $\bar{\delta}_i$ be the estimated value of δ_i, i=x,v,z. Let $\bar{\alpha}$ be:

$$\bar{\alpha} = X\,\bar{\delta}_x + V\,\bar{\delta}_v + Z\,\bar{\delta}_z \qquad (3.14)$$
$$= X_\alpha\,\bar{\delta}$$

So $\bar{\alpha}$ is the estimate of α^*. $\bar{\alpha}$ and the estimate of σ_{uu}, $\bar{\sigma}_{uu}$, are used in the analysis of the wage equation to obtain an instrumental variable for α.

Many data sets report only the primary activity of the individual. No detailed information is provided on how hours are divided. The approach proposed above can easily be modified to accommodate this situation. Assume that the individual reports as his primary activity that one, in which he spends the largest number of working hours. So he is called self-employed (wage market worker) if $H_s > H_m$ ($< H_m$). This translates into a value for α^* less than 0.5 for self-employed workers, and α^* is greater than 0.5 for market workers. The parameters of (3.13)' can be estimated by probit analysis rather than by two-limit tobit, as proposed above; the parameter estimates are standardized by $\sigma_{uu}^{-\frac{1}{2}}$, which is unknown due to the nature of the probit technique.[14)]

This modification brings out the similarity with the index function framework (Heckman and MaCurdy, 1981). If no information on hours in both activities is available, the approach proposed in this section follows that framework strictly. On the other hand, if hours are known, the argumentof the index function (i.e., α^*) is known for workers participating in both activities. This information can be used advantageously in the empirical analysis.

3.3. The Instrumental Variable in the Wage Equation

The most commonly estimated wage equation has a semi-log linear form. For reasons of comparison, we adhere to this practice. Assume that the market wage W_m and self-employment wage W_s depend on the set of exogenous variables X and Z_3 in the following way:

$$\ln W_m = X\,\beta_m + \varepsilon_1 \qquad (3.15)$$

$$\ln W_s = X\,\beta_s + Z_3\,\beta_{sz} + \varepsilon_2 = X_s\,\beta_s' + \varepsilon_2 \qquad (3.16)$$

The vector X_s contains both X and Z. ε_1 and ε_2 are random variables, correlated with u. It is assumed that ε_1, ε_2 and u are jointly normally distributed with mean 0 and covariance matrix Σ where:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{1u} \\ \sigma_{12} & \sigma_{22} & \sigma_{2u} \\ \sigma_{1u} & \sigma_{2u} & \sigma_{uu} \end{pmatrix}$$

To connect (3.15) and (3.16) with the definitional relationship (3.2), take the logarithm of (3.2) and approximate the resulting expression around $W_m^o = 1$ and $W_s^o = 1$:

$$\begin{aligned} \ln W &= \alpha\, W_m + (1 - \alpha)\, W_s - 1 \\ &= \alpha\, (W_m - 1) \;+\; (1 - \alpha)\, (W_s - 1) \end{aligned} \qquad (3.17)$$

Realizing that $\ln W_m \simeq W_m - 1$ and $\ln W_s \simeq W_s - 1$, we get, by combining (3.15) to (3.17):

$$\ln W = \alpha\, X\, \beta_m + (1 - \alpha)\, X_s\, \beta_s' + \varepsilon \qquad (3.18)$$

where $\varepsilon = \alpha\, \varepsilon_1 + (1 - \alpha)\, \varepsilon_2$.

If α were an exogenous variable, equation (3.18) could be estimated by OLS or, in view of the definition of ε, by GLS. These estimates will be referred to as GLS1 and GLS1 estimates. However, α is a random variable correlated with ε_1 and ε_2. This implies that both OLS1 and GLS1 estimates are biased for two reasons. First, interpreting αX and $(1-\alpha) X$ as regressors, the error term ε is correlated with the regressors. Second, the mean of ε, being a function of $X_\alpha\, \delta$, does not equal 0. Both causes of bias are due to the endogeneity of α. Therefore OLS1 and GLS1 estimation should not be utilized to compare β_m with β_s.

The key to dealing with this estimation bias lies in the recognition of the individual's behavior in choosing α, as represented by (3.12) and (3.13)'. Combining these equations with (3.18) gives one system:

$$\ln W = X\,\beta_m + \varepsilon_1 \qquad \text{if} \quad u \geq 1 - X_\alpha\,\delta$$

$$= X_\alpha\,\delta \quad X\,\beta_m + (1-X_\alpha\,\delta)\,X_s\,\beta_s' + \eta \quad \text{if} \quad -X_\alpha\,\delta < u < 1 - X_\alpha\,\delta$$

$$= X_s\,\beta_s' + \varepsilon_2 \qquad \text{if} \quad u \leq -X_\alpha\,\delta \qquad (3.19)$$

where $\eta = (X\,\beta_m - X_s\,\beta_s')\,u + X_\alpha\,\delta\,\varepsilon_1 + (1-X_\alpha\,\delta)\,\varepsilon_2 + u\,\varepsilon_1 - u\,\varepsilon_2$.

Calculate the mean of $\ln W$ over all values of u, ε_1 and ε_2:

$$E(\ln W|X) = \pi\,X\,\beta_m + (1-\pi)\,X_s\,\beta_s' + (\sigma_{1u} - \sigma_{2u})\,X_c \qquad (3.20)$$

where $\pi = 1 - (1-X_\alpha\,\delta)\,F\left[\sigma_{uu}^{-\frac{1}{2}}(1-X_\alpha\,\delta)\right] - X_\alpha\,\delta\,F\left[-\sigma_{uu}^{-\frac{1}{2}}\,X_\alpha\,\delta\right] -$

$$- \sigma_{uu}^{-\frac{1}{2}}\left\{f\left[\sigma_{uu}^{-\frac{1}{2}}(1-X_\alpha\,\delta)\right] - f\left[-\sigma_{uu}^{-\frac{1}{2}}\,X_\alpha\,\delta\right]\right\}$$

$$X_c = F\left[\sigma_{uu}^{-\frac{1}{2}}(1-X_\alpha\,\delta)\right] - F\left[-\sigma_{uu}^{-\frac{1}{2}}\,X_\alpha\,\delta\right]$$

and where f and F are the standard normal density and distribution function respectively. It appears that the mean value of $\ln W$ is a weighted average of the two wages--market and self-employment--, the weight π being an increasing function of $X_\alpha\,\delta$. For $X_\alpha\,\delta = .5$, π equals .5, while for $X_\alpha\,\delta$ approaching ∞ $(-\infty)$, π approaches 1 (0). The steepness of this S-function depends on σ_{uu}. The better the fit of the choice (i.e., the lower σ_{uu}), the steeper is the function. In addition, the mean value of $\ln W$ is determined by a correction factor X_c, partially stemming from the product of errors $u\,\varepsilon_1 - u\,\varepsilon_2$, and partially due to the division of the error domain in three parts (see (3.19)).

