HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Unobserved Heterogeneity, State Dependence, and Health Plan Choices

Faculty Research Working Paper Series

Ariel Pakes
Harvard University

Jack Porter
University of Wisconsin-Madison

Mark Shepard
Harvard Kennedy School

Sophie Calder-Wang
University of Pennsylvania

July 2021

RWP21-020

Visit the **HKS Faculty Research Working Paper Series** at:
https://www.hks.harvard.edu/research-insights/publications?f%5B0%5D=publication_types%3A121

The views expressed in the **HKS Faculty Research Working Paper Series** are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). Papers may be downloaded for personal use only.

www.hks.harvard.edu

Unobserved Heterogeneity, State Dependence, and Health Plan Choices.

Ariel Pakes, Jack Porter, Mark Shepard, and Sophie Calder-Wang *

June 21, 2021

Abstract

We provide a new method to analyze discrete choice models with state dependence and individual-by-product fixed effects, and use it to analyze consumer choices in a policy-relevant environment (a subsidized health insurance exchange). Moment inequalities are used to infer state dependence from consumers' switching choices in response to changes in product attributes. We infer much smaller switching costs on the health insurance exchange than is inferred from standard logit and/or random effects methods. A counterfactual policy evaluation illustrates that the policy implications of this difference can be substantive.

1 Introduction.

Since Heckman (1978, 1981), distinguishing the impacts of unobserved heterogeneity from those of state dependence has been a central issue in empirical work in economics, as the distinction has implications for the interpretation and policy implications of many observed phenomena. The analysis of unemployment durations seeks to separate out the causal effects of being unemployed on future employment from unobserved heterogeneity in worker employability (see Kroft et al. (2013) and the articles cited therein). Both the marketing and I.O. literatures face the problem of distinguishing switching costs from unobserved preferences in explaining the constancy of individual purchasing patterns over time (see the review by Keane (1997) and more recently, Shin, Misra, and Horsky (2012)). Network models often need to distinguish between common preferences and the causal effects of the network (see for example, Sorensen (2006) and more recently Conley and Udry (2010)). A similar problem arises in distinguishing between the effects of moral hazard and adverse selection in evaluating policies designed to monitor behavior in insurance markets (Abbring et al., 2003).

In this paper, we develop a new method to estimate state dependence in a choice model that allows for flexible unobserved heterogeneity through individual-by-product fixed effects. We apply the method to the issue of understanding persistence in health insurance plan choices. This problem

*Pakes: Harvard University, apakes@fas.harvard.edu. Porter: University of Wisconsin-Madison, jrporter@ssc.wisc.edu. Shepard: Harvard University, mark_shepard@hks.harvard.edu. Calder-Wang: University of Pennsylvania, sophiecw@wharton.upenn.edu. We thank Hanbin Yang for truly outstanding research assistance, and Liran Einav, Jeremy Fox, Ben Handel, and Amanda Starc for comments. We also acknowledge the Massachusetts Health Connector (and especially Marissa Woltmann) for help in providing and interpreting the data. We gratefully acknowledge data funding from Harvard's Lab for Economic Applications and Policy.

is policy-relevant as governments set rules for market-based health insurance programs and the Affordable Care Act exchanges, Medicare Part D, and Medicaid managed care cover more than 75 million people and cost over $700 billion in public spending per annum in the U.S. alone. Recent applied work suggests that choice persistence driven by state dependence (e.g., switching costs); may lead to larger insurance markups (Ho, Hogan, and Scott Morton, 2017), may interact with problems created by adverse selection (Handel, 2013; Polyakova, 2016), and may lead to invest-then-harvest pricing dynamics (Ericson, 2014). It is unsurprising then that regulators often seek to encourage switching through reminders and outreach, with the idea that more active consumer shopping will lead to better market outcomes. However as noted by Dafny, Ho, and Varela (2013), if choice persistence is primarily driven by preference heterogeneity, the policy implications may be simply to encourage more product variety.

Separating preference heterogeneity from state dependence in health insurance demand is challenging. Health plans may differ on a complex bundle of attributes, including financial coverage, customer service, and medical provider networks. Preferences for provider networks (the key plan attribute in our empirical setting) are particularly challenging to predict because of their multi-dimensional nature and the individual-specific ways that preferences vary. For instance, patients may care strongly about provider coverage in the immediate neighborhood of their home or work, or coverage of specific doctors or hospitals with whom they have relationships (Shepard, 2020; Tilipman, 2020). This implies an individual-by-plan specific match factor unlikely to be captured by standard methods for dealing with unobserved heterogeneity, including coarse plan fixed effects or independently distributed random effects. Correctly estimating state dependence requires an econometric approach that allows for very flexible heterogeneity.

To deal with these challenges, we consider a semiparametric discrete choice model with lagged dependent variables and individual-by-product fixed effects. Consumer i at time t faces a set of choices $\mathcal{D}_t$ where $\#\mathcal{D}_t = D_t$. Given price $p_{d,i,t}$ for choice $d \in \mathcal{D}_t$ and last period's choice $y_{i,t-1} \in \mathcal{D}_{t-1}$, the utility associated with choice d at time t is

$$U_{d,i,t} = \Big(- p_{d,i,t} - \mathbf{1}\{y_{i,t-1} \neq d\}\kappa_0 \Big)\beta_i + \lambda_{d,i} + \varepsilon_{d,i,t} \tag{1.1}$$

where κ_0 represents the price equivalent of "switching" for individual i , β_i allows the importance of price to vary by individual, $\lambda_{d,i}$ denotes individual (additive) product preferences, and $\varepsilon_{d,i,t}$ captures the remaining unobserved variation in random utility. I.e. we estimate the importance of switching costs relative to price, allowing for a flexible tradeoff between these variables and other additively separable preferences.

The observed choice (or state) is

$$y_{i,t} = \arg\max_{d \in \mathcal{D}_t} U_{d,i,t}.$$

We consider various approaches to exploiting the information provided by switching in combination with different specifications for the distribution of $\{\varepsilon_{d,i,t}\}_{d \in \mathcal{D}_t, 1 \leq t \leq T}$. The fixed effects paradigm with fixed T is adopted throughout, and no restrictions are placed on the joint distribution of the individual effects and the prices or any other observed individual characteristics. We also do not use or assume any explicit restrictions from initial conditions.

In general, κ_0 is only partially identified (Honoré and Tamer, 2006). We generate moment inequalities with intuitive identifying implications for κ_0, and these inequalities are used to provide confidence bounds on κ_0. The inequalities are based on the relationship between consumer switching and changes in product attributes (in our case, price, as we condition on other attributes). An

example, which we return to in the empirical work, is a product j that experiences a large price increase in period $t-1$ followed by an even larger price reduction in t. Our method infers a lower bound on κ_0 based on the share of consumers who have switched away from j in $t-1$ and who do not switch back in t.

These bounds can then be used in either a deeper analysis of the factors associated with the $(\{\lambda_{d,i}\}_{d,i}, \{\beta_i\}_i)$ and/or, as we do below, in assessing the impact of counterfactual prices[1]. Our empirical work is focused on distinguishing between state dependence and unobserved heterogeneity in health insurance choices and follows other recent work on the same topic (Handel, 2013; Polyakova, 2016; Ho, Hogan, and Scott Morton, 2017). This is a market where switching behavior is "infrequent" (a precise definition is given below), but we will also consider the implications of our approach in empirical settings where switching is more frequent.

Before summarizing our results a word on what we do not do is in order. We do not consider the mechanism that generated the state dependence, which likely would depend on the application. Relatedly, equation (1.1) does not explicitly allow for forward looking agents. While these methods could be adapted to specifications that include additional lags of the dependent variable, we do not pursue that generalization here. The moment inequalities we derive are unlikely to provide a sharp characterization of the identifying information on κ_0, but we exploit variation in choices in a straightforward way that should appeal to practitioners.[2] Note that κ_0 is needed to analyze equilibrium responses to policy and/or environmental change. This is one reason our focus is on it, rather than on the quantiles or averages of utilities, which is the focus of Chernozhukov et al. (2013), or the treatment effect parameters defined in Torgovitsky (2019).

Related Econometric Literature. We build on two strands of the literature: papers that analyze discrete choice models with fixed effects and papers that add state dependence to that problem. The literature on discrete choice with fixed effects provides an analogue of "within" estimation in panel data models with continuous dependent variables where the between/within distinction has been a focus of analysis. Chamberlain (1980) shows how an assumption of "logit" disturbances generates a consistent conditional likelihood estimator for that problem. Manski's (1987) maximum score estimator provides consistent estimates for the binary choice problem with fixed effects and a non-parametric disturbance distribution. Papers by Shi, Shum, and Song (2018) and Pakes and Porter (2016), which we come back to below, use an assumption of stationarity of the marginal distribution of disturbances over time to obtain their estimators for multinomial problems. Also related is work by Tebaldi et al. (2019) that develops a method to estimate static demand for health insurance in a model with flexible, non-parametric preference heterogeneity.

Building on this fixed effects literature, Honoré and Kyriazidou (2000) allow for state dependence and fixed effects in point identified binary and multinomial choice problems. They use a conditioning argument and observations that are matched across periods to obtain their estimators. A recent paper by Honoré and Weidner (2020) considers a binary logit model with state dependence but does not require matching. Honoré and Tamer (2006) examine identified sets from a related model, and Khan et al. (forthcoming) investigate different assumptions on disturbances

[1] More generally there could be more than one observed characteristic of interest whose value changes over time for a given individual, and/or for which the form of its interactions with either β_i or λ_i can be specified a priori, in which case the target parameter would be a vector. In our empirical work we condition on cells with common observed characteristics, so using a single target parameter, which also simplifies the exposition, seems appropriate.

[2] One could analyze the distinction between the identified set defined by our moment inequalities and the sharp set in special cases where the sharp set is known as in the dynamic binary response model with discrete covariates (Khan et al., 2020).

using both conditioning and matching. Torgovitsky (2019) considers state dependence through a nonparametric dynamic binary potential outcome framework and provides an approach to computing sharp bounds on state dependent treatment effects under various sets of assumptions.

Methods Used. We begin by considering models with a nonparametric ε distribution. The nonparametric analysis adapts moment inequalities

- from Pakes and Porter (2016; henceforth $P\&P$), which analyzes discrete choice models with individual-by-product additively separable fixed effects but without state dependence, and
- from revealed preference (without disturbances, fixed effects, or state dependence; this literature dates back to Samuelson, 1938),

to our problem. Revealed preference generates inequalities that can be used either with or without a parametric distribution of disturbances. To generate either a meaningful lower bound ($\underline{\kappa} > 0$) or upper bound ($\overline{\kappa} < \infty$), the non-parametric revealed preference inequalities require a condition on switching behavior in the data (observing a minimum switching rate). This condition is not satisfied in our health insurance choice data where switching is relatively uncommon. So the empirical results from the revealed preference assumptions use the parametric version of the inequalities it generates.

