# Sufficient Statistics for Unobserved Heterogeneity in Structural Dynamic Logit Models

Victor Aguirregabiria* University of Toronto and CEPR

Jiaying Gu* University of Toronto

Yao Luo* University of Toronto

April 9, 2018

**Abstract**

We study the identification and estimation of structural parameters in dynamic panel data logit models where decisions are forward-looking and the joint distribution of unobserved heterogeneity and observable state variables is nonparametric, i.e., fixed-effects model. We consider models with two endogenous state variables: the lagged decision variable, and the time duration in the last choice. This class of models includes as particular cases important economic applications such as models of market entry-exit, occupational choice, machine replacement, inventory and investment decisions, or dynamic demand of differentiated products. The identification of structural parameters requires a sufficient statistic that controls for unobserved heterogeneity not only in current utility but also in the continuation value of the forward-looking decision problem. We obtain the minimal sufficient statistic and prove identification of some structural parameters using a conditional likelihood approach. We apply this estimator to a machine replacement model.

**Keywords:** Panel data discrete choice models; Dynamic structural models; Fixed effects; Unobserved heterogeneity; Structural state dependence; Identification; Sufficient statistic.

**JEL:** C23; C25; C41; C51; C61.

Victor Aguirregabiria. Address: 150 St. George Street. Toronto, ON, M5S 3G7, Canada.
E-mail: `victor.aguirregabiria@utoronto.ca`

Jiaying Gu. Address: 150 St. George Street. Toronto, ON, M5S 3G7, Canada.
E-mail: `jiaying.gu@utoronto.ca`

Yao Luo. Address: 150 St. George Street. Toronto, ON, M5S 3G7, Canada.
E-mail: `yao.luo@utoronto.ca`

*We have benefited from comments and conversations with Bo Honoré, Roger Koenker, Bob Miller, Pedro Mira, Ismael Mourifié, Whitney Newey, Elie Tamer, Yuanyuan Wan, and from participants in seminars at Harvard-MIT, Princeton, Toronto, and at the UCL *Conference on Implementation of Structural Dynamic Models: Methodology and Applications*.

# 1 Introduction

Persistent unobserved heterogeneity is pervasive in the empirical analysis using panel data of individuals, households, or firms. A key econometric issue in dynamic panel data models of economic behavior consists of distinguishing between *true dynamics* due to state dependence and *spurious dynamics* due to unobserved heterogeneity (Heckman, 1981). There are two general approaches to deal with this issue: random effects and fixed effects models/methods. Random-effects models impose restrictions on the distribution of unobserved heterogeneity (e.g., parametric, finite mixture), and on the joint distribution of these unobservables and the initial conditions of the observable explanatory variables. In contrast, fixed-effects methods are more robust because they are fully nonparametric in the specification of the joint distribution of unobserved heterogeneity and exogenous or predetermined explanatory variables.[1]

Among the class of fixed-effects estimators in short panels, the dummy-variables estimator is the simplest of these methods. However, due to the *incidental parameters problem*, this estimator is inconsistent in most nonlinear panel data models when $T$ is fixed (Neyman and Scott, 1948, Lancaster, 2000).[2] Other estimators are based on a transformation of the model that eliminates the fixed effects, e.g., Manski's maximum score estimator (Manski, 1987). However, these estimators, for consistency, require strict exogeneity of the explanatory variables, ruling out dynamic models. Finally, pioneered by Andersen (1970) and extended by Chamberlain (1980), another type of fixed effects methods is based on the derivation of sufficient statistics for the incidental parameters (fixed effects) and the maximization of a likelihood function conditional on these sufficient statistics. This paper deals with the fixed effects - sufficient statistics - conditional maximum likelihood approach (FE-CMLE hereinafter). We study the applicability of this approach to structural dynamic discrete choice models where agents are forward-looking.[3]

There is a wide class of nonlinear panel data models where the FE-CMLE approach cannot identify the structural parameters.[4] In general, a sufficient statistic of the incidental parameters

[1] See Arellano and Honoré (2001), and Arellano and Bonhomme (2012, 2017) for recent surveys on the econometrics of nonlinear panel data models.

[2] Two-step bias reduction methods, both analytical and simulation-based, have been proposed to correct for the asymptotic bias of these dummy-variables fixed-effect estimators (e.g., Hahn and Newey, 2004, Browning and Carro, 2010, and Hahn and Kuersteiner, 2011, among others).

[3] This approach in this paper is also in the spirit of Bonhomme's functional diffencing approach (Bonhomme, 2012).

[4] In this paper, the concepts of identification and consistent estimation, as $N$ goes not infinity and $T$ is fixed, are used as synonymous.

always exists.[5] The identification problem appears when the minimal sufficient statistic is such that the likelihood conditional on this statistic does not depend on the structural parameters. For instance, in the context of binary choice models, Chamberlain (1993, 2010) shows that a necessary and sufficient condition for identification under the FE-CMLE approach is that the distribution of the time-varying unobservable is logistic.[6] Similarly, identification is not possible in discrete choice models where unobserved heterogeneity appears in the slope parameters, interacting with predetermined explanatory variables [7] This has important implications for structural dynamic discrete choice models. In these models, even if the fixed effect is additively separable in the one-period utility function, the solution of the structural model implies that this unobserved variable appears in the continuation value function interacting non-additively with the observable state variables. In general, this interaction between the unobserved heterogeneity and the endogenous state variables implies that structural parameters are not identified in the fixed-effects model.

For non-structural (i.e., myopic) dynamic logit models with unobserved heterogeneity only in the intercept, Chamberlain (1985) and Honoré and Kyriazidou (2000) have shown that the FE-CMLE approach can identify the parameters of interest.[8] In contrast, all the methods and applications for structural dynamic discrete choice models have considered random-effects models with a finite mixture distribution, e.g., Keane and Wolpin (1997), Aguirregabiria and Mira (2007), Kasahara and Shimotsu (2009), Arcidiacono and Miller (2011), among many others. This random-effects approach imposes important restrictions: the number of points in the support of the unobserved heterogeneity is finite and is typically reduced to a small number of points; furthermore, the joint distribution of the unobserved heterogeneity and the initial conditions of the observable state vari-

[5] For instance, we could define as sufficient statistic the complete choice history of an individual. Obviously, the conditional likelihood function based on this sufficient statistic does not depend neither on incidental nor on structural parameters. Though this is an extreme example, it illustrates that the key identification problem is not finding a sufficient statistic for the incidental parameters but showing that there are sufficient statistics for which the conditional likelihood still depends on the structural parameters.

[6] Chamberlain (1993, 2010) considers the model where the time-varying unobservables are independently and identically distributed. Magnac (2004) studies a two-period model where the two time-varying unobservables have a general joint distribution. Honorè and Tamer (2006) study partial identification of the dynamic Probit model and derive sharp bounds on parameters.

[7] Browining and Carro (2014) study the identification of this type of dynamic binary choice model with maximal heterogeneity in short panels. The fixed-effects model (nonparametric specification of the unobserved heterogeneity) is not identified. The consider a finite mixture specification of the heterogeneous parameters. This is in the same spirit as Kasahara and Shimotsu (2009), though these other authors consider a fully nonparametric Markov chain with unobserved heterogeneity.

[8] In the models of these papers, the only endogenous (predetermined) explanatory variable is the lagged decision. For instance, duration since the last change in choice is not an explanatory variable. In our model, we include both lagged decision and duration as state variables.

ables is restricted.[9]

In this paper, we revisit the applicability of FE-CMLE methods to the estimation of structural dynamic discrete choice models. We follow the sufficient statistics approach to study the identification of payoff function parameters in structural dynamic logit models with a fixed-effects specification of the time-invariant unobserved heterogeneity. We consider multinomial models with two types of endogenous state variables: the lagged value of the decision variable, and the time duration in the last choice. The main challenge for the identification of this model comes from the fact that unobserved heterogeneity enters not only in current utility but also in the continuation value of the forward-looking decision problem. In general, this continuation value is a nonlinear function of unobserved heterogeneity and state variables.[10] Therefore, identification requires a sufficient statistic that controls for this continuation value but implies a conditional likelihood that still depends on the structural parameters that capture true state dependence. We derive the minimal sufficient statistic and show that some structural parameters are identified. The forward-looking model where the only state variable is the lagged decision is identified under the same conditions as the myopic version of the model. Instead, with duration dependence, there are some parameters identified in the myopic model but not in the forward-looking model.

Based on our identification results, we consider a conditional maximum likelihood estimator, and a test for the validity of a correlated random effects specification. We apply this estimator and the test to the bus engine model Rust (1987) using both simulated and actual data.

This paper contributes to the literature on structural dynamic discrete choice models. The structure of the payoff function and of the endogenous state variables that we consider in this paper includes as particular cases important economic applications in the literature of dynamic discrete choice structural models, such as models of market entry and exit either binary (Roberts and Tybout, 1997, Aguirregabiria and Mira, 2007) or multinomial (Sweeting, 2013; Caliendo et al, 2015); occupational choice models (Miller, 1984; Keane and Wolpin, 1997); machine replacement models (Rust, 1987; Das, 1992; Kennet, 1993; and Kasahara, 2009); inventory and investment

[9] In general, the joint distribution of the unobserved heterogeneity and the initial conditions is not nonparametrically identified. However, misspecification of this joint distribution can generate important biases in the parameters of interest (Heckman, 1981, Chamberlain, 1985, Lancaster, 2000, among others).

[10] In fact, before solving the model, we do not know how unobserved heterogeneity and state variables enter this continuation value function. Therefore, for fixed-effects estimation, it is as if we had a nonparametric specification of this function.

decision models (Aguirregabiria 1999; Ryan, 2013; Kalouptsidi, 2014); demand of differentiated products with consumer brand switching costs (Erdem, Keane, and Sun, 2008) or storable products (Erdem, Imai, and Keane, 2003; Hendel and Nevo, 2006); and dynamic pricing models with menu costs (Willis, 2006), or with duration dependence due to inflation or other forms of depreciation (Slade, 1998; Aguirregabiria, 1999; Kano, 2013); among others.[11] Our paper also contributes to the literature on nonlinear dynamic panel data models by providing new identification results of fixed effects dynamic logit models with duration dependence (Frederiksen, Honoré, and Hu, 2007).

The rest of the paper is organized as follows. Section 2 describes the class of models that we study in this paper. Section 3 presents our identification results. Section 4 deals with estimation and inference. In section 5, we illustrate our identification results in the context of a bus replacement model. Section 6 summarizes and concludes. Proofs of Lemmas and Propositions are in the Appendix. Also in the Appendix, we show that our identification results extend to a more general version of our model where the endogenous state variables have a stochastic transition rule.

# 2 Model

Time is discrete and indexed by $t$ that belongs to $\{1, 2, ..., \infty\}$.[12] Agents are indexed by $i$. Every period $t$, agent $i$ chooses a value of the discrete variable $y_{it} \in \mathcal{Y} = \{0, 1, ..., J\}$ to maximize her expected and discounted intertemporal utility $\mathbb{E}_t \left[ \sum_{j=0}^{\infty} \delta_i^j \; \Pi_{i,t+j}(y_{i,t+j}) \right]$, where $\delta_i \in (0, 1)$ is agent $i$'s time discount factor, and $\Pi_{it}(y)$ is her one-period utility if she chooses action $y$. This utility is a function of four types of state variables which are known to the agent at period $t$:

$$\Pi_{it}(y) = \alpha\left(y, \boldsymbol{\eta}_i, \mathbf{z}_{it}\right) + \beta\left(y, \mathbf{x}_{it}, \mathbf{z}_{it}\right) + \varepsilon_{it}(y). \tag{1}$$

$\mathbf{z}_{it}$ and $\mathbf{x}_{it}$ are observable to the researcher, and $\varepsilon_{it}$ and $\boldsymbol{\eta}_i$ are unobservable. The vector $\mathbf{z}_{it}$ contains exogenous state variables and it follows a Markov process with transition probability function $f_{\mathbf{z}}(\mathbf{z}_{i,t+1}|\mathbf{z}_{it})$. The vector $\mathbf{x}_{it}$ contains endogenous state variables. We describe below the nature of these endogenous state variables and their transition rules. Both $\mathbf{z}_{it}$ and $\mathbf{x}_{it}$ have discrete supports $\mathcal{Z}$ and $\mathcal{X}$, respectively. The unobservable variables $\{\varepsilon_{it}(y) : y \in \mathcal{Y}\}$ are *i.i.d.* over $(i, t, y)$

[11] Note that most of the empirical applications cited above in this paragraph do not allow for time-invariant unobserved heterogeneity. This is still a common approach in empirical applications. The exceptions, within the cited papers, are Keane and Wolpin (1997), Erdem, Imai, and Keane (2003), Willis (2006), Aguirregabiria and Mira (2007), and Erdem, Keane, and Sun (2008).

[12] The time horizon of the decision problem is infinite.

with an extreme value type I distribution. The vector $\boldsymbol{\eta}_i$ represents time-invariant unobserved heterogeneity from the point of view of the researcher. Let $\theta_i \equiv (\boldsymbol{\eta}_i, \delta_i)$ represent the unobserved heterogeneity from individual $i$. The probability distribution of $\theta_i$ conditional on the history of observable state variables $\{\mathbf{z}_{it}, \mathbf{x}_{it} : t = 1, 2, ...\}$ is unrestricted and nonparametrically specified, i.e., fixed effects model. Functions $\alpha(y, \boldsymbol{\eta}, \mathbf{z})$ and $\beta(y, \mathbf{x}, \mathbf{z})$ are nonparametrically specified but they are bounded.

Our specification of the utility function represents a general semiparametric fixed-effect logit model. It builds on *Rust model* (Rust, 1987, 1994) and generalizes it in two directions. First, Rust assumes that all the unobservables satisfy the conditions of *additive separability* and *conditional independence*, and they have an extreme value distribution. While our time-varying unobservables $\varepsilon_{it}(y)$ satisfy these conditions, our time-invariant unobserved heterogeneity interacts, in an unrestricted way, with the exogenous state variables and the choice, and they do not satisfy the conditional independence assumption. Second, we allow for unobserved heterogeneity in the discount factor.

The assumption of additive separability between $\boldsymbol{\eta}_i$ and the endogenous state variables in $\mathbf{x}_{it}$ is key for the identification results and estimation methods in this paper. This condition does not imply that the conditional-choice value functions, that describe the solution of the dynamic model, are additive separability between $\boldsymbol{\eta}_i$ and $\mathbf{x}_{it}$. In general, the solution of the dynamic programming problem implies a value function that is not additively separable in $\boldsymbol{\eta}_i$ and $\mathbf{x}_{it}$ even when the utility function is additive in these variables.

The model can accommodate two types of endogenous state variables that correspond to two different types of state dependence, $\mathbf{x}_{it} = (y_{i,t-1}, d_{it})$: (a) dependence on the the lagged decision variable, $y_{i,t-1}$; and (b) *duration dependence*, where $d_{it} \in \{1, 2, ..., \infty\}$ is the number of periods since the last change in choice. The lagged decision has the obvious transition rule. The transition rule for the duration variable is $d_{i,t+1} = 1\{y_{it} = y_{i,t-1}\}\ d_{it} + 1$, where $1\{.\}$ is the indicator function.[13]

The term $\beta(y, \mathbf{x}_{it}, \mathbf{z}_{it})$ in the payoff function captures the dynamics, or structural state dependence, in the model. We distinguish in this function two additive components that correspond to

[13] Note that these endogenous state variables follow deterministic transition rules. In the Appendix, we present a version of the model that allows for stochastic transition rules for the endogenous state variables.

the two forms of state dependence in the model:

$$\beta\left(y, \mathbf{x}_{it}, \mathbf{z}_{it}\right) = 1\{y = y_{i,t-1}\}\ \beta_d\left(y, d_{it}, \mathbf{z}_{it}\right) + 1\{y \neq y_{i,t-1}\}\ \beta_y\left(y, y_{i,t-1}, \mathbf{z}_{it}\right) \tag{2}$$

Function $\beta_d\left(y, d_{it}, \mathbf{z}_{it}\right)$ captures duration dependence. For instance, in an occupational choice model, this term captures the return on earnings of job experience in the current occupation. Function $\beta_y\left(y, y_{i,t-1}, \mathbf{z}_{it}\right)$ represents switching costs. In an occupational choice model, this term represents the cost of switching from occupation $y_{i,t-1}$ to occupation $y$. The additive separability between switching costs and "returns to experience" is not without loss of generality. For instance, the cost of switching occupation could depend on experience in the current job not only through the loss of the returns of experience, i.e., $\beta_y(.)$ could depend on $d_{it}$. However, this additive separability facilitates our analysis of identification and the model is still more general than previous fixed-effects discrete choice models.

We impose a restriction on the structural function $\beta_d\left(y, d, \mathbf{z}_{it}\right)$ that play a role in our identification results for this function. We assume that there is not duration dependence in choice alternative $y = 0$, i.e., $\beta_d\left(0, d, \mathbf{z}_{it}\right) = 0$ for any value of $d$. Also, but without loss of generality, we set $\beta_y(y, y, \mathbf{z}_{it}) = 0$, i.e., the switching cost of no-switching is zero.[14] Assumption 1 summarizes our basic conditions on the model. For the rest of the paper, we assume that this assumption holds.

*ASSUMPTION 1. (A) The time horizon is infinite and $\delta_i \in (0,1)$. (B) The utility function has the form given by equations (1) and (2), and functions $\alpha\left(y, \eta, \mathbf{z}\right)$, $\beta_d\left(y, d, \mathbf{z}\right)$, and $\beta_y(y, y_{-1}, \mathbf{z})$ are bounded. (C) $\beta_y(y, y, \mathbf{z}) = 0$, $\beta_d\left(0, d, \mathbf{z}\right) = 0$. (D) $\{\varepsilon_{it}(y) : y \in \mathcal{Y}\}$ are i.i.d. over $(i, t, y)$ with a extreme value type I distribution. (E) $\mathbf{z}_{it}$ has discrete and finite support $\mathcal{Z}$ and follows a time-homogeneous Markov process. (F) The probability distribution of $\theta_i \equiv (\boldsymbol{\eta}_i, \delta_i)$ conditional on $\{\mathbf{z}_{it}, \mathbf{x}_{it} : t = 1, 2, ...\}$ is nonparametrically specified and completely unrestricted.* ■

Since the model does not have duration dependence when at choice alternative 0, it is convenient for notation to make duration equal to zero at state $y_{t-1} = 0$. In other words, we consider the following modification in the transition rule for duration:

$$d_{i,t+1} = \begin{cases} 1\left\{y_{it} = y_{i,t-1}\right\}\ d_{it} + 1 & \text{if} \quad y_{it} > 0 \\ 0 & \text{if} \quad y_{it} = 0 \end{cases} \tag{3}$$

[14] Given the payoff function in equation (2), the parameter $\beta_y(y, y)$ is completely irrelevant for an individual's optimal decision. When $y_{it} = y_{i,t-1} = y$, we have that $\beta\left(y, \mathbf{x}_{it}\right) = \beta_d\left(y, d_{it}\right) + 0$ such that the term $\beta_y\left(y, y\right)$ never enters in the relevant payoff function. Therefore, $\beta_y\left(y, y\right)$ can be normalized to zero without loss of generality.

For our identification results in forward-looking models with duration dependence, we also impose the following assumption.

*ASSUMPTION 2. For any $y \in \mathcal{Y}$ there is a finite value of duration, $d_y^* < \infty$, such that the marginal return of duration is zero for values greater that $d_y^*$:*[15]

$$\beta_d(y, d, \mathbf{z}) = \beta_d\left(y, d_y^*, \mathbf{z}\right) \quad \textit{for any } d \geq d_y^* \qquad \blacksquare \tag{4}$$

For the moment, we assume that the researcher knows the values of $d_y^*$. In section 4, we show that these values $\{d_y^*\}$ are identified from the data, as long as $d_y^* \leq (T-1)/2$.

The following are some examples of models within the class defined by Assumption 1.

