Bounding the Influence of Attrition on Intertemporal Wage Variation in the NLSY[1]

by

Thomas MaCurdy[2] Christopher Timmins[3]

October 2001

Abstract

This paper analyzes wage dynamics in the National Longitudinal Survey of Youth, controlling for the effects of censoring caused by non-random attrition. Non-random attrition, caused by individuals failing to appear for interviews or choosing not to work, is common in longitudinal surveys like they NLSY and can bias statistical analyses. Techniques to control for the effects of non-random censoring on the dynamics of mean wages require a great deal of knowledge of (or assumptions about) the censoring process. We adapt the non-parametric bounding techniques of Manski for use with a newly proposed Smoothed GMM quantile estimator to overcome this problem by studying the dynamics of wage percentiles. Results suggest that non-random attrition does not pose serious problems for the analysis of men's wages, but that the combination of multiple sources of attrition leads to significant potential biases in the study of women's wages.

---

[1] MaCurdy gratefully acknowledges research support from NIH grant HD32055-02 and from the U.S. Department of Labor Bureau of Labor Statistics. Opinions stated in this document do not necessarily represent the official position or policy of any agency funding this research.

[2] Professor, Department of Economics, and Senior Fellow, The Hoover Institution, Stanford University, Stanford, CA 94305

[3] Assistant Professor, Department of Economics, Yale University, New Haven, CT 06520

1. Introduction

The availability of rich longitudinal data sources has greatly enhanced researchers' abilities to study the dynamic properties of wage and earnings growth experienced by individuals over various time periods and stages of their life cycle. Surveys such as the PSID and NLS offer extensive information on many individuals over decades. Conceptually, with a time series of observations on earnings supplied for each member of a random sample, analysts can formulate elaborate models of stochastic processes to summarize the features of intertemporal earnings mobility. Unfortunately, one rarely has a random sample. As panel surveys progress, individuals are lost for a variety of reasons. This attrition is rarely random, implying that the remaining observations are not representative of the original population. Evidence for the PSID and the NLSY, for example, suggests individuals lost in sampling are drawn disproportionately from the lowest and/or upper segments of the wage distribution. [See Gottschalk and Moffitt (1998) for the PSID and MaCurdy, Mroz and Gritz (1998) for the NLSY] Consequently, over time the distribution of wages becomes corrupted for the observations remaining in the sample. Moreover, when analyzing the evolution of wages for women, one must also account for the additional source of data censoring arising from the non-random loss of observations caused by women selecting not to work. Remarkably little is known about the impacts of either attrition or nonparticipation on our inferences about wage dynamics.

This paper exploits recent advances in nonparametric estimation to gain an understanding of how attrition and non-participation biases the values of parameters relating current to past wages in empirical relationships describing the evolution of individuals' wages. One can learn relatively little about the sensitivity of estimated coefficients to attrition in familiar ARMA-type specifications whose parameters link the conditional means and autocovariances of variables, for these parameters are not identified in the presence of censoring without relying on strong and untestable distributional assumptions. Consequently, one cannot calculate useful bounds for either forecasted means or autocovariances; virtually all values are possible for these quantities without considerable knowledge of the sample selection mechanism at work. Such is not the case, however, for specifications based on order statistics describing dynamic relationships between variables. Manski (1995) illustrates a non-parametric method for determining bounds on quantiles of variables drawn from censored samples. His methods are attractive as they forego the intricate assumptions regarding knowledge

of the sample selection mechanism at work, and instead rely on the worst-case consequences of non-random attrition. This paper estimates quantile regressions analogous to the autoregressive relations found in ARMA models and identifies bounds for these specifications following Manski's approach. The results demonstrate how the autoregressive coefficients associated with these bounds differ from autoregressive coefficients that ignore sample selection.

We conduct this study using data from the National Longitudinal Survey of Youth (NLSY), which has become one of the most widely used data sources for investigating the economic and demographic circumstances of young adults during the 1980's. As the NLSY enters its second decade, concern has arisen regarding the representativeness of the sample due to the possibility of non-random attrition, which has plagued other longitudinal data sets. In contrast to most other longitudinal data sources, sample members departing from the NLSY are often recruited back into the sample at a later date; this process of "returning" may also be non-random, possibly offsetting or exacerbating the effects of attrition. MaCurdy, Mroz, and Gritz (1998) present an exhaustive analysis of attrition in the NLSY, investigating: attrition patterns, the characteristics of who departs from and returns to the sample, and how the depiction of youths' labor market outcomes in the NLSY differs from that in the Current Population Survey (CPS). Their analysis documents that those who attrit tend to have higher wages and earnings before they leave the sample and have lower wages and earnings upon their return. Losing the workers with higher compensation from the NLSY means that its compensation profiles will understate the actual profiles, assuming these workers would have held their positions in the distributions in future years. How this phenomena might impact the coefficient estimates of autoregressive relationships in describing the dynamic properties of wages is the question addressed in this paper.

Our empirical methodology should prove useful in a variety of contexts in addition to the dynamic analysis of wages. An obvious potential application in the field of Industrial Organization is in the analysis of how the distribution of firms' productivities evolves over time. In that setting, firms non-randomly "attrit" by exiting the industry, with important implications for the measurement of productivity dynamics. [Pakes and Olley (1996)] In the field of Development Economics, our approach to quantile estimation (i.e., converting the problem to a smoothed GMM procedure) will facilitate the analysis of cross-country income distribution dynamics, allowing one to test more directly, for example, the conditional convergence hypothesis. The technique will likely prove useful

in other fields as well, whenever the evolution of distributions over time is of interest, and especially when non-random censoring is a concern.

The remainder of this paper contains four sections. Section 2 presents the central ideas underlying our approach to formulating and estimating autoregressive analoges of quantile regressions and the bounds for these relations accounting for censoring. Section 3 describes the sample composition and nature of sample attrition in the NLSY. Section 4 reports the empirical results, and Section 5 summarizes our findings.

2. Estimating Wage Growth Relationships with Censoring

Formulating an econometric approach for estimating bounds for quantiles describing wage growth involves four tasks. The first consists of proposing a specification for quantiles associated with the distribution of current wages conditional on past values. The second recognizes the problems created by censoring in estimating statistical relationships. The third surmises how censoring figures into the construction of estimated bounds for specifications of wage quantiles. The fourth identifies a robust and flexible procedure for estimating the parameters of quantile specifications in a panel data setting. After covering these tasks, this section ends by bringing these items together to develop the estimation approach that we will apply in our empirical analysis.

2.1 Models Characterizing Dynamic Properties of Wages

A popular empirical specification for modeling the growth of wages experienced by individuals in longitudinal data takes the form:

$$\begin{aligned} (2.1) \qquad \omega_{i,t} &= \rho_1 \omega_{i,(t-1)} + \ldots + \rho_s \omega_{i,(t-s)} + X_{i,t}\beta + \varepsilon_{i,t} \qquad t = 1,\ldots,T, \quad i = 1,\ldots,N, \\ &\equiv Z_{i,t}\theta + \varepsilon_{i,t} \end{aligned}$$

where $\omega_{i,t}$ is the dependent variable for the i-th individual in the t-th year, $X_{i,t}$ is a vector of measured variables describing that individual, the coefficients ρ_j and β are parameters, and $\varepsilon_{i,t}$ is an error term. The elements of $X_{i,t}$ include year and age effects, measures of educational attainment, and gender and race indicators. Throughout the majority of the subsequent analysis we assume $\varepsilon_{i,t}$ is distributed

independently both across time and individuals. The autoregressive coefficients ρ_j characterize the dynamic properties of wages after removing trends.

A conventional autoregressive formulation of (2.1) invokes the moment restriction:

$$E\,(\,\varepsilon_{i,t} \mid\ \omega_{i,t-\tau}\,,\ X_{i,t}\,) = 0 \tag{2.2}$$

where $\omega_{i,t-\tau}$ signifies the past wages $\omega_{i,t-1}, ..., \omega_{i,t-s}$ appearing in (2.1). This condition implies that (2.1) characterizes how the first moment of the Markov distribution of $\omega_{i,t}$, conditional on $\omega_{i,t-\tau}$ and $X_{i,t}$, evolves over time. One applies least squares or generalized least squares methods to estimate the parameters of such formulations, suitably adjusting for heteroscedasticity or correlation in an individual's errors when appropriate.

Alternatively, one can associate relation (2.1) with an autoregressive formulation of the α-th percent quantile of the Markov distribution of $\omega_{i,t}$ by imposing the restriction:

$$Q_\alpha\,(\,\varepsilon_{i,t} \mid\ \omega_{i,t-\tau}\,,\, X_{i,t}\,) = 0 \tag{2.3}$$

where $Q_\alpha\,(\,\bullet\,)$ designates the α-th percent quantile of the distribution of $\varepsilon_{i,t}$ conditional on $\omega_{i,t-\tau}$ and $X_{i,t}$, where $\alpha \in (0, 100)$. When $\alpha = 50$, equation (2.1) determines how the conditional median of $\omega_{i,t}$ evolves over time. To our knowledge, such relations have not been estimated in a panel data context, but conceptually the application of LAD procedures would produce consistent estimates of the autoregressive coefficients appearing in (2.1).

