

## Appendix B from Scholz et al., “Are Americans Saving “Optimally” for Retirement?”

(JPE, vol. 114, no. 4, p. 607)

### Imputing Earnings in the HRS

We have two problems with the earnings data that we address. For 77 percent of the 1992 HRS sample, we have access to each individual’s social security earnings records from 1951 to 1991. The social security earnings records report wage, salary, and self-employment income up to the earnings maximum (the earnings thresholds at which social security taxes are no longer taken from income). For 93 percent of the respondents with social security earnings records, we also have W-2 earnings records from 1980 to 1991. These W-2 records provide complete earnings information for wage and salary earners and the self-employed. The first difficulty is that 24 percent of positive social security earnings records are top-coded, and 40 percent of respondents with social security earnings records have at least one top-coded observation.

Our second problem is that 23 percent of respondents refused to grant access to social security earnings records. For these households we have self-reported earnings information for their current job (or the most recent job if not employed) and as many as three previous jobs. We need to estimate their earning profiles on the basis of their self-reported earnings information.

The goal is to use all available information to impute top-coded and missing earning observations and as a result obtain complete individual earnings histories. For the imputation, we proceed in two steps. First, on the basis of the social security and W-2 records, we estimate a dynamic-panel Tobit model to obtain individual earning processes. Then, conditional on all available earnings information, we use the estimates to impute the top-coded and missing observations.

#### A. Estimation

We start by describing our approach to estimating earnings for individuals with top-coded earnings.

For simplicity, suppose that we have earnings records of  $N$  individuals from time  $t = 0$  to  $T$ , where 0 is the first period in which these individuals started working full-time. Assume for the moment that earnings are positive in each time period.<sup>41</sup> Denote the logarithmic value of individual  $i$ ’s latent and observed earnings as  $y_{i,t}^*$  and  $y_{i,t}$ , respectively. The relationship between the latent and observed earnings is

$$y_{i,t} = \begin{cases} y_{i,t}^* & \text{if } y_{i,t}^* < y_t^{tc} \\ y_t^{tc} & \text{if } y_{i,t}^* \geq y_t^{tc}, \end{cases}$$

where  $y_t^{tc}$  is the logarithmic value of the social security maximum taxable earnings at time  $t$ .

The individual log earnings process is specified as

$$\begin{aligned} y_{i,0}^* &= \mathbf{x}_{i,0}'\beta_0 + \epsilon_{i,0}, \\ y_{i,t}^* &= \rho y_{i,t-1}^* + \mathbf{x}_{i,t}'\beta + \epsilon_{i,t}, \quad t \in \{1, 2, \dots, T\}, \\ \epsilon_{i,t} &= \alpha_i + u_{i,t}, \end{aligned} \tag{B1}$$

where  $\mathbf{x}_{i,t}$  is the vector of  $i$ ’s characteristics at time  $t$ , and the error term  $\epsilon_{i,t}$  includes an individual-specific component  $\alpha_i$ , which is constant over time and known to the individual before time 0, and the unanticipated

<sup>41</sup> Generalizing this to the case in which the earnings series begins after time 0 and the case in which some earnings observations are zero is straightforward but detail-oriented, so we omit the discussion. We did treat these cases in practice, however.

white-noise component,  $u_{i,t}$ . Notice that parameters  $\beta_0$  and  $\beta$  are allowed to be different. In the following analysis, we employ random-effect assumptions with homoskedastic errors.

ASSUMPTION 1.  $\alpha_i | \underline{x}_i \sim \text{iid } N(0, \sigma_\alpha^2)$ .

ASSUMPTION 2.  $u_{i,t} | \underline{x}_i \sim \text{iid } N(0, \sigma_u^2)$  for all  $t$ .

ASSUMPTION 3.

$$E[u_{i,t} | \underline{x}_i, \alpha_i] = 0, E[u_{i,t}^2 | \underline{x}_i, \alpha_i] = \sigma_u^2, E[u_{i,t} u_{i,k} | \underline{x}_i, \alpha_i] = 0 \quad \forall t \in \{0, 1, \dots, T\}, t \neq k,$$

where  $\underline{x}_i \equiv (x_{i,0}, x_{i,1}, \dots, x_{i,T})$ .

These three assumptions imply that

$$\epsilon_i = (\epsilon_{i,0}, \epsilon_{i,1}, \dots, \epsilon_{i,T})' \sim N(0, \Sigma), \quad (\text{B2})$$

where

$$\Sigma = \begin{bmatrix} \sigma_{0,0}^2 & \sigma_{0,1} & \cdots & \sigma_{0,T} \\ \sigma_{1,0} & \sigma_{1,1}^2 & \cdots & \sigma_{1,T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T,0} & \sigma_{T,1} & \cdots & \sigma_{T,T}^2 \end{bmatrix},$$

with  $\sigma_{j,k}^2 = \sigma_\alpha^2 + \sigma_u^2$  for  $j = k$ , and  $\sigma_{j,k}^2 = \sigma_\alpha^2$  otherwise. Our goal here is to obtain consistent estimates of the true parameters  $\theta^* = (\beta, \beta_0, \rho, \sigma_\alpha^2, \sigma_u^2)$ . We do this by maximum likelihood.

