A MULTIVARIATE MODEL
OF LABOR SUPPLY:
METHODOLOGY FOR ESTIMATION

PREPARED UNDER A GRANT FROM THE DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

GIORA HANOCH

R-1869-HEW
SEPTEMBER 1976
The research reported herein was performed pursuant to Grant No. 016A-7601-P2021 from the Department of Health, Education, and Welfare, Washington, D.C. The opinions and conclusions expressed herein are solely those of the author, and should not be construed as representing the opinions or policy of any agency of the United States Government.
A MULTIVARIATE MODEL
OF LABOR SUPPLY:
METHODOLOGY FOR ESTIMATION

PREPARED UNDER A GRANT FROM THE DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

GIORA HANOCH

R-1869-HEW
SEPTEMBER 1976

Rand
SANTA MONICA CA 90406
This is one of a series of reports from Rand's Labor and Population Studies Program dealing with issues relating to labor supply. The research was performed under Grant No. 016A-7601-F2021 from the Office of the Assistant Secretary for Planning and Evaluation of the Department of Health, Education, and Welfare.

This report should be viewed as the second in a trilogy of reports by the same author. The first (R-1787-HEW) is a theoretical analysis of the multidimensional nature of time allocation. The present report deals with methodological issues confronted in estimating a multivariate labor supply model that incorporates the theory outlined in the first report. The final report (R-2067-HEW) will summarize empirical results obtained when the model was estimated, using the methodology developed here.

An earlier draft of this report, entitled "Theory and Estimation of a Complete Labor Supply Model" (September 1975), included some additional topics, such as estimation of time and money costs, which will be dealt with by the author separately.

Giora Hanoch was a visiting professor at the University of California at Los Angeles during 1974-1975. He is permanently at the Hebrew University of Jerusalem, and is a consultant to the Labor and Population Studies Program at The Rand Corporation.
SUMMARY

The analysis of labor supply has consumed much effort and time of theoretical and empirical economists, and has given rise to many complex conceptual and statistical problems.

A companion report by the present author, Hours and Weeks in the Theory of Labor Supply, R-1787-HEW, August 1976, analyzes several aspects of labor supply that have been overlooked by earlier studies. The present report treats some of these aspects econometrically, providing a model and a feasible method of estimation for these problems, in conjunction with others recently examined elsewhere. In a forthcoming companion report, Work and Wages of Married Women: Estimates of a Multivariate Model, R-2067-HEW, empirical results based on the present model and methods are presented and analyzed with respect to a sample of married women in the 1967 Survey of Economic Opportunities (SEO). These results generally tend to show that neglecting to take any of these considerations into account may cause serious biases and misinterpretations. On the other hand, certain other considerations are ignored in this analysis and require future attention.

This study considers econometrically the following six major theoretical and statistical problems peculiar to labor supply. The first two problems were treated empirically in other recent studies, whereas the remaining four are believed to be treated econometrically for the first time here.

1. The simultaneity bias in regression estimates of labor supply, resulting because the wage rate is endogenous in cross sections.

2. The participation selectivity bias, arising because market wage offers are measured (by actual wages) only for workers. Since the criterion for selecting the working sample is the wage-rate comparison with the individual's own reservation wage, this criterion is endogenous and induces a selectivity bias.

3. The simultaneous joint determination of the supply of annual hours A and annual weeks K of work. Most previous empirical
studies have treated hours and weeks of work as alternative measures of the same dependent supply variable, whereas theoretical discussions have usually regarded the corresponding two types of leisure as perfect substitutes. The present study, however, considers them distinct but jointly determined endogenous variables.

4. The discontinuity of supply at the reservation wage, or the existence of minimum supply quantities. The theoretical study shows that, even in the absence of any fixed costs of participation, one of the variables A or K (and normally K) will be discontinuous, with minimum positive amounts supplied at the reservation wage. But the presence of any fixed time or money costs associated with work induces discontinuity of supply of both A and K. Empirical models that ignore this problem tend to produce large positive biases in estimates of the elasticity of labor supply with respect to wages.

5. The corner solution in terms of full-year work. The actual number of weeks worked in one year cannot exceed a physical limit (taken as 52), but the desired number of weeks, as determined by underlying preferences, may exceed this limit. This restriction causes a truncation of the supply of weeks worked, as well as a discontinuity of responses (slopes) of annual hours supplied, through the effects of the weeks' truncation on the total demand for composite leisure.

6. The survey-week selectivity bias. In data sources such as the SEO and the census, wage and hours-of-work information is available only for individuals who work during the week prior to their survey interviews. For individuals intending to work K weeks in the current year, therefore, the probability of working in a particular week is equal, on the average, to K/52. Since this selection probability is proportional to the endogenous variable K, a bias is introduced into labor supply estimates.

Section I is an introduction describing the problems treated in the study as well as additional problems ignored by it.
Section II presents the variables and the basic model, consisting of a market wage equation, a reservation wage equation, and two supply equations for \( A \) and \( K \). From these are derived additional relations, such as the probability-of-participation equation and its corresponding probit index, a "pseudo reservation wage equation" (in which the unobserved reservation wage is replaced by the market wage, in conjunction with the index of participation), and minimum-supply equations for hours and weeks.

Section III transforms the general model into an operational model, which is applicable to the working sample. The latter model contains only observable variables and equations modified to account for the two types of selectivity bias.

Finally, Sec. IV presents an iterative strategy for estimating the model, proposed here and applied in the empirical study (R-2067-HEW). This section also discusses general estimation problems in greater detail. The final stage of the proposed procedure applies weighted three-stage-least-squares estimation to alternative simultaneous linear systems of modified equations. A few alternative (presumably more efficient) estimation procedures are outlined but not recommended, except for an alternative estimate of the reservation wage equation. The appendix derives the magnitude of the survey-week bias and the method of eliminating it by a weighting scheme.

The methodology presented here is not aimed at achieving maximum efficiency of estimation or at eliminating all sources of specification and estimation biases; it is rather informal and intuitive, preferring methods that are less costly to execute and easier to understand.
ACKNOWLEDGMENTS

I am indebted to Finis Welch, who supported and encouraged this study and discussed it with me regularly, providing many helpful suggestions and comments. I am also grateful to Michael Ward and Zvi Griliches, who commented extensively on an earlier draft ("Theory and Estimation of a Complete Labor Supply Model," September 1975) and contributed many helpful ideas for this revision. I had fruitful discussions with John Cogan, James P. Smith, and many others. Richard Buddin, who assisted throughout with technical work, programming, and computations, helped me with discussions and comments as well. I thank Linda Colbert for a most helpful editing.

 Needless to say, any errors are my own responsibility.
CONTENTS

PREFACE ................................................................. iii
SUMMARY ............................................................... v
ACKNOWLEDGMENTS .................................................... ix

Section
I. INTRODUCTION ..................................................... 1

II. VARIABLES AND THE BASIC MODEL ......................... 8
    Endogenous Variables .................................. 8
    Exogenous Observable Variables .................. 8
    Latent Unobserved (Random) Variables .......... 9
    The Basic Model ........................................ 10

III. BASIC MODEL CORRECTED FOR SELECTIVITY BIASES .... 14
    Case 1: Wage and Hours Data Available
            for All Participants .......................... 16
    Case 2: Wage and Hours Data Available Only
            for Survey-Week Workers .................... 19

IV. STRATEGY FOR ESTIMATING THE BASIC MODEL ........... 22
    Introduction and General Comments .............. 22
    Estimating the Probability of Participation:
        Probit Analysis .................................. 23
        Preliminary Reduced-Form Estimates of Weighting
            Scheme and Market Wage Equation ............ 26
        Tobit Analysis of Supply of Weeks .......... 27
        Modifying the Pseudo Reservation Wage Equation .... 30
        Linear-System Estimates .................... 31
        Alternative Estimate of Reservation Wage Equation .. 35
        Postestimation Analysis .................... 37

Appendix
SURVEY WEEK SELECTIVITY BIAS AND ITS CORRECTION
BY WEIGHTING .................................................. 41

REFERENCES .......................................................... 47
I. INTRODUCTION

The individual labor supply relation is but one, albeit important, of many relations determining general equilibrium of individual units and economic aggregates. In elementary textbooks, it appears to be a simple reservation-demand equation. Nevertheless, theoretical and empirical economists have spent much time and effort analyzing this relation. The more thoroughly they examined it, the less stable and more complex it seemed, with new conceptual and statistical problems surfacing at every stage.

In a previous theoretical report (R-1787-HEW), I have analyzed several aspects of labor supply that were previously overlooked. The purpose of the present report is to treat some of these aspects econometrically—that is, to provide a model and a feasible method of estimation that account for these problems, in conjunction with others recently accounted for elsewhere. Using the methods described here, a forthcoming report (R-2067-HEW) presents and analyzes empirical results of this model with respect to a sample of married women in the 1967 Survey of Economic Opportunities (SEO). These results generally tend to show that all of these considerations are important quantitatively, and that neglecting to take any of them into account may cause serious biases and misinterpretations.

I do not mean to imply that other aspects of labor supply, which are assumed away or ignored by the present model, are negligible or unimportant, or that the model outlined here is free from biases and potentially misleading conclusions. I hope that additional important aspects of the labor supply relation can be treated in the future, possibly generalizing some of the methods and concepts presented here.

These problems are particularly significant with respect to women's labor supply. Men tend to participate in the labor force in much larger proportions, over more extended and uninterrupted periods, and to work full time when employed. Variations in male labor market quantities are thus smaller and determined more by labor demand conditions and restrictions than by individual supply, as compared to women, especially married women.
The theoretical study treated the problem of family labor supply, where substitution and complementarity of the husband's and wife's time may affect each family member's own labor supply behavior. For convenience of presentation, the present analysis applies the model to individuals, ignoring the joint allocation of time in the family. But it could easily be extended to a family, with minor modifications, as shown in the theoretical study. For example, it allows husband's wages and earnings to affect the wife's supply of labor, but ignores, at this stage, the endogenous nature of these variables.

Of the theoretical and statistical problems peculiar to labor supply that the present model does recognize and resolve econometrically, the following six major elements may be cited. The first two of these were treated empirically in other recent studies,† whereas the last four are related to problems discussed in the theoretical study, and are believed to be treated econometrically for the first time here.‡

1. *The simultaneity bias in regression estimates of labor supply.* This bias occurs because the wage rate is an endogenous variable in cross-sectional data—that is, because some individual market productivity factors, unaccounted for by the included explanatory variables, tend to be correlated with the individual labor supply residuals, which in turn are associated with individual productivity of nonmarket time as well as individual preferences.

2. *The participation selectivity bias.* This bias occurs because market wage offers are measured only for workers (where they are assumed to equal the actual wage paid), but not for nonparticipants (who refuse these wage offers). Since the criterion for selecting the working sample is the wage-rate

---

* The simultaneity bias was treated, for example, by Heckman (1974a) and Hall. The selectivity bias was analyzed by Gronau (1973), Lewis, and Heckman (1974a and R-1984).