Now substitute the values $\bar{\sigma}_{uu}$ and $\bar{\alpha} = X_\alpha\,\bar{\delta}$, which were estimated in the first stage, into the expressions of π and X_c, obtaining $\bar{\pi}$ and $\bar{X}_c$, which obviously vary between individuals. Equation (3.20) implies that β_m, β_s' and $\sigma_{1u} - \sigma_{2u}$ can be estimated consistently by the OLS technique from the following regression:

$$\ln W = (\pi X)\,\beta_m + ((1-\pi)\,X)\,\beta_s' + (\sigma_{1u}-\sigma_{2u})\,\bar{X}_c + \theta_1 \qquad (3.21)$$

where θ_1 is an error with mean 0. This estimator will be referred to as OLS2.

The variance of $\ln W$ can be calculated as a large expression in terms of $X_\alpha\,\delta$, σ_{uu}, $(X\,\beta_m - X_s\,\beta_s')$, σ_{11}, σ_{12}, σ_{22}, σ_{1u}, and σ_{2u}. The error of a regression that involves the <u>true</u> values of α_1 and X_c is heteroskedastic. This suggests a GLS estimator (by substituting $\bar{\sigma}_{uu}$ and $\bar{\alpha} = X_\alpha\,\bar{\delta}$ for σ_{uu} and $X_\alpha\,\delta$ in π, X_c and this variance) that is likely to be more efficient than the OLS2 estimator. But since it does not take the difference between $(\bar{\delta}, \bar{\sigma}_{uu})$ and (δ, σ_{uu}) into account, this estimator, referred to as GLS2, still will not be most efficient (see a parallel problem in Heckman, 1979).

In the appendix to this paper, another estimation procedure (called GLS3) is suggested, that is of much use in problems where individuals are forced to choose only one alternative, rather than participate in both alternatives at the same time. When the latter option is available, information on individuals choosing this option is left unused by the GLS3 estimates. Therefore we shall concentrate on the OLS2 and GLS2 estimates.

One might suggest that it would be preferable to obtain the exact distribution of the random error of (3.19) and to perform estimation on the likelihood function based on this distribution. But (3.19) involves the product of two normal variables, the distribution of which does not seem to have an analytical solution to simpler problems than considered here (Kendall and Stuart (1977), pp. 286, 424).

The OLS2 and GLS2 estimation procedure are relatively simple to implement. Testing for equality in the determinants of W_m and W_s is done simply by restricting β_m to be equal to β_s. We can also test for equality of a subset of parameters, e.g., those related to human capital accumulation. The estimated parameters β_m and β_s are readily comparable with those obtained in other studies using the common semi-log linear equation approach. So while unfortunately we need to make two approximations to reach equation (3.18), the approach proves quite versatile.

Before turning to the empirical implementation we want to point out the difference between the estimation method proposed in this paper and the well-known switching regression technique. According to this technique, if one knows how the sample is divided between the two (or more) alternative regimes, OLS regressions are run on the sample in each regime and the Chow (1960) test is used to test for similarity of the coefficients. This causes no problem if the sample is divided in some exogenous way (i.e., "nature chooses").[15] But when the individual chooses, the random factor in his choice is likely to be correlated with the error term of the regression equation, causing biased parameter estimates as demonstrated with equation (3.18). This bias is eliminated using the instrumental variable $\bar{\pi}$, or by using the procedure suggested in the appendix.

4. Data and Variable Specification

Data used for this study are obtained from the Malaysian Family Life Survey (MFLS), conducted in 1976 by the Rand Corporation. The MFLS survey includes interview responses from 1261 ever-married females 50 years of age or younger and, if available, their husbands (See Jones and Spoelstra, 1978). A multi-stage stratified sampling design was employed whereby 51 communities were chosen from which eligible females were chosen for interviews. The primary reason for this survey was to gather data to study patterns and correlates of fertility (hence the particular sampling frame), but a wide variety of other data were collected, such as time allocation variables, characteristics of the communities, income and asset variables, inputs and outputs of family businesses.

The variables of importance to our wage equation are distinguished in three categories. The first contains human capital variables: education (i.e., years of schooling), and experience defined as (age - education - 6) in the Mincerian fashion. Experience enters the wage equation as a linear and a squared term. The second category refers to ethnic variables. Two dummy variables take on the value of unity for Malays and Chinese respectively; the excluded group is Indians and others. The third category contains rural-urban variables. One dummy indicates urban market centers (Kuala Lumpur, Ipoh and Penang); another indicates other urban areas. The excluded group is rural areas. In addition to these three categories of variables we expect land (measured in acres) to affect self-employment wages as a fixed factor of production (i.e., Z' in section 3.3)

The choice between market work and self-employment is also affected, according to the model of section 3.2, by income from sources other than labor (V). "Other income" includes payments and gifts from children to parents, pension income, dowries, etc.

The endogenous variables of our study are α, the fraction of hours of work spent in the labor market, and wage rates, defined as the sum of earnings from market labor and self-employment divided by total hours of work. The MFLS survey distinguishes a number of activities that can all be described as self-employment. This subdivision created some problems, however, since no earnings were reported for some of those activities. Our strategy is the following. There is one group of activities that is described as self-employed or worker on own account, worker in a family business, and employer. A second group is described as home products and services for sale, and home products and services for own use. If within a group wages are reported for one activity but not for another, we assign a wage rate to the missing one equal to the average wage rate for the other activities within the same group.[16)] If no wage can be calculated for one of the two activity groups, the observation is omitted.

Table 2 contains the descriptive statistics of the variables used for this study. The sample contains 45.9 percent Malays, 41.2 percent Chinese and 12.9 percent Indian and others. 20.9 percent of the observations come from urban market centers, 24.1 percent from other urban areas and 55 percent from rural areas. The mean of α indicates that on average 54.7 percent of the hours worked are spent in the wage labor market.

The mean of the log of market wages of those individuals that spend time in the market exceeds the mean of the log of self-employment wages. Note, however, that the mean of market wages (rather than the mean of the logarithm) equals 1.854 Malaysian dollar, and thus is less than the mean of self-employment wages, equal to 4.339 Malaysian dollar.

To find the difference in the means between those individuals working in the market and those who are self-employed we weighted the variables with $\bar{\pi}$ and $1 - \bar{\pi}$ respectively. Wage market workers prove to be more educated, less experienced (younger) and more concentrated in urban areas. Relatively more Malays are self-employed. Self-employed workers appear to own more land.[17)]

5. Empirical Results

5.1 First-Stage Estimates: Choice of Activity

In the first stage of the two-stage estimation procedure we analyze the choice of activities, estimating the model given by equations (3.12) and (3.13)'. Table 3 reports the results. A positive parameter indicates that the particular variable raises the likelihood of being a wage market worker. So education and experience (up to 24.1 years) increases this likelihood, suggesting that human capital accumulation tends to raise an individual's productivity more as a market worker than as a self-employed worker. Living in urban areas raises the likelihood of being a market worker. Malays and Chinese are more likely to be self-employed than Indians.

Income received from non-productive sources (V) has a negative coefficient in Table 3. It is easy to show that α is inversely related to L, as well as H_s, so the coefficient of V is evidence of a positive income effect on leisure, consistent with other research.

