1982

Sample Selection Bias and the Nature of Unemployment

Joshua David Angrist
Oberlin College

SAMPLE SELECTION BIAS AND THE NATURE OF UNEMPLOYMENT

by

Joshua D. Angrist

April 1982 Oberlin College

A. Introduction (1)

The most disturbing and difficult empirical problems of labor economics revolve around the absence of crucial information; the wage an unemployed person would receive if he or she were working. The most controversial policy problem of labor economics is embodied in the question; when is unemployment a problem? The goal of this paper is to propose a methodology for studying the first problem that sheds some light on the second.

Most economists agree that some level of unemployment is a necessary and socially efficent characteristic of the labor market. This necessary level of unemployment is often thought to be around 1 to 2% and an important part of the process of efficient allocation of human resources. What about 5% unemployment? A case can be made that at 5% unemployment a portion of the unemployed have had their incentives to work eroded by transfer payments and the progressive tax structure. Some might say that such a level of unemployment is to be expected when the government interferes in the labor market by raising the opportunity cost of working. And 10% unemplyment? Surely some of these people are "involuntarily" unemployed. They are not unemployed because they are looking for better jobs or doing so well on unemployment compensation or some other sort of transfer income that it is not worth their while to take a job. Rather, they have looked for work, are willing to work at

---

(1) This paper would have been impossible without the assistance and patience of the Oberlin College Economics Department and the staff of the Oberlin College Computing Center.

prevailing wages and are capable of productive activity. How does one find out who these people are or prove to the sceptic that they exist at all? Sociologists might ask them but in economics we are all behaviorists.

Something we do have an idea about is the type of information relevant ot the individual's decision to work or not to work. In fact, we can precisely formulate some theoretical decision rules and then ask how well our theoretical decision rules explain the behavior that we actually observe in the labor market. Any economic characterization of an individual's decision making process takes into account the costs and benefits of an action taken, where these costs and benefits are evaluated at the margin. The relevant marginal benefit in the case of the work decision is clearly the hourly wage. Of course, this may not be a single known number. Workers may have expectations about the disrtibution of wage rates over time as in the literature on job search or intertemporal substitution. Another interesting modification to the simple notion of a single hourly wage is the implicit contracts model wherein workers "sell" job security in return for a higher hourly wage. Lucas and Rapping laid the foundations for the search theory model in 1970 and interested readers may find a recent examination of it in Ashenfelter and Altonji (1980). Azariades (1975) provides an example of an implicit contracts model. Although these approaches are not incorporated here it is not because they are thought to be without merit. In fact they are very appealing in that they treat information as any other good; one which economic agents behave

purposefully and rationally in collection and usage (McCullum, 1980 p.717). In our model of the labor market and wage determination it is the asking wage which represents the counterpart of the offered wage; the offered wage is the marginal benefit from working another hour while the asking wage is the marginal cost, the opportunity cost of working as measured by the value of leisure time, loss of transfer income, other work opportunities and consumption pressure. The opportunity cost of working is analagous to Gronau's (1973) "value of time" and will be used to develop an econometric specification for what will be called the participation wage in the rest of the analysis presented here.

In this paper we would like to ask what sorts of decision rules will best explain the observed distribution of wage rates paid. If we understand something about the way the observed wage distribution is drawn from the population of wage rates; i.e., the wages paid to those who work and the wages the unemployed could expect if they were working, then we may be in a position to make some inferences about the nature of unemployment. In particular we ask: is unemployment the result of utility maximizing decision makers who, in equilibrium, have chosen not to work because the costs (the participation wage) exceed the benefits (the wage they could expect from working) ? If in fact this is the true nature of unemployment then clearly the observed wage distribution cannot be an unbiased sample of the true wage distribution because the people for whom we observe a wage are "special" in the sense that they have an equilibrium position in

the labor market which is a consequence of their particular expected return to work and opportunity cost of working. The bias in the observed wage distribution is a by-product of the nature of the equilibrium which generated that distribution. Of course it is also possible that the unemployed (and correspondingly the employed) may not be in utility maximizing equilibrium, yet they may still be special in some sense, that is, the observed wage distribution may still suffer from sample selection. In the econometric work to follow in this paper it will be assumed that if the unemployed are not "voluntarily" out of work in the sense that they are in the equilibrium discussed above then they are a random subsample of the population in the labor force. Naturally this is a highly unrealistic assumption and even within narrowly defined demographic groups the observed wage data will still have some sample selection bias, i.e., not reflect the true wage distribution for that demographic group, even if all the unemployed within that group are not in equilibrium. However, if the unemployed are not in equilibrium in the sense that they would be better off by working (where better off means that the wage they could expect to receive exceeds the wage at which they would be willing to offer positive hours) then the sample selection rule which determines the observed wage distribution cannot be the one that says: " the people whose wage we don't observe are those whose participation wage is in excess of their expected wage ". Therefore, to test for the presence of involuntary unemployment we will compare the predictive ability of the model of equilibrium voluntary unemployment with the

simplest alternative; unemployment occurs with a constant unconditional probability p, the best estimator of which is the relative frequency of unemployment. This sort of comparison is conceptually similar to the standard F-test in multivariate regression analysis where a detailed specification is compared with the naive estimator, the mean of the dependent variable, to determine the legitimacy of the detailed specification. Unfortunately the statistics of the process are not at all comparable to the F-test procedure as will be shown in later sections of the paper. The natural choice for the model wherein the unemployed are in utility maximizing equilibrium is Nelson's (1974) Censored Regression Model, designed specifically to overcome the problem of sample selection bias that arises when observed data is not an unbiased sample of the population for the reasons discussed here (2).

The Censored Regression Model proposed by Nelson is shown below:

$$We(i) = B1X1 + u(i)$$

$$Wp(i) = B2X2 + v(i)$$

$$W(i) = We(i)$$

$$\text{iff } We(i) > Wp(i)$$

$$W(i) = 0 \text{ otherwise.}$$

where

---

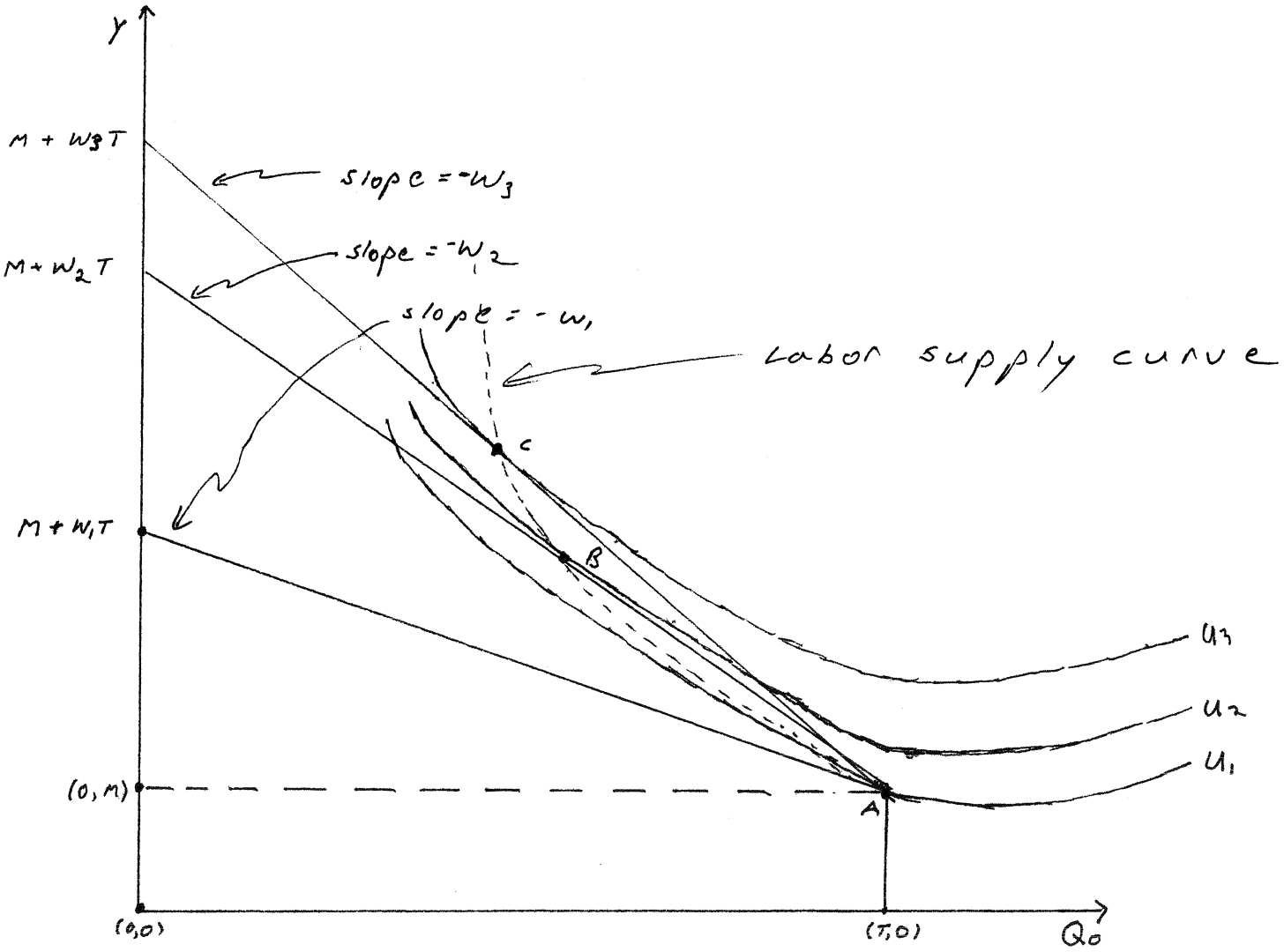
(2)This model can be seen to be similar to another model of truncation, namely the Tobit model (Tobin, 1958). The main difference between this model and the Tobit model is that Tobit takes the point of truncation as known and fixed whereas in the Censored Regression Model the point of truncation is a stochastic variable to be estimated.

