I'm an adult, now what?
Attrition and time-variant unobserved heterogeneity in an education-labor dynamic discrete choice model

José Alfonso Muñoz-Alvarado

Toulouse School of Economics

August 16, 2021

Motivation 1

- DCDP models: incorporation of time-invariant unobserved heterogeneity.
- Acknowledge the heterogeneity with respect to the ability to accumulate skills at school or at work.
- Unobserved heterogeneity: cognitive skills (intelligence) and non-cognitive skills (motivation or locus of control, or personality traits).
- **But**, these components evolve with age and can be altered by education (Cunha, Heckman, and Schennach, 2010)

Motivation 1

- Hu & Shum (2012); Hu, Shum, Tan & Xiao (2017): Identification of time-variant stochastic process.
- Todd & Zhang (2020): estimation modeling personality traits as time-variant proxy for the unobserved heterogeneity.
    - No comparison between time-variance and invariant counterfactuals.
- **This paper:**
    - Estimate time-variant without proxy, using only choice data.
    - Test the fit of the data between a model with time-variant vs time-invariant unobserved heterogeneity.

Motivation 2

- DCDP models use panel data $\longrightarrow$ attrition.
- **But,** not considered in the literature:
  - NLSY79: 33% sample loses between 1979 and 2014.
  - HILDA: 40% between the 1st and 16th wave.
  - Génération 98: 67% between 1st and 4th wave.
- Observations leaving sample drawn randomly $\longrightarrow$ no problem.
- Reasons for leaving the sample are endogenous $\longrightarrow$ selection bias.
- **This paper:**
  - Incorporates a selection mechanism to test and correct for attrition bias.

This paper

Research questions

- How time-varying unobserved heterogeneity and endogenous attrition alter structural estimation and counterfactuals?

- How education affect labor outcomes through the indirect effect of modifying unobserved skills with age?

Model: Dynamic discrete choice model à la Keane & Wolpin (1997)

- Decisions: education, labor, staying at home.
- Attrition decision.
- Time-variant unobserved heterogeneity.

Data: Génération 98

- 16,717 observations in 4 waves over 10 years.
- 67% attrition between 1st and 4th wave.

Model

- Age 16 to 65 (retirement).
- Individuals make a decision at period t, $d_j(t)$, out of 3 alternatives j:
  1. Schooling ($j = 1$):

$$u_1(t) = \beta_1 X_s' + \beta_2 \mathbb{1}\{college\} + \beta_3 \mathbb{1}\{graduate\} + \theta_s(k(t)) + \varepsilon_s(t)$$

  2. Work ($j = 2$):

$$u_2(t) = w(t) + \varepsilon_w(t)$$

$$\log(w(t)) = \gamma_1 X_w' + \gamma_2 g(t) + \gamma_3 expe(t) + \gamma_4 expe(t)^2 + \theta_w(k(t)) + \eta(t)$$

     $g(t)$ years of schooling, $expe(t)$ work experience and $\eta(t) \sim \mathcal{N}(0, \sigma^2)$

  3. Staying at home ($j = 3$):

$$u_3(t) = \delta_1 age + \delta_2 age^2 + \theta_h(k(t)) + \varepsilon_h(t)$$

Attrition

- Attrition decision happens at 3 different ages for each i.

$$a(t) = \phi X_a^{'} + \theta_a(k(t)) + \varepsilon_a(t)$$

$\varepsilon_a(t)$ is an i.i.d type I extreme value distribution.

- Under this specification: *attrition conditionally random*.
- $X_a^{'}$: couple, kids, urban zone, distance between middle school and current place.

Law of Motion

- Time-varying component in state space: $g(t)$, $expe(t)$ and $k(t)$.
- Years of education and work experience:

$$\begin{aligned} g(t+1) &= g(t)+1, \quad \text{if } d_1(t)=1 \\ expe(t+1) &= expe(t)+1, \quad \text{if } d_2(t)=1 \end{aligned}$$

- Types:

$$M(t) = (1-p(t))\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + p(t)\begin{bmatrix} q_{k=1}(t) & q_{k=1}(t) \\ q_{k=2}(t) & q_{k=2}(t) \end{bmatrix}$$

where

$p(t)$ prob of changing types. Depends on age at t.

