DOCUMENT RESUME

ED 199 457 CE 028 222

AUTHOR Bassi, Laurie J.
TITLE Estimating the Effect of Training Programs with Nonrandom Selection. Final Report, July-November 1980.
SPONS AGENCY Office of the Assistant Secretary for Policy, Evaluation and Research (DOL), Washington, D.C.
PUB DATE Nov 80
CONTRACT DOL-B-9-M-0-1064
NOTE 30p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; *Employment Programs; *Evaluation Criteria; Evaluation Needs; Federal Programs; *Job Training; Participant Characteristics; *Program Effectiveness; Program Evaluation; Research Methodology; *Research Problems; *Selection
IDENTIFIERS Comprehensive Employment and Training Act; *Nonrandom Selection

ABSTRACT

A study was conducted to develop a theoretical framework for unbiased estimation of the dynamic net impact of Comprehensive Employment and Training Act (CETA) programs on participants' earnings. (The possibility of selectivity bias arises from the non-random nature of participation in the program. That is, if participation is a function of unobservable variables such as ability or motivation, and these variables are also determinants of earnings, it will be impossible to distinguish the effects of the unobservables from the effects of the program without controlling for the selection process.) In the study, selectivity biases were controlled for through the use of an error components model. The model relies on the longitudinal nature of the data to eliminate the effects of the unobservables by differencing them away. The estimation techniques developed allow for participation to be a non-random function of individual, unobserved variables that are both fixed and changing over time, and temporary fluctuation in earnings prior to the program. This framework represents an advance in the state of the art in impact estimation for employment and training programs, since it places fewer restrictions on the nature and type of comparison group necessary for unbiased estimation and, thereby, contributes to solving a problem that has long plagued evaluations of employment and training programs. (Author/KC)

ESTIMATING THE EFFECT OF TRAINING PROGRAMS WITH NONRANDOM SELECTION

by

Laurie J. Bassi*

November, 1980

This report was prepared for the Office of the Assistant Secretary for Policy, Evaluation and Research, U.S. Department of Labor, under contract/purchase order No. B-9-M-O-1064. Since contractors conducting research and development projects under Government sponsorship are encouraged to express their own judgement freely, this report does not necessarily represent the official opinion or policy of the Department of Labor. The contractor is solely responsible for the contents of this report.

*The author wishes to acknowledge helpful discussions with Gary Chamberlain and Phil Moss. Particular thanks go to Orley Ashenfelter.

50272-101

REPORT DOCUMENTATION PAGE

1. REPORT NO. ASPER/PUR-80/1064/A

2.

3. Recipient's Accession No.

4. Title and Subtitle

Estimating the Effect of Training Programs with Nonrandom Selection

5. Report Date November 1980

6.

7. Author(s)

Laurie J. Bassi

8. Performing Organization Rept. No.

9. Performing Organization Name and Address

Laurie J. Bassi
Box 17--Graduate College
Princeton, New Jersey 08544

10. Project/Task/Work Unit No.

11. Contract(C) or Grant(G) No.

(C) B-9-M-0-1064

(G)

12. Sponsoring Organization Name and Address

U.S. Department of Labor/ASPER
200 Constitution Avenue, N.W.
Washington, D.C. 20210

13. Type of Report & Period Covered

Final
7/80 to 11/80

14.

15. Supplementary Notes

16. Abstract (Limit: 200 words)

The purpose of this study was to develop consistent econometric estimators for measuring the net impact of employment and training programs on participants' post-training earnings. The estimators derived here will provide unbiased measures of net impact when participation in the program is a non-random function of individual, unobservables variables. These unobserved variables may be: fixed, changing over time, or purely transitory. The most generalized version of the estimation technique allows for participation to be simultaneously determined by all three types of unobservables.

ant Analysis a. Descriptors

Econometrics, Economic mode cation, Evaluation, cal models

b. Identifiers/Open-Ended Terms

Selectivity Bias

c. COSATI Field/Group 5C

18. Availability Statement

Release unlimited. Available from National Technical Information Service, Spr ngfield, Va. 22161

19. Security Class (This Report)

UNCLASSIFIED

20. Security Class

UNCLASSIF D

21. No. of Pages

22. Price

TABLE OF CONTENTS

	Page
Executive Summary	i
Introduction	1
Methods for Consistent Estimation	2
An Earnings Function with Heterogeneity and Time Effects	4
Data and Estimation Agenda	15
Appendix A	17
Appendix B	19
References	21

EXECUTIVE SUMMARY

The purpose of this study is to develop a theoretical framework for unbiased estimation of the dynamic net impact of CETA on participants' earnings. The framework is developed for empirical application to the Continuous Longitudinal Manpower Survey (CLMS).

This paper presents a method of estimating CETA's effect, free of selectivity bias. The possibility of selectivity bias arises due to the non-random nature of participation in the program. That is, if participation is a function of unobservable variables (such as ability or motivation), and these variables are also determinants of earnings, it will be impossible to distinguish the effects of the unobservables from the effects of the program without controlling for the selection process.

In this paper, selectivity biases are controlled for through the use of an error components model. The model relies on the longitudinal nature of the data to eliminate the effects of the unobservables by differencing them away. The estimation techniques developed here allow for participation to be a non-random function of: individual, unobserved variables that are both fixed and changing over time, and temporary fluctuations in earnings prior to the program.