† Recent works by Heckman (R-1984) and Cogan (written after the draft of the present report was completed) also treat one of these elements—namely, the discontinuity of supply due to fixed costs (see point 4, next page).
comparison with the individual's own reservation wage, this
criterion is endogenous and induces a selectivity bias, with
either workers' wages higher than expected (given the measured
variables), or reservation wages lower than expected, or both.*

3. The simultaneous joint determination of the supply of annual
hours A and annual weeks K of work—that is, the time supplied
has more than one dimension and individuals are not indifferent
to variations in the allocation of any given total amount of
time supplied with respect to the corresponding dimensions such
as hours and weeks. Thus, the two dimensions correspond to
demand for two types of leisure—workweek leisure and nonwork-
week leisure—which are jointly determined in accordance with
individual preferences and market restrictions; and which de-
termine, in turn, the supply of hours and weeks. Most previous
empirical studies have treated hours and weeks of work as al-
ternative measures of the same dependent supply variable, whereas
theoretical discussions have usually regarded the two types of
leisure associated with A and K as perfect substitutes.

4. The discontinuity of supply at the reservation wage, or the
existence of minimum supply quantities. The theoretical study
shows that, even in the absence of any fixed costs of partici-
pation, one of the variables A or K (and normally K) will be
discontinuous, with minimum positive amounts supplied at the
reservation wage (i.e., at the point of entry into the market).
But the presence of any fixed time or money costs associated
with work (except if costs are proportional to hours of work,
when they can be viewed as a reduction in the wage rate) in-
duces a gap between the price of time at home (if not partici-
pating) and the reservation wage; and, correspondingly, induces
a discontinuity of supply of both A and K at the reservation
wage W_o. Although theoretical analyses of labor supply recognize

*However, a positive correlation between wage and reservation wage
residuals (due to a positive correlation between market and home pro-
ductivity) may decrease or eliminate this bias. See Eq. (15) below,
p. 17.
this problem, it has been ignored in empirical models, which therefore tend to produce large positive biases in estimates of the wage elasticity of labor supply.*

5. The corner solution in terms of full-year work. The actual number of weeks worked in one year K cannot exceed a physical limit (taken as exactly 52 for simplicity), but the desired number of weeks, as determined by underlying preferences (assumed to depend on the two types of leisure and to be well defined and continuous for a domain exceeding available time), may exceed this limit. This restriction causes a corner solution at 52 in the supply of weeks worked, but it may also cause the responses (slopes) of annual hours supplied to be discontinuous, through the effects of the corner solution in weeks on the total demand for composite leisure.† Quantitatively, about 40 percent of the women who participate in the labor force are full-year workers, so that the effect of this constraint on their labor supply may be quite significant.

6. The survey-week selectivity bias. In data sources such as the SEO and the census, wage and hours-of-work information is available only for individuals who work (any positive number of hours) during the week prior to their survey interviews. For individuals intending to work K weeks in the current year, therefore, the probability of working in a particular week is equal, on the average, to K/52 (ignoring seasonal effects, which are assumed to average out among individuals). Since the selection probability is proportional to the endogenous variable K, a bias is introduced into labor supply estimates based on this type of data. The present analysis shows the magnitude of this bias and suggests statistical methods to eliminate it.

A simple examination of the data shows the validity of this argument and its quantitative importance. For example, in an SEO sample of white married women, out of 3044 women

* For example, in Heckman (1974a and b). See also Hanoch (R-1787), pp. 18-20.
† See Hanoch (R-1787), Sec. III.
participating in the labor force during 1966 (i.e., with \( K \) positive), only 2064 (67.8 percent) worked during the survey week (in March 1967). Since the mean number of weeks worked by participants was 35.3, the proportion working in a given week is predicted to be \( 35.3 \div 52 \), or 67.9 percent! On the other hand, the biased mean of \( K \), as measured for survey-week workers, is 41.6 weeks, compared with the mean of 22.2 weeks for the excluded group (participants in 1966 who did not work in the survey week).

The model presented here ignores or abstracts from certain important elements, in addition to the forementioned individual/family distinction. First, in choosing annual hours and annual weeks as the time dimensions, and one year as the basic decision period, the model abstracts from other time dimensions that may be relevant, such as weekly days of work (and, correspondingly, daily hours of work), or a planning horizon longer than one year (which could explain variations of labor supply from year to year and over the life cycle). Formally, one could easily extend the model to additional dimensions and a longer horizon. However, common limitations of data on individuals, particularly from the large cross-sectional sources that this model is designed to use, dictate the present specification.

Second, the problem of daily hours is strongly related to another consideration that is ignored by the present model—namely, variation in the hourly wage rate as a function of hours, such as a lower wage for part-time work or a higher wage for overtime. The model assumes that the hourly wage rate \( W \) offered to an individual is independent of the number of weekly hours of work (as well as of weeks worked annually). In reality, employers may have strong preferences regarding the length of the workday, due to its effect on productivity and on work organization and scheduling. On the other hand, individuals may change employers, thereby varying their daily hours, perhaps without significantly affecting their hourly wage. Nevertheless, genuine changes in productivity with

\*See Hanoch (R-1787), Sec. VI.
number of hours worked daily and lack of sufficient variety among em-
ployers will generally limit the employees' ability to substitute, and
will make the maximum achievable wage rate depend on daily hours, as
well as on total supply of hours.

The allocation of weekly hours of work among days is thus largely
determined by demand constraints, the effects of which are difficult
to separate from pure supply effects related to preferences and produc-
tivity at home. In view of this, the present specification dealing
with weekly rather than daily hours seems preferable, since the depen-
dence of wage offers on the weekly total \( H \) is expected to be weaker
than on daily hours.

Third, other equalizing differences in wages are also ignored
here. These may be connected with work conditions, opportunities for
accumulation of human capital (on-the-job training), and nonpecuniary
costs and benefits. It is hoped that such elements can be treated in
a future extended model, if related data are available.

Finally, the role of income taxes is also ignored here. Clearly,
if marginal and average tax rates differ among individuals, and if labor
supply decisions depend on after-tax wages and incomes, an analysis
based on gross (before-tax) wages and income is deficient. At present,
data on effective rates of individual tax payments are scarce. In
addition, the modifications required in the model to account for tax
effects are complex, given the existing current structure of income
taxes, with their variable degree of progressivity, credits, deductions,
and numerous loopholes.

Section II presents the variables and the basic model, consisting
of a market wage equation, a reservation wage equation, and two supply
equations for \( A \) and \( K \). From these are derived additional relations such
as the probability-of-participation equation and its corresponding probit
index, a "pseudo reservation wage equation" (in which the unobserved
reservation wage is replaced by the market wage, in conjunction with the
index of participation), and minimum-supply equations for hours and weeks.

Section III transforms the general model into an operational model,
which is applicable to the working sample. The latter model contains
only observable variables and equations modified to account for the two
types of selectivity bias.
Finally, Sec. IV presents an iterative strategy for estimating the model, proposed here and applied in the empirical study (R-2067). This section also discusses estimation problems in greater detail. The final stage of the proposed procedure applies a weighted three-stage-least-squares estimation to alternative simultaneous linear systems of modified equations. A few alternative (presumably more efficient) estimation procedures are outlined but not recommended, except for an alternative estimate of the reservation wage equation, using predicted wages from the system estimates. The appendix derives the magnitude of the survey-week bias, and the method of eliminating it by a weighting scheme.

The methodology presented here is not aimed at achieving maximum efficiency of estimation or at eliminating all sources of specification and estimation biases; it is rather informal and intuitive, preferring methods that are less costly to execute and easier to understand.
II. VARIABLES AND THE BASIC MODEL

The following definitions and assumptions are adopted, with respect to individuals.

**ENDOGENOUS VARIABLES**

\[ DP = \text{A dichotomous dummy variable determining participation in the labor market: } DP = 1 \text{ for participants and 0 for nonparticipants.} \]

\[ A = \text{Annual hours of work by participants.} \]

\[ K = \text{Annual weeks of work by participants}--\text{i.e., the number of weeks in which a positive number of hours were worked.} \]

\[ W = \text{Market hourly wage rate. } W \text{ is observed for workers, but not for nonparticipants. } W \text{ is assumed positive.} \]

\[ w = \ln W \text{ (natural logarithm).} \]

\[ W_0 = \text{Reservation wage}--\text{i.e., the minimum hourly wage that induces participation. } W_0 \text{ is unobserved, but positive.} \]

\[ w_0 = \ln W_0. \]

It is assumed that decisions on \( DP, A, \) and \( K \) are made jointly and simultaneously, on a one-year basis. The condition for participation (\( DP = 1 \)) is equivalent to the condition \( W \geq W_0, \) or \( w \geq w_0. \)

**EXOGENOUS OBSERVABLE VARIABLES**

\[ x = \text{A set of variables that influence market wage offers}--\text{i.e., variables related to market productivity and labor demand conditions.} \]

\[ z = \text{A set of variables that influence individual preferences between nonmarket time (leisure) and other consumption goods}--\text{i.e., variables determining } A \text{ and } K, \text{ conditional on participation and wage rate. These include personal tastes, productivity at home, and the availability of time and money in the family.} \]

\[ y = \text{A set of variables that influence the reservation wage } W_0. \]

In general, \( y \) includes all \( z \) variables, as well as certain variables that may affect time and money costs of participation, but have a negligible effect on quantities of labor supplied by participants.
It is assumed (and required for identification of the model) that some \( x \) variables are not in \( z \) or \( y \), and some \( z \) and \( y \) variables are not in \( x \). By assumption, \( x \) and \( y \) (and therefore \( z \)) are observable for everyone, including nonparticipants.

**Latent Unobserved (Random) Variables**

The set \( \mathbf{u}^o = (u^o_0, u^o_1, u^o_2, u^o_3, u^o_4) \) denotes additive latent components of the following five endogenous variables, respectively:

\[
(\omega_0, \nu^o_0 - \nu, w, A, K).
\]

Thus \( u^o_0 = u^o_1 + u^o_2 \). The four random variables \( u = (u^o_1, u^o_2, u^o_3, u^o_4) \) are assumed to have a nondegenerate joint-normal distribution with zero means: \( \mathbf{E}u = 0 \); and a variance matrix \( \mathbf{Eu}u' = \mathbf{\Sigma} \).

\( \mathbf{\Sigma} \) is generally nondiagonal and nonsingular.

A crucial assumption made here, as in all econometric models, is that \( u \) are distributed independently of the exogenous variables \( (x, y) \), therefore \( \mathbf{E}(x, y)u' = 0 \) (zero correlations); and that \( u \) are independent among individuals. Similar assumptions apply to \( u^o \), since \( u^o_0 = u^o_1 + u^o_2 \).

It is also assumed that \( \mathbf{\Sigma} \) is constant among individuals (homo-skedasticity). Denote also \( \mathbf{\Sigma}^o = \mathbf{Eu}^o u'^o = \{\sigma_{ij}^o \}; i, j = 0, \ldots, 4 \} \) and \( \sigma_{ii}^o = \sigma_{ii}^2 \). Thus, \( \sigma_{i0}^o = \sigma_{i1}^2 + \sigma_{i2}^2 + 2\sigma_{i12}^2 \).

The latent variables defined here combine different types of components that are not distinguished in a cross-sectional model like this, but are conceptually distinguishable: (1) permanent individual components, associated with various unobserved permanent exogenous variables, such as ability, personal tastes, motivation; (2) transitory components, which vary for a given individual from year to year (or from week to week if related to a variable measured on a weekly basis, such as the wage rate in the SEO sample); (3) measurement errors in observed endogenous variables \( (W, A, K, \text{and DP}) \); and (4) errors in specification of functional form, such as nonlinearity components, assumed either negligible or orthogonal to the right-side variables.

