The Covariance of Earnings and Hours Revisited*

John M. Abowd, Gary Benedetto, Martha H. Stinson†

December 2007

Abstract

In this paper we examine the earnings covariance matrix generated from a ten-year time series and estimate a variance components model that parameterizes the process generating earnings. We use our estimated variance components to test key hypotheses concerning life-cycle human capital investment and labor supply separately for men and women. Human capital investment models predict that individuals with higher initial earnings have lower growth rates of earnings and that earnings follow a random growth model with individual specific rates of growth due to experience. Life-cycle labor supply models predict that variation in individual productivity affects earnings more than hours supplied. In order to test these hypotheses, we look for permanent individual variance components in the growth rate of earnings and significant auto-correlation in earnings over time. We also test for the presence of a common component of variation between hours and earnings and explore how this component contributes to earnings relative to hours. We look for evidence to support or contradict the predictions of the models using a new data source – a set of SIPP panels linked to administrative tax data on labor market earnings. Our data contain Survey of Income and Program Participation (SIPP) respondents from the five panels conducted by the Census Bureau in the 1990s with linked W-2 wage records filed by employers with the IRS. The sum of these wage records for a given year provides an uncapped annual earnings measure. We use survey information on the number of weeks worked full-time and part-time in a year to estimate annual hours worked. Because of the length of the time period covered (1990-1999), the size of the sample (approximately 230,000 individuals), and the high quality of the earnings measure, these data offer a unique opportunity to re-visit several classic labor economics questions and provide fresh evidence for on-going debates.

*We would like to thank Anja Decressin, Lisa Dragoset, Bryan Ricchetti, Karen Masken, Sam Hawala, Arthur Jones, Jr., Simon Woodcock, Howard Iames, Brian Greenberg, Dawn Haines, and Susan Grad for their help and support in creating the SIPP Synthetic Beta.

†John M. Abowd is the Edmund Ezra Day Professor of Industrial and Labor Relations at Cornell University, Distinguished Senior Research Fellow at the U.S. Census Bureau, Research Associate NBER, Research Affiliate CREST/INSEE, and Research Fellow IZA. Martha Stinson and Gary Benedetto are economists at the U.S. Census Bureau. Contacts: john.abowd@cornell.edu, martha.stinson@census.gov, gary.l.benedetto@census.gov

This research is part of a program jointly sponsored by the Census Bureau, the Social Security Administration, and the Internal Revenue Service to produce public use files that integrate elements of survey and administrative data. The main goal of the program is to develop an innovative confidentiality protection technique that will allow the release of microdata by preserving the confidential identities of survey respondents while maintaining the analytic usefulness of the data. Because the proposed public use files contain lifetime earnings histories from the Social Security Administration's master earnings file (W-2 data under IRS stewardship), the data are likely to attract considerable professional interest once released. Hence it is important to assess their analytic validity by considering questions and models where there is considerable scientific evidence already in the public record.

1 Introduction[1]

In this paper we re-visit a classic labor economics question–the intertemporal labor supply of men–using a new source of panel data. Our purpose is two-fold. First we intend to contribute to the general knowledge of the how individuals make dynamic labor supply decisions and how their earnings evolve over time. Second, we test a new type of public-use data–partially synthetic micro-data without topcoding–in order to determine whether these data can be used in standard micro-data analyses. It is because of this second purpose that we are particularly interested in re-visiting a classic labor economics model. We wish to see if this new type of data will give results that are consistent with what has commonly been found in the past.

The particular model we will re-visit is the individual life-cycle labor supply model. The main tenet of this model is that individuals respond to changes in their wages (*i.e.*, productivity) by changing their hours supplied to the labor market and hence changing their earnings. However productivity variation affects earnings more than hours supplied. We will test this hypothesis following the general method used by Abowd and Card (1989). In particular we are interested in determining whether the random component in the earnings model will have a coefficient relative to the same component in the hours model that is greater than one. We will also investigate the general covariance structure of hours and earnings residuals and investigate whether they appear to contain measurement error and whether there is significant autocorrelation across years,

[1]Disclaimer and acknowledgements: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau, Cornell University, or any of the project sponsors. This work was partially supported by the National Science Foundation Grants SES-9978093, ITR-0427889, and SES-0339191 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging, and the Alfred P. Sloan Foundation. All of the data used in this paper are confidential data. The U.S. Census Bureau supports external researchers' use of these data items; please visit www.sipp.census.gov/sipp/ and click on "Access SIPP Synthetic Data."

supporting the random walk hypothesis for the path of earnings residuals.

Our data stem from a well-known source, the Survey of Income and Program Participation (SIPP), but have been expanded to include administrative earnings records from the Social Security Administration (SSA) and the Internal Revenue Service (IRS). These administrative records are matched to individual survey respondents and include uncapped total earnings for every employer of the matched respondents for the years 1978 to 2003. However because of the sensitive and confidential nature of these data, they are not publicly releasable in their original form. Hence, working in conjunction with IRS and SSA, the U.S. Census Bureau has developed new disclosure protection methods that involve creating synthetic data based on the methodology briefly described in section 3. A trial version of these linked survey-administrative records, called the SIPP Synthetic Beta (SSB) is available for public access.[2] Since the release of synthetic data as a public-use product is still relatively uncommon, the agencies involved in creating the SSB are undertaking significant testing of the data in order to determine whether they actually reproduce the fundamental qualities of the original data. This paper is part of that effort.

2 Background

2.1 Life-cycle Labor Supply Model

We consider a life-cycle labor supply model as developed by MaCurdy (1981). The consumer chooses consumption and leisure in order to maximize lifetime utility subjected to the constraint that he cannot consume more than his earnings and assets over the course of his life. This problem is formally stated as:

$$\max \sum_{t=0}^{T} \frac{1}{(1+\rho)^t} U\left[C(t), L(t)\right]$$

subject to

$$\begin{aligned} A(0) + \sum_{t=0}^{T} R(t)N(t)W(t) &= \sum_{t=0}^{T} R(t)C(t) \\ A(0) &= \text{initial assets} \\ R(t) &= \text{market interest rate} \\ N(t) &= L^* - L(t) \\ L^* &= \text{total time in each period} \\ W(t) &= \text{real wage rate} \\ C(t) &= \text{consumption} \\ L(t) &= \text{leisure} \\ \rho &= \text{subjective discount rate} \end{aligned}$$

[2] Details are available here: http://www.bls.census.gov/sipp/synth_data.html.

The optimal choices of $C(t)$ and $L(t)$ satisfy the budget constraint as well as the following conditions:

$$\begin{aligned} MU_C &= R(t)(1+\rho)^t\lambda \\ MU_L &\geq R(t)(1+\rho)^t\lambda W(t) \\ \lambda &= \text{Lagrange multiplier for the budget constraint} \\ &\quad \text{and marginal utility of wealth in period 0} \end{aligned}$$

The first condition states that the marginal utility of consumption must equal the marginal utility of wealth in period 0, appropriately discounted. The second condition says that the marginal utility of leisure must be at least equal to the marginal utility of the value of the wage. If this second condition holds with equality then $N(t) > 0$ and at least some labor will be supplied. These conditions give rise to functions for $C(t)$ and $N(t)$ that give the optimal values conditional on the marginal utility of wealth, λ, the market interest rate, the discount rate, and the wage rate, $W(t)$. MaCurdy refers to these as the "lamda-constant" functions because λ summarizes everything about a person's initial wealth, lifetime path of wage rates, and preferences that are needed in order to determine labor supply and consumption at any point in time. Blundell et al. (2006) and others commonly refer to these equations as Frisch demand (labor supply) functions because they hold the marginal utility of wealth, and not wealth itself, constant. Besides λ, no information from outside the time period is needed to make decisions about how much to work and how much to consume. If one assumes a particular form of the utility function, an estimable equation for $N(t)$ can be developed. MaCurdy suggests the following utility function which gives rise to the following equation for $N(t)$.

$$\begin{aligned} U_i &= Y_{1i}(t)\left[C_i(t)\right]^{\omega_1} - Y_{2i}(t)\left[N_i(t)\right]^{\omega_2} \\ \ln N_i(t) &= F_i + \eta\sum_{k=0}^{t}\left[\rho - r(k)\right] + \eta\ln W_i(t) + u_i(t) \\ Y_{2i}(t) &= \sigma_i + u_i^*(t) \\ F_i &= \eta * \left\{\ln\lambda_i - \sigma_i - \ln\omega_2\right\} \\ \eta &= \frac{1}{(\omega_2 - 1)} \\ u_i(t) &= \eta u_i^*(t) \\ r(0) &= \rho, \ln(1+r(t)) \approx r(t), \ln(1+\rho) \approx \rho \\ \text{if } r(t) &= r \text{ for all } t, \text{ then } \eta(\rho - r)t = \eta\sum_{k=0}^{t}\left[\rho - r(k)\right] \end{aligned}$$

Thus, an estimable equation for $N_i(t)$ contains a person effect, a time effect, a wage effect, and an error term. The term of interest is η which is the intertemporal substitution elasticity. This term tells how an individual will change her supply of labor across time periods in response to observing changes in wage

rates. MaCurdy describes these as "evolutionary" changes, or changes along a wage path over the lifetime of the individual.