Empirical Results. Our empirical results analyze health insurance choices in the Commonwealth Care (or "CommCare") program in Massachusetts, enacted as part of the "RomneyCare" reform. The program provided subsidized health insurance to low-income citizens via an insurance exchange that let consumers choose among competing private plans. The program started in 2007 and grew steadily during 2007 and 2008. We begin our analysis in 2009 at the time of the first large price change (conditioning on choices prior to this) and use plan switching behavior from 2009 to 2014, the year of transition to the Affordable Care Act exchange, for our empirical estimates.

The non-parametric extension of the P&P inequalities applied to our full sample generate an inordinate number of inequalities and only generates a lower bound for κ_0 at a significance level (an "α") of .05. The lower bound estimate $\hat{\underline{\kappa}} = \$20$ per month is about one-third of average (after-subsidy) consumer premiums in the market. The theory underlying the P&P inequalities indicate that there are four subsamples that should generate bounds and that we can use those subsamples without creating a selection problem, thus mitigating the impact of slack moments on the test statistics from the full sample. When we use these subsamples, we generate an upper bound estimate $\hat{\overline{\kappa}} = \$57$. The parametric revealed preference results we turn to next sharpens the lower bound, moving it up to \$32, but produces an upper bound of \$56 almost identical to the P&P results. For comparison the average premium per member per month of the different plans varied between \$48 and \$62.

To see how allowing for fixed effects impacts these results, we also provide estimates of comparison models that allow for state dependence but do not allow for fixed effects. In particular we estimate logit models with fixed effects interacted with: (i) increasing numbers of individual characteristics; (ii) random effects that condition on the initial choice, and (iii) random initial conditions (the initial condition estimator). The point estimates of κ_0 obtained range from \$118 to \$77. Prior results on this data are consistent with these findings (see Shepard (2020), who finds a point estimate of $\kappa_0 \approx \$100$). So the estimates that omit fixed effects are two to three times larger than those that allow for them. We examine the implications of this difference for a counterfactual pricing policy of interest, and find that they are likely to be rather dramatic.

Outline of Paper. We begin with the nonparametric results, first adapting the P&P inequalities to allow for state dependence and providing the associated empirical results, and then adapting the revealed preference inequalities. The revealed preference parametric inequalities for the parametric case are provided next, first without and then with the additional structure of extreme value disturbances. The latter are exceptionally easy to implement. Before going to the parametric revealed preference empirical results, we present the results from the parametric comparison models that do not allow for fixed effects. The empirical results for the parametric revealed preference bounds with fixed effects are provided next together with the counterfactual analysis. We conclude with a brief summary. All proofs are provided in Appendix C.

Notation. Let $\epsilon_{i,t} \equiv [\epsilon_{1,i,t}, \ldots \epsilon_{\mathcal{D},i,t}]$, $\epsilon_i \equiv [\epsilon_{i,1}, \ldots \epsilon_{i,T}]$, $\lambda_i \equiv [\lambda_{1,i}, \ldots, \lambda_{\mathcal{D},i}]$, $p_{i,t} \equiv [p_{1,i,t}, \ldots p_{\mathcal{D},i,t}]$, and $p_i \equiv [p_{i,1}, \ldots, p_{i,T}]$. While p_i denotes price in our application, it could include any time-varying observed covariates more generally.

2 P&P Approach

2.1 Identifying Inequalities

The following assumption underlies our first set of inequalities.

Assumption 2.1. *For any t, the disturbance $\epsilon_{i,t}$ is: (i) conditionally independent of p_i and $y_{i,s}$ for all $s < t$ given β_i and λ_i, and (ii) is stationary over time, that is*

$$\epsilon_{i,t} \,|\, p_i, y_{i,t-1}, y_{i,t-2}, \ldots, y_{i,0}, \beta_i, \lambda_i \;\sim\; \epsilon_{i,t} \,|\, \beta_i, \lambda_i \;\sim\; \epsilon_{i,1} \,|\, \beta_i, \lambda_i. \quad \square$$

Assumption 2.1 is common in dynamic panel settings. It includes strict exogeneity of the time-varying covariates p_i while placing no restrictions on the correlation between λ_i and p_i, or λ_i and β_i. Nor does it impose any restriction on the distribution of $\epsilon_{d,i,t}$ across choices d, so $\epsilon_{d,i,t}$ can be freely correlated with $\epsilon_{c,i,t}$, $\forall (c,d) \in \mathcal{D}^2$. However, Assumption 2.1 does not allow for dependence over time in $\{\epsilon_{i,t}\}$; the model filters the individual-specific dependence over time through $y_{i,t-1}$. We will assume the ϵ_i are identically distributed across individuals i, though, in principle, the identification results could be re-written to allow for non-identical distributions.

We now adapt the inequalities from the discrete choice model without state dependence provided in P&P (Lemma 2.2 below), to the model with state dependence. Consider any two periods t and s, with $t > s$. For a fixed κ, order choices by the *difference* in the structural part of the utility function between t and s, where the structural part of utility is defined as $U_{d,i,t} - \epsilon_{d,i,t}$ from equation (1.1). Since the λ_i and the β_i do not vary over time, the choice with the largest structural utility difference also does not depend on their values:

$$d_1(y_{i,t-1}, y_{i,s-1}, p_{i,t}, p_{i,s}; \kappa) = max_{d \in \mathcal{D}} \left[(-p_{d,i,t} - \{y_{i,t-1} \neq d\}\kappa) - (-p_{d,i,s} - \{y_{i,s-1} \neq d\}\kappa) \right], \quad (2.1)$$

while for $j = 2, \ldots D$, the choice with j^{th} largest difference is

$$d_j(y_{i,t-1}, y_{i,s-1}, p_{i,t}, p_{i,s}; \kappa) = max_{d \notin \{d_1, \ldots d_{j-1}\}} \left[(-p_{d,i,t} - \{y_{i,t-1} \neq d\}\kappa) - (-p_{d,i,s} - \{y_{i,s-1} \neq d\}\kappa) \right].$$

In words, d_1 is the choice whose structural component of random utility improves most between periods s and t (conditional on lagged choices), d_2 improves the next most, and so on. So if $y_{i,t-1} = y_{i,s-1}$, d_1 would be the option whose price falls most between s and t. In general, however,

d_1 depends on both price changes and the lagged choice in each period. This dependence on lagged choices is what allows us to generate moment inequalities that help bound κ_0.

Since $\epsilon_{i,t}|z_i, y_{i,t-1}, \beta_i, \lambda_i \sim \epsilon_{i,s}|z_i, y_{i,s-1}, \beta_i, \lambda_i$, the relative magnitude of the conditional probabilities for $y_{i,t} = d_1$ and $y_{i,s} = d_1$ depends only on the difference in the structural part of the utility. Moreover since the fixed effects are the same across periods, the difference in the structural part of utility over time for a given choice does not depend on the fixed effects. So when $\kappa = \kappa_0$ equation (2.1) ensures that the conditional probability of observing $d_1(\cdot;\kappa_0)$ in period t is greater than in period s. More generally, this argument leads to the following result.

Lemma 2.2. *Suppose Assumption 2.1 holds. Assume $t > s$, and $D_0 \subset \mathcal{D}$. If*

$$\min_{d\in D_0}\Big[- p_{d,i,t} - \{y_{i,t-1} \neq d\}\kappa_0 - \Big(- p_{d,i,s} - \{y_{i,s-1} \neq d\}\kappa_0\Big)\Big]$$
$$\geq \max_{c\notin D_0}\Big[- p_{c,i,t} - \{y_{i,t-1} \neq c\}\kappa_0 - \Big(- p_{c,i,s} - \{y_{i,s-1} \neq c\}\kappa_0\Big)\Big]$$

then

$$\Pr(y_{i,t} \in D_0|p_i, y_{i,t-1}, \beta_i, \lambda_i) \geq \Pr(y_{i,s} \in D_0|p_i, y_{i,s-1}, \beta_i, \lambda_i). \ \square$$

Let d_j^0 denote $d_j(y_{i,t-1}, y_{i,s-1}, p_{i,t}, p_{i,s}; \kappa_0)$. Then the choice sets (or the D_0) that satisfy the supposition of this lemma are $D_0 = \{d_1^0\}$, $\{d_1^0, d_2^0\}$, ..., and $\{d_1^0, \ldots, d_{D-1}^0\}$.

In words, the lemma states the following. If the structural utility for individual i (including any switching costs) for all options $d \in D_0$ improves by more than all options $c \notin D_0$ between periods s and t, then an individual will be more likely to choose an option $d \in D_0$ at time t than at time s. This is true regardless of the value of the unobserved fixed effects in λ_i (or of β_i provided it is positive).

Notice that the conditioning sets for the probabilities in the concluding inequality differ due to the presence of the lagged dependent variable at different points in time. This feature distinguishes the dynamic model from the static model, and underlies the need for a different argument for identification than that used in P & P.

To see how Lemma 2.2 can yield an inequality for identification in the dynamic model, consider the special case where $s = t - 1$ and D_0 is a singleton, e.g. $y_{i,t-1} = d_1^0$ and $D_0 = \{d_1^0\}$ in the supposition of Lemma 2.2. Applying Lemma 2.2 we obtain $\Pr(y_{i,t} = d_1^0|p_i, y_{i,t-1} = d_1^0, \beta_i, \lambda_i) \geq \Pr(y_{i,t-1} = d_1^0|p_i, y_{i,t-2}, \beta_i, \lambda_i)$. Multiplying both sides by $\Pr(y_{i,t-1} = d_1^0|p_i, y_{i,t-2}, \beta_i, \lambda_i)$ we obtain

$$\begin{aligned}\Pr(y_{i,t} = d_1^0, y_{i,t-1} = d_1^0|p_i, y_{i,t-2}) &= E[\Pr(y_{i,t} = d_1^0, y_{i,t-1} = d_1^0|p_i, y_{i,t-2}, \beta_i, \lambda_i)|p_i, y_{i,t-2}] \\ &\geq E[\big(\Pr(y_{i,t-1} = d_1^0|p_i, y_{i,t-2}, \beta_i, \lambda_i)\big)^2 |p_i, y_{i,t-2}] \\ &\geq \big(\Pr(y_{i,t-1} = d_1^0|p_i, y_{i,t-2})\big)^2. \end{aligned} \quad (2.2)$$

Here the first inequality follows from the lemma, and the second from Jensen's Inequality. Dividing both sides of (2.2) by $\Pr(y_{i,t-1} = d_1^0|p_i, y_{i,t-2})$ yields

$$\Pr(y_{i,t} = d_1^0|y_{i,t-1} = d_1^0, p_i, y_{i,t-2}) \geq \Pr(y_{i,t-1} = d_1^0|p_i, y_{i,t-2}), \quad (2.3)$$

as in Theorem 2.4 below.

Notice that the slackness in the last inequality generating (2.2) is due to the fact that it ignores the variance in $\Pr(y_{i,t-1} = d_1^0|p_i, y_{i,t-2}, \beta_i, \lambda_i)$ conditional only on $(p_i, y_{i,t-2})$. This conditional variance, in turn, depends on the variance of the λ_i and β_i. The $\{\beta_i, \lambda_i\}_i$ are explained by both

the observable and the unobservable determinants of utility, and the richer the set of observable characteristics that the analyst can condition on, the lower the conditional variance of the λ_i and β_i in the data generating the inequality, and the more powerful the inequality in (2.2). This motivates our decision to form moments from cells with common observable characteristics and $y_{i,t-2}$ in the empirical analysis which follows.