This framework represents an advancement in the state of the art in impact estimation for employment and training programs. Further, it places fewer restrictions on the nature and type of comparison group necessary for unbiased estimation and, thereby, contributes to solving a problem that has long plagued evaluations of employment and training programs.

I. INTRODUCTION

Federally funded employment and training programs have become a permanent and major feature of labor market policy in the United States. Most of these programs are now funded through the Comprehensive Employment and Training Act of 1973 (CETA). By fiscal year 1979 CETA funding had reached $9.4 billion.

Despite the magnitude of employment and training programs, little is known about their effect on participants after leaving the program. Our inability to isolate program impacts is a result of the massive amount of data that are necessary. Without specifically controlling for: time, social factors, demographic variables, the economic climate, and unobservable factors associated with each individual - it is impossible to isolate the independent effect of the program.

Recent availability of the Continuous Longitudinal Manpower Survey will eliminate many of these insurmountable data problems of the past. These data contain a representative sample of CETA participants, as well as comparison groups which have been drawn from the Current Population Survey. The availability of the comparison groups will permit an isolation of the effect of the program on participants' earnings, independent from the effects of time and the state of the economy. The longitudinal nature of the data enable an examination of the dynamic impact of the program. Perhaps most importantly, the longitudinal data also provide us with a much better opportunity to control for individial, unobservable characteristics than would be available from purely cross-sectional data.

The plan of the paper is as follows. Section II describes techniques to control for selectivity bias. Section III develops consistent estimators for the net impact of employment and training programs on participants' earnings, allowing for the possibility that participation in the program is correlated with individual, unobservable characteristics that are both fixed and changing over time. These estimators are then generalized to allow for participation to be a function of transitory earnings fluctuations prior to participation. Section IV describes the data that will be used, and outlines the estimation agenda.

II. METHODS FOR CONSISTENT ESTIMATION

Increased awareness among economists as to the pervasiveness of nonrandom selection has spawned a rapidly growing literature attempting to deal with the econometric problems that it introduces. If the selection process is a function of unobserved variables that are correlated with the independent regressors, ordinary least squares will yield inconsistent parameter estimates. The literature, to date, has developed two strategies for breaking the correlation between the error term (the unobservables) and the independent regressors.

The first approach (developed by Heckman and Maddala and Lee) attempts to estimate the value of the unobserved (latent) variables by first predicting the outcome of the selection process using probit analysis. By assuming that the error terms are normally distributed, the expected value of the latent variables can be calculated by taking the inverse of the Mill's Ratio of the predicted value of the participation variable. By including this estimated value in the earnings equation, the omitted variables problem is eliminated. The resultant ordinary least squares estimates are consistent.

This approach has a great deal of appeal since the adjustment for selectivity bias is based on a model of the economic decision to enter a program. Unfortunately, in the problem being studied here, there is a second selection process in addition to self-selection. This selection is based on program administrators' decisions about who will be permitted to participate in the program from the pool of eligible applicants. Since this decision is likely to be a function of many non-economic considerations which are unknown to an analyst, it is impossible to estimate the probit without introducing additional latent variables. Under these conditions, the Heckman technique will produce inconsistent estimates.

Given this problem it seems likely that an error components model, the second method of controlling for selectivity bias, is the appropriate technique. These models generally specify the latent variable as including both a permanent component which is associated with an individual, as well as a transitory component which is common across individuals. For the purpose of net impact estimation, the existence of a transitory component makes it necessary to have data on both participants and nonparticipants. Otherwise it is impossible to isolate the independent effects of the program from that of the transitory error component.

If the permanent component of the error (the latent variable) is constant over time, then first differencing of the data will eliminate any correlation between the remaining error and the independent regressors. If, however, the latent variable changes over time, then simple first differencing will not be sufficient to eliminate the correlation. The next section develops a model which allows for latent variables which are both fixed and changing over time, as well as transitory unobserved components. Consistent estimates are derived, allowing for the possibility that both types of unobserved components are correlated with the independent regressors.

III. AN EARNINGS FUNCTION WITH CHANGING HETEROGENEITY AND TIME EFFECTS

A unique characteristic of longitudinal data is that it enables us to control for individual unobserved characteristics (heterogeneity) to an extent that is not possible using purely cross-sectional data. A frequently used specification of earnings determination is

$$(1)\quad Y_{it} = \beta_i X_i + \beta_2 Z_{it} + \varepsilon_i + \varepsilon_t + \varepsilon_{it}$$

where Y_{it} is the earnings of the i^{th} individual in time t, X_i is a vector of background variables for the i^{th} individual that remains fixed over time, and Z_{it} is a vector of variables that changes over time. The error term consists of the components: ε_i which is unique to the individual and fixed over time, ε_t which is common across all individuals at time t, and ε_{it} which is specific to individual i at time t.

In recent work, Ashenfelter has used this model to estimate the impact of employment and training programs on participants' earnings. He assumed that Z_{it} included a participation dummy variable and a cubic in age, and that ε_{it} was a random error term with zero expectation. Using longitudinal data on a sample of 1964 MDTA participants, as well as a comparison group drawn from the CPS, Ashenfelter was able to estimate

$$(2)\quad Y_{it} - Y_{is} = (\varepsilon_t - \varepsilon_s) + \beta_2(Z_{it} - Z_{is}) + (\varepsilon_{it} - \varepsilon_{is})$$

where s was the base (pre-program) year and t represents a series of post-program years.