2.2 Our Statistical Model Following the Abowd and Card Model

Abowd and Card (1989) further refine this model of life-cycle labor supply. Specifically they begin with hours and earnings equations as follows:

$$\begin{aligned} \log h_{it} &= a_{it} + \eta \log \theta_{it} + \eta \log \lambda_{it} \\ \log g_{it} &= a_{it} + (1+\eta) \log \theta_{it} + \eta \log \lambda_{it} \end{aligned}$$

where:

$$\begin{aligned} h_{it} &= \text{hours of individual } i \text{ in period } t \\ g_{it} &= \text{earnings of individual } i \text{ in period } t \\ \theta_{it} &= \text{wage rate for individual } i \text{ in period } t \\ a_{it} &= \text{component controlling for individual} \\ &\quad \text{and time period specific preference variation} \\ \lambda_{it} &= \text{marginal utility of consumption, equal to} \\ &\quad \text{marginal utility of wealth appropriately discounted} \\ \eta &= \text{intertemporal elasticity of substitution} \end{aligned}$$

Next, they specify how the marginal utility of consumption develops over time and substitute this into the hours and earnings equations:

$$\log \lambda_{it} - \log \lambda_{it-1} = \log(\frac{1+\rho}{1+r_t}) + \phi_{it}$$

where:

$$\begin{aligned} \rho &= \text{subjective discount rate} \\ r_t &= \text{real rate of return on assets between period } t \text{ and } t+1 \\ \phi_{it} &= \text{one period ahead prediction error} \\ &\quad \text{in log marginal utility of consumption} \end{aligned}$$

Substituting for the change in the marginal utility of consumption and adding measurement error terms gives us:

$$\begin{aligned} \Delta \log g_{it} &= \eta \log \left(\frac{1+\rho}{1+r_t} \right) + \Delta a_{it} + (1+\eta) \Delta \log \theta_{it} + \eta \phi_{it} + \Delta u_{it} \\ \Delta \log h_{it} &= \eta \log \left(\frac{1+\rho}{1+r_t} \right) + \Delta a_{it} + \eta \Delta \log \theta_{it} + \eta \phi_{it} + \Delta v_{it} \end{aligned}$$

to describe observed changes in log earnings and log hours, where:

$$\begin{aligned} u_{it} &= \text{measurement error in earnings of individual } i \text{ in period } t \\ v_{it} &= \text{measurement error in hours of individual } i \text{ in period } t \end{aligned}$$

We prefilter the data following the same procedures as Abowd and Card (1989). The raw change data are adjusted by estimating a regression of changes in log earnings and log hours on time-varying person characteristics to estimate the fixed effects. The residuals are then taken to be the adjusted changes in log earnings and log hours which follow:

$$\begin{aligned}\Delta \log \widetilde{g}_{it} &= \Delta \widetilde{a}_{it} + (1+\eta)\Delta \log \widetilde{\theta}_{it} + \eta \phi_{it} + \Delta u_{it} \\ \Delta \log \widetilde{h}_{it} &= \Delta \widetilde{a}_{it} + \eta \Delta \log \widetilde{\theta}_{it} + \eta \phi_{it} + \Delta v_{it}\end{aligned}$$

where the tilde denotes that the measure is adjusted for the observable characteristics in the original regression. We create a balanced panel of 9 periods (1991-1999) of observed changes in log earnings and log hours for each individual (some individuals have missing data for one or more rows of the data matrix in order to make the balanced panel). We use the assumptions from Abowd and Card about ϕ_{it} and $\Delta \widetilde{a}_{it}$, namely:

$$\begin{aligned}Cov(\phi_{it}, \phi_{it-j}) &= 0 \text{ for } j \neq 0 \\ \Delta \widetilde{a}_{it} &= \Delta \psi_{it} + \Delta \zeta_{it}\end{aligned}$$

where ψ_{it} are transitory preference shocks and ζ_{it} are permanent preference shocks such that:

$$\begin{aligned}Var(\psi_{it}) &= \sigma_{\psi}^2 \\ Cov(\psi_{it}, \psi_{is}) &= 0 \text{ for } s \neq t \\ Cov(\zeta_{it} - \zeta_{it-1}, \zeta_{it-j} - \zeta_{it-j-1}) &= 0 \text{ for } j > 0\end{aligned}$$

Since η is restricted by the model to be strictly positive, the ratio $\frac{(1+\eta)}{\eta}$ is restricted to be greater than one. Abowd and Card normalize the coefficient on the log wage rate in the hours equation to be one and define $\mu = \frac{(1+\eta)}{\eta}$ as the coefficient on the log wage rate in the earnings equation. Therefore, we can decompose the change in each of log earnings and log hours into three identifiable additive components:

$$\begin{aligned}\Delta \log \widetilde{g}_{it} &= \mu z_{it} + \omega_{1it} + \varepsilon_{it} \\ \Delta \log \widetilde{h}_{it} &= z_{it} + \omega_{2it} + \varepsilon_{it}\end{aligned}$$

where:

$$\begin{aligned}\mu &= (1+\eta)/\eta \\ z_{it} &= \eta \Delta \log \theta_{it} \\ \omega_{1it} &= \Delta u_{it} + \Delta \psi_{it} \\ \omega_{2it} &= \Delta v_{it} + \Delta \psi_{it} \\ \varepsilon_{it} &= \eta \phi_{it} + \Delta \zeta_{it}\end{aligned}$$

The covariance structure of ω_{1it} and ω_{2it} are, by construction, first-order-one moving average processes (MA(1)) with MA(1) parameter equal to -1. The ε_{it}

term contributes only to the contemporaneous variances and covariances. This leaves us with a block-diagonal covariance matrix of dimension $18N$ where N is the number of individuals used in the estimation and the 18×18 non-zero matrix along the diagonal has the following parameterization:

$$\Sigma = \begin{array}{c} \begin{array}{cccccc} 1 & & ... & 10 & & ... \end{array} \\ \left[\begin{array}{cccccc} \mu^2 Var(z_{it}) + 2(\sigma_\psi^2 + \sigma_u^2) + \sigma_\varepsilon^2 & ... & \mu Var(z_{it}) + 2\sigma_\psi^2 + \sigma_\varepsilon^2 & ... & 1 \\ \mu^2 Cov(z_{it}, z_{it-1}) - (\sigma_\psi^2 + \sigma_u^2) & ... & \mu Cov(z_{it}, z_{it-1}) - \sigma_\psi^2 & ... & 2 \\ \mu^2 Cov(z_{it}, z_{it-2}) & ... & \mu Cov(z_{it}, z_{it-2}) & ... & 3 \\ ... & ... & ... & ... & ... \\ \mu Var(z_{it}) + 2\sigma_\psi^2 + \sigma_\varepsilon^2 & ... & Var(z_{it}) + 2(\sigma_\psi^2 + \sigma_v^2) + \sigma_\varepsilon^2 & ... & 10 \\ \mu Cov(z_{it}, z_{it-1}) - \sigma_\psi^2 & ... & Cov(z_{it}, z_{it-1}) - (\sigma_\psi^2 + \sigma_v^2) & ... & 11 \\ \mu Cov(z_{it}, z_{it-2}) & ... & Cov(z_{it}, z_{it-2}) & ... & 12 \\ ... & ... & ... & ... & ... \end{array}\right] \end{array}$$

After reviewing the sample estimates of Σ shown in tables 2 and 3, we decided to use a first-order-two (MA(2)) process to estimate the process that the wage rate (z_{it}) follows. The coefficients of this MA(2) process are estimated as part of the statistical model for the complete covariance matrix of earnings and hours changes and are reported in Tables 4 and 5.

3 Data Description: SIPP Synthetic Beta (SSB)

The project to create a new SIPP public use file began in 2001 when a special Federal Treasury Regulation went into effect that allowed the U.S. Census Bureau to obtain administrative earnings data from the Social Security Administration and the Internal Revenue Service that matched to respondents in certain SIPP panels. Census's primary mission was to integrate the administrative and survey data and make a product that would be available to researchers interested in studying earnings and federal benefits issues. To this end, we created an extract from the five SIPP panels conducted in the 1990s (beginning years of 1990, 1991, 1992, 1993, 1996 respectively) and merged on the available administrative data. We refer to these data as the gold-standard because they represent the kind of data that would be compiled for analysis by a researcher working in a confidential protected area at either SSA or Census.