2.2 Problems Induced by Censoring

Serious complications arise in the estimation of (2.1) under either restriction (2.2) or (2.3) when data are missing in a non-random manner. Two sources contribute to this non-random sampling: (i) individuals depart from the sample, and (ii) persons fail to work during the specified period. Both of these phenomena represent a form of data censoring that leads to biased estimation of dynamic relationships.

Attrition is inherently a dynamic phenomenon – an individual observed in one period may for some reason be unobserved in a later year, only to be brought back to the sample at some later

date. Evidence suggests [MaCurdy, Gritz, and Mroz (1996)] that individuals lost in sampling are not randomly drawn from the wage distribution; rather, these individuals tend to come from the upper segments of that distribution. Consequently, over time the distribution of wages becomes perturbed for the observations remaining in the sample. Biases in estimating wage equations attributable to persons selecting not to work are well known in the labor economics literature, having been studied for almost 30 years in the case of women's hours-of-work behavior. To the extent that low-wage women choose not to work, their wages are missing in the observed wage distribution corresponding to the employed population.

Observations making up the sample available for estimating intertemporal wage relationships must, therefore, not have departed from the sample and have worked positive hours. To evaluate the effects of this censoring, let the variable $\delta_{i,t} = 1$ indicate when an observation meets this criteria and $\delta_{i,t} = 0$ when it fails to do so. To account for biases in the estimation of equation (2.1) with moment condition (2.2), the corrected version of (2.1) takes the form:

$$\omega_{i,t} = \rho_1 \omega_{i,t-1} + \ldots + \rho_s \omega_{i,t-s} + X_{i,t}\beta + \lambda(\omega_{i,t-v}, X_{i,t}) + \zeta_{i,t} \tag{2.4}$$

where

$$\lambda(\omega_{i,t-v}, X_{i,t}) = E(\varepsilon_{i,t} \mid \omega_{i,t-v}, X_{i,t}, \delta_{i,t} = 1) \tag{2.5}$$

Developing formulae for the conditional moment λ requires the introduction of extensive distributional assumptions for $\varepsilon_{i,t}$, and, in a longitudinal setting, typically involves considerable computational burden.

2.3 Nonparametric Bounds for Quantiles

One can construct upper and lower bounds for quantiles to account for sample censoring. The idea underlying these bounds is simple. Suppose one wants to estimate the value of the α-th percent quantile of a variable $y_{i,t}^*$. One does not have data on all observations of $y_{i,t}^*$, but instead observes data on the variable $y_{i,t}$, where $\delta_{i,t} = 1$ implies $y_{i,t} = y_{i,t}^*$ and $\delta_{i,t} = 0$ implies that $y_{i,t}^*$ is not seen. Let $P = \text{Prob}(\delta_{i,t} = 0)$ represent the proportion of the sample that is missing, and let π designate the π-th percentile of the distribution of $y_{i,t}$. To form bounds for α based on measures of π, one can imagine two extreme circumstances. First, all missing observations come from the very bottom of the distribution. In this case, π may actually represent as high as $\alpha = P + \pi\,(1\text{-}P)$ of the true

percentile. Analogously, if at the other extreme, all missing observations come from the very top of the distribution, then (1-π) may actually represent as high as the true percentile 1-$\alpha = P + (1-\pi)(1-P)$. Combing these insights, one readily infers that α must lie in the interval:

(2.6) $$\pi_L \leq \alpha \leq \pi_U$$

where

(2.7) $$\pi_L = \frac{\alpha - P}{1 - P} \qquad \pi_U = \frac{\alpha}{1 - P}$$

Given a particular value of α, equations (2.6) and (2.7) show how to use information on the percentiles of $y_{i,t}$ to form bounds for α.

These bounds are equivalent to ones proposed by Manski (1995). Without prior assumptions regarding the conditional distribution of the censored values of $y_{i,t}$, Manski notes the following relationship:[4]

(2.8) $$P(y_{i,t} \leq t \mid Z_{i,t}, \delta_{i,t} = 1)\, P(\delta_{i,t} = 1 \mid Z_{i,t}) \leq P(y_{i,t} \leq t \mid Z_{i,t}) \leq P(y_{i,t} \leq t \mid Z_{i,t}, \delta_{i,t} = 1)\, P(\delta_{i,t} = 1 \mid Z_{i,t}) + P(\delta_{i,t} = 0 \mid Z_{i,t})$$

Manipulation of these inequalities implies that the inverse of $P(y_{i,t} \leq t \mid Z_{i,t})$ – i.e., the α^{th} percentile of $y_{i,t}$ given $X_{i,t}$ – falls within the bounds:

(2.9) $$\ell(\alpha, Z_{i,t}) \leq \alpha \leq \mu(\alpha, Z_{i,t})$$

where

---

[4] This inequality comes from the the Law of iterated expectations:

$$P(y \leq t \mid x) = P(y \leq t \mid x,z=0) \cdot P(z=0) + P(y \leq t \mid x,z=1) \cdot P(z=1) \quad .$$

P(y≤t |x,z=1), P(z=0), and P(z=1) are all observed in the censored data. Without any additional assumptions, we know that 0 ≤ P(y≤t|x,z=0) ≤ 1. This yields the bounds shown here for the percentile of y.

$$(2.10) \qquad \ell(\alpha, Z_{i,t}) = \left[\frac{\alpha - P(\delta_{i,t} = 0 \mid Z_{i,t})}{P(\delta_{i,t} = 1 \mid Z_{i,t})} \right]^{th} - \text{percentile of } P(y_{i,t} \mid Z_{i,t}, \delta_{i,t} = 1)$$

and

$$(2.11) \qquad \mu(\alpha, Z_{i,t}) = \left[\frac{\alpha}{P(\delta_{i,t} = 1 \mid Z_{i,t})} \right]^{th} - \text{percentile of } P(y_{i,t} \mid Z_{i,t}, \delta_{i,t} = 1) .$$

Hence, the α^{th} percentile of the uncensored distribution of $y_{i,t}$ given $Z_{i,t}$ is bounded by two percentiles calculated from the conditional distribution of the censored variable. The width of these bounds is proportional to the degree of censoring in the sample, or $P(\delta_{i,t} = 0 \mid Z_{i,t})$.

2.4 Estimating Quantiles Using Nonlinear Instrumental Procedures without Censoring

MaCurdy and Hong (1999) propose a class of quantile estimators for systems of simultaneous equation models that provides a flexible and non-cumbersome procedure for estimating parameters of the dynamic wage growth equation introduced above. In essence, assuming specifications for the quantiles of structural error distributions conditional on exogenous or predetermined instruments, the estimators formulate these conditional quantiles into moment conditions capable of being estimated within a conventional nonlinear instrumental variables or Generalized Method of Moments (GMM) framework. This apparatus matches the sample analog of the conditional quantiles against their population values, employing a smoothing procedure familiar in various problems found in non-parametric inference and simulation estimation. The analysis applies standard arguments to demonstrate consistency and asymptotic normality of the resulting Smoothed GMM quantile estimator. Simulation exercises reveal that this procedure accurately produces estimators and test statistics generated by conventional quantile estimation approaches.

To apply this GMM quantile procedure, let $\omega_{i,t}$ denote the log of hourly wages in year t for individual i, and let $X_{i,t}$ denote demographic characteristics. We are interested in obtaining information about the distribution of $\omega_{i,t}$ conditional on $X_{i,t}$ and $\omega_{i,t-\tau}$ (past hourly wages). We will use $Q_\alpha(\omega_{i,t-\tau}, X_{i,t})$ to represent the α^{th} percent quantile of this conditional distribution, where $\alpha \in (0, 100)$. Our Smoothed GMM quantile estimator makes use of the following moment conditions, which underlie the construction of most quantile estimation procedures:

(2.12) $$P(\omega_{i,t} < Q_\alpha(\omega_{i,t-\tau}, X) \mid \omega_{i,t-\tau}, X_{i,t}) = \alpha$$

This relation implies the condition

(2.13) $$E[1(\omega_{i,t} < Q_\alpha(\omega_{i,t-\tau}, X_{i,t})) - \alpha \mid \omega_{i,t-\tau}, X_{i,t}] = 0$$

where *1(•)* represents the indicator function which takes value 1 when the condition expressed in the parentheses is true, and 0 otherwise. The indicator function inside the moment condition is neither continuous nor differentiable. To incorporate this moment condition into the standard framework of nonlinear method of moments estimation, MaCurdy and Hong (1999) propose to use the modified smooth version of this condition:

(2.14) $$E\left[lim_{N\to\infty} \Phi\left(\frac{\omega_{i,t} - Q_\alpha(\omega_{i,t-\tau}, X_{i,t})}{s_N} \right) - (1 - \alpha) \right] = 0$$

where N represents the sample size, and Φ is a continuously differentiable distribution function with bounded symmetric density function φ. The following analysis selects Φ to be the cumulative standard normal distribution function; a natural alternative would be the logit distribution function. The quantity s_N is a bandwidth parameter that converges to 0 as N goes to ∞ at a rate slower than that of $N^{1/2}$. Formally, one may choose $s_N = N^{-d}$, for $0 < d < 1/2$. One can readily verify that when $s_N \to 0$, $\Phi(\cdot)$ converges almost surely to the indicator function $1(\omega_{i,t} > Q_\alpha(\omega_{i,t-\tau}, X_{i,t}))$. Since Φ is a bounded function, one can exchange expectation and limit to obtain the above smoothed moment condition. The condition imposed on the convergence rate $0 < d < ½$ is needed for the proof of asymptotic normality. A generalized nonlinear two-stage least squares estimation routine can be directly applied to this asymptotic moment condition. MaCurdy and Hong (1999) explore the performance of various choices for the bandwidth parameter in a simulation study; the estimation analysis below relies on the results of this exercise. The estimation approach selects instrumental variables that are conditionally independent of the error terms defined by $1(\omega_{i,t} > Q_\alpha(\omega_{i,t-\tau}, X_{i,t})) - \alpha$.

This estimation framework extends to consider a set of quantile relations, which may either describe several percentiles of a single conditional distribution or characterize the same quantile for marginal distributions of a variable in different time periods. To estimate any finite and fixed number of quantiles of the conditional wage distribution jointly in an efficient way, let $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_K < 1$ be the K quantiles of interest. Define a system of K simultaneous equations through the following asymptotic moment conditions:

$$(2.15) \qquad E\left[lim_{N\to\infty} \Phi\left(\frac{\omega_{i,t} - Q_{\alpha_k}(\omega_{i,t-\tau}, X_{i,t})}{s_N} \right) - (1-\alpha_k) \right] = 0, \qquad k=1,\ldots,K$$

Each of these equations can be separately estimated using single equation two-stage least squares methods. To improve efficiency given the available instruments, one can apply a three-stage nonlinear least squares or joint-equation GMM estimation procedure by weighting the K equations optimally. The optimal weighting matrix, will be determined by the variance-covariance matrix of the K sign-variables:

$$(2.16) \qquad 1[\ \omega_{i,t} > Q_{\alpha_k}(\omega_{i,t-\tau}, X_{i,t})\ ] \qquad k=1,\ldots,K$$

This matrix depends only on the α's associated with the specific distribution of the error term. In particular, $Var[1(\omega_{i,t} > Q_\alpha(\omega_{i,t-\tau}, X_{i,t})] = \alpha(1-\alpha)$, and for $\alpha_i > \alpha_j$:

$$(2.17) \qquad Cov[\ 1(\omega_{i,t} > Q_{\alpha_i}(\omega_{i,t-\tau}, X_{i,t}))\ ,\ 1(\omega_{i,t} > Q_{\alpha_j}(\omega_{i,t-\tau}, X_{i,t}))\ ] = 1-\alpha_i-(1-\alpha_i)(1-\alpha_j)$$

One can permit flexible and unknown forms of heteroscedasticity in calculating the optimal weighting matrix used in GMM estimation, as well as an unbalanced number of equations corresponding to the different observations. Incorporating these generalizations involves implementing the conventional approach utilized in multiple-equation GMM procedures. The conditional quantile, $Q_\alpha(\omega_{i,t-\tau}, X_{i,t})$, can be chosen to be any flexible nonlinear function.

Finally, the Smoothed GMM quantile procedure readily allows for weighting in estimation to account for stratified sampling, which is present in most data sets. With $Y_{i,t}$ representing the

weight supplied by the data set for individual *i* in period *t*, replace the system of structural equations (2.15) by:

$$(2.18) \qquad E\left[Y_{i,t} \left(lim_{N\to\infty} \Phi\left(\frac{\omega_{i,t} - Q_{\alpha_k}(\omega_{i,t-\tau}, X_{i,t})}{s_N} \right) - (1-\alpha_k) \right) \right] = 0, \qquad k=1,...,K.$$

The inclusion of weights in these moment conditions now defines the value of $Q_\alpha(\omega_{i,t-\tau}, X_{i,t})$ associated with the appropriate population conditional quantile. As described in the above discussion, these equations can be estimated separately using single equation two-stage least square methods, or estimated jointly using a multiple-equation GMM procedure with an optimally computed weighting matrix.

2.5 Estimating Bounds for Autoregressive Parameters

The specification of the conditional quantile function adopted in our characterization of wage dynamics is the linear distributed lag relation:

$$(2.19) \qquad Q_{\alpha_k}(\omega_{i,t-\tau}, X_{i,t}) = \rho_1 \omega_{i,t-1} + ... + \rho_s \omega_{i,t-s} + \beta X_{i,t}$$

If one had a random sample available to estimate this conditional quantile, the variant of the nonlinear simultaneous equation implied by (2.14) takes the form:

$$(2.20) \qquad \Phi\left(\frac{\omega_{i,t} - \rho_1 \omega_{i,t-1} - ... - \rho_s \omega_{i,t-s} - \beta X_{i,t}}{s_N} \right) - (1-\alpha_k) = v_{i,t}$$

where $v_{i,t}$ is treated as the error with $E(v_{i,t} \mid \omega_{i,t-1},..., \omega_{i,t-s}, X_{i,t}) = 0$.

Specifications of nonlinear simultaneous equations implied by relation (2.14), applied to estimating coefficients corresponding to the bounds (2.10) and (2.11), take the form:

$$\Phi\left(\frac{\omega_{i,t} - \rho_{L1}\,\omega_{i,t-1} - \ldots - \rho_{Ls}\,\omega_{i,t-s} - \beta_L\, X_{it}}{s_N} \right) - (1 - \ell(\alpha,\ Z_{i,t})) = v_{L,i,t}$$

(2.21)

$$\Phi\left(\frac{\omega_{i,t} - \rho_{U1}\,\omega_{i,t-1} - \ldots - \rho_{Us}\,\omega_{i,t-s} - \beta_U\, X_{it}}{s_N} \right) - (1 - \mu(\alpha,\ Z_{i,t})) = v_{U,i,t}$$

where one interprets the errors $v_{L,i,t}$ and $v_{U,i,t}$ as possessing the conditional means $E\,(v_{L,i,t} \mid \omega_{i,t-1},\ldots, \omega_{i,t-s}, X_{i,t}, \delta_{i,t} = 1) = 0$ and $E\,(v_{U,i,t} \mid \omega_{i,t-1},\ldots, \omega_{i,t-s}, X_{i,t}, \delta_{i,t} = 1) = 0$. In both (2.20) and (2.21), one sets the denominator s_N to an appropriately small number.

Reliance on the Smoothed GMM quantile approach implies that one can estimate the coefficients ρ and β by implementing conventional nonlinear IV or 2SLS/3SLS procedures to equations (2.20) and (2.21). The resulting estimators are consistent and asymptotically normally distributed with standard errors computed using robust methods. In the subsequent analysis we estimate variants of these equations separately for the conditional 25, 50, and 75th percent quantiles. While we could jointly estimate specifications for several α_i's simultaneously, we do not do so here. Conceptually, one can generalize equations (2.20) and (2.21) to allow parameters to be year (or age) dependent. Estimation in this instance would require the introduction of an equation for each quantile for each year (or age) in which a person has current and past wage observations. Within- and cross-equation restrictions on the quantile regression coefficients could be imposed in the standard way using the multi-equation GMM framework given by (2.15). In the estimation of these wage growth profiles, the proper choice of instruments include the elements of $X_{i,t}$, lagged values of the natural log of the real wage, and possibly interactions and higher order powers of these variables. One would not use the same set of instrumental variables for each equation, since the wage variables $\omega_{i,t-1},\ldots, \omega_{i,t-s}$ are predetermined for period t but not for previous years. Using different instruments is easily accomplished in GMM estimation. Finally, if weighting is required to adjust for the stratified character of a data set, then equations (2.20) and (2.21) are modified by simply multiplying these relations by a weight $Y_{i,t}$ analogous to (2.18) and then apply conventional NIV estimation methods. (This is what we mean by weighting in the subsequent analysis.)

The bounds for quantiles specified in (2.21) do not, of course, imply that estimation of the these relations creates bounds for the autoregressive coefficients ρ_j.[5] In the subsequent discussion, when we refer to bounds for the autoregressive coefficients we merely mean estimates of the values of ρ_{Lj} and ρ_{Uj}. We infer that a large discrepancy between these values suggests that results are sensitive to the presence of censoring.