X1 and X2 are vectors of personal characteristics, We(i) is the iTH agent's expected wage, i.e., the wage he or she would receive if working, Wp(i) is the iTH agent's participation wage, i.e., the wage below which he or she will offer zero hours, u(i) and v(i) are classical error terms, W(i) is the wage observed for agent i and B1 and B2 are vectors of parameters to be estimated. Gronau (1974) was the first to point out that if this is the way observed wages, W(i), are generated then Ordinary Least Squares estimation of the first equation above,

$$W(i) = We(i) = B1X1 + u(i)$$

will be biased. Heckman (1976 and 1979) discusses this sample selection bias as a specification error in the expected wage equation and shows (1979) in particular that the specification error in this case is the failure to include the mean of the error term, conditional on the sample selection rule. Following Nelson we will estimate the CRM as a system by the method of maximum likelihood since ordinary least squares is clearly inappropriate for either of the wage equations shown. The important thing to notice about this model is that, once We and Wp are properly and precisely defined, as they hopefully will be in the next section of this paper, then the CRM is the most basic and obvious specification for the generation of observed wage data in the neo-classical theory of labor supply. Consider diagram 1 on the following page wherein the partcicpation decision is illustrated for the standard case of a linear budget constraint in the income-leisure plane. Here Qo is leisure, Y is income and T is the total time available for allocation between

DIAGRAM 1



W_1 = the participation wage, w/tangency at point A. The dotted line is the labor supply curve found by increasing the wage rate.

Q_0 = Leisure
T = time
M = non-Labor income
y = Total income

income and leisure. Then, if M is the amount of non-labor and transfer income, we see the budget constraint to be:

$$M + W(T-Qo) = Y.$$

At wage rate W1 we find indifference curve U1 tangent at point A, where zero hours of work are offered and T hours of leisure are consumed. Since the marginal rate of substitution equals the wage rate W1 at such a tangency then W1 must be the value of time or the participation wage. Thus, a wage greater than W1 is necessary to induce participation. Additional points of equilibrium are seen at points B and C, each corresponding to a higher wage rate and generating points on the labor supply curve where positive hours are offered since W3 > W2 > W1. Even if non-labor income M is equal to zero the participation wage is simply zero and we may continue to think of the budget constraint above as a fairly general representation of a linear budget constraint in the neo-classical model where markets clear and there is no involuntary unemployment. Clearly it is a more difficult task to choose a model for the generation of the observed wage distribution under the hypothesis of involuntary unemployment. Certainly involuntary unemployment carries with it the notion of some additional constraint other than the budget constraint and any complete specification must describe this constraint. In this paper we will avoid the issue of what the constraint might be in our model of involuntary unemployment. We have chosen an alternative that precludes "voluntary" unemployment and that is about all we can say about the constant probability model of unemployment, with one important exception.

Ordinary Least Squares is an appropriate estimator for the expected wage equation in our model of involuntary unemployment since we have assumed that there is no sample selection bias. Thus our only wage equation in the constant probability model is the first wage equation of the CRM above:

$$W(i) = We(i) = B1X1 + u(i)$$

iff i is employed.

$$W(i) = 0 \text{ otherwise.}$$

The iTH agent is unemployed with constant probability p, and therefore this equation may be used to provide unbiased estimates of the parameters B1. We will call this constant probability model the OLS model; by OLS model we will mean the probability p of unemployment, the above wage equation and a likelihood equation to be derived later for predicting employment status for a particular sample. Likewise, by the CRM model we will mean the system of equations previously introduced, the corresponding implicit selection rule determining employment status and probability of unemployment for a given agent and a corresponding likelihood equation, also to be derived later, for predicting employment status for a particular sample.

To apply the notions of how one might describe the nature of unemployment based on information about the observed wage distribution both the CRM and OLS models will be estimated for the two racial groups, black and white. "A Priori" we would expect the CRM model to do a better job of predicting employment status than the OLS model for whites and we would expect the OLS constant probability model to be better than the CRM for blacks.

The results presented in the final section of this paper tend to support that hypothesis. Unfortunately, the statistics involved in a comparison such as this are not always well defined and certainly are not fully developed in this paper. Only one sample was used for estimation and clearly multiple sampling is in order if one wishes to use this procedure to make any strong statements about the nature of unemployment. Consequently the work presented here is offered as a methodological approach to the study of unemployment. Even readers sympathetic to the conclusions of the particular estimation to follow should not interpret the results as a conclusive statement about the nature of unemployment for the particular sample used in estimation.

Having given the necessary caveats with regard to interpretation we turn in the next section to the origins of the participation and expected wages in the neoclassical theory of labor supply. Section C discusses some of the econometrics involved in estimating the OLS and CRM models using the functional forms suggested by section B. In section D the results of the various estimations and test procedures are presented and evaluated in light of the original goals and propositions. Section E suggests some conclusions about the procedure and some directions for future work. Also included are a number of appendices containing some earlier empirical work on the OLS wage equation and discussions of technical issues in estimation.

B. A MODEL OF LABOR SUPPLY

In order to precisely identify the participation and expected wages it is necessary to develop a model of labor supply. In general there are three ways of doing this. Only one will be used here. For an excellent summary of these three general approaches see Abbot and Ashenfelter (1976). Our approach starts with a direct utility function in income and leisure and then uses the first-order conditions for constrained maximization to derive labor supply. This approach was chosen for what are largely intuitive reasons. It seems to make sense that if one is going to discuss voluntary and involuntary behavior one ought to begin with a description of preferences.

In this case our description of preferences is taken from Stone's (1954) Linear Expenditure System. It is the only expenditure system whose corresponding utility function satisfies the theoretical restrictions of adding-up, homogeneity and symmetry (Deaton and Muellbauer, 1980, p.65). It is also the only linear expenditure system that can be derived from a classical utility function (Goldberger, 1967). The general form of the LES direct utility function is,

$$U = V(Q) = \prod_i (Qi-Hi)^{Bi}, \quad \sum_i Bi = 1$$

where

$Q = (Qo, Q1, \ldots, Qn)$ is the agent's consumption vector

$H = (Ho, H1, \ldots, Hn)$ and $B = (Bo, B1, \ldots, Bn)$ are parameter vectors.