Any selectivity bias present in equation (1) due to correlation between the fixed unobservables, ε_i, and the independent regressors, Z_{it}, has been eliminated from equation (2) by first differencing. So if participation in the program is a function of unobservables that are constant over time (such as innate ability), equation (2) will yield unbiased net impact estimates. If, however, participation in the program is a function of individual, unobservable variables that are not fixed over time, the ε_{it} term, equation (2) will produce biased estimates since the assumption that $E(\varepsilon_{it} - \varepsilon_{is}, Z_{it} - Z_{is}) = 0$ is violated.

An example of an such a variable might be local labor market conditions. Since employment and training programs are generally funded at higher levels in areas experiencing unusually high levels of structural or cyclical unemployment, participation is likely to be a function of local labor market conditions. Since these conditions are not fixed over time, any selectivity bias created by them will not be eliminated by equation (2).

It is possible to generalize the specification of the earnings function in equation (1) to allow for correlation between the latent variables that are not fixed over time (changing heterogeneity) and program participation. By assuming that these variables change over time at a constant rate $\bar{\rho}$, a generalized version of equation (1) is rewritten below.

$$(3)\quad Y_{ik} = \beta_1 X_i + \beta_2 Z_{ik} + \varepsilon_i + \varepsilon_k + \varepsilon_{ik}$$

$$(4)\quad \varepsilon_{ik} = \rho\varepsilon_{ik-1} + \upsilon_{ik}$$

and $$(5)\quad \upsilon_{ik} = Q_{ek} + Q_{nk}$$

where Q_{ek} is a random (transitory) error for program participants and Q_{nk} is a random error term for non-participants. The sum of these two terms, υ_{ik} is a random error term with zero expectation and no serial correlation. υ_{ik} has been written in this way to allow for the possibility that participation is a function of transitory fluctuations in earnings prior to the program. This would capture the possibility that cyclically unemployed workers are more likely to participate in the program. The Q term could also incorporate the effect of "creaming". If program administrators "cream" from the pool of eligible applicants, then

$$E(Q_{ek}) < E(Q_{nk}) \quad , \quad k < t$$

$$E(Q_{ek}) = E(Q_{nk}) \quad , \quad k \geq t$$

where t is the period of participation. This would mean that while the participants met the eligibility requirements for the program, their eligibility was only temporary because of an unlucky year prior to program participation.[1] By allowing for the possibility that $E(Q_k) \neq 0$ prior to the program, any correlation between υ_{k-1} and Z_k can be removed. As we will see below, this is critical if $\rho \neq 0$.

For the sake of expositional simplicity, we will begin by assuming that $E(Q_k) = 0$, for all t. Towards the end of this section, this assumption will be relaxed. The availability of an "appropriate" comparison group is assumed.[2] Also it will be assumed throughout that the β_1, β_2, and ρ are constant over time, although these parameters may vary between the participants and non-participants.[3]

Using this assumption, equation (3) is rewritten below in first differences form.

$$(6)\quad Y_{it} - Y_{it-1} = (\varepsilon_t - \varepsilon_{t-1}) + \beta_2(Z_{it} - Z_{it-1}) + (\varepsilon_{it} - \varepsilon_{it-1})$$

[1] - These requirements generally restrict participation in the program to individuals who: have experienced a spell of unemployment, have very low earnings, or are welfare recipients.

[2] - Techniques for choosing an "appropriate" comparison groups are described in the next section.

[3] - The model is written here using the assumption that the parameter values are constant across the two groups. This assumption can be easily relaxed. It is also possible to generalize the model to allow for a higher order autoregressive scheme, although no such generalization will be reported in this paper. It is not possible, however, to relax the assumption that ρ is constant and simultaneously maintain identification of program impacts.

The impact of the program on participants' earnings in period t is captured in β_2.[4] Note that the fixed heterogeneity (ε_i) has been eliminated by differencing. If, however, the heterogeneity changes over time, then simple differencing will not completely eliminate the possibility of selectivity bias. Any remaining inconsistency can be eliminated by using equation (4) to rewrite the error term in equation (6).

$$\varepsilon_{it} - \varepsilon_{it-1} = \rho(\varepsilon_{it-1} - \varepsilon_{it-2}) + \upsilon_{it} - \upsilon_{it-1}$$

$$\varepsilon_{it} - \varepsilon_{it-1} = \rho[(Y_{it-1} - \beta_1 X_i - \beta_2 Z_{it-1} - \varepsilon_i - \varepsilon_{t-1}) - (Y_{it-2} - \beta_1 X_i - \beta_2 Z_{it-2} - \varepsilon_i - \varepsilon_{t-2})] + (\upsilon_{it} - \upsilon_{it-1})$$

$$(7) \quad \varepsilon_{it} - \varepsilon_{it-1} = \rho(\varepsilon_{t-2} - \varepsilon_{t-1}) - \rho\beta_2(Z_{it-1} - Z_{it-2}) + \rho(Y_{it-1} - Y_{it-2}) + (\upsilon_{it} - \upsilon_{it-1})$$

It is not possible to use (7) as the basis for consistent estimation of ρ since there is a negative correlation of ρ between the error term and $\rho(Y_{it-1} - Y_{it-2})$. This problem cannot be eliminated by differencing over two periods.[5] To see this consider equation (8).

$$(8) \quad \varepsilon_{it} - \varepsilon_{it-2} = \rho(\varepsilon_{t-3} - \varepsilon_{t-1}) - \rho\beta_2(Z_{it-1} - Z_{it-3}) + \rho(Y_{it-1} - Y_{it-3}) + (\upsilon_{it} - \upsilon_{it-2})$$

[4] - Note that if $\varepsilon_t = \varepsilon_{t-1} = 0$, no comparison group would be necessary since the intercept would then measure the program's impact. In a dynamic economy, however, this condition is unlikely to hold.