*It is not assumed, however, that \( \mathbf{\Sigma} \) is diagonal--i.e., that residuals are uncorrelated across equations. The homoskedasticity assumption is made as a convenient approximation only, and may be modified without greatly affecting the model. See Eq. (30) below, p. 35.*
THE BASIC MODEL

The following four equations summarize the basic model, applicable to every individual, including nonparticipants, in terms of the variables defined above and the constant parameters $\alpha$, $\beta$, $\gamma_1$, $\gamma_2$, $\delta_1$, $\delta_2$, and $\mu$.

Market wage equation:

$$w = \alpha'x + u_2.$$  \hspace{1cm} (1)

Reservation wage equation:

$$w_o = \beta'y + u_0.$$ \hspace{1cm} (2)

Annual hours supply equation:

$$A = \begin{cases} 
\gamma_1'z + \delta_1 w + u_3 & \text{if } w \geq w_o \\
0 & \text{if } w < w_o
\end{cases}$$

or

$$A = (\gamma_1'z + \delta_1 w + u_3) \cdot DP.$$ \hspace{1cm} (3)

Annual weeks supply equation:

$$K = \begin{cases} 
\gamma_2'z + \delta_2 w + u_4 & \text{if } w \geq w_o \\
0 & \text{if } w < w_o
\end{cases}$$

or

$$K = (\gamma_2'z + \delta_2 w + u_4) \cdot DP.$$ \hspace{1cm} (4)
Additional inequality constraints are imposed on the variables: $W > 0$, $W_e > 0$, $A > 0$, and $0 < K \leq 52$, for participants. The positivity requirements on $A$ and $K$ are ignored here, assuming that the negative correlation between the reservation wage and the supply residuals is sufficiently high to make negligible the normal-distribution tail probabilities corresponding to negative supply quantities. (This problem could be eliminated if $A$ and $K$ were defined as logarithms of the corresponding quantities, but such definitions would imply inferior functional-form specifications of the supply equations.)

The possibility of a corner solution corresponding to full-time work in terms of weeks (owing to the restriction $K \leq 52$) is important. To simplify the discussion, this restriction is ignored here, but is fully recognized and treated in Sec. IV below, pp. 27-30.

An equation determining the probability of participation $p$ may be specified by combining Eqs. (1) and (2). Denote the unit-normal cumulative probability by

$$P(\chi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\chi} e^{-t^2/2} \, dt,$$

and define the participation index as the following exogenous variable:

$$I = \frac{1}{\sigma_1^2} (\mathbf{a}' \mathbf{x} - \mathbf{b}' \mathbf{y}),$$

(5)

where $\sigma_1^2$ is the variance of $u_1$. Then the probability of participation is

$$\text{prob} \{w \geq w_o\} = \text{prob} \{u_o - u_2 \leq \mathbf{a}' \mathbf{x} - \mathbf{b}' \mathbf{y}\} = \text{prob} \{u_1 \leq \sigma_1 I\} = P(I),$$

since $(u_1/\sigma_1)$ is unit-normal. We thus get the

*Probability-of-participation equation:*

$$p = E(DP) = P(I) = P \left( \frac{1}{\sigma_1} \left( \mathbf{a}' \mathbf{x} - \mathbf{b}' \mathbf{y} \right) \right).$$

(6)
Since the reservation wage is unobservable, a "pseudo reservation wage equation" may be specified in terms of the market wage and the participation index $I$ by substituting from the identity equation (5) into Eq. (1).

**Pseudo reservation wage equation:**

$$w = \beta' \gamma + \sigma_1 I + u_2 .$$  \hspace{1cm} (7)

Equation (2) of the basic model may thus be replaced by Eq. (6) or (7) by substituting the participation probability $p$ or the participation index $I = F^{-1}(p)$, respectively, for the reservation wage variable $w_o$.

The minimum supply quantities for participants, $A_o$ and $K_o$ (assumed nonnegative), are derived by substituting $w_o$ from Eq. (2) into Eqs. (3) and (4):

**Minimum-supply equations:**

$$A_o = \gamma_1' z + \delta_1 w_o + u_3 = \gamma_1' z + \delta_1 \beta' \gamma + (\delta_1 u_o + u_3) ,$$  \hspace{1cm} (8)

$$K_o = \gamma_2' z + \delta_2 w_o + u_4 = \gamma_2' z + \delta_2 \beta' \gamma + (\delta_2 u_o + u_4) .$$

If $\delta_1 > 0$ and $\delta_2 > 0$, the condition for participation ($DP = 1$, or $w \geq w_o$) is seen to be equivalent to each of the conditions $A \geq A_o$ and $K \geq K_o$ by comparing Eqs. (3) and (4) with Eq. (8).

With no time or money costs of participation, either $A_o$ or $K_o$ is zero.* However, any fixed costs associated with market work (such as travel time to work and search costs) imply that $A_o$ and $K_o$ are both positive, manifesting discontinuity of both supply curves at the point of minimum participation. (Simulations with various functional forms for the underlying utility function have shown that relatively small fixed costs tend to induce large magnitudes of $A_o$ and $K_o$.)† The present

---

*As Hanoch (R-1787) shows, $K_o$ is more likely to be nonzero if both types of leisure are normal goods.

†See Hanoch (R-1787), Fig. 3, p. 19.
model permits such discontinuities in the supply curves at \( W_0 \), thus allowing for fixed costs of participation. These costs may constitute a component of the latent random variables, or a component of the effects of the observed variables \( y \), and are therefore permitted to vary among individuals.

This generalization should be contrasted with the model proposed and estimated by Heckman* and others, which assumes (identically) zero minimum hours, implying zero costs of participation and no discontinuity of supply at the reservation wage. Translating Heckman's model into the present notation, his restrictive assumptions are seen to be

(i) \( y \) and \( z \) are identical,

(ii) \( \beta = -\frac{1}{\delta_1} \delta_1 \),

(iii) \( u_o = -\frac{1}{\delta_1} u_3 \), or \( u_3 = -\delta_1 (u_1 + u_2) \),

implying \( A_0 (\equiv K_0) \equiv 0 \). In the present model, however, these assumptions may be tested after the model is estimated, but are not imposed on the estimation.

*Heckman (1974a and b).
III. BASIC MODEL CORRECTED FOR SELECTIVITY BIASES

The model outlined in Sec. II applies to everyone, including nonparticipants. However, it is not operational in the given form, since in general observed data do not include information about market wage offers for nonparticipants. Wages are considered endogenous, including an unobserved individual component \( u_2 \), which is correlated, in the general case, with the other endogenous variables (DP, A, and K). Therefore, the application of the model as specified to a sample of participants (i.e., selecting the sample using DP = 1, which involves the component \( u_1 \) as a criterion) implies a well-known participation selectivity bias. *

In addition, observations in some cross-sectional samples (such as in census data and in the SEO) involve a selectivity bias of another type, here termed the survey-week selectivity bias. In these surveys, information on hourly wage rates and on weekly hours worked \( H \) is available only for individuals working a positive number of hours during one particular week—namely, that prior to the survey interview; whereas information on number of weeks worked (during the previous year) and on \( x \) and \( y \) exists for all sample observations. If the survey week may be considered a random drawing from the current year's 52 weeks, then the probability of finding at work during the survey week an individual who decides to work \( K \) weeks in the survey year is simply \( K/52 \). The sample of workers during the survey week thus overrepresents individuals whose desired \( K \) exceeds the average, and underrepresents part-year workers whose \( K \) is below average. Since the probability of entry into the survey-week working sample is associated with the endogenous variable \( K \), and since \( K \) (and its latent component \( u_4 \)) is generally correlated with other endogenous variables (\( W \) and \( A \), given \( DP = 1 \)), a selectivity bias occurs if the model is applied as specified to the subsample of survey-week workers.

* See Gronau (1973 and 1974) and Heckman (1974a). See also the footnote on p. 3.
The present section indicates the magnitude of these biases and reformulates the model to eliminate them.

The following lemma will be useful for further discussion. First, some notation: Let \( u \) denote \( r \) joint-normal random variables with zero means and a variance matrix \( \Sigma = \{\sigma_{jk}\} = \mathbb{E}u'u' \). Denote \( b_j = -(1/\sigma_1)\sigma_{1j} \), \( j = 1, \ldots, r \), where \( \sigma_1 = \sqrt{\sigma_{11}} \). Note that \( b_1 = -\sigma_1 \); in vector notation, \( b = -(1/\sigma_1)\Sigma_1 \), where \( \Sigma_1 \) is the 1st column of \( \Sigma \). The function \( f(x) \) denotes the unit-normal density:

\[
f(x) = \left(\frac{1}{\sqrt{2\pi}}\right)e^{-x^2/2} = \frac{d}{dx}P(x),
\]

where \( P(x) \) denotes the probability (as in Sec. II, p. 11). Define also \( \lambda(x) = f(x)/P(x) \), and \( \bar{\lambda}(x) = -f(x)/(1 = P(x)) = -\lambda p/(1 - p) \). The magnitude \( \lambda(x) \) is the conditional density of the truncated unit-normal variable \( x \) for an upper limit (left tail) truncation, whereas \( -\bar{\lambda} \) is the conditional density of \( x \) corresponding to a lower limit (right tail) truncation.*

Lemma: The \( u_1 \)-truncated joint distributions (i.e., truncated with respect to \( u_1 \) only) of \( u \) have the following first and second moments \( (j, k = 1, \ldots, r) \):

\[
E(u_j | u_1 \leq \sigma_1 I) = b_j \lambda(I),
\]

\[
E(u_j | u_1 \geq \sigma_1 I) = b_j \bar{\lambda}(i);
\]

\[
E(u_j u_k | u_1 \leq \sigma_1 I) = \sigma_{jk} - b_j b_k \lambda(I),
\]

\[
E(u_j u_k | u_1 \geq \sigma_1 I) = \sigma_{jk} + b_j b_k \bar{\lambda}(I);
\]

*As shown in Eq. (9) below, \( \lambda \) and \( \bar{\lambda} \) are also (minus) the conditional means for upper or lower truncation, respectively, of the unit-normal truncated variable (substitute \( \sigma_1 = 1 \), and \( j = 1 \), and therefore \( b_1 = -1 \), in Eq. (9)).
\[
\text{Cov}(u_j, u_k | u_1 \leq \sigma_1 I) = \sigma_{jk} - b_j b_k [\lambda(\bar{I})^2 + I\alpha(I)] = \sigma_{jk} - b_j b_k D(I),
\]
\[
\text{Cov}(u_j, u_k | u_1 \geq \sigma_1 I) = \sigma_{jk} - b_j b_k [\bar{\lambda}(\bar{I})^2 - \bar{I}\bar{\alpha}(I)] = \sigma_{jk} - b_j b_k \bar{D}(I),
\]

where \( D = \lambda^2 + I\alpha, \) and \( \bar{D} = \bar{\lambda}^2 - \bar{I}\bar{\alpha}. \)

The proof is immediate, applying well-known formulas* corresponding to the general truncated-normal distribution in \( \mathbb{R}^n \) to this particular case and notation.