The result in the theorem to follow extends the argument in (2.2) and (2.3) in two ways. First, we broaden the argument to apply to choice probabilities of non-singleton sets. Lemma 2.2 considers the case where the conditioning set at time $t-1$ includes the lagged value $y_{i,t-1}$. To extend the argument in (2.2) from a singleton value d_1^0 to a set of choices D_0, we apply Lemma 2.2 for each value $y_{i,t-1} \in D_0$, which requires the supposition in Lemma 2.2 to hold for each value $y_{i,t-1} \in D_0$. Second, when $s < t-1$, $y_{i,t-1}$ only enters the inequality in the lemma through the conditional probability for $y_{i,t}$, which means $y_{i,t-1}$ can be allowed to take values different than $y_{i,t}$. The condition below generalizes the supposition of Lemma 2.2 to accommodate these cases.

Condition 2.3. *Assume $t > s$, and $D_0, D_1 \subset \mathcal{D}$. Given $p_{i,t}$, $p_{i,s}$, $y_{i,s-1}$, and κ_0, for all $d' \in D_1$,*

$$\min_{d \in D_0} \Big[- p_{d,i,t} - \{d' \neq d\}\kappa_0 - \Big(- p_{d,i,s} - \{y_{i,s-1} \neq d\}\kappa_0 \Big) \Big] \\ \geq \max_{c \notin D_0} \Big[- p_{c,i,t} - \{d' \neq c\}\kappa_0 - \Big(- p_{c,i,s} - \{y_{i,s-1} \neq c\}\kappa_0 \Big) \Big].$$

This condition ensures that structural utility differences for the choices in D_0 are larger at time t for any value of the lagged dependent variable in D_1 than at time s with a lagged dependent value of $y_{i,s-1}$.

Theorem 2.4. *Suppose Assumption 2.1 holds.*

(a) For $s = t-1$, for any choice set $D_0 = D_1$ satisfying Condition 2.3,

$$\Pr(y_{i,t} \in D_0 \,|\, p_i, y_{i,t-1} \in D_0, y_{i,t-2}) \;\geq\; \Pr(y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2})$$

(b) For $s < t-1$, for any choice sets D_0 and D_1 satisfying Condition 2.3,

$$\Pr(y_{i,t} \in D_0 \,|\, p_i, y_{i,t-1} \in D_1, y_{i,s} \in D_0, y_{i,s-1}) \;\geq\; \Pr(y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}) \quad \square$$

Remark 2.5. If $s = t-1$, Lemma 2.2 will generate $D-1$ inequalities. However since Condition 2.3 with $D_0 = D_1$ requires the assumption of the lemma to hold for every $y_{i,t-1} \in D_0$, the amount of inequalities Theorem 2.4(a) generates will depend on κ_0, with the maximum number equal to $D-1$ and the minimum of one. In contrast if, as in Theorem 2.4(b), $s < t-1$ the fact that Condition 2.3 allows the choice set D_1 for $y_{i,t-1}$ to be distinct from the choice set D_0 leads to potentially many more inequalities than D, though again the actual number will depend on κ. Also inequalities can be built for temporal sequences of sets each of which satisfies the appropriate condition.

Generating Empirical Inequalities: Examples

To see the usefulness of Theorem 2.4 in application, consider two simple cases with a singleton $D_0 = \{d\}$, and a single alternate option $c \neq d$. The theorem generates inequalities which can be implemented empirically. The simplest way to see the information these inequalities provide about the parameter κ_0 is to find the values of κ_0 that are ruled out by violation of the inequalities. That is, it is useful to examine the implications of the contrapositive of the theorem.

In the first example, we compare the probability of switching from choice c to choice d at time $t-1$, $\Pr(y_{i,t-1} = d \,|\, y_{i,t-2} = c, p_i)$, with the probability of staying with choice d at time t, $\Pr(y_{i,t} = d \,|\, y_{i,t-1} = d, y_{i,t-2} = c, p_i)$. With a positive cost of switching κ_0, the probability of "staying" would generally be expected to be larger than the probability of "switching", even if the relative price of choice d increased by a small amount from $t-1$ to t. Theorem 2.4(a) makes precise this intuition: if the relative price of d increases by less than twice the switching cost, then staying at time t is more likely than switching at time $t-1$.

Assume the relative price of d does increase between $t-1$ and t. Suppose, however, that the fraction switching from choice c to d at time $t-1$ is larger than the fraction staying with choice d at time t. That is, suppose the inequality in the conclusion of Theorem 2.4(a) fails:

$$\Pr(y_{i,t} = d \,|\, y_{i,t-1} = d, y_{i,t-2} = c, p_i) \;<\; \Pr(y_{i,t-1} = d \,|\, y_{i,t-2} = c, p_i).$$

This can only happen if the supposition of the theorem also fails, i.e. the inequality in Condition 2.3 is reversed. With a bit of rearrangement, this reversed inequality can be written as

$$2\kappa_0 \;<\; \Delta p_{d,i} - \Delta p_{c,i}, \quad \text{where } \Delta p_{d,i} \equiv p_{d,i,t} - p_{d,i,t-1}, \quad \text{and} \quad \Delta p_{c,i} \equiv p_{c,i,t} - p_{c,i,t-1}.$$

So a sufficiently large likelihood of switching implies an upper bound on how high switching costs can be.

In the second example, we obtain a lower bound on switching costs by comparing the probability of staying with choice d at time $t-2$ to switching from choice c to choice d at time t. This requires the subscriber to have switched to c in $t-1$, so we are considering a situation where the relative price of d rose in $t-1$ and then fell sharply in t. If the fall was sharp enough for the relative price of d to be considerably lower at t than in $t-2$, then the probability of switching into d at time t for the subscriber should be higher than the probability of staying with choice d at time $t-2$. The reversal of this inequality underlies our lower bound.

To formalize this intuition it is useful to start with Condition 2.3 and the implications of Lemma 2.2. Set $D_0 = \{d\}$, $D_1 = \{c\}$ and $y_{i,t-3} = d$, the inequality in Condition 2.3 can be written as:

$$\Delta\bar{p}_{c,i} - \Delta\bar{p}_{d,i} \geq 2\kappa_0, \quad \text{where } \Delta\bar{p}_{d,i} \equiv p_{d,i,t} - p_{d,i,t-2}, \quad \textit{and} \quad \Delta\bar{p}_{c,i} \equiv p_{c,i,t} - p_{c,i,t-2}.$$

Under this condition, Lemma 2.2 and Assumption 2.1 yield that

$$\Pr(y_{i,t} = d \,|\, y_{i,t-1} = c, y_{i,t-2} = d, y_{i,t-3} = d, p_i, \beta_i, \lambda_i) \;\geq\; \Pr(y_{i,t-2} = d \,|\, y_{i,t-3} = d, p_i, \beta_i, \lambda_i).$$

The lemma shows that if the relative price of choice d has gone down by more than twice the cost of switching, then the probability of switching at t will be greater than the probability of staying at $t-2$. Unfortunately, these probabilities are not directly useful empirically since we do not observe the individual effects λ_i or β_i. Theorem 2.4(b) aggregates this probability inequality to a more empirically useful conclusion:

$$\Pr(y_{i,t} = d \,|\, y_{i,t-1} = c, y_{i,t-2} = d, y_{i,t-3} = d, p_i) \;\geq\; \Pr(y_{i,t-1} = c, y_{i,t-2} = d \,|\, y_{i,t-3} = d, p_i). \quad (2.4)$$

Notice that the righthand side probability is smaller than the "aggregated" probability of staying at time $t-2$, $\Pr(y_{i,t-2} = d \,|\, y_{i,t-3} = d, p_i)$. This is the slackness introduced to enable aggregation in the theorem.

Finally, we can consider the contrapositive in this example. Suppose we observe that the relative price of d has decreased between times $t-2$ and t and yet the conclusion of the theorem in (2.4) fails:

$$\Pr(y_{i,t} = d \,|\, y_{i,t-1} = c, y_{i,t-2} = d, y_{i,t-3} = d, p_i) \;<\; \Pr(y_{i,t-1} = c, y_{i,t-2} = d \,|\, y_{i,t-3} = d, p_i).$$

Then, it must follow that Condition 2.3 also fails and the decline in the relative price of d must have been less than twice the switching costs:

$$2\kappa_0 \;>\; \Delta\bar{p}_{c,i} - \Delta\bar{p}_{d,i}. \tag{2.5}$$

To summarize, if a decline in the relative price of d does not induce sufficient switching into choice d, then switching costs must not be too low.

These examples show how Theorem 2.4 can be used to generate bounds on κ_0. In both examples we compare events at different points in time for the same individuals. So we hold the individual's preferences fixed. Then we consider events that force a trade-off between changes in relative prices and switching costs. These events must include switching to distinguish state dependence from heterogeneity. We turn now to showing how this reveals bound on κ in our data.

2.2 Empirical Results: Massachusetts Health Insurance Data.

We analyze health insurance plan choices made by enrollees in the Commonwealth Care ("CommCare") program in Massachusetts between 2009-2014. The program provided heavily subsidized insurance to low-income adults (earning less than 300% of the Federal Poverty Level) via a market featuring competing private health insurers. Five insurers participate in the market during our data period, with each insurer (by rule) offering a single plan. Program rules required each enrollee to make a separate choice; there was no family coverage, and kids were covered in the separate Medicaid program. Individuals make plan choices at two times: (1) when they join the market as a new enrollee, and (2) during an annual open enrollment month when they are allowed to switch plans. Because our focus is on switching costs, we study open enrollment choices, setting the prior choice (the state, $y_{i,t-1}$) equal to the individual's plan in the month prior to open enrollment.[3] For more detail on the data and the CommCare program see Shepard (2020); Finkelstein, Hendren, and Shepard (2019); McIntyre, Shepard, and Wagner (2021).

We want to capture switching costs that are not induced by changes in the individual's choice environment, just by prices, and this requires the choice set to be the same in the two periods we compare. We therefore remove comparisons for individuals who changed regions (there are five in the data), or who faced different plan offerings in the comparison periods, and use separate inequalities for each pair of income groups in years t and s. We distinguish income groups because subsidies—and therefore post-subsidy premiums—vary across five income groups (0-100% of poverty, and four 50% of poverty groups from 100-300%). Lower-income groups both pay lower premiums overall and have narrower premium differences across plans. This generates substantial price variation that we can use to estimate κ_0. Besides variation across income groups due to subsidies, price variation was limited by regulations. Prices could vary by region in 2009-2010 but not from 2011-on. No variation was allowed on other factors including age, gender, health status, or any other characteristics.