[5] - Thanks are due to Gary Chamberlain for pointing out a mistake in this section of an earlier draft of the paper.

From equations (3) and (4) we have

$$(9)\quad Y_{it-1} = \beta_1 X_i + \beta_2 Z_{it-1} + \varepsilon_i + \varepsilon_{t-1} + \rho^2 \varepsilon_{it-3} + \rho \upsilon_{it-2} + \upsilon_{it-1}$$

By substituting equation (9) into (8) we would be left with a negative correlation of ρ^2 between the error term and $\rho(Y_{it-1} - Y_{it-3})$. Higher order differencing would reduce the magnitude of the bias that is created by this correlation, but it would never be eliminated.

The only way to consistently estimate ρ is by using an instrument, $\hat{Y}_{it-1}$, for Y_{it-1} in equation (7). By substituting equation (7) into (6) we are then left with

$$(10)\quad Y_{it} - Y_{it-1} = [\varepsilon_t - \varepsilon_{t-1} - \rho(\varepsilon_{t-1} - \varepsilon_{t-2})] + \beta_2 [Z_{it} - Z_{it-1} - \rho(Z_{it-1} - Z_{it-2})] + \rho(\hat{Y}_{it-1} - Y_{it-2}) + (\upsilon_{it} - \upsilon_{it-1})$$

Now suppose that *t* is the period of program participation, and the program results in a shift in earnings of β_t during that time. Equation (9) can be rewritten as

$$(11)\quad Y_t - Y_{t-1} = \beta_o + \beta_2{}^* Z_t{}^* + \beta_t P + \rho(\hat{Y}_{t-1} - Y_{t-2}) + (\upsilon_t - \upsilon_{t-1})$$

Some notational simplification has been introduced here which will be maintained throughout. The *i* subscript has been dropped. β_o indicates a combination of transitory error components, $\beta_2{}^*$ represents a non-linear combination of β_2 and ρ, and $Z_t{}^*$ represents differencing of the Z_t vector, where Z_t no longer contains the participation variable. P is a dummy variable measuring program particiation, and β_t is an unbiased estimate of the programs' effect in year *t*. Program impacts in period *t+1* are calculated below in a similar manner.

$$(12)\quad Y_{t+1} - Y_t = \beta_o + \beta_2{}^*Z^*{}_{t+1} + (\beta_{t+1}-\rho\beta_t)P + \rho(\hat{Y}_t - Y_{t-1}) + (\upsilon_{t+1}-\upsilon_t)$$

In order to consistently estimate the cumulative impact of the program, β_{t+1}, it is necessary to first estimate ρ from equation (11). Then multiplying equation (11) by ρ and adding it to (12), leaves

$$(13)\quad Y_{t+1} - (1+\rho)Y_t-\rho\hat{Y}_t - \rho^2(\hat{Y}_{t-1}-Y_{t-2}) = \beta_o + \beta_2{}^*Z^*{}_{t+1} + \beta_{t+1}P + [\upsilon_{t+1}-(1-\rho)\upsilon_t - \rho\upsilon_{t-1}]$$

It is possible to continue to solve for the cumulative program impacts in later years in a similar manner. However, the left hand side of the equation becomes extremely complicated. A more straight forward method is to solve for the impacts recursively, using the estimated values in period *(t+j)* and *(t+j+1)* to solve for the impact in period *(t+j+2)*. Both solution methods are consistent. The cumulative impact in any period can be calculated by the use of equation (14).

$$(14)\quad Y_{t+n} - Y_{t+n-1} = \beta_o+\beta_2{}^*Z^*{}_{t+n} + \beta_{t+n}P + \rho(\hat{Y}_{t+n-1} - \beta_{t+n-1} - Y_{t+n-2} + \beta_{t+n-2}) + (\upsilon_{t+n} - \upsilon_{t+n-1})$$

Once ρ has been estimated from equation (11), and β_{t+n-1} and β_{t+n-2} have been solved for, it would then be possible to estimate β_{t+n} by re-writing equation (14) as

$$(15)\quad Y_{t+n} - Y_{t+n-1} - \hat{\rho}(\hat{Y}_{t+n-1}-\hat{\beta}_{t+n-1} - Y_{t+n-2} + \hat{\beta}_{t+n-2}) = \beta_o+ \beta^*{}_2Z^*{}_{t+n} + \beta_{t+n}P + (\upsilon_{t+n}-\upsilon_{t+n-1})$$

The unique feature of this technique is that it provides us with a framework for consistent estimation of the dynamic impact of the program on participants' earnings under a very broad range of non-random selection processes. Equation (15) is sufficiently general to allow for participation to be correlated with latent variables that are both fixed and changing over time. Previous estimates have allowed for either fixed heterogeneity or changing heterogeneity, but not both simultaneously.[6]