*(a) Market entry-exit models.* In its simpler version, there is only one market, and the choice variable is binary and represents a firm's decision of being active in the market ($y_{it} = 1$) or not ($y_{it} = 0$), e.g., Dunne et al. (2013). The only endogenous state variable is the lagged decision, $y_{i,t-1}$. The parameter $-\beta_y(1, 0)$ represents the cost of entry in the market. Similarly, the parameter $-\beta_y(0, 1, \mathbf{z})$ represents the cost of exit from the market. An extension of the basic entry model includes as an endogenous state variable the number of periods of experience since last entry in the market, $d_{it}$, which follows the transition rule $d_{i,t+1} = d_{it} + 1$ if $y_{it} = 1$ and $d_{i,t+1} = 0$ if $y_{it} = 0$. The parameter $\beta_d(1, d, \mathbf{z})$ represents the effect of market experience on the firm's profit (Roberts and Tybout, 1997). The model can be extended to $J$ markets (Sweeting, 2013; Caliendo et al, 2015). The two endogenous state variables are the index of the market where the firm was active at the previous period ($y_{i,t-1}$) and the number of periods of experience in the current market ($d_{it}$). The parameter $\beta_y(y, y_{-1}, \mathbf{z})$ represents the cost of switching from market $y_{-1}$ to market $y$. There is not duration dependence if a firm is not active in any market (if $y = 0$), and the marginal return to experience in market $y$ is zero after $d_y^*$ periods in the market.

*(b) Occupational choice models (Miller, 1984; Keane and Wolpin, 1997).* A worker chooses between $J$ occupations and the choice alternative of not working ($y = 0$). There are costs of switching occupations such that a worker's occupation at previous period, $y_{it-1}$, is a state variable of the model. There is (passive) learning that increases productivity in the current occupation. There is not duration dependence if the worker is unemployed.

[15] The assumption of no duration dependence in choice alternative $y = 0$ is equivalent to assuming $d_0^* = 1$.

*(c) Machine replacement models (Rust, 1987; Das, 1992; Kennet, 1993; and Kasahara, 2009).* The choice variable is binary and it represents the decision of keeping a machine ($y_{it} = 1$) or replacing it ($y_{it} = 0$). The only endogenous state variable is the number of periods since the last replacement, $d_{it}$, i.e., the machine age. The evolution of the machine age is $d_{i,t+1} = d_{it} + 1$ if $y_{it} = 1$ and $d_{i,t+1} = 0$ if $y_{it} = 0$. The parameter $\beta_d (1, d, \mathbf{z})$ represents the effect of age on the firm's profit, e.g., productivity declines and maintenance costs increase with age.[16] More generally, the class of models in this paper includes binary choice models of investment in capital, inventory, or capacity (Aguirregabiria 1999; Ryan, 2013; Kalouptsidi, 2014), as long as the depreciation of the stock is deterministic.

*(d) Dynamic demand of differentiated products (Erdem, Imai, and Keane, 2003; Hendel and Nevo, 2006).* A differentiated product has $J$ varieties and a consumer chooses which one, if any, to purchase (no purchase is represented by $y = 0$). Brand switching costs imply that the brand in the last purchase is a state variable (Erdem, Keane, and Sun, 2008). For storable products, the duration since last purchase, $d_{it}$, represents (or proxies) the consumer's level of inventory that is an endogenous state variable. Function $\beta_d (y, d, \mathbf{z})$ captures the effect of inventory on the consumer's utility, and function $\beta_y (y, y_{-d}, , \mathbf{z})$ represents brand switching costs.

*(e) Menu costs models of pricing (Slade, 1998; Aguirregabiria, 1999; Willis, 2006; Kano, 2013).* A firm sells a product and chooses its price to maximize intertemporal profits. The firm's profit has two components: a variable profit that depends on the real price (in logarithms), $r_{it}$; and a fixed menu cost that is paid only if the firm changes its nominal price. There is a constant inflation rate, $\pi$, that erodes the real price. Every period, the firm decides whether to keep its nominal price ($y_{it} = 1$) or to adjust it ($y_{it} = 0$) such that current real price becomes $r^*$. The evolution of log-real-price is: $r_{it+1} = r_{it} - \pi$ if $y_{it} = 1$, and $r_{it+1} = r^* - \pi$ if $y_{it} = 0$. If $d_{it}$ represents the time duration since the last nominal price change, we can represent the transition rule of the real price as follows: $(r_{it+1} - r^*)/\pi = d_{it} + 1$ if $y_{it} = 1$, and $(r_{it+1} - r^*)/\pi = 0$ if $y_{it} = 0$. This model has a similar structure as the machine replacement models described above. ■

We now derive the optimal decision rule and the conditional choice probabilities in this model.

[16] In some versions of this model, such as Rust (1987), the endogenous state variable represents cumulative usage of the machine and it can follow a stochastic transition rule. We consider this stochastic version of the model in the Appendix.

Agent $i$ chooses $y_{it}$ to maximize its expected and discounted intertemporal utility. Given the infinite horizon and the time-homogeneous utility and transition probability functions, Blackwell's Theorem establishes that the value function and the optimal decision rule are time-invariant (Blackwell, 1965). Let $V_{\theta_i}(y_t, d_t, \mathbf{z}_t)$ be the integrated (or smoothed) value function for agent type $\theta_i$, as defined by Rust (1994).[17] The optimal choice at period $t$ can be represented as:

$$y_{it} = \arg\max_{y \in \mathcal{Y}} \left\{ \alpha(y, \boldsymbol{\eta}_i, \mathbf{z}_{it}) + \beta(y, \mathbf{x}_{it}, \mathbf{z}_{it}) + \varepsilon_{it}(y) + \delta_i \, \mathbb{E}\left[V_{\theta_i}(y, d_{i,t+1}, \mathbf{z}_{i,t+1}) \mid y, \mathbf{x}_{it}, \mathbf{z}_{it}\right] \right\} \tag{5}$$

Note that $d_{i,t+1}$ is a deterministic function of $(y, \mathbf{x}_{it})$, i.e., conditional on $y_{it} = y$. Therefore, we can represent the continuation value $\mathbb{E}[V_{\theta_i}(y, d_{i,t+1}, \mathbf{z}_{i,t+1}) \mid y, \mathbf{x}_{it}, \mathbf{z}_{it}]$ using a function $v_{\theta_i}(y, d_{it+1}(y), \mathbf{z}_{it})$ with $d_{it+1}(y) = 0$ if $y = 0$ and $d_{it+1}(y) = 1\{y = y_{it-1}\}d_{it} + 1$ if $y > 0$. The extreme value type 1 distribution of the unobservables $\varepsilon$ implies that the *conditional choice probability* (CCP) function has the following form:

$$P_{\theta_i}(y \mid \mathbf{x}_{it}, \mathbf{z}_{it}) = \frac{\exp\left\{ \alpha(y, \boldsymbol{\eta}_i, \mathbf{z}_{it}) + \beta(y, \mathbf{x}_{it}, \mathbf{z}_{it}) + v_{\theta_i}(y, d_{it+1}(y), \mathbf{z}_{it}) \right\}}{\sum_{j \in \mathcal{Y}} \exp\left\{ \alpha(j, \boldsymbol{\eta}_i, \mathbf{z}_{it}) + \beta(j, \mathbf{x}_{it}, \mathbf{z}_{it}) + v_{\theta_i}(j, d_{it+1}(j), \mathbf{z}_{it}) \right\}} \tag{6}$$

The continuation value function $v_{\theta_i}$ has two properties which play an important role in our identification results. First, for any duration $d_{it+1}(y) \geq d_y^*$, we have that $v_{\theta_i}(y, d_{it+1}(y), \mathbf{z}_{it}) = v_{\theta_i}(y, d_y^*, \mathbf{z}_{it})$. Second, in a model without duration dependence (i.e., $\beta_d = 0$), the continuation value function becomes $v_{\theta_i}(y, \mathbf{z}_{it})$ that does not depend on the state variable, $y_{it-1}$.

# 3 Identification

## 3.1 Preliminaries

The researcher has a panel dataset of $N$ individuals over $T$ periods of time, $\{y_{it}, \mathbf{x}_{it}, \mathbf{z}_{it} : i = 1, 2, ..., N \, ; \, t = 1, 2, ..., T\}$. We consider microeconometric applications where $N$ is large and $T$ is small. More precisely, our identification results and the asymptotic properties of the proposed estimators assume that $N$ goes to infinity and $T$ is small and fixed.[18] We are interested in the

[17] The integrated value function is defined as the integral of the value function over the distribution of the i.i.d. unobservable state variables $\varepsilon$.

[18] Note that $T$ represents the number of periods with data on the decision variable and the state variables for all the individuals. The set of observable state variables includes the endogenous state variables $y_{i,t-1}$ and $d_{it}$. Knowing the values of these state variables at the initial period $t = 1$ (i.e., $y_{i0}$ and $d_{i1}$) may require data on the individual's choices for periods before $t = 1$. Therefore, the time dimension $T$ may not correspond to the actual time dimension of the required panel dataset.

identification of the functions $\beta_y$ and $\beta_d$ that represent the dependence of utility with respect to the endogenous state variables.

For the rest of this section, we omit the individual subindex $i$ in most of the expressions, and instead we include $\theta$ as an argument (or subindex) in those functions that depend on the time-invariant unobserved heterogeneity, i.e., $\alpha_\theta(y, \mathbf{z})$ and $v_\theta(\mathbf{x}, \mathbf{z})$. We use $\beta$ to represent the vector of structural parameters that define the functions $\beta_y$ and $\beta_d$.[19]

As in Honoré and Kyriazidou (2000), our sufficient statistics include the condition that the exogenous state variables, $\mathbf{z}$, remains constant over several periods. For notational simplicity, we omit $\mathbf{z}$ as an argument in most of the expressions for the rest of this section. In section 4, we explain how to deal with this condition in the implementation of the conditional maximum likelihood estimator.

Let $\mathbf{y}^T = \{y_1, y_2, ..., y_T\}$ be an individual's observed history of choices and exogenous state variables, respectively. The model implies that:

$$\mathbb{P}\left(\mathbf{y}^T \mid \mathbf{x}_1, \theta, \beta\right) = \prod_{t=1}^{T} \frac{\exp\{\ \alpha_\theta(y_t) + \beta(y_t, \mathbf{x}_t) + v_\theta(y_t, d_{t+1}(y_t))\ \}}{\sum\limits_{j\in\mathcal{Y}} \exp\{\ \alpha_\theta(j) + \beta(j, \mathbf{x}_t) + v_\theta(j, d_{t+1}(j))\ \}} \tag{7}$$

Our identification results, for different versions of the model, have the following common features. First, we show that the log-probability function $\ln \mathbb{P}\left(\mathbf{y}^T \mid \mathbf{x}_1, \theta, \beta\right)$ has the following structure:

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid \mathbf{x}_1, \theta, \beta\right) = U(\mathbf{y}^T, \mathbf{x}_1)' g_\theta + S(\mathbf{y}^T, \mathbf{x}_1)' \beta^* \tag{8}$$

where $U = U(\mathbf{y}^T, \mathbf{x}_1)$ and $S = S(\mathbf{y}^T, \mathbf{x}_1)$ are vector of statistics (i.e., deterministic functions of the history $(\mathbf{y}^T, \mathbf{x}_1)$), $g_\theta$ is a vector of functions of $\theta$, and $\beta^*$ is a vector of linear combinations of the original vector of structural parameters $\beta$. This representation is such that each of the vectors, $U$, $g_\theta$, $S$, and $\beta^*$, has elements which are linearly independent.[20] Based on this representation of the log-probability of a choice history, we establish the following results.

*(i) Sufficiency.* $U = U(\mathbf{y}^T, \mathbf{x}_1)$ is a sufficient statistic for $\theta$, i.e., for any $(\mathbf{y}^T, \mathbf{x}_1)$ and $\theta$, $\ln \mathbb{P}\left(\mathbf{y}^T | \mathbf{x}_1, \theta, U\right)$ does not depend on $\theta$. Note that, by definition, $\ln \mathbb{P}\left(\mathbf{y}^T | \mathbf{x}_1, \theta, U\right)$ is equal to $\ln \mathbb{P}\left(\mathbf{y}^T | \mathbf{x}_1, \theta\right) -$

[19] Since $(y_t, \mathbf{x}_t, \mathbf{z}_t)$ has finite support, we can represent the structural functions $\beta_y(y_t, y_{t-1}, \mathbf{z}_t)$ and $\beta_d(y_t, d_t, \mathbf{z}_t)$ using a vector of parameters in a finite dimensional space.

[20] Suppose that $S$ and $\beta$ are $K \times 1$ vectors, and only $K^* < K$ elements in $S$ are linearly independent. Then, $S = [S_a, S_b]$ where $S_a$ contains $K^*$ linearly independent elements, and $S_b = \mathbf{A}\ S_a$ where $\mathbf{A}$ is a $(K - K^*) \times K^*$ matrix. This implies that $S'\beta = S_a'\beta^*$ with $\beta^* = [\mathbf{I} \vdots \mathbf{A}]'\beta$, such that $S_a$ and $\beta^*$ are vectors with linearly independent elements.

$\ln \mathbb{P}(U|\mathbf{x}_1, \theta)$, and taking into account the form of the log-probability in equation (8), we have:

$$\begin{aligned}\ln \mathbb{P}\left(\mathbf{y}^T|\mathbf{x}_1, U, \beta^*\right) &= U(\mathbf{y}^T, \mathbf{x}_1)'g_\theta + S(\mathbf{y}^T, \mathbf{x}_1)'\beta^* - \ln\left(\sum_{j:U(j)=U(\mathbf{y}^T,\mathbf{x}_1)} \exp\left\{U(j)'g(\theta) + S(j)'\beta^*\right\}\right) \\ &= S(\mathbf{y}^T, \mathbf{x}_1)'\beta^* - \ln\left(\sum_{j:U(j)=U(\mathbf{y}^T,\mathbf{x}_1)} \exp\left\{S(j)'\beta^*\right\}\right)\end{aligned} \tag{9}$$

where we use $j$ to index all the possible histories $(\mathbf{y}^T, \mathbf{x}_1)$, and $\sum_{j:U(j)=U}$ represents the sum over all the histories that imply a the same value $U$ of the vector of statistics. Equation (9) shows that the structure of the log-probability in (8) implies that $U$ is a sufficient statistic for $\theta$.

*(ii) Minimal sufficiency.* $U(\mathbf{y}^T, \mathbf{x}_1)$ is a *minimal sufficient statistic*, i.e., it does not contain redundant information. More formally, let $\mathbf{U}$ be a matrix where each row corresponds to a value of the choice history $(\mathbf{y}^T, \mathbf{x}_1)$. Then, $U(\mathbf{y}^T, \mathbf{x}_1)$ is *minimal* if and only if matrix $\mathbf{U}$ is full-column rank.

*(iii) Identification.* Define the conditional log-likelihood function, in the population, $\ell(\beta^*) \equiv \mathbb{E}_{(\mathbf{y}^T,\mathbf{x}_1)}\left[\ln \mathbb{P}\left(\mathbf{y}^T|\mathbf{x}_1, U, \beta^*\right)\right]$. We show that this population likelihood is uniquely maximized at the true value of $\beta^*$, i.e., $\beta^*$ is point identified. Lemma 1 establishes a necessary and sufficient condition for identification that is simple to verify. Let $K$ be the dimension of the vector of statistics $S$ and of the vector of parameters $\beta^*$.

*LEMMA 1. Given $K+1$ histories $(\mathbf{y}^T, \mathbf{x}_1)$, say $\{A_j : j = 0, 1, 2..., K\}$, define a $K \times K$ matrix $\mathbf{S}$ such that every row $j$ is associated to a history and contains the vector of statistics $S(A_j)' - S(A_0)'$. The vector of parameters $\beta^*$ is identified if and only if there exist $K+1$ histories $\{A_j : j = 0, 1, 2..., K\}$ with the same value of the statistic $U$ and matrix $\mathbf{S}$ that is non-singular.* ■

*Corollary: If $K = 1$, parameter $\beta^*$ is identified iff there are two histories, $A$ and $B$, such that $U(A) = U(B)$ and $S(A) \neq S(B)$.*

The derivation of these sufficient statistics should deal with two issues that do not appear in the previous literature on FE-CMLE of non-structural (or myopic) nonlinear panel data models. First, we consider models with duration dependence. Second, we should take into account that unobserved heterogeneity enters in the continuation value function, $v_\theta$. This implies that the sufficient statistic $U$ should control not only for $\alpha_\theta(y)$ but also for the continuation values $v_\theta(y, d)$. This is challenging because, in general, these continuation values depend on the endogenous state variables. We cannot fully control for (or condition on) the value of the state variables because the identification condition (iii) would not hold. We show that there are states where the continuation value does not depend on current state variables once we condition on current choices.

The presentation of our identification results tries to emphasize both the links and extensions with previous results in the literature. For this reason, we start presenting identification results for the binary choice model, that is the model more extensively studied in the literature of nonlinear dynamic panel data. For this binary choice model, we present new identification results for the myopic model with duration dependence and for the forward-looking model with and without duration dependence. Then, we present our identification results for multinomial models.

*Some useful statistics.* We show below that, in our model, the log-probability of a choice history, $\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta, \beta\right)$, can be written in terms of several sets of statistics or functions of $(y_0, d_1, \mathbf{y}^T)$: the initial and final choices, $\{y_0, y_T\}$; the initial and final durations, $\{d_1, d_{T+1}\}$; and the statistics that we denote as *hits*, *dyads*, *switches*, *histogram of durations*, and *extended histogram of durations*. We now define these statistics. Note that each of these statistics is for a single history $(y_0, d_1, \mathbf{y}^T)$.

*Hit statistics.* For any choice alternative $y \in \mathcal{Y}$, the *hit* statistic $T^{(y)}$ represents the number of times that alternative $y$ is visited (or *hit*) during the choice history $\mathbf{y}^T$, i.e., $T^{(y)} \equiv \sum_{t=1}^{T} 1\{y_t = y\}$.

*Dyad statistics.* For $y_{-1}$ and $y$ in $\mathcal{Y}$, the *dyad* statistic $D^{(y_{-1},y)}$ is the number of times that the sequence $(y_{-1}, y)$ is observed at two consecutive periods in the choice history $(y_0, \mathbf{y}^T)$, i.e., $D^{(y_{-1},y)} \equiv \sum_{t=1}^{T} 1\{y_{t-1} = y_{-1},\, y_t = y\}$.

*Histogram of durations.* Given a history $(y_0, d_1, \mathbf{y}^T)$, the statistic $H^{(y)}(d)$ (for $y \in \mathcal{Y}$ and $d \geq 0$) is the number of times that we observe duration $d$ under choice alternative $y$ between periods 1 and $T$, i.e., $H^{(y)}(d) = \sum_{t=1}^{T} 1\{y_{t-1} = y,\, d_t = d\}$.

*Extended histogram of durations.* For any $y \in \mathcal{Y}$ and $d \geq 0$, the statistic $X^{(y)}(d)$ represents the number of times that we observe duration $d$ under choice alternative $y$ and the individual decides to continue one more period in that choice. By definition, $X^{(y)}(d) = \sum_{t=1}^{T} 1\{y_{t-1} = y_t = y,\, d_t = d\}$.

*Difference between final and initial state variables.* For any $y \in \mathcal{Y}$ and $d \geq 0$, the statistic $\Delta^{(y)}(d)$ is defined as $1\{y_T = y,\, d_{T+1} = d\} - 1\{y_0 = y,\, d_1 = d\}$. When the difference applies only to the choice variable, we represent it as $\Delta^{(y)} \equiv 1\{y_T = y\} - 1\{y_0 = y\}$.

Table 1 summarizes our definition of statistics. The following Lemma 2 establishes several properties of these statistics that we apply in our derivations.