Our model assumes that individual-level unobserved plan preferences ($\lambda_{d,i}$) are stable over time. This is a sensible assumption given the nature of plans in the CommCare market. Coverage is heavily regulated, with all cost sharing and covered medical services completely standardized

[3]Individuals are also allowed to switch plans when they change income group or move across regions. We treat changes occurring at these times as separate choices for estimation purposes. There are a small number of instances where individuals switch plans outside of the standard open enrollment, and we drop these observations from the data.

across insurers. The only flexible plan attributes are provider networks.[4] These were largely stable during our sample period with one major exception. Network Health (one of our plans) drops Partners HealthCare (the state's largest medical system) from its hospital network at the start of 2012. To account for this, we treat Network as two different plans, one before and one after 2012, and apply the rules above with that understanding. There were no other major changes in the networks of the plans during our study period. However, one plan enters mid-sample (Celticare in 2010), and one plan (Fallon) exits several areas in 2011.

As noted to form the sample analogues of the inequalities in Theorem 2.4 we form cells with the same observed characteristics and $y_{i,s-1}$. The observed characteristics of a cell are denoted by x_i and are defined by the Cartesian product of; a) couple of years, b) region c) plan availability and d) income group. So the $\{\lambda_i, \beta_i\}_i$ represent differences in tastes among consumers with the same x_i and $y_{i,s-1}$. Recall that though the ratio of price sensitivity to switching cost is held constant, the price sensitivities themselves (the β_i in equation (1.1)) can vary in an arbitrary way, and are not identified in the non-parametric analysis.

Table 1 provides summary statistics on the data used, which is constructed from all the cells defined above that have more than 20 members. We then sum the inequalities generated by these cells across regions and plan availabilities to obtain our groups, and use these in estimation. There are 242 groups (defined by couple of years, income, and prior choice) and about 75,000 individual comparisons. Given Assumption 2.1 the variables defining our groups are condiionally independent of the values of the disturbances in the comparison periods, so we can use subsets of the groups in estimation without incurring a selection bias. As noted, the number of inequalities obtained in this way depends on the value of κ being tested but, when using the full sample is always large, varying between nine and twenty-four thousand, all but two of which will be slack.

Table 1: Summary Statistics for the Non-Parametric Estimator

(s, t)	Number of Members	Number of Groups	Number of Members Above Cutoff	Number of Groups Above Cutoff	Minimum Number of Moments	Maximum Number of Moments
(2009, 2010)	19,550	96	17,349	66	1,494	3,671
(2010, 2011)	13,989	96	13,181	76	3,181	8,748
(2012, 2013)	47,266	120	44,438	100	4,251	11,225
Total	80,805	312	74,968	242	8,926	23,644

Notes: The cell size cutoff is 20.

These inequalities are divided by their estimated standard errors, stacked, and the inequality with the largest negative value was calculated for each candidate κ. As suggested by Armstrong (2014), this became the sample value of the test statistic for that κ. The simulated value of the test statistic for the given κ was obtained once without any adjustment for slack moments, and once using the adjustment proposed in Romano, Shaikh, and Wolf (2014, henceforth RSW), and the variance covariance required for these calculations was obtained via a bootstrap.

[4]Preferences across plan provider networks are likely to be quite heterogeneous across individuals in hard-to-observe ways, making them a good candidate for our flexible fixed effects approach. For instance, Shepard (2020) and Tilipman (2020) find that consumers value provider networks heavily based on existing relationships with doctors and hospitals, which vary substantially even across enrollees living in the same location.

Figure 1: Choices and Premiums

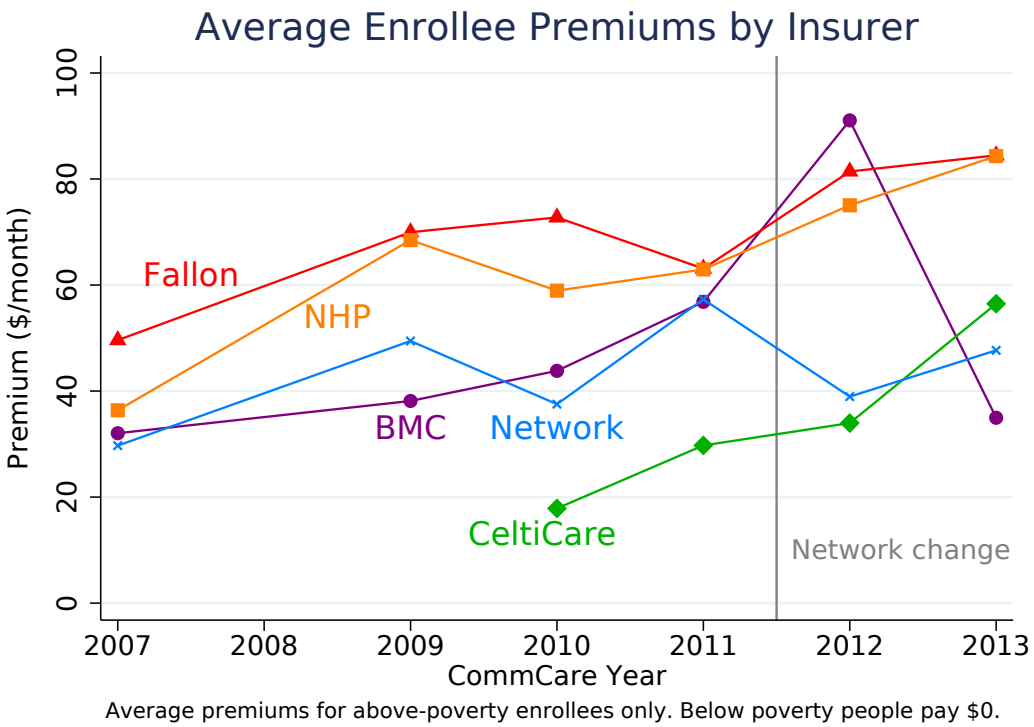


Figure 1 provides the average prices paid by consumers (i.e. after subsidy) by year and plan.[5] Our inequalities compare changes in choices to changes in prices, so the within-plan variance in prices over time underlies our set estimator of κ_0. The variance in prices within plan *and* time period is largely tied to income (though up to 2010 there was also regional variation in prices), which is the reason we separated groups by income.

The figure indicates that prices do go both up and down within a plan over time. This is largely a result of perceived changes in demand and plans adapting pricing strategies as they learned about the market and as regulatory rules changed. Importantly, with one exception, these do not appear to reflect changes in plan quality. As noted above, most insurance attributes, including cost sharing and covered medical services, were standardized by regulators so that all plans provided identical benefits within a given income group. The only significant non-standardized attribute is the covered network of doctors and hospitals, which are relatively stable over time (see Appendix A Figure 6) with one key exception: a major network narrowing by Network Health in 2012. We denote this narrowing by the vertical line in Figure 1 and account for this explicitly in our analysis by treating Network Health as a different plan on either side of this cutoff.

The largest price change in Figure 1 is the increase and then decrease in price of BMC in 2012-2013 relative to other plans. As noted there were no major changes in BMC's network or other quality attributes at this time. Instead, this change appears to reflect BMC's strategic response to a new competitive rule introduced by market regulators in 2012. The rule specified that new enrollees with incomes below the poverty line – who are fully subsidized, so pay $0 for any plan – were limited to choosing one of the two lowest-price plans (to the state). This created an auction-like dynamic in which the two lowest-bidding plans "won" access to this large group,

[5]Prices set in 2007 were locked in from 2007-08, which is why there are no separate points shown for 2008.

representing about half of new enrollees. In 2012, Network Health and CeltiCare bid low and won this auction, while BMC chose to raise its price and earn a larger margin on those members who did not move (BMC had the largest share in 2011). However, having lost almost half of its market share during 2012, BMC reversed course in 2013 and undercut both its competitors. This allowed it to rebuild its market share during 2013 leading into the important transition of CommCare into an Affordable Care Act exchange in 2014.

Table 2 summarizes the path of enrollment between 2011 and 2013 for people enrolled in CommCare in 2011. The top panel indicates that about two thirds of those enrolled in CommCare in 2011 had moved out of CommCare by 2013. This is a market with a lot of "churn" (partly induced by movements in and out of low-income eligibility due to employment changes). The bottom panel reports on switching behavior among subscribers who stayed in CommCare between 2011 and 2013. The fraction who switch plans in 2012 is 14.5%; BMC, the plan with the largest price increase, loses 19% of its 2011 subscribers. Though all prices changed in 2013, BMC is the only plan whose average price decreased. It saw 36% of those who switched out in 2012 switch back in 2013, while only 4% of those who switched out of other plans in 2012 switched back in 2013. These changes underlie much of the empirical work that follows.

Table 2: Statistics on Enrollment and Switching for 2011 Enrollees over 2011-2013

	All 2011 Enrollees	**By 2011 Plan**	
		BMC	All Other Plans
Number of Enrollees			
Total Enrollees in 2011	111,226	36,235	74,991
Leave Market before 2013	76,007	24,812	51,195
Stay in Market 2011-13	35,219	11,423	23,796
Switching Rates (among stayers in market)			
Switch Plans from 2011-2012	14.5%	19.0%	12.3%
Switch in 2012, Switch Back in 2013	2.5%	6.8%	0.5%
Switch in 2012, Do Not Switch Back 2013	11.9%	12.1%	11.8%

Note: The table shows statistics on enrollment and swithing rates over the 2011-13 period. The sample is people enrolled in CommCare in 2011 who are not in the below-poverty income group (who do not pay premiums so do not experience the premium changes shown in Figure 1), and the columns separate this group by plan in 2011. The top panel shows enrollment numbers, and the bottom panel shows switching rates among people who stay in the market from 2011-13.

Estimates. We begin with estimates that use the full dataset (Figure 2), then explain how theory directs us to choose particular subsets of the data, and finally show the results for those subsets (Figure 3).

The blue line in Figure 2 provides the value of the test statistic that the data generates for alternative values of κ. The long-dashed red line provides the five percent critical value for the κ values obtained from a test statistic that does not use a correction for slack moments, and the dotted red line is the test statistic when we use the RSW correction. Acceptable values of κ are all values where the blue line is lower than the red lines. The blue line crosses the red lines at $\kappa = \$19.6$ and since it remains below it thereafter, \$19.6 becomes $\hat{\underline{\kappa}}$, the lower bound of the 95% confidence set for κ_0 based on the full dataset. Notice, however, that the blue line is always noticeably above

zero, indicating that there are groups in the data which violate one or more inequalities at higher levels of κ, but the violations are not significant at $\alpha = .05$ when we use all the moments counted in Table 1.