As was mentioned earlier, it is also important to consider the possibility that participation in the program may be correlated with the pre-program transitory error component, υ_{t-1}. For instance, if program administrators "creamed" from the pool of eligible applicants, then $E(\upsilon_{t-1})<0$ for participants. This would create correlation between υ_{t-1} and participation in the program. Since υ_{t-1} is an element of the error component in the net impact estimation equations developed here [see equation (13)], this will create an additional source of inconsistency unless specifically controlled for. Once it is recognized that the Q variable has the same role in the pre-program years as the P variable does in the post-program years, it follows that equations (13) and (15) are sufficiently general for consistent net

[6] - Ashenfelter's estimates of equation (2) have eliminated any bias from non-random selection due to fixed latent variables. He also estimated equation (1) where Z_{it} included lagged values of Y. This method is sufficient to eliminate any bias created from non-random selection due to changing latent variables. It will not, however, eliminate fixed effects bias.

impact estimation in the case when $Q_{t-k} \neq 0$.[7] For instance, if we knew the value of Q_{t-k}, we could estimate the net impact of the program in period t from equation (16) below.[8]

$$(16)\ Y_t - Y_{t-1} - \hat{\rho}(\hat{Y}_{t-1} - Y_{t-2}) + \hat{\rho}(\hat{Q}_{t-1} - \hat{Q}_{t-2}) = \beta_o + \beta_2 {*} Z_t{*} + \beta_t P + (\upsilon_t - \upsilon_{t-1})$$

As in the case of net impact estimation, where it was necessary to first estimate β_t before estimating β_{t+n}, we must first estimate Q_{t-1}. In order to be able to do this, however, it is necessary to assume that $Q_{t-j} = 0$, for $j \geq 2$. To see this, consider the following pre-program earnings equation where Q_{t-j} is not constrained to be zero.

$$(17)\ Y_{t-1} - Y_{t-2} = \beta_o + \beta_2 {*} Z{*}_{t-1} + (Q_{t-1} - Q_{t-2}) + \rho(\hat{Y}_{t-2} - Q_{t-2} - Y_{t-3} + Q_{t-3}) + (\upsilon_{t-1} - \upsilon_{t-2})$$

Without assuming that $Q_{t-j}=0$ for $j \geq 2$, it is not possible to estimate Q_{t-1}. However by imposing this assumption, we are left with

$$(18)\ Y_{t-1} - Y_{t-2} = \beta_o + \beta_2 {*} Z{*}_{t-1} + \rho(\hat{Y}_{t-2} - Y_{t-3}) + (Q_{t-1} + \upsilon_{t-1} - \upsilon_{t-2})$$

The residual from equation (18) gives us an estimate of Q_{t-1} since $E(\upsilon_{t-1} - \upsilon_{t-3}) = 0$. By including $\hat{Q}_{t-1}$ in the net impact estimation equation, any bias that results from the program

[7] - It is assumed that ρ created a shift in the earning's function of participants after the program. The Q variable assumes the same role prior to the program.

[8] - This simply rewrites equation (15).

on the basis of the pre-program transitory error component has been eliminated. The most general version of equation (15) is given below.[9]

$$(19)\quad Y_{t+j+2}-Y_{t+j+1}-\hat{\rho}(\hat{Y}_{t+j+1}-Y_{t+j})+\hat{\rho}(R_{t+j+1}-R_{t+j}) =$$

$$\beta_o+\beta_2{}^*Z^*{}_{t+j+3}+R_{t+j+2}+\upsilon_{t+j+2}-\upsilon_{t+j+1}$$

where $R_l = \hat{\beta}_l \quad , \quad l \geq t$

$R_l = \hat{Q}_l \quad , \quad l = t-1$

$R_l = O \quad , \quad l < t-1$

Estimation of equation (19) will provide consistent estimates under very general conditions. Its validity, however, is dependent upon the following key assumptions: (1) ρ is constant over the entire time span being considered, (2) the earnings equations to be estimated are the same for the comparison group as for the participants except for the P and Q terms, (3) $Q_{t-j}=0$, for $j \geq 2$, and (4) both P and Q enter the earnings equations linearly. The validity of the first three assumptions is easily tested. It is possible to relax assumption (4) by doing separate net impact estimation for the different age/race/sex groups.[10]

---

[9] - Unlike β, it is necessary to estimate $\hat{Q}$ for non-participants as well as participants. However, a separate value of Q is estimated for each group.

[10]- The analysis could be done separately by whatever variables in X or Y seem likely to interact with P or Q in the earnings equations.

Perhaps the most serious drawback of this approach is that it is not appropriate for very young participants of employment and training programs. Since very few of the youngest participants would have any earnings for the three years prior to program participation, it is impossible to estimate ρ for this group. This is especially unfortunate since employment and training programs generally have a large number of young participants. Nevertheless, a thorough examination of this problem is beyond the scope of the current analysis. The empirical analysis will be done only for participants who were at least twenty-three years of age upon entering the program.