*LEMMA 2. For any history* $(y_0, d_1, \mathbf{y}^T)$ *and value* $y > 0$ *the following properties apply: (i)* $H^{(y)}(0) = 0$*; (ii)* $X^{(y)}(0) = 0$*; (iii)* $\sum_{d \geq 1} H^{(y)}(d) = T^{(y)} - \Delta^{(y)}$*; (iv)* $\sum_{d \geq 1} X^{(y)}(d) = D^{(y,y)}$*; (v) for* $d \geq 1$, $X^{(y)}(d) = H^{(y)}(d+1) + \Delta^{(y)}(d+1)$*; (vi)* $\sum_{d \geq 1} \Delta^{(y)}(d) = \Delta^{(y)}$*; and (vii) for* $y \geq 1$, $\sum_{y_{-1} \neq y} D^{(y_{-1},y)} = H^{(y)}(1) + \Delta^{(y)}$. ■

**Table 1**
**Definition of statistics for a choice history** $\{y_0, d_1 \mid \mathbf{y}^T\}$

| *Name: Symbol* | *Definition* |
|---:|---|
| *Hits:* $T^{(y)}$ | $\sum_{t=1}^{T} 1\{y_t = y\}$ |
| *Dyad:* $D^{(y_{-1},y)}$ | $\sum_{t=1}^{T} 1\{y_{t-1} = y_{-1},\, y_t = y\}$ |
| *Duration histogram:* $H^{(y)}(d)$ | $\sum_{t=1}^{T} 1\{y_{t-1} = y,\, d_t = d\}$ |
| *Extended duration histogram:* $X^{(y)}(d)$ | $\sum_{t=1}^{T} 1\{y_{t-1} = y_t = y,\, d_t = d\}$ |
| *Diff. final-initial states:* $\Delta^{(y)}(d)$ | $1\{y_T = y, d_{T+1} = d\} - 1\{y_0 = y, d_1 = d\}$ |
| $\Delta^{(y)}$ | $1\{y_T = y\} - 1\{y_0 = y\}$ |

### 3.2 Binary choice models

Consider the binary choice version of the model characterized by Assumption 1. The optimal decision rule in this model is:

$$y_t = 1\left\{ \begin{array}{l} \alpha_\theta(1) - \alpha_\theta(0) + \beta(1, y_{t-1}, d_t) - \beta(0, y_{t-1}, d_t) \\ \\ +v_\theta(1, d_t + 1) - v_\theta(0) + \varepsilon_t(1) - \varepsilon_t(0) \geq 0 \end{array} \right\} \tag{10}$$

where for choice $y = 0$ we use $v_\theta(0)$ instead $v_\theta(0,0)$ to emphasize that there is not duration dependence when the state is $y = 0$. We now present identification results for different versions of this model, starting with the myopic model without duration dependence that has been studied by Chamberlain (1985) and Honoré and Kyriazidou (2000).

#### 3.2.1 Myopic dynamic model without duration dependence

Consider the model in equation (10) under the restrictions of myopic behavior (i.e., $\delta = 0$) and no duration dependence (i.e., $\beta_d(y, d) = 0$). These restrictions imply that the continuation values,

$v_\theta(1, d_t+1)$ and $v_\theta(0)$, become zero, and the term $\beta(1, y_{t-1}, d_t) - \beta(0, y_{t-1}, d_t)$ becomes equal to $\beta_y(1,0) - y_{t-1}\ [\beta_y(1,0) + \beta_y(0,1)]$. We can present this model using the more standard representation,

$$y_t = 1\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y\ y_{t-1} + \widetilde{\varepsilon}_t \geq 0\right\} \tag{11}$$

with $\widetilde{\alpha}_\theta \equiv \alpha_\theta(1) - \alpha_\theta(0) + \beta_y(1,0)$, $\widetilde{\beta}_y \equiv -\beta_y(1,0) - \beta_y(0,1)$, and $\widetilde{\varepsilon}_t \equiv \varepsilon_t(1) - \varepsilon_t(0)$. In a model of market entry-exit, the parameter $\widetilde{\beta}_y$ represents the sum of the costs of entry and exit, or equivalently the sunk cost of entry. This is an important structural parameter.

Define function $\sigma_\theta(y_{t-1}) \equiv -\ln\left(1 + \exp\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1}\right\}\right)$. The log-probability of the choice history $\mathbf{y}^T$ conditional on $(y_0, \theta)$ is:

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, \theta\right) \quad = \quad \sum_{t=1}^{T} y_t\left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1}\right] + (1 - y_{t-1})\ \sigma_\theta(0) + y_{t-1}\ \sigma_\theta(1) \tag{12}$$

Proposition 1 establishes (i) the sufficient statistic, (ii) minimal sufficiency, and (iii) identification for this model.

*PROPOSITION 1. In the myopic binary choice model without duration dependence the log-probability* $\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, \theta, \beta\right)$ *has the form*

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, \theta\right) \quad = \quad T^{(1)}\, g_{\theta,1} + \Delta^{(1)}\, g_{\theta,2} + \widetilde{\beta}_y\ D^{(1,1)} \tag{13}$$

*with* $g_{\theta,1} \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1) - \sigma_\theta(0)$, *and* $g_{\theta,2} \equiv \sigma_\theta(0) - \sigma_\theta(1)$, *such that* $U = \{T^{(1)},\ \Delta^{(1)}\}$, $S = D^{(1,1)}$, *and* $\beta^* = \widetilde{\beta}_y$. *We have that: (i)* $U = \{T^{(1)},\ \Delta^{(1)}\}$ *is a sufficient statistic; (ii)* $T^{(1)}$ *and* $\Delta^{(1)}$ *are linearly independent such that* $U$ *is a minimal sufficient statistic; and (iii) for* $T \geq 3$ *there is a pair of histories* $\{y_0|\mathbf{y}^T\}$, *say* $A$ *and* $B$, *with* $U(A) = U(B)$ *and* $S(A) \neq S(B)$ *such that the parameter* $\widetilde{\beta}_y$ *is identified as* $[\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)]\ /\ \left[D_A^{(1,1)} - D_B^{(1,1)}\right]$. *For instance,* $A = \{0|0,1,1\}$ *and* $B = \{0|1,0,1\}$. ■

This Proposition 1 is almost identical to the identification result in Chamberlain (1985). Chamberlain shows that the vector of statistics $\{T^{(1)}, y_0,\ y_T\}$ is sufficient for $\theta$, and conditional on this vector the parameter $\widetilde{\beta}_y$ is identified. Our Proposition 1 shows that Chamberlain's sufficient statistic is not minimal and the minimal statistic is $\{T^{(1)}, y_T - y_0\}$. However, it turns out that, in this model, the extra variation left by the minimal sufficient statistic does not help in the identification of $\widetilde{\beta}_y$, so the two CMLEs are equivalent.

### 3.2.2 Forward-looking dynamic model without duration dependence

Consider a forward-looking version of the model in equation (10) but still without duration dependence. Since the model is of forward-looking behavior, now we have the continuation values $v_\theta(1, d_t + 1) - v_\theta(0)$. However, there is not duration dependence, and the only state variable is $y_{t-1}$. Therefore, for this version of the model we have that $v_\theta(1, d_t + 1) - v_\theta(0)$ becomes $v_\theta(1) - v_\theta(0) \equiv \widetilde{v}_\theta$, i.e., continuation values depend on current choices but not on the current state variable $y_{t-1}$. This is a key property for the derivation of sufficient statistics in this model of forward-looking behavior. We can represent this model as,

$$y_t = 1\{\widetilde{\alpha}_\theta + \widetilde{v}_\theta + \widetilde{\beta}_y y_{t-1} + \widetilde{\varepsilon}_t \geq 0\} \tag{14}$$

The only difference between this model and the myopic model is that now the fixed effect has two components: $\widetilde{\alpha}_\theta$ that comes from current profit, and $\widetilde{v}_\theta$ that comes from the continuation values. However, from the point of view of fixed-effects estimation, the two models are observationally equivalent.

A key feature of this model, that determines the observational equivalence with the myopic model, is the property that the state variable at period $t+1$ depends on the choice at period $t$ but not on the state variable at period $t$, i.e., $x_{t+1} = y_t$.

Proposition 2 establishes this equivalence.

*PROPOSITION 2. In the forward-looking binary choice model without duration dependence the log-probability* $\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, \theta, \beta\right)$ *has the form*

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, \theta\right) \;=\; T^{(1)}\, g_{\theta,1} + \Delta^{(1)}\, g_{\theta,2} + \widetilde{\beta}_y\; D^{(1,1)} \tag{15}$$

*with* $g_{\theta,1} \equiv \widetilde{\alpha}_\theta + \widetilde{v}_\theta + \sigma_\theta(1) - \sigma_\theta(0)$, *and* $g_{\theta,2} \equiv \sigma_\theta(0) - \sigma_\theta(1)$, *such that* $U = \{T^{(1)}, \Delta^{(1)}\}$, $S = D^{(1,1)}$, *and* $\beta^* = \widetilde{\beta}_y$. *We have that: (i)* $U = \{T^{(1)}, \Delta^{(1)}\}$ *is a sufficient statistic; (ii)* $T^{(1)}$ *and* $\Delta^{(1)}$ *are linearly independent such that* $U$ *is a minimal sufficient statistic; and (iii) for* $T \geq 3$ *there is a pair of histories* $\{y_0|\mathbf{y}^T\}$, *say* $A$ *and* $B$, *with* $U(A) = U(B)$ *and* $S(A) \neq S(B)$ *such that the parameter* $\widetilde{\beta}_y$ *is identified as* $[\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)] \,/\, \left[D_A^{(1,1)} - D_B^{(1,1)}\right]$. *For instance,* $A = \{0|0, 1, 1\}$ *and* $B = \{0|1, 0, 1\}$. ■

### 3.2.3 Myopic dynamic model with duration dependence

The continuation values are zero, and the term $\beta(1, y_{t-1}, d_t) - \beta(0, y_{t-1}, d_t)$ that captures state dependence is equal to $(1 - y_{t-1})\ \beta_y(1,0)+\ y_{t-1}\ \beta_d(1, d_t)-\ y_{t-1}\beta_y(0,1)$, and it can be represented as $\beta_y(1,0) + \widetilde{\beta}_y\ y_{t-1} + \beta_d(1, d_t)\ y_{t-1}$. Therefore, we can present this model as

$$y_t = 1\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1} + \beta_d(1, d_t)\ y_{t-1} + \widetilde{\varepsilon}_t \geq 0\} \tag{16}$$

For this model, the log-probability of the choice history $\mathbf{y}^T$ conditional on $(y_0, d_1, \theta)$ is:

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta\right) \quad = \quad \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y\ y_{t-1} + \beta_d(1, d_t)\ y_{t-1}\right] + \sigma_\theta(y_{t-1}, d_t) \tag{17}$$

where $\sigma_\theta(y_{t-1}, d_t) \equiv -\ln\left(1 + \exp\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y\ y_{t-1} + \beta_d(1, d_t)\ y_{t-1}\right\}\right)$. In order to emphasize that $\sigma_\theta(y_{t-1}, d_t)$ does not depend on $d_t$ when $y_{t-1} = 0$, we use the notation $\sigma_\theta(0)$ to represent $\sigma_\theta(0,0)$. Now, with duration dependence, the log-probability of a choice history includes the terms $\sum_{t=1}^{T} y_{t-1}$ $1\{d_t = d\}\ \sigma_\theta(1, d)$, and $\sum_{t=1}^{T} y_t\ y_{t-1}\ 1\{d_t = d\}\ \beta_d(1, d_t)$ for $d \geq 1$.

Proposition 3 establishes the minimal sufficient statistic and identification of structural parameters in this model.

*PROPOSITION 3. In the myopic binary choice model with duration dependence under Assumption 1, the log-probability* $\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta, \beta\right)$ *has the form*

$$\begin{aligned} \ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1\right) \quad &= \quad \sum_{d \geq 1} H^{(1)}(d)\ g_{\theta,1}(d) + \Delta^{(1)}\ g_{\theta,2} \\ &+ \quad \sum_{d \geq 1} \Delta^{(1)}(d)\ \gamma(d-1) \end{aligned} \tag{18}$$

*with* $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1, d) - \sigma_\theta(0) + \gamma(d-1)$, $g_{\theta,2} \equiv \widetilde{\alpha}_\theta$, $\gamma(d) \equiv \widetilde{\beta}_y + \beta_d(1, d)$, *and* $\gamma(0) = 0$, *such that* $U = \{H^{(1)}(d) : d \geq 1,\ \Delta^{(1)}\}$, $S = \{\Delta^{(1)}(d) : d \geq 1\}$, *and* $\beta^* = \{\gamma(d) : d \geq 1\}$. *Then, we have that: (i)* $U$ *is a sufficient statistic. (ii) The elements in the vector* $U$ *are linearly independent such that* $U$ *is a minimal sufficient statistic. (iii) Conditional on* $U$, *the statistics* $\{\Delta^{(1)}(d) : d \geq 1\}$ *have variation and the structural parameters* $\{\gamma(d) : 1 \leq d \leq T-2\}$ *are identified, i.e., for any* $1 \leq d \leq T-2$, *there is a pair of histories,* $A$ *and* $B$, *such that* $U(A) = U(B)$ *and* $\gamma(d) = \ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)$. ■

*Proof.* The derivation of equation (18) is in the Appendix. Proof of (iii). For any duration $n$, with $1 \leq n \leq T-2$, define a sub-history $\{y_0, d_1 \mid y^{n+2}\}$, and consider the sub-histories $A =$

$\{0, 0 \mid 0, \mathbf{1}_{n+1}\}$ and $B = \{0, 0 \mid \mathbf{1}_n, 0, 1\}$, where $\mathbf{1}_n$ represents a sequence of $n$ consecutive $1'$s. The corresponding histories of durations $\{d_t : t = 1, ..., n+2\}$ are: for $A$, $\{0, 0, 1, ..., n\}$; and for $B$, $\{0, 1, ..., n, 0\}$. It is clear that the histogram of durations is the same under the two histories: $H_A^{(1)}(d) = H_B^{(1)}(d) = 1$ for any $1 \leq d \leq n$, and $H_A^{(1)}(d) = H_B^{(1)}(d) = 0$ for $d \geq n+1$. Also, $\Delta_A^{(1)} = y_{n+2,A} - y_{0,A} = 1$ and $\Delta_B^{(1)} = y_{n+2,B} - y_{0,B} = 1$. Therefore, we conclude that $U(A) = U(B)$. For the statistics associated to the structural parameters: $d_{1,A} = 0$ and $d_{n+3,A} = n+1$, such that $\Delta_A^{(1)}(n+1) = 1$ and $\Delta_A^{(1)}(d) = 0$ for any $d \neq n+1$; $d_{1,B} = 0$ and $d_{n+3,B} = 1$, such that $\Delta_B^{(1)}(d) = 0$ for any $d \geq 2$. Therefore, $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = [\Delta_A^{(1)}(n+1) - \Delta_B^{(1)}(n+1)]\ \gamma(n) = \gamma(n)$, and this structural parameter is identified. ■

For this model, the vector of sufficient statistics include the histogram of durations, $\{H^{(1)}(d) : d \geq 1\}$. Conditional on this statistics, the identification of the structural parameter $\gamma(d)$ comes from the difference between the final and the initial value of duration, $\Delta^{(1)}(d+1) = 1\{d_{T+1} = d+1\} - 1\{d_1 = d+1\}$. The identification result in Proposition 3 for the myopic model with duration dependence does not depend on Assumption 2.

In this binary choice model, the parameters $\widetilde{\beta}_y$ and $\beta_d(1,n)$ cannot be separately identified. However, given the parameters $\{\gamma(d) : 1 \leq d \leq T-2\}$, we can identify the marginal returns to experience $\beta_d(1,d) - \beta_d(1,d-1)$ as $\gamma(d) - \gamma(d-1)$ for any value $d$ between 2 and $T-2$.[21]

#### 3.2.4 Forward-looking dynamic model with duration dependence

Now, the optimal decision rule includes the difference of continuation values $v_\theta(1, d_t+1) - v_\theta(0)$. Therefore, the model is:

$$y_t = 1\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y\ y_{t-1} + \beta_d(1,d_t)\ y_{t-1} + v_\theta(1, d_t+1) + \widetilde{\varepsilon}_t \geq 0\right\} \tag{19}$$

where now $\widetilde{\alpha}_\theta \equiv \alpha_\theta(1) - \alpha_\theta(0) + \beta_y(1,0) - v_\theta(0)$. For this model, the log-probability of the choice history $\mathbf{y}^T$ conditional on $(y_0, d_1, \theta)$ is:

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta\right) = \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1} + \beta_d(1,d_t) y_{t-1} + v_\theta(1, d_t+1)\right] + \sigma_\theta(y_{t-1}, d_t) \tag{20}$$

[21] In this binary choice model with both switching costs and duration dependence, it is not possible to separately identify the switching cost parameter $\widetilde{\beta}_y$ and the level of the return to experience $\beta_d(1,d)$. This result resembles the under-identification of the autoregressive of the order two model studied by Chamberlain (1985). In that model, we have $y_{it} = 1\{\widetilde{\alpha}_i + \beta_1\ y_{i,t-1} + \beta_2\ y_{i,t-2} + \widetilde{\varepsilon}_{it} \geq 0\}$. Chamberlain shows that the parameter $\beta_2$ is identified but the parameter $\beta_1$ is not.

with $\sigma_\theta(y_{t-1}, d_t) \equiv -\ln(1+ \exp\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1} + \beta_d(1, d_t) y_{t-1} + v_\theta(1, d_t+1)\})$. Comparing equation (20) with (17) we can see the forward looking model has the additional term $\sum_{t=1}^{T} y_t \; v_\theta(1, d_t+1)$.

Proposition 4 establishes that under Assumption 1 only there is not identification of any structural parameter.

*PROPOSITION 4. In the forward-looking binary choice model with duration dependence under Assumption 1, the log-probability has the following form*

$$\ln \mathbb{P}\left(\mathbf{y}^T | y_0, \theta\right) = \sum_{d \geq 1} H^{(1)}(d) \; g_{\theta,1}(d) + \sum_{d \geq 1} \Delta^{(1)}(d) \; g_{\theta,2}(d) \tag{21}$$

*with* $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1, d) - \sigma_\theta(0) + \gamma(d-1) + v_\theta(1, d)$, $g_{\theta,2}(d) \equiv \widetilde{\alpha}_\theta + v_\theta(1, d) + \gamma(d-1)$, $\gamma(d) \equiv \widetilde{\beta}_y + \beta_d(1, d)$, *and* $\gamma(0) = 0$, *such that* $S = \{\Delta^{(1)}(d) : d \geq 1\}$ *and* $U = \{H^{(1)}(d) : d \geq 1, \Delta^{(1)}(d) : d \geq 1\}$. *The minimal sufficient statistic* $U$ *includes the whole vector* $S$*, and therefore, the structural parameters* $\gamma(d)$ *are not identified.* ■

The difference between the minimal sufficient statistic for this forward-looking model and for its myopic model version is that now we need to control for the difference between final and initial duration, $\Delta^{(1)}(d+1)$. These additional statistics are also the only statistics associated with the structural parameter $\gamma(d)$. Therefore, after controlling for the vector of sufficient statistics $U$, there is not variation left that can identify structural parameters in this model.

The under-identification result in Proposition 4 applies to the model under Assumption 1 but without Assumption 2. Under Assumption 2, continuation values are such that $v_\theta(1, d+1) = v_\theta(1, d^*)$ for any $d \geq d^* - 1$. This restriction provides identification of structural parameters. Proposition 5 establishes this result.

