Stock Sampling with Interval-Censored Elapsed Duration: A Monte Carlo Analysis

Michael P. Babington* and Javier Cano-Urbina†

August 31, 2018

Abstract

Duration data obtained from a given stock of individuals can fail to observe those with relatively short spells. The bias generated by this sampling scheme is known as length-biased sampling because long spells are more likely to be sampled than short spells. Accounting for this sample bias requires knowledge of the exact starting time of each duration spell. Unfortunately, it is common in economic duration data to have coarse measures for starting times, complicating the resolution of this sampling bias. This paper investigates three alternatives for overcoming this coarseness by imputing interval-censored starting times. These three imputation procedures produce estimates that are very close to the true parameter values and are substantially easier and computationally simpler than using the exact likelihood function.

Keywords: Duration Data, Survival Analysis, Stock Sampling, Monte Carlo Simulation

*Department of Economics, The Florida State University. E-mail: mb13m@my.fsu.edu.

†Corresponding author. Department of Economics, The Florida State University. 113 Collegiate Loop, 257 Bellamy Building, Tallahassee, Florida 32306. E-mail: jcanourbina@fsu.edu.

1 Introduction

Duration data that measure the length of time that individuals spend in a certain state before transition to another state is often at the center of applied studies in economics. For example, the length of a employment spell before an individual moves to another job or before the individual becomes unemployed is crucial to pin down key parameters in equilibrium search models.[1] When working with duration data it is necessary to account for the sampling scheme used to collect these data in order to obtain consistent estimators. A commonly used sampling procedure consists of including individuals in the sample who are currently in the state of interest at certain point in time and follow their duration in the state of interest from that point on. For example we only include individuals that are currently employed in January 2010 and at this point we start following their employment spell. This sampling scheme is known as *stock sampling*.[2]

Figure 1 illustrates the typical stock sampling scheme. On the sampling date, A, we obtain information about the date each sampled individual started their current job, and then we follow all sampled individuals over a fixed period of time. In the picture, individual I_1 started her job spell on date s_1 and finished her job spell on date x_1. As indicated in the picture, the duration of employment of individual I_1 is the sum of two components: (i) the elapsed duration, $e_1 = (A - s_1)$, and (ii) the residual duration, $u_1 = (x_1 - A)$. Next, notice in Figure 1 that individual I_2 has not finished her job spell when we stop following individuals at date Z. In this case, the duration of the employment spell of individual I_2 is *right censored*. The elapsed and residual durations of individual I_2 are $e_2 = (A - s_2)$ and $u_2 = (Z - A)$ respectively. Finally, individual I_3 is not included in the sample since she started and finished her employment spell before the sampling date A. This is because individual I_3 started and finished her employment spell before the sampling date A. This feature of stock sampling is typically referred to as *left truncation* (Kalbfleisch and Prentice, 1980; Wooldridge, 2002), and the bias it generates as *length-biased sampling* because long spells are more likely to be sampled than short spells (Kiefer, 1988; Lancaster, 1990).

In order to obtain consistent estimators with a stock sample, we need to incorporate in the likelihood function the fact that long spells are sampled more often than short spells. Let the duration of employment of individuals I_1 and I_2 be t_1 and t_2, respectively, where $t_i = e_i + u_i$ for $i = 1, 2$. The likelihood function for this sample of two individuals is:

$$L = \left(\frac{f(t_1)}{1 - F(e_1)}\right) \times \left(\frac{1 - F(t_2)}{1 - F(e_2)}\right) \tag{1}$$

where f is the p.d.f. and F is the c.d.f. of random variable T representing duration spells. The denominator of each term in the likelihood is $P(t_i > e_i) = 1 - F(e_i)$ and corrects for the

[1] The duration of job spells provide valuable information to identify parameters such as the arrival rate of wage offers while employed and the exogenous destruction rate.

[2] Wooldridge (2002) refers to this scheme as *stock sampling* and this paper follows this convention. However, other authors use different denominations. For example, Kalbfleisch and Prentice (1980) refer to this sampling scheme as *delayed entry* and Lancaster (1990) refers to it as *observation over a fixed interval* (see chap. 8, sect. 3.1). Also, Lancaster (1990) and Murphy (1996) define stock sampling a scheme in which only the elapsed duration of the individuals in the sample is observed, but there is no follow up of the individual after the sampling date (see Lancaster, 1990, chap. 8, sect. 3.3).

fact that long spells are sampled more often in stock sampling, and that short spells, such as t_3, are not included in the sample.