The gold standard contains many variables not used in our analysis. For the purposes of brevity, we will describe only the variables used here but a complete description of the publicly available variables can be found on the SIPP home page.[3] Individuals had to be at least 15 years of age by the time of their second SIPP interview in order to be included in the gold standard. We used self-reported race, coded as black/non-black, gender, marital status at the time

[3] Please see www.sipp.census.gov/sipp/ and click on "Access SIPP Synthetic Data." Variable descriptions can be found in "Technical Description SIPP Synthetic Beta."

of the second interview or closest available interview, birthdate, and education, where we took the highest level ever reported during the survey. These variables were available for every individual in the gold standard. We also used indicators for disabled, foreign born, Hispanic and continuous variables for number of kids in the family at the time of the second interview, collected in a topical module in different waves depending on the panel. These variables were sometimes missing due to item non-response within the survey.

The hours variable was significantly more complicated. The SIPP collects information on usual weekly hours for at most two jobs. We created a monthly total hours worked variable by multiplying the usual weekly hours for each job by 4.33 if the person worked the entire month at that job or else the fraction of the month worked if it was a beginning or ending month. We then summed the hours across jobs to create a monthly total. For individuals who did not miss any of the interviews during the panel, we were able to create an annual hours measure for years completely covered by the survey by summing these monthly totals. However for individuals who missed interviews within a panel and for years not covered, either in full or part, by the survey, these annual hours data were incomplete.

The earnings variable was taken from the Detailed Earnings Records (DER) extract from SSA's Master Earnings File (MEF). These records included historical reports of annual earnings, by employer, from 1978-2003 and were the amounts the employer reported to the IRS in box 1 of the W-2 tax form. SIPP respondents who either provided a Social Security Number (SSN) or did not refuse to give one were matched against several master Census databases to either check the validity of their SSN or attempt to fill in a missing SSN. Those who explicitly refused to provide the survey with an SSN were not included in this exercise. People who passed the validation process and had a "validated" SSN were then matched against the MEF at SSA to create the earnings history. Hence unlike hours, SIPP respondents either had a complete earnings history or no history at all. For individuals with DER data, we created annual earnings measures from 1990-1999 by summing the totals earned from each employer in a year.

Finally labor market experience as of 1989 was created using the life-time earnings histories taken from the Summary Earnings Records (SER), a separate extract from the Master Earnings File. These records contain total wages and salary paid to an individual from all employers, up to the maximum taxable under FICA. These records have the advantage of beginning in 1951 and hence are ideal for creating a number of years worked variable. Labor market experience is then increased from 1990-1999 every time an individual has a positive earnings report in the DER.

After creating the gold standard file, we used multiple imputation methods to complete (i.e. impute) all the missing values. There were three main reasons for missing data. First, failure to answer the hours question within a panel; second, lack of hours data in a year not covered by a particular panel; third, failure to provide a valid SSN that linked to earnings and labor force experience. After completing missing values, we synthesized the variables in the gold standard,

with a few exceptions.[4] Of the variables we use for this paper, gender and marital status remained unsynthesized but all other variables were synthesized.

In order to preserve exact logical relations among the variables, the first step of the missing data imputation process, and the first step of the data synthesizing process, is to implement a binary tree of parent-child relations among all the variables. This tree guides the execution of first, the missing data imputation and then, the synthetic data phase. We created the binary tree to organize the data processing by summarizing all of the assumptions and logical restrictions that must be preserved in the final data product.

The top level of this binary tree contains all variables that exhibit no logical dependencies on any other variables in the file, for example birth date. The tree has nine levels. At each level below the top, variables depend upon their parents, and are only processed when appropriate. In the intermediate levels of the tree, a variable can be both a parent and a child, for example, whether or not there is a second marriage is a child of the same variable for the first marriage and a parent of the variable for the third marriage. The terminal level and all leaves of the binary tree contain only child variables.

For each iteration of the missing data imputation phase and again during the synthesis phase, we estimate a joint posterior predictive distribution for all of the required variables according to the following protocol. At each node of the parent/child tree, a statistical model is estimated for each of the variables at the same level. The statistical model is a Bayesian bootstrap, logistic regression, or linear regression (possibly with transformed inputs). All statistical models are estimated separately for detailed groups of individuals based on the values of categorical variables that include both demographic and economic controls. Logistic and linear regressions also include additional linear controls that are selected from a long list of potential right-hand-side control variables on the basis of the Bayes Information Criterion. Once the analyst specifies the grouping variables and their associated control variables, the estimation of a proper posterior predictive distribution from which to impute or synthesize, as appropriate, is fully automated. On the basis of the estimated models, and taking proper account of parameter uncertainty, each variable is imputed (missing data phase) or synthesized (synthetic data phase) conditional on all values of all other variables for that individual. The missing data phase included nine iterations of estimation. The posterior predictive density, when estimated using the Sequential Regression Multivariate Imputation (SRMI) method, must be fit by iterative re-estimation of each of the equations within each of the conditioning groups. The first iteration uses arbitrary starting values for all of the missing data, generally computed from an easily implemented Bayesian Bootstrap.

[4] Four variables were not synthesized: gender, marital status, type of initial OASDI benefit (TOB initial), and type of OASDI benefit in the year 2000 (TOB 2000). We also did not perturb the relationship between spouses. The synthetic data contains the original link between a married individual and his or her spouse. Thus in actuality, each individual's record contains four additional unsynthesized variables: spouse gender, spouse marital status, spouse TOB initial, and spouse TOB 2000. Spouse gender and marital status, by definition, do not provide any additional information beyond the gender and marital status of the respondent since same-sex couples were not recorded in the 1990s SIPP panels.

Subsequent iterations use values sampled from the previous iteration's PPD. There is no formal test for convergence; however, experience has shown that the process generally settles down after a few iterations. Since computation time was a significant constraint, we did not restart the missing data PPD estimation every time we made a minor adjustment to the data specification; instead, we added additional iterations of the SRMI using missing data sampled from the previous iteration's estimated PPD. The process terminated at the ninth iteration, whose PPD was used to sample all of the missing data The synthetic data phase occurred on the tenth iteration and did not require any iteration. Four missing data implicates were created. These constitute the completed data files that are the inputs to the syntihesis phase. Four synthetic implicates were created for each missing data implicate. Thus, there are a total of sixteen synthetic implicates in the SSB. Further details about the data completion and synthesis used to create the SSB can be found in Abowd, Stinson and Benedetto (2006).

The source SIPP records for the SSB were derived from five different SIPP panels, each with its own sampling frame. In order to combine data from these five sources to provide meaningful population estimates, a new weight was created. The process is described in detail in Abowd, Stinson and Benedetto (2006). For completeness of this exposition, we summarize the process here. The 1996 SIPP sampling frame (unit sample only) was recreated using the micro-data from Census 2000. Every record in the SSB was matched to the same individual in Census 2000–either exactly, based on Census internal identifiers, or approximately based on probabilisitic record linking. We created provisional design weights that were representative of the civilian non-institutional population aged 18 or older as of April 1, 2000. These weights were then raked to the standard Census population controls, which are based on sex, race, ethnicity, and age groups as of April 2000. The final weight from this process appears on all four missing data implicates. A synthetic version of this weight appears on all sixteen synthetic implicates. These weights may be used to construct estimates that are representative of the civilian, non-institutional population age 18 and older as of April 1, 2000. No other weights are provided on the SSB; however, synthetic birth and death dates are available and can be used to create samples representative of other dates if appropriate control totals are available. In keeping with the comparative exercise that motivates this paper no weights were used in the estimation of any of the covariance matrices reported below; however, the weights were used in all of the analytical validity testing reported in Abowd, Stinson and Benedetto (2006).

After the creation of the gold standard, the completion of missing data, and the synthesis of most of the variables, we proceeded to estimate the model described earlier in section 2. We first-differenced annual log earnings and log total hours and then regressed these variables on marital status, education level, male, black, disabled, foreign born, Hispanic, number of kids in the family, a quartic in age, unrestricted time effects, and a quartic in labor force experience. We then created the variance-covariance matrix of the residuals for each dependent variable from each year from this regression and fit our model to this 18x18 matrix (the dimension of 18 comes from 9 years of data for each of 2 dependent

variables).