It seems likely, however, that the advantages of this approach more than outweigh its shortcomings. It provides a framework for consistent estimation of the dynamic impact of employment and training programs on participants' earnings. The model allows for unobserved variables which are both fixed (such as ability), and changing over time (such as health or local labor market conditions). These unobservables, as well as the transitory error component, may be correlated with participation in the program. Finally, the model is consistent with a very general specification of the earnings function. It is consistent with a model in which: (1) lagged values of earnings affect current earnings (above and beyond the extent to which they represent fixed effects) or (2) disturbances in earnings are correlated over time.[11]

IV. DATA AND ESTIMATION AGENDA

The Continuous Longitudinal Manpower Survey (CLMS) public use tapes will be used for the estimation suggested by Section III. The CLMS represents a major data development effort of the Department of Labor for evaluation of CETA funded employment and training programs. For a representative sample of 6,700 individuals that participated in the program during 1975 and 13,300 who participated during fiscal year 1976, the file contains: a four year record of labor force experience beginning one year prior to CETA enrollment, basic demographic characteristics, a history of public benefits received by the individual and/or the individual's family, family-related variables, and reported annual social security earnings for 1951-1977. For purposes of comparison, the March CPS (the annual demographic file) has been included with reported annual social security earnings and counts of quarters worked appended.

[11] - In general, it is not possible to empirically distinguish between these two models. See Appendix A for a discussion of the necessary conditions for identification.

Data collection and preparation have been carried out by the Bureau of Census. Westat, Inc. has been responsible for data management, preliminary analysis, and development of comparison groups for the CETA participants from the CPS. The comparison groups have been generated by "matching" CETA participants with their CPS counterparts on a variety of socio-demographic and past earnings variables, using different priorities in the matching process. These procedures has been used to generate six (6) comparison groups for the 1975 participants and three (3) comparison groups for the 1976 participants.

The results derived in Section III suggest a framework for the empirical work. It is first necessary to estimate ρ for each of several years before and after the participation year. Next, the data should be pooled in order to estimate the aggregate value of ρ.[12] This should be done separately for the participants, the comparison group, and the two combined. Chow tests should be performed within each group, and across the two groups in order to determine: (1) the appropriateness of pooling ρ, and (2) the appropriateness of the comparison group.[13] At this point it will be possible to re-estimate the pre-program earnings equations in order to estimate Q_{t-j}. By including a dummy variable for participation in the pre-program equations, we will be able to determine if $E(Q_{t-j})=0$, for $j \leq 2$. This entire process will be done with each comparson group in order to determine which comes closest to matching the participants.

[12] - See Appendix B for a consistent estimator of an aggregate ρ.

[13] - Note that it is not necessary that β_2 and ρ are the same across the two groups. The Chow tests can be done by constraining only the β_0 term to be the same for the participants and non-participants.

APPENDIX A

Identification of Alternative Models

For the purposes of this Appendix, we will be using the following specification of earnings determination[1]

$$(1) \quad Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t$$

$$(2) \quad \varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \upsilon_t$$

where υ_t is a random variable with an expected value of zero. We will consider the following cases:

(A) $\beta_2 = \rho_1 = \rho_2 = 0$

(B) $\beta_1 = \beta_2 = \rho_2 = 0$

(C) $\beta_2 = \rho_2 = 0$

(D) $\rho_1 = \rho_2 = 0$

(E) $\beta_1 = \beta_2 = 0$

First differencing of equation (1) for cases (A)-(C) gives the following equations:

$$(A1) \quad Y_t - Y_{t-1} = \beta_1(\hat{Y}_{t-1} - Y_{t-1}) + \upsilon_t - \upsilon_{t-1}$$

$$(B1) \quad Y_t - Y_{t-1} = \rho_1(\hat{Y}_{t-1} - Y_{t-1}) + \upsilon_t - \upsilon_{t-1}$$

$$(C1) \quad Y_t - Y_{t-1} = (\beta_1+\rho_1)(\hat{Y}_{t-1}-Y_{t-2}) - \beta_1\rho_1(Y_{t-2}-Y_{t-3})+\upsilon_t-\upsilon_{t-1}$$

[1] - The X and Z vectors used in the text have been excluded for the sake of expositional simplicity. In general, adding them to equation (1) does not affect any of the results.

Equations (A1) and (B1) are not empirically distinguishable[2]. If we are unable to reject the null hypothesis that $\beta_{\hat{1}}\rho_1 = 0$, it would then be necessary to generalize the model in the text to allow for a second order autoregressive error structure. This follows since it is impossible to distinguish (D), a second order autoregressive structure from (C) or (E). To see this, equation (1) is written below in first differences form for (D) and (E).

$$\text{(D1)} \quad Y_t - Y_{t-1} = \beta_1(\hat{Y}_{t-1} - Y_{t-2}) + \beta_2 (Y_{t-2}-Y_{t-3}) + \upsilon_t-\upsilon_{t-1}$$

$$\text{(E1)} \quad Y_t - Y_{t-1} = \rho_1(\hat{Y}_{t-1} - Y_{t-2}) + \rho_2(Y_{t-2}-Y_{t-3}) + \upsilon_t-\upsilon_{t-1}$$

So if we are able to accept the null hypothesis that $\beta_{\hat{1}}\rho_1$ is zero in equation (C1), then we can conclude that either: (1) the specification of the earnings function should include a lagged value of the dependent variable, or (2) the disturbances in the earnings function follow a first order autoregressive scheme. If we reject the null hypothesis, then we can conclude that either: (a) the previous two alternatives are simultaneously true, or (b) the specification of the earnings function should include two lagged values of the dependent variable, or (3) the disturbances in the earnings function follow a second order autoregressive scheme. All of these alternatives are consistent with heterogeneity that changes over time.

These findings can be generalized to handle more complicated specifications of equations (1) and (2).