*PROPOSITION 5. In the forward-looking binary choice model with duration dependence under Assumptions 1 and 2, the log-probability has the following form*

$$\begin{aligned} \ln \mathbb{P}\left(\mathbf{y}^T | y_0, \theta\right) &= \sum_{d \leq d^*-1} H^{(1)}(d) \; g_{\theta,1}(d) + \left[\sum_{d \geq d^*} H^{(1)}(d)\right] g_{\theta,1}(d^*) \\ &+ \sum_{d \leq d^*-1} \Delta^{(1)}(d) \; g_{\theta,2}(d) + \left[\sum_{d \geq d^*} \Delta^{(1)}(d)\right] g_{\theta,2}(d^*) \\ &+ \Delta^{(1)}(d^*) \; [\beta_d(1, d^*-1) - \beta_d(1, d^*)] \end{aligned} \tag{22}$$

*with* $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1, d) - \sigma_\theta(0) + \gamma(d-1) + v_\theta(1, d)$, *and* $g_{\theta,2}(d) \equiv \widetilde{\alpha}_\theta + v_\theta(1, d) + \gamma(d-1)$. *Then, we have that: (i)* $U = \{H^{(1)}(d) : d \leq d^*-1, \sum_{d \geq d^*} H^{(1)}(d), \Delta^{(1)}(d) : d \leq d^*-1, \sum_{d \geq d^*} \Delta^{(1)}(d)\}$ *is* *a sufficient statistic for $\theta$. (ii) The elements in the vector $U$ are linearly independent such that $U$ is a minimal sufficient statistic. (iii) Conditional on $U$, the statistic $\Delta^{(1)}(d^*)$ has variation and the structural parameter $\Delta\beta_d(d^*) \equiv \beta_d(1, d^*) - \beta_d(1, d^* - 1)$ is identified, i.e., there is a pair of histories, $A$ and $B$, such that $U(A) = U(B)$ and parameter $\Delta\beta_d(d^*) = \ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)$.* ■

*Proof.* The derivation of equation (22) is in the Appendix. Proof of (iii). Given a choice history $\{y_0, d_1 \mid \mathbf{y}^T\}$ consider sub-histories $\{y_0, d_1 \mid \mathbf{y}^{2d^*+1}\}$. Consider the choice histories $A = \{0, 0 \mid \mathbf{1}_{d^*-1}, 0, \mathbf{1}_{d^*+1}\}$ and $B = \{0, 0 \mid \mathbf{1}_{d^*}, 0, \mathbf{1}_{d^*}\}$. The corresponding histories of durations $\{d_t : t = 1, ..., 2d^* + 1\}$ are: for $A$, $\{0,\ 1,\ 2, ...,\ d^* - 1,\ 0,\ 1,\ 2,\ ..., d^*\}$; and for $B$, $\{0,\ 1,\ 2,\ ...,\ d^*,\ 0,\ 1,\ 2,\ ...,\ d^* - 1\}$. We verify that $U(A) = U(B)$: (a) for any $d \leq d^* - 1$, $H_A(d) = H_B(d) = 2$; (b) $\sum_{d \geq d^*} H_A(d) = \sum_{d \geq d^*} H_B(d) = 1$; (c) for any $d \leq d^* - 1$, $\Delta_A^{(1)}(d) = \Delta_B^{(1)}(d) = 0$; and (d) $\sum_{d \geq d^*} \Delta_A^{(1)}(d) = \sum_{d \geq d^*} \Delta_B^{(1)}(d) = 1$. However, the two histories have different values for the statistic $\Delta^{(1)}(d^*)$, i.e., $\Delta_A^{(1)}(d^*) = 0$ and $\Delta_B^{(1)}(d^*) = 1$. Therefore, $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = [\Delta_A^{(1)}(d^*) - \Delta_B^{(1)}(d^*)]\ [-\Delta\beta_d(d^*)] = \Delta\beta_d(d^*)$. ■

In the forward-looking binary choice model with duration dependence, only $\Delta\beta_d(d^*)$ is identified. This result contrasts with the myopic model where we can identify $\Delta\beta_d(d)$ for any duration $2 \leq d \leq T - 1$ (Proposition 3). Table 2 summarizes the identification results for the binary choice model.

**Table 2**
**Identification of Dynamic Binary Logit Models**

**Panel 1: Models without duration dependence**

| *Myopic Model* | | | *Forward-Looking Model* | | |
|---|---|---|---|---|---|
| Minimal sufficient stat. | Identified parameters | Identifying statistics | Minimal sufficient stat. | Identified parameters | Identifying statistics |
| $\{T^{(1)}, \Delta^{(1)}\}$ | $\widetilde{\beta}_y$ | $D^{(1,1)}$ | $\{T^{(1)}, \Delta^{(1)}\}$ | $\widetilde{\beta}_y$ | $D^{(1,1)}$ |

**Panel 2: Models with duration dependence**

| *Myopic Model* | | | *Forward-Looking Model* | | |
|---|---|---|---|---|---|
| Minimal sufficient stat. | Identified parameters | Identifying statistics | Minimal sufficient stat. | Identified parameters | Identifying statistics |
| $\{\Delta^{(1)},$ $H^{(1)}(d) : d \geq 1\}$ | $\widetilde{\beta}_y + \beta_d(1, d)$ for $d \leq T-2$ | $\Delta^{(1)}(d)$ | $\{H^{(1)}(d) : d \leq d^*{-}1,$ $\sum_{d \geq d^*} H^{(1)}(d),$ $\Delta^{(1)}(d) : d \leq d^*{-}1,$ $\sum_{d \geq d^*} \Delta^{(1)}(d)\ \}$ | $\Delta\beta_d(d^*) \equiv$ $\beta_d(1, d^*)$ $-\beta_d(1, d^*{-}1)$ | $\Delta^{(1)}(d^*)$ |

*Identification of $d^*$ in the forward-looking model.* We have assumed so far that the value of $d^*$ is known to the researcher. We now establish the identification of $d^*$. Let $n$ be any duration such that $2n + 1 \leq T$. Consider the pair of histories $A_n = \{0, 0 \mid \mathbf{1}_{n-1}, 0, \mathbf{1}_{n+1}\}$ and $B_n = \{0, 0 \mid \mathbf{1}_n, 0, \mathbf{1}_n\}$. We have that:

$$\begin{aligned} &\text{For } n > d^*, \quad U(A_n) = U(B_n), \text{ and } \ln \mathbb{P}(A_n|U) - \ln \mathbb{P}(B_n|U) = \Delta\beta_d(n) = 0 \\ &\text{For } n = d^*, \quad U(A_n) = U(B_n), \text{ and } \ln \mathbb{P}(A_n|U) - \ln \mathbb{P}(B_n|U) = \Delta\beta_d(d^*) \neq 0 \\ &\text{For } n < d^*, \quad U(A_n) \neq U(B_n) \end{aligned} \tag{23}$$

Note that $\ln \mathbb{P}(A_n|U_n) - \ln \mathbb{P}(B_n|U_n)$ identifies the parameter $\Delta\beta_d(n)$ only if $n \geq d^*$. Given a dataset with $T$ time periods, we can construct histories $A_n$ and $B_n$ only if $2n + 1 \leq T$. Putting these two conditions together, the identification of the value of $d^*$ requires that $T \geq 2d^* + 1$ or equivalently, $d^* \leq (T-1)/2$. Under this condition, we can describe the parameter $d^*$ as the maximum value of $n$ such that $\ln \mathbb{P}(A_n|U_n) - \ln \mathbb{P}(B_n|U_n) \neq 0$. This condition uniquely identifies $d^*$.

*PROPOSITION 6. Consider the forward-looking binary choice model with duration dependence* *under Assumptions 1 and 2. For any duration $n$ with $2n+1 \leq T$, define the pair of histories $A_n = \{0, 0 \mid \mathbf{1}_{n-1}, 0, \mathbf{1}_{n+1}\}$ and $B_n = \{0, 0 \mid \mathbf{1}_n, 0, \mathbf{1}_n\}$. Then, if $d^* \leq (T-1)/2$, we have that the value of $d^*$ is point identified as:*

$$d^* = \max\left\{n : \ \ln \mathbb{P}\left(A_n | U_n\right) - \ln \mathbb{P}\left(B_n | U_n\right) \neq 0\right\} \qquad \blacksquare \tag{24}$$

## 3.3 Multinomial choice models

### 3.3.1 Multinomial myopic model without duration dependence

We can represent this model as $y_t = \arg\max_{y \in \mathcal{Y}}\{\alpha_\theta(y) + \beta_y(y, y_{t-1}) + \varepsilon_t(y)\}$. The log-probability of the choice history $\mathbf{y}^T$ conditional on $(y_0, \theta)$ is:

$$\ln \mathbb{P}\left(\mathbf{y}^T | y_0, \theta\right) \ = \ \sum_{t=1}^{T} \left[\alpha_\theta(y_t) + \beta_y(y_t, y_{t-1})\right] + \sigma_\theta(y_{t-1}) \tag{25}$$

where $\sigma_\theta(y_{t-1}) \equiv -\ln\left[\sum_{y=0}^{J} \exp\{\alpha_\theta(y) + \beta_y(y, y_{t-1})\}\right]$. Proposition 7 presents our identification result for this model.

*PROPOSITION 7. In the myopic multinomial model without duration dependence under Assumption 1, the log-probability has the following form*

$$\ln \mathbb{P}\left(\mathbf{y}^T | y_0, \theta\right) \ = \ \sum_{y=1}^{J} T^{(y)} \ g_{\theta,1}(y) + \sum_{y=1}^{J} \Delta^{(y)} \ g_{\theta,2}(y) + \sum_{y_{-1}=1}^{J}\sum_{y=1}^{J} D^{(y_{-1},y)} \ \widetilde{\beta}_y(y, y_{-1}) \tag{26}$$

*where $g_{\theta,1}(y) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \sigma_\theta(y) - \sigma_\theta(0) + \beta_y(0, y) + \beta_y(y, 0)$, $g_{\theta,2}(y) \equiv -\sigma_\theta(y) + \sigma_\theta(0) - \beta_y(0, y)$, $\widetilde{\beta}_y(y, y_{-1}) \equiv \beta_y(y, y_{-1}) - \beta_y(0, y_{-1}) - \beta_y(y, 0)$ for any $y, y_{-1} \in \mathcal{Y}$, and $\Delta^{(y)} \equiv 1\{y_T = y\} - 1\{y_0 = y\}$. Then: (i) $U = \{T^{(y)} : y \geq 1, \ \Delta^{(y)} : y \geq 1\}$ is a sufficient statistic for $\theta$. (ii) The elements in the vector $U$ are linearly independent such that $U$ is a minimal sufficient statistic. (iii) Conditional on $U$, the vector of statistics $\{D^{(y_{-1},y)} : y_{-1}, y \in \mathcal{Y} - \{0\}\}$ are linearly independent such that they can identify the vector of parameters $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1}, y \in \mathcal{Y} - \{0\}\}$, i.e., for every pair of choices $y_{-1}, y \in \mathcal{Y} - \{0\}$, there is a pair of histories, $A$ and $B$, such that $U(A) = U(B)$ and parameter $\delta(y, y_{-1}) = [\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)] / \ [D_A^{(y_{-1},y)} - D_B^{(y_{-1},y)}]$.* $\blacksquare$

As in the binary choice model, we cannot identify the whole switching cost function $\beta_y$. With $J+1$ choice alternatives, we can identify $J^2$ switching cost parameters. However, the structural parameter $\widetilde{\beta}_y(y, y_{-1}) \equiv \beta_y(y, y_{-1}) - \beta_y(0, y_{-1}) - \beta_y(y, 0)$ has a clear interpretation: it is different between the cost of a *direct switch* from $y_{-1}$ to $y$ and an *indirect switch* via alternative 0. For this identification result, there is nothing special with alternative 0 and we could choose any other alternative as the baseline. Note also that the set of identified structural parameters $\widetilde{\beta}_y(y, y_{-1})$ includes the *sunk cost of entry* in market $y$, i.e., for any $y > 1$, $\widetilde{\beta}_y(y, y) = -\beta_y(0, y) - \beta_y(y, 0)$, because $\beta_y(y, y) = 0$.

The following example illustrates a pair of histories that identifies $\widetilde{\beta}_y(y, y_{-1})$.

*EXAMPLE 1.* Suppose that $T = 3$ and consider the following two realizations of the history $(y_0|\mathbf{y}^T)$: $A = \{0 \mid 0, j, k\}$ and $B = \{0 \mid j, 0, k\}$ with $j, k \neq 0$. We first confirm that $U(A) = U(B)$: $T_A^{(j)} = T_B^{(j)} = 1$, $T_A^{(k)} = T_B^{(k)} = 1$, and $T_A^{(y)} = T_B^{(y)} = 0$ for any $y \neq 0, j, k$; and $T_A^{(k)} = T_B^{(k)} = 1$, and $T_A^{(y)} = T_B^{(y)} = 0$ for any $y \neq 0, k$. The identifying statistics $D^{(y_{-1}, y)}$ take the following values: $D_A^{(j,k)} - D_B^{(j,k)} = 1$, $D_A^{(j,0)} - D_B^{(j,0)} = -1$, $D_A^{(0,k)} - D_B^{(0,k)} = -1$, and $D_A^{(y_{-1},y)} - D_B^{(y_{-1},y)} = 0$ for any other pair $(y_{-1}, y)$. Therefore, we have that $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \widetilde{\beta}_y(k, j) - \widetilde{\beta}_y(0, j) - \widetilde{\beta}_y(k, 0) = \widetilde{\beta}_y(k, j)$. A particular case of this example is when $j = k$, such that $A = \{0 \mid 0, j, j\}$ and $B = \{0 \mid j, 0, j\}$. In this case, $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)$ identifies $\widetilde{\beta}_y(j, j)$ that is equal to the sunk cost $-\beta_y(0, y) - \beta_y(y, 0)$. ■

### 3.3.2 Multinomial forward-looking model without duration dependence

The optimal decision rule for this model is $y_t = \arg\max_{y \in \mathcal{Y}} \{\alpha_\theta(y) + v_\theta(y) + \beta_y(y, y_{t-1}) + \varepsilon_t(y)\}$, where $v_\theta(y)$ is the continuation value of choosing alternative $y$. The log-probability of the choice history $\mathbf{y}^T$ conditional on $(y_0, \theta)$ has a similar form as in the myopic model, but now the incidental parameter $\theta$ enters through the function $\alpha_\theta(y) + v_\theta(y)$.

$$\ln \mathbb{P}\left(\mathbf{y}^T|y_0, \theta\right) = \sum_{t=1}^{T} [\alpha_\theta(y_t) + v_\theta(y_t) + \beta_y(y_t, y_{t-1})] + \sigma_\theta(y_{t-1}) \tag{27}$$

Therefore, the identification of the structural parameters is the same as in the myopic model without duration dependence.

*PROPOSITION 8. In the multinomial forward-looking model without duration dependence under Assumption 1, the log-probability has the following form*

$$\ln \mathbb{P}\left(\mathbf{y}^T|y_0, \theta\right) = \sum_{y=1}^{J} T^{(y)} \; g_{\theta,1}(y) + \sum_{y=1}^{J} \Delta^{(y)} \; g_{\theta,2}(y) + \sum_{y_{-1}=1}^{J} \sum_{y=1}^{J} D^{(y_{-1},y)} \; \widetilde{\beta}_y(y, y_{-1}) \tag{28}$$

*where* $g_{\theta,1}(y) \equiv \alpha_\theta(y) - \alpha_\theta(0) + v_\theta(y) - v_\theta(0) + \sigma_\theta(y) - \sigma_\theta(0) + \beta_y(0, y) + \beta_y(y, 0)$, *and* $g_{\theta,2}(y) \equiv \sigma_\theta(y) - \sigma_\theta(0) - \beta_y(0, y)$. *Then: (i)* $U = \{T^{(y)} : y \geq 1,\ \Delta^{(y)} : y \geq 1\}$ *is a sufficient statistic for*

*$\theta$. (ii) The elements in the vector $U$ are linearly independent such that $U$ is a minimal sufficient statistic. (iii) Conditional on $U$, the vector of statistics $\{D^{(y_{-1},y)} : y_{-1}, y \in \mathcal{Y} - \{0\}\}$ are linearly independent such that they can identify the vector of parameters $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1}, y \in \mathcal{Y} - \{0\}\}$.* ■

### 3.3.3 Multinomial myopic model with duration dependence

The model is $y_t = \arg\max_{y\in\mathcal{Y}}\{\alpha_\theta(y) + 1\{y \neq y_{t-1}\}\ \beta_y(y, y_{t-1}) + 1\{y = y_{t-1}\}\ \beta_d(y, d_t) + \varepsilon_t(y)\}$, and the log-probability of a choice history $\mathbf{y}^T$ conditional on $(y_0, d_1, \theta)$ is:

$$\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1, \theta\right) = \sum_{t=1}^{T} \left[\alpha_\theta(y_t) + 1\{y_t \neq y_{t-1}\}\beta_y(y_t, y_{t-1}) + 1\{y_t = y_{t-1}\}\beta_d(y_t, d_t)\right] + \sigma_\theta(y_{t-1}) \tag{29}$$

Proposition 9 presents identification results for the structural parameters $\beta_y$ and $\beta_d$.

*PROPOSITION 9. In the multinomial myopic model with duration dependence under Assumption 1, the log-probability $\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta, \beta\right)$ has the form*

$$\begin{aligned} \ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1\right) &= \sum_{y=1}^{J}\sum_{d\geq 1} H^{(y)}(d)\ g_{\theta,1}(y, d) + \sum_{y=1}^{J} \Delta^{(y)}\ g_{\theta,2}(y) \\ &+ \sum_{y_{-1}=1}^{J}\ \sum_{y=1, y\neq y_{-1}}^{J} D^{(y_{-1},y)}\ \widetilde{\beta}_y(y, y_{-1}) + \sum_{y=1}^{J}\sum_{d\geq 1} \Delta^{(y)}(d)\ \gamma(y, d-1) \end{aligned} \tag{30}$$

*with $g_{\theta,1}(y, d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \sigma_\theta(y, d) - \sigma_\theta(0) + \beta_y(0, y) + \beta_y(y, 0) + \gamma(y, d-1)$, $g_{\theta,2}(y) \equiv \alpha_\theta(y) + \beta_y(y, 0)$, $\widetilde{\beta}_y(y, y_{-1}) \equiv \beta_y(y, y_{-1}) - \beta_y(0, y_{-1}) - \beta_y(y, 0)$, and $\gamma(y, d) \equiv \beta_d(y, d) - \beta_y(y, 0) - \beta_y(0, y)$. Then: (i) $U = \{H^{(y)}(d) : y \geq 1,\ d \geq 1,\ \Delta^{(y)} : y \geq 1\}$ is a sufficient statistic of $\theta$. (ii) The elements in the vector $U$ are linearly independent such that $U$ is a minimal sufficient statistic. (iii) Conditional on $U$, the vector of statistics $\{D^{(y_{-1},y)} : y_{-1},\ y \geq 1;\ \Delta^{(y)}(d) : y \geq 1,\ d \geq 1\}$ are linearly independent and they identify the vectors of structural parameters $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1},\ y \geq 1,$ $y \neq y_{-1};\ \gamma(y, d) : y \geq 1,\ d \geq 1\}$.* ■

The following examples present choice histories that identify structural parameters $\widetilde{\beta}_y(y, y_{-1})$ and $\gamma(y, d)$ according to Proposition 9.

*EXAMPLE 2.* Suppose that $T = 3$ and consider two realizations of the history $\left(y_0, d_1 | \mathbf{y}^T\right)$: for $j \neq k$, $A = \{0, 0 \mid 0, j, k\}$ and $B = \{0, 0 \mid j, 0, k\}$. It is straightforward to verify that $U(A) = U(B)$ and that $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \beta_y(k, j) - \beta_y(k, 0) - \beta_y(0, j) = \widetilde{\beta}_y(k, j)$. ■

*EXAMPLE 3.* Given an arbitrary positive integer $n$, consider the pair of choice histories $\left(y_0, d_1 | \mathbf{y}^T\right)$ with $T = n + 2$: $A = \{0, 0 \mid 0, \mathbf{y}_{n+1}\}$ and $B = \{0, 0 \mid \mathbf{y}_n, 0, y\}$, where $\mathbf{y}_n$ represents a vector of dimension $n$ with all its elements equal to $y$. It is simple to verify that $U(A) = U(B)$ (i.e., same values for $H^{(y)}(d)$ and $\Delta^{(y)}$). Furthermore, $\Delta_A^{(y)}(n+1) = 1$ and $\Delta_A^{(y)}(n+1) = 0$, and we have that $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \gamma(y, n)$. ■

### 3.3.4 Multinomial forward-looking model with duration dependence

The model is $y_t = \arg\max_{y \in \mathcal{Y}} \{\alpha_\theta(y) + \beta_y(y, y_{t-1}) + 1\{y = y_{t-1}\}\beta_d(y, d_t) + v_\theta(y, d_{t+1}[y, y_{t-1}, d_t]) + \varepsilon_t(y)\}$, where $d_{t+1}[y, y_{t-1}, d_t] = 0$ if $y = 0$, and $d_{t+1}[y, y_{t-1}, d_t] = 1\{y = y_{t-1}\}d_t + 1$ if $y \neq 0$. In contrast to the binary choice model, in the multinomial choice model it is possible to identify switching cost parameters without imposing Assumption 2. Proposition 10 establishes the identification of switching costs parameters under Assumption 1.