Individuals owning land are much more likely to be self-employed. This observation can be interpreted in two ways. First, land may affect hours of self-employment work positively, an indication that land and (own) labor are complements in (agricultural) production ($\frac{\partial H_s}{\partial Z}$ in equation (3.10)). Second, since land raises self-employment production and income, land will increase leisure through the income effect, which is positive as established above ($\frac{\partial L}{\partial Z}$ in equation (3.9)). This second interpretation ties in with a finding by Hay (1980) that owning land, as a wealth measure that is a proxy for a stream of rural incomes, deters migration from rural to urban areas.

The results are quite similar to those obtained by Huffman (1980). He found, using aggregated (by county) farm data of three U.S. states that the odds of farmers participating in the labor market increase with the husband's wage rate, his education and age (but a squared age variable has a negative sign), and decreases with the number of children younger than age 5. His "other income" variable carries a positive sign, indicating a negative income effect on leisure, contrary to our finding.

Table 4 shows how the observed value of α corresponds to the predicted value of $\bar{\alpha}$. On the diagonal of the table we find the number of correct predictions. Percentages of row and column are indicated in parentheses. So 145 self-employed individuals are predicted to be self-employed: this is 42.1 percent of those who are actually self-employed and 67.4 percent of those predicted to be self-employed. The model has 456 (= 145 + 92 + 219) correct predictions, i.e., 52.4 percent of the sample. There are 112 (= 78 + 35) "bad" predictions (12.9 percent), where individuals were predicted to be full-time self-employed while they are in reality full-time market workers, or vice versa.

5.2 Second-Stage Estimates: The Wage Equation

In addition to regional, ethnic and human capital differences in activity-specific wages, unmeasured differences are expressed in different intercepts and random variables. One can test for different intercepts only when the parameters on the other variables are the same.[19)] Random differences imply that ε_1 is not identical to ε_2 (equations (3.15) and (3.16)), so that the error of the regression equation is heteroskedastic (see (3.18), (3.21) and (3.24)). Therefore analyzing heteroskedasticity reveals random differences between self-employment and market wages.

The versatility of the specification of the wage equation leads to a large number of equations to be estimated for each estimation strategy suggested in section 3.3. To set the stage for the discussion, Table 5 contains the results of OLS regressions, in which the parameters β_m and β_s in the market and self-employment wage equations are assumed equal. The first column then shows estimates of a specification that suggests that both land and work status "have something to do" with wages (see section 2), and therefore are entered rather indiscriminately into the regression equation. The work status variable would indicate that market wages are 10 percent lower than self-employment wages. The other coefficients are quite standard, except that there is no sign of any experience profile, and that land has a significant impact.

The second column recognizes that land affects only self-employment wages. This column thus presents OLS1 estimates of equation (3.18). The coefficients are very similar to those of column 1.

The third column contains OLS2 estimates of equation (3.21). Ethnic differences in wage rates appear to increase somewhat, while the effect of education, region and land diminishes, compared to column 1. The changes are due to both the disappearance of the bias caused by the presence of α among the regressors, and the addition of $\bar{X}_c$, so no conclusion can be drawn about the direction of the bias caused by α.

Table 6 focuses on the difference between β_m and β_s. Among the many specifications that were estimated for each estimation strategy, the "best" ones are reported in Table 6. "Best" means the most parsimonious specification that cannot be rejected in favor of a more elaborate one. Column 1 shows estimates of equation (3.18), the GLS1 version. The variance of ε equals $\alpha^2 \sigma_{11} + 2\alpha(1-\alpha)\sigma_{12} + (1-\alpha)^2 \sigma_{22}$, provided that α is exogenous. So these variance terms can be estimated by maximum likelihood. Human capital variables appear to lead to differences between market and self-employment wages. The difference in the rate of return to education is over 12 percent. Market wages show an inverted-U shaped experience profile peaking at 57 years of experience (although the quadratic term is not significant), while self-employment wages indicate a weak U-shaped profile. There is a strong differential ethnic effect. Chinese and Malay self-employed workers earn 103 and 57 percent more than do Indians, while Chinese market workers receive 20 percent more than others. Regional variables have neither a differential nor an absolute effect on wages. The variance of the market wage equation (σ_{11}) is quite a bit smaller than the variance of the self-employment wage equation (σ_{22}). This result is hardly surprising in view of the risk that self-employed workers bear (see section 2). Note that these estimates supposedly suffer from the bias introduced by the endogeneity of α.

Next consider the OLS2 estimates, that use the instrumental variable $\bar{\pi}$ and the correction factor $\bar{X}_c$, in column 2 of Table 6. The difference with column 1 is rather dramatic. First, while the education parameters are qualitatively the same, the market experience profile is steeper, peaking at 24.5 year of experience, and the U-shaped profile for self-employment wages disappears. This suggests that there is some self-selection of self-employed workers according to experience, as suggested in section 1. Second, the coefficients for Malays are spread apart farther than in column 1. The coefficients for Chinese are lower. Third, the impact of urbanization is now large and significant, without showing a differential effect. Fourth, the coefficient of land is reduced to insignificance. So the conclusion from comparing columns 1 and 2 is that, while the uncorrected estimates may indicate the same differential effect qualitatively, the endogeneity of is an important source of bias in measuring the qualitative impacts.

Moving on to column 3 of Table 6, these conclusions are generally supported by the GLS2 estimates, that take into account the heteroskedasticity of the error of the regression equation. The regional variables now provide a differential impact on market and self-employment wages that is significant at the two percent level. In a quantitative sense, the GLS2 human capital parameters are similar to the OLS2 estimates. The ethnic variables indicate higher earnings in the wage labor market for Chinese and Malays compared to Indians, and lower earnings in self-employment. This differential is in the opposite direction from that found in column 2. The regional variables do not indicate a significant urban-rural market wage difference. Urban self-employment wages are significantly higher than rural self-employment wages. The coefficient of land, though negative, is insignificant again.

As to the covariance matrix Σ that can be estimated with GLS2, the correlation coefficient between u and ε_1 (ε_2) equals -1.045 (1.018), and the correlation coefficient between ε_1 and ε_2 equals -8.960. In spite of the fact that none of these coefficients fall between -1 and 1, I do not interpret this as evidence of misspecification of the model, since bounds are not imposed in the estimation procedure, and the elements of the covariance matrix (especially σ_{12}) are not estimated very precisely. Moreover, the estimates are sensitive to whether restrictions are imposed on parameters of the wage equation, and in such regressions are more well-behaved.

The empirical evidence presented in this section on the parameters of the wage equations can be summarized as follows. First, the endogeneity of α , the choice of activity, appears to bias the parameters of the wage equations. While the bias cannot be determined analytically, human capital parameters seem to be biased downward as are regional parameters; the land coefficient is biased upward; and the effect on ethnic parameters is uncertain.