Figure 2 illustrates the stock-sampling scheme analyzed on this paper. Once again, let A be the stock-sampling date and s_i be the spell's starting date. Now, suppose that when we ask sampled individuals about the starting date of their job we record the starting date in two formats depending on when the job started. If the job started: (i) during the previous calendar year then we record the exact starting date of the job, (ii) before the previous calendar year then we only record the year when the job started. In Figure 2, the previous calendar year is marked at time B. For individuals that started their job after B, such as individual I_2, the starting time s_2 is observed exactly, and so their elapsed duration $e_2 = A - s_2$ is also observed exactly. For individuals that started their job before B, such as I_1, the starting time s_1 is not observed, only the interval $[S_1^L, S_1^R]$ containing s_1 is observed, and so the elapsed duration e_1 is only known to be contained in the interval $[E_1^L, E_1^R$, where $E_1^R = A - S_1^L$ and $E_1^L = A - S_1^R$. The contribution to the likelihood of individual I_2 is $f(t_2)/[1 - F(e_2)]$, similar to the first term in equation (1). However, the contribution to the likelihood of individual I_1 is complicated by the fact that we do not know the elapsed duration e_1. As a result we cannot correct the bias introduced by stock sampling, that is, since we do not know e_1 we do not know the term $[1 - F(e_1)]$ in the likelihood.[3]

The goal of this paper is to explore alternative methods for overcoming the coarseness of information about e_2. We perform a Monte Carlo analysis to gauge the properties of the estimators. The interval-censored elapsed duration is imputed using: (a) the lower bound of the interval containing the elapsed duration, (b) the upper bound of the interval, and (c) the midpoint of the interval. The Monte Carlo analysis indicates that the three methods produce estimates that are very close to the true parameter values, but using the midpoint or the upper bound of the interval tend to perform better than using the lower bound. We consider specifying the exact likelihood function under interval-censored elapsed durations, however our findings suggest that the alternative of imputing the missing elapsed duration have very good performance and is substantially simpler.

Section 2 provides some examples of stock sampling with interval-censored elapsed durations. Section 3 discusses the exact likelihood. Sections 4 and 5 discuss the imputation and simulation procedures. Section 6 presents the results and Section 7 presents an empirical application of our imputation procedure and concludes.

2 Examples of Stock Sampling with Interval-Censored Starting Times

In practice, the sampling scheme depicted in Figure 2 occurs when collecting job duration data from surveys that are implemented as rotating panels. In these surveys, job duration

[3]Note that the problem arises because the elapsed duration is interval censored. If the residual duration were interval censored we would not have any problem. Recall that the full spell of an individual is the sum of the residual and the elapsed duration, that is $t_i = u_i + e_i$. If the elapsed duration were observed exactly, but the residual duration u_i were interval censored, i.e. we only know that $u_i \in [U_i^L, U_i^R]$, the full spell would be interval censored, e.g. $t_i \in [L_i, R_i]$, where $L_i = U_i^L + e_i$ and $R_i = U_i^R + e_i$. In this case, the contribution to the likelihood would be: $[F(R_i) - F(L_i)]/[1 - F(e_i)]$, which can be easily implemented. However, when the elapsed duration is interval censored we cannot find a term such as $[1 - F(e_i)]$ to correct for the bias introduced by stock sampling, because e_i is only known to be within some interval.

data are mostly obtained from the stock of individuals who are already employed at the time of the first interview and the starting time of the job is obtained as retrospective information. We often only observe a coarse measure of the starting time for individuals that started long time before the first interview. For instance, the National Survey of Occupation and Employment from Mexico (ENOE) is a rotating panel in which individuals are visited five times over the course of a year. Individuals that are employed at the time of the first visit provide the exact starting date of their job only if it began in the previous calendar year, otherwise they simply state the year the job began. The Monthly Employment Survey from Brazil (PME) follows a similar structure, visiting individuals eight times over the course of a year. In this survey, individuals provide the exact starting date in months if their current job started within the last two years, otherwise they only provide the number of years they have been employed in their current job.

Table I shows that a significant percentage of individuals in both the ENOE and PME do not provide the exact date their current job began. Given the frequency of spells with interval-censored starting times, they cannot be ignored. To deal with interval-censored starting times we have two alternatives: (i) drop all observations with interval-censored starting times or (ii) specify the exact likelihood. The first alternative is not ideal as it can introduce a selection sample problem because from the sampled duration spells we are systematically dropping those with longer duration. In Section 7, we provide an empirical application that illustrates the problems of dropping duration spells with interval-censored elapsed duration. The second alternative is explained in the next section.

3 Exact Likelihood

Consider a stock-sampling scheme where A represents the stock sampling date.[4] Let T be a random variable that represents the duration of some event and t a realization of T. The density of T is given by $f(t|x;\theta)$ where x is a set of time-invariant covariates and θ is the vector of parameters of interest characterizing the duration model. Let S be a random variable representing the starting time of the event of interest and s a realization of S with density $k(s|x;\eta)$, where η is the vector of parameters characterizing the distribution of S. If starting times are independent of the duration variable conditional on covariates x then the joint density of T and S is given by $g(t,s|x;\theta,\eta) = f(t|x;\theta)k(s|x;\eta)$.