This function is usually transformed using natural logarithms into the log-linear form

$$U = LnV(Q) = \sum_i BiLn(Qi-Hi).$$

When one of the commodities, Qi is defined to be leisure, say Qo, then the above function is referred to as an augmented (for leisure) Stone-Geary Utility Function (Goldberger, 1967). If all non-leisure commodities are aggregated into one, called income, then the function is usually written

$$U = V(Q) = B_o \text{Ln}(Q_o - H_o) + B\text{Ln}(Q - H)$$

where Q is income and H is a single valued parameter.
This utility function has a number of interesting properties that are consistent with some intuitive notions about how preferences might be mapped into a cardinal ordering. The first thing to notice is that the logarithmic terms give the function diminishing marginal utility. In addition it is not defined where $Q_o < H_o$ or $Q < H$ because the arguments of the log terms would then be negative. The quantities Ho and H can therefore be interpreted as "subsistence" quantities of leisure and income respectively, and they serve as additional parameters which make the functional form more flexible. Consider for example the parameter H as a function of personal characteristics,

$$H(i) = C_o + CZ(i) + e(i)$$

where Z is a relevant set of i's characteristics

C_o, C are parameters and

$e(i)$ is a classical error term $\sim N(0, \sigma^2)$.

Following Gronau (1973) this approach will be incorporated into our estimation of the participation wage.

In particular our maximization procedure parallels that of Leuthold (1968), expressing the budget constraint in terms of income and time allocation identities and therefore bypassing issues of commodity demand. For this to be a valid procedure it is necessary to assume that relative prices are constant across the consumption bundles represented in our sample.
Let the utility function be[3]

$$U = V(Q_o, Q) = A\text{Ln}(Q_o - H_o) + B\text{Ln}(Q - H)$$

where

$Q_o = T - L$ T=total time available, L=labor supply

---

[3]The subscript i will be dropped for clarity of notation.

$Q=WL+M$ $M=$ non-labor income, $W=$the actual wage

$A+B=1$, Ho, H are parameters[4].

Substituting the constraints directly into the utility function gives

$$U= A\text{Ln}[(T-L)-Ho] + B\text{Ln}[(WL+M)-H]$$

with the corresponding first order condition for maximization

$$\partial U/\partial L = -A/(T-L-Ho) + (BW)/(WL+M-H) = 0.$$

Re-arranging and solving for L gives the labor supply function

$$L= B(T-Ho) - (A/W)(M-H).$$

Labor supply is seen to reach a maximum at $B(T-Ho)$, i.e., B times the time remaining after subtracting 'committed' leisure time. It varies up to this point as a function of W and M, increasing in W and decreasing in M. The participation wage is found by setting L to zero:

$$B(T-Ho)-(A/W)(M-H)=0$$

and solving for the participation wage Wp, giving

$$Wp= [A(M-H)]/[B(T-Ho)].$$

Thus our true labor supply function is now discontinuous at Wp, the participation wage, and is given by

$$L= B(T-Ho)-(A/W)(M-H) \quad \text{for } W>Wp$$

$$L= 0 \text{ otherwise.}$$

Now if one is not particularly interested in identifying the constants A,B and T-Ho then it is possible to linearize the functional form of the participation wage

$$Wp= [A(M-H)]/[B(T-Ho)]$$

$$=\{A/[B(T-Ho)]\}(M-H)$$

where estimation will only allow identification of the coefficient $A/[B(T-Ho)]$ as one number, say K.

---

[4]The restriction $A+B=1$ is not particularly important, only serving to make the indifference curves associated with this utility function rectangular hyperbolas, assymptotic to the two positive asymptotes, Ho and H. The lower A is relative to B the flatter the indifference curves will be; reducing the marginal rate of substitution.

Allowing H to be a linear function of personal characteristics with a classical error term

$$H = Co + CZ + e$$

then

$$Wp = K(M - Co - CZ - e)$$
$$= -KCo + KM - KCZ - Ke$$
$$= Eo + KM + EZ + u$$

where if

$e \sim N(0, \sigma^2)$ is a classical error term then

$u \sim N(0, K^2\sigma^2)$ is also a classical error term.

Thus by sacrificing identification of some parameters we have a linear form for the participation wage that is a function of non-labor income and personal characteristics.

Having derived a simple linear form for the participation wage the next task is to justify the proposed characterization of the expected wage. The procedure here follows closely that of Hall (1973) although the interpretation is somewhat different. Hall applies the concept of an expected wage to estimation of labor supply as a solution to the missing variables problem and a correction for measurement error, but does not justify his imputation of fitted values to the unemployed members of his sample. Justification of this procedure requires some rather strong assumptions about the nature of unemployment that will be explored here. For an econometric justification of the expected wage as an instrumental variables estimator of the actual wage for the employed allowing unbiased estimation of their labor supply see Appendix I where Hall's proof is duplicated. Recall the labor supply equation

$$Ls = B(T-Ho) - (A/W)(M-H).$$

This equation must be part of a two equation system describing supply and demand in the labor market,

$$Ls = Ls(W,R)$$

$$Ld = Ld(W,S)$$

$$L=Ls=Ld$$

where R and S are vectors of exogenous variables that may overlap and W is the actual wage.

Then

$$Ls(W,R)=Ld(W,S)=L$$

may be solved for the actual wage, W. By fitting the resulting equation over the members of the sample who have an observed non-zero wage we will obtain estimates of the expected wage for the employed[5]. From this we can see that our expected wage equation may be thought of as a reduced form from a labor supply and demand system.

Having more fully characterized the expected and participation wages the next task is to discuss procedures for their estimation in the context of the OLS and CRM models.

---

[5]Actually, the appearance of the wage as a reciprocal in the labor supply equation makes solving for the expected wage a little less casual than described above. This non-linearity in the labor supply equation prohibits a linear form for the expected wage as a reduced form of the simultaneous system. Consequently, when choosing a linear (in the parameters) form for the expected wage one may prefer to think of the expected wage as a general instrumental variables estimator for the system. This does not imply that Hall's proof of unbiasedness in application of this estimator to the expected wage for the employed justifies imputation of an expected wage to the unemployed by retaining the coefficients from this procedure.

C. ESTIMATING THE WAGE EQUATIONS

Estimating the wage equation for the OLS model is straightforward since the assumption is that the expected wage is similarly determined for both employed and unemployed. Therefore it is sufficient to fit the observed wage for workers to a linear combination of observed characteristics, then take the coefficients estimated and use them to find a 'fitted' wage for non-workers. In this case a log-linear functional form was chosen with the log of the wage on the left hand side. This is the form predominant in the literature, fitting percent changes instead of levels. Two versions of this wage equation were estimated. The first is a more detailed specification with eight regressors. Results of this fully specified wage equation are reported in Appendix II. The equation used in the body of the work was estimated on a subset of the regressors from the fully specifed equation that were chosen according to their significance in the fully specified model. It was necessary to trim down the wage equation for the OLS so that the OLS expected wage equation would be comparable to the CRM expected wage equation. Cost in terms of computer resources prohibited using the full specification in the CRM model. Space and time requirements were found to rise rapidly with the number of parameters to be estimated in the CRM. In fact the model estimated was the largest possible with the technique used on the Oberlin College Xerox Sigma-9 operating system[6].

The OLS model also has a more interesting interpretation than that of a simple linear regression. It can be expressed in terms of a probabilistic model of labor force participation with a corresponding likelihood function. Before this interpretation can be given it is helpful to derive the likelihood function for the CRM and detail the assumptions under which it is

---

[6]Other methods of estimation may have proven to be more efficient. For a more detailed discussion of computer techniques, software available and hardware resource constraints with regard to the mechanics of estimating the CRM likelihood function see Appendix III.

estimated.

Recall the CRM model of wage determination

$$We = B1X1 + u$$

$$Wp = B2X2 + v$$

$$W = We \text{ iff } We > Wp$$

$$= 0 \text{ otherwise.}$$

Ordinary least squares is clearly not the appropriate estimating technique even for the expected wage equation. Because of the sample selection bias which is assumed to exist in this model, u will be correlated with X1 when We is close to Wp in the non-censored sample. When We, which is a function of X1, is close to Wp, more of the observed wage distribution is truncated than when it is farther away. Consequently the level of truncation, which affects the error term, will be a function of X1. Heckman (1976) points out that in general one cannot sign the direction of the bias introduced by this correlation. Goldberger (1975) has shown that under certain conditions one can identify the sign of the bias if one also has 'a priori' information about signs of the coefficients in the model's structural equations. Ordinary least squares is impossible for the participation wage equation for the simple reason that there are no observations on the dependent variable. However we may use the inequality determining participation to derive a likelihood function incorporating all the information available. First divide the sample into two sub-samples, S1 containing the unemployed and S2 containing the employed. Then in sample S1 we would like to maximize the probability for each i that $W(i)=0$. In S2 we would like to maximize the probability for each i that $W(i)>Wp(i) <=> W(i)>0$ and $W(i)=We(i)=B1X1(i)+u(i)$. In sample S1,

$$Pr(W=0) = Pr(We<Wp) = Pr(B1X1+u<B2X2+v)$$

$$= Pr(u-v<B2X2-B1X1).$$

Now if

$u \sim N(0, \sigma_1^2)$ and $v \sim N(0, \sigma_2^2)$

then

$u - v \sim N(0, \sigma^2)$ where $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2COV(u,v)$.