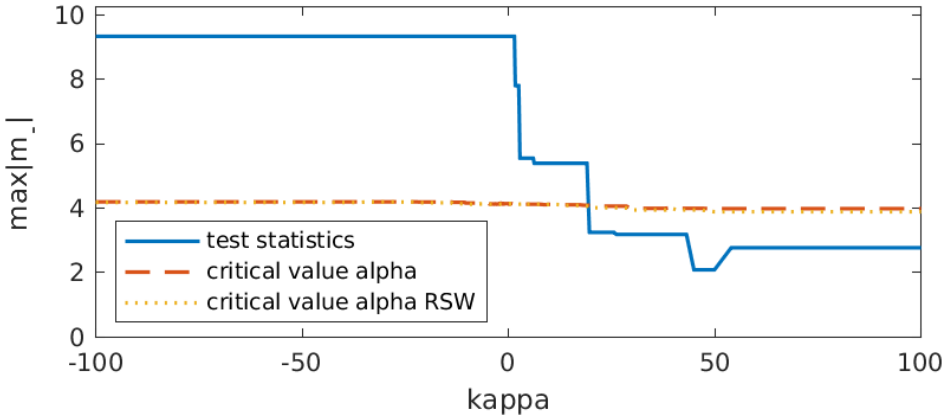


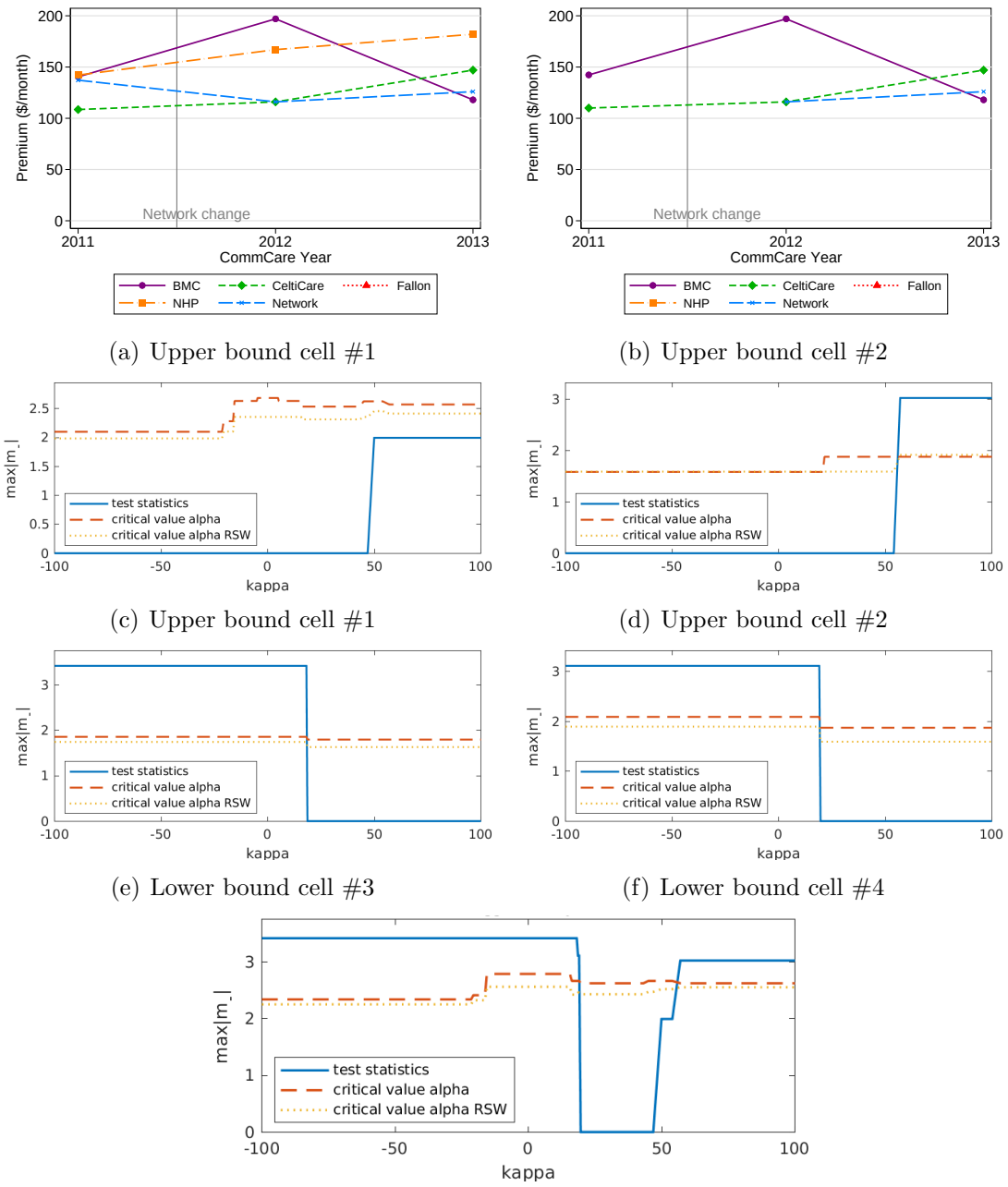
Figure 2: Non-parametric Estimates of Switching Costs (κ_0)

Note: The non-parametric estimator restricts the comparisons to individuals with the same choice sets at time s and t. Cell level moments are constructed, and then aggregated into groups. The lower bound identified is at \$19.6.

While the full dataset is a useful starting point, we also consider using only the conditional moment inequalities associated with groups expected to be most informative about the cost of switching. Since prices appear only in the conditioning sets of the moments, we use price information to determine these groups. The examples introduced following Theorem 2.4 illustrate how theory leads us to expect certain price movements to be especially helpful in providing upper and lower bound information on κ_0. Recall that the first example yielded an upper bound when switching to choice d in the previous period was more likely than staying with d in the current period. This circumstance is expected when a sharp relative price drop in choice d at $t-1$ induces sufficient switching into d at $t-1$, and is followed by a relative price increase of d at time t which induces sufficient switching back out of d. Generally, this is the situation for BMC's competitors from 2011-13.

The second example showed how a lower bound could be achieved when the relative price of d declines over two periods yet the likelihood of staying with choice d two periods ago followed by a switch away from d in the next period, is larger than the likelihood of a switch back to choice d in the current period. This might occur if there is a relative price increase in choice d at time $t-1$ and an even larger price fall in t. If a sufficient number of the subscribers who shifted out of d in $t-1$ did not shift back when its relative price was lower in t then when they chose it in $t-2$, we would infer a lower bound to κ. This describes the situation for BMC pricing over 2010-2013.

Figure 1 shows the large BMC price increase between 2010 to 2012 followed by the even larger decrease in 2013 that yielded the lower bound in Figure 2 via (2.5). Panel (a) of Figure 3 shows prices facing a particular income group in the Boston area, and panel (b) shows the prices for an income group in Western Massachusetts. These two panels show that there are groups with price changes in those years that satisfy the condition for the upper bound, that is for which $(p_{d,i,t} - p_{c,i,t}) - (p_{d,i,t-1} - p_{c,i,t-1}) > 0$ (with choice $c =$ BMC). The bottom panels summarize the test statistics for: these two groups (panels c and d), the two groups for which $0 < (p_{d,i,t-2} - p_{c,i,t-2}) - (p_{d,i,t} - p_{c,i,t})$ and hence might generate lower bounds (panels e and f), and finally in panel (g) the test when we use the inequalities generated by all four of the selected groups.



(a) Upper bound cell #1

(b) Upper bound cell #2

(c) Upper bound cell #1

(d) Upper bound cell #2

(e) Lower bound cell #3

(f) Lower bound cell #4

(g) Combining four cells

Figure 3: Average premiums and non-parametric estimators in two selected cells

Note: The figures show cells in the data that have price changes that satisfy the conditions for a bound. Cells are defined by individuals who are in a given region and follow a specific income group path across the years 2011-2013. They focus on the year pair (2012, 2013) for individuals whose lagged choice in 2011 is BMC to capture the large increase in BMC premium in 2012 and then the large drop in 2013. Plan prices differ across income groups due to subsidy schedules, and choice sets differ across regions. Panels (a) to (d) focus on two cells where we obtain upper bounds. Panels (a) and (b) show premiums of available plans. Panels (c) and (d) show test statistics and critical values. Panels (e) and (f) show test statistics and critical values based on two cells where we obtain lower bounds. Lastly, panel (g) shows test statistics and critical values combining the four cells in panels (c) to (f). Panels (a) and (c) correspond to the cell with income 250-300% of poverty in 2012 and 2013, in Boston with choice set: {BMC, CeltiCare, NHP, Network}. Panels (b) and (d) correspond to the cell with income 250-300% of poverty in 2012 and 2013, in Western MA with choice set: {BMC, CeltiCare, Network}. Panel (e) corresponds to the cell with income 100-150% of poverty in both 2012 and 2013, in Western MA with choice set: {BMC, Network}. Panel (f) corresponds to the cell with income 100-150% of poverty in 2012 and 150-200% of poverty in 2013, in Western MA with choice set: {BMC, Network}.

Panels (e) and (f) correspond to cells that underlie the lower bound in Figure 2. Panel (c) shows an indication of an upper bound at $\kappa = 49$ though it is not significant at our significance level of $\alpha = .05$, but panel (e) shows an upper bound of $\kappa = 54$ which is significant at $\alpha = .05$. When we put the four groups which satisfy our exogenous selection criteria together in panel (g) and use an $\alpha = .05$ we obtain the confidence interval $\kappa_0 \in [\hat{\underline{\kappa}} = 19.6, \hat{\overline{\kappa}} = 57]$.

2.3 Non-Parametric Revealed Preference.

Revealed preference and the random utility model in (1.1) imply that the probabilities of sequences of choices that difference out the fixed effects are determined by κ_0 and the distribution of disturbances. The bounds on κ_0 that this generates depend on the assumption made on the distribution of the disturbances. The next section considers parametric distributional assumptions. This section focuses on nonparametric assumptions on the disturbance distribution.