$q_k(t)$ prob of being type k. Depends on initial conditions at age 16, $g(t)$ and $expe(t)$.

Equations

Solution, Likelihood & Identification

Solution: Solution

- Backward induction from age $T = 65$.
- Value functions include the probability of changing types.

Likelihood: Likelihood

- Depends on event on t:
  - Attrition decision is present or not & if i leaves or not the sample.
- Likelihood of i is the product over all t and over the unobserved state $k(t)$, with initial condition variables at age 16.

Identification: Identification

- Time-variance unobserved heterogeneity: Hu, Shum, Tan & Xiao (2017).
- Attrition: exogenous restrictions and Little and Robin (2019).

Data

- *Génération 98*: *i*s entering labor market for 1st time after education in 1998.
- 10 years of data collected in 4 waves.
- Educational information since high school.

Male attrition

Wave	N	% Attrition	% Cumulative
1998-2001	16,717		
2002-2003	10,966	34.40%	34.40%
2004-2005	8,026	26.81%	48.01%
2006-2008	5,439	32.23%	67.46%

Individuals' characteristics

Results

- Preliminary results ▸ Results
- Model fit not good.

Next steps

- Time-invariant:
  - Adjust to fit better.
  - Testing between attrition and no attrition models.
- Time-varying estimation results:
  - Model fit.
  - Testing between attrition and no attrition models.
  - Testing between time-varying and invariant models.
- Counterfactual analysis.

Discussion

- Time-variance parametric specification:
  - Numerical issues as value functions are too big.
  - Potential collinearity from age variable.
  - Use of another, less restrictive specification?
  - Use time-variant variables to extract the time variation:
    - Number of children, type of education, skill-job missmatch.
    - Last two are choice specific.

Thanks!

Law of Motion

- Types:

$$M(t) = (1 - p(t)) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + p(t) \begin{bmatrix} q_{k=1}(t) & q_{k=1}(t) \\ q_{k=2}(t) & q_{k=2}(t) \end{bmatrix}$$

where

$p(t)$ prob of changing types:

$$p(t) = \frac{1}{1 + \exp(\varphi_1 + \varphi_2(\mathrm{t} - 16) + \varphi_3(\mathrm{t} - 16)^2)}$$

$q_k(t)$ prob of being type k:

$$q_k(t) = \frac{\exp(v_k(t))}{\sum_{k=1}^{K=2} \left(\exp(v_k(t))\right)}$$

$$v_k(t) = \varphi_{4,k} + \varphi_{5,k} X_{16} + \varphi_{6,k} g(t) + \varphi_{7,k} expe(t)$$

Back

Solution

- The model is solved by using backward induction.
- At T, the value function of i of type k is:

$$V_k(s(T), \Omega) = \max_{\{d_j(T)\}} \left[\sum_{j=1}^{J} u_j(T) d_j(T) \,\middle|\, s(T) \right]$$

- At each $t < T$:

$$V_k(s(t), \Omega) = \max_{\{d_j(t)\}} \mathbb{E}\left[\sum_{\tau=t}^{T} \delta^{\tau-t} \sum_{j=1}^{J} u_j(t) d_j(t) \,\middle|\, s(t) \right]$$

- Bellman equation form:

$$V_{k,j}(s(t), \Omega) = \tilde{u}_j(t) + \delta\, \mathbb{E}_{k,\varepsilon_j(t)}\left[V(s(t+1), \Omega) \,\middle|\, s(t), d_j(t) \right]$$