Consider a duration spell where the starting time is not observed exactly but only the interval $[S^L, S^R]$ containing s is observed. The residual duration U is defined as $U = T - A + S$. Then, using the change of variable technique, it is straightforward to show that the density of the residual duration U is given by $h(u|x;\theta,\eta) = \int_{S^L}^{S^R} f(u + A - s|x;\theta)k(s|x;\eta)ds$.

Next, to account for stock sampling, recall that a duration spell t is observed if and only if $t > A - S$ so the probability that a spell with interval-censored starting time is included in the sample is given by $\Pr\{T > A - S|x\} = \int_{S^L}^{S^R}[1 - F(A - s|x;\theta)]k(s|x;\eta)ds$, where $F(\cdot|\cdot)$ is the c.d.f. of T.

Finally, to account for right censoring suppose that after the sampling date individuals in the stock sample are only followed during a fixed interval of time, C. If $U > C$ the

[4]The discussion in this section follows the same pattern of exercise 20.8 of Wooldridge (2002).

spell will be right-censored and the probability that the spell is right-censored is given by $\Pr\{U > C|x\} = 1 - \int_0^C h(u|x;\theta,\eta)du$.

Therefore contribution to the likelihood of a spell with interval-censored starting time is:

$$L_i(\theta,\eta|x_i) = \frac{h(u_i|x_i;\theta,\eta)^{d_i}\left[1 - \int_0^C h(u|x;\theta,\eta)du\right]^{(1-d_i)}}{\int_{S^L}^{S^R}[1 - F(A-s|x;\theta)]k(s|x;\eta)ds}, \tag{2}$$

where d_i is an indicator equal to 1 for completed spells and 0 for right-censored spells.

Hence, with knowledge of $k(s|x;\eta)$ it is possible to estimate the vector of parameters (θ,η). However, estimation is not trivial. In Appendix A we present the case of where duration spells follow a Weibull distribution and starting times follow a Uniform distribution over the interval that contains them. We attempted to maximize this likelihood function however, even in this simple case we encountered frequent occurrences where portions of the likelihood function would evaluate to 0, which made solving the optimization problem impossible. In the following section we show that a simple imputation procedure produces estimate of θ that are comparable to estimates of θ when the exact start time is know.

4 Imputing Starting Times

As in the previous section, let T be a random variable with density $f(t|x;\theta)$ representing the duration of the event of interest, where x is a set of time-invariant covariates and θ is the vector of parameters of interest characterizing the duration model. A realization of T from a stock-sampling scheme is a duration spell $t = e + u$ where e and u are the realizations of the elapsed and residual duration, respectively.

Focusing on spells with interval-censored elapsed duration let $\hat{e}$ be the imputed elapsed duration. We explore three imputation methods: (a) the lower bound $\hat{e} = E^L$, (b) the upper bound $\hat{e} = E^R$, and (c) the midpoint $\hat{e} = (1/2) \times (E^L + E^R)$. Next, suppose that, after the stock-sampling date individuals are only followed during a fixed interval of time, C. If $u > C$ the spell is right-censored. The contribution to the likelihood of a spell with imputed elapsed duration is given by:

$$L_i(\theta|x_i) = \frac{f(\hat{t}_i|x_i;\theta)^{d_i}\left[1 - F(\hat{e}_i + C|x_i;\theta)\right]^{(1-d_i)}}{1 - F(\hat{e}_i|x_i;\theta)} \tag{3}$$

where d_i is an indicator equal to 1 for completed spells and 0 for right-censored spells, $\hat{e}_i$ is the imputed elapsed duration, and $\hat{t}_i = u_i + \hat{e}_i$ is the imputed duration.

5 Data Simulation

We simulate 100 data sets of $1,000$ observations using a Weibull-gamma mixture with a hazard function given by:

$$\lambda(t|x,\nu) = \mu\alpha t^{\alpha-1}\nu, \tag{4}$$

where α is the measure of duration dependence; $\mu = \exp\{\beta_0 + \beta_1 x\}$ and x is a vector of time-invariant covariates to account for observed heterogeneity; and the parameter ν represents unobserved heterogeneity, which is assumed to follow a Gamma distribution with $E(\nu) = 1$ and $V(\nu) = 1/\delta$. This implies that small(large) values of δ imply that a large(small) portion of the variation in the duration variable is due to unobserved heterogeneity.[5] To generate a stock sample, the duration data are simulated as a *renewal process* as described in Lancaster (1990) (see ch. 5 sec. 3) and its adaptation to our simulation is explained in Appendix B.