Second, human capital has a differential effect on market and self-employment wages. The effect of human capital on self-employment wages is insignificant. It may be appropriate to measure experience in each activity separately. It is quite plausible, as argued in section 2, that self-employment productivity benefits from specific (rather than general) forms of human capital investment, which are less likely to be accumulated either in school or when one starts out his working life as an employee. If efficiently supervising and managing employees and successfully choosing among profitable opportunities are skills that can be acquired, learning by doing is probably the best way to acquire them.

Third, there is evidence of a differential ethnic effect. But there is a dilemma, since the OLS2 and GLS2 estimates present different pictures on the comparison of Chinese and Malay with Indian workers. In view of the attempts by the Malaysian government to improve the economic well-being of the Malay population relative to the Chinese, we are more interested in comparing Malay with Chinese wage differences, which we shall do in section 5.3.

Fourth, regional variables may have a differential effect on the two types of wages. The OLS2 estimates suggest equalization of relative prices within regions for all form of economic activity, with urban wages more than 50 prcent higher than rural wages. The GLS2 estimates suggest an active wage labor market across regions, but absense of rural-urban wage equalization in self-employment. If the rural-urban distinction is a measure for local amenities, this result indicates that, while self-employed and market workers may consume the amenities in the same manner, self-employment production benefits from urban amenities; to what degree firms bear the costs of providing such amenities is not registered in the MFLS data set. On the other hand, if the rural-urban distinction measures regional development differences, the differential in rural and urban self-employment wages may measure costs of relocation that do not exist for wage market workers. Such differential can only exist in the short run however.

Fifth, the variance of the labor market equation (σ_{11}) is smaller than the variance of the self-employment wage equation (σ_{22}). This is more evident in the GLS2 specifications not reported here than from Table 6. As indicated in section 2, this result is expected. It replicates results of other studies (e.g., Friedman (1957), Fields and Schultz (1982)).

Finally, these conclusions are not evident from the customarily estimated specifications like those reported in Table 5, columns 1 and 2.

5.3 Combined Impact of Exogenous Variables on Wages

Exogenous variables affect wages W directly through the market and self-employment wage equations ((3.15) and (3.16), combined in (3.19)), as well as through the choice of activity represented by α ((3.12) and (3.13)'). Therefore the analysis of the empirical results is not complete without examining these two effects together.

The total effect of an exogenous variable X_i on the log of wages is found by differentiating equation (3.21) with respect to X_i :

$$\frac{\partial E(\ln W|X)}{\partial X_i} = \bar{\pi}\,\beta_{mi} + (1-\bar{\pi})\,\beta'_{si} + \{(X\beta_m - X_s\beta'_s)\frac{\partial\bar{\pi}}{\partial\bar{\alpha}} + (\sigma_{1u}-\sigma_{2u})\frac{\partial\bar{X}_c}{\partial\bar{\alpha}}\}\,\delta_i \quad (5.1)$$

The first and second term of (5.1) are called the direct effect. The third term represents the choice effect. Clearly the total effect is highly nonlinear. To get an insight into the range of the effect one needs to distinguish different values of α .

Table 7 reports the marginal impact of exogenous variables on the log of wages, for "low", "high" and "mean" values of α . These values represent the first and third quartile and the mean of α , i.e., .010, 1.355, and .509, respectively.[20)] A low (high) value of α indicates an individual who is more likely to be a self-employed (market) worker. The direct effect of education increases as one becomes more likely to be a market worker. The choice effect of education is negative: as education rises, one is more likely to become a market worker, giving up a higher average self-employment wage for a lower market wage (see section 4). Thus the total effect of education is less than the direct effect, but still positive.

In a similar way one can explain the impact of other variables. Of particular interest is the behavior of Malays and Chinese, who, compared to Indians, improved their economic well-being by consciously opting for self-employment. The effect of land appears to work mainly through the choice of activity, rather than through a direct effect on wages. Thus, owning land enables one to receive the higher self-employment wages, but within the class of self-employment activities land does not seem to be a productive input. This could be due to the broad definition of self-employment used in this paper, which aggregates over widely varying "production technologies."

Equation (5.1) can be used to decompose wage differentials between Chinese and Malay workers.[21)] Table 8 indicates that Chinese workers have more education and less experience, are more urbanized, own less land, and receive more exogenous income. Each of these factors, plus the ethnic difference, generates a wage difference. Since the OLS2 and GLS2 results are roughly the same, I will discuss only the GLS2 estimates. These show that the ethnic effect is the strongest determinant of the wage differential. This ethnic effect is to a large extent choice-determined. In other words, because Chinese are more sensitive to the choice of activity, they are able to receive 16 percent higher wages (i.e., 44 percent of the predicted differential). Human capital differences account for 23 percent of the differential, urbanization for 14 percent, exogenous income for 4 percent, and land actually decreases the differential by 10 percent.

Is there an economic basis to the program of the Malaysian government, that attempts to close the wage gap between Chinese and Malay workers? On basis of our results we can say that there is some merit, but that simply measuring the wage differential grossly overstates the problem. The purely ethnic wage gap is only one fourth of the measured wage gap (i.e., .0923/.3693). Other programs that focus on educating the population and raising the economic well-being of the rural population, especially self-employed workers, will take care of the remaining 75 percent of the Chinese-Malay wage gap.

Finally, Table 7 suggests that Indians receive lower wages than Malays and Chinese. It is an interesting exercise to subject the Indian-Malay and Indian-Chinese wage gap to a similar decomposition to the one performed above. However, the sample of Indians is quite a bit smaller than those of Chinese and Malays. Since policy recommendations would be subject to more uncertainty, this decomposition is better left to a study that has access to a larger sample.

6. Summary and Concluding Remarks

This paper analyzes, estimates and compares wage equations of different activities, where the individual chooses between the alternative activities. This choice thus becomes an endogenous factor in the wage equations, and must be accounted for to obtain unbiased estimates. A single estimation equation is derived that suggests how to take care of the simultaneity induced by the choice of activity. This adjustment has its base in the index function framework. In addition, the estimation equation permits direct comparison of coefficients of the wage equations considered.

The approach is applied to a comparison of market wages and self-employment wages. The choice between these activities is dealt with in a upper-and-lower limit tobit model, yielding plausible results that are broadly consistent with previous work. The wage equation results indicate differences between market and self-employment wages due to ethnic variables and human capital variables. The latter could be due to the specific nature of the human capital used in self-employment production, so that a better measure of self-employment experience is desirable. One set of estimates (GLS2) also indicates a differential regional effect.

In the introduction three reasons were noted why wage equations may not be stable across different types of income-earning activities, dealing with (1) the character of the labor market, (2) non-pecuniary benefits of employment, and (3) kinds of returns included in the measured earnings. Our results indicate, first of all, that labor market rigidities do not dominate. Economic incentives (exogenous income and productive endowments) affect the choice of activity in the expected fashion. Our measures of human capital, imperfect as they are, are more important in determining market wages than self-employment earnings. Thus they are expected to, and found to, increase the likelihood of being a market worker. The ethnic effect on wages is to a large extent choice-determined (see both Table 7 and 8). Although the effect of regional variables is somewhat ambiguous (Table 6), notice that the positive pure wage effect of urbanization is almost balanced by the negative choice effect (Table 7), so that there is some form of wage equalizing process at work across the national labor market. All this taken together, this leads to the conclusion that labor markets tend to operate smoothly. This is evidence against the hypothesis of dualism, that associates an formal sector with wage labor, and equates an informal sector with "unproductive" self-employment. Such generalization appears invalid.