3. Sample Composition and Attrition in the NLSY

This section summarizes the structure of the NLSY and describes how attrition alters its make-up during the years of the survey. The NLSY permits analyses of a number of different sample compositions that exclude non-respondents according to a variety of rules. This discussion explores the consequences of using these alternative samples to estimate wage distributions conditional upon prior observed wages and other demographic variables of interest.

3.1 Description of the NLSY

The NLSY is a multistage, stratified, clustered probability sample designed to represent the entire population of youth residing in the United States on January 1, 1979. Three independent probability samples make up the NLSY, each comprised of youth born January 1, 1957 through December 31, 1964: (i) a cross-sectional sample designed to be nationally representative of the non-institutionalized civilian population in the United States; (ii) a supplemental sample of the non-institutionalized civilian population comprised of Hispanics, Blacks, and economically-disadvantaged White (i.e., non-Hispanic, non-Black) youth; and (iii) a military sample representing men and women who served in the military on September 30, 1978. Households served as the interview unit for the civilian samples, and the military sample was drawn from rosters of active duty military personnel. The civilian samples were selected after the completion of initial screening interviews of approximately 75,000 dwelling units, and first-round interviews were completed for about 90 percent of the civilian youths designated for the base-year interviewing.

---

[5] Note, however, that in tests designed to infer the behavior of ρ_j between ρ_{Lj} and ρ_{Uj} (i.e., by calculating the values of these parameters at, for example, 20 different quantiles between $\ell(\alpha,Z)$ and $\mu(\alpha,Z)$), we found that, in every case, the value of ρ_j moved monotonically from ρ_{Lj} to ρ_{Uj}. In practice, therefore, ρ_{Lj} and ρ_{Uj} do behave like bounds on ρ_j in our particular application, although we have no reason to expect this property to necessarily carry-over to other applications.

For the purposes of our discussion, it is convenient to consider the NLSY as made up of five distinct components: (1) the cross-sectional sample comprised of 6,111 youths; (2) the supplemental sample of Hispanics with 1,480 youths; (3) the supplemental sample of Blacks containing 2,172 youths; (4) the supplemental sample of economically-disadvantaged Whites with 1,643 youths; and (5) the military sample of 1,280 personnel. We restrict our analysis to the 9,763 young men and women included in the first three components of the sample. Two factors lead us to exclude the economically-disadvantaged White supplemental sample. First, this particular supplemental sample was discontinued from the NLSY in 1991. Second, this action was taken due to serious suspicions about the representativeness of this supplemental sample; the primary criteria used to screen households into the sample was based solely on household income in 1978 – not necessarily parents' income – relative to the poverty level. We do not consider the military sample either because the vast majority of its members were not interviewed after 1983.[6]

Tables 1 (a) and (b) present simple summary statistics for the men and women included in the random sample and the supplemental samples of Hispanics and Blacks. The table reports sample sizes and statistics calculated separately for the three race-ethnic categories: Whites, Blacks, and Hispanics. The variables summarized include average hourly earnings (wages), annual reported earnings, annual imputed earnings, and annual hours of work, all measured for the calendar year preceding the year of the interview in 1990 dollars. Annual reported earnings in the NLSY corresponds to a CPS-type measure of annual earnings. Average hourly earnings equals annual reported earnings divided by annual hours of work, all referring to the previous calendar year. The variable used for the annual hours of work measure is a key variable created by the Center for Human Resource Research (CHRR) from the work history data.

A unique feature of the NLSY is the reference period used to collect the work-history data. This period extends back to the date of the last interview completed by a respondent and can span two or more calendar years. In sharp contrast, the reference period for the annual reported measures spans only the previous calendar year. Because the work-history data includes both hours and earnings information on a weekly basis, one can construct alternative measures of annual earnings. Moreover, not only can one calculate a second measure of annual earnings in the calendar year preceding an interview, one can also impute this measure in the calendar years with missed

[6] A total of 201 military respondents were retained from the original military sample of 1,280.

interviews for respondents who eventually return to the sample. We construct such "imputed earnings" for all respondents in all calendar years for which information is available.[7]

The calculation of the descriptive statistics reported in Tables 1 (a) and (b) does not use weights. The first column presents the percentage of noninterviews for the samples relevant for each row, summed over the first thirteen rounds of the NLSY. According to this column, well over ninety percent of the sample is interviewed in each year on average, and Hispanic men and women are more likely to miss interviews compared to White and Black respondents. The next four groups of columns present statistics for average hourly earnings, annual reported earnings, annual imputed earnings, and annual hours of work. The first column of each group presents the percentage of observations with missing information.[8] For imputed earnings, the first column in the group indicates the extent to which we are able to impute annual earnings during the calendar years preceding a missed interview for those respondents that return to the sample. For example, in the case of Hispanic men, data are available to calculate an imputed earnings measure for 3.8 percent of all calendar years, which translates into 46 percent of all calendar years preceding a missing interview by this subgroup (i.e., 3.8/8.2). The second column of each group reports the percentage of nonmissing observations with a value of 0. The third column presents means and standard deviations calculated over observations with positive values.[9] Since these statistics account for no weighting or sample composition of the NLSY, we do not discuss them here.

---

[7] We impute our earnings measure by first calculating an earnings measure for each week in which the respondent was working at a job, and then summing weekly earnings over the relevant weeks in each calendar year. Weekly earnings are inferred by multiplying usual hours worked per week and usual hourly wage rate for each job held by a respondent during the specific week and summing across all jobs held during this week. As a matter of convention, if for any reason either usual hours worked or usual wage is missing for a job, the job does not contribute to weekly earnings.

[8] Missing information occurs because of item non-response (refusal), a "don't know" response, a valid skip, or an invalid skip.

[9] The relatively large standard deviations result from extreme values of truncated variables or misreported/misrecorded information; these large standard deviations point out the importance of using statistics that are not unduly influenced by these extreme values. Extreme values for reported earnings result from the replacement of values above $100,001 with the average earnings of respondents who are U.S. residents and who reported values above this threshold in the latter years of the sample. The extreme values for imputed earnings stem from misreported/misrecorded values for hourly wage rates in the work-history data. Hours of work are also truncated in the data at 96 hours per week. In our subsequent analysis, we top code all earnings data to 100001, which is the level used in the earlier years of the NLSY.

3.2 Patterns of Attrition

Tables 2 (a) - (d) summarize the cumulative effects of attrition on the sample composition of the NLSY, with Tables 2 (a) and (b) reporting results for men and Tables 2 (c) and (d) listing findings for women. The first set of rows shows the fraction of original respondents who are not interviewed in each year, and the second set reports the fraction of interviewees who are returnees (i.e., who formerly departed from the sample) in each year. The tables show that:

- Approximately 10 percent of the original respondents are regularly missing in the latest years of the NLSY. (This figure is somewhat higher for men and lower for women.).

- Approximately 20 percent of the male interviewees are returnees in the latest years, while less than 15 percent of the women are returnees.

Tables 2 (a) - (d) provide more detail on the patterns of initial attrition and returning to the sample. The columns in this table represent the year of first attrition. The top set of rows shows the percentage of individuals who miss an interview for the first time in the year designated in the column, with percentages computed using those who have continuously remained in the sample as the baseline. (Thus, these percentages are estimates for the hazard rates for first leaving the NLSY.) The results are given for the entire sample and for race-ethnic groups and age cohorts. The second set of rows gives the fraction of these initial attritions in each year who never return to the sample by 1991, by age cohort. The third set of rows reports the fraction of these attritions who miss more than one interview but who return in at least one year after initial attrition, again by age cohort. The fourth set of rows gives the average number of spells experienced by individuals who attrit for multiple-years (identified in the previous set of rows), along with the average number of years that they are missing during the year brackets 1980-83, 1984-87, and 1988-91.

We see that initial attrition rates never exceed six percent for men or women, and these rates are typically in the one percent to three percent range. Among the race-ethnic groups, Hispanics experience the highest rates of initial attrition, especially in the early years of the survey. Among the cohorts, attrition rates tend to rise with age, once again particularly early in the survey. The second set of rows (percentage of those who attrit and never return) show that of the persons who depart in the early years, around 10 to 25 percent are never seen again by 1991. This range rises to

as much as 50 percent in the later years reflecting, in part, the shorter length of time available for finding these individuals and recruiting them back into the sample. The results presented in the third set of rows show that around 50 percent of those who miss at least one interview and who are not lost forever (i.e., who return at some time before 1991) end up missing two or more interviews. The figures reported for the number of spells experienced by those who attrit for multiple-years indicate that intermittent periods of absence are common. Finally, the findings for the average numbers of missing years during the various time horizons, listed in the bottom rows of the tables, further show that periods of absence for the multi-year attritions involve about the same numbers of years in the early in the NLSY as in its later interviews.