We choose six different parameter sets for the data generating process to account for every combination of three cases of duration dependence and two cases of unobserved heterogeneity. For duration dependence we use $\alpha = 0.5$ for negative, $\alpha = 1.5$ for positive, and $\alpha = 1$ for no duration dependence. For unobserved heterogeneity we use $1/\delta = 1/10000$ for no unobserved heterogeneity and $1/\delta = 1$ for data with unobserved heterogeneity. The parameter β_0 is chosen such that the mean duration from the parameter sets without unobserved heterogeneity matches the average duration in the ENOE (approximately 19 months), and $\beta_1 = 1$.

Only one covariate x is considered and it is held fixed across all simulated samples and parameter sets. This covariate is drawn from a $\mathrm{N}(\mu_x, \sigma_x^2)$. The mean of this distribution is set to $\mu_x = 0$, and its variance is set to $\sigma_x^2 = 0.25$. The choice of variance follows Baker and Melino (2000): σ_x^2 is chosen so that the R^2 from a regression of the simulated $\ln(t)$ on the simulated x is similar to the R^2 of a similar regression using the duration data and a set of covariates from the ENOE.[6]

6 Simulation Results

The results from model estimations with imputed elapsed duration are provided in Table II. Regardless of the parameter set all models are estimated using a Weibull-gamma mixture. Table II is divided in two blocks. The top block considers parameterizations with no unobserved heterogeneity and the bottom block with unobserved heterogeneity. Subsequent rows display the average of the parameter estimates (and its standard deviation in parenthesis). While all three choices for imputation produce estimates that are close to the true parameter values both the upper bound (E^R) and midpoint (E^M) tend to preform better than the lower bound E^L. The results in Table II for parameter set 4 representing negative duration dependence, present the higher challenge to recover the true parameters, however in most cases, the average estimate is within one standard deviation of the true parameter.

Tables III and IV present the mean squared error (MSE) and the mean absolute deviation (MAD) for the simulated data. The three imputation methods produce similar MSE and MAD for the six parameter sets. The MSE for E^R is usually the smallest all parameter sets. The MAD shows a similar pattern. In all cases the MSE and MAD of E^R is very close to these measures for E^M and E^L.

[5] When δ grows indefinitely, the distribution of T converges to a Weibull distribution without unobserved heterogeneity (see Cameron and Trivedi, 2005).

[6] The duration data from the ENOE contains some interval-censored spells. In order to fit this regression, these spells are imputed as $\widetilde{t}_i = L_i + u \cdot (R_i - L_i)$, where u is drawn from a uniform distribution in $[0, 1]$ and $(L_i, R_i]$ is the interval containing the actual duration.

7 Empirical Application

In this section we present an empirical application using the ENOE survey from Mexico. Motivated by the work of Cano-Urbina (2015) we use job duration in the informal sector before making a transition to the formal sector.[7]

We estimate a Weibull-gamma mixture using two alternatives: (i) dropping job spells with interval-censored elapsed duration, and (ii) using all job spells and imputing interval-censored elapsed durations with the midpoint of the interval E^M.[8]

We use years of education and age for the covariates. The results in Table V clearly suggests different estimates for the duration model. When we drop job spells with interval-censored elapsed duration the estimates suggest that unobserved heterogeneity is very important in generating variation in the data while including all observations and imputation suggest a modest role for unobserved heterogeneity. Dropping these observations also suggest that there is no duration dependence as the estimate of α contains 1 in the 95% confidence interval, while including all observations and imputation suggest that there is negative duration dependence. Equally important are the differences in the estimated effects of education and age. Finally, note that using all the sample and imputation provides estimates with smaller standard errors as a result of having a larger sample size.

Figure 3 presents the plots of the unconditional hazards resulting from each of these estimation samples. While both unconditional hazard functions present a similar shape, dropping observations with interval-censored elapsed duration suggest a fastest drop in the likelihood of making a transition from an informal-sector job to a formal-sector job. This result is due to the higher role of unobserved heterogeneity suggested by the estimates with this sample.

8 Conclusion

In this paper we study how different imputation choices for duration data with interval-censored start times affects estimation results. This censoring mechanism occurs when obtaining duration data from a stock of individuals who are employed at the time of the interview and the starting time of employment is obtained as retrospective information. All three proposed choices for imputing duration (left endpoint, midpoint, or right endpoint) are easily implemented and produce point estimates that are close to the true parameter values.

[7]The informal sector is composed of jobs that do not comply with some or all of the labor regulations. The formal sector is composed by jobs that comply with all labor regulations.

[8]The ENOE also contains observations in which the residual duration is only known to be contained in an interval. In these cases the duration of the spell t_i is included within the interval $[L_i, R_i]$. The contribution of these observations to the likelihood function is $[F(R_i) - F(L_i)]/[1 - F(e_i)]$. See Footnote 3 for further details.