With at least one switch in choice across time, we show that information is available on κ_0 for arbitrary $\{\lambda_{d,i}\}_{d,i}$. At least three time periods of data are needed to observe a lagged dependent value prior to a switch in choices. For choices $c \neq d$, suppose that $y_{i,t} = d$ and $y_{i,s} = c$. Equation (1.1) tells us that if $(d,c) \in \mathcal{D}_t$ and $y_{i,t} = d$, then $U_{d,i,t} \geq U_{c,i,t}$. If $y_{i,t-1} = r$ this implies

$$\big(- [p_{d,i,t} - p_{c,i,t}] - [\mathbf{1}\{r \neq d\} - \mathbf{1}\{r \neq c\}]\kappa_0\big)\beta_i + [\lambda_{d,i} - \lambda_{c,i}] + [\epsilon_{d,i,t} - \epsilon_{c,i,t}] \geq 0.$$

Analogously if $(d,c) \in \mathcal{D}_s$ and $y_{i,s} = c$, then $U_{c,i,s} \geq U_{d,i,s}$ and

$$\big(- [p_{c,i,s} - p_{d,i,s}] - [\mathbf{1}\{y_{i,s-1} \neq c\} - \mathbf{1}\{y_{i,s-1} \neq d\}]\kappa_0\big)\beta_i + [\lambda_{c,i} - \lambda_{d,i}] + [\epsilon_{c,i,s} - \epsilon_{d,i,s}] \geq 0.$$

Note that when $s = t-1$, $r = c$ in the above inequalities. If we add these two inequalities, the fixed effects cancel and we obtain

$$\begin{aligned} 0 \;\leq\; & \big(- [(p_{d,i,t} - p_{c,i,t}) - (p_{d,i,s} - p_{c,i,s})] - [\mathbf{1}\{r \neq d\} - \mathbf{1}\{r \neq c\} \\ & -(\mathbf{1}\{y_{i,s-1} \neq d\} - \mathbf{1}\{y_{i,s-1} \neq c\})]\kappa_0\big)\beta_i + [\epsilon_{d,i,t} - \epsilon_{c,i,t} - (\epsilon_{d,i,s} - \epsilon_{c,i,s})] \end{aligned} \tag{2.6}$$

Since this result depends only on the double differences of certain variables we introduce notation for those differences,

$$\begin{aligned} \Delta\Delta p^{a,b}_{i,\tau,\tau'} \;&\equiv\; p_{a,i,\tau} - p_{b,i,\tau} - (p_{a,i,\tau'} - p_{b,i,\tau'}), \quad \Delta\Delta \epsilon^{a,b}_{i,\tau,\tau'} \;\equiv\; \epsilon_{a,i,\tau} - \epsilon_{b,i,\tau} - (\epsilon_{a,i,\tau'} - \epsilon_{b,i,\tau'}), \\ \Delta\Delta \mathbf{1}^{a,b}(r,r') \;&\equiv\; \mathbf{1}\{r \neq a\} - \mathbf{1}\{r \neq b\} - [\mathbf{1}\{r' \neq a\} - \mathbf{1}\{r' \neq b\}] \end{aligned}$$

and rewrite the revealed preference inequality in (2.6) as

$$\big(\Delta\Delta p^{d,c}_{i,t,s} + \kappa_0 \cdot \Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \big)\beta_i \leq \Delta\Delta \epsilon^{d,c}_{i,t,s}. \tag{2.7}$$

Note that $\Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \in \{-2,-1,0,1,2\}$. If $s = t-1$, then to obtain inequality (2.6) we set $r = c$ in which case $\Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2}) \geq 0$. If $s < t-1$, since $\beta_i \geq 0$, then all r with the same value of $\Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1})$ can be pooled together to form an inequality which bounds κ_0.

Theorem 2.6. *Let $\mathcal{F}^{d,c}_{t,s}(\cdot)$ denote the conditional distribution function of $\Delta\Delta\epsilon^{d,c}_{i,t,s}$ given $(p_i, y_{i,s-1}, \beta_i)$ and assume $d, c \in \mathcal{D}_t \cap \mathcal{D}_s$ and $d \neq c$. Then,*

(a) $\Pr(y_{i,t} = d, y_{i,t-1} = c | p_i, y_{i,t-2}, \beta_i) \leq 1 - \mathcal{F}^{d,c}_{t,t-1}((\Delta\Delta p^{d,c}_{i,t,t-1} + \kappa_0 \Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2}))\beta_i)$, *and*

(b) *Assuming $\kappa_0 \geq 0$,*

$$\sum_{r : \Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \geq \mathsf{y}} \Pr(y_{i,t} = d, y_{i,t-1} = r, y_{i,s} = c | p_i, y_{i,s-1}, \beta_i) \leq 1 - \mathcal{F}^{d,c}_{t,s}((\Delta\Delta p^{d,c}_{t,s} + \kappa_0 \mathsf{y})\beta_i). \quad \square$$

Part (a) of Theorem 2.6 follows directly from equation (2.7). For part (b) notice that the event defined by (2.7) yields a bound for all r where $\Delta\Delta\mathbf{1}^{d,c}(r, y_{i,s-1}) = \mathsf{y}$ and will contain the upper bound event for all r' such that $\Delta\Delta\mathbf{1}^{d,c}(r', y_{i,s-1}) \geq \mathsf{y}$, as long as $\kappa_0 \geq 0$. The inequality in (b) follows, and we show below why adding the associated probabilities is likely to be helpful.

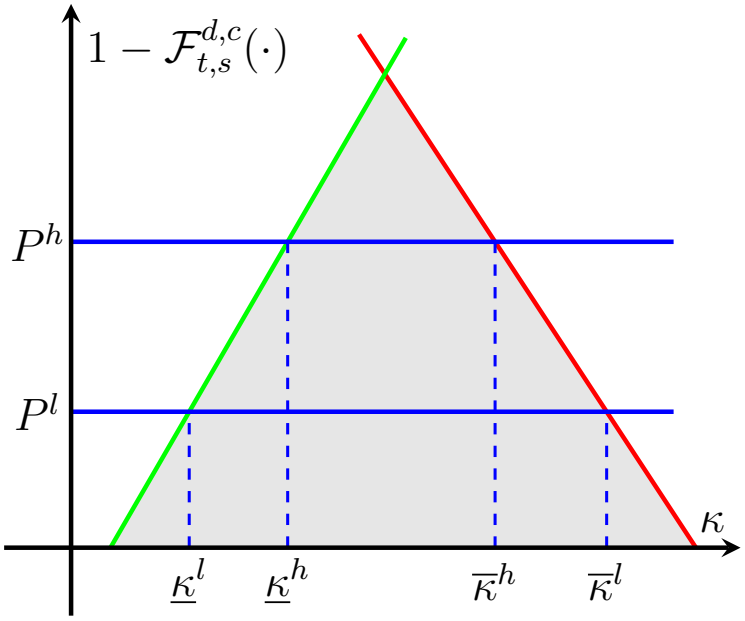


Figure 4: Identified set for κ_0

To see the potential usefulness of the Theorem 2.6 inequalities, Figure 4 illustrates the bounds on the parameter κ_0 in the case where $\mathcal{F}^{d,c}_{t,s}$ and β_i are *known*. Fix $\Delta\Delta p^{d,c}_{i,t,t-1}$ and suppose $\Delta\Delta\mathbf{1}^{d,c}(c, y_{i,s-1}) > 0$. Plotting $1 - \mathcal{F}^{d,c}_{t,s}(\Delta\Delta p^{d,c}_{i,t,s} + \kappa\Delta\Delta\mathbf{1}^{d,c}(c, y_{i,s-1}))$ as a function of κ, we obtain the red curve that is decreasing in κ. Then, the values of κ satisfying an inequality of the form $P \leq 1 - \mathcal{F}^{d,c}_{t,s}(\Delta\Delta p^{d,c}_{i,t,s} + \kappa\Delta\Delta\mathbf{1}^{d,c}(c, y_{i,s-1}))$ will be the κ values to the left of the red line at P illustrating how an upper bound on κ_0 would be implied by Theorem 2.6. When $\Delta\Delta\mathbf{1}^{d,c}(c, y_{i,s-1}) < 0$, $1 - \mathcal{F}^{d,c}_{t,s}(\Delta\Delta p^{d,c}_{i,t,s} + \kappa\Delta\Delta\mathbf{1}^{d,c}(c, y_{i,s-1}))$ is an increasing function of κ as illustrated by the upward sloping green line. In this case, the inequality $P \leq 1 - \mathcal{F}^{d,c}_{t,s}(\Delta\Delta p^{d,c}_{i,t,s} + \kappa\Delta\Delta\mathbf{1}^{d,c}(c, y_{i,s-1}))$ yields a lower bound on κ_0. Finally, the different values of P show that larger P values yield tighter bounds on κ_0, which explains how the accumulation of probabilities through the sum in part (b) can sharpen the information on κ_0.

The ability to make use of these potential bounds on κ_0 rests on the available information about the conditional distribution $\mathcal{F}^{d,c}_{t,s}$ and β_i. We consider two approaches. First, we take a nonparametric approach, and show that common stochastic assumptions on $\varepsilon_{i,t}$ lead to the conclusion that the median of $\Delta\Delta\epsilon^{d,c}_{i,t,s}$ is zero. With just this assumption and no assumptions on the β_i beyond nonnegativity, we show that the inequalities in Theorem 2.6 provide identifying information on κ_0 when the choice probabilities on the lefthand side of the inequalities in that theorem exceeds one half. Second, we assume that $\mathcal{F}^{d,c}_{t,s}$ comes from a given parametric distribution, and β_i is assumed to be a function of observables x_i. After providing the inequalities implied by Theorem 2.6 for a general $\mathcal{F}^{d,c}_{t,s}$, we consider the additional inequalities that result from assuming $\mathcal{F}^{d,c}_{t,s}$ is a logistic distribution.

The nonparametric assumption on the random utility errors considered thus far is Assumption 2.1. This assumes both stationarity and conditional independence across time. It is straightforward to show that the conditional distribution of $\Delta\Delta\epsilon_{i,t,s}^{d,c}$ is then symmetrically distributed around zero and $\mathcal{F}_{t,s}^{d,c}(0) = 0.5$. This median zero restriction is sufficient to generate a bound on κ_0 using the inequalities in Theorem 2.6. Moreover, this median zero restriction can also be obtained under stochastic assumptions that weaken the stationarity requirement. The extension we consider in our empirical work and which is likely relevant for other applied problems is the case where a pair of choices c and d are available in different time periods, but the other choices could change across time. Then it is sufficient to assume that the disturbance in random utility for this couple of choices is stationary and conditionally independent across time. Alternatively, the median zero restriction can be obtained under stochastic assumptions where stationarity is completely relaxed. It suffices to assume $\varepsilon_{d,i,t} - \varepsilon_{c,i,t}$ and $\varepsilon_{d,i,s} - \varepsilon_{c,i,s}$ are symmetrically distributed about zero. These could have different distributions, but given conditional independence across time, the conditional distribution of $\Delta\Delta\epsilon_{i,t,s}^{d,c}$ is symmetrically distributed around zero, so that $\mathcal{F}_{t,s}^{d,c}(0) = 0.5$. Note that the assumption that $\varepsilon_{d,i,t} - \varepsilon_{c,i,t}$ and $\varepsilon_{d,i,s} - \varepsilon_{c,i,s}$ are symmetrically distributed about zero would follow from an assumption of exchangeability of the disturbances across choices (Manski, 1975; Fox, 2007; Yan, 2013).

Next we state the implication of a median zero assumption when combined with the inequalities of Theorem 2.6.

Corollary 2.7. *Suppose* $\mathcal{F}_{t,s}^{d,c}(0) = 0.5$ *and* $\beta_i > 0$.

(a) If $\Pr(y_{i,t} = d, y_{i,t-1} = c \,|\, p_i, y_{i,t-2}) \geq 0.5$, *then* $\Delta\Delta p_{i,t,t-1}^{d,c} + \kappa_0 \Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2}) \leq 0$; *and*

(b) If $\sum_{r:\Delta\Delta \mathbf{1}^{d,c}(r,y_{i,s-1})\geq \mathsf{y}} \Pr(y_{i,t} = d, y_{i,t-1} = r, y_{i,s} = c \,|\, p_i, y_{i,s-1}) \geq 0.5$, *then* $\Delta\Delta p_{t,s}^{d,c} + \kappa_0 \mathsf{y} \leq 0$.

In this corollary, the revealed preference implications of the zero median restriction generate inequalities which condition on only observable variables, in contrast with the statement Theorem 2.6, and give the corollary its identifying power.