*PROPOSITION 10. In the multinomial forward-looking model with duration dependence under Assumption 1, the log-probability* $\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta, \beta\right)$ *has the form*

$$\begin{aligned} \ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1\right) &= \sum_{y=1}^{J} \sum_{d \geq 1} H^{(y)}(d)\ g_{\theta,1}(y, d) + \sum_{y=1}^{J} \sum_{d \geq 1} \Delta^{(y)}(d)\ g_{\theta,2}(y, d) \\ &+ \sum_{y_{-1}=1}^{J} \sum_{y=1}^{J} D^{(y_{-1}, y)}\ \widetilde{\beta}_y(y, y_{-1}) \end{aligned} \tag{31}$$

*with* $g_{\theta,1}(y, d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \sigma_\theta(y, d) - \sigma_\theta(0) + \beta_y(0, y) + \beta_y(y, 0) + v_\theta(y, d) - v_\theta(0) + \gamma(y, d-1)$, *and* $g_{\theta,2}(y, d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \beta_y(0, y) + v_\theta(y, d) + \gamma(y, d-1)$. *Then: (i)* $U = \{H^{(y)}(d) : y \geq 1, d \geq 1, \Delta^{(y)}(d) : y \geq 1, d \geq 1\}$ *is a sufficient statistic of* $\theta$. *(ii) The elements in the vector* $U$ *are linearly independent such that* $U$ *is a minimal sufficient statistic. (iii) Conditional on* $U$, *the vector of statistics* $\{D^{(y_{-1}, y)} : y_{-1}, y \geq 1\}$ *are linearly independent and they identify the vectors of structural parameters* $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1}, y \geq 1\}$. *The duration dependence parameters* $\gamma(y, d)$ *are not identified.* ■

For instance, the pair of choice histories in Example 2, $A = \{0, 0 \mid 0, j, k\}$ and $B = \{0, 0 \mid j, 0, k\}$, are such the have the same history of continuation values. In this forward-looking model, it is simple to very that these histories satisfy the conditions in Proposition 10 such that $U(A) = U(B)$ and $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \widetilde{\beta}_y(k, j)$.

For the identification of duration dependence parameters, we impose the restriction in Assumption 2. Proposition 11 presents this identification result.

*PROPOSITION 11. In the multinomial forward-looking model with duration dependence under Assumptions 1 and 2, the log-probability* $\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta, \beta\right)$ *has the form*

$$\begin{aligned}
\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1\right) \quad = \quad & \sum_{y=1}^{J} \sum_{d \leq d_y^* - 1} H^{(y)}(d)\ g_{\theta,1}(y,d) + \left[\sum_{y=1}^{J} \sum_{d \geq d_y^*} H^{(y)}(d)\right] g_{\theta,1}(y, d_y^*) \\
& \sum_{y=1}^{J} \sum_{d \leq d_y^* - 1} \Delta^{(y)}(d)\ g_{\theta,2}(y,d) + \left[\sum_{y=1}^{J} \sum_{d \geq d_y^*} \Delta^{(y)}(d)\right] g_{\theta,2}(y, d_y^*) \\
+ \quad & \sum_{y_{-1}=1}^{J} \sum_{y=1, y \neq y_{-1}}^{J} D^{(y_{-1},y)}\ \widetilde{\beta}_y(y, y_{-1}) - \sum_{y=1}^{J} \Delta^{(y)}(d_y^*)\ \Delta\beta_d(y, d_y^*)
\end{aligned} \tag{32}$$

*with* $g_{\theta,1}(y,d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \sigma_\theta(y,d) - \sigma_\theta(0) + \beta_y(0,y) + \beta_y(y,0) + v_\theta\left(y,d\right) - v_\theta\left(0\right) + \gamma(y, d-1)$*, and* $g_{\theta,2}(y,d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \beta_y(y,0) + v_\theta\left(y,d\right) - v_\theta\left(0\right) + \gamma(y, d-1)$*, and* $\Delta\beta_d(y, d_y^*) \equiv \beta_d(y, d_y^*) - \beta_d(y, d_y^* - 1)$*. (i)* $U = \{H^{(y)}(d) : y \geq 1,\ d \leq d_y^* - 1,\ \sum_{d \geq d_y^*} H^{(y)}(d),\ \Delta^{(y)}(d) : y \geq 1,$ $d \leq d_y^* - 1,\ \sum_{d \geq d_y^*} \Delta^{(y)}(d)\}$ *is a sufficient statistic of* $\theta$*. (ii) The elements in the vector* $U$ *are linearly independent such that* $U$ *is a minimal sufficient statistic. (iii) Conditional on* $U$*, the vector of statistics* $\{D^{(y_{-1},y)} : y_{-1},\ y \geq 1\}$ *are linearly independent and they identify the vector of structural parameters* $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1},\ y, y \neq y_{-1} \geq 1\}$*. Furthermore, the vector of statistics* $\{\Delta^{(y)}(d^*) : y \geq 1\}$ *are also linearly independent and they identify the vector of structural parameters* $\{\Delta^{(y)}(d^*) : y \geq 1\}$*.* ■

*EXAMPLE 4.* Given $y \geq 1$ with $d_y^* \geq 2$, consider the pair of choice histories $A = \{0, 0 \mid \mathbf{y}_{d^*-1}, 0, \mathbf{y}_{d^*+1}\}$ and $B = \{0, 0 \mid \mathbf{y}_{d^*}, 0, \mathbf{y}_{d^*}\}$. The two choice histories have the same statistics $H^{(y)}(d)$ for all $1 \leq d \leq d_y^* - 1$ and $\sum_{d \geq d_y^*} H^{(y)}(d)$, and $\min\{d_1, d^*\}$ and $\min\{d_{T+1}, d^*\}$ agrees between $A$ and $B$. Therefore, we have that $U(A) = U(B)$. It is straightforward to show that $\Delta_A^{(y)}(d_y^*) = 0$ and $\Delta_B^{(y)}(d_y^*) = 1$, and this implies that $\ln \mathbb{P}\left(A|U\right) - \ln \mathbb{P}\left(B|U\right) = \Delta\beta_d(y, d_y^*)$. ■

## 4 Estimation and Inference

Since the identification is based on the conditional MLE approach, the estimator for the structural parameters of interest $(\beta_y, \beta_d)$ will be an Anderson (1970) type of estimator and the inference is standard. We illustrate the estimator for the forward-looking multinomial choice model with duration dependence under Assumption 1 and 2, since estimators for the structural parameters in the other models can be defined in a similar fashion.

Let $\beta^* = \{\widetilde{\beta}_y', \gamma'\}'$ be the vector of identified structural parameters (or linear combinations of the original structural parameters). Let $U_i \equiv U(y_{i0}, d_{i1}, \mathbf{y}_i^T)$ be the vector of sufficient statistics (associated to $\theta$), and let and $S_i \equiv S(y_{i0}, d_{i1}, \mathbf{y}_i^T)$ be the vector of identifying statistics, associated to $\beta^*$. Then, the conditional MLE for $\beta^*$is defined as the maximizer of the conditional log-likelihood function:

$$\ell^c(\beta^*) = \sum_{i=1}^{N} \ell_i^c(\widehat{\beta}^*) = \sum_{i=1}^{N} S_i'\beta^* - \left( \sum_{j:U(j)=U_i} \exp\left\{S(j)'\beta^*\right\} \right) \tag{33}$$

where the condition $\{j : U(j) = U_i\}$ represents all the choice histories $(y_0, d_1, \mathbf{y}^T)$ with the sample value of $U$ as the history in observation $i$. This log-likelihood function is globally concave in $\beta^*$, and therefore the computation of the CMLE is straightforward using Newton-Raphson or BHHH algorithm. Using standard arguments (Newey and McFadden (1994)), we have

$$\sqrt{N}(\widehat{\beta}^* - \beta^*) \Rightarrow \mathcal{N}(0, J(\beta^*)^{-1})$$

The consistent estimator for the Fisher information is $J_N(\widehat{\beta}^*) = -N^{-1} \sum_{i=1}^{N} \nabla_{\beta\beta} \ell_i^c(\widehat{\beta}^*)$.

For the estimation of the parameter $d^*$ we consider a sequential testing procedure. We start with the highest value of duration for which we can estimate consistently parameter $\Delta\beta_d(n) \equiv \beta_d(n) - \beta_d(n-1)$. Under the assumption that $d^* \leq (T-1)/2$, this maximum value of duration if $n = int[(T-1)/2]$. We estimate $\Delta\beta_d(n)$ and test for the null hypothesis that this parameter estimate is equal to zero. If we reject the null hypothesis, then we conclude that $d^* = n$. Otherwise, we conclude that $d^* < n$ and then apply the same procedure to duration $n - 1$. The procedure stops when we find a value of duration $n$ for which we reject the null hypothesis $\Delta\beta_d(n) = 0$. If we cannot reject the null hypothesis for any duration $n$, including $n = 1$, we conclude that $\Delta\beta_d(n) = 0$ for any $n \geq 1$, i.e., the model does not have duration dependence (i.e., $\beta_d(y, n)$ is constant and equal to $\beta_d(y, 0)$).

For the implementation of this sequential testing procedure, a key decision for the researcher is the choice of the significance level for testing the null hypothesis $\Delta\beta_d(n) = 0$. As we show in the Monte Carlo experiments in section 5.2, this choice has important implications for the properties of the estimator of $d^*$ and, importantly, for the estimator of the structural parameters $\beta^*$. In general, a higher significance level (i.e., higher probability of rejecting the null) implies that we stop the procedure at a higher estimate for $d^*$.

# 5 Empirical Application

Here we revisit the model and data in the seminal article by Rust (1987). The model belongs to the class of *machine replacement models* that we have briefly described in section 2. The superintendent of maintenance at the Madison (Wisconsin) Metropolitan Bus Company has a fleet of $N$ buses indexed by $i$. For every bus $i$ and at every period $t$, the superintendent decides whether to keep the bus engine ($y_{it} = 1$) or to replace it ($y_{it} = 0$). In Rust's model, if the engine is replaced, the payoff is equal to $-RC + \varepsilon_{it}(0)$, where $RC$ is a parameter that represents the replacement cost. If the manager decides to keep the engine, the payoff is equal to $-c_0 - c_1(m_{it}) + \varepsilon_{it}(1)$, where $m_{it}$ is a state variable that represents the engine cumulative mileage, and $c_0 + c_1(m_{it})$ is the maintenance cost. We incorporate two modifications in this model. First, we replace cumulative mileage $m_{it}$ with duration since last replacement, $d_{it}$. The transition rule for this state variable is $d_{it+1} = y_{it}[d_{it} + 1]$, such that $d_{it} \in \{0, 1, 2, ...\}$. Using Rust's actual data, the correlation between the variables $m_{it}$ and $d_{it}$ is 0.9552. Second, we allow for time-invariant unobserved heterogeneity in the replacement cost, $RC_i$, and in the constant term in the maintenance cost function, $c_{0i}$. Using our notation, the payoff function is $\alpha_i(0) + \varepsilon_{it}(0)$ if $y_{it} = 0$ (replacing the engine), and $\alpha_i(1) + \beta_d(d_{it}) + \varepsilon_{it}(1)$ if $y_{it} = 1$ (keeping the engine), where $\alpha_i(0) = -RC_i$, $\alpha_i(1) = -c_{0i}$, and $\beta_d(d_{it}) = -c_1(d_{it})$.

In section 5.1, we present evidence from several Monte Carlo experiments using this model. The purpose of these experiments is threefold. First, showing that the FE-CMLE provides precise and robust estimates of structural parameters, even when the sample size is not large. Second, showing that the bias of misspecifying the distribution of the unobserved heterogeneity. And third, showing that the Hausman test, based on the comparison of the FE-CMLE and a CRE-MLE, has enough power to reject specifications that wrongly ignore unobserved heterogeneity, or that misspecified its probability distribution or its joint distribution with the initial conditions of the state variables.

In section 5.2 we use our Monte Carlo experiments to investigate the finite-sample bias associated to the pre-estimation (or pre-testing) of the parameter $d^*$. Finally, in section 5.3, we apply the FE-CMLE method, our procedure to estimate $d^*$, and the Hausman test to the actual dataset in Rust (1987).

## 5.1 Monte Carlo experiments

We present experiments using simulated data from four different Data Generating Processes (DGPs). Table 3 describes these DGPs. The difference between the four DGPs is in the specification of the distribution of the unobserved heterogeneity for the replacement cost $RC_i$. In DGP 1, the distribution of the replacement cost is normal with mean 8 and standard deviation 2. In DGPs 2 and 3, this distribution has only two types. Finally, DGP 4 is a model without unobserved heterogeneity.

**Table 3**
**Description of DGPs in the Monte Carlo experiments**

| *Parameter / Constant* | *DGP 1* | *DGP 2* | *DGP 3* | *DGP 4* |
|---|---|---|---|---|
| $\alpha_i(0) = -RC_i$ | $N(\mu, \sigma^2)$ | Two types | Two types | 1 type |
| Random draws from: | $\mu = 8, \sigma = 2$ | $RC_1 = 4.5, RC_2 = 9$ | $RC_1 = 8, RC_2 = 9$ | $RC = 8$ |
| | | $\lambda_1 = \lambda_2 = 0.5$ | $\lambda_1 = \lambda_2 = 0.5$ | |
| $\alpha_i(1) = -c_{0i}$ | 0 | 0 | 0 | 0 |
| $\beta_d(d) = \beta\ d$ if $d \leq d^*$ | $\beta = 1$ | $\beta = 1$ | $\beta = 1$ | $\beta = 1$ |
| $d^*$ | 3 | 3 | 3 | 3 |
| *Discount factor* ($\delta$) | 0.95 | 0.95 | 0.95 | 0.95 |
| *Initial* $y_0, d_1$ | $0, 0$ | $0, 0$ | $0, 0$ | $0, 0$ |
| *Maximum* $T$ | 25 | 25 | 25 | 25 |
| *N (number of buses)* | 1000 | 1000 | 1000 | 1000 |
| *# simulated samples* | 1000 | 1000 | 1000 | 1000 |

For each of these DGPs, we do not estimate the model using the whole sample of $T = 25$ periods. Instead, we construct three samples: sample A, from period 1 to 7; sample B, from period 1 to 14; and Sample C, from period 8 to 21. Therefore, we present results from 12 Monte Carlo experiments, i.e., four DGPs times 3 samples. We analyze the effect of increasing the number of time periods $T$, by comparing the experiments with sample A (with $T = 7$) and sample B (with $T = 14$). We study the effect of the initial conditions problem by comparing the experiments for sample B (where at $t = 0$ all the buses have the same initial condition, $(y_{i0}, d_{i1}) = (0, 0)$) and sample C, that is subject to the initial conditions problem.

The structural parameter of interest is parameter $\beta$ in the maintenance cost function, $\beta_d(d) = \beta\ d$. We apply three estimators to each of the samples: the *FE-CMLE*, an MLE that imposes the restriction of no unobserved heterogeneity (that we denote as *MLE-noUH*), and an MLE that assumes that there are two types of replacement costs and ignores the potential initial conditions problem (that we denote as *MLE-2types*). We compare the bias and variance of these estimators.[22] We also implement two Hausman tests: a test of the null hypothesis of no unobserved heterogeneity, that compares estimators *FE-CMLE* and *MLE-noUH*; and a test of the null hypothesis of two-types, that compares estimators *FE-CMLE* and *MLE-2types*.

**Table 4**
**Monte Carlo Experiments with DGP 1 (Normal RCs)**

| Estimator of $\beta$ | **Sample A** ($t = 1$ to 7) Estimate$^{(1)}$ Mean | Median | St. dev. | **Sample B** ($t = 1$ to 14) Estimate$^{(1)}$ Mean | Median | St. dev. | **Sample C** ($t = 8$ to 21) Estimate$^{(1)}$ Mean | Median | St. dev. |
|---|---|---|---|---|---|---|---|---|---|
| FE-CMLE | 0.9996 | 0.9924 | 0.1481 | 1.0021 | 1.0017 | 0.0741 | 0.9958 | 0.9936 | 0.0716 |
| MLE-2types | 0.9625 | 0.9612 | 0.0478 | 0.8955 | 0.8955 | 0.0265 | 0.8687 | 0.8688 | 0.0289 |
| MLE-noUH | 0.6175 | 0.6173 | 0.0277 | 0.5922 | 0.5920 | 0.0210 | 0.5653 | 0.5654 | 0.0213 |
| **Testing null hypothesis** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** |
| No Unob. Het. | 0.504 | 0.753 | 0.854 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Two types | 0.007 | 0.046 | 0.095 | 0.138 | 0.348 | 0.466 | 0.244 | 0.481 | 0.610 |

Note (1): Mean, Median, and Standard deviation of estimated parameter over the distribution of 1,000 replications.

We present the results of these experiments in tables 4 to 7, one table for each DGP. Table 4 deals with DGP 1, with normally distributed replacement costs. The two MLE estimators are substantially biased, and their bias is particularly large in sample $C$ (with the initial conditions problem) and sample $B$ (with large $T$ such that there are multiple spells per bus, and therefore, more room for stronger correlation between durations and unobserved heterogeneity). Instead, the bias of the *FE-CMLE* is negligible. As expected, the variance of the CMLE is larger than the variances of the MLEs. However, the Mean Square Error (MSE, variance plus square bias) of the

[22] The code for this experiment is in Matlab. For the two ML estimators, we use the Nested Fixed Point Algorithm. The maximization of the log-likelihood function applies a quasi-newton method (procedure `fminunc`) using the true value of the vector of parameters as the starting value. For the MLE with 2-types, during the search algorithm we often get a singular Hessian matrix. When this happens, we switch to the BHHH method.

CMLE is substantially smaller than the one of the *MLE-noUH* in the three samples, and of the *MLE-2types* in samples B and C. Interestingly, in sample A the *MLE-2types* has a MSE comparable to the one of the *FE-CMLE.* That is, without initial conditions problem and one duration spell for most of the buses, a misspecified random effects model with only two types has good properties. However, this is not longer the case in samples B and C. Hausman test has very strong power to reject the model without unobserved heterogeneity. It has also substantial power to reject the model with two types in samples B and C. However, the rejection rates for the model with two types in sample A are practically equal to the nominal size or significance level of the test. In sample A, the model with two types is almost observationally equivalent to the true model.

It is important to note that in DGP 1, though the distribution of types is continuous, the level of unobserved heterogeneity is modest. In the distribution of the replacement costs $RC_i$, the ratio between the standard deviation and the mean (coefficients of variation) is only 25%. Continuous distributions with higher variance imply higher rejection rates for the two types model, even in sample A.

**Table 5**
**Monte Carlo Experiments with DGP 2 (Two types: RC = 4.5, 9)**

| Estimator of $\beta$ | **Sample A** ($t = 1$ to 7) Estimate[(1)] Mean | Median | St. dev. | **Sample B** ($t = 1$ to 14) Estimate[(1)] Mean | Median | St. dev. | **Sample C** ($t = 8$ to 21) Estimate[(1)] Mean | Median | St. dev. |
|---|---|---|---|---|---|---|---|---|---|
| FE-CMLE | 1.0122 | 1.0078 | 0.1653 | 1.0043 | 1.0055 | 0.0805 | 1.0028 | 0.9993 | 0.0784 |
| MLE-2types | 1.0040 | 1.0019 | 0.0477 | 1.0003 | 0.9991 | 0.0251 | 0.9948 | 0.9946 | 0.0261 |
| MLE-noUH | 0.5606 | 0.5603 | 0.0221 | 0.5429 | 0.5428 | 0.0145 | 0.5259 | 0.5255 | 0.0143 |
| **Testing null hypothesis** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** |
| No Unob. Het. | 0.555 | 0.797 | 0.874 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Two types | 0.009 | 0.046 | 0.088 | 0.009 | 0.050 | 0.099 | 0.006 | 0.051 | 0.103 |

Note (1): Mean, Median, and Standard deviation of estimated parameter over the distribution of 1,000 replications.