Thus the probability that W(i) equals zero is given by

$$Pr(W=0) = Pr(u-v<B2X2-B1X1) = F\left(\frac{B2X2-B1X1}{\sigma}\right)$$

where F is the N(0,1) cumulative distribution function.

In the set S2 we wish to take advantage of the information that not only is W(i) non-zero but it is observed and equal to We(i).
Let $G(u,v) = G(We-B1X1, Wp-B2X2)$ be the bivariate normal density of u and v.
We are interested in

$Pr(W=We>Wp$ AND $W=We=B1X1+u)$ for i element of S2

$= Pr(v<W-B2X2$ AND $u=W-B1X1)$.

This is given by

$$\int^{W-B2X2} G(W-B1X1, v)\, dv.$$

The likelihood function for the whole sample is found by combining the likelihoods for each sub-sample giving

$$L(B1, B2, \sigma_1, \sigma_2, COV(u,v) \mid X1, X2, W) = \prod_{S1} F\left(\frac{B2X2-B1X1}{\sigma}\right) \prod_{S2} \int^{W-B2X2} G(W-B1X1, v)\, dv .$$

This is not an easy function to evaluate. One simplifying assumption is that of zero covariance. In the case of zero covariance $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ and the likelihood function of the second sub-sample factors because the assumption of zero covariance means that u and v are independent.
Then,

$$Pr(W>Wp \text{ AND } W=B1X1+u) = Pr(W>Wp)\, Pr(W=B1X1+u)$$

$$= \left[1/\sigma_1\right] F'\left(\frac{W-B1X1}{\sigma_1}\right) F\left(\frac{W-B2X2}{\sigma_2}\right)$$

where F' is the density function corresponding to the distribution function F. Thus our sample likelihood becomes

$$L(B1, B2, \sigma_1, \sigma_2 \mid X1, X2, W) =$$

$$= \int_{S1} \int F\left(\frac{B2X2 - B1X1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \int_{S2} \int F\left(\frac{W - B2X2}{\sigma_2}\right) F'\left(\frac{W - B1X1}{\sigma_1}\right) (1/\sigma_1).$$

This simpler likelihood function was the one employed in estimation of B1, B2, σ_1 and σ_2 consequently some justification of the assumption of zero covariance is in order. If the model is correctly specified there is no obvious reason why the errors of the participation and expected wage equations should be correlated. For example, if u(i) is positive and we have explained all systematic variance in We(i) then i is simply a 'lucky' individual. There is no reason to think that luck in the labor market should cause us to systematically under or overestimate a lucky individual's participation wage. A similar informal argument can be made for the case of a negative u(i). Likewise we can proceed from errors in the Wp equation and argue that a given error in the Wp equation would not allow one to predict the error in the We equation. In addition, the assumption of zero covariance allows us to choose overlapping sets of regressors for the participation and expected wage equations without sacrificing identification of the parameters B2 (Nelson, 1974, p.19).

To see how the OLS model may be characterized as a probabilistic model of employment it is helpful to first consider the CRM model when we discard observations on W(i) and simply retain knowledge of employment status. Then our model says, where Y(i) is i's employment status;

$$Y(i) = \begin{cases} 1 \text{ with probability } P(X) & \text{unemployed} \\ 0 \text{ with probability } 1-P(X) & \text{employed} \end{cases}$$

where P(X) is given as before, $P(X) = Pr(We < Wp) = Pr(W = 0) =$

$$F\left(\frac{B2X2 - B1X1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right).$$

The likelihood function for this simplified version of the CRM, continuing to assume COV(u,v)=0 is then[7]

$$L(B1, B2, \sigma_1, \sigma_2 | X1, X2) = \prod_{S1} F\left(\frac{B2X2 - B1X1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \prod_{S2} \left[1 - F\left(\frac{B2X2 - B1X1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)\right] .$$

Now, picturing the OLS model in a similar manner we see that the model can be described in terms of employment status as

$$Y(i) = \begin{cases} 1 \text{ with probability } P \text{ unemployed} \\ 0 \text{ with probability } 1-P \text{ employed} \end{cases}$$

where P is constant for all i. Then employment status in the OLS model is a simple Bernoulli trial with fixed probability P, the best estimator of which is the relative frequency of unemployment, S1/(S1+S2), or more briefly n/N, where N is sample size and n is the number unemployed. The likelihood of the sample in the OLS model is then

$$(n/N)^n \left[1 - (n/N)\right]^{N-n} .$$

Incorporating the information on the observed wage into the OLS likelihood function allows us to characterize the OLS and CRM likelihood functions as similarly as possible. For the OLS model,

$$L(B1, \sigma_1, n, N | X1, X2) = \prod_{S1} (n/N) \prod_{S2} \left[1 - (n/N)\right] F'\left(\frac{W - B1X1}{\sigma_1}\right)(1/\sigma_1)$$

$$= (n/N)^n \left[1 - (n/N)\right]^{N-n} \prod_{S2} F'\left(\frac{W - B1X1}{\sigma_1}\right)(1/\sigma_1) .$$

---

[7]This function is a superficially appealing approach to estimation of the CRM because of its relative simplicity, especially if we are only interested in the participation decision. However, inspection shows the parameters to be identified only up to a scale factor of proportionality (recall that σ_1 and σ_2 must be estimated). For more on the problems of indentification in this class of models see Nelson (1974). Something we will do with this likelihood is to plug in the parameter estimates from the full CRM likelihood in order to evaluate the likelihood of employment status alone without including data on the observed wage for the employed.

Ideally we would like to be able to compare the CRM and the OLS models with the aid of a formal hypothesis test. The most appealing approach would be to consider the OLS model as a constrained subset of the CRM model, with corresponding null hypothesis that OLS is true against the alternative hypothesis that the CRM is true. The test statistic would then be the likelihood ratio statistic. Therefore it is necessary to ask; is the OLS model in fact a subset of the CRM model? It turns out that it is not. Consider the relevant probability measures in the CRM model

$$Pr(W=0) = F\left(\frac{B2X2-B1X1}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$$

and

$$Pr(W>0 \text{ AND } W=We) = Pr(W>Wp \text{ AND } W=We)$$

$$=F\left(\frac{W-B2X2}{\sigma_2}\right) F'\left(\frac{W-B1X1}{\sigma_1}\right)(1/\sigma_1)$$

and in the OLS model

$$Pr(W=0) = (n/N)$$

$$Pr(W>0 \text{ AND } W=We) = [1-(n/N)]F'\left(\frac{W-B1X1}{\sigma_1}\right)(1/\sigma_1).$$

Letting n/N be equal to the constant probability P we see that for OLS to be a constrained subset of the CRM there must exist some parameters B1, B2, σ_1, σ_2 such that

$$F\left(\frac{B2X2-B1X1}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) = P \text{ for all i}$$

and

$$F\left(\frac{W-B2X2}{\sigma_2}\right) = 1-P \text{ for all i.}$$

Remembering that X1 and X2 are large matrices of dimension number of coefficients by number of observations it becomes apparent that the first equality will be true only for B1=0 and B2=0 for all B's in B1 and B2. If B2 is 0 then the second equality requires that W be constant for all i since the second equality degenerates to $F(W/\sigma_2)$ equaling a constant 1-P and σ_2 is

constant for all i. Consequently, at least for the time being, it is necessary to look for some informal means of comparison between the OLS and the CRM models. Since the two models were estimated separately for racial groups these tests are more informative than one might think. Consider the hypothesis that for blacks unemployment is random and for whites unemployment is voluntary and given by the CRM decision rule, i.e., those who are not working have chosen to do so because their expected wage fails to exceed their participation wage. To examine this hypothesis the following statistics were computed for each racial group and informal comparisons were made. These statistics are the mean squared error of the wage equations, goodness of fit of the CRM functional probabilities to the OLS constant probability and the simplified log likelihood for employment status at estimated parameter values. A fourth statistic for goodness of fit of the CRM is[8]

$$Z = \sum_i \{Y(i) - P(X)\} \Big/ \sqrt{\sum_i \{P(X)[1 - P(X)]\}}$$

where Y(i) is 1 or 0 for i's employment status.