4. Empirical Analysis of the Effects of Censoring on Wage Dynamics in the NLSY

This section applies the ideas described in Section 2 to assess the effects of attrition and non-work-participation on the estimates of autoregressive coefficients characterizing wage dynamics in the NLSY. The analysis considers both men and women, and accounts for nonrandom selection attributable to several sources of censoring.

4.1 A Two-Step Estimation Approach

Application of our technique practically involves a two-step estimation procedure. The first step estimates the probabilities of censoring needed to construct the quantile bounds given by (2.10) and (2.11). The second inserts these estimated bounds in the specification of the nonlinear structural equations appropriate to estimate particular conditional quantiles and estimates the resulting relationships using NIV methods.

To calculate $\ell(\alpha, Z_{i,t})$ and $\mu(\alpha, Z_{i,t})$ we apply a probit estimation procedure to derive fitted values for $P(\delta_{i,t} = 1 \mid Z_{i,t})$; i.e., the probability that an individual i appears in the sample in year t given attributes $Z_{i,t}$. We apply weighted maximum likelihood estimation, using the $Y_{i,t}$ weights provided by the NLSY to account for stratified sampling. We include the following variables in $Z_{i,t}$ in carrying out this analysis:

(4.1) $$Z_{it} = [\textit{year effects, age, age}^2\textit{, education, black, Hispanic, } \delta_{i,t-\tau}]$$

where the education variable signifies to the highest year of education completed by the individual, and $\delta_{i,t-\tau}$ refers to i's attrition status in year t-τ. With fitted values of the conditional probability of appearing in the sample, we compute fitted Manski bounds on the α^{th} quantile of the conditional wage distribution according to:

$$(4.2) \qquad \hat{\ell}(\alpha,\ Z_{i,t}) = \left[\frac{\alpha - \hat{P}(\delta_{i,t}=0 \mid Z_{i,t})}{\hat{P}(\delta_{i,t}=1 \mid Z_{i,t})} \right] \qquad \hat{\mu}(\alpha,\ Z_{i,t}) = \left[\frac{\alpha}{\hat{P}(\delta_{i,t}=1 \mid Z_{i,t})} \right].$$

In the second stage of the estimation, we apply weighted GMM procedures to compute values for the parameters appearing in the nonlinear simultaneous equations:

$$(4.3) \qquad Y_{i,t}\left[\Phi\left(\frac{\omega_{i,t} - \rho_1 \omega_{i,t-1} - \ldots - \rho_s \omega_{i,t-s} - \beta X_{i,t}}{s_N} \right) - (1 - \alpha_k) \right] = v_{i,t}$$

and

$$Y_{i,t}\left[\Phi\left(\frac{\omega_{i,t} - \rho_{L1} \omega_{i,t-1} - \ldots - \rho_{Ls} \omega_{i,t-s} - \beta_L X_{i,t}}{s_N} \right) - (1 - \hat{\ell}(\alpha_k,\ Z_{it})) \right] = v_{L,i,t}$$

(4.4)

$$Y_{i,t}\left[\Phi\left(\frac{\omega_{i,t} - \rho_{U1} \omega_{i,t-1} - \ldots - \rho_{Us} \omega_{i,t-s} - \beta_U X_{i,t}}{s_N} \right) - (1 - \hat{\mu}(\alpha_k,\ Z_{i,t})) \right] = v_{U,i,t}.$$

using only data satisfying the criterion $\delta_{i,t} = 1$.

The following analysis estimates variants of (4.2) - (4.4) for the $\alpha_k = 25$, 50, and 75th percent quantiles of the conditional wage distribution, although we cannot estimate bounds the 25th and 75th percent quantiles when censoring exceeds 25%. In estimating (4.2) - (4.4), we consider three sets of instrumental variables in the implementation of NIV methods:

(4.5) *(i)* $\omega_{i,t-1}$ *and* $X_{i,t}$ *= [year effects, age, age^2, education, black, Hispanic]*
(ii) $\omega_{i,t-2}$ *and* $X_{i,t}$ *= [year effects, age, age^2, education, black, Hispanic]*
(iii) $\omega_{i,t-2}$, $\omega_{i,t-2}$, *and* $X_{i,t}$ *= [year effects, age, age^2, education, black, Hispanic]*

We estimate each equation separately using single-equation GMM procedures. Reported standard errors from the second-stage do not account for estimation error introduced by the fitted probabilities of censoring.

The data used for this analysis come from the original NLSY 1979-1991, with the selection based on two criteria: (i) any individual-year observations reporting missing values for completed grades are dropped, and (ii) individuals are sampled only after leaving school and entering the labor force. Practically, the second criterion is defined as being satisfied when an individual's reported number of completed grades stops growing. Altogether, selecting based on these criteria reduced the sample from its original 126,919 individual-year observations to 64,767 individual-year observations.

4.2 Estimation of Quantile Coefficients

We do not report estimates obtained from our probit maximum likelihood for the sake of brevity. Nearly all parameter estimates were individually statistically significant and had signs corresponding to the description of the attrition data in Section 3.

Tables 3 (a) - (f) report our findings for the autoregressive coefficient estimates calculated using the second-stage smoothed GMM quantile regressions for α_k = 50 (i.e., the conditional medians) and α_k = 25 and 75. Tables 3 (a) - (c) present results for men and Tables 3 (d) - (f) report estimates for women. These tables report results for the three specifications described above. Tables 3 (a) and (d) list estimates of ρ_1 for Specification #1, which assumes equations (4.3) - (4.4) have a 1st-order autoregressive structure (i.e., restrict $\rho_2 = ... = \rho_s = 0$) and uses instrument set (i) from (4.5). Tables 3 (b) and (e) present estimates from Specification #2, which assumes the same 1st-order autoregressive formulations for (4.3) - (4.4), but which employs instrument set (ii) from (4.5). Finally, Tables 3 (c) and (f) show estimates for Specification #3, which relies on instrument set (iii) to estimate variants of (4.3) - (4.4) taking the form of a 2nd-order autoregressive model. In each table, the results listed in the columns "ignoring censoring" are for the autoregressive coefficients associated with equation (4.3), and the estimates given in the columns "upper bound" and "lower

bound" are for the autoregressive coefficients associated with equations (4.4). As a check of whether the autoregressive coefficients are time invariant (as has been assumed throughout this analysis), each table presents three sets of results. The top set of rows lists findings for all years pooled, and the lower two sets report coefficient values for separate specifications estimated using only data for the years 1985 and 1990, respectively.

The estimates account for two sources of censoring: (i) an individual departs from the sample in any portion of the periods covered by the data appearing in a specification; and (ii) an individual does not work in any period covered by the specification. Unless an individual does not attrit according to both of these criteria, he/she cannot be included among the observations used in estimation because there is missing data on his or her wages. In the case of men, the first source is the principal reason for censoring because men typically work continuously after leaving school; i.e., the estimated bounds for men primarily reflect the consequences of attrition from the survey. On the other hand, both sources are relevant for women. Thus, the estimated bounds for women show the combined effects of both types of sample censoring.

4.3 Interpretation of Results

The findings for men in Tables 3 (a) - (c) tell a straightforward story. The estimated autoregressive coefficients are virtually identical for the upper-bound, ignoring-censoring, and lower-bound specifications, regardless of whether one considers the 1st- or 2nd-order autoregressive formulations. Thus, censoring appears to play no role in biasing coefficient estimates determining how the conditional quantiles of future wages relate to current wages. In sharp contrast, evidence suggests that autoregressive coefficients vary over time and across cohorts. Not only do the values of ρ_1 vary across the pooled, 1985 and 1990 data, but so do the values of ρ_2. The estimates of ρ_2 are statistically significant but relatively small for the pooled and 1990 samples. However, this is not the case for the 1985 sample. Irrespective of which autoregressive specification one selects, the findings indicate that a high degree of dependence of current wages on past wages; the 1st-order coefficients regularly reach the value of 0.9 which implies a slow decay in the effects of current wages on the distribution of future earnings.

The results for women in Tables 3 (d) - (f) reveal that the ranges for estimated autoregressive coefficients are wider than those found for men, with large bounds found for the 25th and 75th

percentiles than for the median. This is hardly surprising given the generally higher conditional probability of censoring found for women, attributable to more of them being lost due to non-participation in work. The bounds on the autoregressive coefficients derived from the median regressions are, however, especially tight, suggesting that not much has been lost in the way of accuracy arising from censoring when fitting this quantile of women's wages. In exactly the same way as was seen for men, the results for women convincingly reveal that autoregressive coefficients vary over time or across cohorts. Also as was true for men, the estimated values of these coefficients attest to a high degree of dependence of current wages on past wages.