We deal first with the case where wage-rate and hours data are available for all participants (as in Parnes' data†), and therefore only the first (participation) selectivity bias is present.‡

**CASE 1: WAGE AND HOURS DATA AVAILABLE FOR ALL PARTICIPANTS**

Define participants as individuals working during at least one week in a given year (\( K > 0 \)). It is assumed here that data are given for a sample of \( N \) individuals, \( n \) of which are participants (\( n < N \); and participants are ordered as the first \( n \) observations). The variables \( x_1, y_1, K_1, \) and \( A_1 \) are observed for all \( N \), whereas the hourly wage \( W_1 \) is measured only for \( n \) participants (but not for \( i = n+1, \ldots, N \)). The selectivity bias is manifested by the fact that the expected value of the variables \( u = (u_1, u_2, u_3, u_4) \) within the sample of participants is given, using Eq. (9), by

\[
E(u_i | u_1 \leq \sigma_1 I) = \lambda(I_i) b = \lambda_i b , \quad (i = 1, \ldots, n)
\]

---

*For example, see Johnson and Kotz, pp. 112-114.
†That is, the National Longitudinal Survey, 1967, described in Shea, Spitz, and Zeller. This is the data base used by Heckman (1974a).
‡The magnitude of the selectivity bias was shown by Amemiya. A method essentially similar to the one presented here for correcting the selectivity bias was first proposed by Heckman (R-1984).
since $DP_i = 1$ implies $u_{li} \leq \sigma_1 I_1$, by Eq. (6), and $\sigma_1$ and $b$ are independent of $i$ ($\Sigma$ is assumed constant among individuals). Define the following random variables $\nu_i = (v_{1i}, v_{2i}, v_{3i}, v_{4i}), i = 1,\ldots,n$:

$$\nu_i = u_i - \lambda_i b.$$ (12)

Then the joint distribution of $\nu_i$ (for $i = n$) has zero means, by Eq. (9); $\nu_i$ are uncorrelated with the exogenous variables $E[\nu_i (x_i, y_i)'] = 0$, using Eqs. (9) and (12); and $\nu_i$ are independently distributed across individuals.

The joint distribution of $\nu_i$ is not normal, however, and its variance matrix $\sum_i^{\nu}$ depends on $i$ (heteroskedastic) and is given, using Eq. (11), by

$$E(v_{ji} v_{ki}) = \text{cov}(u_{ji} u_{ki} | u_{li} \leq \sigma_1 I_i)$$

$$= \sigma_{jk} - b_j b_k D_i, \quad (j, k = 1,\ldots,4)$$ (13)

where $D_i = D[\lambda(I_i)] = \lambda_i^2 + I_i \lambda_i$, as above.

Eq. (13) is expressed in matrix notation as follows:

$$\sum_i^{\nu} = \sum - b b' D_i = \sum - \frac{1}{\sigma_1^2} \sum_1^{'} \sum_1^{'} D_i.$$ (14)

Using these definitions and results, the model as applied to the working sample consists of the following equations (in addition to Eq. (6)) in this case ($i = 1,\ldots,n$).

**Market wage** (modifying Eq. (1)):

$$w_i = a x_i + b_2 \lambda_i + v_{2i}.$$ (15)

**Annual hours supply** (modifying Eq. (3)):

$$A_i = \gamma z_i + \delta w_i + b_3 \lambda_i + v_{3i}.$$ (16)
Weeks supply (modifying Eq. (4)): 

\[ K_i = Y_2'Z_i + \delta_2 W_i + b_4 \lambda_i + v_{4i} \cdot \tag{17} \]

The variables \( v_i = (v_{1i}, v_{2i}, v_{3i}, v_{4i}) \) satisfy \( E v_i = 0 \); \( \text{var} v_i \) is given in Eq. (14), in particular, \( \text{var} v_{1i} = \sigma_1^2 (1 - D_i) \), since \( b_1 = -\sigma_1 \). Also \( v_i \) are uncorrelated with the exogenous variables, including \( \lambda_i \).

Substitution from Eq. (5) into Eq. (15) gives the modified pseudo reservation wage equation (modifying Eq. (7)): 

\[ w_i = \varphi' v_i + \sigma_1 I_i + b_2 \lambda_i + v_{2i} \cdot \tag{18} \]

Denote by \( w_i^*, A_i^* \), and \( K_i^* \) the expected values of the corresponding variables in the working sample (that is, conditional on \( u_{1i} \leq \sigma_1 I_i \) or \( D_{p_i} = 1 \)). Substitution of \( w_i = w_i^* + v_{2i} \) into the supply equations gives

\[ A_i = Y_1'Z_i + \delta_1 w_i^* + b_3 \lambda_i + (\delta_1 v_{2i} + v_{3i}) = A_i^* + (\delta_1 v_{2i} + v_{3i}) , \tag{19} \]

\[ K_i = Y_2'Z_i + \delta_2 w_i^* + b_4 \lambda_i + (\delta_1 v_{2i} + v_{4i}) = K_i^* + (\delta_1 v_{2i} + v_{4i}) . \]

Further substitution of \( w_i^* \) from Eq. (15) into Eq. (19) gives the reduced-form supply equations:

\[ A_i = Y_1'Z_i + \delta_1 \alpha' x_i + (\delta_1 b_2 + b_3) \lambda_i + (\delta_1 v_{2i} + v_{3i}) , \tag{20} \]

\[ K_i = Y_2'Z_i + \delta_2 \alpha' x_i + (\delta_2 b_2 + b_4) \lambda_i + (\delta_2 v_{2i} + v_{4i}) . \]

Equations (20) combined with Eqs. (6) and (15) constitute, therefore, a complete reduced-form model for this case. The model specified above, which corrects for the participation bias, lends itself to known consistent estimation techniques, provided the participation index \( I_i \) (and
hence \( \lambda_i = \lambda(I_i) \) is given or measured with negligible errors (see, however, pp. 35-37), and provided all other variables are measured for all participants.

The coefficients in Eqs. (15) to (18) identify all the coefficients appearing in the original model, as well as the elements of the first column \( \Sigma_1 \) of the original variance matrix \( \Sigma \). Additional elements of \( \Sigma \) (and \( \Sigma^0 \)) are identifiable through the average variance matrix

\[
\bar{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \hat{\Sigma}_i
\]

of residuals in Eqs. (15) to (17) or through moments of reduced-form residuals in Eqs. (20), combined with Eq. (14).

**CASE 2: WAGE AND HOURS DATA AVAILABLE ONLY FOR SURVEY-WEEK WORKERS**

Assume that \( m \) participants (\( m < n \)) work during the survey week (ordered first among the \( n \) participants). An estimate of the hourly wage \( W_i \) (applicable to hours in the current survey year) is provided by the hourly wage rate measured for \( H_i \) hours worked in the survey week.* Annual hours are estimated by \( A_i = H_i K_i \), where \( K_i \) corresponds to the number of weeks worked (assumed known) in the year preceding the survey.

The residuals (random variables) \( u_2, u_3, \) and \( u_4 \) of Eqs. (1), (3), and (4) now include additional error components, relative to Case 1: \( u_{2i} \) includes week-to-week variations in the mean hourly wage rate, \( u_{4i} \) includes year-to-year variations in weeks (for a given individual \( i \)), and \( u_{3i} \) compounds week-to-week variations in \( H_i \) with the year-to-year components in \( K_i \) (i.e., in \( u_{4i} \)). In addition, measurement-error

---

*In surveys such as the SEO, workers are not asked directly about their hourly wage rates, but rather about their total weekly earnings \( E_i \), with \( W_i \) computed as \( E_i / H_i \). See Hanoch (R-2067) for a discussion of this definition as a source of error.
components could be larger in this case.* However, if all these components may be assumed transitory, such that the deterministic part of the model is identical to that of Eqs. (1) to (4), and if these transitory components are also independent of the exogenous variables \( x \) and \( y \), then the applicable underlying model is formally the same, except that the variance matrix \( \sum \) now incorporates these additional transitory error components. However, the probability of working during the survey week \( (i \leq m) \) is proportional to the endogenous variable \( K_i \), as explained above, p. 14. Therefore, a survey-week bias is implied for Eqs. (15) to (18), if applied to these workers only.

These biases and their magnitudes are derived explicitly in the appendix. Equation (A-3) shows that the expected values of \( v_{-i} \) in the survey-week working sample (where \( H_i > 0 \) are generally not zero, and are given by

\[
E(v_{-i} \mid H_i > 0) = \frac{1}{K_i^*} E(\delta_2 v_{21} + v_{41} v_{-i})
\]

\[
= \frac{1}{K_i^*} \left[ \left( \delta_2 \sum_{22} + \sum_{44} \right) - (\delta_2 b_2 + b_4) b D_i \right], \tag{21}
\]

where \( K_i^* \) is expected weeks (among participants, Eq. (19)), and the bracketed term is the covariance between \( (K_i - K_i^*) \) and \( v_{-i} \), as derived by Eq. (14).

Since this bias arises because the selection probability is proportional to the endogenous variable \( K_i \), the bias is eliminated by transforming all the variables (including the constant) by the multiplicative factor \( (K_i^*/K_i)^{1/2} \) for each individual \( i \). Alternatively and equivalently, all the first and second moments may be weighted by the individual weights \( K_i^*/K_i \) in estimating Eqs. (15) to (18) by linear methods. (See appendix.)

*The SEO instructions for interviewers seem to induce such measurement errors by treating hours and earnings information differently and inconsistently with regard to unusual circumstances such as sickness, vacation, overtime. See also the previous footnote.
Since $K^*_1$ is not observed, it may be estimated consistently from the weeks reduced-form equation in Eqs. (20), which is estimable for the total sample of participants. Alternatively, $K^*_1$ may be replaced by $E[K]_1$, the expected value of $K_1$ in a reduced-form estimate using Tobit, which allows for the constraint $K_1 \leq 52$.

If the individual residuals $(K_1 - K^*_1)$ include no transitory components and are therefore constant in adjacent years, this method gives unbiased estimates even if $K_1$ is measured in the year preceding the survey year. However, if a year-to-year transitory variation in $K_1$ exists, the bias is only partially eliminated, since the selection probability is proportional to the desired $K^*_1$ in the survey year $t$, whereas the weights use $K_{t-1}$ measured in year $t-1$. In any case, the remaining bias is small relative to the bias in the unweighted case, since the permanent component in the weeks residuals accounts for most of the total residual variance in weeks.\footnote{See appendix, and Nanoch (R-1787), p. 5, footnote *.}

To summarize, the correction for the survey-week bias is performed by replacing each variable $X_i$ in the equations of Case 1 by its transform $\hat{X}_i = (\hat{K}^*_i/K_i)^{1/2} X_i$ and then applying the model to the subsample of survey-week workers.

\footnote{See Tobin, and also the discussion below in Sec. IV, pp. 27-30.}
IV. STRATEGY FOR ESTIMATING THE BASIC MODEL

INTRODUCTION AND GENERAL COMMENTS

In principle, the basic model (Eqs. (1), (3), (4) and (6)), subject to the modifications required because of insufficient information and to the weeks corner-solution truncation, could be estimated as a complete simultaneous model by a full-information maximum likelihood method. That is, the likelihood function corresponding to the sample in either Case 1 or 2 may be explicitly derived, and then maximized with respect to all the unknown parameters.

However, prohibitive complexity and costs preclude this method. The joint distribution of the random variables of the model involves a four-dimensional, multivariate, doubly truncated, normal distribution, where each of the variables is a function of many unknown parameters. Maximizing this likelihood seems infeasible at the present state of the art. In addition, it is doubtful that such full-information maximum likelihood estimates, which are asymptotically fully efficient if the model is specified perfectly, are sufficiently robust against various misspecifications of the model, such as variable exclusions, heteroskedasticity (of the underlying $\sigma^2$), non-normality, nonlinearities, and inhomogeneity of expected behavior among individuals.