If we knew the true parameter vectors this statistic would be distributed N(0,1). However we only have estimates for B1,B2 σ_1 and σ_2 of which P(X) is a function. Because of this a fairly complex correction to this statistic is required using the variance-covariance matrix of the estimated parameter values. For reasons of time and diminishing marginal returns this test is set aside for future work[9]. The statistic for goodness of fit of the CRM probabilities of unemployment to the constant OLS probability of unemployment is constructed like a χ^2 statistic, however it is not

---

[8]Note that the notation P(X) is a function of each characteristic vector X(i) and is meant to imply P[X(i)].

[9]Actually this statistic was computed without the correction for estimates of the parameters. For the record the results are reported in the next section. The value of this statistic for both white and black is out in the tail of the distribution. It is likely however that introduction of the correction would bring the values closer to the center of the distribution.

distributed χ^2. The formula used here is

$$\sum_i \{[P(X)-P]^2/P\} = Q.$$

The mean squared error of the OLS model is given by

$$\sum_i [Y(i)-P]^2/N \qquad \text{since } E[Y(i)]= P$$

where Y(i) is employment status.

For the CRM model the mean squared error is given by

$$\sum_i [Y(i)-P(X)]^2/N \qquad \text{since } E[Y(i)]= P(X).$$

Often two mean squared errors may be compared formally with an F-test on their ratio as each is distributed χ^2. However, as before with the likelihood ratio test, in this case we run into the problem of identifying the hypothesis to be tested. Because the OLS model is not nested in the CRM model an F-test is inappropriate. A possible solution to this problem is the estimation of a "supermodel" in which both the CRM and OLS models are nested, if such a model exists. Another approach that may bear fruit in the future is the derivation of the distribution of the likelihoods for the two models so that their difference or ratio may be compared. The last method used here to evaluate the two models in terms of their relative performance across racial groups is the calculation of log likelihoods for the probability of employment status from the two models. For the OLS model this required calculation of

$$\text{Log } L(n,N) = n\text{Log}(n/N) + (N-n)\,\text{Log}[1-(n/N)]$$

and for the CRM,

$$\text{Log } L(B1, B2, \sigma_1, \sigma_2 | X1, X2) =$$

$$\sum_{S1} \text{Log } F\left(\frac{B2X2-B1X1}{N\sigma_1^2+\sigma_2^2}\right) + \sum_{S2} \text{Log}\left[1-F\left(\frac{B2X2-B1X1}{N\sigma_1^2+\sigma_2^2}\right)\right]$$

at estimated parameter values.

Results for these procedures along with coefficient estimates for the OLS

and CRM models are reported in the next section.

D. RESULTS OF THE ESTIMATIONS

Before describing the results a discussion of the data is in order. The sample is a subset of the National Longitudinal survey data on men aged 45-59 in 1966, for whom observations on the wage were available or who were listed as unemployed by Current Population Survey definition. The sample excluded men who never worked, men paid by piece rate, men self-employed and volunteer workers. This selection process also removed a number of unexplained missing sample points and reduced the uncensored sample size of 5020 to 3705. Of the 1315 missing observations, about 430 are missing because their labor force status did not describe them as eligible for an hourly wage and 885 are missing because of other factors in the selection process as described above. Estimation of the model further reduced the sample size, as cases were removed for which one or more variables was missing. The final sample size was 2429 white and 1076 black for a grand total of 3505. The mean and median wages for the employed members of this hardy group are given below:

GROUP	MEAN	MEDIAN
white	3.51	3.12
black	2.22	2.10 (black and 'other')
Total	3.12	2.84

Except in the case of the fully specified wage equation discussed in Appendix II the sample was further reduced by random sampling so that the core space requirements for the program that estimated the CRM would not be unmanageable (see Appendix III). The final estimations of the CRM and the OLS were done on two samples, one of 847 whites, 36 of whom were unemployed, and one of 846 blacks, 29 of whom were unemployed. Sampling weights were introduced into the likelihood functions so that they would estimate in the context of the correct ratio of unemployed to employed. Likewise, any time

the relative frequency of unemployment was calculated it was done on the basis of the relative frequency in the uncensored sample.

The results for the expected wage equation are presented in Table 1 below[10]. On the following page Table 2 presents the results of the corresponding CRM estimation of the participation wage. Table 3 contains the statistics for comaprison of the CRM and OLS models described in the last section.

---

[10]Note that coefficients are of the form LOG(B).

Table 1

EXPECTED WAGE EQUATION FOR WHITES:

	OLS		CRM	
Variable	Coefficient	T-statistic	Coefficient	T-statisic
CONSTANT	5.185	----	5.177	160.59
HEALTH	0.1520	3.85	0.1618	12.23
(some effect->no effect on work)				
AGE	-0.00646	-1.73	-0.00658	-5.22
(45->59)				
EDUCATION	0.0551	11.75	0.05617	37.70
AREA RES	-0.0345	-6.06	-0.03717	-18.60
(urban->rural)				
JOB TENURE	0.01267	8.17	0.01273	23.84
	F-STATISTIC= 66.48 R-SQUARED= .29 STD ERROR REG=.44		ESTIMATED M.S.E.=3.685 MAXIMUM LIKELIHOOD ESTIMATE OF STD ERROR OF REG= .4384 (T-statistic=160.59)	

EXPECTED WAGE EQUATION FOR BLACKS:

CONSTANT	5.837	----	6.107	39.17
HEALTH	-0.00347	-0.09	-0.2856	-10.62
AGE	-0.0108	-3.03	-0.00687	-2.58
EDUCATION	0.02728	6.77	0.01604	5.833
AREA RES	-0.0845	-14.40	-0.0848	-20.88
JOB TENURE	0.01136	7.28	0.01022	9.97
	F-STATISTIC= 91.95 R-SQUARED= .36 STD ERROR REG=.41		ESTIMATED M.S.E.=1.892 MAXIMUM LIKELIHOOD ESTIMATE OF STD ERROR OF REG=.4075 (T-statistic=73.69)	

Table 2

PARTICIPATION WAGE EQUATION FOR WHITES

Variable	Coefficient	T-statistic
CONSTANT	2.182	1.08
MAR.STAT.	-0.0555	-0.109
(1=married,2=not married)		
NO.DEP'S	-0.00254	-0.0296
(dependents)		
TRANSFERS	0.00221	1.19
(food stamps, unemployment compensation, welfare and public assis.)		
ASSETS	-0.000055	-0.00841
(Family and Business)		
EDUCATION	0.02616	0.58500

ESTIMATED M.S.E.=3.685
MAXIMUM LIKELIHOOD ESTIMATE
OF STD ERROR OF REG=1.31 (T-statistic=1.27)

PARTICIPATION WAGE EQUATION FOR BLACKS

CONSTANT	-8.264	-0.48
MAR.STAT	0.8962	0.634
NO.DEP'S	-0.00168	-0.114
TRANSFERS	0.0009	0.763
ASSETS	0.00402	0.0785
EDUCATION	-0.1033	-0.609

ESTIMATED M.S.E.=1.892
MAXIMUM LIKELIHOOD ESTIMATE
OF STD ERROR OF REG= 5.0 (converged at a boundary,T=.78)

Table 3[11]

	WHITE		BLACK	
STATISTIC	OLS	CRM	OLS	CRM
1) GOODNESS OF FIT (OLS with CRM for probabilities)		113.7		15.7
2) MEAN SQUARED ERROR (for probabilities)	.0414	.0376	.0331	.03377
3) LOG LIKELIHOOD (of probabilities)	-186.2	-155.6	-132.7	-152.7
4) LOG LIKELIHOOD (from full CRM)		-1534.0		-787.0
5) UNCORRECTED NORMAL TEST (CRM only)		6.4		11.8

Regardless of the hypothesis about how the CRM results should compare with the OLS results there is no getting around the fact that the estimates of the participation wage are not very good. No doubt this is partly a result of the CRM estimating procedure which is very sensitive to parameters used in the numerical algorithm such as increment halving and boundary constraints. In addition choice of initial values was very important. The initial values procedure used here was cumbersome, involving preliminary OLS and Probit estimation in the manner suggested by Nelson (1974). For a more detailed discussion of this procedure see Appendix III. Another factor contributing to the high standard errors of the coeffcient estimates in the

---

[11] There are a lot of mean squared errors floating about and it is important to distinguish them. The estimated M.S.E. of tables 1 and 2 is computed by the CRM regression program and is the estimated M.S.E. of the entire CRM system. Also in Tables 1 and 2 the values given as maximum likelihood estimates of standard error of the regression are the estimated parameters, σ_1 and σ_2, one for each equation. The mean squared errors in table 3 are for differences in actual and expected probabilities of unemployment predicted by the two models. They are given by the formulas shown earlier.

participation wage equation is simply the meager information available on which to base an estimation. The CRM had no trouble estimating the expected wage with a high degree of confidence.