In an effort to gain some understanding of whether the similarity of the autoregressive coefficients found in Tables 3(a) - (f) translate into tight bounds for the future wage quantiles forecasted using current and past wages, Tables 4 (a) and (b) report averages of the fitted values of the bounds for the 25, 50, and 75th percent quantiles, for all estimated specifications and instrument sets. These fitted values are computed across all individuals in the sample who have non-missing data for all variables appearing in estimated autoregressive relations. In general, the bounds are quite tight and are narrower for men than for women. The closeness of the averaged upper and lower bounds indicates that percentiles for the conditional distributions of wages are themselves close in value, at least in the middle 50% of the distribution. Thus, much of the variation we see in wages reflects variation across persons rather than variation over time for particular individuals. This finding alone would readily explain why the autoregressive coefficients are relatively insensitive to the effects of censoring.

5. Concluding Remarks

The empirical technique outlined in this paper for assessing the effects of censoring on the estimation of conditional quantiles has many potential applications. Nonrandom sample selection is a particularly relevant concern given its documented prevalence in many popular longitudinal data sets. Moreover, the loss of observations due to individuals self-selecting themselves out of the sample (e.g., by choosing not to work) is an additional source of censoring that possibly contaminates the randomness of samples and biases estimates. The analysis presented here demonstrates that estimating bounds for conditional quantiles, accounting for such censoring, can produce relatively robust values for estimated coefficients.

In general, under a conservative definition of attrition, estimation of a relatively assumption-free set of bounds on a quantile of the uncensored conditional distribution does a good job of identifying autoregressive parameters determining conditional wage distributions. A comparison of findings between men and women suggests that censoring due to persons not working induces larger biases than sample losses due to attrition alone. Moreover, our results indicate greater sensitivity at the lower quantiles of the wage distribution. All considered, our application to the NLSY suggests that not much is lost in the way of accuracy in moving from a traditional method of ignoring sample selection to the robust econometric procedure introduced here.

Our findings reveal that autoregressive coefficients of conditional wage quantiles change over time or across cohorts, so one must allow for this feature in order to model wage dynamics accurately. One further needs to explore the way in which parameterizations might vary across education and race groups. In further developing our econometric method, our approach will be generalized to admit more sophisticated serial dependence in wages and disturbances, and to recognize adjustments for estimation error encountered in our two-step econometric approach. Finally, while in principle applicable, our methods have not been tested to determine how well they would work in the presence of more severe forms of attrition (e.g., when persons are dropped forever after they miss their first year in a longitudinal sample). We intend to pursue all of these generalizations in future research.

References

Frankel, Martin R., Harold A. McWilliams and Bruce D. Spencer. 1983. "National Longitudinal Survey of Labor Force Behavior, Youth Survey: Technical Sampling Report." NORC, University of Chicago.

Gottschalk, P. and R. Moffitt. 1998. "Earnings and Wage Distributions in the NLS, CPS and PSID." *Journal of Human Resources*. 33(2):251-99.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. "What Do We Really Know About Wages? The Importance of Non-reporting and Census Imputation." *Journal of Political Economy*. 94(3): 489-506.

MaCurdy, T., and H. Hong. 1999. "Smoothed Quantile Regression in Generalized Method of Moments." Mimeo, Stanford University.

MaCurdy, T., T. Mroz, and R.M. Gritz. 1998. "An Evaluation of the NLSY: A Comparison with CPS and an Examination of Attrition." *Journal of Human Resources*. 33:345-436.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences.* Harvard University Press. Cambridge, MA.

Olley, Steven and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica*. 64 (6): 1263-97.

Table 1 (a)
Basic Summary Statistics for Various Demographic Samples -- Men
Nominal Values, Unweighted

Sample Composition (size)	Descriptive Statistics												
	% Not Interviewed	Average Hourly Earnings			Reported Earnings			Imputed Earnings			Annual Hours of Work		
		% Missing[1]	% Zero[2]	Mean (standard deviation)[3]	% Missing	% Zero	Mean (standard deviation)	% Imputed	% Zero	Mean (standard deviation)	% Missing	% Zero	Mean (standard deviation)
Entire Sample (4837)	7.1	8.8	10.9	7.24 (9.16)	4.9	10.9	12356 (14420)	3.0	16.9	17733 (335963)	5.9	10.8	1703 (863)
Random Sample (3003)	6.9	8.4	8.8	7.60 (9.86)	4.1	8.8	13332 (16077)	2.9	14.4	18259 (285520)	5.9	8.4	1754 (866)
Whites (2439)	6.9	8.2	8.0	7.73 (9.56)	3.7	8.0	13816 (16837)	2.9	13.5	17574 (200710)	6.0	7.5	1785 (868)
Blacks (1451)	6.8	9.7	16.4	6.37 (8.11)	6.5	16.4	9843 (10339)	2.9	23.3	16765 (464529)	5.7	17.0	1545 (864)
Hispanics (947)	8.2	9.1	10.2	7.11 (9.31)	5.2	10.2	11892 (11318)	3.8	15.8	19513 (402974)	6.0	9.9	1706 (816)

1 For variables other than imputed earnings, this column shows the percentage of values associated with negative codings reflecting "don't know", refusal to answer, and valid and invalid skips. For imputed earnings, this column shows the percentage of observations for which no retrospective information is available to infer imputed earnings.

2 This column shows the percentage of observations with non-missing information (due to non-interview or missing for other reasons) that equal zero.

3 The mean and standard deviation are calculated for samples incorporating only those observations with positive values of the specified variable.

Table 1 (b)
Basic Summary Statistics for Various Demographic Samples -- Women
Nominal Values, Unweighted

Sample Composition (size)	Descriptive Statistics												
	% Not Interviewed	Average Hourly Earnings			Reported Earnings			Imputed Earnings			Annual Hours of Work		
		% Missing[1]	% Zero[2]	Mean (standard deviation)[3]	% Missing	% Zero	Mean (standard deviation)	% Imputed	% Zero	Mean (standard deviation)	% Missing	% Zero	Mean (standard deviation)
Entire Sample (4926)	5.5	8.2	20.3	6.20 (29.18)	5.3	20.3	8855 (11666)	2.0	25.8	16496 (431652)	5.1	19.8	1421 (788)
Random Sample (3108)	5.5	7.9	17.2	6.43 (35.42)	4.6	17.2	9137 (12036)	2.1	22.4	17191 (450383)	5.1	16.5	1443 (781)
Whites (2477)	5.3	7.6	14.9	6.58 (38.77)	4.0	14.9	9351 (12114)	2.1	19.8	18606 (493019)	5.2	14.1	1461 (777)
Blacks (1472)	4.6	8.9	26.9	5.59 (7.64)	6.9	26.9	8113 (12674)	1.5	33.1	10301 (133135)	4.8	27.0	1367 (811)
Hispanics (977)	7.4	8.6	24.0	5.92 (6.40)	5.9	24.0	8432 (8094)	2.5	29.7	19307 (514628)	5.2	23.3	1380 (778)

1 For variables other than imputed earnings, this column shows the percentage of values associated with negative codings reflecting "don't know", refusal to answer, and valid and invalid skips. For imputed earnings, this column shows the percentage of observations for which no retrospective information is available to infer imputed earnings.

2 This column shows the percentage of observations with non-missing information (due to non-interview or missing for other reasons) that equal zero.

3 The mean and standard deviation are calculated for samples incorporating only those observations with positive values of the specified variable.

Table 2 (a)
Attrition Patterns for Men -- First Attrition

Attrition Measure	Attrition Year											
	80	81	82	83	84	85	86	87	88	89	90	91
First Attrition												
% attrition												
cohort:												
14-15	2.4	0.9	1.6	1.2	2.1	2.0	3.1	3.0	3.3	2.0	2.5	2.6
16-17	3.9	1.9	1.9	1.2	2.0	2.9	3.6	4.5	2.6	2.3	2.7	1.9
18-19	5.3	2.1	1.9	1.5	2.8	3.5	2.6	4.3	3.1	2.9	3.2	2.0
20-21	5.9	3.1	2.6	1.2	2.6	3.4	3.8	3.8	3.4	1.7	2.2	1.9
% never return												
cohort:												
14-15	11.1	10.0	38.9	0.0	13.6	14.3	18.8	20.0	12.5	47.4	30.4	-
16-17	3.7	24.0	8.0	13.3	23.1	30.6	22.7	20.8	20.7	20.0	27.6	-
18-19	7.6	20.8	22.7	17.6	19.4	28.9	14.8	14.0	36.7	40.7	34.5	-
20-21	16.1	35.7	8.7	30.0	27.3	25.0	33.3	13.8	36.0	41.7	26.7	-
% multiple attrition												
cohort:												
14-15	55.6	60.0	27.8	53.8	50.0	52.4	43.8	30.0	21.9	26.3	-	-
16-17	63.0	48.0	56.0	46.7	61.5	52.8	31.8	32.1	10.3	24.0	-	-
18-19	56.1	45.8	54.5	76.5	54.8	50.0	55.6	27.9	16.7	22.2	-	-
20-21	53.6	50.0	60.9	30.0	45.5	42.9	23.3	37.9	20.0	25.0	-	-
Multiple Attrition												
Average # attrition spells[1]	2.1	2.0	2.2	1.9	1.9	1.6	1.7	1.7	1.4	1.3	-	-
Average missing years in:												
1980-83	2.4	2.0	1.4	1.0	-	-	-	-	-	-	-	-
1984-87	1.5	1.7	1.8	2.1	2.5	2.3	1.6	1.0	-	-	-	-
1988-91	1.6	1.7	1.7	1.5	1.8	1.4	1.4	1.8	2.4	2.0	-	-
% attrition												
Race-Ethnic Groups:												
All	4.3	2.0	2.0	1.2	2.4	2.9	3.3	3.9	3.1	2.2	2.7	2.1
Whites	3.6	2.2	1.3	0.8	2.8	2.5	3.1	3.3	2.8	2.3	2.3	1.5
Blacks	4.2	1.4	2.4	1.7	1.7	3.3	3.5	4.1	2.9	2.7	3.4	3.3
Hispanics	6.3	2.5	3.0	1.5	2.5	3.5	3.6	5.3	4.2	1.3	2.5	2.0