The strategy proposed here for estimating the basic model is designed to handle all six types of problems unique to it, as outlined above in the Introduction, and to yield approximately consistent estimates for all its parameters with relatively low computation and programming costs. Although these estimates are not fully efficient, they are satisfactory since they use efficient methods and the information contained in the total sample, including that for the nonparticipants. In addition, the methods proposed seem robust against many of the misspecification problems mentioned above, particularly heteroskedasticity and non-normality. More efficient estimates may be achieved, at some

*The information is not used in each stage, however. For example, in the probit analysis, pp. 23-26, information on wages and hours of workers is ignored.
of the proposed steps in this scheme, by allowing for heteroskedasticity (as estimated from the residuals, in combination with that implied by the theory) and by iteratively repeating certain steps in the scheme. But these measures are more complex and therefore more costly, and yield unknown, possibly minor benefits.

The foregoing discussion of the proposed estimation scheme is mostly nonrigorous and intuitive, since a rigorous analysis of problems such as relative efficiency and robustness is extremely difficult. Intuition, like personal taste, varies among researchers, however. The expert econometrician who reads this might therefore choose to apply a different strategy to estimate the same model.

The discussion here is aimed at estimating the model using a large (several thousands) cross-sectional sample, which falls under Case 2 in Sec. III—that is, a survey-type sample (census or SEO data) where hours of work and hourly wage rates are measured only for workers in the survey week. Since estimation of the model under Case 1 is similar but simpler, the more complex strategy could be modified to fit Case 1 by omitting the steps designed to eliminate the survey-week bias through the weighting scheme explained in the previous section.

Some sources, however, give panel-type data (e.g., the Income Dynamics sample), which combine cross-sectional with time-series information. Utilizing such data may considerably improve the estimates, as well as provide additional information and insights regarding the stability of the behavioral relations, the sizes of the transitory and permanent components, and the validity of various specifications. But the analysis of such data requires extension and completion of the model itself, in addition to alternative estimation methods, and is outside the scope of this report.

ESTIMATING THE PROBABILITY OF PARTICIPATION: PROBIT ANALYSIS

In this first step, Eq. (6) is estimated by maximum likelihood (probit analysis), using data on $X$, $Y$, and DP (participation) for the total sample, to get estimates of the unit-normal index $I_i$ and to derive from it (using exact functional relations) estimates of $\lambda_i$, $D_i = \lambda_i (\lambda_i + I_i)$, and the participation probability $p_i$ for each individual $i$. 

The likelihood function is given by

\[ L = \prod_{i=1}^{n} P(I_i) \prod_{i=n+1}^{N} [(1 - P(I_i))] , \]

where \( I_i = (\alpha'x_i - \beta'y_i)/\sigma_i \), and \( P(I) \) is the unit-normal probability function.

In following steps, the estimated \( I_i \), \( \lambda_i \), and \( D_i \) are substituted for their true values without distinguishing them in notation. Moreover, the error in the actual participation equation--namely, \([DP_i - E(DP_i)]\)--must not be confused with the relatively small errors of estimation--namely, \((I_i - EI_i)\) or \((\lambda_i - E\lambda_i)\). Since the method is maximum likelihood, it is asymptotically fully efficient relative to the information utilized—that is, the information on all exogenous variables for the total sample and on previous-year participation (ignoring at this stage the information pertaining to quantities supplied by participants and to measured wage rates).

Nevertheless, some estimation errors in \( \lambda_i \) and \( I_i \) exist, and are thus incorporated into the residuals in each equation that includes estimated \( \lambda_i \) or \( I_i \). Since the estimated values are functions of the exogenous variables, \( \lambda_i \) and \( I_i \) are treated as exogenous variables (instruments) in these equations, thus providing a "two-stage" procedure, where the probit first stage is nonlinear (maximum likelihood).

Equation (6) is treated here as a reduced-form equation, not as a structural equation, in the procedure for estimating the simultaneous model. That is, the probit step is used for estimating the values of \( \lambda_i \) and \( I_i \) for each worker, but not for directly estimating the parameters \((\alpha, \beta)/\sigma_i\), since these are estimated directly at later steps using data on measured wage rates in Eqs. (15) and (5) or (18). In principle, this step gives an estimate of the coefficient \( \alpha_j/\sigma_1 \).

---

* Computer programs for probit analysis are available and their application is inexpensive. See Hanoch (R-2067).
† Except in the section below, pp. 35-37, where \( \hat{I} \) denotes the estimate.
for every variable \( x_j \) in \( \mathbf{x} \) and not in \( \mathbf{y} \). Since \( \alpha_j \) is estimated directly in the wage equation (15), this step provides multiple estimates of \( \sigma_1 \) (by taking ratios of coefficients for each such \( x_j \)) and of \( \alpha_j \) (by multiplying the probit coefficients by each alternative estimate of \( \sigma_1 \)). This procedure can also estimate coefficients \( \beta_j \), corresponding to variables \( \mathbf{y}_j \) in \( \mathbf{y} \) and not in \( \mathbf{x} \); and \( (\alpha_k - \beta_k) \) for variables common to \( \mathbf{x} \) and \( \mathbf{y} \), providing multiple indirect estimates of \( \beta_k \) for such variables (by using again the corresponding estimate of \( \alpha_k \) from the wage equation). These possibilities thus indicate overidentification of parameters whenever there exists more than one variable \( x_j \) included in \( \mathbf{x} \) and not in \( \mathbf{y} \).

In addition to sampling errors, however, the probit estimates of \( \alpha_j \) (up to a factor of proportionately \( \sigma_1 \)) may also differ conceptually from \( \hat{\alpha}_j \) estimated in the wage equation. The "concealed" market wage rate influencing the individual's decision on participation in the labor force (which is associated with the probit coefficients) may be biased, relative to both actual wage offers and the wage rate conceived by workers in their decisions concerning quantities supplied. Since the latter is the focus of interest in this model, the estimated individual probit coefficients are ignored, except as they contribute to the linear combination index \( I_{1} \). This index, in turn, is used for correcting the selectivity bias (through \( \lambda(I_{1}) \)), as well as for estimating the reservation wage coefficients, by introducing \( I \) as a variable in Eq. (5) or (18), together with \( \mathbf{y} \), and using as a dependent variable either the actual wage or the estimated expected wage (as explained on pp. 30-31 and 35-37 below).

Finally, experience with alternative models (such as linear probability or logit*) and alternative methods for estimating the probability of participation suggests that probit estimates are efficient and robust against non-normality and heteroskedasticity,† although normality and

---

* See Berkson (1944 and 1955) regarding the logistic distribution and the logit estimation.

† That is, heteroskedasticity of the underlying normal distribution (of \( u_{11} \), in this case). The binomial deviations (\( DP_1 - EDP_1 \)) are heteroskedastic even if \( \text{var} u_{11} \) is constant.
homoskedasticity may appear to be crucial in formulating the likelihood function of Eq. (22). This observation seems to hold for a variety of problems and types of data.

PRELIMINARY REDUCED-FORM ESTIMATES OF WEIGHTING SCHEME AND MARKET WAGE EQUATION

In order to apply the weighting scheme (described in Sec. III and the appendix), which nearly eliminates the survey-week bias, an estimate of the weight $K^*/K$ is required for each survey-week worker (denoted simply as "worker" in the following discussion, as distinguished from "participant"). Since information on weeks but not on wages is available for all participants, $K^*$ is estimated by ordinary least squares (OLS) regression of $K$ in the reduced-form equation for weeks in Eq. (20), where the independent variables are $\lambda$, $x$, and $z$.

Note that $\lambda$ is often highly correlated with $I$ over the participants' sample range, although $\lambda(I)$ is a nonlinear convex and decreasing function. In addition, if $z = y$, then $I$ is an exact linear combination of $x$ and $z$, and Eq. (20) is highly multicollinear. This multicollinearity affects the reliability of individual coefficients, but not the coefficient of determination $R^2$ or the predicted value $K^*$ (as long as numerical errors are not too large, which might be a problem if the independent-variables moment matrix is too nearly singular).

This predicted value is now used to compute weights $K^*/K_i$ for each worker ($i = m$). These weights are later revised by allowing for the truncation effect in weeks, but are required at this stage for estimating the predicted wage in the next step, which is required in turn for executing this weeks-truncation correction itself.

Next, first-round estimates of $w^*$ are derived, applying weighted least squares (WLS) linear regression of $w = \ln W$ on the variables $x$ and $\lambda$ (Eq. (15)) to the working sample, using the weights ($K^*/K$) estimated above to correct for the survey-week bias. These estimates of $w^*$ are also revised in the final stage, but are needed for the next step.

*The correlation is typically around -.98.*
TOBIT ANALYSIS OF SUPPLY OF WEEKS

Given the estimates of Eq. (15) above, an estimate of \( w^*_1 \) may now be imputed to each participant \((i \leq n, \text{ or } K_1 > 0)\), although market wage data are missing for nonworking participants. The weeks supply equation in Eq. (19) is reproduced here, allowing for the truncation at 52 weeks:

\[
\begin{align*}
(i) \quad \tilde{K}_1 &= \gamma_{2i} \tilde{z}_1 + \delta_2 w^*_1 + b_4 \lambda_1 + \varepsilon_1 \\
(ii) \quad K_1 &= \tilde{K}_1 \quad \text{if} \quad \tilde{K}_1 \leq 52 \quad ; \quad K_1 = 52 \quad \text{if} \quad \tilde{K}_1 > 52 .
\end{align*}
\]

\( \tilde{K}_1 \) is "desired weeks supply" and \( \varepsilon_1 = \delta_2 \tilde{v}_2 + \nu_{4i} \). This truncated weeks supply equation may now be calculated by Tobin's method* of estimating a linear equation with a limited dependent variable (denoted Tobit). This method yields maximum likelihood estimates, assuming normality and homoskedastic disturbances \( \varepsilon_1 \). However, since \( \varepsilon_1 \) is normal but heteroskedastic in this model (see Eq. (14)), applying this method here gives only an approximation for maximum likelihood.†

This equation may be chosen as a final estimate of the weeks supply equation. Alternatively, a similar equation may be estimated using only the working sample of \( m \), and using a weighted-Tobit method with \((K^*/K)\) as weights.‡ One justification for the latter alternative is comparability with the hours supply and wage equations, which are estimable only for workers. Another is the assertion that, since \( w^* \) was estimated in the working sample, the imputation of \( w^* \) to nonworkers may involve biases (if transitory components in \( \varepsilon \) are important and highly correlated with \( v_2 \)); therefore, the internal consistency of the equation might be improved if restricted to the working sample.

---

*See Tobin, Amemiy.

†Computer programs for executing Tobit estimation are readily available and usually converge in three to five iterations. In terms of costs, each iteration is equivalent to about two OLS regressions on equivalent numbers of observations and variables. See Hanoch (R-2067).

‡Existing Tobit computer programs can handle the weighted case as well, at negligible additional costs.
However, the Tobit step is used here not as a vehicle for yielding final estimates, but rather as an intermediary step, for revising the weights as well as for creating a dichotomous variable that distinguishes between full-year and part-year workers. Several considerations lead to this suggestion. First, as explained above, the Tobit maximum likelihood method is predicated on normality and homoskedasticity, and may be sensitive to deviations from these assumptions. Second, this analysis cannot be easily incorporated into a simultaneous-equations model as part of a system. Therefore, although possibly efficient as a single-equation (limited information) technique, it ignores relations across equations. Empirical evidence supports the belief that the weeks and hours residuals are highly correlated; also, both may be correlated with the wage equation residuals.* Thus, the weeks equation should preferably be estimated as part of a simultaneous system, as suggested for the final stage, using residual correlations across equations to improve efficiency.