Inspection of the results in terms of the original hypothesis shows the evidence to be mixed. The signs of the coefficients on the expected wage equation do not change between CRM and OLS estimation for either race. However, for whites the magnitude of the coefficients for the CRM and OLS estimation of the expected wage is much closer than it is for blacks. This might indicate that sample selection is more important for blacks than for whites, contrary to the original hypothesis that black unemployment is randomly drawn while white unemployment is voluntary. This difference in magnitude is fairly weak evidence especially when one considers the fact that the greatest differences in OLS and CRM estimates for blacks are found for those coefficient estimates in which the least confidence is warranted. These are the health and age coefficients. The black OLS health coefficient is negative and this clearly is a nonsensical result since the health variable ranged from 1= some health effect to 2=no health effect on work. The T-statistic for this estimate is almost zero and the sign in this case can be attributed to sampling error. In Appendix II the fully specified large sample model produced a positive health coefficient for blacks, though again it is not significantly different from zero. There is no denying that the coefficient remains negative in the CRM estimation and the level of signficance increases dramatically. This suggests another explanation, that the model is mis-specified. The other main difference between OLS and CRM estimation of the expected wage equation for blacks is in the age and education coefficients. The signs don't change but the CRM estimates are somewhat smaller in absolute value. In both models the age coefficient estimates are less significant than the other regressors except for some health coefficient estimates.

The negative sign on age for both races in both models is an interesting result. It was robust under a number of OLS specifications. It seems that whatever returns to age there are for hourly wage earners come from increased job tenure, and that once this variable is added age is free to show its true effects. Variations in specification designed to test this result are discussed in Appendix II. The remaining estimates for the expected wage equations seem to make sense in both the CRM and OLS models. The overall pattern in the coefficients point up some interesting differences in black and white ~~age~~ wage determination. While education and job tenure are significant positive factors for both races they are less so for blacks, where type of area of residence is the driving factor.

The participation wage estimates in Table 2 are very difficult to interpret and a number of stories may be told about them. For both blacks and whites the participation wage is low relative to the expected wage. In fact there are no sample points, unemployed or otherwise, for whom the CRM was successful in fitting the log of the participation wage above the log of the expected wage. This does not necessarily imply that there is no one for whom Wp > We. To determine the fitted We and Wp values requires more than taking the exponential of log values because of the bias introduced by the non-linearity of the logarithmic transformation. The correction is slightly complex (see Neyman and Scott, 1960) and requires estimates of σ_1 and σ_2 to get corrected estimates even at the means of the independent variables. It was not possible to get any reliable corrected estimates of black participation and expected wages because the estimate of σ_2 converged to a boundary constraint and it is therefore unlikely that the estimate is very close to the true value.

A number of variations in the numerical algorithm parameters were tried for both races to evaluate the sensitivity of the results to these factors.

The white results were not very sensitive to these variations. However, in the black participation wage equation it was found that either the σ_2 term always converged to the positive maximum boundary or the constant term always converged to the negative minimum boundary. This suggests the interesting interpretation that for blacks the participation wage is virtually zero. If the participation wage is zero for blacks the constant term should converge to a very negative number as the wage tries to reach zero. Likewise if the constant is free to go as low as it wants then the error of the equation could be expected to converge at its maximum since the wage can only reach zero in the logarithmic transformation when the determining variables reach negative infinity. This interpretation also helps explain the low T-ratios for the coefficients in the black participation wage equation. We cannot reject the hypothesis that all of the coefficients are zero. The zero participation wage explanation also helps explain the counter-intuitive result that for blacks more education decreases the participation wage. If the true coeffecients are zero and the constant would be infinitely negative if we let it then the signs on the estimated coefficients are meaningless.

The parameter estimates for the white participation equation suggest that the white participation wage may be positive, but there is no conclusive evidence either way. The constant is positive and along with the coefficient on transfer income is the most significant determinant of participation wages. Like the equation for blacks the T-ratios are low and do not encourage confidence in the estimates. An encouraging difference is the better estimate of σ_2 for whites than for blacks. Though not statistically significant, the results for whites are somewhat better overall than for blacks in the participation wage. The most disturbing result for white participation wages is the negative coefficient on assets. However the T-ratio for this estimate is the worst of all the estimates and

is virtually zero. Once again the evidence is weak, but it seems to suggest that if the participation wage is significant for anyone it is significant for whites. If the participation wage is indeed relevant for whites and irrelevant for blacks then the hypothesis that white unemployment is more a product of voluntary decisions than black has received some support.

Turning to the comparative statistics of Table 3 a definite pattern presents itself. The first statistic indicates that the fit of estimated probabilities from the CRM to the OLS relative frequencies is much better for blacks. Interpretation of this result is highly problematic; are the black CRM results close to the 'true' OLS results or vice versa? With regard to the mean squared errors of the probabilities from the two models the expected inequality holds. For whites the CRM did a better job of predicting employment status than did the OLS model. For blacks the opposite is true, the OLS has a lower mean squared error than the CRM. The log likelihoods for employment status are consistent with the first two results. For whites CRM predictions of employment status are more 'likely' then the OLS predictions and for blacks the opposite is true. However, the likelihood of the full CRM is higher for blacks than it is for whites. One explanation for this is the better overall fit of the black expected wage equation which appears in the CRM. It is also possible for the CRM to do a better job of explaining wages for blacks than for whites and still do a worse job relative to the OLS model. That is, the CRM is a better explainer for blacks than for whites but the OLS is a better explainer for blacks than the CRM. The ~~fourth~~ fifth statistic is the uncorrected normal test of the CRM. For both races the statistic's values are high enough to reject the CRM, however the rejection is stronger for blacks than for whites. Introduction of the correction for parameter estimates mentioned in footnote nine would probably bring both races' statistic value closer to the center of the distribution without changing their relationship to each other.

These results are admittedly inconclusive. Nonetheless it seems that what evidence there is supports a case for the notion that white adult unemployment in 1966 was a product of deliberate decision while this is less likely for blacks. Conversely, it seems more likely that black unemployment was a random, involuntary phenomenon than it does for whites.

E. CONCLUSIONS AND FUTURE WORK

Although it is not entirely satisfactory, the procedure proposed here for characterizing unemployment appears to be an interesting and useful approach. Some weak evidence has been given for differences in the nature of unemployment across races. Clearly, more work is necessary before anything can be said with confidence. The econometrics of the censored regression model are complex and cumbersome. In order to improve this situation some work has been done on simpler estimators for this class of models. See, for example, Heckman (1976) wherein a simplified estimator is proposed that allows estimation by least squares and Probit analysis. A recent revision and clarification of this article is Heckman (1979) wherein a two-stage estimator is applied to the problem.

In Heckman (1976) a censored variable estimation is performed for white married women. Heckman compares his estimator with the maximum likelihood and OLS estimators. He does not, however investigate a breakdown by racial groups. Heckman (1979) points out that while his results do not allow rejection of the null hypothesis that sample selection is an unimportant phenomenon Gronau (1974) found significant selectivity bias. Both Gronau and Heckman concentrate on selectivity bias as an econometric missing variable or specification problem and not as a tool for understanding the type of unemployment observed. Heckman's estimates of the sample likelihood for the hypothesis of sample selection and the hypothesis of no sample selection are almost identical[12]. If one considers women to be a group that suffers discrimination this tends to support the notion that for groups for whom unemployment is an involuntary phenomenon the OLS will do at least as good a job of predicting employment status as the censored regression model.