Table 5 presents results under DGP 2, with two types of replacement costs, $RC_1 = 4.5$ and $RC_2 = 9$, with equal probabilities. In this case, both the *MLE-2types* and our *FE-CMLE* are consistent estimators. Both estimators have negligible finite-sample biases in the three samples. As expected, the *MLE-2types* has smaller variance, especially in sample A. In the three samples, the *MLE-noUH* is still extremely biased and Hausman test has strong power to reject the model without unobserved heterogeneity. For the rejection of the true model with two types, Hausman test exhibits a rejection rate that is practically identical to the nominal size or significance level.

Table 6 deals with DGP 3, that has also two types of replacement costs, but now these types are very similar: $RC_1 = 8$ and $RC_2 = 9$, with equal probabilities. The main purpose of the experiments with this DGP is to investigate the bias of the *MLE-noUH* and the power of this Hausman test in an scenario with a very modest amount of unobserved heterogeneity. Even in this scenario, the bias of the *MLE-noUH* is approximately 4% and Hausman test rejects the null hypothesis of no unobserved heterogeneity with probability that is more than twice the nominal size of the test.

**Table 6**
**Monte Carlo Experiments with DGP 3 (Two types: RC = 8, 9)**

| Estimator of $\beta$ | **Sample A** ($t = 1$ to 7) Estimate[1] Mean | Median | St. dev. | **Sample B** ($t = 1$ to 14) Estimate[1] Mean | Median | St. dev. | **Sample C** ($t = 8$ to 21) Estimate[1] Mean | Median | St. dev. |
|---|---|---|---|---|---|---|---|---|---|
| FE-CMLE | 1.0013 | 0.9976 | 0.1307 | 1.0023 | 1.0034 | 0.0652 | 1.0001 | 1.0027 | 0.0639 |
| MLE-2types | 1.0079 | 1.0058 | 0.0491 | 1.0013 | 1.0016 | 0.0301 | 1.0008 | 0.9988 | 0.0316 |
| MLE-noUH | 0.9690 | 0.9683 | 0.0351 | 0.9631 | 0.9636 | 0.0240 | 0.9620 | 0.9619 | 0.0259 |
| **Testing null hypothesis** | Frequency of Ho rejection with significance level 1% | 5% | 10% | Frequency of Ho rejection with significance level 1% | 5% | 10% | Frequency of Ho rejection with significance level 1% | 5% | 10% |
| No Unob. Het. | 0.011 | 0.062 | 0.111 | 0.029 | 0.099 | 0.174 | 0.028 | 0.096 | 0.170 |
| Two types | 0.010 | 0.049 | 0.105 | 0.008 | 0.050 | 0.111 | 0.011 | 0.060 | 0.104 |

Note (1): Mean, Median, and Standard deviation of estimated parameter over the distribution of 1,000 replications.

Finally, Table 7 presents results of experiments under DGP 4 where there is not unobserved heterogeneity and $RC = 8$. The purpose of these experiments is to study possible biases in the size of Hausman test for the null hypothesis of no unobserved heterogeneity. We can see that, for the three samples, the size of this test is very close to the nominal size.

**Table 7**
**Monte Carlo Experiments with DGP 4 (No UH, RC = 8)**

| Estimator of $\beta$ | **Sample A** ($t = 1$ to 7) Estimate[(1)] Mean | Median | St. dev. | **Sample B** ($t = 1$ to 14) Estimate[(1)] Mean | Median | St. dev. | **Sample C** ($t = 8$ to 21) Estimate[(1)] Mean | Median | St. dev. |
|---|---|---|---|---|---|---|---|---|---|
| FE-CMLE | 0.9986 | 1.0007 | 0.1283 | 1.0020 | 1.0011 | 0.0623 | 1.0026 | 0.9995 | 0.0593 |
| MLE-2types | 1.0167 | 1.0159 | 0.0410 | 1.0082 | 1.0071 | 0.0254 | 1.0083 | 1.0087 | 0.0259 |
| MLE-noUH | 1.0010 | 0.9996 | 0.0336 | 1.0018 | 1.0008 | 0.0233 | 1.0025 | 1.0022 | 0.0244 |
| **Testing null hypothesis** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** | **Frequency of Ho rejection with significance level 1%** | **5%** | **10%** |
| No Unob. Het. | 0.011 | 0.053 | 0.109 | 0.010 | 0.050 | 0.099 | 0.007 | 0.055 | 0.101 |
| Two types | 0.012 | 0.053 | 0.125 | 0.011 | 0.052 | 0.106 | 0.014 | 0.061 | 0.112 |

Note (1): Mean, Median, and Standard deviation of estimated parameter over the distribution of 1,000 replications.

## 5.2 Estimation of d*

In all the Monte Carlo experiments in the previous section we have assumed that the researcher knows the true value of $d^*$. Now, we present Monte Carlo experiments where we study the implications of estimating the value of $d^*$. We are interested in two questions: how precise is the estimation of $d^*$; and, more substantially, how the estimation of $d^*$ in a first step affects the of our estimation of $\beta$ in a second step. Remember that our estimation of $d^*$ is based on a sequential testing procedure. Starting with a high value of duration (i.e., $n = T/2 - 1$) we estimate $\Delta\beta_d(n) \equiv \beta_d(n) - \beta_d(n-1)$ and test for the null hypothesis that this value is equal to zero. If we reject the null hypothesis, we conclude $d^* = n$. If we do not reject the null hypothesis, then we conclude that

$d^* < n$ and then apply the same procedure to duration $n-1$. Different choices for the significance level of the testing procedure imply different estimates of $d^*$. Intuitively, a higher significance level (i.e., higher probability of rejecting the null) implies that we stop the procedure at a higher estimate for $d^*$.

Table 8 presents the results of our Monte Carlo experiments for the FE-CMLE when $d^*$ is estimated. The results are for DGP 1 and Sample B (periods $t = 1$ to 14). The results for the other DGPs are very similar. We report results for four different estimators of $d^*$ and the corresponding estimator of $\beta$: for three different significance levels (1%, 5%, and 10%), and using the true value of $d^*$. The value of $d^*$ is precisely estimated. There is an upward bias in the estimation of $d^*$ and this bias increases with the significance level. Interestingly, even a small bias (and variance) in the estimation of $d^*$ can generate a substantial bias (and large variance) in the estimation of $\beta$ in the second step. The bias in the estimation of $\beta$ is downward: a high significance level in the estimation of $d^*$ generates a downward bias in the estimation of duration dependence.

**Table 8**
**Monte Carlo experiments with estimation of d***

| | **DGP 1 (Normal RC)** | | | | | | |
|---|---|---|---|---|---|---|---|
| Estimator | Distribution estimate d* | | | | Estimate of $\beta$ | | |
| of d* | 4 | 5 | 6 | 7 | Mean | Median | St. dev. |
| True $d^*$ | 1000 | 1000 | 1000 | 1000 | 0.9983 | 0.9962 | 0.0758 |
| $\alpha = 1\%$ | 958 | 13 | 16 | 13 | 0.9548 | 0.9935 | 0.2705 |
| $\alpha = 5\%$ | 859 | 38 | 49 | 54 | 0.8519 | 0.9833 | 0.4465 |
| $\alpha = 10\%$ | 727 | 72 | 92 | 109 | 0.7201 | 0.9623 | 0.5532 |

## 5.3 Estimation using Rust's dataset

Rust's full sample contains a total of 124 buses that are classified in eight groups according to the bus size, and the engine manufacturer, model and year. For the estimation of the structural model, Rust focuses on groups 1 to 4 that account for 104 buses: 15 buses in group 1; 4 buses in group 2; 48 buses in group 3; and 37 buses in group 4. For each bus engine, the choice history in the data includes the actual initial condition of the engine, i.e., the first month where the engine was installed.

For these 104 buses, the empirical distribution of the number of engine replacements per bus is the following: 0 engine replacements for 44 buses; 1 replacement for 59 buses; and 2 replacements for 1 bus. For our FE-CMLE of the structural parameters $\beta_d$, choice histories with zero replacement do not contain any useful information. Therefore, for the CMLE we have only 61 complete spells until replacement. For our analysis, we consider that the frequency of the superintendent's decisions is at the annual level. Table 9 presents the empirical distribution of choice histories with a replacement.

**Table 9**
**Bus Engine Replacement (Rust, 1987)**
**Empirical distribution of choice histories with replacement**

| | *Frequency* | | |
|---|---|---|---|
| *Choice history* | *Absolute* | *%* | *% cumulative* |
| 1101111111 | 3 | 5.17 | 5.17 |
| 1110111111 | 11 | 18.96 | 24.13 |
| 1111011111 | 9 | 15.51 | 39.64 |
| 1111101111 | 18 | 31.03 | 70.67 |
| 1111110111 | 7 | 12.07 | 82.74 |
| 1111111011 | 5 | 8.62 | 91.36 |
| 1111111101 | 3 | 5.17 | 96.53 |
| 1111111110 | 2 | 3.45 | 100.00 |

Table 10 presents ML estimates of the model with three different specifications of the maintenance cost function $\beta_d(d)$ (linear, quadratic, and square-root), and different values for $d^*$. We consider a model with two unobserved types. However, for all the specifications, we always converge to a model with a single type. We have tried thousands of initial values for the vector of parameters (i.e., $RC_1$, $RC_2$, $\lambda$, and $\beta_d$), and we have also estimated the model using grid search. Regardless the computational strategy, we always converge to the same estimate with only one type. In terms of the value of $d^*$, the log-likelihood function is maximized at $d^* = 6$ for the square-root and linear specifications, and at $d^* = 5$ for the quadratic specification.

**Table 10**
**Bus Engine Replacement (Rust, 1987)**

**Maximum Likelihood Estimates**

| *Model* | | *RC* | | $\beta_d^* \equiv -\Delta\beta_d(d^*)$ | | |
|---|---|---|---|---|---|---|
| $\beta_d(d)$ | $d^*$ | $\widehat{RC}$ | $se\left(\widehat{RC}\right)$ | $\widehat{\beta_d^*}$ | $se\left(\widehat{\beta_d^*}\right)$ | log-*likelihood* |
| *Square root* | 3 | 36.0730 | 8.9966 | 2.0205 | 0.5218 | -162.7225 |
| $\beta_d(d) = \beta\sqrt{d}$ | 4 | 19.7583 | 3.6604 | 0.7939 | 0.1570 | -160.8052 |
| | 5 | 14.7558 | 2.3188 | 0.4608 | 0.0794 | -158.6263 |
| | **6** | **12.1821** | **1.7405** | **0.3145** | **0.0513** | **-158.2941**** |
| | 7 | 10.7006 | 1.4500 | 0.2391 | 0.0387 | -158.8309 |
| | 8 | 9.8510 | 1.3094 | 0.1975 | 0.0328 | -159.6416 |
| | 9 | 9.4309 | 1.2597 | 0.1745 | 0.0301 | -160.2893 |
| *Linear* | 3 | 22.2706 | 5.1700 | 2.0384 | 0.5008 | -162.7526 |
| $\beta_d(d) = \beta\ d$ | 4 | 13.0569 | 2.2277 | 0.8391 | 0.1580 | -160.9506 |
| | 5 | 10.1939 | 1.4449 | 0.5077 | 0.0824 | -158.8128 |
| | **6** | **8.6395** | **1.0906** | **0.3601** | **0.0552** | **-158.7414** |
| | 7 | 7.6966 | 0.9054 | 0.2837 | 0.0436 | -159.6630 |
| | 8 | 7.1362 | 0.8131 | 0.2432 | 0.0389 | -160.8675 |
| | 9 | 6.8582 | 0.7800 | 0.2243 | 0.0378 | -161.8001 |
| *Square* | 3 | 15.1987 | 3.2751 | 2.0778 | 0.4867 | 162.8254 |
| $\beta_d(d) = \beta\ d^2$ | 4 | 9.6031 | 1.4608 | 0.9282 | 0.1617 | 161.3203 |
| | **5** | **7.8125** | **0.9786** | **0.5998** | **0.0901** | **159.2938** |
| | 6 | 6.7404 | 0.7458 | 0.4496 | 0.0650 | 159.7556 |
| | 7 | 6.0324 | 0.6159 | 0.3723 | 0.0557 | 161.4468 |
| | 8 | 5.5874 | 0.5478 | 0.3364 | 0.0547 | 163.4387 |
| | 9 | 5.3611 | 0.5225 | 0.3329 | 0.0594 | 164.9528 |

Table 11 presents estimates of parameter $\beta$ using the FE-CMLE and under different values of $d^*$. Using a significance level of 10%, the estimate of $d^*$ is 3. The estimate of $\beta$ is 1.3218 (se = 0.5470).

**Table 11**
**Bus Engine Replacement (Rust, 1987)**

**Fixed-Effects-Conditional Maximum Likelihood**

| $d^*$ | $\widehat{\beta}$ | $se\left(\widehat{\beta}\right)$ | *p-value* $H_0 : \beta = 0$ | log-*likelihood* |
|---|---|---|---|---|
| 4 | 0.2048 | 0.3576 | 0.5670 | -22.864 |
| **3** | **1.3218** | **0.5470** | **0.0160** | **-11.222**$^{***}$ |

Table 12 compares the CMLE estimate of the parameter $\Delta\beta_d(d^*) \equiv \beta_d(d^*) - \beta_d(d^* - 1)$ with the corresponding MLE using the estimates in Table 9. Given the very small sample size and the corresponding large standard error of the CMLE estimate, we cannot reject the null hypothesis of no unobserved heterogeneity, despite the magnitude of the difference between MLE and CMLE estimates is important and it generates important differences in distribution of durations.

**Table 12**
**Bus Engine Replacement (Rust, 1987)**

**Hausman Test of Unobserved Heterogeneity**

| *Model* | $\widehat{\Delta\beta_d(d^*)}$ $(se)$ *MLE* | $\widehat{\Delta\beta_d(d^*)}$ $(se)$ *CMLE* | *Hausman* | *p-value* |
|---|---|---|---|---|
| Square root | 0.4294 (0.0700) | 1.3218 (0.5470) | 1.5479 | 0.2134 |
| Linear | 0.3601 (0.0552) | 1.3218 (0.5470) | 1.7181 | 0.1899 |
| Square | 0.3817 (0.0573) | 1.3218 (0.5470) | 1.6636 | 0.1971 |

# 6 Conclusions

To our knowledge, this paper presents the first identification results of structural parameters in forward-looking dynamic discrete choice models where the joint distribution of time-invariant unobserved heterogeneity and endogenous state variables is nonparametrically. This unobserved heterogeneity can have multiple components and can have continuous support. The distribution of the initial conditions is also nonparametrically specified. We consider models with two endogenous state variables: the lagged decision variable, and the time duration in the last choice. We show that structural parameters that capture switching costs are identified under mild conditions. The identification of structural parameters that capture duration dependence require additional restrictions. In particular, to obtain identification of these parameters we assume that the marginal return of an additional period of experience (duration) becomes equal to zero after a finite number of periods.

Based on our identification results, we propose tests for the null hypothesis of no unobserved heterogeneity and for a particular parametric specification of random effects and the initial conditions problem. Our Monte Carlo experiments show that the Conditional MLE provides precise estimates of structural parameters and the specification test has strong power in models with non-neglible amount of heterogeneity.

For models with duration dependence, we propose a sequential testing procedure to estimate the value of duration for which the return of experience becomes zero. The estimation of this parameter plays an important role in the finite-sample properties of the estimates of structural parameters. So far, we have considered a two-step procedure. It would be interesting to considered the joint estimation of the two sets of parameters.

## Appendix 1. Proofs

**Proof of Lemma 2.**

**(i)** For any $y > 0$, we have that $1\{y_{t-1} = y,\ d_t = 0\} = 0$ because $y_{t-1} > 0$ implies $d_t > 0$. Therefore, $H^{(y)}(0) = \sum_{t=1}^{T} 1\{y_{t-1} = y,\ d_t = 0\} = 0$.

**(ii)** For any $y > 0$, we have that $1\{y_{t-1} = y_t = y,\ d_t = 0\} = 0$ because $y_{t-1} > 0$ implies $d_t > 0$. Therefore, $X^{(y)}(0) = \sum_{t=1}^{T} 1\{y_{t-1} = y_t = y, d_t = 0\} = 0$.

**(iii)** For any $y > 0$, $\sum_{d\geq 1} H^{(y)}(d) = \sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_{t-1} = y,\ d_t = d\} = \sum_{t=1}^{T} 1\{y_{t-1} = y\} = T^{(y)} + 1\{y_0 = y\} - 1\{y_T = y\}$.

**(iv)** For any $y > 0$, $\sum_{d\geq 1} X^{(y)}(d) = \sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_{t-1} = y_t = y, d_t = d\} = \sum_{t=1}^{T} 1\{y_{t-1} = y_t = y\} = T^{(y)} + 1\{y_0 = y\} - 1\{y_T = y\}$.

**(v)** First, note that $y_{t-1} = y > 0$ implies $d_t \geq 1$. Therefore, for any $y > 0$ and $d \geq 1$, the event $\{y_{t-1} = y_t = y, d_t = d\}$ is equivalent to the event $\{y_t = y,\ d_{t+1} = d + 1\}$ for any $1 \leq t \leq T$. Therefore, $X^{(y)}(d) = \sum_{t=1}^{T}\{y_t = y, d_{t+1} = d + 1\} = \sum_{t=2}^{T+1} 1\{y_{t-1} = y, d_t = d + 1\} = H^{(y)}(d + 1) - 1\{y_0 = y, d_1 = d + 1\} + 1\{y_T = y, d_{T+1} = d + 1\}$. ■

**Proof of Propositions 1 and 2.** Remember that $T^{(y)}$ represents the number of times that choice alternative $y$ is visited in the choice history $\mathbf{y}^T$, and $D^{(y)}$ is the number of times that choice alternative $y$ is observed at two consecutive periods over the history $(y_0, \mathbf{y}^T)$. For the binary choice model, we have that $\sum_{t=1}^{T} y_t = T^{(1)}$, $\sum_{t=1}^{T} y_{t-1} y_t = D^{(1,1)}$, and $y_T - y_0 = \Delta^{(1)}$.