[2] - If Z changes over time in a non-trend like fashion, it may be possible to identify (A) from (B) through Z. In general, the observable changes in Z (such as age or experience) will not be sufficient to enable us to distinguish between (A) and (B).

APPENDIX B

Consistent Estimation of Aggregate ρ

Continuing to use the notation developed in the text.

(1) $Y_t=\beta_1X+\beta_2Z_t+R_t+\varepsilon_i+\varepsilon_t+\rho(Y_{t-1}-\beta_1X-\beta_2Z_{t-1}-R_{t-1}-\varepsilon_i-\varepsilon_{t-1})+\upsilon_t$

Rewriting (1) by first differencing gives

(2) $$Y_t-Y_{t-1}=[\varepsilon_t-(1+\rho)\varepsilon_{t-1}+\rho\varepsilon_{t-2}] + [R_t - (1+\rho)R_{t-1}+\rho R_{t-2}] + \beta_2[Z_t-(1+p)Z_{t-1} + \rho Z_{t-2}] + \rho(\hat{Y}_{t-1}-Y_{t-2}) + \upsilon_t - \upsilon_{t-1}$$

Introducing some notional simplification

(3) $Y_t-Y_{t-1} = \beta_t + R_t + \beta_2Z_t{*} + \rho(\hat{Y}_{t-1} - Y_{t-2}) + \upsilon_t-\upsilon_{t-1}$

where β_t is an intercept term common to the comparison group and the participants, and R_t is a shift in the intercept term for participants. Both of these terms are unique to period t. This suggests that if a pooled version of (3) were to be estimated, it would be necessary to allow for a separate intercept term for each year, as well as for each of the two groups. The nature of the error structure in equation (3) makes it impossible to obtain consistent estimates of an aggregate ρ by simply pooling (3) over time. To see this, consider equations (4)-(7) below. Here the specific years that we are concerned with have been used in the subscripting.

(4) $Y_{77}-Y_{76} = \beta_{77}+R_{77}+\beta_2Z^*_{77}+\rho(\hat{Y}_{76}-Y_{75}) + \upsilon_{77}-\upsilon_{76}$

(5) $Y_{76}-Y_{75} = \beta_{76}+R_{76}+\beta_2Z^*_{76}+\rho(\hat{Y}_{75}-Y_{74}) + \upsilon_{76}-\upsilon_{75}$

(6) $Y_{75}-Y_{74} = \beta_{75}+R_{75}+\beta_2Z^*_{75}+\rho(\hat{Y}_{74}-Y_{73}) + \upsilon_{75}-\upsilon_{74}$

(7) $Y_{74}-Y_{73} = \beta_{74}+R_{74}+\beta_2Z^*_{74}+\rho\ (\hat{Y}_{73}-Y_{72}) + \upsilon_{74}-\upsilon_{73}$

Note that it is not possible to pool any two consecutive equations since this would create negative correlation between the error term and $\rho(Y_n - Y_{n-1})$. In fact, no linear combination of equations (4)-(7) can completely eliminate this negative correlation. It is, therefore, necessary to estimate equations (4)-(7) as a system of equations, imposing constraints across the equations to obtain aggregate estimates of ρ and β_2. Ordinary least squares estimates will be consistent but inefficient. Zeller Seemingly Unrelated Regression techniques will produce efficient as well as consistent estimates.

REFERENCES

1. Ameniya, Takeshi, "The Estimation of a Simultaneous Equation Generalized Probit Model," Econometrica, pgs. 1193-1205, Vol. 46, No. 5, September 1978.

2. Ameniya, Takeshi, "Regression Analysis when the Dependent Variable is Truncated Normal," Econometrica, pgs. 997-1016, Vol. 41, No. 6, November 1973.

3. Ashenfelter, Orley, "Estimating the Effect of Training Programs on Earnings," Review of Economics and Statistics, pgs. 47-57, Vol. 60, February 1978.

4. Avery, Robert and Watts, Harold W., "The Application of an Error Components Model to Experimental Panel Data," in The New Jersey Income-Maintenance Experiment; Harold W. Watts and Albert Rees (eds.), Academic Press, Inc., New York, 1977.

5. Barnes, William, "Target Groups," in CETA: An Analysis of the Issues, Special Report No. 23, National Commission for Employment Policy, May 1978.

6. Barnow, Burt S.; Cain, Glen G.; and Goldberger, Arthur S., "Issues in The Analysis of Selection Bias," unpublished paper, August 1978.

7. Chamberlain, Gary, "Heterogeneity, Omitted Variable Bias, and Duration Dependence," Discussion Paper Number 691, Harvard Institute of Economic Research, March 1979.

8. Chamberlain, Gary, "On the Use of Panel Data," mimeo, October 1978.

9. Duncan, Greg J. and Hoffman, Saul D., "Dynamics of Wage Change," mimeo, Institute for Social Science Research, The University of Michigan, July 1980.

10. Duncan, Gregory and Leigh, Duane, "Wage Determination in the Union Sectors: A Sample Selectivity Approach," Industrial and Labor Relations Review, pgs. 24-34, Vol. 34, No. 1, October 1980.

11. Ehrenberg, Ronald G. and Schumann, Paul L., "Compensating Wage Differentials for Mandatory Overtime?" mimeo, Cornell University, May 1980.