Figure 1: Typical Stock Sampling

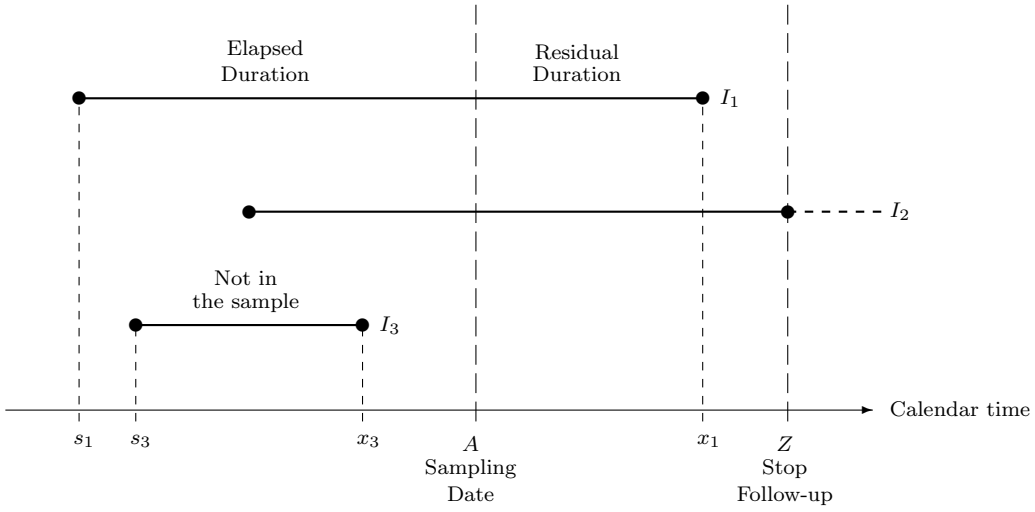


Figure 2: Stock Sampling with Interval-Censored Starting Time

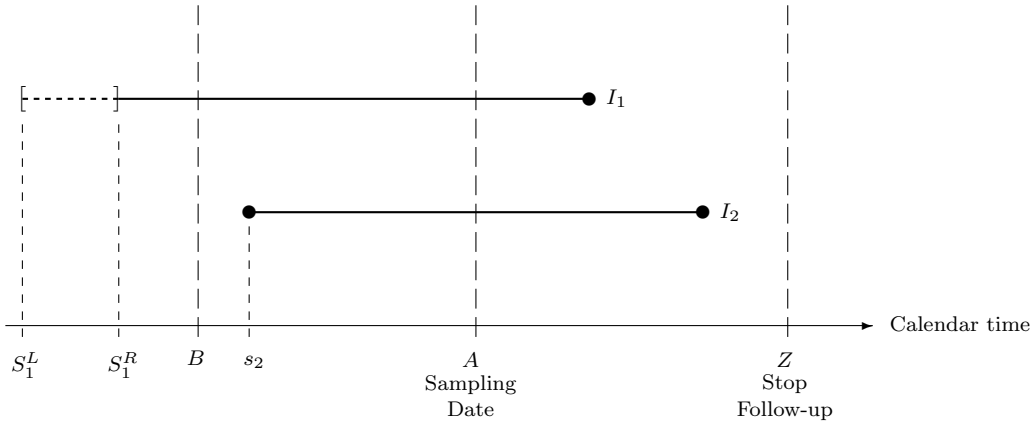


Figure 3: Webibull-Gamma Unconditional Hazard

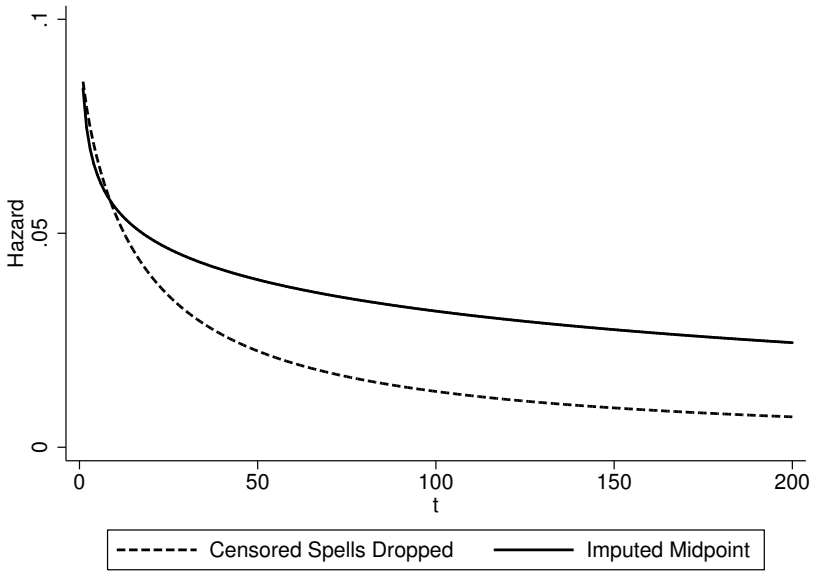


Table I: Elapsed Duration in the ENOE and PME

Elapsed Duration	ENOE		PME	
	Num. Obs.	%	Num. Obs.	%
Exact	20,499	27.37	12,875	39.51
Interval-Censored	54,399	72.63	19,713	60.49
Total	74,898		32,588	

Source: INEGI for ENOE, IBGE for PME. Data from the ENOE is for the first quarter of 2010. Data from PME is for January of 2012. The table only includes paid employees.