4 Analytical Validity

One of the main purposes of this paper is to test whether these synthetic variables closely mimic the characteristics of their gold standard counterparts, in other words whether analyses using these variables will produce statistically valid results. We define statistical validity according to Rubin as:

> First and foremost, for statistical validity for scientific estimands, point estimation must be approximately unbiased for the scientific estimands averaging over the sampling and posited nonresponse mechanisms. ... Second, interval estimation and hypothesis testing must be valid in the sense that nominal levels describe operating characteristics over sampling and posited nonresponse mechanisms. (1996, p. 474)

This definition should be modified to include the phrase "confidentiality protection mechanisms" wherever "nonresponse mechanisms" appears.

Thus we run our analysis multiple times using the multiple implicates created by the completion and synthesis processes. We combine the results from the four completed data implicates and from the 16 synthetic implicates using the statistical formulae described below to calculate the variances. We then compare our estimates from the synthetic data to our estimates from the completed data, which we will euphemistically call the "truth" since it is the best available comparison data. If the estimates are unbiased and the variances of the estimates are such that inferences drawn about the estimates are the similar to the inferences in the completed (*i.e.*, "true") data, then the data are statistically valid[5]. We now give the formulae that we use to combine the separate estimates from each implicate.

4.1 Missing Data Only

In our use of the classic Rubin (1987) missing data application, Y_{mis} is imputed m times by sampling from $p\left(Y_{mis}\left|D\right.\right)$, the posterior predictive distribution of Y_{mis} given D, the non-missing data. The completed data consist of m sets $D^{(\ell)} = \left\{D, Y_{mis}^{(\ell)}\right\}$, where $Y_{mis}^{(\ell)}$ is the ℓ^{th} draw from $p\left(Y_{mis}\left|D\right.\right)$ and is called the ℓ^{th} implicate. To create the SSB, we sampled four times and created four implicates $D^{(1)}$, $D^{(2)}$, $D^{(3)}$, and $D^{(4)}$. Inference is based on the following

[5] We compare the synthetic data to the completed data in order to separately determine how well the synthesizer performs. If we compared the synthetic data directly to the underlying gold standard data, the effects of completing missing values would be confounded with the effects of synthesizing variables.

formulae:

$$\text{statistic calculated on each implicate file}$$
$$q^{(\ell)} = q\left(D^{(\ell)}\right).$$

The statistic is calculated separately for each implicate and then averaged across implicates as the next formula indicates:

$$\text{average of the statistic across implicates}$$
$$\bar{q}_m = \sum_{\ell=1}^{m} \frac{q^{(\ell)}}{m}.$$

The statistic $\bar{q}_m$ is the new quantity of interest and will serve as the basis for comparison with the synthetic data. Analytic validity requires that synthetic data reproduce $\bar{q}_m$, on average, and that inferences made about $\bar{q}_m$ remain the same, as expressed by the confidence interval associated with $\bar{q}_m$. In order to draw proper inferences, the correct variance measure must be used. The variance of $\bar{q}_m$ has two parts. The first part is commonly referred to as the "between-implicate" variance, defined by the following formula:

$$\text{variance of the statistic across implicates}$$
$$b_m = \sum_{\ell=1}^{m} \frac{\left(q^{(\ell)} - \bar{q}_m\right)\left(q^{(\ell)} - \bar{q}_m\right)'}{m-1}$$

The measure b_m tells how much variation has been introduced by the multiple draws from the posterior predictive distribution. The second component of the overall variance of $\bar{q}_m$ is calculated by averaging the within implicate variance across implicates. We define the variance of $q^{(\ell)}$ for each implicate ℓ and the average across implicates as follows:

$$\text{variance of the statistic on each implicate file}$$
$$u^{(\ell)} = u\left(D^{(\ell)}\right)$$

and

$$\text{average variance of the statistic across implicates}$$
$$\bar{u}_m = \sum_{\ell=1}^{m} \frac{u^{(\ell)}}{m}.$$

The total variance is then calculated as a weighted sum of the between implicate variance and the average within implicate variance, defined as follows:

$$\text{total variance of the average statistic across implicates}$$
$$T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

When n and m are large, inference is based on $(\bar{q}_m - Q) \sim N(0, T_m)$. When m is moderate and the estimator $\bar{q}_m$ is univariate (*i.e.,* $c = 1$), inference is based on $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, where the degrees of freedom ν_m are defined as

$$\nu_m = (m-1)\left(1 + \frac{\bar{u}_m}{\left(1 + \frac{1}{m}\right) b_m}\right)^2$$

Proofs and further details can be found in Rubin (1987, 1996).

4.2 Missing and Partially Synthetic Data

For each completed data set, we partially synthesized r implicates by sampling from $p\left(Y_{rep}|D^{(\ell)}\right)$. Denote the r completed partially synthetic data sets as $D^{(\ell,k)}$. For the SSB, $r = 4$ and hence there were 16 synthetic implicates: $D^{(1,1)}, D^{(1,2)}, D^{(1,3)}, D^{(1,4)} ... D^{(4,1)}, D^{(4,2)}, D^{(4,3)}, D^{(4,4)}$. In order to compare to the completed data, we first calculate the statistic of interest for each of the 16 implicates:

statistic calculated on each implicate file

$$q^{(\ell,k)} = q\left(D^{(\ell,k)}\right).$$

Then, we average across the four synthetic implicates that correspond to a given missing data implicate creating $\bar{q}^{(1)}$, $\bar{q}^{(2)}$, $\bar{q}^{(3)}$, $\bar{q}^{(4)}$ according to the formula:

average of the statistic across the synthetic implicates

$$\bar{q}^{(\ell)} = \sum_{k=1}^{r} \frac{q^{(\ell,k)}}{r}$$

Finally, we average across all 16 implicates to create $\bar{q}_M$. This final average can then be compared to the $\bar{q}_m$ created from the missing data implicates only:

average of the statistic across synthetic and missing data implicates

$$\bar{q}_M = \sum_{\ell=1}^{m}\sum_{k=1}^{r} \frac{q^{(\ell,k)}}{mr} = \sum_{\ell=1}^{m} \frac{\bar{q}^{(\ell)}}{m}.$$

The variance calculations for data that have been completed and synthesized must also account for the additional source of variation that comes from synthesizing. Thus, we calculate the "between synthetic implicate" variance using the following formula:

variance of the statistic due to variation in synthetic implicates

$$b^{(\ell)} = \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{r-1}.$$

This formula quantifies the variation introduced by differences between two synthetic implicates that were generated from the same missing data implicate,

i.e., deviations of the synthetic implicate from the average across both synthetic implicates $q^{(\ell,k)} - \bar{q}^{(\ell)}$. We then average this variance over the missing data implicates:

$$\text{average of } b^{(\ell)} \text{ over missing data implicates}$$
$$b_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{m\,(r-1)} = \sum_{\ell=1}^{m} \frac{b^{(\ell)}}{m}.$$

The next source of variation comes from the multiple implicates due to missing data completion. This variance is calculated using the deviations of the average for a missing data implicate from the overall average, *i.e.,* $\bar{q}^{(\ell)} - \bar{q}_M$. This is the "between missing data implicate" variance:

$$\text{variance of the statistic due to variation in missing data implicates}$$
$$B_M = \sum_{\ell=1}^{m} \frac{\left(\bar{q}^{(\ell)} - \bar{q}_M\right)\left(\bar{q}^{(\ell)} - \bar{q}_M\right)'}{m-1}.$$

Finally, the last source of variance comes from the within implicate variance, which is averaged across the synthetic implicates for a given missing data implicate and then averaged across all the implicates according to the formulae:

$$\text{variance of the statistic on each implicate file}$$
$$u^{(\ell,k)} = u\left(D^{(\ell,k)}\right),$$

$$\text{average variance of the statistic across synthetic implicates}$$
$$\bar{u}^{(\ell)} = \sum_{k=1}^{r} \frac{u^{(\ell,k)}}{r}$$

and

$$\text{average variance of the statistic across synthetic and missing data implicates}$$
$$\bar{u}_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{u^{(\ell,k)}}{mr} = \sum_{\ell=1}^{m} \frac{\bar{u}^{(\ell)}}{m}$$

The total variance is, once again, a weighted sum of the different sources of variation–between synthetic implicate, between missing data implicate, and within implicate:

$$\text{total variance of the average statistic across implicates}$$
$$T_M = \left(1 + \frac{1}{m}\right) B_M - \frac{b_M}{r} + \bar{u}_M.$$

T_M is the variance used to draw inferences about $\bar{q}_M$ and variation introduced by the synthetic and missing data implicates must not be so large that the

inferences will be substantially different from those drawn using $\bar{q}_m$ and T_m. When n, m and r are large, inference is based on $(\bar{q}_M - Q) \sim N(0, T_M)$. When m and r are moderate and the estimator $\bar{q}_M$ is univariate (*i.e.,* $c = 1$), inference is based on $(\bar{q}_M - Q) \sim t_{\nu_M}(0, T_M)$ where the degrees of freedom ν_M are defined as

$$\nu_M = \frac{1}{\left(\frac{\left(\left(1+\frac{1}{m}\right)B_M\right)^2}{(m-1)T_M^2} + \frac{(b_M/r)^2}{m(r-1)T_M^2}\right)}$$

Proofs and details can be found in Reiter (2004).