To construct the likelihood function for each individual's earnings series, notice that we can write the joint probability density function of each pair of random variables  $(y_{i,t}, y_{i,t}^*)$  as

$$g(y_{i,t}, y_{i,t}^* | y_{i,t-1}, y_{i,t-1}^*, y_{i,t-2}, y_{i,t-2}^*, \dots, y_{i,0}, y_{i,0}^*; \underline{x}_i, \theta).$$

From the AR(1) assumption on earnings made in (B1), it follows that

$$g(y_{i,t}, y_{i,t}^* | y_{i,t-1}, y_{i,t-1}^*, y_{i,t-2}, y_{i,t-2}^*, \dots, y_{i,0}, y_{i,0}^*; \underline{x}_i, \theta) = g(y_{i,t}, y_{i,t}^* | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta).$$

In other words, of all the information about past realized and observed earnings, only information from the previous period matters. As a special case, the conditional likelihood of the pair  $(y_{i,t}, y_{i,t}^*)$  is  $g_0(y_{i,0}, y_{i,0}^* | \underline{x}_i, \theta)$  because there is no information about earnings before period 0.

Applying Bayes' rules to the density  $g(\cdot)$ , we have

$$g(y_{i,t}, y_{i,t}^* | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta) = h(y_{i,t}^* | y_{i,t}, y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta) q(y_{i,t} | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta), \quad (\text{B3})$$

where the density for the log of observed earnings conditional on the past information is a conventional Tobit likelihood function

$$q(y_{i,t} | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta) = [f(y_{i,t} | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta)]^{1(y_{i,t}^* < y_{i,t}^c)} [\Pr(y_{i,t}^* \geq y_{i,t}^c | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta)]^{1(y_{i,t}^* \geq y_{i,t}^c)},$$

where  $f(\cdot)$  and  $\Pr(\cdot)$  are a probability density and a cumulative distribution function, respectively, and the conditional density  $h(\cdot)$  for noncensored observations is the probability mass function

$$h(y_{i,t}^* | y_{i,t} < y_{i,t}^c, y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta) = \begin{cases} 1 & \text{if } y_{i,t}^* = y_{i,t} \\ 0 & \text{if } y_{i,t}^* \neq y_{i,t} \end{cases}$$

and the conditional density is simply  $h(y_{i,t}^* | y_{i,t} = y_{i,t}^c, y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta)$  for censored observations.

Similarly, we can write

$$g_0(y_{i,0}, y_{i,0}^* | \underline{x}_i, \theta) = h_0(y_{i,0}^* | y_{i,0}; \underline{x}_i, \theta) q_0(y_{i,0} | \underline{x}_i, \theta),$$

where the conditional density  $h_0(\cdot)$  for noncensored observations is the probability mass function

$$h_0(y_{i,0}^* | \underline{x}_i, \theta) = \begin{cases} 1 & \text{if } y_{i,0}^* = y_{i,0} \\ 0 & \text{if } y_{i,0}^* \neq y_{i,0} \end{cases}$$

and the conditional density is  $h_0(y_{i,0}^* | y_{i,0} = y_0^{tc}; \underline{x}_i, \theta)$  for censored observations. In addition, the density for the time 0 log of observed earnings conditional on known information is

$$q_0(y_{i,0} | \underline{x}_i, \theta) = [f(y_{i,0} | \underline{x}_i, \theta)]^{1(y_{i,0} < y_0^{tc})} [\Pr(y_{i,0}^* \geq y_0^{tc} | \underline{x}_i, \theta)]^{1(y_{i,0} \geq y_0^{tc})}.$$

From (B1) and (B2), it is apparent that the functions  $f(y_{i,0} | \underline{x}_i, \theta)$  and  $f(y_{i,t} | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta)$ ,  $t > 0$ , are normal probability density functions and  $\Pr(y_{i,0}^* \geq y_0^{tc} | \underline{x}_i, \theta)$  and  $\Pr(y_{i,t}^* \geq y_t^{tc} | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta)$ ,  $t > 0$ , are normal cumulative distribution functions. For expositional convenience, define

$$\begin{aligned} h^{tc}(0; \underline{x}_i, \theta) &\equiv h_0(y_{i,0}^* | y_{i,0} = y_0^{tc}; \underline{x}_i, \theta), \\ h^{tc}(t; \underline{x}_i, \theta) &\equiv h(y_{i,t}^* | y_{i,t} = y_t^{tc}, y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta), \quad t > 0, \\ q(0; \underline{x}_i, \theta) &\equiv q_0(y_{i,0}; \underline{x}_i, \theta), \\ q(t; \underline{x}_i, \theta) &\equiv q(y_{i,t} | y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta), \quad t > 0. \end{aligned}$$