Table II: Imputed Elapsed Duration

	Parameter Set 1				Parameter Set 2				Parameter Set 3			
	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$
	-1.6	1	0.5	1/10000	-4.2	1	1	1/10000	-6.6	1	1.5	1/1000
E^L	-1.70	1.11	0.53	0.07	-4.37	1.06	1.04	0.04	-7.02	1.00	1.61	0.05
	(0.44)	(0.30)	(0.09)	(0.10)	(0.41)	(0.18)	(0.09)	(0.06)	(0.50)	(0.18)	(0.11)	(0.07)
E^M	-1.80	1.11	0.55	0.06	-4.36	1.06	1.04	0.04	-6.76	1.00	1.54	0.03
	(0.43)	(0.28)	(0.09)	(0.09)	(0.39)	(0.18)	(0.08)	(0.06)	(0.43)	(0.18)	(0.09)	(0.05)
E^R	-1.88	1.09	0.57	0.05	-4.36	1.06	1.04	0.03	-6.52	0.99	1.48	0.03
	(0.41)	(0.27)	(0.09)	(0.08)	(0.37)	(0.18)	(0.08)	(0.05)	(0.41)	(0.18)	(0.09)	(0.04)
	Parameter Set 4				Parameter Set 5				Parameter Set 6			
	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$
	-1.6	1	0.5	1	-4.2	1	1	1	-6.6	1	1.5	1
E^L	-0.91	1.86	0.77	0.88	-4.03	1.17	1.18	0.83	-6.39	1.06	1.60	0.79
	(0.55)	(0.97)	(0.22)	(0.33)	(0.67)	(0.43)	(0.19)	(0.23)	(0.81)	(0.35)	(0.21)	(0.22)
E^M	-0.90	1.65	0.71	0.72	-4.02	1.08	1.16	0.70	-6.37	1.05	1.58	0.69
	(0.51)	(0.86)	(0.19)	(0.26)	(0.60)	(0.39)	(0.16)	(0.17)	(0.76)	(0.34)	(0.19)	(0.18)
E^R	-0.87	1.46	0.66	0.58	-3.95	1.00	1.12	0.59	-6.32	1.02	1.56	0.59
	(0.48)	(0.77)	(0.16)	(0.22)	(0.55)	(0.36)	(0.14)	(0.13)	(0.70)	(0.32)	(0.17)	(0.14)

Table III: Mean Squared Error

	Parameter Set 1				Parameter Set 2				Parameter Set 3			
	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$
E^L	0.0134	0.0292	0.0029	0.0017	0.1637	0.0333	0.0105	0.0079	0.4133	0.0294	0.0245	0.0068
E^M	0.0125	0.0260	0.0024	0.0013	0.1445	0.0314	0.0089	0.0052	0.2363	0.0311	0.0127	0.0045
E^R	0.0117	0.0236	0.0020	0.0010	0.1289	0.0301	0.0076	0.0035	0.2039	0.0340	0.0109	0.0044
	Parameter Set 4				Parameter Set 5				Parameter Set 6			
	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$
E^L	0.7859	1.6697	0.1234	0.1218	0.4715	0.2074	0.0663	0.0792	0.6988	0.1238	0.0516	0.0927
E^M	0.7426	1.1574	0.0803	0.1473	0.3923	0.1567	0.0497	0.1156	0.6211	0.1169	0.0420	0.1287
E^R	0.7640	0.7975	0.0511	0.2199	0.3625	0.1267	0.0341	0.1877	0.5662	0.1041	0.0322	0.1860

Table IV: Mean Absolute Deviation

	Parameter Set 1				Parameter Set 2				Parameter Set 3			
	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$
E^L	0.3541	0.2385	0.0719	0.0653	0.3534	0.1501	0.0782	0.0398	0.5134	0.1404	0.1212	0.0479
E^M	0.3638	0.2318	0.0749	0.0607	0.3396	0.1512	0.0743	0.0361	0.3611	0.1393	0.0785	0.0318
E^R	0.3767	0.2202	0.0786	0.0489	0.3286	0.1509	0.0709	0.0327	0.3258	0.1393	0.0724	0.0279
	Parameter Set 4				Parameter Set 5				Parameter Set 6			
	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$	β_0	β_1	α	$1/\delta$
E^L	0.7357	1.0128	0.2886	0.2942	0.5574	0.3675	0.2056	0.2321	0.6760	0.2696	0.1788	0.2668
E^M	0.7219	0.8095	0.2314	0.3296	0.5011	0.3204	0.1808	0.3070	0.6384	0.262	0.1629	0.3253
E^R	0.7456	0.6452	0.1817	0.4196	0.4822	0.2893	0.1496	0.4147	0.6123	0.2519	0.1438	0.4090