[1] This row reports the average number of attrition spells experienced by individuals who initially attrit in the specified year, who attrit for more than one year, and who return at sometime after this initial attrition.

Table 2 (b)
Attrition Patterns for Men

Total Attrition												
% Attrition												
Race-Ethnic Groups:												
All	4.3	3.6	4.5	4.1	5.3	6.8	9.2	11.2	10.7	10.0	11.5	11.0
Whites	3.6	3.8	3.8	3.4	5.6	6.7	9.0	10.6	10.5	10.3	11.6	10.5
Blacks	4.2	2.8	4.8	4.1	4.6	7.0	9.1	10.6	9.5	9.1	10.6	11.7
Hispanics	6.3	4.5	6.2	5.5	5.5	7.1	9.8	13.7	13.3	10.7	12.8	11.3
Fraction of Interviewed Population Who are Returnees												
% Returnees												
Race-Ethnic Groups:												
All	-	2.7	3.7	5.4	6.5	7.7	8.5	10	13.3	15.9	16.8	19
Whites	-	2	3.3	4.5	5	6.3	6.9	8.3	11.1	13.2	14	16.3
Blacks	-	2.8	3.2	5.5	6.6	7.4	8.6	10.9	14.5	17.1	18.6	20.3
Hispanics	-	4.3	5.5	7.7	10.1	11.7	12.3	13.2	17.3	20.8	20.9	23.8

Table 2 (c)
Attrition Patterns for Women -- First Attrition

Attrition Measure	Attrition Year											
	80	81	82	83	84	85	86	87	88	89	90	91
First Attrition												
% attrition												
cohort:												
14-15	2.8	0.7	0.8	0.7	1.3	1.2	2.7	3.3	3.4	0.9	1.8	1.7
16-17	2.7	1.3	1.4	1.1	1.3	1.8	2.4	1.5	2.8	0.9	2.3	0.8
18-19	4.5	2.0	1.8	0.8	1.8	2.0	2.2	2.3	2.4	1.6	1.7	0.9
20-21	5.2	1.8	1.6	1.5	1.6	2.7	2.2	2.8	2.5	1.0	1.7	1.1
% never return												
cohort:												
14-15	16.7	0.0	0.0	42.9	23.1	41.7	7.7	29.0	32.3	50.0	25.0	-
16-17	21.6	17.6	11.1	33.3	23.5	21.7	30.0	5.3	20.6	40.0	29.6	-
18-19	12.1	20.8	9.1	22.2	23.8	26.1	28.0	12.0	26.9	35.3	38.9	-
20-21	12.3	10.5	18.8	20.0	25.0	19.2	19.0	15.4	21.7	44.4	26.7	-
% multiple attrition												
cohort:												
14-15	60.0	57.1	62.5	42.9	46.2	50.0	69.2	25.8	12.9	12.5	-	-
16-17	51.4	70.6	38.9	53.3	58.8	56.5	43.3	47.4	23.5	30.0	-	-
18-19	55.2	50.0	63.6	55.6	57.1	47.8	56.0	40.0	34.6	29.4	-	-
20-21	64.9	68.4	37.5	46.7	43.8	53.8	61.9	34.6	26.1	33.3	-	-
Multiple Attrition												
Average # attrition spells[1]	1.9	1.9	2.1	1.9	1.9	1.6	1.6	1.6	1.4	1.6	-	-
Average missing years in:												
1980-83	2.5	2.0	1.4	1.0	-	-	-	-	-	-	-	-
1984-87	1.4	1.6	2.3	1.7	2.5	2.3	1.7	1.0	-	-	-	-
1988-91	1.2	1.5	2.1	1.4	1.7	1.5	1.6	1.9	2.6	2.0	-	-
% Attrition												
Race-Ethnic Groups:												
All	3.8	1.5	1.4	1.0	1.5	2.0	2.4	2.4	2.8	1.1	1.9	1.1
Whites	3.8	1.4	1.1	0.9	1.7	1.7	2.3	1.9	2.2	0.9	1.8	0.7
Blacks	3.3	1.1	0.9	1.3	1.2	1.2	1.6	2.4	3.0	1.6	1.6	1.7
Hispanics	4.3	2.4	3.0	1.0	1.6	3.8	3.6	3.8	3.9	1.1	2.7	1.0

[1] This row reports the average number of attrition spells experienced by individuals who initially attrit in the specified year, who attrit for more than one year, and who return at sometime after this initial attrition.

Table 2 (d)
Attrition Patterns for Women

Total Attrition												
% Attrition												
Race-Ethnic Groups:												
All	3.8	3.3	3.9	3.2	4.1	5.2	7.0	8.1	9.0	7.3	8.9	8.0
Whites	3.8	3.2	3.7	3.3	4.5	5.3	7.0	7.7	8.1	7.1	8.3	7.3
Blacks	3.3	2.7	2.9	2.6	3.1	3.6	5.0	6.0	7.5	6.3	8.4	8.4
Hispanics	4.3	4.3	5.6	4.1	4.8	7.3	10.3	12.1	13.4	9.3	11.0	9.1
Fraction of Interviewed Population Who are Returnees												
% Returnees												
Race-Ethnic Groups:												
All	-	2	2.8	4.4	5	5.8	6.2	7.4	9.1	11.8	12	13.7
Whites	-	2.1	2.7	4	4.4	5.2	5.8	6.9	8.5	10.3	10.7	12.3
Blacks	-	1.6	2.2	3.8	4.5	5.1	5.4	6.6	7.9	10.6	10.1	11.6
Hispanics	-	2.4	3.9	6.4	7.2	8.4	8.7	10.4	12.5	17.4	18.2	20.6

Table 3 (a)
Estimated Bounds for Autoregressive Quantile Coefficients for Men, Specification #1
Asymptotic standard errors in parentheses

Sample Period	Recognized Source of Censoring	α	Autoregressive Coefficients					
			1st Order			2nd Order		
			Upper Bound	Lower Bound	Ignoring Censoring	Upper Bound	Lower Bound	Ignoring Censoring
Pooled (All Years)	Attrition and Working in Contiguous Years	25	0.852672 (0.008434)	0.845286 (0.009241)	0.850772 (0.008777)			
		50	0.834229 (0.007706)	0.840362 (0.007643)	0.837325 (0.007836)			
		75	0.702084 (0.011372)	0.718560 (0.010236)	0.714496 (0.010015)			
1985		25	0.791278 (0.023406)	0.791385 (0.032539)	0.791629 (0.024903)			
		50	0.764129 (0.029762)	0.768963 (0.030301)	0.766684 (0.030110)			
		75	0.602281 (0.033647)	0.609330 (0.033797)	0.607361 (0.031988)			
1990		25	0.824670 (0.032482)	0.805601 (0.034980)	0.820043 (0.033479)			
		50	0.855091 (0.019947)	0.859759 (0.019470)	0.857423 (0.020051)			
		75	0.789570 (0.020710)	0.799214 (0.016954)	0.797815 (0.017046)			

Table 3 (b)
Estimated Bounds for Autoregressive Quantile Coefficients for Men, Specification #2
Asymptotic standard errors in parentheses