The likelihood function for  $i$ 's series of observed log earnings is

$$\begin{aligned} L_i(y_{i,T}, y_{i,T-1}, \dots, y_{i,1}, y_{i,0}; \underline{x}_i, \theta) &= \int_{y_{i,T}^*}^{\infty} \dots \int_{y_{i,0}^*}^{\infty} \left[ \prod_{t=0}^T q(t; \underline{x}_i, \theta) \right] \cdot \left[ h(y_{i,0}^*; \underline{x}_i, \theta) \prod_{t=1}^T h(y_{i,t}^* | y_{i,t}, y_{i,t-1}, y_{i,t-1}^*; \underline{x}_i, \theta) \right] dy_{i,0}^* \dots dy_{i,T}^* \\ &= \int_{y_{i,c_n}^*}^{\infty} \dots \int_{y_{i,c_1}^*}^{\infty} \left[ \prod_{t=0}^T q(t; \underline{x}_i, \theta) \right] \cdot \left[ \prod_{k=c_1}^{c_n} h^{tc}(k; \underline{x}_i, \theta) \right] dy_{i,c_1}^* \dots dy_{i,c_n}^* \\ &= E_{y_{i,c_1}^*, \dots, y_{i,c_n}^*} \left[ \prod_{t=0}^T q(t; \underline{x}_i, \theta) \right], \end{aligned} \tag{B4}$$

where  $c_1, c_2, \dots, c_n$  are the periods in which the log observed earnings are censored, that is, equal to their corresponding top-coded limits. Notice that, since we do not observe  $y_{i,t}^*$  when it is censored, we integrate out  $y_{i,t}^*$  for censored observations. Unfortunately, the integration does not yield any analytical solution, nor is direct numerical evaluation of the integral computationally feasible in this case. As an alternative, Chang (2002) proposes using a Geweke-Hajivassiliou-Keane (GHK) (probit) simulator to deal with the computational burden of the integration.<sup>42</sup> The estimation results are given in table B1.

## B. Imputation

The idea is to impute top-coded and missing earnings observations with their conditional expectations, where the conditioning variables include both the individual's characteristics and observed earnings. The conditional expectations are calculated numerically on the basis of the dynamic earnings model (B1) and the distributional assumption (B2). The imputation scheme is similar for top-coded and missing observations; therefore, we discuss only the scheme for top-coded observations here.<sup>43</sup>

<sup>42</sup> The GHK simulator gives a numerical approximation of a probit probability of interest. It is a popular choice of probit simulators because of its relative accuracy; see Geweke, Keane, and Runkle (1994) for details.

<sup>43</sup> We can think of a missing earnings observation as an observation with a top-coded value of zero, which is equivalent to saying that we know nothing about earnings in that period (as opposed to the case in which we observe top-coded earnings and know that the actual earnings are at least as

To be concrete, notice that (B1) implies that

$$\begin{aligned} E[y_0^* | \underline{y}, \underline{x}, \theta] &= x'_0 \beta_0 + E[\epsilon_0 | \underline{y}, \underline{x}, \theta], \\ E[y_t^* | \underline{y}, \underline{x}, \theta] &= \rho E[y_{t-1}^* | \underline{y}, \underline{x}, \theta] + x'_t \beta + E[\epsilon_t | \underline{y}, \underline{x}, \theta], \quad t \in \{1, 2, \dots, T\}, \end{aligned} \quad (\text{B5})$$

where  $\underline{y} = (y_0, y_1, \dots, y_T)$  is the series of individual  $i$ 's log of observed earnings (the individual subscript  $i$  is omitted throughout this subsection). By construction, given information about the individual's characteristics and observed earnings,  $E[y_t^* | \underline{y}, \underline{x}, \theta]$  is on average the best guess for  $y_t^*$ . In other words, for top-coded observations, equation (B1) suggests the imputed values

$$y_t^{\text{imp}} = E[y_t^* | y_t = y_t^{tc}; \underline{y}, \underline{x}, \theta],$$

which requires knowledge of  $E[\epsilon_t | y_t = y_t^{tc}; \underline{y}, \underline{x}, \theta]$  for every period  $t$  in which  $y_t = y_t^{tc}$ . The analytical form of  $E[\epsilon_t | y_t = y_t^{tc}; \underline{y}, \underline{x}, \theta]$  is not available in our case; therefore, we calculate this object numerically using the Gibbs sampling procedure.<sup>44</sup>