On the other hand, in specifying the supply equations within the linear simultaneous system, it would be desirable to use the Tobit estimates to allow for the effect of the upper-limit restriction on weeks, as explained below.

The Tobit analysis produces two key estimates (which may be computed for every individual as part of the Tobit computer program output, at negligible marginal costs): (1) An estimate of $\tilde{K}^*$, the predicted value of desired weeks ("the index"), as given by the deterministic part of Eq. (23); (2) the expected value locus $E_K$, which is a nonlinear function of the index (and the limit). For truncation at an upper limit, $E_K$ is a concave function of $\tilde{K}^*$, smaller than $\tilde{K}^*$ and approaching the limit asymptotically from below.†

To introduce the effect of this truncation into the linearized supply equations, a shift is allowed in the level at the point where predicted desired weeks exceed the limit, and a change in slope is allowed with respect to the two primary variables—the wage variable

---

*See Hanoch (R-2067).
†See Tobin, Amemiya, Hanoch (R-2067).
w (= ln W) and nonwage income Y. This linearization in the weeks supply approximates the expected value locus EK by two linear segments, such that the kink occurs where the individual’s desired number of weeks exceeds the physical limit of 52, as predicted by the Tobit index K*=—that is, where the probability of full-year work exceeds 0.5.

The actual method for executing this modification is simple. Define the dichotomous variable D^K:

\[
D^K_i = \begin{cases} 
0 & \text{if } \tilde{K}^* \leq 52 \\
1 & \text{if } \tilde{K}^* > 52 
\end{cases}.
\]

Then introduce (linearly) into the two supply equations these three variables:

(i) \( D^K \);  
(ii) \( D^{KW} = D^K \cdot w \);  
(iii) \( D^{KY} = D^K \cdot Y \).

The corresponding coefficients are then interpreted as follows:

1. The coefficient of \( D^K \) estimates the change in the level of supply between the two groups (i.e., workers with desired K exceeding the limit, compared with other workers).
2. The coefficient of \( D^{KW} \) estimates the change in the slope of the supply curve with respect to \( w \) between the two groups. In other words, for desired full-time workers the slope \( \partial K/\partial w \) now consists of the sum \( \partial K/\partial w + \partial K/\partial D^{KW} \), compared with only \( \partial K/\partial w \) for others.
3. Similarly, the coefficient of \( D^{KY} \) estimates the change in the slope of the income effect between the two groups.

One could not get consistent estimates by the same method with an alternative dichotomous variable (and its interactions with w and Y).
that distinguishes between full-year and part-year workers by the actual rather than predicted number of weeks worked, since this variable is an endogenous variable associated with the dependent variable K itself.

On the other hand, D^K is exogenous, because the combination of exogenous variables given by the Tobit estimate K* is used as the criterion for defining it. * D^KY is also exogenous (since Y is considered exogenous), whereas D^KW is an endogenous variable in the linear model, since w itself is endogenous.

An additional use of the Tobit analysis is to redefine the weights used for survey-week bias correction to be EK/K (rather than K*/K), where EK is the estimated expected value of K. These weights seem preferable to K*/K because the numerator of the weight is designed to equalize, or preserve the scale of, the variances of residuals across individuals. This preservation of scale results because the numerator's expected value is equal to that of the denominator K. But the Tobit EK is a consistent and more efficient estimate of this mean, compared with the OLS estimate K*, since the estimating equation for K* ignores the truncation problem and the implied nonlinearity of the expected value locus.

To summarize, the Tobit step produces two modifications of the model. First, EK/K are used as weights to correct the survey-week bias. Second, the estimated index \( \tilde{K} \) is used to define the truncation dummy variable D^K as well as the two associated interaction variables D^KW and D^KY to be included in each of the supply equations (16) and (17). This procedure permits a shift in the supply level, as well as a change in the wage and income coefficients, between the groups of predicted full-year and part-year workers.

**MODIFYING THE PSEUDO RESERVATION WAGE EQUATION**

Equation (18), which is used as an alternative for estimating the coefficients \( \beta \) of the reservation wage equation, includes the two independent variables I and \( \lambda \). As mentioned above, p. 26, I and \( \lambda \) tend to be highly correlated negatively over the working sample range,

* A possible but probably inefficient alternative is to treat the actual full-year dichotomous variable, as well as its interaction with Y, as additional endogenous variables in the linear model.
since \( \lambda \) is a monotone decreasing function of \( I \) with a small degree of nonlinearity.* To avoid this multicollinearity problem, it is preferable to substitute for \( I \) a linear function of \( \lambda \), as estimated by the following regression in the working sample:

\[
I_i = I_o - I_{1 \lambda i} + e_i .
\]  

(24)

The modified equation (18) is now

\[
\hat{w}_i = (\beta_o + \sigma_{1 I_0}) + \beta_{11}v_{1 i} + (b_2 - \sigma_{1 I_1})\lambda_i + \hat{\nu}_{2 i} ,
\]

where \( \beta_o \) is the constant coefficient in \( \beta \); and \( v_{1 i} \) and \( \beta_{11} \) are, respectively, the \( v \) variables without the constant and their \( \beta \)-coefficients. The variable \( \hat{\nu}_{2 i} \) includes \( \sigma_{1 e_i} \), the residual generated by the nonlinearity in \( I(\lambda) \), in addition to the wage equation residual \( v_{2 i} \). †

The coefficient \( b_2 \) is estimated separately in the market wage equation (15), and \( \lambda_i \) and \( I_o \) are estimated separately by Eq. (24), so that \( \sigma_{1 I_1} \) and \( \beta_o \) can be estimated from the coefficient of \( \lambda \) and the intercept, respectively, in Eq. (25).

In modifying this equation, the function \( I(\lambda) \) rather than \( \lambda(I) \) should be linearized, if the same variable \( \lambda \) is to appear in all equations of the operational model. Omitting \( I \) as an instrument in the model involves no loss of power, since it is an exact linear combination of the other instruments \( x \) and \( y \).

**LINEAR-SYSTEM ESTIMATES**

The various modifications of the basic model may now be assembled into a system of simultaneous linear equations defined for the working sample of \( m \) individuals (where \( H > 0 \)).

---

*Estimates of the unit-normal index \( I \) typically range between \(-3\) and \(+3\), yielding a correlation coefficient of about \(-.98\) to \(-.99\) between \( \lambda \) and \( I \).

†Estimation errors in \( b_2 \hat{\lambda_i} \) (from probit) are common to the estimated residuals in both equations. See p. 33 below.
Since the recommended computer programs handle the estimation by using a moment matrix as an input, rather than individual observations, the variables in each observation need not be transformed for application of the weighting scheme of Sec. III, Case 2. Instead, a preliminary program creates the weighted moment matrix,

\[
\hat{X}'\hat{X} = \left\{ \sum_{i} \left( X_{j1} X_{k1} \cdot \frac{EK_i}{K_i} \right) \right\},
\]

where \( X \) contains all variables of the model, including the constant \( 1 \), and the weights \( EK/K \) are estimated by Tobit analysis above. Collecting all previous results and omitting the weighting operator, the linear system consists of the following four equations applicable to the working sample \( (i = 1, \ldots, m) \).

**Market wage equation:**

\[
\ln W_i = w_i = a_1 x_i + b_2^\lambda_1 + v_{2i}.
\]  

(26)

**Annual hours supply equation:**

\[
A_i = \gamma_1 z_i + \delta_1 w_i + a_1 D_1^K + a_2 D_1^{KW} + a_3 D_1^{KY} + b_3 \lambda_i + v_{3i}.
\]  

(27)

**Weeks supply equation:**

\[
K_i = \gamma_2 z_i + \delta_2 w_i + a_1 D_1^K + a_2 D_1^{KW} + k_3 D_1^{KY} + b_4 \lambda_i + v_{4i}.
\]  

(28)

**Pseudo reservation wage equation:**

\[
\ln W_i = w_i = (\beta_0 + \sigma I_0) + (b_2^\lambda_1) z_{1i} + (b_3 - \sigma I_1) \lambda_i + \hat{v}_{2i}.
\]  

(29)

*The sums \( \sum_{i} X_{j1} (EK_i/K_i) \) are thus also weighted by \( EK_i/K_i \). The sum of the weights modifies approximately the number of observations \( m \), to achieve correctly scaled means, covariances, and other test statistics.*
The endogenous variables in these four equations are $W, A, K,$ and $D_{KW}$; $Y_{1}$ is the vector of exogenous variables $Y$, omitting the constant $1$; and $B_{1}$ are the corresponding coefficients, omitting the intercept $B_{0}$.

In terms of the order condition for identification (exclusion restrictions), this system is identified if at least two distinct $X$ or $Y$ variables are not included among the variables $Z$, since each of the supply equations (27) and (28) includes two right-side endogenous variables, $W$ and $D_{KW}$ (if $\lambda$ is viewed as exogenous).

Consistent single-equation estimates of these linear equations may now be derived inexpensively by using weighted two-stage-least-squares (TSLS) estimation. A more efficient set of system estimates is provided, however, by applying the weighted three-stage-least-squares (3SLS) estimation technique. The increased efficiency results from taking into account covariances of residuals across equations to yield generalized least squares (GLS) estimates, using the TSLS residuals to calculate the variance-covariance matrix.

However, the 3SLS technique should not be applied to all four equations (Eqs. (26 to 29)) in one sweep, because of the high correlation between residuals in Eqs. (26) and (29). The difference between these residuals, $\hat{\nu}_{2}$ and $\nu_{2}$, consists of the following four elements:

1. $\sigma_{1}e$, the nonlinearity error component of Eq. (24);
2. the error associated with the estimation of $I$ in the probit analysis (in addition to the error in $b_{2}\lambda$, which is common to both $\nu_{2}$ and $\hat{\nu}_{2}$);
3. differences between measurement errors in $X$ and $Y$, if any;
4. possibly some behavioral errors due to differences between expected actual wages and expected conceived wages used in the participation decision, as explained above, p. 25.

---

* It is possible to treat $\lambda$ as endogenous, because of its nonlinear relation to $I$. In this case, three variables should be excluded from $Z$. See Hanoch (R-2067).

† See Zellner and Theil.

‡ See Aitken, Theil.

** Computer programs for executing this method using the weighted moment matrix as input are also available (in the same package with the TSLS program), and inexpensive to apply. See Hanoch (R-2067).
Nevertheless, these components are small relative to the cross-
sectional wage residual variance \( \hat{\sigma}_2^2 \), implying a high correlation be-
tween \( v_2 \) and \( \hat{v}_2 \). Since Eqs. (26) and (29) have the same dependent
variable, this correlation also implies a high correlation between the
deterministic parts of the two equations. In other words, the identity
equation (5) holds with a high degree of approximation for the estimated
I and the measured \( x \) and \( y \). Since the 3SLS method is equivalent to GLS
estimates of a single equation created by "stacking" all equations of
the model, \( \dagger \) the high correlation between these two equations would in-
roduce high multicollinearity into the stacked equation, with the known
result of drastically diminishing the accuracy of estimates in each
equation separately. In addition, the adverse effects of multicol-
linearity would be magnified by specification errors in either or
both of these two equations.

The alternative suggested here avoids this problem, by estimating
separately two systems of three equations each, using 3SLS in each case:

1. A three-equation system consisting of Eqs. (26) to (28), used
mainly for efficient estimates of the wage equation (26), and
of the residual variance matrix.