The case for the existence of non-pecuniary benefits is not as clear. One may argue that workers are free to choose an activity and that, among Malays and Chinese, self-employed workers are not as well off as market workers, in comparison to the differential among Indians. This could indicate that Malays and Chinese receive non-pecuniary benefits in being market workers. On the other hand, ethnic variables may stand for some unobserved characteristic systematically varying between ethnic groups. For example, suppose that ability, being unobserved, affects market and self-employment differently, and that the population is heterogeneous in ability. The results reported would then suggest that Malay and Chinese wage market workers are more able than their self-employed counterparts, compared to Indians.

Self-employment earnings contain returns other than labor income. Most notable is the return to risk. Returns to productive endowments, land in this study, are indirect through the effect on the choice of activity; this result may be due to the aggregative definition of the activity of self-employment. Evidence of an effect of organizational human capital (section 2) is not strong.

Finally, market wages rise more with education than do self-employment wages. One of the characteristics of economic development is the general rise in the population's educational level. This rise will induce a shift from self-employment activities toward market (wage) employment. It is interesting that this prediction holds for a cross-sectional comparison of those countries tabulated in Table 1.

The approach of this paper can readily be applied to the analysis of wages earned in other activities. For example, one could think of differences in the wage equations of blue collar versus white collar workers, between workers in the manufacturing sector versus the service sector, or between migrants and non-migrants. As indicated in section 3.2, the choice of activity in these examples would be characterized by a probit model rather than a two-limit tobit model. Similarly, one could analyze three or more alternative activities (using multinomial probit), but the number of parameters increases rapidly, roughly linearly, since for each activity one estimates a separate wage equation. An appropriate instrumental variable can be developed along the same lines as in section 3.3. One could test the equality of parameters across pairs of activities. The approach suggested in this paper is therefore quite versatile.

One important area of future research is the contribution of family workers to self-employment earnings. The allocation of time of household members other than the head is presumably a family choice along with self-employment, home production (e.g., child care) and market employment.[22] A family is a unit of observation with varying size, both cross-sectionally and over time. The family time allocation choice therefore has varying dimensions. Integrating this choice, which by itself has not yet been adequately analyzed, with self-employment earnings is a double challenge.

Appendix

Individuals have the option to divide their time between the wage labor market and self-employment. In other problems one may find that individuals could not choose such mixed strategies, e.g., migration, and, conceivably, occupational choice. This appendix illustrates that the basic ideas of this paper, i.e., that the choice of alternatives must be integrated with the wage equation, apply to such problems as well.

Let there be two alternatives. For individuals choosing the first (second) alternative, the value of the index function α^* is greater (less) than a certain bound A ; the index d takes on a value $d = 1$ $(d = 0)$; the wage equation is written as $\ln W_m = X\beta_m + \varepsilon_1$ $(\ln W_s = X\beta_s + \varepsilon_2)$. The expectation of $\ln W$, conditional on the value of d , is written as (see Heckman, 1979):

$$E(\ln W|d) = X\beta_m + \frac{\sigma_{1u}}{\sigma_{uu}^{-\frac{1}{2}}} \frac{f(\sigma_{uu}^{-\frac{1}{2}}(A - X_\alpha\delta))}{1 - F(\sigma_{uu}^{-\frac{1}{2}}(A - X_\alpha\delta))} \quad \text{if} \quad d=1$$

$$= X\beta_s - \frac{\sigma_{2u}}{\sigma_{uu}^{-\frac{1}{2}}} \frac{f(\sigma_{uu}^{-\frac{1}{2}}(A - X_\alpha\delta))}{F(\sigma_{uu}^{-\frac{1}{2}}(A - X_\alpha\delta))} \quad \text{if} \quad d=0 \qquad \text{(A.1)}$$

where σ_{1u}, σ_{2u} and $X_\alpha\delta$ are the same as defined in section 3. One can combine the two parts of (A.1) in regression analysis and impose restrictions that (parts of) the vectors β_m and β_s are the same. The error $\eta = d\varepsilon_1 + (1-d)\varepsilon_2$ of the resulting regression equation is heteroskedastic with variance $d\,s_{11} + (1-d)\,s_{22}$. From s_{11} and s_{22} one can derive consistent estimates of σ_{11} and σ_{22} (Heckman, 1979, p. 157). This estimator is referred to as GLS3.

In Table A1 results of this regression strategy are shown for the sample of full-time market workers (i.e., A=1 for d=1) and full-time self-employed workers (i.e., A=0 for d=0). This sample is a subsample of the one used for Tables 2 to 7 of this paper: individuals participating in both alternatives are omitted. Due to this loss of information one can expect that tests for differential effects will be weaker than for the OLS2 and GLS2 specifications. This is evident in the fact that the hypotheses of differential ethnic and regional effects are rejected. Thus the estimates of Table A1 are inferior to those of Table 6[†] and therefore must be considered as an illustration of the GLS3 strategy only. It is left to the reader to compare those estimates.

---

† It should be repeated that this is due to the fact that workers have the option of a mixed strategy of working in the wage market and in self-employment.

Table 1

International Comparison of Labor Force Status,

Percentage Distribution of Males†

Country	Year	Self-employed workers††	Employees	Unpaid family workers	Not classifiable
Indonesia	1976	49.65	28.86	18.55	2.94
Egypt	1978	32.00	49.73	15.59	2.68
Thailand	1978	44.94	23.39	30.74	.92
Philippines	1977	44.21	42.91	11.40	14.79
South Korea	1979	37.38	50.79	7.17	4.66
Iran	1976	34.79	50.28	6.20	8.73
Spain	1979	21.76	65.16	3.36	9.72
Italy	1979	24.59	68.14	2.38	4.89
U.K.	1977	9.33	84.20	-	6.47
Japan	1979	19.56	74.68	3.43	2.33
W. Germany	1979	11.77	87.46	7.66	-
U.S.	1979	10.09	89.13	.23	.56

† Source: Yearbook of Labour Statistics 1980, International Labour Office; those countries with a population exceeding 30 million people, for which statistics were reported in the 1980 Yearbook, are tabulated; countries are ordered according to the 1977 per capita GNP (in dollars) as reported in the World Development Report, The World Bank, 1979.

†† indicates employers and workers on own account.