To facilitate the discussion about details of the procedure, denote  $\underline{\epsilon}_{<t} = (\epsilon_0, \epsilon_1, \dots, \epsilon_{t-1})'$ ,  $\underline{\epsilon}_{>t} = (\epsilon_{t+1}, \epsilon_{t+2}, \dots, \epsilon_T)'$ , and  $\underline{\epsilon}_{-t} = (\epsilon_0, \epsilon_1, \dots, \epsilon_{t-1}, \epsilon_{t+1}, \dots, \epsilon_T)'$  for any vector  $\underline{\epsilon} = (\epsilon_0, \epsilon_1, \dots, \epsilon_T)'$ . Here, we want to simulate  $R$  sets of  $\underline{\epsilon}$  that are consistent with the observed  $\underline{y}$  and  $\underline{x}$  given  $\theta$ . The Gibbs sampling procedure does this in two steps for each round of simulation.

1. In the  $r$ th round of simulation,  $r = 1, 2, \dots, R$ , generate a “random” initial value  $\underline{\epsilon}_0^{(r)} = (\epsilon_{0,0}^{(r)}, \epsilon_{1,0}^{(r)}, \dots, \epsilon_{t,0}^{(r)}, \dots, \epsilon_{T,0}^{(r)})$  that satisfies (B1) given  $\underline{y}, \underline{x}$ , and  $\theta$ . Notice that  $\epsilon_{t,0}^{(r)}$  is not identified when  $y_t = y_t^{tc}$ . In this case,  $\epsilon_{t,0}^{(r)}$  is chosen randomly under the restriction that  $y_{t,0}^{*(r)} \equiv (y_t^* | \underline{\epsilon}_{<t} = \underline{\epsilon}_{<t,0}^{(r)}, \epsilon_t = \epsilon_{t,0}^{(r)}; \underline{x}, \theta) \geq y_t^{tc}$ . If  $y_{t-1} < y_{t-1}^{tc}$  and  $y_t < y_t^{tc}$ ,  $\epsilon_{t,0}^{(r)}$  is defined by (B1); that is, it is actually not random.

2. Starting with  $m = 1$ , draw a random number  $\epsilon_{t,m}^{(r)}$  from  $t = 0, \dots, T$  from the distribution of  $\epsilon_t | \underline{\epsilon}_{-t}; \underline{x}, \theta$  such that

$$y_{t,m}^{*(r)} = (y_t^* | \underline{\epsilon}_{>t} = \underline{\epsilon}_{>t,m-1}^{(r)}, \epsilon_{<t} = \epsilon_{<t,m}^{(r)}, \epsilon_t = \epsilon_{t,m}^{(r)}; \underline{x}, \theta) = y_t \quad \text{if } y_t < y_t^{tc}$$

and

$$y_{t,m}^{*(r)} = (y_t^* | \underline{\epsilon}_{>t} = \underline{\epsilon}_{>t,m-1}^{(r)}, \epsilon_{<t} = \epsilon_{<t,m}^{(r)}, \epsilon_t = \epsilon_{t,m}^{(r)}; \underline{x}, \theta) \geq y_t \quad \text{if } y_t = y_t^{tc}.$$

(This is equivalent to drawing  $\epsilon_{t,m}^{(r)}$  from  $\epsilon_t | \underline{\epsilon}_{-t}; \underline{y}, \underline{x}, \theta$ .) Then, continue from  $m = 2$  to  $m = M$ .

With  $\underline{\epsilon}_M^{(r)}$ ,  $r = 1, 2, \dots, R$ , an estimate of  $E[\epsilon_t | \underline{y}, \underline{x}, \theta]$  is

$$\hat{E}[\epsilon_t | \underline{y}, \underline{x}, \theta] = \frac{1}{R} \sum_{r=1}^R \epsilon_{t,M}^{(r)}. \quad (\text{B6})$$

Given the estimate  $\hat{E}[\epsilon_t | \underline{y}, \underline{x}, \theta]$ , we calculate the imputed value of earnings as

$$y_0^{\text{imp}} = x'_0 \beta_0 + \hat{E}[\epsilon_0 | \underline{y}, \underline{x}, \theta]$$

and

$$y_t^{\text{imp}} = \rho y_{t-1}^{\text{imp}} + x'_t \beta + \hat{E}[\epsilon_t | \underline{y}, \underline{x}, \theta], \quad t \in \{1, 2, \dots, T\}. \quad (\text{B7})$$

Notice that, by construction,

large as the top-coded earnings).