Table V: Estimation Example from ENOE

	Dropping Observations with Interval-Censored Elapsed Duration	Using All Sample and Imputed Elapsed Duration
Education	0.1402***	0.0825***
	(0.0231)	(0.0114)
Age	0.0236**	0.0162***
	(0.0105)	(0.0059)
Constant	-3.9862***	-3.2865***
	(0.4379)	(0.2156)
α	0.9677***	0.8388***
	(0.1046)	(0.0345)
δ	1.6184**	18.2374*
	(0.7238)	(10.4474)
Log-Likelihood	-3,674.14	-7,002.43
Observations	4,121	8,336

Notes: Standard errors are in parenthesis. The sample only includes males ages 16-30 with less than 12 years of education. We use all ENOE samples from the first quarter of 2005 to the fourth quarter of 2010. *Significant at 10%, **Significant at 5%, ***Significant at 1%.

Appendix

A Weibull Distribution with Uniformly Distributed Starting Times

Consider the case in which the random variable T measuring duration follows a Weibull distribution but we have some duration spells with interval-censored starting times, in which case the starting time is only known to be contained in the interval $[S^L, S^R]$. Suppose a stock sample is obtained at the sampling date A and individuals are only followed for a fixed period of time C after the sampling date. Duration spells that have not finished by date $A+C$ are right censored.

Then, in this case, the contribution to the likelihood of a duration spell with interval-censored starting time is given by equation (2), which is reproduced here for convenience:

$$L_i(\theta,\eta|x_i) = \frac{h(u_i|x_i;\theta,\eta)^{d_i}\big[\Pr\{U>C|x\}\big]^{(1-d_i)}}{\Pr\{T>A-S|x\}} = \frac{h(u_i|x_i;\theta,\eta)^{d_i}\big[1-\int_0^C h(u|x;\theta,\eta)du\big]^{(1-d_i)}}{\int_{S^L}^{S^R}[1-F(A-s|x;\theta)]k(s|x;\eta)ds},$$

For the simple case of a Weibull distribution with uniformly distributed starting times the components of the likelihood function are given by:

$$h(u|x;\theta,\eta) = \frac{1}{S^R-S^L}\Big[\exp(-\mu(u+A-S^R)^\alpha) - \exp(-\mu(u+A-S^L)^\alpha)\Big]$$

$$\Pr\{U>C|x\} = 1 - \frac{\mu^{-\alpha}}{S^R-S^L}\Big[\frac{1}{\alpha}\gamma\Big(\frac{1}{\alpha},\mu(u+A-S^R)^\alpha\Big) - \frac{1}{\alpha}\gamma\Big(\frac{1}{\alpha},\mu(u+A-S^L)^\alpha\Big)\Big]$$

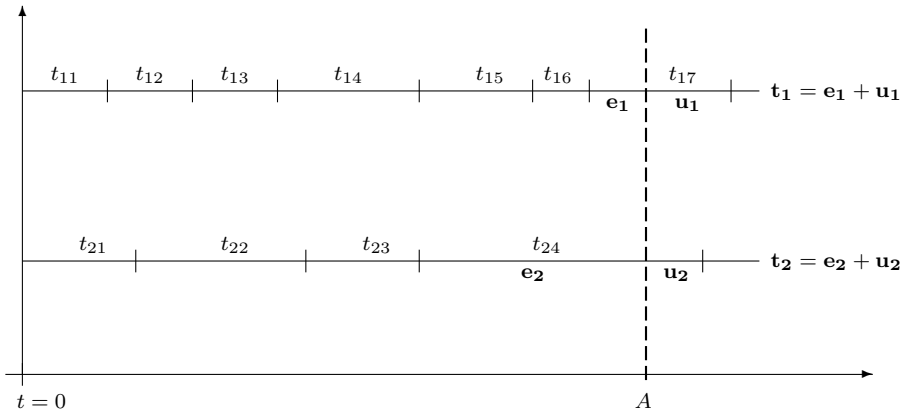
$$\Pr\{T>A-S|x\} = \frac{-\mu^{-\alpha}}{S^R-S^L}\Big[\frac{1}{\alpha}\gamma\Big(\frac{1}{\alpha},\mu(A-S^R)^\alpha\Big) - \frac{1}{\alpha}\gamma\Big(\frac{1}{\alpha},\mu(A-S^L)^\alpha\Big)\Big]$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function, and α and $\mu = \exp(\beta_0+\beta_1 x)$ are the duration dependence and observed heterogeneity parameters of the Weibull distribution, respectively. We attempted to maximize the likelihood function however, even in this simple case we encountered frequent occurrences where portions of the likelihood function would evaluate to 0, which made solving the optimization problem impossible. Specifically, this occurs when calculating the difference between incomplete gamma functions, whenever the second argument is sufficiently large the difference is calculated as 0. In this simple case, we are ignoring the unobserved heterogeneity component. However, if we were to include unobserved heterogeneity, the problem would only get more complicated.