---

[12]The likelihood for the hypothesis of sample selection is -5,778 and for the hypothesis of no sample selection is -5,783 (Heckman, 1976, Table 3).

An important contribution to the type of study undertaken here would be the development of statistics enabling one to compare apples and oranges like the OLS and censored regression model in a rigorous and meaningful way. To this end two suggestions are offered. A useful approach may be the development of a distribution theory for the likelihood equations generated by this type of comparison. Their individual distributions should give rise to a distribution useful for the construction of hypothesis tests on their differences. Another valuable contribution would be the identification of a more general model in which both the CRM and OLS are nested.

A number of interesting modifications should be made to the economic foundations of the model. One of the most interesting would be the introduction of time preference and search behavior. In this cases the CRM might describe the unemployed as making decisions based on the difference between the present value of the expected wage and the present value of the participation wage, over time.

APPENDIX 1: The Expected Wage as Instrumental Variables Estimator.

Hall's procedure for imputing a wage to the unemployed and removing the bias from measurement error begins by considering a simple labor supply function. For example,

$$L = Bo + B1W + u$$

where W is the actual wage for agent i and u is the iTH stochastic error term. In many cross-section studies the observed wage differs from the true wage by some random measurement error:

$$w = W + v$$

where w is the iTH observed wage and v is the error of measurement. We assume that v is uncorrelated with W , i.e., that there are no systematic errors in measuring the wage. Substituting the latter equation into the former gives

$$L = Bo + B1w + u - B1v \text{ because } W = w - v$$

$$= Bo + B1w + e \text{ where } e = u - B1v$$

which is what we would estimate without correcting for measurement error. The problem of measurement bias arises because v is positive in w = W + v and negative in e = u - B1v causing w and e to be negatively correlated, violating one of the classical assumptions that errors are uncorrelated with regressors. This negative correlation will bias B1 downwards, underestimating the labor supply response to the wage (Hall,1973). However, this problem as well as the problem of a lack of observations on the wage rate for the unemployed may be resolved in one procedure by adding an instrumental variables estimator to the labor supply equation, where the instruments are the (exogenous) determinants of the wage rate. The wage determination equation says that the observed wage is a function of personal characteristics, a random disturbance for measurement error, v and a second error term, p . The resulting wage equation is

$$W = Ao + A1X1 + ...AnXn + p + v$$

$$= w' + p + v$$

where X = X1,.....,Xn is a vector of observed personal characteristics. Applying OLSQ to this equation yields an equation that will provide an imputed wage:

$$w' = Ao' + A1'X1 + ... + An'Xn$$

where the A' s are estimated coefficients.

From the two equations

$$L = Bo + B1w + u - B1v \text{ and}$$

$$w = w' + p + v$$

we get

$$L = Bo + B1[\ w' + p + v\ ] + u - B1v$$

$$= Bo + B1[\ w' + p\ ] + u$$

$$= Bo + B1w' + z$$

where

$$z = B1p + u.$$

Estimation of the last labor supply equation,

$$L = Bo + B1w' + z$$

by OLSQ is appropriate if the imputed wage is used for all observations in the sample[13] and will provide consistent and unbiased estimates of the parameters of the labor supply equation (Hall, 1973, pp. 110-111).

---

[13] If the imputed wage were to be used only for those without an observed wage and the observed wage for those with an observed wage, then the appropriate estimator would be weighted least squares, with less weight on those sample points using the imputed wage because VAR(z) > VAR(u).

APPENDIX II: The Fully Specified Expected Wage Equation.

The following model was estimated for each racial group:

$$lnW=AX1+BX2+CX3+DX4+EX5+FX6+GX7+HX8+e$$

where

X1=1	HEALTH limits amount or kind of work
=2	HEALTH has no effect on work
X2=AGE	45-59
X3=EDUCation	years, 1-18
X4=SKILL	content of current or most recent job by training time required
X5=RESindency	by population density
X6=RCAN	Duncan index of socioeconomic prestige for current or most recent job
X7=TENURE	at current or most recent job
X8=UNEMployment	rate in respondent's labor market.

The results for this regression are reported separately for black and white below:

(A) FOR WHITES, working or with a job not working by CPS definitions

Dependent Variable=Ln(hourly wage) F-statistic=197.5

R-squared=.36

Std. Error Reqr.=.42

VARIABLE	COEFFICIENT	STD.ERROR COEF	F-statistics
HEALTH	.09	.02	18.1
AGE	(-.07)	.002	10.5
EDUC	.03	.003	67.0
SKILL	.05	.006	55.0
RES	(-.03)	.003	63.1
RCAN	.005	.0005	97.6
TENURE	.008	.0008	101.3
UNEM	.002	.0004	35.2

The regression is highly significant overall and all coefficients are significantly different from zero at any reasonable level of significance. The results for blacks had a similar basic pattern:

(B) FOR BLACKS, same labor force group as in above table

Dependent Variable=Ln(hrlywage) F-statistic=97.0

R-squared=.39

Std.Error Reqr=.40

VARIABLE	COEFFICIENT	STD.ERROR COEF	F-statistic
HEALTH	.04	.03	1.7
AGE	(-.01)	.003	14.0
EDUC	.015	.004	16.0
SKILL	.03	.009	9.3
RES	(-.09)	.005	288.7
RCAN	.004	.001	10.0
TENURE	.009	.001	51.9
UNEM	.002	.0006	16.2

In this set of regressions there are a number of interesting results. Education is highly significant but certainly not the most significant determinant of wages, especially in the case of black workers. It would seem that the most important factor in the determination of black wages in 1966 was residency by population density. Another interesting result is the significance of the tenure variable in both regressions. A surprising result is the sign of the coefficient on the unemployment variable which has a robustly positive sign under all variations in regressors and functional form tried. This suggests the counter-intuitive conclusion that the higher the unemployment rate in the respondent's labor market, the higher a wage he could expect. One explanation for this is a leftward shift in supply in the high unemployment areas, in which case some observed unemployment must be voluntary. Another explanation is the negative zero-order correlation between population density and unemployment (-.12), that is, as population density decreased, unemployment fell, therefore high unemployment is correlated with high population density which is correlated with high wages and the two variables, UNEM and RES are picking up some of the same effects. Retention of UNEM in the regression and removal of the population density variable, RES raises the significance of the coefficient on UNEM, supporting the idea that these two variables proxy some of the same effects. There were some problems with multi-collinearity, though nothing severe enough to warrant re-specification. The highest zero-order correlations between dependent variables were around .67, between the skill and the Duncan index variable. This is admittedly high, but the Duncan index for current or most recent job and the skill variable probably pick up a lot of the variation in wages that are not captured by education or job tenure. The introduction of variables for father's education or Duncan index when the respondent was 15 produced no significant results, not surprising for this age group at this time. Perhaps the most interesting result from this procedure is the

significantly negative coefficient on age. Other studies have found this, but not usually until around age 60. This result was robust under all specifications attempted including the introduction of an interaction term for age and tenure, AGE*TENURE. However, the interaction term estimation was uninformative because it was 99% collinear with the tenure variable. There is an interesting result that sheds a little light on the negative age coefficient. Removal of the tenure variable, which was zero-order correlated with age between .1-.2, resulted in the sign of the age coefficient remaining negative, but becoming insignificant. This indicates that whatever positive effects age has on wages over this range is largely because of the correlation of age with job tenure. When tenure is introduced into the analysis, it takes on a significantly positive coefficient and allows the effect of age to show shows its true colors[14]. To wrap up this discussion zero-order correlations for the dependent variable with all the independent variables are reported below:

ln(wage)	HEALTH	AGE	EDUC	SKILL	RES	RCAN
White	.14	-.07	.43	.44	-.24	.51
Black	.05	-.11	.36	.27	-.52	.34

ln(wage)	TENURE	UNEM	DADCAN (Duncan index father's job)
White	.24	.12	.25
Black	.26	.09	.08

For the most part these results are consistant with the regression results, particularly with respect to Black/White differences.

---

[14]This last discussion applies only to the white results, the coefficient on age for blacks remains negative and significant no matter what is done.