Solution

- $\varepsilon_d(t) \sim$ i.i.d.T1EV distribution (location parameter λ and Euler's constant ζ):

$$\mathbb{E}\Big[V(s(t+1),\Omega) \;\Big|\; s(t), d_j(t)\Big] = \mathbb{E}_{k,\varepsilon_j(t)}\left[\max_{d_j(t)} \sum_{j=1}^{J} d_j(t)\Big\{\tilde{V}_{k,j}(s(t),\Omega) + \varepsilon_j(t)\Big\}\right] \tag{1}$$

$$= \mathbb{E}_k\left[\lambda\log\Big(\sum_{j=1}^{3}\exp(\tilde{V}_{k,j}(s(t),\Omega)/\lambda)\Big) + \lambda\zeta\right] \tag{2}$$

$$= \mathbb{E}_k\Big[Z(k(t), s_{-k}(t),\Omega)\Big] \tag{3}$$

- $\mathbb{E}_k$ is the expectation over the time-variant types:

$$\mathbb{E}_k\Big[Z(k(t), s_{-k}(t),\Omega\Big] = \Big(1 - p(t+1)\Big)\Big(Z(k(t+1), s_{-k}(t+1),\Omega\Big) + \\ p(t+1)\sum_{m=1}^{K=2}\Big(q_m(t+1)Z(m(t+1), s_{-m}(t+1),\Omega)\Big) \tag{4}$$

Back

Likelihood

Two possibilities at period t:

1. Attrition decision **is not** present for i_k:

$$L_{i,k,t}(\Omega|d,w,a,s(t)) = Pr\Big(d_{i,j}(t)|\theta_j(k(t)),s(t),\Omega\Big)\times Pr\Big(w_i(t)|\theta_w(k(t)),s(t),\Omega\Big)^{\mathbb{1}\{d_{i,2}(t)=1\}}$$

2. Attrition decision **is** present for i_k:
   - Individual i leaves the sample:

$$L_{i,k,t}(\Omega|d,w,a,s(t)) = Pr\Big(a_i(t)|d_i,w_i,\theta_a(k(t)),s(t),\Omega\Big)$$

   - Individual i leaves **does not** the sample:

$$\begin{aligned} L_{i,k,t}(\Omega|d,w,a,s(t)) =& Pr\Big(d_{i,j}(t)|\theta_j(k(t)),s(t),\Omega\Big)\times \\ & Pr\Big(w_i(t)|\theta_w(k(t)),s(t),\Omega\Big)^{\mathbb{1}\{d_{i,2}(t)=1\}}\times \\ & \Big(1-Pr\Big(a_i(t)|d_i,w_i,\theta_a(k(t)),s(t),\Omega\Big)\Big) \end{aligned}$$

Likelihood

- Likelihood of i is the product over all the periods and the integration over the unobserved state $k(t)$, with $q_k(16)$ as the initial type probability:

$$L_i(\Omega|d,w,a,s(16)) = \sum_{k=1}^{K=2} \Bigg[q_k(16) L_{i,k,16}(\Omega|d,w,a,s(16)) \times \ldots$$

$$\ldots \times \prod_{t=17}^{T} \Bigg((1-p(t)) L_{i,k,t}\big(\Omega|d,w,a,s(t)\big) +$$

$$p(t) \sum_{m=1}^{K=2} q_m(t) L_{i,m,t}\big(\Omega|d,w,a,s(t)\big) \Bigg) \Bigg]$$

- Log likelihood of the sample:

$$\sum_{n=1}^{N} \ln \Big(L_i(\Omega|d,w,a,s(16)) \Big)$$

▸ Back

Identification

- **Time-variant unobserved heterogeneity:** by Hu, Shum, Tan & Xiao (2017).
  1. **Limited feedback**: limit to necessary periods on conditional probabilities.
  2. **Full rank**: joint probability matrix of decision and state space is invertible.
  3. **Distinctive types**: probability of the same decision at the same period and state space is different for different types.
  4. **First order stochastic dominance**: for a fixed state space, the probability of a decision is stochastically increasing in the sense of first-order stochastic increasing in the previous type.