B Renewal Process for Simulation of a Stock Sample

To generate a stock sample, the duration data are simulated as the *renewal process* depicted in Figure 4. In Figure 4, a job spell is denoted t_{ij} where the subindex i identifies a type of individuals with observable and unobservable characteristics (x_i,ν_i), and the subindex j identifies a particular member of type i. At each point in time, only one member of type i is employed and when this member exits is replaced by another member of the same type i that has the same characteristics (x_i,ν_i). Hence, starting at time $t=0$ for each type i we

Figure 4: Simulation as a Renewal Process



have a sequence of job spells $\{t_{i1}, t_{i2}, t_{i3}, \ldots\}$. The job spells for each type are realizations of a random variable T_{ij} with cumulative distribution function $F(t|x_i, \nu_i)$. For the particular case of the Weibull-gamma mixture we use in our simulations the c.d.f. is given by:

$$F(t|x_i, \nu_i) = 1 - \exp\{-(\beta_0 + \beta_1 x_i)t^{\alpha}\nu_i\}. \tag{5}$$

To generate a stock sample, we fix a time which is denoted $t = A$ in Figure 4 and represents the stock-sampling date. Figure 4 presents two hypothetical types. At the stock sampling date, A, the first type is in its seventh generation, t_{17}, while the second type is in its fourth, t_{24}. And so for these two types we include in the stock sample the job spells $\mathbf{t_1} = t_{17}$ and $\mathbf{t_2} = t_{24}$. For these job spells in the sample the elapsed duration is $\mathbf{e_i}$ and the residual duration is $\mathbf{u_i}$ for $i = 1, 2$.

For each sequence of generations i, the stock sample data generation uses the following steps:

1. Draw x_i from a $\mathrm{N}(\mu_x, \sigma_x^2)$, to obtain $\mu_i = \exp\{\beta_0 + \beta_1 x_i\}$.
2. Draw ν_i from a gamma distribution with mean 1 and variance $1/\delta$.
3. Start with $j = 1$, compute the duration spell t_{ij} as follows:
   (a) Draw Y from a Uniform[0,1].
   (b) Compute t_{ij} using the inverse of the cumulative distribution function (5) as follows:
   $$t_{ij} = F^{-1}(Y|x_i, \nu_i) = \left[-\frac{\ln(1-Y)}{\mu_i \nu_i}\right]^{1/\alpha}.$$
4. Compute the cumulative duration for the sequence of spells up to generation j:
   $$\overline{\overline{T}}_{ij} = \sum_{k=1}^{j} t_{ik}.$$

5. If $\overline{\overline{T}}_{ij} > A$, then stop and go to 6, otherwise go back to 3, increase j by 1, and repeat process.

6. Once the stock sampling date is reached, compute the residual, the elapsed, and the complete duration, respectively, as:

$$\mathbf{u_i} = \overline{\overline{T}}_{ij} - A$$
$$\mathbf{e_i} = \overline{\overline{T}}_{ij} - u_i^*$$
$$\mathbf{t_i} = e_i^* + u_i^*$$

This process is repeated for $i = 1, 2, \ldots,$ 1,000; that is, the sample size is $N =$ 1,000. Notice that the draw of the observed and unobserved heterogeneity components, x_i and ν_i respectively, is done only once for each sequence of generations of the population and stays constant during the repeated draws from the uniform distribution Y.

References

Baker, M. and Melino, A. (2000). Duration dependence and nonparametric heterogeneity: A monte carlo study. *Journal of Econometrics* 96: 357 – 393.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics : Methods and Applications*. Cambridge ; New York: Cambridge University Press.

Cano-Urbina, J. (2015). The role of the informal sector in the early careers of less-educated workers. *Journal of Development Economics* 112: 33 – 55.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of Economic Literature* 26: 646–679.

Lancaster, T. (1990). *The econometric analysis of transition data*. New York: Cambridge University Press.

Murphy, A. (1996). A piecewise-constant hazard-rate model for the duration of unemployment in single-interview samples of the stock of unemployed. *Economics Letters* 51: 177 – 183.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass.: The MIT Press.