APPENDIX III: Technical Issues in Estimation

a. Programming the Censored Regression Model

The censored regression model was estimated by an adaptation of the BMDP (1977) proprietary software subrogram PAR for derivative free non-linear regression. Ordinarily the PAR program will estimate a non-linear equation specified in a user written FORTRAN subroutine, the format of which is given in the BMDP manual. Users of IBM operating systems may specify the location of the subroutine in system job control language. For XEROX users like Oberlin College it is necessary to modify the source code of the PAR program directly, inserting the user written subroutine as an internal subroutine to be compiled with the rest of the program. In addition, if it is necessary to introduce sampling weights into the likelihood function (as opposed to the data) because of biased sampling (in this case the unemployed were oversampled) an additional labelled COMMON block should be created containing the weights. The main program may then pass this information to the subroutine without cumbersome, CPU time consuming read statements. PAR generates parameter estimates by least squares using a psuedo Gauss-Newton algorithm. To adapt the program to do maximum likelihood estimation the least squares criteria for convergence may be turned off and replaced with a user supplied loss function in the FORTRAN subroutine. The loss function, in this case -2 times the log likelihood , will be minimized at convergence. (See pp. 499-513 in BMDP-77.) There is an alternative method mentioned in the manual that is somewhat less transparent.

Users of small and medium sized operating systems may expect to encounter several difficulties with the use of PAR for maximum likelihood estimation with cross-section data. The PAR program loads all the data into core memory and its core requirements rise rapidly with the number of parameters to be estimated. For the CRM estimated here there were twelve

parameters. It was found that the regular sized PAR program of 15,000 words could only estimate that many parameters for about 200 cases. With the program workspace increased to 55,000 words and a number of program overlays the total core required became 72,000 words, the maximum usable core on the Oberlin XEROX Sigma-9. At this size the program was capable of estimating twelve parameters for about 900 cases. The BMDP P3R program is also capable of maximum likelihood estimation by the methods described above and in the BMDP-77 manual. It is more space efficient but requires that the user supply partial derivatives for all parameters in the FORTRAN subroutine. The PAR program is slow, requiring about 60 minutes of CPU time for each 100 iterations.

The advantage of using the PAR program is that little complex programming is required of the user. An alternative, relatively labor intensive approach is a user written main program manipulating the Newton-Raphson minimization and matrix inversion subroutines from the FORTRAN Scientific Subroutine Package[15]. This approach is probably more efficient in terms of computer resource consumption as the programmer need not instruct his main program to store the data in core during execution. Consumption of CPU time would probably go up with this method because of additional input and output tasks. In addition, more programming is required for this approach than with the adaptation of relatively complete, user- friendly proprietary software.

A number of parameters affecting the numerical algorithm in PAR may be modified by the user. For the CRM estimates done here boundary constraints were imposed on all coefficients of (-10,+10) and constants of (-20,+20) and standard deviations, σ_1 and σ_2 of (0,5). The only estimation to converge

---

[15]The evaluation of the normal density and distribution in the FORTRAN likelihood function subroutine was done using the FORTRAN Scientific Subroutine Package subroutine NDTR in double precision arithmetic.

at a boundary point was the black σ_2 term in the participation wage equation. Some possible reasons for this boundary convergence are given in the body of the paper. Another user controlled parameter is the number of increment halvings between iterations in the search for a minimum of the loss function. The program defaults to a maximum of five. It was found that increasing the maximum number of increment halvings would sometimes reduce the number of iterations required for convergence. Variations in increment halvings and boundary constraints did not have much effect on coefficient estimates but did effect the standard errors of the estimates to some degree. For blacks the reported estimates were chosen on the basis of the lowest loss function for which the constant in the participation wage equation did not converge to a boundary. (The equation with a slightly lower loss function wherein the participation wage constant term did converge to boundary gave a constant of -20!) For whites no parameter estimates converged at boundary points and different estimations gave very similar results.

b. Initial Values Procedure

Good initial values were very important. The procedure used here follows that suggested by Nelson (1974). The steps are described below:

1. Estimate the expected wage equation for the employed by ordinary least squares.
2. Retain the coefficients from step 1 and impute a fitted wage to all members of the sample, including the unemployed.
3. Using the fitted wage as the observed wage = expected wage, apply Probit analysis to estimate the probability that We>Wp=B2X2+v. Gronau (1973, p.S178) provides a more detailed explanation of the application of Probit to this type of threshold problem and the interpretation of Probit coefficients.

Nelson (1974) points out that the Probit likelihood function is actually a special case of the CRM likelihood function when the COV(u,v) is zero and the expected wage is observed for the entire sample. Then,

$$Pr(W=0) = Pr(We<Wp) = Pr(v<B2X2-We)$$

$$= F\left(\frac{B2X2-We}{\sigma_2}\right).$$

and

$$Pr(w>0) = Pr(We>Wp) = 1-Pr(v<B2X2-We)$$

$$= 1-F\left(\frac{B2X2-We}{\sigma_2}\right)$$

giving sample likelihood where Y(i) is employment status;

$$L(B2, \sigma_2 \mid Y, We, X2) =$$

$$\prod_{S1} F\left(\frac{B2X2-We}{\sigma_2}\right) \prod_{S2} \left[1-F\left(\frac{B2X2-We}{\sigma_2}\right)\right]$$

which is the Probit likelihood function.

REFERENCES

1) Abbot,M., and O. Ashenfelter (1976), "Labour Supply, Commodity Demand, and the Allocation of Time," Review of Economic Studies, Vol. 43, pp. 389-411.
2) Ashenfelter, O. and J. Altonji (1980), "Wage Movements and the Labor Market Equilibrium Hypothesis," Economica, Vol. 47, pp. 217-245.
3) Azariades,C. (1975), "Implicit Contracts and Underemployment Equilibria," Journal of Political Economy, Vol.83, pp. 1183-202.
4) Deaton, A., and J. Muellbauer (1980), ECONOMICS AND CONSUMER BEHAVIOR, Cambridge: Cambridge University Press.
5) Dixon, W.J., Ed. (1977), BMDP-77, BMD BIOMEDICAL COMPUTER PROGRAMS P-SERIES 1977 Los Angeles: University of California Press. Program revision date 1978.
6) Goldberger,A.S. (1967), "Functional Form and Utility: A Review of Consumer Demand Theory," Unpublished paper, Social Systems Research Institute, University of Wisconsin, October 1967.
7) (1975), "Linear Regression in Truncated Samples," Social Systems Research Institute, University of Wisconsin-Madison, May 23, 1975.
8) Gronau, R. (1973), "The Effects of Children on the Housewife's Value of Time," Journal of Political Economy, Vol. 81, (Supplement) pp. S168-99.
9) (1974), "Wage Comparisons-A Selectivity Bias," Journal of Political Economy, Vol. 82, pp. 1119-43.
10) Hall, R.E. (1970), "Why is Unemployment so High at Full Employment?," Brookings Papers on Economic Activity, No. 3, pp.369-410.
11) (1973), "Wages, Income and Hours of Work in the U.S. Labor Force," in G. Cain and H. Watts (eds.), INCOME MAINTAINANCE AND LABOR SUPPLY, Chicago: Markham.
12) Heckman, J.J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," Annals of Economic and Social Measurement, Vol. 5, pp. 475-492.
13) (1979), "Sample Selection Bias as a Specification Error," Econometrica, Vol. 47, pp. 153-61.
14) Leuthold (1968), "An Empirical Study of Formula Income Transfers and the Work Decision of the Poor," The Journal of Human Resources, No. 3, 1970, pp.312-323.
15) Lucas, R.E. and L. Rapping (1970), "Real Wages, Employment and Inflation," in E.S. Phelps, et al. MICROECONOMIC FOUNDATIONS OF EMPLOYMENT AND INFLATION THEORY, New York: Norton.
16) McCallum, B.T. (1980), "Rational Expectations and Macroeconomic Stabilization Policy: An Overview," Journal of Money, Credit and Banking, November 1980, Part 2, pp. 716-746.
17) Nelson, F.D. (1974), "Censored Regression Models with Unobserved Stochastic Censoring Thresholds," National Bureau of Economic Research Working Paper No. 63, December 1974.
18) Neyman,J., and E.L. Scott (1960), "Correction for Bias Introduced by a Transformation of Variables," Annals of Mathematical Statistics, Vol. 31, pp. 643-655.
19) Stone, R.N.J. (1954), "Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand," Economic Journal, Vol. 64, pp. 511-527.