When estimating the first and second moments that enter the Abowd and Card covariance models, we used the conventional estimators of means and variance/covariances. When estimating the variance covariance matrix of these moments, we replaced the conventional fourth-moment based estimator that Abowd and Card used with a bootstrap estimator based on 100 bootstrap samples of the same size as the original, sampled with replacement.

5 Statistical Results

We begin by providing a brief summary of the data sample we used. We then describe the results from estimating our model using the four completed implicates. We view these results as most closely corresponding to the previous analysis done by Abowd and Card. After discussing the implications of our results for the economic model, we then compare the results from the completed data to the results from the synthetic data. We end with discussions about the analytic validity of the synthetic data.

5.1 Summary Statistics

Table 1 presents summary statistics for men who were continuously employed from 1990 to 1999 for both the completed and synthetic data. This sample most closely corresponds to the one analyzed by Abowd and Card; hence, it facilitates comparisons to their work. The first thing to notice is that all of the means agree very closely between the completed and synthetic samples except for the annual change in log hours, which is very different in the synthetic data. This difference signals that the models we fit below may show some discrepancies between the completed and synthetic data, as is indeed the case.

Table 2 presents the entire variance-covariance matrix of log earnings and log hours changes, adjusted for fixed experience and other human capital effects for the same sample of continuously employed men estimated from the completed data. The variances and covariances of log earnings and log hours, respectively, are displayed on the diagonal and in the lower triangle of the panels. Correlations are shown in the upper triangle. The cross-covariances of log earnings with log hours and the cross correlations are shown in separate panels. Table 6 displays the average distance in probability measure of the t-statistics from zero formed from independent null hypotheses that a given element of the

covariance matrix estimated on the completed data is the same as the corresponding element from the matrices estimated in Abowd & Card. Using table 6 and table 2, one can see evidence that the pattern of data in these matrices is remarkably similar to the pattern for all three of the data sources that Abowd and Card originally examined. We did not formally test for similarity; however, neither the variance of log earnings nor the variance of log hours is stationary. Both variables display a strong negative serial correlation at the first lag, which is much smaller at lag 2, and is essentially zero thereafter. The cross-correlations of log earnings and log hours, however, are much weaker in the completed data than in any of the three sources originally studied by Abowd and Card. It appears that the requirement that the completed data fill-in the missing SIPP hours data (in order to prevent confidentiality compromises in the synthetic data arising from reverse engineering of the panel source) has led to some attenuation of the relation between earnings and hours that one finds when using only sample individuals with complete data (as Abowd and Card, and MaCurdy both did).

Table 3 presents the entire variance-covariance matrix of log earnings and log hours changes estimated from the synthetic data. Table 7 gives the percentage overlap of 95% confidence intervals around the point estimates of the elements of the variance-covariance matrices from the completed and synthetic data (where the size of the confidence interval from the completed data serves as the denominator). Examining table 7, a couple patterns become clear. First, the log earnings covariance/correlation matrix is essentially identical to the one in Table 2. The log hours covariance/correlation matrix is very similar to the one in table 2. But, the cross-covariance/cross-correlation matrix in the synthetic data is essentially zero. The synthetic data did not preserve the cross-correlation relation between these two individual-level time series. As of this writing, we do not have an explanation for this finding.

5.2 Structural Results from the Completed Data

Table 4 presents the results of the key structural parameters estimated for the same sample of continuously employed men using the completed data. This table is directly comparable to the two-component models fit by Abowd and Card. We used the two-component model because the third component (the variance of the contemporaneous shocks to changes in log earnings and log hours) was consistently going to the lower boundary (zero) in our optimization routines for both the completed and synthetic data. The critical structural parameter, μ, has an estimated value of about 1.3 (implying a point estimate of the intertemporal labor supply elasticity of about 3.3). The 95% confidence interval around this estimate contains the corresponding point estimates from Abowd and Card on the PSID and SIME/DIME for the two-component model and on the PSID, PSID SEO excluded, and SIME/DIME for the three-component model. The values of the combined measurement error and transitory shock components are quite precisely estimated and imply a first-order serial correlation which is positive but not significantly different from zero.

5.3 Comparing Results from Synthetic Implicates to Completed Implicates

Table 5 presents results from the synthetic implicates. In the synthetic data, the point estimate of μ is closer to one than in the completed data, and the two-standard deviation confidence interval is actually smaller, and does not cover the interval from Table 4. These results, which indicate that an analyst would infer a larger intertemporal elasticity of labor supply from the synthetic data, are due entirely to the failure of the synthesizing process to capture the properties of the cross-covariance matrix of log earnings and log hours changes. In particular, it suggests that the synthetic data consistently underestimate the covariance of earnings and hours. Thus, an important goal for any re-estimation of the data synthesizer would be to diagnose and repair this problem. On the bright side, the 95% confidence interval around μ from the completed data entirely covers the corresponding interval from the synthetic data, and the synthetic data interval accounts for 46% of the completed data interval. Moreover, the synthetic data gave qualititatively identical results on the ω components of variance (the combined effects of measurement error and transitory preference shocks) and had the same difficulty distinguishing the remaining component of variance, the common economic shock. Hence, a researcher using these synthetic data to do model selection would have arrived at the same set of models to estimate from the completed data.

6 Conclusion

Clearly, the completion and synthesis did a very good job of preserving means and distributions of many variables, but struggled quite a bit with the hours arrays from the SIPP. These struggles seem largely due to the fact that there was so much missing data with which to work and, as a result, less data on which to build the models used in the multiple imputation process. Of course, part of the issue is that we are still very much in the early stages of learning how to build good models to estimate the posterior predictive distributions used in multiple imputation for data completion and synthesis. This issue is compounded with the fact that the SIPP Synthetic Beta project attempted such a large scale synthesis that we were unable to dedicate a sufficient amount of variable-specific attention to every variable in the list. Our hope is that the broader research community will use the SSB data for a large variety of analyses with the idea that their analyses can then also be performed by Census staff on the confidential, completed data. The researchers would then get the benefit of receiving the disclosable results of their analysis on the confidential data, and we would receive the benefit of thoroughly documenting the strengths and weaknesses of the current version of the SSB. We can then use the ever-improving synthesis techniques and a more definitive list of what needs to be improved to produce more analytically reliable, partially synthetic, public-use micro-datasets in the future.

References

[1] Abowd, John M. and David Card (1989). "On the Covariance Structure of Earnings and Hours Changes." *Econometrica* Vol. 57, No. 2, 411-445.

[2] Abowd, John M., Martha H. Stinson, Gary Benedetto (2006). "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project" at http://www.bls.census.gov/sipp/synth_data.html (cited November 26, 2007).

[3] Blundell, Richard, Luigi Pistaferri, and Ian Preston (2006) "Consumption Inequality and Partial Insurance," Institute for Fiscal Studies, University College London working paper (December).

[4] MaCurdy, Thomas E. (1981). "An Empirical Model of Labor Supply in a Life-Cycle Setting." *The Journal of Political Economy* Vol. 89, No. 6, 1059-1085.

[5] Reiter, J. P. (2004). "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation," *Survey Methodology* 30, 235-242.

[6] Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys* (Hoboken, NJ, Wiley Classics Library).

[7] Rubin, D.B. (1996) "Multiple Imputation after 18+ Years," *Journal of the American Statistical Association*, 91, 473-489.