<sup>44</sup> Briefly, Gibbs sampling is a procedure to draw a set of numbers randomly from a (valid) joint distribution. Then, the random draws are used to estimate properties of any marginal distribution of interest, which is difficult to derive analytically from the joint distribution. The procedure relies on the law of large numbers, i.e., that moments of a distribution can be estimated consistently from a set of random draws from that distribution.

$$y_0^{\text{imp}} = \frac{1}{R} \sum_{r=1}^R y_{0,M}^{*(r)}$$

and

$$y_t^{\text{imp}} = \frac{1}{R} \sum_{r=1}^R y_{t,M}^{*(r)}, \quad t \in \{1, 2, \dots, T\},$$

and that  $y_t^{\text{imp}} = y_t$  if  $y_t < y_t^{tc}$  and  $y_t^{\text{imp}} \geq y_t$  if  $y_t = y_t^{tc}$ .

The remaining parts of this subsection (i) construct the functional form for the conditional distribution of  $\epsilon_t | \underline{\epsilon}_{-t}; \underline{x}, \theta$  and (ii) show how to draw a random number  $\epsilon_{t,m}^{(r)}$  from this conditional distribution to satisfy (B1) given  $\underline{y}, \underline{x}$ , and  $\theta$ . More notation is required. For any matrix  $\Sigma$ , denote  $\Sigma_{t,t}$  as the element of  $\Sigma$  on the  $t$ th row and  $t$ th column,  $\Sigma_{t,-t}$  as the  $t$ th row of  $\Sigma$  with the element  $\Sigma_{t,t}$  removed,  $\Sigma_{-t,t}$  as the  $t$ th column of  $\Sigma$  with the element  $\Sigma_{t,t}$  removed, and  $\Sigma_{-t,-t}$  as the matrix  $\Sigma$  with the  $t$ th row and  $t$ th column removed.

Recall the property of a joint-normal vector that

$$\begin{aligned} \underline{\epsilon} &\sim N(E[\underline{\epsilon}], \Sigma) \Rightarrow \epsilon_t | \underline{\epsilon}_{-t} \sim N(\mu_{t|-t}, \Sigma_{t|-t}), \\ \mu_{t|-t} &= E[\epsilon_t] + \Sigma_{t,-t} \Sigma_{-t,-t}^{-1} \{\underline{\epsilon}_{-t} - E[\underline{\epsilon}_{-t}]\}, \\ \Sigma_{t|-t} &= \Sigma_{t,t} - \Sigma_{t,-t} \Sigma_{-t,-t}^{-1} \Sigma_{-t,t} \end{aligned} \quad (\text{B8})$$

(see, e.g., Goldberger 1991, 196–97). Recall from (B2) that  $E[\epsilon_j] = 0$  and  $\Sigma = (1 - \rho)\sigma_\epsilon^2 I_{T+1} + \rho\sigma_\epsilon^2 1_{T+1}$ , where  $1_{T+1}$  is a  $(T + 1) \times (T + 1)$  matrix whose elements are all 1. Thus  $\mu_{t|-t} = \Sigma_{t,-t} \Sigma_{-t,-t}^{-1} \underline{\epsilon}_{-t}$ , and  $\Sigma_{t,-t} \Sigma_{-t,-t}^{-1} = \Sigma_{s,-s} \Sigma_{-s,-s}^{-1}$  and  $\Sigma_{t|-t} = \Sigma_{s|-s}$  for any  $t = 0, \dots, T$  and  $s = 0, \dots, T$ .

Recall that we draw a value for  $\epsilon_{t,m}^{(r)}$  randomly from the conditional distribution  $\epsilon_t | \underline{\epsilon}_{-t}$  such that, given  $\underline{\epsilon}_{>t,m-1}^{(r)}$ ,  $\epsilon_{<t,m}^{(r)}$ ,  $\epsilon_{t,m}^{(r)}$ ,  $\underline{x}$ , and  $\theta$ ,

$$y_{t,m}^{*(r)} \begin{cases} = y_t & \text{if } y_t < y_t^{tc} \\ \in [y_t^{tc}, \infty) & \text{if } y_t = y_t^{tc}. \end{cases} \quad (\text{B9})$$

In practice, it is more convenient to work with the standard-normal transformation of  $\epsilon_t | \underline{\epsilon}_{-t}$ ,

$$z_{t|-t} \equiv \frac{(\epsilon_t | \underline{\epsilon}_{-t}) - \mu_{t|-t}}{\sigma_{t|-t}} \sim N(0, 1), \quad \sigma_{t|-t} = \sqrt{\Sigma_{t|-t}}. \quad (\text{B10})$$

From (B1),  $\epsilon_0 | \underline{\epsilon}_{-0} = (y_0^* | \underline{\epsilon}_{-0}) - x_0' \beta_0$  and  $\epsilon_t | \underline{\epsilon}_{-t} = (y_t^* - \rho y_{t-1}^* | \underline{\epsilon}_{-t}) - x_t' \beta$ ,  $t \in \{1, 2, \dots, T\}$ . Also, since  $(y_t^* | \underline{\epsilon}_{-t}; \underline{x}, \theta) = (y_t^* | \underline{\epsilon}_{<t}; \underline{x}, \theta)$ ,  $(y_{t-1}^* | \underline{\epsilon}_{>t} = \underline{\epsilon}_{>t,m-1}^{(r)}, \underline{\epsilon}_{<t} = \underline{\epsilon}_{<t,m}^{(r)}; \underline{x}, \theta) = y_{t-1,m}^{*(r)}$ .