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, \theta\right) \quad = \quad T^{(1)}\left[\widetilde{\alpha}_\theta + \sigma_\theta(1) - \sigma_\theta(0)\right] + \Delta^{(1)}\ \left[\sigma_\theta(0) - \sigma_\theta(1)\right] + \widetilde{\beta}_y\ D^{(1,1)} \tag{A.1}$$

where we have omitted the term $T\ \sigma_\theta(0)$ because $T$ is constant over all the histories. Consider choice histories $A = \{0|0, 1, 1\}$ and $B = \{0|1, 0, 1\}$. It is clear that $T_A^{(1)} = T_B^{(1)} = 2$, and $\Delta_A^{(1)} = \Delta_B^{(1)} = 1$, such that $U_A = U_B$. Also, $D_A^{(1,1)} = 1$ and $D_B^{(1,1)} = 0$. Therefore, $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \widetilde{\beta}_y$. ■

**Proof of Proposition 3.** The log-probability of this model is:

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta\right) \quad = \quad \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y\ y_{t-1} + \beta_d(1, d_t)\ y_{t-1}\right] + \sigma_\theta(y_{t-1}, d_t) \tag{A.2}$$

We have that $\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta\right) = T^{(1)}\widetilde{\alpha}_\theta + [T^{(0)} - \Delta^{(0)}]\sigma_\theta(0) + D^{(1,1)}\widetilde{\beta}_y + \sum_{d\geq 1} X^{(1)}(d)\ \beta_d(1, d) + \sum_{d\geq 1} H^{(1)}(d)\ \sigma_\theta(1, d)$. Taking into account that $\sum_{d\geq 1} H^{(1)}(d) = T^{(1)} - \Delta^{(1)}$ and $D^{(1,1)} = \sum_{d\geq 1} X^{(1)}(d)$, we obtain:

$$\begin{aligned}\ln \mathbb{P}\left(\mathbf{y}^T|y_0,\theta\right) &= \sum_{d\geq 1} H^{(1)}(d)\ \left[\widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0)\right] + \Delta^{(1)}\widetilde{\alpha}_\theta \\ &+ \sum_{d\geq 1} X^{(1)}(d)\ \gamma(d)\end{aligned} \tag{A.3}$$

where we have omitted the term $T\ \sigma_\theta(0)$ because $T$ is constant over all the histories, and we define $\gamma(d) \equiv \widetilde{\beta}_y + \beta_d(1,d)$. Now, Lemma 2(v) establishes that $X^{(1)}(d) = H^{(1)}(d+1) + \Delta^{(1)}(d+1)$. Note that $\sum_{d\geq 1}\left[H^{(1)}(d+1) + \Delta^{(1)}(d+1)\right]\ \gamma(d)$ is equal to $\sum_{d\geq 1}\left[H^{(1)}(d) + \Delta^{(1)}(d)\right]\ \gamma(d-1)$, if we define $\gamma(0) = 0$. Then, we have that,

$$\begin{aligned}\ln \mathbb{P}\left(\mathbf{y}^T|y_0,\theta\right) &= \sum_{d\geq 1} H^{(1)}(d)\ \left[\widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0)\right] + \Delta^{(1)}\widetilde{\alpha}_\theta \\ &+ \sum_{d\geq 1}\left[H^{(1)}(d) + \Delta^{(1)}(d)\right]\ \gamma(d-1) \\ &= \sum_{d\geq 1} H^{(1)}(d)\ \left[\widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1)\right] + \Delta^{(1)}\widetilde{\alpha}_\theta \\ &+ \sum_{d\geq 1}\Delta^{(1)}(d)\ \gamma(d-1) \qquad \blacksquare\end{aligned} \tag{A.4}$$

**Proof of Proposition 4.** The log-probability of this model is:

$$\ln \mathbb{P}\left(\mathbf{y}^T \mid y_0, d_1, \theta\right) = \sum_{t=1}^{T} y_t\left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1} + \beta_d(1,d_t)y_{t-1} + v_\theta\left(1, d_t+1\right)\right] + \sigma_\theta(y_{t-1}, d_t) \tag{A.5}$$

Comparing this log-probability with the one for the myopic model with duration, we can see that the only difference is in the term $\sum_{t=1}^T y_t\ v_\theta\left(1,d_t+1\right)$, that can be written as $\sum_{d\geq 0}\sum_{t=1}^T y_t\ 1\{d_t = d\}$ $v_\theta\left(1,d+1\right)$. For the statistic $\sum_{t=1}^T y_t\ 1\{d_t = d\}$ we can distinguish two cases: (a) if $d=0$, then $\sum_{t=1}^T y_t\ 1\{d_t = 0\} = \sum_{t=1}^T y_t\ (1-y_{t-1}) = T^{(1)} - D^{(1,1)}$; and (b) if $d \geq 1$, then $\sum_{t=1}^T y_t\ 1\{d_t = d\} = \sum_{t=1}^T y_t\ y_{t-1}\ 1\{d_t = d\} = X^{(1)}(d)$. Therefore,

$$\begin{aligned}\sum_{d\geq 0}\sum_{t=1}^{T} y_t 1\{d_t = d\} v_\theta\left(1,d+1\right) &= \left[T^{(1)} - D^{(1,1)}\right]\ v_\theta\left(1,1\right) + \sum_{d\geq 1} X^{(1)}(d)\ v_\theta\left(1,d+1\right) \\ &= T^{(1)}\ v_\theta\left(1,1\right) + \sum_{d\geq 1} X^{(1)}(d)\ \left[v_\theta\left(1,d+1\right) - v_\theta\left(1,1\right)\right]\end{aligned} \tag{A.6}$$

where for the second equality we have applied Lemma 2(iv), $D^{(1,1)} = \sum_{d\geq 1} X^{(1)}(d)$. Then, the log-probability is equal to

$$\begin{aligned}\ln \mathbb{P}\left(\mathbf{y}^{T}|y_{0},\theta\right) &= \sum_{d\geq 1} H^{(1)}(d) \;\; [\widetilde{\alpha}_{\theta}+\sigma_{\theta}(1,d)-\sigma_{\theta}(0)+\gamma(d-1)] + \Delta^{(1)}\widetilde{\alpha}_{\theta} \\ &+ \sum_{d\geq 1} \Delta^{(1)}(d) \; \gamma(d-1) \\ &+ T^{(1)} \;\; v_{\theta}\left(1,1\right) + \sum_{d\geq 1} X^{(1)}(d) \;\; [v_{\theta}\left(1,d+1\right) - v_{\theta}\left(1,1\right)]\end{aligned} \tag{A.7}$$

From Lemma 2, we have that: (iii) $T^{(1)}=\sum_{d\geq 1}H^{(1)}(d)+\Delta^{(1)}$; and (v) $X^{(1)}(d)=H^{(1)}(d+1)+$ $\Delta^{(1)}(d+1)$, and solving these expressions in (A.13), we have that:

$$\begin{aligned}\ln \mathbb{P}\left(\mathbf{y}^{T}|y_{0},\theta\right) &= \sum_{d\geq 1} H^{(1)}(d) \;\; [\widetilde{\alpha}_{\theta}+\sigma_{\theta}(1,d)-\sigma_{\theta}(0)+\gamma(d-1)+v_{\theta}\left(1,d\right)] \\ &+ \Delta^{(1)} \;\; [\widetilde{\alpha}_{\theta}+v_{\theta}\left(1,1\right)] \\ &+ \sum_{d\geq 1} \Delta^{(1)}(d) \;\; [v_{\theta}\left(1,d\right)-v_{\theta}\left(1,1\right)+\gamma(d-1)]\end{aligned} \tag{A.8}$$

Taking into account that $\Delta^{(1)}=\sum_{d\geq 1}\Delta^{(1)}(d)$, we have:

$$\ln \mathbb{P}\left(\mathbf{y}^{T}|y_{0},\theta\right) = \sum_{d\geq 1} H^{(1)}(d) \; g_{\theta,1}(d) + \sum_{d\geq 1} \Delta^{(1)}(d) \;\; [\widetilde{\alpha}_{\theta}+v_{\theta}\left(1,d\right)+\gamma(d-1)] \tag{A.9}$$

with $g_{\theta,1}(d)\equiv\widetilde{\alpha}_{\theta}+\sigma_{\theta}(1,d)-\sigma_{\theta}(0)+\gamma(d-1)+v_{\theta}\left(1,d\right)$. ■

**Proof of Proposition 5.** Define $Z\equiv\sum_{d\geq 1}\Delta^{(1)}(d)\ [v_{\theta}\left(1,d\right)+\gamma(d-1)]$. Under Assumption 2, we have that $v_{\theta}\left(1,d\right)=v_{\theta}\left(1,d^{*}\right)$ for any $d\geq d^{*}$, and $\gamma\left(d-1\right)=\gamma\left(d^{*}\right)$ for any $d\geq d^{*}+1$. Therefore, we have:

$$\begin{aligned} Z &= \sum_{d\leq d^{*}-1} \Delta^{(1)}(d) \; v_{\theta}\left(1,d\right) + \left[\sum_{d\geq d^{*}} \Delta^{(1)}(d)\right] v_{\theta}\left(1,d^{*}\right) \\ &+ \sum_{d\leq d^{*}} \Delta^{(1)}(d) \; \gamma(d-1) + \left[\sum_{d\geq d^{*}+1} \Delta^{(1)}(d)\right] \gamma(d^{*}) \\ &= \sum_{d\leq d^{*}-1} \Delta^{(1)}(d) \;\; [v_{\theta}\left(1,d\right)+\gamma(d-1)] + \left[\sum_{d\geq d^{*}} \Delta^{(1)}(d)\right] [v_{\theta}\left(1,d^{*}\right)+\gamma(d^{*})] \\ &+ \Delta^{(1)}(d^{*}) \;\; [\gamma(d^{*}-1)-\gamma(d^{*})]\end{aligned} \tag{A.10}$$

Then, the log-probability becomes:

$$\begin{aligned}\ln \mathbb{P}\left(\mathbf{y}^T | y_0, \theta\right) &= \sum_{d \geq 1} H^{(1)}(d)\ g_{\theta,1}(d) \\ &+ \sum_{d \leq d^*-1} \Delta^{(1)}(d)\ g_{\theta,2}(d) + \left[\sum_{d \geq d^*} \Delta^{(1)}(d)\right] g_{\theta,2}(d^*) \\ &+ \Delta^{(1)}(d^*)\ \left[\gamma(d^*-1) - \gamma(d^*)\right]\end{aligned} \tag{A.11}$$

with $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1) + v_\theta\,(1,d)$, and $g_{\theta,2}(d) \equiv \widetilde{\alpha}_\theta + v_\theta\,(1,d) + \gamma(d-1)$. Note that $g_{\theta,1}(d) = g_{\theta,1}(d^*)$ for any $d \geq d^*$. Therefore, we have $\sum_{d \geq 1} H^{(1)}(d)\ g_{\theta,1}(d) = \sum_{d \leq d^*-1} H^{(1)}(d)$ $g_{\theta,1}(d) +$ $\left[\sum_{d \geq d^*} H^{(1)}(d)\right] g_{\theta,1}(d^*)$, such that

$$\begin{aligned}\ln \mathbb{P}\left(\mathbf{y}^T | y_0, \theta\right) &= \sum_{d \leq d^*-1} H^{(1)}(d)\ g_{\theta,1}(d) + \left[\sum_{d \geq d^*} H^{(1)}(d)\right] g_{\theta,1}(d^*) \\ &+ \sum_{d \leq d^*-1} \Delta^{(1)}(d)\ g_{\theta,2}(d) + \left[\sum_{d \geq d^*} \Delta^{(1)}(d)\right] g_{\theta,2}(d^*) \\ &+ \Delta^{(1)}(d^*)\ \left[\gamma(d^*-1) - \gamma(d^*)\right] \qquad \blacksquare\end{aligned} \tag{A.12}$$

**Proof of Propositions 7 and 8.** For this model, the log probability is $\sum_{j=0}^{J} \sum_{t=1}^{T} 1\{y_t = j\}$ $\alpha_\theta(j) +$ $\sum_{j=0}^{J} \sum_{k=0}^{J} \sum_{t=1}^{T} 1\{y_{t-1} = j,\ y_t = k\}\ \beta_y(k,j) +$ $\sum_{j=0}^{J} \sum_{t=1}^{T} 1\{y_{t-1} = j\}\ \sigma_\theta(j)$. Using the definitions of our statistics, we have that:

$$\ln \mathbb{P}\left(\mathbf{y}^T | y_0, \theta\right) = \sum_{j=0}^{J} T^{(j)} \alpha_\theta(j) + \sum_{j=0}^{J} \sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j) + \sum_{j=0}^{J} \left[T^{(j)} - \Delta^{(j)}\right] \sigma_\theta(j) \tag{A.13}$$

where $\Delta^{(j)} \equiv 1\{y_T = j\} - 1\{y_0 = j\}$. Note that $T^{(0)} = T - \sum_{j=1}^{J} T^{(j)}$, and $\Delta^{(0)} = 1 - \sum_{j=1}^{J} \Delta^{(j)}$, such that:

$$\begin{aligned}\ln \mathbb{P}\left(\mathbf{y}^T | y_0, \theta\right) &= \sum_{j=1}^{J} T^{(j)} \left[\alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j) - \sigma_\theta(0)\right] + \sum_{j=1}^{J} \Delta^{(j)}\ \left[-\sigma_\theta(j) + \sigma_\theta(0)\right] \\ &+ \sum_{j=0}^{J} \sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j)\end{aligned} \tag{A.14}$$

where we have omitted the term $T\alpha_\theta(0) + \sigma_\theta(0)$ because it is constant over all the choice histories. For the term, $\sum_{j=0}^{J} \sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j)$, note that: $\sum_{j=0}^{J} D^{(j,k)} = T^{(k)}$ such that $D^{(0,k)} = T^{(k)} -$

$\sum_{j=1}^{J} D^{(j,k)}$; and $\sum_{k=0}^{J} D^{(j,k)} = T^{(j)} - \Delta^{(j)}$ such that $D^{(j,0)} = T^{(j)} - \Delta^{(j)} - \sum_{k=1}^{J} D^{(j,k)}$. Therefore,

$$\begin{aligned} \sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j) &= \sum_{j=0}^{J} \left[ \sum_{k=1}^{J} D^{(j,k)} \beta_y(k,j) + \left[ T^{(j)} - \Delta^{(j)} - \sum_{k=1}^{J} D^{(j,k)} \right] \beta_y(0,j) \right] \\ &= \sum_{j=0}^{J}\sum_{k=1}^{J} D^{(j,k)} \left[ \beta_y(k,j) - \beta_y(0,j) \right] + \sum_{j=1}^{J} \left[ T^{(j)} - \Delta^{(j)} \right] \beta_y(0,j) \end{aligned} \tag{A.15}$$

where we have omitted the term $(T-1)\beta_y(0,0)$ because it is constant over every choice history (also we have normalized $\beta_y(y,y) = 0$ for every $y$). Now, applying a similar property to the term $\sum_{j=0}^{J}\sum_{k=1}^{J} D^{(j,k)} \; [\beta_y(k,j) - \beta_y(0,j)]$, we have:

$$\begin{aligned} \sum_{j=0}^{J}\sum_{k=1}^{J} D^{(j,k)} \left[ \beta_y(k,j) - \beta_y(0,j) \right] &= \sum_{k=1}^{J} \left[ \sum_{j=1}^{J} D^{(j,k)} \left[ \beta_y(k,j) - \beta_y(0,j) \right] + \left[ T^{(k)} - \sum_{j=1}^{J} D^{(j,k)} \right] \left[ \beta_y(k,0) - \beta_y(0,0) \right] \right] \\ &= \sum_{k=1}^{J}\sum_{j=1}^{J} D^{(j,k)} \left[ \beta_y(k,j) - \beta_y(0,j) - \beta_y(k,0) \right] + \sum_{k=1}^{J} T^{(k)} \beta_y(k,0) \end{aligned} \tag{A.16}$$

Putting together (A.15) and (A.16), we have that:

$$\sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j) = \sum_{k=1}^{J}\sum_{j=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} T^{(j)} \left[ \beta_y(0,j) + \beta_y(j,0) \right] - \sum_{j=1}^{J} \Delta^{(j)} \beta_y(0,j) \tag{A.17}$$

where $\widetilde{\beta}_y(k,j) \equiv \beta_y(k,j) - \beta_y(0,j) - \beta_y(k,0)$. And plugging this expression into equation (A.14) for the log-probability, we obtain:

$$\ln \mathbb{P}\left( \mathbf{y}^T | y_0, \theta \right) = \sum_{j=1}^{J} T^{(j)} \; g_{\theta,1}(j) + \sum_{j=1}^{J} \Delta^{(j)} \; g_{\theta,2}(j) + \sum_{j=1}^{J}\sum_{k=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) \tag{A.18}$$

where $g_{\theta,1}(j) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j) - \sigma_\theta(0) + \beta_y(0,j) + \beta_y(j,0)$, and $g_{\theta,2}(j) \equiv -\sigma_\theta(j) + \sigma_\theta(0) - \beta_y(0,j)$. $\blacksquare$

**Proof of Proposition 9.** For this model, the log probability is $\sum_{j=0}^{J}\sum_{t=1}^{T} 1\{y_t = j\} \; \alpha_\theta(j) + \sum_{j=0}^{J}\sum_{k=0}^{J}\sum_{t=1}^{T} 1\{y_{t-1} = j, \, y_t = k\} \; \beta_y(k,j) + \sum_{j=1}^{J}\sum_{d=1}^{J}\sum_{t=1}^{T} 1\{y_{t-1} = y_t = j, \, d_t = d\} \; \beta_d(j,d) + \sum_{j=0}^{J}\sum_{d=0}^{J}\sum_{t=1}^{T} 1\{y_{t-1} = j, \, d_t = d\} \; \sigma_\theta(j,d)$. Using the definition of the statistics in Table 1, we can write this log-probability as follows:

$$\begin{aligned} \ln \mathbb{P}\left( \mathbf{y}^T | y_0, d_1, \theta \right) &= \sum_{j=0}^{J} T^{(j)} \alpha_\theta(j) + \left[ T^{(0)} - \Delta^{(0)} \right] \sigma_\theta(0) + \sum_{j=1}^{J}\sum_{d \geq 1} H^{(j)}(d) \sigma_\theta(j,d) \\ &+ \sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j) + \sum_{j=1}^{J}\sum_{d \geq 1} X^{(j)}(d) \beta_d(j,d) \end{aligned} \tag{A.19}$$

Taking into account that: $T^{(0)} = T - \sum_{j=1}^{J} T^{(j)}$, we have that $\sum_{j=0}^{J} T^{(j)} \alpha_\theta(j) +$ $T^{(0)}\sigma_\theta(0) =$ $T[\alpha_\theta(0) + \sigma_\theta(0)] +$ $\sum_{j=0}^{J} T^{(j)}[\alpha_\theta(j) - \alpha_\theta(0) - \sigma_\theta(0)]$. And using equation (A.17) from the proof of Propositions 7-8, we have:

$$
\begin{aligned}
\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1, \theta\right) &= \sum_{j=0}^{J} T^{(j)} \left[\alpha_\theta(j) - \alpha_\theta(0) - \sigma_\theta(0) + \beta_y(0,j) + \beta_y(j,0)\right] + \sum_{j=1}^{J} \Delta^{(j)} \left[\sigma_\theta(0) - \beta_y(0,j)\right] \\
&+ \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d) \sigma_\theta(j,d) \\
&+ \sum_{k=1}^{J} \sum_{j=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} \sum_{d \geq 1} X^{(j)}(d) \beta_d(j,d)
\end{aligned} \tag{A.20}
$$

where we have omitted the term $T\alpha_\theta(0) + (T-1)\sigma_\theta(0)$ because the are constant across all the histories. Given that $T^{(j)} = \Delta^{(j)} + \sum_{d \geq 1} H^{(j)}(d)$ and $D^{(j,j)} = \sum_{d \geq 1} X^{(j)}(d)$ and $\tilde{\beta}_y(j,j) = -\beta_y(j,0) - \beta_y(0,j)$ by construction, we get:

$$
\begin{aligned}
\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1, \theta\right) &= \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d) \; \left[\alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j,d) - \sigma_\theta(0) + \beta_y(0,j) + \beta_y(j,0)\right] \\
&+ \sum_{j=1}^{J} \Delta^{(j)} \left[\alpha_\theta(j) - \alpha_\theta(0) + \beta_y(j,0)\right] \\
&+ \sum_{k=1}^{J} \sum_{j=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} \sum_{d \geq 1} X^{(j)}(d) \gamma(j,d)
\end{aligned} \tag{A.21}
$$

Now, consider the term $\sum_{j=1}^{J} \sum_{d \geq 1} X^{(j)}(d) \; \beta_d(j,d)$. By Lemma 2, for $d \geq 1$, $X^{(j)}(d) = H^{(j)}(d+1) - \Delta^{(j)}(d+1)$. Therefore,

$$
\begin{aligned}
\sum_{j=1}^{J} \sum_{d \geq 1} X^{(j)}(d) \beta_d(j,d) &= \sum_{j=1}^{J} \sum_{d \geq 1} \left[H^{(j)}(d+1) + \Delta^{(j)}(d+1)\right] \gamma(j,d) \\
&= \sum_{j=1}^{J} \sum_{d \geq 1} \left[H^{(j)}(d) + \Delta^{(j)}(d)\right] \; \gamma(j,d-1)
\end{aligned} \tag{A.22}
$$

where, for the second equality, we take into account the normalization $\beta_d(j,0) = 0$ for any $j \geq 1(*)$. Solving equation (A.22) into (A.21), and taking into account that $\sum_{d \geq 1} \Delta^{(j)}(d) = \Delta^{(j)}$, we obtain:

$$
\begin{aligned}
\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1, \theta\right) &= \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d) \; g_{\theta,1}(j,d) + \sum_{j=1}^{J} \Delta^{(j)} g_{\theta,2}(j) \\
&+ \sum_{k=1}^{J} \sum_{j=1, j \neq k}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} \sum_{d \geq 1} \Delta^{(j)}(d) \; \gamma(j,d-1)
\end{aligned} \tag{A.23}
$$

with $g_{\theta,1}(j,d) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j,d) - \sigma_\theta(0) + \beta_y(0,j) + \beta_y(j,0) + \gamma(j,d-1)$, $g_{\theta,2}(j) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \beta_y(j,0)$, $\widetilde{\beta}_y(y,y_{-1}) \equiv \beta_y(y,y_{-1}) - \beta_y(0,y_{-1}) - \beta_y(y,0)$, and $\gamma(j,d) \equiv \beta_d(j,d) - \beta_y(j,0) - \beta_y(0,j)$.