$$
\begin{aligned}
Pr\Big(d_j(t), s_{-k}(t), k(t)\Big| d_j(t-1), s_{-k}(t-1), k(t-1)\Big) = \\
& Pr\Big(d_j(t)\Big| s_{-k}(t), k(t)\Big) \times \\
& Pr\Big(s_{-k}(t)\Big| d_j(t-1), s_{-k}(t-1), k(t-1)\Big) \times \\
& Pr\Big(k(t)\Big| s_{-k}(t), s_{-k}(t-1), k(t-1)\Big)
\end{aligned}
$$

Back

Identification

- **Attrition:**
  - Little & Rubin (2019): when attrition is conditionally random $\longrightarrow$ separate the joint probability of outcomes and attrition:

$$Pr\big(D = d_j(t), W = w(t), A = a(t)|\beta, \gamma, \phi, \theta_j(k(t)), \theta_a(k), s(t)\big) =$$
$$Pr\big(d_j(t)|\beta, \gamma, \theta_j(k(t)), s(t)\big) \times Pr\big(w(t)|\gamma, \theta_w(k(t)), s(t)\big)$$
$$\times Pr\big(a(t)|d, w, \phi, \theta_a(t), s(t)\big)$$

  - Exogenous variables for attrition: kids, couple, distance to hometown.
- Normalization of one of the utility values: non-pecuniary benefits of working.
- Wages observed in data $\longrightarrow$ wage parameters $\longrightarrow$ choices parameters.

Back

Statistics of individuals' characteristics

Table: **Summary statistics**

Variable	**Percentage**
French	95.71
Late junior school	28.37
Both parents French	80.75
One parent French	7.81
None parents French	11.44
N	16,717

Back

Statistics of individuals' characteristics

Variable	Mean	Std. Dev.	Min	Max
Age in 1998	20.977	2.749	16	30
Years of Education in 1998	14.933	2.632	10	21
Middle school and 1st interview	56.313	139.843	0	1,073.948
Middle school and 2nd interview	65.498	148.747	0	1,018.842
Middle school and 3rd interview	71.078	151.739	0	1,176.063
Middle school and 4th interview	79.121	160.500	0	1,048.287

Back

Table: **Summary Statistics of personal characteristics in each attrition decision**

Variable	**1st (2002)**	**2nd (2004)**	**3rd (2006)**	**3rd (2008)**
Lives in couple	23.23%	39.57%	53.04%	65.53%
No kids	93.31%	84.40%	71.92%	52.34%
One Kid	5.33%	11.47%	17.81%	22.80%
Two or more kids	1.36%	4.13%	10.27%	24.86%
Urban zone	81.98%	81.62%	80.93%	79.22 %
N	16,717	10,966	8,026	5,439

Back

Preliminary results: Two types invariant

Table: **Schooling**

	Attrition	**No Attrition**
Age over 18	-2.222	-2.915
	(0.031)	(0.034)
Age over 20	-2.140	-4.167
	(0.017)	(0.027)
Urban	0.255	0.173
	(0.013)	(0.013)
Constant		
Type 1	16.181	4.856
	(0.038)	(0.050)
Type 2	24.137	15.001
	(0.067)	(0.064)
N	16,717	

Back

Preliminary results: Two types invariant

Table: **Working**

	Attrition	**No Attrition**
School years	0.010	0.068
	(0.000)	(0.000)
Experience	0.038	0.063
	(0.000)	(0.001)
Experience$\hat{2}$	-0.001	-0.007
	(0.000)	(0.000)
Urban	0.001	0.008
	(0.000)	(0.001)
Standard error	0.205	0.253
	(0.003)	(0.003)
Constant		
Type 1	2.307	1.600
	(0.002)	(0.004)
Type 2	2.789	1.367
	(0.003)	(0.005)
N	16,717	

Back

Preliminary results: Two types invariant

Table: **Home**

	Attrition	**No Attrition**
Age over 18	-0.259	-0.218
	(0.008)	(0.009)
Age over 20	0.005	0.001
	(0.000)	(0.000)
Constant		
Type 1	15.774	11.376
	(0.117)	(0.134)
Type 2	24.830	16.562
	(0.127)	(0.143)
N	16,717	