Thus, with the transformation (B10), drawing  $\epsilon_{t,m}^{(r)}$  from (B8) to satisfy (B9) is equivalent to drawing  $z_{t|-t,m}^{(r)}$  such that

$$z_{0|-0,m}^{(r)} = \begin{cases} \frac{y_0 - x_0' \beta_0 - \mu_{0|-0,m}^{(r)}}{\sigma_{0|-0}} & \text{if } y_0 < y_0^{tc} \\ \Phi^{-1} \left( \xi_{0,m}^{(r)} + [1 - \xi_{0,m}^{(r)}] \Phi \left( \frac{y_0^{tc} - x_0' \beta_0 - \mu_{0|-0,m}^{(r)}}{\sigma_{0|-0}} \right) \right) & \text{if } y_0 = y_0^{tc} \end{cases}$$

for  $t = 0$ , and

$$z_{t|t-m}^{(r)} = \begin{cases} \frac{y_t - \rho y_{t-1,m}^{*(r)} - x_t' \beta - \mu_{t|t-m}^{(r)}}{\sigma_{t|t-m}} & \text{if } y_t < y_t^{tc} \\ \Phi^{-1} \left( \xi_{t,m}^{(r)} + [1 - \xi_{t,m}^{(r)}] \Phi \left( \frac{y_t^{tc} - \rho y_{t-1,m}^{*(r)} - x_t' \beta - \mu_{t|t-m}^{(r)}}{\sigma_{t|t-m}} \right) \right) & \text{if } y_t = y_t^{tc} \end{cases}$$

for  $t > 0$ ,<sup>45</sup> with

$$y_{0,m}^{*(r)} = x_0' \beta_0 + [\sigma_{0|-0} z_{0|-0,m}^{(r)} + \mu_{0|-0,m}^{(r)}]$$

and

$$y_{t,m}^{*(r)} = \rho y_{t-1,m}^{*(r)} + x_t' \beta + [\sigma_{t|-t} z_{t|-t,m}^{(r)} + \mu_{t|-t,m}^{(r)}], \quad t \in \{1, 2, \dots, T\},$$

where

$$\mu_{t|-t,m}^{(r)} = \Sigma_{t,-t} \Sigma_{-t,-t}^{-1} \begin{pmatrix} \epsilon_{<,m}^{(r)} \\ \epsilon_{>,m-1}^{(r)} \end{pmatrix},$$

$\epsilon_{t,m}^{(r)} = \sigma_{t|-t} z_{t|-t,m}^{(r)} + \mu_{t|-t,m}^{(r)}$ ,  $\sigma_{t|-t} = \sqrt{\Sigma_{t|-t}}$  and  $\Sigma_{t|-t} = \Sigma_{t,t} - \Sigma_{t,-t} \Sigma_{-t,-t}^{-1} \Sigma_{-t,t}$ , and  $\xi_{t,m}^{(r)}$  is a random draw from a  $[0, 1]$  uniform distribution. Notice that  $y_{t,m}^{*(r)} = y_t$  if  $y_t < y_t^{tc}$  and  $y_{t,m}^{*(r)} \geq y_t$  if  $y_t = y_t^{tc}$  by construction.

### C. Social Security Function

From the expected earnings profiles, we can calculate the *lifetime* summation of household earnings up to the year of retirement as  $E_R \equiv \sum_{j=S}^R e_j$ , where  $e_j$  denotes the household earnings at age  $j$  in a common base year unit, and  $S$  and  $R$  denote the first and the last working ages, respectively.<sup>46</sup> Denote  $\bar{\phi}^h$  and  $\bar{\phi}^w$  as the fractions of  $E_R$  that are contributed by the husband and wife of the household, respectively.<sup>47</sup> On the basis of  $E_R$ ,  $\bar{\phi}^h$ , and  $\bar{\phi}^w$ , we can approximate the household annual social security benefits as follows.