**Proof of Proposition 10.** The expression of the log-probability is similar as in Proposition 9 but now we have the additional term $\sum_{t=1}^{T} v_\theta(y_t, d_{t+1}[y, y_{t-1}, d_t])$. This term is equal to $T^{(0)} v_\theta(0) + \sum_{j=1}^{J} \sum_{d \geq 1} \sum_{t=1}^{T} 1\{y_t = y,\ d_{t\check{d}\check{L}\text{ffi}1} = d\}\ v_\theta(y,d) = T^{(0)} v_\theta(0) + \sum_{j=1}^{T} \sum_{d \geq 1} v_\theta(y,d) \Big[ H^{(y)}(d) + \Delta^{(y)}(d) \Big]$. Given $T^{(0)} = T - \sum_{j=1}^{T} T^{(y)} = T - \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d) - \sum_{j=1}^{T} \sum_{d \geq 1} \Delta^{(j)}(d)$ and using equation (A.23) from the proof of Proposition 9, we have

$$
\begin{aligned}
\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1, \theta\right) &= \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d)\ g_{\theta,1}(j,d) + \sum_{j=1}^{J} \Delta^{(j)} g_{\theta,2}(j) \\
&+ \sum_{k=1}^{J} \sum_{j=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} \sum_{d \geq 1} \Delta^{(j)}(d)\ \left(\gamma(j,d-1) + v_\theta(j,d)\right)
\end{aligned}
\tag{A.24}
$$

with $g_{\theta,1}(j,d) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j,d) - \sigma_\theta(0) + \beta_y(0,y) + \beta_y(y,0) + \gamma(y,d-1) + v_\theta(j,d) - v_\theta(0)$, $g_{\theta,2}(j) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \beta_y(y,0) - v_\theta(0)$

Taking into account that $\sum_{d \geq 1} \Delta^{(j)}(d) = \Delta^{(j)}$ for any $j \geq 1$, we have

$$
\begin{aligned}
\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1, \theta\right) &= \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d)\ g_{\theta,1}(j,d) \sum_{k=1}^{J} \sum_{j=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) \\
&+ \sum_{j=1}^{J} \sum_{d \geq 1} \Delta^{(j)}(d)\ \left(\gamma(j,d-1) + v_\theta(j,d) + \alpha_\theta(j) - \alpha_\theta(0) + \beta_y(y,0) - v_\theta(0)\right)
\end{aligned}
\tag{A.25}
$$

**Proof of Proposition 11.** Define $Z^{(j)} \equiv \sum_{d \geq 1} \Delta^{(j)}(d)[v_\theta(j,d) + \gamma(j,d-1)]$. Under Assumption 2, we have $v_\theta(j,d) = v_\theta(j,d^*)$ for any $d \geq d_j^*$, and $\gamma(j,d-1) = \gamma(j,d^*)$ for any $d \geq d_j^* + 1$. Therefore, we have for all $j \geq 1$,

$$
\begin{aligned}
Z^{(j)} &= \textstyle\sum_{d \leq d_j^* - 1} \Delta^{(j)}(d) v_\theta(j,d) + \left[ \sum_{d \geq d_j^*} \Delta^{(j)}(d) \right] v_\theta(j, d_j^*) \\
&+ \textstyle\sum_{d \leq d_j^*} \Delta^{(j)}(d) \gamma(j,d-1) + \left[ \sum_{d \geq d_j^* + 1} \Delta^{(j)}(d) \right] \gamma(j, d_j^*) \\
&= \textstyle\sum_{d \leq d_j^* - 1} \Delta^{(j)}(d) [v_\theta(j,d) + \gamma(j,d-1)] + \left[ \sum_{d \geq d_j^*} \Delta^{(j)}(d) \right] [v_\theta(j,d_j^*) + \gamma(j,d_j^*)] \\
&+ \Delta^{(j)}(d_j^*) [\gamma(j, d_j^* - 1) - \gamma(j, d_j^*)]
\end{aligned}
\tag{A.26}
$$

Then the log-probability becomes:

$$\begin{aligned}
\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1, \theta\right) &= \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d)\ g_{\theta,1}(j,d) \sum_{k=1}^{J} \sum_{j=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) \\
&+ \textstyle\sum_{j=1}^{J} \sum_{d \geq d_j^*-1} \Delta^{(j)}(d) g_{\theta,2}(d) + \sum_{j=1}^{J} \left[ \sum_{d \geq d_j^*} \Delta^{(j)}(d) \right] g_{\theta,2}(d_j^*) \\
&+ \sum_{j=1}^{J} \Delta^{(j)}(d_j^*)\ \left(\gamma(j, d_j^*-1) - \gamma(j, d_j^*)\right)
\end{aligned} \tag{A.27}$$

with $g_{\theta,1}(j,d) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j,d) - \sigma_\theta(0) + \beta_y(0,y) + \beta_y(y,0) + \gamma(y,d-1) + v_\theta(j,d) - v_\theta(0)$ and $g_{\theta,2} \equiv \alpha_\theta(j) - \alpha_\theta(0) + \beta_y(y,0) - v_\theta(0) + v_\theta(j,d) + \gamma(j,d-1)$. Note that $g_{\theta,1}(j,d) = g_{\theta,1}(j,d_j^*)$ for $d \geq d_j^*$. Therefore, we have $\sum_{d \geq 1} H^{(j)}(d) g_{\theta,1}(d) = \sum_{d \leq d_j^*-1} H^{(j)}(d) g_{\theta,1}(d) + \left[ \sum_{d \geq d_j^*} H^{(j)}(d) \right] g_{\theta,1}(d_j^*)$, such that

$$\begin{aligned}
\ln \mathbb{P}\left(\mathbf{y}^T | y_0, d_1, \theta\right) &= \sum_{j=1}^{J} \sum_{d \leq d_j^*-1} H^{(j)}(d)\ g_{\theta,1}(j,d) + \textstyle\sum_{j=1}^{J} \left[ \sum_{d \geq d_j^*} H^{(j)}(d) \right] g_{\theta,1}(d_j^*) \\
&+ \sum_{k=1}^{J} \sum_{j=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) \\
&+ \textstyle\sum_{j=1}^{J} \sum_{d \geq d_j^*-1} \Delta^{(j)}(d) g_{\theta,2}(d) + \sum_{j=1}^{J} \left[ \sum_{d \geq d_j^*} \Delta^{(j)}(d) \right] g_{\theta,2}(d_j^*) \\
&+ \sum_{j=1}^{J} \Delta^{(j)}(d_j^*)\ \left(\gamma(j, d_j^*-1) - \gamma(j, d_j^*)\right)
\end{aligned} \tag{A.28}$$

## Appendix 2. Model with stochastic transition of the endogenous state variables

Consider a model with the same structure as the model in Section 2 and Assumption 1 but now the vector of endogenous state variables is $\mathbf{x}_t = (x_t^y, x_t^d)$ and variables $x_t^y$ and $x_t^d$ *stochastic versions* of the variables $y_{t-1}$ and $d_t$, respectively. We now describe precisely the stochastic process of these variables.

The support of state variable $x_t^y$ is the choice set $\mathcal{Y}$, and its transition rule is $x_{t+1}^y = f_y(y_t, \xi_{t+1}^y)$ where $\xi_{t+1}^y$ is i.i.d. over time and independent of $\mathbf{x}_t$. The support of state variable $x_t^d$ is the set of possible durations, $\{1, 2, ..., \infty\}$, and its transition rule is $x_{t+1}^d = 1\{y_t > 0\}[\ 1\{y_t = x_t^y\}\ x_t^d + 1 + \xi_{t+1}^d]$, where $\xi_{t+1}^d$ has support $\{0, 1, ..., \infty\}$, and it is i.i.d. over time and independent of $\mathbf{x}_t$. Importantly, the stochastic shocks $\xi_{t+1}^y$ and $\xi_{t+1}^d$ are not known to the agent when she makes her decision at period $t$. Note that this model becomes our model in the main text when these shocks have a degenerate probability distribution with $p(\xi_{t+1}^y = 0) = p(\xi_{t+1}^d = 0) = 1$.

Assumption 1' below is simply an extension of our Assumption 1 to this stochastic version of the model. We omit the exogenous state variables $\mathbf{z}_t$ for notational simplicity.

*ASSUMPTION 1'. (A) The time horizon is infinite and $\delta \in (0,1)$. (B) The utility function is $\Pi_t(y) = \alpha_\theta(y) + 1\{y = x_t^y\}\ \beta_d\left(y, x_t^d\right) + 1\{y \neq x_t^y\}\ \beta_y\left(y, x_t^y\right) + \varepsilon_t(y)$, and functions $\alpha_\theta(y)$, $\beta_d\left(y, x_t^d\right)$, and $\beta_y(y, x_t^y)$ are bounded. (C) $\beta_y(y,y) = 0$, $\beta_d\left(0, x^d\right) = 0$. (D) $\{\varepsilon_t(y) : y \in \mathcal{Y}\}$ are i.i.d. over $(i,t,y)$ with a extreme value type I distribution. (E) $\mathbf{z}_t$ has discrete and finite support $\mathcal{Z}$ and follows a time-homogeneous Markov process. (F) The probability distribution of $\theta$ conditional on $\{\mathbf{z}_t, \mathbf{x}_t : t = 1, 2, ...\}$ is nonparametrically specified and completely unrestricted. (G) $x_t^y \in \mathcal{Y}$, and $x_{t+1}^y = f_y(y_t, \xi_{t+1}^y)$ where $\xi_{t+1}^y$ is i.i.d. over time and independent of $\mathbf{x}_t$; $x_t^d \in \{0, 1, ..., \infty\}$, and $x_{t+1}^d = 1\{y_t > 0\}[\ 1\{y_t = x_t^y\}\ x_t^d + 1 + \xi_{t+1}^d]$, where $\xi_{t+1}^d$ has support $\{0, 1, ..., \infty\}$, and it is i.i.d. over time and independent of $\mathbf{x}_t$.* ■

The model has the following integrated Bellman equation:

$$V_\theta(\mathbf{x}_t) = \ln\left(\sum_{y \in \mathcal{Y}} \exp\left\{\ \alpha_\theta(y) + \beta(y, \mathbf{x}_t) + \delta\ \mathbb{E}_{\xi_{t+1}}\left[V_\theta\left(f_y(y_t, \xi_{t+1}^y),\ 1\{y_t = x_t^y\}\, x_t^d + 1 + \xi_{t+1}^d\right)\right]\ \right\}\right)$$

where $\mathbb{E}_{\xi_{t+1}}(.)$ the expectation over the distribution of $(\xi_{t+1}^y, \xi_{t+1}^d)$. Let $v_{\theta,t}$ be the continuation value function $\delta\ \mathbb{E}_{\xi_{t+1}}[V_\theta\left(f_y(y_t, \xi_{t+1}^y),\ 1\{y_t = x_t^y\}\, x_t^d +\ 1 + \xi_{t+1}^d\right)]$ . Under our assumptions on the distribution of $(\xi_{t+1}^y, \xi_{t+1}^d)$, the continuation value function has very similar properties as in the model with a deterministic transition of the endogenous state variables. More specifically, (a) it depends only $y_t$ and $1\{y_t = x_t^y\}\, x_t^d + 1$, i.e., $v_{\theta,t} = v_\theta(y_t, 1\{y_t = x_t^y\}\, x_t^d + 1)$; (b) If $y_t \neq x_t^y$, then $v_{\theta,t} = v_\theta(y_t, 1)$; (c) If $y_t = x_t^y$, then $v_{\theta,t} = v_\theta(y_t, x_t^d + 1)$; and (D) if $x_t^d \geq d_y^* - 1$ and $y_t = x_t^y$, then $v_{\theta,t} = v_\theta(y_t, d_y^*)$.

# References


[1] Aguirregabiria, V. (1999): "The dynamics of markups and inventories in retailing firms," *The Review of Economic Studies*, 66, 275-308.

[2] Aguirregabiria, V. and P. Mira (2007): "Sequential Estimation of Dynamic Discrete Games," *Econometrica*, 75, 1-53.

[3] Andersen, E (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society, Series B*, 32, 283 301.

[4] Arcidiacono, P., and R. Miller (2011): "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, 79(6), 1823-1867.

[5] Arellano, M., and S. Bonhomme (2012): "Nonlinear Panel Data Analysis", *Annual Review of Economics*, 3, 395-424.

[6] Arellano, M., and S. Bonhomme (2017): "Nonlinear Panel Data Methods for Dynamic Heterogeneous Agent Models, *Annual Review of Economics*, forthcoming.

[7] Arellano, M., and B. Honoré (2001): "Panel Data Models: Some Recent Developments," in J. J. Heckman and E. Leamer (eds.) *Handbook of Econometrics*, Volume 5, Chapter 53, North-Holland, 3229-3296.

[8] Blackwell, D. (1965): "Discounted dynamic programming," *The Annals of Mathematical Statistics*, 36(1), 226-235.

[9] Bonhomme, S. (2012): "Functional differencing," *Econometrica*, 80(4), 1337-1385.

[10] Browning, M., and J. Carro (2010): "Heterogeneity in dynamic discrete choice models," *Econometrics Journal*, 13(1), 1-39.

[11] Browning, M., and J. Carro (2014): "Dynamic binary outcome models with maximal heterogeneity," *Journal of Econometrics*, 178(2), 805-823.

[12] Caliendo, L., M. Dvorkin, and F. Parro (2015): "Trade and Labor Market Dynamics," NBER WP No. 21149.

[13] Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47(1), 225-238.

[14] Chamberlain, G. (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in Longitudinal Analysis of Labor Market Data, edited by J. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

[15] Chamberlain, G. (1993): "Feedback in Panel Data Models," unpublished manuscript, No. 1656. Harvard-Institute of Economic Research.

[16] Chamberlain, G. (2010): "Binary response models for panel data: Identification and information," *Econometrica*, 78(1), 159-168.

[17] Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535-580.

[18] Chintagunta, P., E. Kyriazidou, and P. Perktold (2001): "Panel Data Analysis of Household Brand Choices," *Journal of Econometrics*, 103(1), 111-153.

[19] Das, M. (1992): "A Micro-econometric Model of Capital Utilization and Retirement: The Case of the Cement Industry," *Review of Economic Studies*, 59, 277-297.

[20] Dunne, T., S. Klimek, M. Roberts, and D. Xu (2013): "Entry, Exit and the Determinants of Market Structure," *RAND Journal of Economics*, 44(3), 462-487.

[21] Erdem, T., S. Imai and M. P. Keane (2003): "Brand and Quantity Choice Dynamics under Price Uncertainty," *Quantitative Marketing and Economics*, 1, 5-64.

[22] Erdem, T., M. Keane, and B. Sun (2008): "A dynamic model of brand choice when price and advertising signal product quality," *Marketing Science*, 27(6), 1111-1125.

[23] Ericson, R. and A. Pakes (1995): "Markov Perfect Industry Dynamics: A Framework for Empirical Work," *Review of Economic Studies*, 62, 53-82.

[24] Frederiksen, A., B. Honoré, B., and L. Hu, L. (2007): "Discrete time duration models with group-level heterogeneity," *Journal of Econometrics*, 141(2), 1014-1043.

[25] Hahn, J.,and W. Newey (2004): "Jacknife and analytical bias reduction for nonlinear panel data models," *Econometrica*, 72(4), 1295-1319.

[26] Hahn, J. anf G. Kuersteiner (2011): "Bias reduction for dynamic nonlinear panel models with fixed effects," *Econometric Theory*, 27(6), 1152-1191.

[27] Heckman, J. (1981): "The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process," in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press.

[28] Hendel, I., and A. Nevo (2006): "Measuring the Implications of Sales and Consumer Inventory Behavior. Econometrica 74, 1637-1674.

[29] Honoré, B. , and E. Kyriazidou (2000): "Panel data discrete choice models with lagged dependent variables," *Econometrica*, 68(4), 839-874.

[30] Honoré, B. , and E. Kyriazidou (2017): "Panel Vector Autoregressions with Binary Data," manuscript. Princeton University.

[31] Honorè, B., and E. Tamer (2006): "Bounds on parameters in panel dynamic discrete choice models," *Econometrica*, 74(3), 611-629.

[32] Kalouptsidi, M. (2014): "Time to build and fluctuations in bulk shipping," *American Economic Review*, 104(2), 564-608.

[33] Kalouptsidi, M., P. Scott, and E. Souza-Rodrigues (2015): "Identification of counterfactuals in dynamic discrete choice models," NBER working paper, No. 21527.

[34] Kano, K. (2013): "Menu Costs and Dynamic Duopoly," *International Journal of Industrial Organization*, 31(1), 102-118.

[35] Kasahara, H. (2009): "Temporary Increases in Tariffs and Investment: The Chilean Case," Journal of Business and Economic Statistics, 27(1), 113-127.

[36] Kasahara, H. and K. Shimotsu (2009): "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, 77(1), 135-175.

[37] Keane, M. and K. Wolpin (1997): "The career decisions of young men," *Journal of Political Economy*, 105, 473-522.

[38] Kennet, M. (1993) "Did Deregulation Affect Aircraft Engine Maintenance? An Empirical Policy Analysis," *The RAND Journal of Economics,* 24, 542-558.

[39] Lancaster, T. (2000): "The incidental parameter problem since 1948," *Journal of Econometrics*, 95(2), 391-413.

[40] Magnac, T. (2004): "Panel Binary Variables and Sufficiency: Generalizing Conditional Logit," *Econometrica*, 72(6), 1859-1876.

[41] Magnac, T., and D. Thesmar (2002): "Identifying dynamic discrete decision processes," *Econometrica*, 70(2), 801-816.

[42] Manski, C. (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica*, 55(2), 357-362.

[43] Miller, R. (1984): "Job Matching and Occupational Choice", *Journal of Political Economy*, 92, 1086-1120.

[44] Neyman, J. and E. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16(1), 1-32.

[45] Roberts, M. and J. Tybout (1997): "The Decision to Export in Colombia: An Empirical Model of Entry with Sunk Costs," *American Economic Review*, 87(4), 545-564.

[46] Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica,* 55, 999-1033.

[47] Rust, J. (1994): "Structural estimation of Markov decision processes," in R. E. Engle and McFadden (eds.) *Handbook of Econometrics Volume 4*, North-Holland. Amsterdam.

[48] Ryan, S. (2013) "The costs of environmental regulation in a concentrated industry," *Econometrica,* 80(3), 1019-1061.

[49] Slade, M. (1998): "Optimal Pricing with Costly Adjustment: Evidence from Retail Grocery Stores", *Review of Economic Studies*, 65, 87-108.

[50] Sweeting, A. (2013): "Dynamic product positioning in differentiated product markets: The effect of fees for musical performance rights on the commercial radio industry," *Econometrica*, 81(5), 1763-1803.

[51] Willis, J. (2006): "Magazine prices revisited," *Journal of Applied Econometrics*, 21(3), 337-344.