Back

Preliminary results: Two types invariant

Table: **Attrition**

	Coeff	**SE**
Age	-0.377	0.055
Age square	0.006	0.001
Couple	-0.071	0.030
One kid	0.138	0.047
2 kids or more	0.175	0.069
Urban	0.156	0.031
Distance MS-Interview	0.009	0.010
Constant		
Type 1	4.620	0.690
Type 2	4.703	0.687
N	16,717	

Back

Model Fit: two types invariant

Figure: Attrition: Schooling

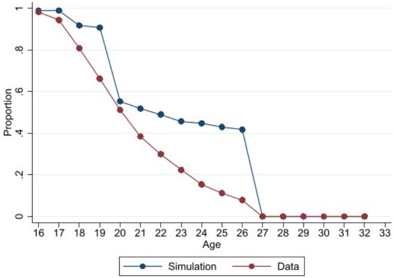
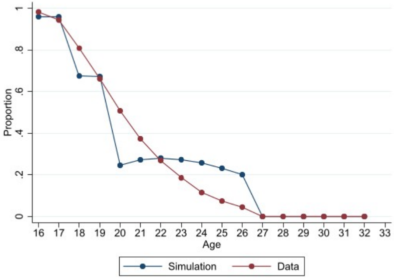


Figure: No Attrition: Schooling



Back

Model Fit: two types invariant

Figure: Attrition: Work

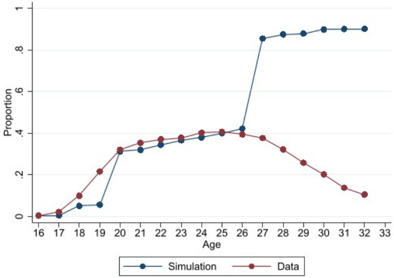
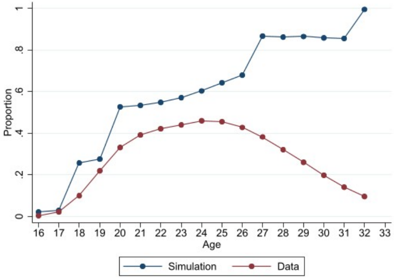


Figure: No Attrition: Work



Back

Model Fit: two types invariant

Figure: Attrition: Home

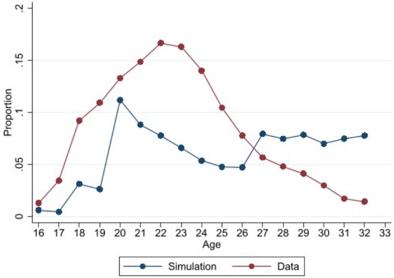
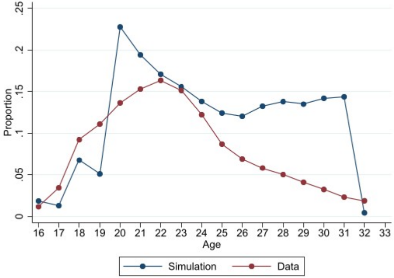


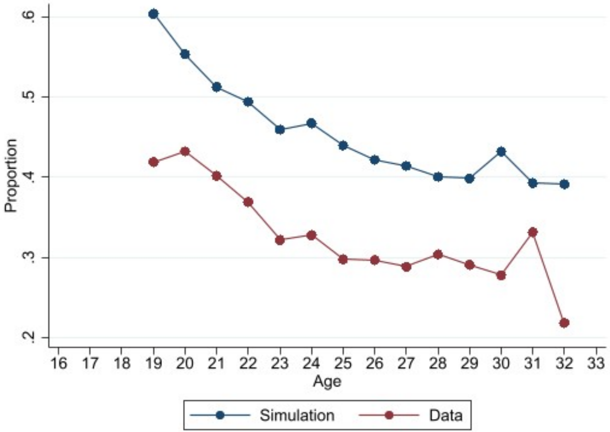
Figure: No Attrition: Home



Back

Model Fit: two types invariant

Figure: Attrition



Back