Table 2
Descriptive Statistics of Variables Used[†]

Variable	Unweighted Mean	Unweighted Standard Deviation	Weighted Mean, Weight = $\bar{\pi}$	Weighted Mean, Weight = $1-\bar{\pi}$
Education	5.731	4.029	6.476	4.920
Experience	27.372	11.530	25.464	29.451
Experience^2	882.009	682.580	762.369	1012.398
Malay	.459	.499	.421	.501
Chinese	.412	.493	.404	.421
Urban Center	.209	.407	.263	.150
Other Urban	.241	.428	.327	.147
Other Income	513.137	2683.677	-	-
Land Owned	1.908	6.633	.573	3.364
α	.547	.468	-	-
$\bar{\alpha}$	.509	1.497	-	-
$\bar{\pi}$	.521	.216	-	-
$\bar{X}_c$	.201	.047	-	-
Log of Wage Rate (W)	.333	1.092	-	-
Log of Market Wage (W_m)	.226	.885	-	-
Log of S.E. Wage (W_s)	.178	1.606	-	-

[†]Number of observations is 871, except for the statistics related to $\ln W_m$ (527 observations) and $\ln W_s$ (518 observations).

Table 3
Estimates of First-Stage Regression on the Choice of Activity or α

Parameter	Estimate	Asymptotic t-statistic
Education	.0596	2.697
Experience	.0674	2.888
Experience^2	-.0014	-3.591
Malay	-.6155	-2.686
Chinese	-1.0787	-4.658
Urban Center	.8486	4.637
Other Urban	1.2459	6.787
Other Income	-.00005	-2.997
Land Owned	-.1766	-6.542
Intercept	.1599	0.353
$\sigma_{uu}^{1/2}$	1.6440	14.935

Value of log-likelihood: -807.026

Number of Observations: 871

Table 4

Comparison of Predicted and Observed Categories

Value of $\bar{\alpha}$ \ Observed α	$\alpha = 0$ Self-employed	$0 < \alpha < 1$ Mixed	$\alpha = 1$ Employee	Total
$\bar{\alpha} \leq 0$ Self-employed	145 (42.1/67.4)	35 (20.1/16.3)	35 (9.9/16.1)	215 (24.7)
$0 < \bar{\alpha} < 1$ Mixed	122 (35.5/39.0)	92 (52.9/29.4)	99 (28.0/31.6)	313 (35.9)
$\bar{\alpha} \geq 1$ Employee	77 (22.4/22.4)	47 (27.0/13.7)	219 (63.8/62.0)	343 (39.4)
Total	344 (39.4)	174 (20.0)	353 (40.5)	871

Format of cells of the table:

Number of observations

(Percentage of column/Percentage of row)

Format of fringe of the tabel:

Number of observations

(Percentage of total)

Table 5
Parameter Estimates of Wage Equation with β_m Assumed Equal to β_s
(t-statistics in parentheses)

Parameter	Including Work Status Variable α	Correcting Land By Work Status α (OLS1 Estimates)	OLS2 Estimates
Education	.0775 (6.58)	.0774 (6.57)	.0706 (6.01)
Experience	.0045 (0.33)	.0042 (0.31)	.0112 (0.83)
Experience^2	.0001 (0.26)	.0001 (0.35)	-.0001 (-0.44)
Malay	.1859 (1.64)	.2018 (1.81)	.2993 (2.65)
Chinese	.3933 (3.49)	.4162 (3.72)	.5741 (4.89)
Urban Center	.3179 (3.21)	.2868 (2.94)	.2622 (2.71)
Other Urban	.1508 (1.64)	.1212 (1.36)	-.137 (-0.15)
Land	.0186 (3.41)	-	-
Work Status α	-.0997 (-1.21)	-	-
α · Land	-	.0172 (3.17)	-
$\bar{\pi}$ · Land	-	-	.0006 (0.08)
$\sigma_{1u} - \sigma_{2u}$	-	-	-4.2508 (-4.13)
Intercept	-.6150	-.6714	.1328
SEE	907.99	913.50	893.82
R^2	.1250	.1197	.1386

Table 6
Parameter Estimates of Wage Equation,
Subject to Tests of Equality Among Groups of Parameters.
(Asymptotic t-statistics in parentheses)

Parameter	GLS1 Estimates M.W.	GLS1 Estimates S.E.	OLS2 Estimates M.W.	OLS2 Estimates S.E.	GLS2 Estimates M.W.	GLS2 Estimates S.E.
Education	.1241 (13.48)	-.0128 (-0.52)	.1428 (4.90)	.0103 (0.28)	.1470 (8.09)	.0221 (1.05)
Experience	.0342 (2.73)	-.0295 (-1.20)	.0883 (2.45)	.0091 (0.30)	.0579 (2.88)	.0217 (0.86)
$Experience^2$	-.0003 (-1.32)	.0005 (1.31)	-.0018 (-2.70)	.0000 (0.01)	-.0011 (-2.68)	-.0003 (-0.57)
Malay	-.0026 (-0.03)	.5750 (2.41)	-.4886 (-1.83)	.6915 (1.59)	.1963 (1.38)	-.2392 (-1.12)
Chinese	.2019 (2.29)	1.0332 (4.24)	-.0798 (-0.24)	.5233 (1.18)	1.0111 (3.96)	-.7298 (-2.50)
Urban Center	.0797 (1.18)		.6406 (4.26)		-.1230 (-0.54)	1.1474 (4.42)
Other Urban	.0069 (0.11)		.5664 (2.97)		-.2635 (-0.91)	1.0064 (3.78)
Land	-	.0254 (3.34)	-	.0031 (0.39)	-	-.0169 (-1.19)
Intercept	-1.0586 (-4.75)	-.3425 (-0.71)	-2.2559 (-2.71)	-0.2564 (-0.35)	-.9595 (-1.40)	1.6296 (2.74)
$\sigma_{1u} - \sigma_{2u}$	-	-	.6780 (0.44)		-	-
σ_{1u}	-	-	-	-	-2.6831 (-2.25)	
σ_{2u}	-	-	-	-	2.9945 (3.19)	
σ_{11}	.3126 (11.91)		-	-	2.4375 (1.51)	

Table 6 continued

Parameter	GLS1 Estimates M.W.	GLS1 Estimates S.E.	OLS2 Estimates M.W.	OLS2 Estimates S.E.	GLS2 Estimates M.W.	GLS2 Estimates S.E.
σ_{12}	.2985 (1.70)		-	-	-25.0237 (-1.48)	
σ_{22}	1.9559 (13.39)		-	-	3.2002 (2.65)	
Log-Likelihood	-1097.62		-1234.31		-1167.24	
R^2	-		.2347		-	
N. of Obs.	871		871		871	