To examine the direction of the bounds on κ_0 provided by Corollary 2.7, suppose, for some y, $\sum_{r:\Delta\Delta \mathbf{1}^{d,c}(r,y_{i,s-1})\geq \mathsf{y}} \Pr(y_{i,t} = d, y_{i,t-1} = r, y_{i,s} = c \,|\, p_i, y_{i,s-1}) \geq 0.5$ as in part (b). When $\mathsf{y} > 0$, the corollary implies the upper bound

$$\kappa_0 \leq -\frac{\Delta\Delta p_{i,t,t-1}^{d,c}}{\mathsf{y}}.$$

When $\mathsf{y} < 0$ the corollary implies the lower bound

$$-\frac{\Delta\Delta p_{i,t,t-1}^{d,c}}{\mathsf{y}} \leq \kappa_0,$$

which will be useful when $\Delta\Delta p_{i,t,t-1}^{d,c} > 0$.

Without further restrictions on $\mathcal{F}_{t,s}^{d,c}(\cdot)$, when $\Pr(y_{i,t} = d, y_{i,t-1} = c | p_i, y_{i,t-2}) < 0.5$, no identifying information on κ_0 is provided by Theorem 2.6 and the additional assumptions in Corollary 2.7. Unfortunately, none of the groups in our data corresponding to different conditioning sets satisfy the probability condition required to obtain the non-parametric bounds given in Corollary 2.7.

Summary: Non-Parametric Bounds. The median zero condition on the twice differenced random utility disturbance does not require a disturbance distribution that is stationary over time. This median restriction

- only generates bounds if there is a group with a switching probability greater than one half.[6]

This condition is not satisfied in our (and we expect most) health insurance data, but it likely is satisfied in some retail markets (particularly those with both regular and"sale" prices).

The $P\&P$ bounds do require the stationarity assumption but can generate bounds with switching probabilities less than a half. In particular it generates

- a lower bound if price first rises and then falls to lower than its initial level, and not all people who switched out switch back, and
- an upper bound if people switch out because of a relative price rise and then do move back when its price falls sufficiently.

Using an $\alpha = .05$ the P&P bounds for our data generate the confidence interval $\kappa_0 \in [\$20, \$57]$.

3 Bounds from a Parametric ϵ Distribution.

The parametric model requires a distributional assumption for the random utility disturbances and a functional form for β_i. We will assume that β_i can be written as a function of observed variables x_i that does not vary over time, or $\beta_i = \beta(x_i)$. One could parametrize this function, but we do not pursue this option explicitly here. To parametrize the disturbance distribution, Assumption 2.1 is replaced by

Assumption 3.1. *Assume the conditional distribution of* $\Delta\Delta\epsilon^{d,c}_{i,t,s}$ *given* $(p_i, y_{i,s-1}, x_i)$ *takes a known parametric form with c.d.f.* $\mathcal{F}^{d,c}_{t,s}(\cdot;\sigma)$.

Typically Assumption 3.1 would be satisfied by assuming that $\varepsilon_{d,i,t}$ is conditionally independent across time and specifying parametric joint distributions across choices to yield $\mathcal{F}^{d,c}_{t,s}(\cdot;\sigma)$.

With Assumption 3.1, we can identify separate coefficients on price and the lagged dependent variable, so we re-write the utility equation (1.1) as

$$U_{d,i,t} = \left(-p_{d,it}\gamma_0 - \mathbf{1}\{y_{i,t-1} \neq d\}\delta_0\right)\beta(x_i) + \lambda_{d,i} + \epsilon_{d,i,t}, \tag{3.1}$$

$$\text{with} \qquad \gamma_0 \geq 0 \text{ and } \delta_0 \geq 0.$$

We remain focused on estimating the tradeoff between price sensitivity and switching costs, i.e. $\kappa_0 \equiv \delta_0/\gamma_0$, and as in the non-parametric analysis, assume that κ_0 does not dependent on x_i. We begin by examining the revealed preference inequalities that hold for any parametric distribution of the disturbances. Then we add the extra information we gain by assuming the disturbances at each t are extreme value ("logistic") distributions[7] (in which case $\mathcal{F}^{d,c}_{t,s}(\cdot;\sigma)$ has the analytic form given in section 3.1). Inequality (2.7) does not depend on the form of the distribution function,

[6]The actual condition is somewhat different for the upper and lower bound. For an upper bound we need two switches (three periods of data) with a probability greater than .5. For a lower bound we only need one switch but also a choice not to switch (so four periods of data) with a probability greater than .5.

[7]Other specific distributions would likely add different information. We use the extreme value distribution because it was used extensively in prior work and generates a transparent set of restrictions.

so provided $(d,c) \in \mathcal{D}_t \cap \mathcal{D}_s$, it is still the case that Theorem 2.6 holds, only now with the parameterization in equation (3.1) which gives us the following corollary.[8]

Corollary 3.2. *Suppose Assumption 3.1 holds, and assume $d, c \in \mathcal{D}_t \cap \mathcal{D}_s$ and $d \neq c$. Then,*

$$(a) \quad \Pr(y_{i,t} = d, y_{i,t-1} = c | p_i, y_{i,t-2}, x_i) \leq 1 - \mathcal{F}^{d,c}_{t,t-1}((\Delta\Delta p^{d,c}_{i,t,t-1}\gamma_0 + \Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2})\delta_0)\beta(x_i); \sigma), \quad \textit{and}$$

$$(b) \quad \sum_{r:\Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \geq \mathsf{y}} \Pr(y_{i,t} = d, y_{i,t-1} = r, y_{i,s} = c | p_i, y_{i,s-1}, x_i) \leq 1 - \mathcal{F}^{d,c}_{t,s}((\Delta\Delta p^{d,c}_{t,s}\gamma_0 + \mathsf{y}\delta_0)\beta(x_i); \sigma). \quad \square$$

Anatomy of the Parametric Inequalities. Each inequality generated by the parametric model generates a line which divides the (γ, δ) plane into acceptable and non-acceptable half-spaces. The slope and/or quadrant of the acceptable half-space differs with

- the sign of $\Delta\Delta y^{t,s}_{c,d}$ (which can be greater than, less than, or equal to zero), and
- the sign of $\Delta\Delta p^{d,c}_{t,s}$ (which can be greater than or less than zero).

In addition, if the the median of $\mathcal{F}^{d,c}_{t,s}(\cdot) = 0$, then $\mathcal{F}^{d,c}_{t,s}(\Pr(\cdot|\cdot))^{-1}$ differs in sign according as $\Pr(\cdot|\cdot) \lesseqgtr 1/2$, and the six inequalities have different implications for the two cases. Accordingly there are twelve cases to consider. Appendix B considers each with explanatory graphs.

Since none of the relevant probabilities are greater than a half in our data, there are only six cases to consider and Appendix B shows that only four generate restrictions in the appropriate quadrants.[9] We use all inequalities generated by these four in our empirical analysis.

3.1 The Magic of Logits

Implementation of the parametric revealed preference approach in empirical work requires functional forms for $\mathcal{F}^{d,c}_{t,s}(\cdot)$, and any particular choice may well generate additional restrictions on the parameters. As in much of the prior empirical work and the comparison models we turn to below, we assume a Gumbel (logistic) distribution for ϵ and explore its additional implications.

Assumption 3.3. *(a) Assume $\varepsilon_{i,t}$ is independent of the conditioning set $(p_i, y_{i,t-1}, \ldots, y_{i,0}, x_i, \lambda_i)$, and (b) $\varepsilon_{1,i,t}, \ldots, \varepsilon_{\mathcal{D},i,t}$ are independent (and identically distributed) across choices, where $\varepsilon_{1,i,t}$ has a standard Gumbel distribution.*

Assumption 3.3 yields the traditional logit form for the choice probabilities,

$$\begin{aligned}\mathcal{P}_{d,i,t|y_{i,t-1}} &\equiv \Pr(y_{i,t} = d \,|\, p_i, y_{i,t-1}, x_i, \lambda_i) \\ &= \frac{exp[(-p_{d,i,t}\gamma_0 - \mathbf{1}\{y_{i,t-1} \neq d\}\delta_0)\,\beta(x_i) + \lambda_{d,i}]}{\sum_r exp[(-p_{r,i,t}\gamma_0 - \mathbf{1}\{y_{i,t-1} \neq r\}\delta_0)\,\beta(x_i) + \lambda_{r,i}]} \equiv \frac{\mathcal{N}_t(d, y_{i,t-1}) e^{\lambda_{d,i}}}{\mathcal{M}_t(y_{i,t-1}, \lambda_i)}.\end{aligned}$$

[8] Corollary 3.2 does not exhaust the inequalities generated by the model. When $s < t-1$, it derives an inequality for the probability of $(y_{i,t} = d, y_{i,t-1} = r, y_{i,s} = c)$; a different inequality can be obtained by integrating out $y_{i,t-1}$ conditional on $y_{i,s-1}$. To use that inequality that we have to account for the fact that we do not know what $y_{i,t-1}$ would have been had $y_{i,s} \neq c$, and replace the integrand for that case with the max operator. We experimented with this inequality in our early empirical work but it did not generate binding constraints.

[9] These are the inequalities for the following cases: all three cases corresponding to a positive price change ($\Delta\Delta p^{d,c}_{t,s} > 0$), and the case with a negative price change ($\Delta\Delta p^{d,c}_{t,s} < 0$) and $\Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2}) > 0$ (or $\mathsf{y} > 0$), see Appendix B for details.

This implies that (conditional on $y_{i,t-2}$) the ratio of the probability of choosing d at t and c at $t-1$, to the probability of choosing c at t and d at $t-1$ is

$$\frac{\mathcal{P}_{d,i,t|c}}{\mathcal{P}_{d,i,t-1|y_{i,t-2}}}\frac{\mathcal{P}_{c,i,t-1|y_{i,t-2}}}{\mathcal{P}_{c,i,t|d}} = \frac{\mathcal{N}_t(d,y_{i,t-1}=c)}{\mathcal{N}_{t-1}(d,y_{i,t-2})}\frac{\mathcal{N}_{t-1}(c,y_{i,t-2})}{\mathcal{N}_t(c,y_{i,t-1}=d)} \times \frac{\mathcal{M}_t(y_{i,t-1}=d,\lambda_i)}{\mathcal{M}_t(y_{i,t-1}=c,\lambda_i)}$$

And, for this particular comparison of choices, the fixed effects in the numerator terms cancel. Moreover, the ratio of denominators can be bounded by functions that do not depend on the fixed effects: $exp(\delta_0\beta(x_i)) \geq \mathcal{M}_t(y_{i,t-1}=d,\lambda_i)/\mathcal{M}_t(y_{i,t-1}=c,\lambda_i) \geq exp[-\delta_0\beta(x_i)]$. This gives us the inequalities

$$\begin{aligned} & exp[\delta_0\beta(x_i)]\frac{\mathcal{N}_t(d,y_{i,t-1}=c)}{\mathcal{N}_{t-1}(d,y_{i,t-2})}\frac{\mathcal{N}_{t-1}(c,y_{i,t-2})}{\mathcal{N}_t(c,y_{i,t-1}=d)} \\ \geq \quad & \frac{\mathcal{P}_{d,i,t|c}}{\mathcal{P}_{d,i,t-1|y_{i,t-2}}}\frac{\mathcal{P}_{c,i,t-1|y_{i,t-2}}}{\mathcal{P}_{c,i,t|d}} \quad \geq \quad exp[-\delta_0\beta(x_i)]\frac{\mathcal{N}_t(d,y_{i,t-1}=c)}{\mathcal{N}_{t-1}(d,y_{i,t-2})}\frac{\mathcal{N}_{t-1}(c,y_{i,t-2})}{\mathcal{N}_t(c,y_{i,t-1}=d)}. \end{aligned}$$

So the ratio of the odds of choosing $(y_{i,t-1}=c, y_{i,t}=d)$ to $(y_{i,t-1}=d, y_{i,t}=c)$ is bounded by functions that are independent of the $\{\lambda_i\}_i$. This finding is reminiscent of Chamberlain (1980) who derived a conditional likelihood that did not depend on the $\{\lambda_i\}_i$ for the static multinomial logit panel case (the model with no lagged dependent variable). Once we allow for a lagged dependent variable and rearrange terms we get the inequalities in the theorem that follows for choice probabilities at t and $t-1$. A more detailed argument shows that similar inequalities are valid for the odds ratio at t and $t-2$ (see Remark 3.6).