Sample Period	Recognized Source of Censoring	α	Autoregressive Coefficients					
			1st Order			2nd Order		
			Upper Bound	Lower Bound	Ignoring Censoring	Upper Bound	Lower Bound	Ignoring Censoring
Pooled (All Years)	Attrition and Working in Contiguous Years	25	0.898478 (0.008976)	0.893413 (0.010580)	0.897527 (0.009508)			
		50	0.889850 (0.007736)	0.894278 (0.007507)	0.892194 (0.007615)			
		75	0.856922 (0.055738)	0.810676 (0.009727)	1.808574 (0.009720)			
1985		25	0.955998 (0.029673)	0.959542 (0.036471)	0.956439 (0.030574)			
		50	0.904360 (0.026521)	0.906227 (0.024886)	0.905515 (0.025327)			
		75	0.856922 (0.055738)	0.859978 (0.058097)	0.859297 (0.055596)			
1990		25	0.808629 (0.038549)	0.796716 (0.040744)	0.806405 (0.038393)			
		50	0.834992 (0.028609)	0.836821 (0.037590)	0.835992 (0.033320)			
		75	0.770239 (0.024417)	0.778537 (0.021376)	0.776669 (0.022773)			

Table 3 (c)
Estimated Bounds for Autoregressive Quantile Coefficients for Men, Specification #3
Asymptotic standard errors in parentheses

Sample Period	Recognized Source of Censoring	α	Autoregressive Coefficients					
			1st Order			2nd Order		
			Upper Bound	Lower Bound	Ignoring Censoring	Upper Bound	Lower Bound	Ignoring Censoring
Pooled (All Years)	Attrition and Working in Contiguous Years	25	0.835438 (0.020573)	0.830271 (0.019237)	0.834375 (0.020603)	0.040305 (0.018770)	0.041150 (0.015401)	0.040334 (0.018023)
		50	0.818207 (0.010291)	0.821169 (0.010491)	0.819703 (0.010367)	0.050554 (0.005653)	0.050556 (0.006040)	0.050568 (0.005818)
		75	0.691058 (0.024398)	0.701930 (0.029725)	0.698913 (0.028190)	0.071133 (0.021579)	0.069582 (0.025328)	0.070078 (0.024537)
1985		25	0.610191 (0.074545)	0.612329 (0.096092)	0.610914 (0.078296)	0.259218 (0.092934)	0.253845 (0.109836)	0.257495 (0.093720)
		50	0.628226 (0.061926)	0.627653 (0.059134)	0.627941 (0.060633)	0.178970 (0.060377)	0.181589 (0.057385)	0.486481 (0.089291)
		75	0.478980 (0.087990)	0.486481 (0.089291)	0.484530 (0.088699)	0.208537 (0.063754)	0.208116 (0.063335)	0.209024 (0.063525)
1990		25	0.901473 (0.061022)	0.882508 (0.061426)	0.897136 (0.061072)	-0.072214 (0.047487)	-0.063236 (0.048672)	-0.070073 (0.046670)
		50	0.897092 (0.041202)	0.902314 (0.042372)	0.899620 (0.042581)	-0.048821 (0.025980)	-0.050591 (0.025227)	-0.049657 (0.025829)
		75	0.843824 (0.038140)	0.841442 (0.042345)	0.842058 (0.041260)	-0.052712 (0.037544)	-0.046024 (0.042789)	-0.047608 (0.041999)

Table 3 (d)
Estimated Bounds for Autoregressive Quantile Coefficients for Women, Specification #1
Asymptotic standard errors in parentheses

Sample Period	Recognized Source of Censoring	α	Autoregressive Coefficients					
			1st Order			2nd Order		
			Upper Bound	Lower Bound	Ignoring Censoring	Upper Bound	Lower Bound	Ignoring Censoring
Pooled (All Years)	Attrition and Working in Contiguous Years	25	0.894144 (0.009077)	0.874546 (0.013406)	0.888980 (0.009999)			
		50	0.838520 (0.009377)	0.864776 (0.008324)	0.853740 (0.007440)			
		75	0.639189 (0.013003)	0.712737 (0.013439)	0.696472 (0.012797)			
1985		25	0.774688 (0.069445)	0.803268 (0.035838)	0.0782338 (0.076248)			
		50	0.626360 (0.040396)	0.670796 (0.017898)	0.651863 (0.054066)			
		75	0.441419 (0.024952)	0.499531 (0.019266)	0.494723 (0.056771)			
1990		25	0.918881 (0.027193)	0.893843 (0.015483)	0.913279 (0.024131)			
		50	0.925118 (0.014541)	0.929544 (0.012908)	0.930493 (0.012024)			
		75	0.818567 (0.032232)	0.862047 (0.020172)	0.853032 (0.017232)			

Table 3 (e)
Estimated Bounds for Autoregressive Quantile Coefficients for Women, Specification #2
Asymptotic standard errors in parentheses

Sample Period	Recognized Source of Censoring	α	Autoregressive Coefficients					
			1st Order			2nd Order		
			Upper Bound	Lower Bound	Ignoring Censoring	Upper Bound	Lower Bound	Ignoring Censoring
Pooled (All Years)	Attrition and Working in Contiguous Years	25	0.948260 (0.011316)	0.933461 (0.017123)	0.945042 (0.013747)			
		50	0.909419 (0.008928)	0.929802 (0.007466)	0.921565 (0.008290)			
		75	0.766675 (0.014538)	0.825381 (0.009791)	0.813894 (0.011167)			
1985		25	0.950810 (0.056570)	1.021600 (0.135190)	0.950116 (0.062244)			
		50	0.875832 (0.072970)	0.918084 (0.045469)	0.902622 (0.051327)			
		75	0.681774 (0.044099)	0.719380 (0.075816)	0.704034 (0.045803)			
1990		25	0.871659 (0.025304)	0.829001 (0.036390)	0.863954 (0.034735)			
		50	0.915512 (0.019367)	0.916755 (0.017751)	0.919119 (0.019951)			
		75	0.800699 (0.036082)	0.855194 (0.021042)	0.845489 (0.015898)			

Table 3 (f)
Estimated Bounds for Autoregressive Quantile Coefficients for Women, Specification #3
Asymptotic standard errors in parentheses

Sample Period	Recognized Source of Censoring	α	Autoregressive Coefficients					
			1st Order			2nd Order		
			Upper Bound	Lower Bound	Ignoring Censoring	Upper Bound	Lower Bound	Ignoring Censoring
Pooled (All Years)	Attrition and Working in Contiguous Years	25	0.907281 (0.014537)	0.894997 (0.016630)	0.903545 (0.017484)	0.024384 (0.010384)	0.023522 (0.012845)	0.024619 (0.014256)
		50	0.857159 (0.012081)	0.875828 (0.015863)	0.866739 (0.013888)	0.036310 (0.010629)	0.033032 (0.013309)	0.034930 (0.033826)
		75	0.653650 (0.021563)	0.728371 (0.018822)	0.711111 (0.017872)	0.072097 (0.014209)	0.055307 (0.010788)	0.059022 (0.009891)
1985		25	0.732090 (0.066894)	0.761420 (0.121165)	0.735394 (0.060763)	0.108641 (0.080187)	0.097725 (0.096267)	0.106765 (0.060424)
		50	0.535072 (0.073386)	0.564793 (0.085427)	0.538561 (0.038166)	0.188120 (0.059047)	0.185370 (0.070442)	0.190785 (0.054961)
		75	0.241154 (0.038473)	0.342378 (0.071691)	0.332611 (0.028270)	0.302818 (0.052382)	0.248765 (0.035403)	0.255491 (0.030853)
1990		25	0.967804 (0.041614)	0.970496 (0.032784)	0.971485 (0.044511)	-0.067866 (0.031152)	-0.114877 (0.045597)	-0.077600 (0.037993)
		50	0.950122 (0.015669)	0.955434 (0.020092)	0.953728 (0.015173)	-0.021172 (0.005323)	-0.020142 (0.009770)	-0.019825 (0.004929)
		75	0.871606 (0.024644)	0.940416 (0.060933)	0.904293 (0.044973)	-0.040013 (0.006422)	-0.068031 (0.052959)	-0.050164 (0.032567)

Table 4 (a)
Average Fitted Bounds on Quantiles of the Conditional Log Wage Distribution
Men, $\delta_{i,t} = 0$ if Attrition or Hours = 0

Specification	Bound	25th-Percentile	50th-Percentile	75th-Percentile
#1	Upper	1.82939	1.95860	2.11027
	Lower	1.81599	1.95326	2.10537
#2	Upper	1.89227	2.00457	2.13250
	Lower	1.88048	2.00029	2.12898
#3	Upper	1.93969	2.05453	2.18506
	Lower	1.93057	2.05138	2.18182

Table 4 (b)
Average Fitted Bounds on Quantiles of the Conditional Log Wage Distribution
Women, $\delta_{i,t} = 0$ if Attrition or Hours = 0

Specification	Bound	25th-Percentile	50th-Percentile	75th-Percentile
#1	Upper	1.65110	1.76577	1.91350
	Lower	1.59881	1.74941	1.89790
#2	Upper	1.71692	1.81443	1.93900
	Lower	1.66606	1.80069	1.92546
#3	Upper	1.76594	1.86534	1.99390
	Lower	1.72519	1.85352	1.98105