Table1: Summary Statistics

Annual Hours and Annual Earnings (1999 dollars)	Year	Completed Data Earnings	Completed Data Hours	Synthetic Data Earnings	Synthetic Data Hours
	1990	14862	1097	15139	1033
		(286)	(65)	(35)	(37)
	1991	15072	1058	15417	1025
		(295)	(68)	(205)	(33)
	1992	15923	1094	16138	1044
		(509)	(66)	(423)	(26)
	1993	16054	1117	16551	1062
		(325)	(47)	(354)	(29)
	1994	16371	1101	16771	1056
		(345)	(11)	(236)	(16)
	1995	16981	1066	17254	1028
		(291)	(78)	(180)	(71)
	1996	17444	1148	17818	1101
		(322)	(84)	(335)	(16)
	1997	18287	1149	18625	1088
		(311)	(82)	(289)	(24)
	1998	19395	1147	19693	1091
		(518)	(26)	(363)	(17)
	1999	19661	1161	21133	1068
		(293)	(36)	(2213)	(12)
Changes in Log Annual Earnings (x100) and Log Annual Hours (x100)	Change	Change in Earnings	Change in Hours	Change in Earnings	Change in Hours
	1990-1991	8.50	-9.09	8.61	4.33
		(0.46)	(9.15)	(0.84)	(1.44)
	1991-1992	10.07	2.44	10.13	2.35
		(0.32)	(3.94)	(0.86)	(4.59)
	1992-1993	9.32	7.68	7.13	2.36
		(0.38)	(8.72)	(3.63)	(4.81)
	1993-1994	7.71	2.11	7.31	0.17
		(0.48)	(5.36)	(1.69)	(2.86)
	1994-1995	6.83	-5.59	7.78	-6.61
		(0.33)	(6.78)	(0.83)	(6.53)
	1995-1996	6.55	-4.75	7.08	0.63
		(0.36)	(24.17)	(0.20)	(5.33)
	1996-1997	7.37	1.34	6.96	-1.79
		(0.31)	(17.21)	(1.05)	(3.06)
	1997-1998	6.00	-2.77	4.51	-3.47
		(0.36)	(6.91)	(0.84)	(5.96)
	1998-1999	1.51	-2.11	2.14	-8.96
		(0.32)	(3.85)	(0.84)	(2.69)
3. Demographic Characteristics					
Average Age in 1991		33.50		33.92	
		(0.21)		(0.10)	
Proportion Non-white		0.07		0.08	
		(0.002)		(0.003)	
Proportion Hispanic		0.09		0.09	
		(0.002)		(0.002)	
Proportion Married		0.64		0.66	
		(0.004)		(0.003)	
Proportion No High School		0.10		0.11	
		(0.003)		(0.005)	
Proportion High School Diploma		0.33		0.33	
		(0.004)		(0.006)	
Proportion Some College		0.30		0.27	
		(0.006)		(0.012)	
Proportion College Degree		0.15		0.16	
		(0.005)		(0.008)	
Proportion Graduate Degree		0.12		0.13	
		(0.002)		(0.005)	
Avg. # Years w/ FICA Earnings		16.43		16.71	
		(0.19)		(0.10)	

Table 2, Panel 1:
Covariance/Variance Matrix of Earnings and Hours Residuals, Completed Implicates

Earnings Variances/Covariances Along/Below Diagonal with Standard Errors in Parentheses
Correlation Coefficients Above Diagonal

		1990-1991	1991-1992	1992-1993	1993-1994	1994-1995	1995-1996	1996-1997	1997-1998	1998-1999
Earnings	**1990-1991**	0.3530	-0.2982	-0.1051	-0.0416	-0.0160	-0.0078	-0.0215	-0.0002	-0.0029
		(2.084)								
	1991-1992	-0.1026	0.3352	-0.3101	-0.0955	-0.0330	-0.0165	-0.0088	-0.0137	-0.0051
		(1.261)	(2.168)							
	1992-1993	-0.0361	-0.1039	0.3348	-0.3148	-0.0933	-0.0249	-0.0061	-0.0048	-0.0081
		(0.816)	(1.279)	(2.088)						
	1993-1994	-0.0132	-0.0295	-0.0970	0.2838	-0.2221	-0.0989	-0.0301	-0.0050	-0.0100
		(0.744)	(0.791)	(1.226)	(2.067)					
	1994-1995	-0.0046	-0.0093	-0.0263	-0.0576	0.2370	-0.2778	-0.0945	-0.0173	-0.0081
		(0.699)	(0.723)	(0.756)	(1.188)	(2.011)				
	1995-1996	-0.0023	-0.0047	-0.0070	-0.0257	-0.0660	0.2382	-0.2464	-0.0951	-0.0027
		(0.691)	(0.701)	(0.724)	(0.730)	(1.132)	(1.974)			
	1996-1997	-0.0061	-0.0024	-0.0017	-0.0076	-0.0219	-0.0572	0.2264	-0.2386	-0.0780
		(0.694)	(0.689)	(0.694)	(0.711)	(0.717)	(1.122)	(1.954)		
	1997-1998	0.0000	-0.0038	-0.0013	-0.0013	-0.0040	-0.0221	-0.0540	0.2262	-0.2092
		(0.647)	(0.677)	(0.665)	(0.662)	(0.691)	(0.681)	(1.112)	(1.841)	
	1998-1999	-0.0009	-0.0016	-0.0025	-0.0029	-0.0021	-0.0007	-0.0200	-0.0536	0.2904
		(0.653)	(0.693)	(0.678)	(0.666)	(0.671)	(0.663)	(0.700)	(1.057)	(1.938)

Table 2, Panel 2:
Covariance/Variance Matrix of Earnings and Hours Residuals, Completed Implicates

Earnings Covariances with Standard Deviations in Parentheses

		1990-1991	1991-1992	1992-1993	1993-1994	1994-1995	1995-1996	1996-1997	1997-1998	1998-1999
Hours	**1990-1991**	0.0270	-0.0061	-0.0076	-0.0070	-0.0013	-0.0013	-0.0017	-0.0016	-0.0012
		(0.839)	(0.774)	(0.740)	(0.723)	(0.700)	(0.671)	(0.662)	(0.655)	(0.631)
	1991-1992	-0.0066	0.0271	-0.0035	-0.0021	0.0013	-0.0009	0.0012	0.0006	0.0030
		(0.881)	(0.955)	(0.879)	(0.791)	(0.773)	(0.753)	(0.762)	(0.733)	(0.708)
	1992-1993	-0.0021	-0.0026	0.0285	-0.0035	-0.0046	-0.0007	-0.0011	0.0006	0.0005
		(0.765)	(0.810)	(0.880)	(0.789)	(0.730)	(0.707)	(0.724)	(0.700)	(0.701)
	1993-1994	-0.0014	-0.0083	-0.0054	0.0219	0.0003	-0.0022	0.0003	0.0013	-0.0003
		(0.580)	(0.577)	(0.618)	(0.707)	(0.588)	(0.563)	(0.563)	(0.535)	(0.534)
	1994-1995	-0.0016	0.0035	-0.0051	-0.0070	0.0051	0.0028	0.0009	0.0003	-0.0004
		(0.591)	(0.581)	(0.596)	(0.627)	(0.625)	(0.577)	(0.563)	(0.553)	(0.558)
	1995-1996	0.0015	-0.0044	-0.0002	-0.0001	0.0053	0.0068	-0.0010	-0.0010	0.0012
		(0.821)	(0.823)	(0.827)	(0.811)	(0.778)	(0.799)	(0.779)	(0.757)	(0.773)
	1996-1997	-0.0017	-0.0021	0.0007	-0.0001	0.0002	-0.0008	0.0096	-0.0018	-0.0006
		(0.711)	(0.706)	(0.695)	(0.668)	(0.644)	(0.665)	(0.712)	(0.649)	(0.650)
	1997-1998	0.0005	0.0015	0.0001	0.0011	-0.0017	-0.0013	-0.0016	0.0078	-0.0024
		(0.665)	(0.660)	(0.665)	(0.633)	(0.604)	(0.605)	(0.608)	(0.662)	(0.605)
	1998-1999	-0.0020	0.0002	-0.0005	0.0010	-0.0010	-0.0008	-0.0007	-0.0021	0.0097
		(0.635)	(0.634)	(0.636)	(0.617)	(0.613)	(0.652)	(0.635)	(0.600)	(0.667)

Earnings Correlation Coefficients

		1990-1991	1991-1992	1992-1993	1993-1994	1994-1995	1995-1996	1996-1997	1997-1998	1998-1999
Hours	**1990-1991**	0.0702	-0.0153	-0.0053	-0.0047	-0.0056	0.0031	-0.0037	0.0013	-0.0053
	1991-1992	-0.0163	0.0648	-0.0067	-0.0294	0.0128	-0.0091	-0.0047	0.0043	0.0005
	1992-1993	-0.0202	-0.0084	0.0726	-0.0194	-0.0186	-0.0004	0.0016	0.0002	-0.0015
	1993-1994	-0.0202	-0.0055	-0.0097	0.0845	-0.0276	-0.0002	-0.0002	0.0034	0.0030
	1994-1995	-0.0041	0.0038	-0.0140	0.0011	0.0220	0.0133	0.0006	-0.0056	-0.0035
	1995-1996	-0.0042	-0.0025	-0.0023	-0.0095	0.0121	0.0169	-0.0020	-0.0042	-0.0026
	1996-1997	-0.0054	0.0036	-0.0033	0.0014	0.0040	-0.0025	0.0262	-0.0055	-0.0025
	1997-1998	-0.0052	0.0019	0.0019	0.0057	0.0011	-0.0026	-0.0050	0.0265	-0.0069
	1998-1999	-0.0035	0.0078	0.0013	-0.0013	-0.0015	0.0027	-0.0014	-0.0072	0.0289