a. Calculate *Individual Primary Insurance Amount (PIA)*<sup>48</sup>

Individual  $i$ 's annual indexed monthly earnings (AIME) can be approximated as

$$\text{AIME}^i \approx \frac{\bar{\phi}^i E_R}{L^i}, \quad (\text{B11})$$

with

<sup>45</sup> To see how this works, note first that for  $\epsilon \sim N(\mu, \sigma^2)$ ,  $f(\epsilon) = (2\pi\sigma^2)^{-1/2} \exp[-0.5(\epsilon - \mu)/\sigma]$ . Define  $z \equiv (\epsilon - \mu)/\sigma$ . It follows that  $F(\epsilon) = \Phi(z(\epsilon))$ , where  $\Phi$  is the standard normal cumulative distribution function. Thus

$$\Phi(z(\epsilon^{(r)})) = F(\epsilon^{(r)}) = \xi_t^{(r)} + [1 - \xi_t^{(r)}] F(\epsilon^{tc}) = \xi_t^{(r)} + [1 - \xi_t^{(r)}] \Phi(z(\epsilon^{tc})).$$

In other words, drawing  $\epsilon^{(r)}$  from a truncated distribution of  $\{\epsilon | \epsilon \geq \epsilon^{tc}\}$  is equivalent to drawing  $z^{(r)} = z(\epsilon^{(r)})$  from a truncated distribution of  $\{z | z \geq z^{tc}\}$  and then transforming  $z^{(r)}$  back to  $\epsilon^{(r)}$ .

<sup>46</sup> As opposed to a *discounted* present value of earnings, the summation is a straightforward summation of earnings in a common base year currency unit, which is the concept employed by the Social Security Administration.

<sup>47</sup> The terminologies “husband” and “wife” are not literal. In particular, we call a single male respondent “husband” and a single female respondent “wife.” Without this simplification, we need separate treatments for married and single households. Under this generalization,  $\bar{\phi}_i^h = 1$  and  $\bar{\phi}_i^w = 0$  for single-male households, and  $\bar{\phi}_i^h = 0$  and  $\bar{\phi}_i^w = 1$  for single-female households.

<sup>48</sup> Social security benefits derived from the calculations in this section are not precise because the calculated AIME may be smaller than the actual AIME and, conditional on AIME being correctly calculated, the calculated household benefits may be larger than the actual ones. For the former, the reasons are that (i) we do not exclude the five years of lowest earnings from the calculation, (ii) we use base year (i.e., real) values of earnings after age 60 instead of nominal values, and (iii) we do not take into account earnings in retirement if respondents work beyond their household retirement dates. For the latter, the reason is that we assume that both the husband and the wife of a married household are eligible to collect benefits at the household retirement date. If one of them is not eligible at the retirement date, the approximation will overstate the benefits. Nevertheless, by virtue of having complete earnings histories for most individuals, our calculations are considerably more accurate than those in other life cycle simulation models of wealth accumulation.

$$L^i = 12 \times \max \{R^i - 22, 40\},$$

where  $i = h$  (husband) or  $w$  (wife), and  $L^i$  is the number of months of  $i$ 's covered period.<sup>49</sup> Without loss of generality, we set  $L^w = 40$  for single-male households and  $L^h = 40$  for single-female households.

*Individual* PIA can be calculated as

$$\begin{aligned} \text{PIA}^i &= 0.90 \times \min \{AIME^i, b_0\} + 0.32 \times \min \{\max \{AIME^i - b_0, 0\}, b_1 - b_0\} \\ &+ 0.15 \times \max \{AIME^i - b_1, 0\}, \end{aligned} \quad (\text{B12})$$

where  $b_0$  and  $b_1$  are the bend points. For the 1992 formula,  $b_0 = \$387$  and  $b_1 = \$2,333$ .

*b. Calculate Household Annual Social Security Benefits*

First, the *individual* monthly social security benefits are calculated as

$$ssb^i = \max \{d_{\text{own}}^i \text{PIA}^i, d_{\text{spouse}}^i \text{PIA}^{i\text{spouse}}, ssx^i\}, \quad (\text{B13})$$

where  $i$ 's spouse =  $h$  ( $w$ ) if  $i = w$  ( $h$ ),  $d_{\text{own}}^i$  is the fraction of  $i$ 's PIA that  $i$  would get if  $i$  collected benefits based on  $i$ 's PIA,  $d_{\text{spouse}}^i$  is the fraction of PIA of  $i$ 's spouse that  $i$  would get if  $i$  collected benefits based on PIA of  $i$ 's spouse, and  $ssx^i$  is the monthly benefit that  $i$  would get if  $i$  collected benefits based on the PIA of  $i$ 's ex-spouse.<sup>50</sup> Without loss of generality, for single-male households,

$$d_{\text{spouse}}^h = d_{\text{own}}^w = d_{\text{spouse}}^w = ssx^w = 0$$

and

$$d_{\text{spouse}}^w = d_{\text{own}}^h = d_{\text{spouse}}^h = ssx^h = 0$$

for single-female households. In addition, we set  $ssx^h = ssx^w = 0$  for married households because we do not have any information to determine  $ssx^i$ . Similarly,  $ssx^i = 0$  for any single households without information to determine their ex-spouses' PIA.