2. An alternative three-equation system consisting of Eqs. (27)
to (29), and constituting linear system estimates of the supply
model—that is, the two quantity-supply equations for annual
hours and weeks, and the pseudo reservation wage equation.

Since \( v_2 \) of Eq. (26) includes fewer error components than \( \hat{v}_2 \), the
first three-equation system should be used for estimating the variance-
covariance matrix \( \hat{\Sigma}_v \) as well. Recall that \( \hat{\Sigma}_v \) is heteroskedastic when
the original variance matrix \( \Sigma \) is constant. The 3SLS method assumes
constant \( \hat{\Sigma}_v \) but tends to be robust against its nonconstancy. The estimated
\( \hat{\Sigma}_v \) provided as part of the 3SLS output is thus an estimate of the

\[ \text{The empirically estimated correlation is .94, in Hanoch (R-2067).} \]
\[ \text{† See Theil.} \]
mean $\bar{V}$ among individuals in the working sample. Estimates of the underlying matrix $\bar{\Sigma}$ (if assumed constant; or $\bar{\Sigma}$ if allowing for original heteroskedasticity) may be derived from $\bar{V}$ by applying the relation

$$\bar{\Sigma} = \bar{V} + \hat{b}\hat{b}' \bar{D}. \quad (30)$$

Equation (30) aggregates Eq. (14) over the working sample, and $\hat{b} = (-\hat{c}_1, \hat{b}_2, \hat{b}_3, \hat{b}_4)$ are estimated coefficients of the model. $\bar{D}$ is the mean of $D_1 = \lambda_1^2 + \lambda_1 I_1$ in the working sample.

The second system of Eqs. (27) to (29) provides final estimates for the supply equations using the 3SLS method, which in this case allows for the covariances of the pseudo reservation wage residuals $\hat{v}_2$ with the supply residuals $v_3$ and $v_4$, as well as the covariance between the two supply residuals.

An alternative method for estimating the reservation wage equation, as explained in the next section, makes use of the results provided by these two 3SLS systems in conjunction with additional information on the total sample.

**ALTERNATIVE ESTIMATE OF RESERVATION WAGE EQUATION**

The coefficients $\hat{\alpha}$ of Eq. (26) are estimated consistently and relatively efficiently by the 3SLS method described in the previous section. Using these estimates $\hat{\alpha}$, it is now possible to construct an estimate $\hat{w}_i = \hat{\alpha}'x_i$ of the predicted log wage for each individual, given his measured variables $x_i$, such that $\hat{w}_i$ is free from selectivity biases:

$$\hat{w}_i = \hat{\alpha}'x_i = \alpha'x_i + \varepsilon_{1i}, \quad (i = 1, \ldots, N) \quad (31)$$

with $\varepsilon_{1i}$ equal to $(\hat{\alpha} - \alpha)'x_i$, having zero mean, and uncorrelated with $x_i$ in the limiting distribution within the total sample of both participants and nonparticipants. Substituting from the identity equation (5) into Eq. (31) gives

$$\hat{w}_i = \sigma_{1i} + \hat{\beta}'y_i + \varepsilon_{1i} = \sigma_{1i} + \beta'y_i + (\varepsilon_{1i} + \varepsilon_{2i}), \quad (i = 1, \ldots, N) \quad (32)$$
where $\hat{\beta}_1$ is the estimated probit index and $\varepsilon_{2i} = \sigma_1 (I_i - \hat{I}_i)$. Since the residuals $\varepsilon_1$ and $\varepsilon_2$ are associated with estimation errors, they tend to vanish with increased sample size. The OLS estimate of Eq. (29) on the total sample thus gives consistent estimates of $\sigma_1$ and $\beta$.

The advantages of this method, compared with the 2SLS estimates of Eq. (29), are as follows: (1) It utilizes additional information on $y$ for nonworkers; (2) it is simpler to execute, since no selection bias corrections are required—thus obviating the need to linearize (when both $\lambda$ and $I$ are present in the same equation), or to weigh the moments or the observations, as in Eq. (29); (3) it seems more efficient as explained below.

Substituting from Eq. (31) into the original wage equation (1)

$$w_i = \alpha'x_i + u_i = \hat{\omega}_i - \varepsilon_{1i} + u_{2i}.$$ 

Substituting from Eq. (32) gives

$$w_i = \sigma_1 \hat{\beta}_1 + \beta'\gamma_i + (u_{2i} + \varepsilon_{2i}).$$

Clearly, the variance of the residual $(u_2 + \varepsilon_2)$ approaches $\sigma_2^2$ as the sample size increases, and does not vanish—in contrast with the var of $(\varepsilon_1 + \varepsilon_2)$ in Eq. (32). Even if wages could be measured for every worker, the OLS estimate of Eq. (32) would tend to be more efficient than an estimate of Eq. (33), since both equations have the same right-side variables and var $(\varepsilon_1 + \varepsilon_2)$ tends to be smaller than var $(u_1 + \varepsilon_2)$.

The estimates of var $(\hat{\sigma}_1, \hat{\beta})$ as computed by an OLS output of Eq. (32) are biased downward considerably, since the residual variance is estimated under the assumption that observations on $\hat{\omega}_i$ are $N$ independent observations, whereas in fact they are associated with only $k$ independent
to estimate \( \text{Var} \varepsilon_1 \); and by estimating \( \text{Var} \varepsilon_2 \) as the difference between residual variances in Eqs. (26) and (29).*

In conclusion, both simplicity and efficiency suggest the use of OLS in Eq. (32) for final estimates of the reservation wage equation. In addition, similarity of results from the two alternative methods would help to increase confidence in the internal consistency of this model (although it is not clear exactly what assumptions are tested by this comparison).

**POSTESTIMATION ANALYSIS**

The estimates described in the previous sections may be used to derive other estimates of parameters and variables of the basic model. First, since the residual in the unobserved reservation wage equation is given by \( u_o = u_1 + u_2 \) (Eq. (2)), its estimated variance and covariance with \( u = (u_1, u_2, u_3, u_4) \) are easily derived from \( \hat{\Sigma} \).

Second, the coefficients may be used to construct expected individual predictions of the unobserved variables of interest, such as wage offers \( \hat{W} \), reservation wage \( \hat{W}_o \), predicted participation \( \hat{DP} \) (using the comparison \( \hat{W} \geq \hat{W}_o \) as a criterion), supply quantities (Eq. (3)), predicted full-year work (if \( \hat{K} \geq 52 \)), and minimum supply quantities \( A_o, K_o \) (Eq. (8)). These may be compared with actual values (when given) and with other sources of information to obtain evaluations of the model in terms of reliability and reasonableness.

Third, formal statistical tests of various hypotheses may be carried out, using both actual and predicted variables. Finally, additional variables and measures may be estimated, such as aggregate labor supply elasticities (including effects on participation rates) or predicted changes in labor force composition resulting from various policy measures affecting participation and supply.†

When constructing these predictions, however, one should be careful to distinguish between two types of predictions:

---

* Hanoch (R-2067) gives a more detailed description of these corrected variance estimates.
† Hanoch (R-2067) includes some examples.
1. **Unconditional** predictions, in which the selectivity bias variable $\lambda$ (and its coefficients in the four equations) is omitted entirely, using only the other variables appearing in each equation.

2. **Conditional** predictions, which use the information on actual participation. In the case of predicting $w$ and $w_o$, for example, the coefficient of $\lambda$ in each equation is multiplied for each participant in the sample by $\lambda_i$ and added to the unconditional predicted value; but the same coefficient should be multiplied by the estimated (negative) quantity $\bar{\lambda}_i = -\lambda_i p_i / (1 - p_i)$ (Eq. (9)) for each nonparticipant in the total sample. Since $p_i \lambda_i + (1 - p_i) \bar{\lambda}_i = 0$, these conditional wage-offer and reservation-wage predictions are unbiased when applied to the total sample population; they give the same means as the unconditional predictions, but different, more powerful individual predictions. For predicting expected supplies, however, which are conditional on participation, the positive value $\lambda$ should be used for everybody.

Finally, at increased costs, the estimation procedure outlined above may be iterated to yield possibly improved estimates. For example, the Tobit analysis described above, pp. 27-30, may now be repeated, using the imputed wage estimates provided by the 3SLS method. This iteration generates new estimates for $EK$ and for $D^K$, which may now be employed to repeat the system estimates, modifying both the weights $EK/K$ and the three variables $D^K$, $D^W$, and $D^Y$.

Another possible iteration is to use a Tobit-type maximum likelihood analysis, accounting for the lower truncation of both $A$ and $K$ (i.e., at the minimum quantities $A_o$ and $K_o$ corresponding to $W_o$) by incorporating the lower-limit individual estimates of $A_{oi}$ and $K_{oi}$ from the previous step. The weeks equation now requires a "two-limit Tobit" analysis, with a fixed upper limit of 52 and a variable but known lower limit $K_{oi}$. The system of linear equations of the final stage (for

---

*Computer programs for this type of analysis are available, but the cost of computations is increased.*
each iteration of the model may itself be iterated by using residuals in each iteration for the 3LS estimate of the next iteration. However, it has been shown * that such iteration does not necessarily increase efficiency, since the asymptotic properties of the first 3LS estimates are identical to those of the following steps and to those of full-information maximum likelihood (in a perfectly specified unrestricted linear model), whereas the small sample properties are unknown.

Additional gains in efficiency may be sought by other methods, as discussed above, pp. 22-23. The heteroskedastic structure of residuals may be estimated by regressing the squares of residuals in each equation on the predicted value of the dependent variable. This variance structure may then be employed to modify the weighting scheme used (in each equation separately), so as to approximately equalize the variances. However, small sample or limiting properties of such estimates are not known, and efficiency may not be increased. On the other hand, this suggestion involves much higher computation costs, since the observations must be weighted separately for each equation, thus multiplying the order of the weighted moment matrix by the number of equations.

As explained in the Introduction, none of these additional steps are recommended. Investing the same time and money in improving the specification of the model or in applying it to additional sources of data would doubtless yield higher returns.

* See Dhrymes, pp. 371-372.
APPENDIX

SURVEY-WEEK SELECTIVITY BIAS AND ITS CORRECTION BY WEIGHTING

Denote the joint density function corresponding to the random variables \( v_i \) of Eqs. (15) to (17) by \( g^i(v_i) \). The function \( g^i(\cdot) \) differs among individuals (see Eq. (13)), and corresponds to the \( u_{1i} \)-truncated joint-normal distribution of \( u_i \) where \( u_{1i} \leq \sigma_{1i} \). For each participant \((i \leq n\) in the sample), \( v_i \) have zero means:

\[
E_{v_i} = \int g^i(v_i) v_i dv_i = 0 \quad \text{(A-1)}
\]

as shown in Eqs. (9) and (12). Denote by \( \epsilon_i \) the reduced-form residual for \( K_i \): \( \epsilon_i = \delta_2 v_{2i} + v_{4i} \) and \( K_i = K^*_i + \epsilon_i \) in Eq. (20). Among participants, \( \epsilon_i \) has zero mean and zero correlations with the exogenous variables, including \( K^*_i \), since each \( \epsilon_i \) is a linear function of \( v_i \). Also, \( \epsilon_i \) are independent across individuals. The covariances of \( \epsilon_i \) with \( v_i \) are given by

\[
E_{\epsilon_i v_i} = \delta_2 E_{v_{2i} v_i} + E_{v_{4i} v_i}
\]

\[
= \delta_2 (\sum_{k=2}^r b_{ki}^2 b_{k1}) + (\sum_{h=2}^r b_{4h} b_{i1})
\]

\[
= (\delta_2 b_{2i} + b_{4i} b_{i1}) \quad \text{(A-2)}
\]

by Eq. (14). The conditional density \( g^{ci}(v_i|H_i > 0) \), given that \( H_i > 0 \), is derived by applying Bayes' formula to the corresponding joint density \( g^i \), using Eq. (A-1):

\[
g^{ci}(v_i|H_i > 0) = \frac{(K_i/52)g^i(v_i)}{\int (K_i/52)g^i(v_i) dv_i} = \frac{(K^*_i + \epsilon_i)g^i(v_i)}{\int (K^*_i + \epsilon_i)g^i(v_i) dv_i}
\]

\[
= \frac{(K^*_i + \epsilon_i)g^i(v_i)}{K_i \int g^i(v_i) dv_i + E_{\epsilon_i}} = \frac{1}{K_i} \frac{(K^*_i + \epsilon_i)g^i(v_i)}{K^*_i \int g^i(v_i) dv_i + E_{\epsilon_i}}
\]
where the last equality follows, since \( \int g^i(v_{1i}) dv_{1i} = 1 \) (\( g^i \) being a true density) and \( E\epsilon_{1i} = 0 \).