Table 2, Panel 3:
Covariance/Variance Matrix of Earnings and Hours Residuals, Completed Implicates

Hours Variances/Covariances Along/Below Diagonal with Standard Errors in Parentheses
Correlation Coefficients Above Diagonal

		1990-1991	1991-1992	1992-1993	1993-1994	1994-1995	1995-1996	1996-1997	1997-1998	1998-1999
Hours	**1990-1991**	0.4185	-0.5159	-0.0491	-0.0082	-0.0020	-0.0175	0.0151	0.0096	-0.0213
		(1.968)								
	1991-1992	-0.2408	0.5207	-0.4042	-0.0249	-0.0120	-0.0454	0.0051	-0.0095	0.0113
		(1.651)	(2.356)							
	1992-1993	-0.0216	-0.1978	0.4601	-0.4232	-0.0976	-0.0728	-0.0297	-0.0109	0.0130
		(0.765)	(1.376)	(2.076)						
	1993-1994	-0.0026	-0.0087	-0.1394	0.2358	-0.2988	-0.0486	0.0197	0.0095	-0.0069
		(0.609)	(0.704)	(1.201)	(1.358)					
	1994-1995	-0.0006	-0.0041	-0.0313	-0.0687	0.2242	-0.2953	-0.0003	-0.0027	-0.0084
		(0.592)	(0.660)	(0.776)	(0.886)	(1.286)				
	1995-1996	-0.0093	-0.0270	-0.0407	-0.0194	-0.1151	0.6773	-0.5180	0.0093	-0.0733
		(0.832)	(0.919)	(0.965)	(0.803)	(1.093)	(2.546)			
	1996-1997	0.0075	0.0028	-0.0154	0.0073	-0.0001	-0.3268	0.5875	-0.3623	-0.0587
		(0.689)	(0.756)	(0.734)	(0.572)	(0.568)	(1.837)	(2.146)		
	1997-1998	0.0039	-0.0043	-0.0046	0.0029	-0.0008	0.0048	-0.1729	0.3878	-0.3476
		(0.698)	(0.741)	(0.644)	(0.516)	(0.537)	(0.927)	(1.223)	(1.839)	
	1998-1999	-0.0086	0.0051	0.0055	-0.0021	-0.0025	-0.0376	-0.0280	-0.1350	0.3886
		(0.703)	(0.724)	(0.645)	(0.515)	(0.533)	(0.884)	(0.872)	(1.299)	(2.207)

Table 3, Panel 1:
Covariance/Variance Matrix of Earnings and Hours Residuals, Synthetic Implicates

Earnings Variances/Covariances Along/Below Diagonal with Standard Errors in Parentheses
Correlation Coefficients Above Diagonal

		1990-1991	1991-1992	1992-1993	1993-1994	1994-1995	1995-1996	1996-1997	1997-1998	1998-1999
Earnings	**1990-1991**	0.3884	-0.2809	-0.0913	-0.0318	-0.0238	-0.0123	-0.0111	-0.0034	-0.0051
		(0.826)								
	1991-1992	-0.1023	0.3414	-0.2905	-0.0849	-0.0339	-0.0213	-0.0108	-0.0107	0.0001
		(0.470)	(0.721)							
	1992-1993	-0.0337	-0.1006	0.3514	-0.3128	-0.0923	-0.0358	-0.0204	-0.0061	-0.0049
		(0.332)	(0.456)	(0.749)						
	1993-1994	-0.0115	-0.0288	-0.1077	0.3375	-0.3153	-0.0881	-0.0345	-0.0232	-0.0099
		(0.298)	(0.304)	(0.469)	(0.731)					
	1994-1995	-0.0078	-0.0104	-0.0286	-0.0959	0.2740	-0.2277	-0.0757	-0.0214	-0.0143
		(0.262)	(0.256)	(0.283)	(0.420)	(0.585)				
	1995-1996	-0.0038	-0.0062	-0.0105	-0.0254	-0.0593	0.2474	-0.2710	-0.0754	-0.0278
		(0.238)	(0.231)	(0.242)	(0.262)	(0.327)	(0.534)			
	1996-1997	-0.0035	-0.0032	-0.0061	-0.0102	-0.0201	-0.0683	0.2569	-0.2702	-0.0666
		(0.238)	(0.229)	(0.237)	(0.241)	(0.238)	(0.333)	(0.551)		
	1997-1998	-0.0011	-0.0033	-0.0019	-0.0071	-0.0059	-0.0197	-0.0720	0.2766	-0.2548
		(0.243)	(0.233)	(0.238)	(0.236)	(0.222)	(0.234)	(0.345)	(0.597)	
	1998-1999	-0.0018	0.0000	-0.0016	-0.0033	-0.0043	-0.0079	-0.0192	-0.0763	0.3241
		(0.254)	(0.244)	(0.246)	(0.244)	(0.227)	(0.222)	(0.248)	(0.399)	(0.715)

Table 3, Panel 2:
Covariance/Variance Matrix of Earnings and Hours Residuals, Synthetic Implicates

		Earnings	Covariances with Standard Deviations in Parentheses							
		1990-1991	**1991-1992**	**1992-1993**	**1993-1994**	**1994-1995**	**1995-1996**	**1996-1997**	**1997-1998**	**1998-1999**
Hours	**1990-1991**	0.0036	0.0016	-0.0009	-0.0004	-0.0006	-0.0005	0.0006	0.0014	-0.0018
		(0.319)	(0.297)	(0.293)	(0.281)	(0.254)	(0.236)	(0.236)	(0.244)	(0.258)
	1991-1992	0.0080	0.0039	0.0008	0.0023	-0.0008	0.0006	-0.0003	0.0002	0.0015
		(0.322)	(0.301)	(0.302)	(0.290)	(0.259)	(0.242)	(0.248)	(0.253)	(0.267)
	1992-1993	-0.0009	0.0050	0.0032	-0.0008	0.0000	0.0012	-0.0015	-0.0005	-0.0001
		(0.309)	(0.295)	(0.295)	(0.284)	(0.255)	(0.239)	(0.245)	(0.251)	(0.265)
	1993-1994	-0.0018	-0.0030	0.0035	0.0023	0.0008	-0.0006	0.0017	0.0008	0.0002
		(0.235)	(0.227)	(0.228)	(0.219)	(0.198)	(0.184)	(0.187)	(0.193)	(0.203)
	1994-1995	-0.0004	0.0011	-0.0002	-0.0006	0.0010	0.0010	0.0005	0.0008	-0.0009
		(0.226)	(0.214)	(0.216)	(0.214)	(0.193)	(0.178)	(0.180)	(0.185)	(0.196)
	1995-1996	-0.0014	-0.0020	-0.0004	0.0026	0.0058	0.0027	0.0025	0.0002	0.0039
		(0.375)	(0.355)	(0.361)	(0.350)	(0.316)	(0.298)	(0.305)	(0.315)	(0.339)
	1996-1997	0.0003	0.0017	-0.0014	-0.0001	-0.0012	0.0028	0.0015	0.0002	-0.0001
		(0.302)	(0.289)	(0.295)	(0.285)	(0.258)	(0.247)	(0.250)	(0.261)	(0.280)
	1997-1998	0.0005	-0.0004	0.0007	0.0004	-0.0011	0.0002	0.0021	0.0030	0.0009
		(0.271)	(0.261)	(0.266)	(0.261)	(0.235)	(0.225)	(0.231)	(0.240)	(0.260)
	1998-1999	0.0004	0.0003	0.0002	0.0015	-0.0003	-0.0001	0.0003	-0.0009	0.0041
		(0.313)	(0.298)	(0.304)	(0.297)	(0.268)	(0.253)	(0.259)	(0.272)	(0.292)