Finally, household  $i$ 's *annual* social security benefits can be approximated as

$$ss_i = 12 \times (ssb_i^h + ssb_i^w), \quad (\text{B14})$$

which, for a married household, is the benefits the household would get when both the husband and wife survive. When one of the spouses in a married household dies, the *annual* social security benefits of the surviving spouse are

$$ss_i^{\text{survive}} = 12 \times \max \{d_{\text{own}}^h \text{PIA}^h, d_{\text{own}}^w \text{PIA}^w\}. \quad (\text{B15})$$

In other words, we approximate the surviving spouse's benefits to be the higher of the husband's and wife's benefits that they would be able to collect on the basis of their own earning histories (which determine their PIAs) and the household retirement date (which determines the factors  $d$ ). This approximates the actual guideline of the Social Security Administration.

<sup>49</sup> Without the lower bound of 40 years in the max operator ( $\max \{R^i - 22, 40\}$ ), AIME would be too high for households whose members retire before age 62. In addition, notice that we use the *household* retirement date ( $R^i$ ) rather than the *individual* retirement date.

<sup>50</sup> To recover the ex-spouse's PIA, we first compute the benefit amount that a single respondent would get based on her own earning history. Then, we compare the amount to the reported amount of social security benefits in the first wave in which the respondent reported collecting the benefits. If the reported benefit amount is higher, we assume that the single respondent collected benefits based on her ex-spouse's records, and the reported amount is used to recover her ex-spouse's PIA.

**TABLE B1**  
**Estimation Results of Individual Earnings Processes**

VARIABLE	SAMPLE			
	Male without College	Male with College	Female without College	Female with College
A. Initial Observations ( $t = 0$ )				
Constant	5.691*** (.204)	3.910*** (.444)	7.407*** (.133)	4.738*** (.434)
Race (white = 1, 0 otherwise)	.282*** (.033)	.330 (.305)	.135*** (.029)	.006 (.087)
Years of schooling/professional postgraduate degree dummy	.044*** (.005)	-.135 (.209)	.017*** (.006)	.058 (.138)
No high school dummy/ postgraduate degree dummy	-.005 (.003)	-.120 (.149)	-.162*** (.041)	.065 (.086)
Marital status in 1992 HRS	.107*** (.030)	.034 (.042)	-.032* (.017)	.033 (.278)
Two-earner household dummy	-.047* (.025)	-.145** (.072)	.164*** (.028)	.187 (.155)
Age	.156*** (.013)	.291*** (.024)	.057*** (.009)	.247*** (.028)
.01 × Age <sup>2</sup>	-.177*** (.020)	-.330*** (.038)	-.056*** (.013)	-.307*** (.042)
B. Subsequent Observations ( $t > 0$ )				
Constant	2.642*** (.042)	2.066*** (.156)	3.296*** (.071)	2.917*** (.141)
Race (white = 1, 0 otherwise)	.106*** (.010)	.061** (.025)	.040*** (.010)	-.009 (.091)
Years of schooling/professional postgraduate degree dummy	.020*** (.002)	.033 (.028)	.025*** (.003)	.066* (.035)
No high school dummy/ postgraduate degree dummy	-.037*** (.010)	.002 (.007)	-.069*** (.018)	.077*** (.019)
Marital status in 1992 HRS	.114*** (.012)	.129** (.056)	-.141*** (.014)	-.166*** (.048)
Two-earner household dummy	-.048*** (.007)	-.078*** (.016)	.210*** (.012)	.197*** (.032)
Age	.039*** (.002)	.032*** (.005)	.018*** (.002)	.019*** (.005)
.01 × Age <sup>2</sup>	-.047*** (.002)	-.037*** (.006)	-.012*** (.003)	-.015*** (.006)
Earnings at the previous period	.629*** (.006)	.737*** (.012)	.568*** (.008)	.659*** (.017)
Variance of the individual-specific effect ( $\sigma_v^2$ )	.029*** (.002)	.012** (.005)	.041*** (.003)	.022*** (.004)
Variance of the gross error term ( $\sigma_\epsilon^2$ )	.213*** (.004)	.223*** (.011)	.240*** (.005)	.213*** (.011)
Number of individual-year observations	86,382	21,286	47,145	8,609
Number of respondents	2,914	720	2,576	446

**Note.**— The dependent variable is the respondents’ natural log earnings. For samples with at most high school, the education variables are (i) years of schooling and (ii) no high school dummy. For samples with at least a bachelor’s degree, the education variables are (i) professional postgraduate degree dummy (master of business administration, doctor of law, doctor of medicine, or doctor of philosophy) and (ii) postgraduate degree dummy. Standard errors are in parentheses.

\* Statistically significant at the 10 percent level.

\*\* Statistically significant at the 5 percent level.

\*\*\* Statistically significant at the 1 percent level.