The conditional expectation of \( v_{1i} \) under \( H_i > 0 \) is therefore

\[
E(v_{1i}|H_i > 0) = \int v_{1i} g^{ci}(v_{1i}|H_i > 0) dv_{1i}
\]

\[
= \frac{1}{K_i^*} \left[ K_i^* \int v_{1i} g^i(v_{1i}) dv_{1i} + \int \epsilon_{1i} v_{1i} g^i(v_{1i}) dv_{1i} \right]
\]

\[
= E \frac{\epsilon_{1i} v_{1i}}{K_i^*} = \frac{1}{K_i^*} \left[ (\delta_{2b_2} + \delta_{3b_3}) - (\delta_{2b_2} + b_4)b_{D_1} \right], \quad (A-3)
\]

where these equalities follow, since \( K_i^* \) is exogenous, \( E\epsilon_{1i} = 0 \) (\( i \leq n \), Eq. (A-1)), and the last equality uses Eq. (A-2) above. More explicitly, the biases (expected values of \( v_{1i} \)) in Eqs. (15) to (18), if applied to survey-week workers, are

\[
E(v_{2i}|H_i > 0) = (\delta_2 \sigma_2^2 + \sigma_24) \cdot \frac{1}{K_i^*} - (\delta_2 b_2^2 + b_4) \cdot \frac{D_{i1}}{K_i^*},
\]

\[
E(v_{3i}|H_i > 0) = (\delta_2 \sigma_23 + \sigma_34) \cdot \frac{1}{K_i^*} - (\delta_2 b_23 + b_4) \cdot \frac{D_{i1}}{K_i^*},
\]

\[
E(v_{4i}|H_i > 0) = (\delta_2 \sigma_24 + \sigma_4^2) \cdot \frac{1}{K_i^*} - (\delta_2 b_24 + b_4^2) \cdot \frac{D_{i1}}{K_i^*}. \quad (A-4)
\]

Expressions for the conditional variance matrix of \( v_{1i} \) may also be derived but are quite tedious.

The biases are inversely proportional to \( K_i^* \), and are present in any of these equations if the corresponding residual \( v_{j1i} \) is correlated with the weeks residual \( \epsilon_{1i} = \delta_2 v_{21i} + v_{4i} \). One seemingly feasible method of modifying the model to eliminate these biases is analogous to the treatment of the participation bias under Case 1 in Sec. III. The
variables $1/K^*$ and $D/K^*$ may be estimated separately* and then introduced into the equations linearly, where their corresponding coefficients (defined in Eq. (A-4)) are to be estimated, along with the other coefficients, by linear methods. The implied residuals (e.g., in the wage equation) $\tilde{v}_2$ are:

$$
\tilde{v}_{21} = v_{21} - (\delta_2 \sigma_2^2 + \sigma_{24}) \frac{1}{K_i^*} + (\delta_2 b_2^2 + b_2 b_4) \frac{D_i}{K_i^*},
$$

and may be shown to have zero means, and to be uncorrelated with the exogenous right-side variables within the selective group of survey-week workers.

However, this suggested modification turns out to be self-defeating, because it introduces severe multicollinearity into the equations. $K_i^*$ is a linear combination of all the exogenous variables (including $\lambda_i$), as shown in Eq. (20). $D_i$ is a function of $\lambda_i$ or, equivalently, a function of the index $I_i = (1/\sigma_1)(a_i'x_i - \beta_i'y_i)$, again combining all the exogenous variables. Clearly, the pair of variables $1/K^*$ and $D/K^*$ are bound to be highly collinear with the other variables in each equation (and may be highly correlated with each other), and to make the estimates of coefficients quite unreliable.

A much simpler alternative route is available for eliminating or drastically reducing these survey-week biases. The sample moments in the biased sample of m survey-week workers may be weighted by the inverse of the selection probabilities $K/52$, so that the weighted moments become consistent estimates of the corresponding moments among all participants. To preserve the scale of these moments, and thus to avoid additional heteroskedasticity, † the moments are weighted by $K^*/K$ (or by $EK/K$, where $EK$ is the estimated expected value of $K$, which takes into account the corner-solution restriction $K \leq 52$; see Sec. IV, pp. 27-30).

*Using the exact functional relation $D(I)$ on the probit estimates of $I$, and estimating $K^*$ in a reduced-form equation, as in Sec. IV, p. 26.

†Since some heteroskedasticity is already present, by Eq. (14). A correction for this is costly, and quantitatively of minor consequence.
Since \( K \) is measured and \( K^* \) (or \( EK \)) may be estimated approximately, this weighting scheme is feasible. The same effect may be achieved alternatively by weighting the observations rather than the moments—that is, transforming all the variables (including the constant) through multiplication by the factor \( \sqrt{\frac{K^*_i}{K_i}} \) for each \( i = 1, \ldots, m \). Replacing each variable \( x_i \) by the transformed variable \( \hat{x}_i = \sqrt{\frac{K^*_i}{K_i}} \cdot x_i \) gives a set of equations completely analogous to Eqs. (15) to (20), in terms of the transformed variables, and applicable to the selective subsample \( \{i \leq m\} \). None of the variables is now exogenous, since the endogenous variable \( K \) enters the definition of each transformed variable. Nevertheless, applying familiar linear methods to these equations yields consistent estimates of the parameters.

These informal statements may now be clarified by a more formal treatment, although a complete analysis is tedious. Consider, for example, the weeks supply equation in Eq. (19), where the method is most likely to produce inconsistencies since the transformation involves the same endogenous variable \( K \) as does the dependent variable. In compact form, the equation is

\[
K_i = K^*_i + \varepsilon_i.
\]

Using Eqs. (A-3) and (A-4) above, \( E(\varepsilon_i | H_i > 0) \) is seen to be generally nonzero. Also,

\[
E(\varepsilon_i K^*_i | H_i > 0) = K^*_i E(\varepsilon_i | H_i > 0) = E\varepsilon_i > 0,
\]

implying that least-squares estimates of this equation within the sample of \( m \) are inconsistent. The transformed equation, using the proposed weighting scheme, is derived by multiplying each variable by \( \sqrt{\frac{K^*_i}{K_i}} \) for each \( i \leq m \):

\[
\hat{K}_i = \hat{K}^*_i + \hat{\varepsilon}_i.
\]  

*Except for a slight bias, due to the difference in transitory effects in \( K \) between adjacent years, if the \( K \) used for weighting is measured in the preceding year. See below, Eq. (A-8).
A necessary and sufficient condition for consistency of the OLS estimate of Eq. (A-5) is \( \hat{\epsilon}_i \hat{K}_i^* = 0 \) (not \( \hat{\epsilon}_i = 0 \), since \( \hat{K}_i^* \) is no longer exogenous; \( \hat{K}_i^* = (K_i^*)^{3/2} K_i^{-1/2} \)). This moment is now evaluated, using methods identical to Eqs. (A-3) and (A-4):

\[
\begin{align*}
E(\hat{\epsilon}_i \hat{K}_i^* | H_i > 0) &= E \left( \frac{K_i^*}{K_i} \cdot \epsilon_i \frac{K_i^*}{H_i} | H_i > 0 \right) \\
&= \int \frac{K_i^*}{K_i} \cdot \frac{K_i}{52} \cdot g_i(v_i) dv_i \\
&= \frac{K_i^* \epsilon_i}{(K_i/52) g_i(v_i) dv_i} \\
&= \frac{K_i^* \epsilon_i}{K_i^* + \epsilon_i} = K_i^* \epsilon_i = 0.
\end{align*}
\]

Equation (A-6) shows that the weighting scheme applied in Eq. (A-5) eliminates the bias, and provides consistent estimates of this equation. Moreover, it preserves the scale of the moments, since \( E(\hat{\epsilon}_i \hat{K}_i^* | H_i > 0) = E(\epsilon_i K_i^*) \), where the last term is the unconditional expectation among all participants. Similar evaluation of other equations and moments clearly gives analogous results, since \( K^* \) is eliminated between the weights \( K_i^*/K \) and the probability \( K_i/52 \).

Next, consider the effects of transitory components. Assume

\[
K_i^t = K_i^* + \epsilon_i^t + \epsilon_i^t
\]

and

\[
K_i^{t-1} = K_i^* + \epsilon_i^p + \epsilon_i^{t-1},
\]

where \( K^* \) and \( \epsilon^p \) are constant (between the two adjacent years), and \( \epsilon^t, \epsilon^{t-1} \) are transitory and assumed uncorrelated with each other and with the permanent component \( \epsilon_i^p \).
Repeating the derivation in Eq. (A-6) for this case, using \((K^t/52)\) for the selection probability and \((K^t/K^{t-1})\) for weights, and remembering that the measured weeks variable is \(K^{t-1}\), gives

\[
E(\hat{e}_iK_i^t|H_i > 0) = K_i^t E \left[ \frac{K_i^t(e_i^p + e_i^{t-1})}{K_i^{t-1}} \right]
\]

\[
= K_i^t E \left[ e_i^p + e_i^{t-1} + \frac{(e_i^t - e_i^{t-1})(e_i^p + e_i^{t-1})}{K_i^{t-1}} \right],
\]

by substitution from Eq. (A-7) and some manipulation. Hence,

\[
E(\hat{e}_iK_i^t|H_i > 0) = K_i^t E \left[ \frac{(e_i^t - e_i^{t-1})(e_i^p + e_i^{t-1})}{K_i^{t-1}} \right]
\]

\[
\approx - \text{var } e_i^{t-1}, \tag{A-8}
\]

where the last equality is an approximation derived by replacing expectations by probability limits. On the other hand, the bias in the same original unweighted equation is given by

\[
E(e_iK_i^t|H_i > 0) = E \left[ (e_i^p + e_i^{t-1}) \cdot K_i^t \right]
\]

\[
= E(e_i^p + e_i^{t-1})(e_i^p + e_i^t)
\]

\[
= \text{var } e_i^p. \tag{A-9}
\]

Since the weeks permanent residual variance is estimated to account for most of the total variance of \(K_i^t\), the bias in the unweighted weeks equation (A-9) is larger than in Eq. (A-8). Applying a similar analysis to other equations and moments tends to give similar results.

\*The proportion of the permanent component was estimated as 0.825 in Hanoch (R-1787), p. 5, footnote.
REFERENCES