		Earnings	Correlation Coefficients							
		1990-1991	**1991-1992**	**1992-1993**	**1993-1994**	**1994-1995**	**1995-1996**	**1996-1997**	**1997-1998**	**1998-1999**
Hours	**1990-1991**	0.0087	0.0191	-0.0023	-0.0056	-0.0012	-0.0027	0.0007	0.0011	0.0010
	1991-1992	0.0042	0.0098	0.0127	-0.0097	0.0039	-0.0041	0.0040	-0.0010	0.0008
	1992-1993	-0.0024	0.0019	0.0082	0.0114	-0.0007	-0.0009	-0.0033	0.0018	0.0006
	1993-1994	-0.0011	0.0059	-0.0020	0.0077	-0.0020	0.0054	-0.0001	0.0009	0.0035
	1994-1995	-0.0018	-0.0023	-0.0001	0.0030	0.0037	0.0133	-0.0033	-0.0031	-0.0007
	1995-1996	-0.0016	0.0017	0.0035	-0.0021	0.0039	0.0064	0.0079	0.0007	-0.0002
	1996-1997	0.0019	-0.0009	-0.0045	0.0064	0.0019	0.0058	0.0042	0.0064	0.0008
	1997-1998	0.0040	0.0006	-0.0013	0.0028	0.0030	0.0004	0.0004	0.0086	-0.0023
	1998-1999	-0.0046	0.0039	-0.0001	0.0006	-0.0033	0.0083	-0.0002	0.0023	0.0099

Table 3, Panel 3:
Covariance/Variance Matrix of Earnings and Hours Residuals, Synthetic Implicates

Hours Variances/Covariances Along/Below Diagonal with Standard Errors in Parentheses
Correlation Coefficients Above Diagonal

		1990-1991	1991-1992	1992-1993	1993-1994	1994-1995	1995-1996	1996-1997	1997-1998	1998-1999
Hours	**1990-1991**	0.4392	-0.2981	-0.0800	-0.0215	0.0067	-0.0176	0.0032	-0.0138	-0.0181
		(0.980)								
	1991-1992	-0.1336	0.4575	-0.3499	-0.0624	-0.0408	-0.0742	0.0084	0.0154	-0.0233
		(0.576)	(0.982)							
	1992-1993	-0.0355	-0.1587	0.4496	-0.3458	-0.0979	-0.1239	0.0072	-0.0096	-0.0102
		(0.361)	(0.630)	(1.027)						
	1993-1994	-0.0075	-0.0222	-0.1219	0.2763	-0.3273	-0.0684	-0.0110	-0.0016	0.0096
		(0.246)	(0.303)	(0.547)	(0.663)					
	1994-1995	0.0022	-0.0137	-0.0327	-0.0857	0.2479	-0.2506	0.0050	0.0045	-0.0276
		(0.236)	(0.267)	(0.351)	(0.380)	(0.535)				
	1995-1996	-0.0097	-0.0419	-0.0694	-0.0300	-0.1042	0.6974	-0.3681	-0.0591	-0.0258
		(0.384)	(0.415)	(0.479)	(0.405)	(0.466)	(1.362)			
	1996-1997	0.0015	0.0041	0.0035	-0.0041	0.0018	-0.2189	0.5073	-0.3689	-0.0769
		(0.315)	(0.317)	(0.315)	(0.241)	(0.235)	(0.812)	(1.146)		
	1997-1998	-0.0061	0.0069	-0.0043	-0.0006	0.0015	-0.0327	-0.1742	0.4399	-0.2651
		(0.284)	(0.289)	(0.294)	(0.224)	(0.215)	(0.523)	(0.738)	(1.021)	
	1998-1999	-0.0088	-0.0116	-0.0050	0.0037	-0.0101	-0.0158	-0.0402	-0.1292	0.5397
		(0.331)	(0.329)	(0.323)	(0.247)	(0.243)	(0.533)	(0.539)	(0.648)	(1.185)

Table 4: Model Parameter Estimates -- Completed Implicates

	Parameter Name	Average Parameter (qbar)	Standard Error [sqrt(T)]	Confidence Interval: 95% Lower Bound	Upper Bound
Parameter of Main Interest *:	**mu**	1.3131	0.1630	0.9006	1.7255
1st Differenced Measurement	**VAR(omega1)**	0.1068	0.0172	0.0521	0.1614
Error and Transitory Preference	**COV(omega1,omega2)**	0.0112	0.0645	-0.1942	0.2166
Shocks	**VAR(omega2)**	0.2293	0.0932	-0.0673	0.5258
	VAR(innovation)	0.0270	0.0260	-0.0558	0.1097
Wage Rate Process MA2	**Coef. on 1st lag of MA2**	0.7261	0.3701	-0.0320	1.4841
	Coef. on 2nd lag of MA2	-0.4120	0.2189	-0.9352	0.1112

*equals (1+eta)/eta where eta is the intertemporal substition elasticity

Table 5: Model Parameter Estimates -- Synthetic Implicates

	Parameter Name	Average Parameter (qbar)	Standard Error [sqrt(T)]	Confidence Interval: 95% Lower Bound	Upper Bound
Parameter of Main Interest *:	**mu**	1.0932	0.0598	0.9028	1.2835
1st Differenced Measurement	**VAR(omega1)**	0.1007	0.0277	0.0125	0.1888
Error and Transitory Preference	**COV(omega1,omega2)**	-0.0653	0.0204	-0.1303	-0.0004
Shocks	**VAR(omega2)**	0.1951	0.0511	0.0326	0.3577
	VAR(innovation)	0.0767	0.0332	-0.0290	0.1824
Wage Rate Process MA2	**Coef. on 1st lag of MA2**	0.5338	0.1328	0.1111	0.9564
	Coef. on 2nd lag of MA2	-0.1326	0.1097	-0.4818	0.2167

*equals (1+eta)/eta where eta is the intertemporal substition elasticity

Table 6: Average Distance of PROBT from Mean

	# rows below diagonal	Completed data compared to: PSID	PSID (no SEO)	NLS	SIME/DIME
Covariance of Earnings					
	0	0.023	0.018	0.027	0.140
	1	0.009	0.007	0.008	0.062
	2	0.005	0.005	0.012	0.023
	3	0.010	0.005	0.014	0.021
Covariance of Earnings and Hours					
	-3	0.008	0.008	0.012	0.007
	-2	0.017	0.014	0.018	0.013
	-1	0.063	0.057	0.102	0.066
	0	0.043	0.041	0.024	0.151
	1	0.015	0.011	0.007	0.083
	2	0.012	0.011	0.021	0.028
	3	0.008	0.010	0.026	0.025
Covariance of Hours					
	0	0.020	0.022	0.027	0.150
	1	0.006	0.009	0.020	0.058
	2	0.008	0.006	0.012	0.018
	3	0.012	0.011	0.011	0.006

* Average Distance in Probability Measure (based on Student-T with appropriate degrees of freedom) from Mean of T-statistic for the Null Hypotheses that an element of the Hours/Earn Covariance matrix based on the Completed implicates is the same as the corresponding element of the matrices estimated in Abowd & Card (1989)

Table 7: Comparison of Earn/Hours COV Matrices Estimated from the Completed and Synthetic Data

	# rows below diagonal	Percentage Overlap of Completed and Synthetic Confidence Intervals			
		Mean	Standard Error	Minimum	Maximum
Covariance of Earnings	0	33.10	4.29	27.06	39.65
	1	34.23	3.78	28.90	38.23
	2	36.49	2.56	33.18	40.63
	3	34.76	2.79	32.19	40.02
	4	34.80	1.81	32.90	37.48
	5	35.04	1.51	33.24	36.69
	6	35.01	1.15	34.25	36.33
	7	36.36	1.74	35.13	37.59
	8	38.84	.	38.84	38.84
Covariance of Earnings and Hours	-8	40.85	.	40.85	40.85
	-7	37.51	0.38	37.24	37.78
	-6	35.98	1.68	34.47	37.79
	-5	35.40	2.27	32.52	38.03
	-4	34.69	1.73	32.11	36.25
	-3	36.11	4.39	33.12	43.87
	-2	37.22	4.32	32.03	43.05
	-1	36.93	3.93	30.83	42.90
	0	35.26	4.19	30.82	43.79
	1	38.14	3.41	34.08	45.30
	2	39.58	2.29	36.23	43.11
	3	40.22	2.58	36.76	43.65
	4	41.76	2.17	38.28	43.77
	5	43.71	3.87	40.04	48.21
	6	43.25	4.18	39.50	47.76
	7	43.85	4.52	40.66	47.04
	8	49.31	.	49.31	49.31
Covariance of Hours	0	49.72	5.11	41.58	55.53
	1	45.75	7.25	34.86	60.32
	2	49.38	7.40	41.44	61.78
	3	45.50	8.07	40.14	60.23
	4	43.41	2.23	39.94	45.56
	5	45.39	2.54	41.90	47.97
	6	44.89	5.60	38.93	50.05
	7	43.08	3.37	40.69	45.46
	8	47.13	.	47.13	47.13