Table 7
Marginal Impact of Exogenous Variables of the Log of wages

Variable		OLS2 Estimates Direct	Choice	Total	GLS2 Estimates Direct	Choice	Total
Education	low	.0522	-.0329	.0192	.0616	-.0107	.0509
	high	.1091	-.0257	.0834	.1153	-.0468	.0684
	mean	.0770	-.0344	.0427	.0850	-.0284	.0566
Experience†	low	.0030	.0051	.0081	.0052	.0017	.0069
	high	-.0053	.0040	-.0013	.0015	.0073	.0087
	mean	-.0006	.0053	.0047	.0014	.0044	.0059
Malay	low	.3185	.3402	.6588	-.1016	.1100	.0084
	high	-.1887	.2659	.0772	.0859	.4838	.5694
	mean	.0973	.3549	.4522	-.0199	.2939	.2739
Chinese	low	.3327	.5963	.9290	-.1796	.1928	.0132
	high	.0734	.4659	.5394	.5686	.8479	1.4164
	mean	.2196	.6220	.8416	.1468	.5150	.6618
Urban Center	low	.6406	-.4691	.1715	.7459	-.1517	.5942
	high	.6406	-.3666	.2740	.1999	-.6670	-.4671
	mean	.6406	-.4893	.1513	.5077	-.4051	.1026
Other Urban	low	.5664	-.6887	-.1223	.5051	-.2227	.3824
	high	.5664	-.5382	.0282	.0593	-.9793	-.9200
	mean	.5664	-.7184	-.1520	.3669	-.5948	-.2279
Other Income (* 1000)	low	-	.0276	.0276	-	.0089	.0089
	high	-	.0216	.0216	-	.0393	.0393
	mean	-	.0288	.0288	-	.0239	.0239
Land	low	.0021	.0976	.0997	-.0116	.0316	.0200
	high	.0008	.0762	.0771	-.0043	.1388	.1345
	mean	.0015	.1018	.1033	-.0084	.0843	.0759

†The results for experience take into account the non-linear effect on wages and choice of activity, evaluated at the mean value of experience (27.372)

Table 8
Decomposition of Wage Differential Between Chinese and Malays†

Variable	Difference (Chinese - Malay)	OLS2 Estimates Direct	Choice	Total	GLS2 Estimates Direct	Choice	Total
Education	1.5483	.1076	-.0568		.1206	-.0323	
Experience	-1.5523	.0238	-.0351		.0171	-.0200	
Total Human Capital Effect		.1314	-.0919	.0395	.1377	-.0522	.0855
Malay	-1.0000	-.1645	-.3787		.0447	-.2152	
Chinese	1.0000	.2540	.6636		.0476	.3772	
Total Ethnic Effect		.0895	.2849	.3744	.0923	.1620	.2543
Urban Center	.1803	.1155	-.0941		.1046	-.0535	
Other Urban	.1133	.0642	-.0868		.0498	-.0494	
Total Regional Effect		.1797	-.1809	-.0013	.1544	-.1029	.0515
Land	-.6820	.0012	-.0741	-.0729	.0064	-.0421	-.0357
Other Income	767.0607	-	.0236	.0236	-	.0134	.0134
Total Difference Predicted		.4018	-.0384	.3633	.3908	-.0215	.3693

† The observed difference in ln W equals .3759

Table A1
GLS3 Parameter Estimates of the Wage Equation
(Asymptotic t-statistics in parentheses)

Parameter	M.W.	S.E.
Education	.1239 (10.57)	.0365 (1.46)
Experience	.0396 (2.48)	.0143 (1.00)
$Experience^2$	-.0007 (-2.23)	-.0002 (-1.05)
Malay	-.1508 (-1.43)	
Chinese	.0264 (0.20)	
Urban Center	.6669 (5.53)	
Other Urban	.6417 (4.63)	
Land	-	.0013 (0.44)
Intercept	-2.1411 (-5.64)	.4301 (1.13)
σ_{1u}	1.6308 (3.78)	
σ_{2u}	1.1015 (2.92)	
σ_{11} †	1.0081	
σ_{22} †	2.1007	
log-likelihood	-937.00	
N. of Obs.	697	

† These are derived estimates according to Heckman (1979, p. 157).

Footnotes

1) Studies on dual labor markets make this type of assumption. See Sen (1966), Fei and Ranis (1964), Desai amd Mazumdar (1970)

2) See section 2 for a short survey on these issues.

3) See the surveys of Psacharopoulos (1973) and Blaug (1976). Some recent studies persisting in the habit of omitting self-employed workers are Liu and Wong (1981, on Singapore), Anderson (1980, on El Salvador), Psacharopoulos (1977, on Morocco), Knight and Sabot (1981, on Tanzania).

4) Huffman (1977) provides evidence that more educated farmers were quicker in applying nitrogen fertilizer in the production of corn when in the early 1960's usage of nitrogen fertilizer increased sharply.

5) Section 3.3 will consider this bias into more detail.

6) The assumption if perfect substitutability was recently tested by Deolalikar and Vijverberg (1982) with district-level agricultural data from India. They estimated an elasticity of substitution equal to 2.5, rejecting an infinite elasticity as implied by perfect substitutability. An often-cited test by Bardhan (1973) accepting perfect substitutability was found to be ill-constructed.

7) The budget equation does not contain W_s directly. W_s can now be defined as $W_s=(P_qQ-P_rR-W_hH_h^*)/H_s$. Thus W_s is the average, rather than marginal, value of time H_s.

8) Huffman refers to earlier work of Lee (1965). This reference was not available to me.

9) This is true if there is an infinitesimal small but non-zero cost of hiring labor (to the employer) or job search (to the employee). If such costs are zero, only net labor hired, defined as total labor hired minus own time sold in the market, is determined.

10) E.g. if W_m exceeds W_hs, it would be optimal for an individual to be a full-time market worker himself, while hiring others to take care of his business.

11) Another corner solution is L =T, or both H_m =0 and H_s = 0. We do not analyze this in this paper, since observations for whom L equals T do not provide information on wage rates.

12) Using Kuhn-Tucker conditions that take the corner constraints into account one can derive essentially the same framework.

13) For a more extensive description of models where the dependent variable has a lower and upper limit as well as an intermediate range, see Rosett and Nelson (1975).

14) The variance of u, σ_{uu}, is an identified parameter in the wage equation estimated, and thus can be retrieved from the second stage of the estimation procedure proposed in this paper.

15) A clear illustration of this statement is found in Goldfeld and Quandt (1976, pp. 40-41), where the error term of the decision rule that divides the sample (i.e. their equation (2.16)) is uncorrelated with the random variables of the regression equations under either regime.

16) Self-employment wages take account of the individual's consumption of own production. For farmers crop prices were not always reported. So to get the value of own consumption we assigned the average price for the particular crop of that farm.

17) The amount of land is an observable variable even for those individuals who are full-time employees and do not engage in self-employment. Some of such individuals were actually found to own some land.

18) Complementarity between labor H_s and land X_3, i.e. $F_{14} > 0$, is not a sufficient condition for $\frac{\partial H_s}{\partial X_3}$ to be positive.

19) For example, suppose the ethnic coefficients are different. Since we distinguish Malays, Chinese and Indians as well as market and self-employed workers, there are six cells, each of which has a different intercept due to thnic reasons.(as assumed) In addition to this there are unmeasured differences between market and self-employed workers, they cannot be distinguish from the ethnic differences.

20) The marginal impact also depends on the value of $X\beta_m - X_s\beta_s')$. Since X and X_s are a subvector of X_α, and $\bar{\alpha} = X_\alpha\bar{\delta}$, if one varies $\bar{\alpha}$, one should consider changes in $(X\beta_m - X_s\beta_s')$ as well. The following values were associated with the low, high and mean values of $\bar{\alpha}$: -1.626, -1.277 and -1.497 (OLS2), and -1.142, -1.376 and -1.229 (GLS2).