12. Ellwood, David, "Teenage Unemployment: Permanent Scars or Temporary Blemishes," mimeo, undated.

13. Farebrother, R.W., "A Remark on the Wu Test," Econometrica, pgs. 475-477, Vol. 44, No. 3, March 1976.

14. Goldberger, Arthur S., "Abnormal Selection Bias," Social Systems Research Institute Working Paper #8006, University of Wisconsin - Madison, May 1980.

15. Griliches, Zvi, "Estimating the Returns to Schooling: Some Econometric Problems," Econometrica, pgs. 1-22, Vol. 45, No. 1, January 1977.

16. Gronau, Reuben, "Wage Comparisons - A Selectivity Bias," Journal of Political Economy, pgs. 1119-1143, Vol. 82, No. 6, 1974.

17. Hausman, Jerry A. and Wise, David A., "Social Experimentation, Truncated Distributions, and Efficient Estimation," Econometrica, pgs. 919-938, Vol. 45, No. 4, May 1977.

18. Heckman, James J., "Sample Selection Bias as a Specification Error," Econometrica, pgs. 153-161, Vol. 47, No. 1, January 1979.

19. Heckman, James J., "Dummy Endogenous Variables in a Simultaneous Equation System," Econometrica, pgs. 931-959, Vol. 46, No. 6, July 1978.

20. Heckman, James J., "Heterogeneity and State Dependence in Dynamic Models of Labor Supply," Report 7837, Center for Mathematical Studies in Business and Economics, University of Chicago, May 1978.

21. Heckman, James J., "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependant Variables and a Simple Estimator for Such Models," Annals of Economic and Social Measurement, pgs. 475-492, Vol. 5, No. 4, Fall 1976.

22. Heckman, James, "Shadow Prices, Market Wages, and Labor Supply," Econometrica, pgs. 679-694, Vol. 42,

23. Kiefer, Nicholas M., "Population Heterogeneity and Inference from Panel Data on the Effects of Vocational Education," Journal of Political Economy, pgs. 5213-5226, Vol. 87, No. 5, pt. 2, October 1979.

24. Kiefer, Nicholas M., "Federally Subsidized Occupational Training and the Employment and Earnings of Male Trainees," Journal of Econometrics, pgs. 111-125, Vol. 8, 1978.

25. Lee, Lung-Fei, "Unionisn and Wage Rates: A Simultaneous Equation Model with Qualitative and Limited Dependent Variables," International Economic Review, pgs. 415-433, Vol. 19, No. 2, June 1978.

26. Lee, Lung-Fei: Maddala, G.S., and Trost, R.P., "Asymptotic Covariance Matrices of Two-Stage Probit and Two-State Tobit Methods for Simultaneous Equations Models with Selectivity," Econometrica, pgs. 491-503, Vol. 48, No. 2, March 1980.

27. Lewis, H. Gregg, "Comments on Selectivity Biases in Wage Comparisons," Journal of Political Economy, pgs. 1145-1155, Vol. 82, No. 6, November 1974.

28. Lillard, Lee and Willis, Robert, "Dynamic Aspects of Earnings Mobility," Econometrica, pgs. 985-1012, Vol. 46, No. 5, September 1978.

29. Maddala, G.S., "Simultaneous Probit and Tobit Models with Latent Structures," Discussion Paper #15, Center for Econometrics and Decision Sciences, University of Florida, April 1980.

30. Maddala, G.S., "The Use of Variance Models in Pooling Cross Section and Time Series Data," Econometrica, pgs. 341-358, Vol. 39, No. 2, March 1971.

31. Maddala, G.S. and Lee, Lung-Fei, "Recursive Models with Qualitative Endogenous Variables," Annals of Economic and Social Measurement, pgs. 525-545, Vol. 5, No. 4, Fall 1976.

32. Mundlak, Yair, "On the Pooling of Time Series and Cross Section Data," Econometrica, pgs. 69-84, Vol. 46, No. 1, January 1978.

33. Nerlove, Marc, "A Note on Error Components Models," Econometrica, pgs. 383-396, Vol. 39, No. 2, March 1971.

34. Olsen, Randall J., "Tests for the Presence of Selectivity Bias and Their Relation to Specifications of Functional Form and Error Distribution," Working Paper #812, Institution for Social and Policy Studies, Yale University, November, 1978.

35. Poirier, Dale J., "Partial Observability in Bivariate Probit Models," Journal of Econometrics, pgs. 209-217, Vol. 12, February 1980.

36. Theil, Henri, Principles of Econometrics, John Wiley & Sons, Inc.; New York, 1971.

37. Tobin, James, "Estimation of Relationships for Limited Dependent Variables," Econometrica, pgs. 24-36, Vol. 26, No. 1, 1958.

38. Wallace, T.D., and Hussain, Ashiq, "The Use of Error Components Models in Combining Cross Section with Time Series Data," Econometrica, pgs. 55-72, Vol. 37, No. 1, January 1969.

39. Westat, Inc., "The Impact of CETA on Particpant Earnings," Working Paper #2, June 1980.

40. Westat, Inc., "The Impact of CETA on Participant Earnings," Working Paper #1, January 1980.

41. Willis, Robert J. and Rosen, Sherwin, "Education and Self-Selection," Journal of Political Economy, pgs. S7-S36, Vol. 87, No. 5, pt. 2, October 1979.

42. Wu, De-Min, "Alternative Tests of Independence Between Stochastic Regressors and Disturbances," Econometrica, pgs. 733-750, Vol. 41, No. 4, July 1973.