21) The average value of $\bar{\alpha}$ for Chinese and Malay workers equals .361. The average value of $(X\beta_m - X_s\beta_s')$ equals -1.629 (OLS2) and -1.090 (GLS2).

22) Chiswick (1977b) assumed that each unpaid family worker contributes a constant fraction of the contribution of a paid worker to self-employment earnings, and took the number of workers per family as exogenous. Along the same line of argument as make in section 3 about the endogeneity α, these assumptions may be an important source of simultaneity bias.

References

Anderson, Lascelles, "Rates of Return to Human Capital: A Test Using El Salvador Data," American Economic Review, May 1980, 70, 2, pp. 138-141.

Bardhan, P.K., "Size, Productivity and Returns to Scale: An Analysis of Farm-level Data in Indian Agriculture," Journal of Political Economy, November 1973, 81, 6, pp. 1370-1380.

Barnum, Howard, and Lyn Squire, A Model of an Agricultural Household: Theory and Evidence (Johns Hopkins Press, Baltimore, 1979)

Becker, Gary S., Human Capital, 2nd ed. (National Bureau of Economic Research, Columbia University Press, 1975)

Ben-Porath, Yoram, "The Production of Human Capital and the Life Cycle of Earnings," Journal of Political Economy, August 1967, 75, 4, pp. 352-365.

Blaug, Mark, "An Economic Analysis of Personal Earnings in Thailand," Economic Development and Cultural Change, October 1974, 23, 1, pp. 1-32.

-------, "The Empirical Status of Human Capital Theory: A Slightly Jaundiced Survey," Journal of Economic Literature, September 1976, 14, 3, pp. 827-855.

Chiswick, Carmel U., "On Estimating Earnings Functions for LDC's," Journal of Development Economics, March 1977, 4, 1, pp. 67-78 (a).

-------, "A Procedure for Estimating Earnings of Unpaid Family Workers," Business and Economic Statistics Section, Proceedings of the American Statistical Association, 1977 (b).

Chow, Gregory C., "Tests of Equality between Sets of Coefficients in Two Linear Regressions," Econometrica, July 1960, 28, 3, pp. 591-605.

Deolalikar, Anil B., and Wim P.M. Vijverberg, "The Heterogeneity of Family and Hired Labor in Agricultural Production: A Test Using District Level Data from India," Center Discussion Paper 411, Yale University, 1982.

Fei, John C.H., and Gus Ranus, Development of the Labor Surplus Economy: Theory and Policy (Irwin, Homewood, Ill., 1964)

Fields, Gary S., and T. Paul Schultz, "Income-Generating Functions in a Low Income Country: Columbia," The Review of Income and Wealth, March 1982, 28, 1, pp. 71-87

Friedman, Milton, A Theory of the Consumption Function (NBER, Princeton University Press, 1957)

Goldfeld, S.M., and R.E. Quandt, Studies in Nonlinear Estimation (Ballinger Publishing Co., Cambridge Mass., 1976)

Hay, Michael J., "A Structural Equations Model of Migration in Tunesia," Economic Development and Cultural Change, January 1980, 28, 2, pp. 345-358

Heckman, James J., "Sample Bias as a Specification Error," Econometrica, January 1979, 47, 1, pp.153-162

------- and Tomas E. Macurdy, "New Methods for Estimating Labor Supply Functions: A Survey," in Research in Labor Economics, Vol 4 (JAI Press Inc., 1981), pp. 65-102

Huffman, Wallace E., "Allocative Efficiency: The Role of Human Capital," Quarterly Journal of Economics, February 1977, 41, 1, pp.59-80

-------, "Farm and Off-farm Work Decisions: The Role of Human Capital," Review and Economics and Statistics, February 1980, 62, 1, pp. 14-23.

Jones, Robert, and Nyle Spoelstra, The Malaysian Family Life Survey: Appendix C, Field and Technical Report (Rand Corporation, Santa Monica, 1978)

Kihlstrom, Richard, and Jean-Jacques Laffont, "A General Equilibrium Entrepreneurial Theory of Firm Formation Based on Risk Aversion," Journal of Political Economy, August 1979, 87, 4, pp. 719-748.

Knight, Frank H., Risk Uncertainty and Profit (Houghton Mifflin Co., New York, 1921)

Knight, J.B., and R.H. Sabot, "The Returns to Education: Increasing with Experience or Decreasing with Expansion?" Oxford Bulletin of Economics and Statistics, February 1981, 43, 1, pp. 51-71.

Kuznets, Simon, Modern Economic Growth, Rate, Structure and Spread (Yale University Press, 1966)

Lee, John E., "Allocating Farm Resources between Farm and Non-Farm Uses," Journal of Farm Economics, February 1965, 47, pp. 83-92.

Liu, Pak-Wei, and Yue-Chim Wong, "Human Capital and Inequality in Singapore," Economic Development and Cultural Change, January 1981, 29, 2, pp. 275-293.

Mazumdar, Dipah, and Desai, Meghnad, "A Test of the Hypothesis of Disguised Unemployment," Economica, February 1970, 37, 145, pp. 39-53

Mincer, Jacob, "The Distribution of Labor Incomes: A Survey with Special Reference to the Human Capital Approach," Journal of Economic Literature, March 1970, 8, 1, pp. 1-26.

-------, Schooling, Experience and Earnings (National Bureau of Economic Research, New York, 1974)

-------, "Progress in Human Capital Analysis of the Distribution of Earnings," in A.B. Atkinson, The Personal Distribution of Incomes (Westview Press, Inc., Boulder, Col., 1976), pp. 136-176.

Prescott, Edward C., and Michael Visscher, "Organization Capital," Journal of Political Economy, June 1980, 88, 3, pp. 446-461.

Psacharopoulos, George, with Keith Hinchliffe, Returns to Education, An International Comparison (Elsevier, Amsterdam, 1973)

-------, "Schooling, Experience and Earnings," Journal of Development Economics, March 1977, 4, 1, pp. 39-48.

Rosen, Sherwin, "Human Capital: A Survey of Empirical Research," in Ronald G. Ehrenberg, ed., Research in Labor Economics, vol. 1 (JAI Press, Greenwich, Conn., 1977)

Rosenzweig, Mark R., "Neoclassical Theory and the Optimizing Peasant: An Econometric Analysis of Market Family Labor Supply in a Developing Country," Quarterly Journal of Economics, February 1980, 94, 1, pp. 31-55.

Rosett, Richard N., and Forrest D. Nelson, "Estimation of the Two-Limit Probit Regression Model," Econometrica, January 1975, 43, 1, pp. 141-146.

Schultz, Theodore W., "Investment in Entrepreneurial Ability," Scandinavian Journal of Economics, 1980, pp. 437-448.

Sen, A.K., "Peasants and Dualism With or Without Surplus Labor," Journal of Political Economy, October 1966, 74, pp. 425-450