Theorem 3.4. *Suppose Assumption 3.3 holds, $s \in \{t-1, t-2\}$, and $(d,c) \in \mathcal{D}_t \cap \mathcal{D}_s$. Then,*

$$exp\left[\gamma_0\left(\Delta\Delta p_{t,s}^{c,d}\right)\beta(x)\right] \leq \frac{\Pr(y_{i,t}=d, y_{i,s}=c \,|\, p_i, y_{i,s-1}=c, x_i=x)}{\Pr(y_{i,t}=c, y_{i,s}=d \,|\, p_i, y_{i,s-1}=c, x_i=x)} \leq exp\left[\left(2\delta_0+\gamma_0\left(\Delta\Delta p_{t,s}^{c,d}\right)\right)\beta(x)\right]. \quad \square$$

Remark 3.5. Theorem 3.4 compares the probability of sequences that start at $y_{i,s-1}=c$ and go to $(y_{i,s}=c, y_{i,t}=d)$ versus those that go to $(y_{i,s}=d, y_{i,t}=c)$. Under the logit assumption, both sequences have positive probability, and the first inequality provides upper [lower] bound information on γ_0 when $\Delta\Delta p_{t,s}^{c,d}$ is positive [negative]. The second inequality yields lower bound information on δ_0, so together these inequalities provide a lower bound to $\kappa_0 = \delta_0/\gamma_0$.

To obtain upper bound information on κ_0, we incorporate the inequalities in Corollary 3.2. For example, when $\Delta\Delta p_{t,t-1}^{c,d} < 0$ and $\Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2}) > 0$, the inequality in Corollary 3.2(a) yields upper bound information on δ_0, and the first logit inequality in Theorem 3.4 yields lower bound information on γ_0. Together these inequalities generate upper bound information on κ_0. If $\Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2}) > 0$, then, regardless of the sign of $\Delta\Delta p_{t,t-1}^{c,d}$, Corollary 3.2(a) provides upper bound information on δ_0, which given the logit inequalities yields upper and lower bounds to κ_0.

Remark 3.6. Theorem 3.4 does not exhaust the additional inequalities available when the disturbance distribution is logistic. For completeness, additional inequalities are stated as Theorem C.2 in Appendix C and cover the cases: (i) $y_{i,s-1} = r \notin \{c,d\}$; and (ii) $s < t-2$. In different applications, these additional inequalities could be quite useful, but here they are relegated to the appendix as our data does not have groups of sufficient size to exploit them.

4 Parametric Empirical Results.

Table 3 provides summary statistics for the data and inequalities used in the parametric analysis. The only difference between the sample used for the non-parametric analysis (described in section 2.2) and that used in the parametric analysis, is that in the parametric analysis we keep groups who face different plan offerings in the comparison periods (recall that our assumptions ruled this out for the non-parametric analysis). This increases the size of the data set considerably. The number of inequalities available for the parametric analysis is, on the other hand, much smaller than in the non-parametric analysis.

Table 3: Summary Statistics for the Parametric Estimator

(s, t)	Number of Members	Number of Groups	Number of Members Above Cutoff	Number of Groups Above Cutoff	Number of Moments
(2009, 2010)	59,322	100	32,738	69	248
(2009, 2011)	39,955	100	24,740	78	296
(2010, 2011)	59,629	100	34,300	83	217
(2012, 2013)	69,441	125	43,138	99	522
Total	228,347	425	134,916	329	1,283

Notes: The cell size cutoff is 20.

We begin with a subsection describing empirical findings from models that allow for state dependence but not flexible fixed effects. This will enable us to compare the results from the inequalities in section 3 to those from the point identified models that have been used to model state dependence in prior work. All results, both from the comparison models and from the models of section 3, assume that the distribution of the disturbances is logistic.[10] Note that in this case the distribution of the double difference of disturbances (for $\Delta\Delta\epsilon_{t,s}^{d,c}$) needed to form the inequalities from Corollary 3.2 is analytic, which simplifies computation of those inequalities.[11]

4.1 Comparison Models: State Dependence Without Fixed Effects.

Table 4 summarizes the results from a number of specifications. The estimate of the switching cost is always obtained as the ratio of the lagged dependent variable coefficient to the price coefficient.[12]

[10] We have also done several of the comparison models assuming normal errors. These generated modestly higher values for the state dependence coefficient.

[11] That distribution and its density have analytic forms for the logit case, which are

$$F(y) = \frac{exp(y)(y-1)+1}{(exp(y)-1)^2}, \text{ and } f(y) = \frac{exp(y)(exp(y)(y-2)+y+2)}{(exp(y)-1)^3}.$$

[12] Its standard error is obtained from a Taylor expansion (i.e., the Delta method), which in this context should be accurate as all the price coefficients are two or more orders of magnitude greater than their standard errors. As in all discrete choice models the comparison models require a normalization. We normalize the variance of the disturbance to one. So both the coefficient of price and of the lagged dependent variable should be thought of as the variable's coefficient divided by this standard error.

Table 4: Multinomial Logit Estimation

	Simple	Plan Dummies	Plan × Region Dum.	Detailed Plan Dum.	Plan Dum. + RE	+ New Enr (Plan Dum.)	+ New Enr (Dum. + RE)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Normalize $\epsilon_{i,j,t}$ to EV1							
Switching Cost (δ)	-4.157	-4.234	-4.184	-4.184	-4.549	-4.028	-4.713
	(0.008)	(0.009)	(0.010)	(0.010)	(0.015)	(0.006)	(0.010)
Price Coefficient (β)	-0.0361	-0.0507	-0.0529	-0.0541	-0.0539	-0.0342	-0.0423
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0004)	(0.0001)	(0.0002)
Normalize β to 1							
Switching Cost ($\kappa = \delta/\beta$)	115.20	83.50	79.03	77.34	84.42	117.60	111.40
	(0.75)	(0.43)	(0.41)	(0.40)	(0.43)	(0.44)	(0.43)
Plan Dummies	—	Yes	Yes	Yes	Yes	Yes	Yes
Plan Dummies × Region	—	—	Yes	Yes	—	—	—
× Area, Age-Sex, Illness	—	—	—	Yes	—	—	—
Plan Random Effects	—	—	—	—	Yes	—	Yes
N Plan Dummies + REs	0	5	22	247	9	5	10
N Individuals × Years	1,457,682	1,457,682	1,457,682	1,457,682	1,457,682	4,255,857	4,255,857

The first four columns of the table present results from specifications in which the individual-specific fixed effects used in the inequality analysis are replaced with increasingly detailed interactions of individual characteristics with plan dummies. Column (1) has no plan dummies; column (2) has simple plan dummies; column (3) interacts the plan dummies with region; and column (4) interacts them with twenty age-sex groups, with thirty eight geographic areas ("service areas" determined by the state), and with three chronic illness groups (where the three sets of interactions are additively separable). After excluding interactions with no observations, this generates 247 dummy variables. The switching cost estimate declines monotonically as we add interactions, from \$115.20 (0.75) to \$77.34 (0.40), where here and below the numbers in parentheses are standard errors. Notably, these estimates of κ_0 are all substantially larger than the upper bound of \$57 generated by the non-parametric P&P inequalities.

Next we replace the fixed effects in the inequality analysis with random effects. That is we interact the plan dummies with agent-specific independent normal random variables that are held constant over the period the individual is observed, and use simulated maximum likelihood to estimate. We begin the random effect analysis by allowing for random effects conditional on the first observed choice. So this analysis assumes both that: (i) the within group variance in the plan specific effects is normal with variances that vary by plan (but not by group), and (ii) is uncorrelated with the initial observed choice. This mimics what researchers have done in related problems when they do not have sufficient information on the actual initial choices of individuals. The results from this specification are provided in column (5) of the table. The random effects model generates a switching cost of \$84.42 (0.43), and estimated dummy variables for the plan and standard errors for the random effects which are both highly significant (the t-values for the standard errors varied from eight to over fifty), for all but the smallest plan (Fallon).

We are in the enviable position of knowing the first time a consumer enters the Massachusetts exchange. So provided we are willing to assume that any pre-exchange health choice of these individuals does not influence their behavior on the exchange, we can implement an "initial conditions estimator" that allows for normal random draws on preferences for the exchange's plans that are known to the consumers before making their first choice. Column (7) provides the simulated max-

imum likelihood estimates for this specification.[13] Since column (7) adds the first choice to the switching choices analyzed in columns (1) to (5), it uses a different data set than those columns did. So for comparison column (6) uses the column (7) data in a model without random effects; i.e. column (6) mimics column (2) but uses the data set used in column (7).

When we include initial choices and simple plan dummies (but no random effects) in column (6), our estimated switching cost is \$117.60 (0.44), noticeably larger than in the analogous specification without initial choices (column (2)). When we also include random effects in column (7), the plan dummies and the estimated standard deviations of the random effects for all plans are estimated to be even larger than those from column (5), with t-values for the estimated variances of the plan specific random effects for all plans (including Fallon) now ranging from twelve to over one hundred. Perhaps more surprising is that the estimates of κ_0 in both columns (6) and (7) are quite similar at \$117.60 (0.44) and \$111.40 (0.43). One interpretation of this in light of the finding of a lower estimate of κ_0 in column (5) is that due to the experience they had in making health insurance choices before entering CommCare, consumers priors when entering the program were that plans differed in their coverage, out of pocket payments, and the like. After entering they learn that regulation requires these features to not vary across plans. As a result they become more price sensitive[14].

We conclude that models that do not allow for individual by product fixed effects generate estimates of κ_0 that lie somewhere between \$77 and \$118. This accords well with published work on this data which generates estimates of about \$100 (see Shepard (2020)). Recall from Figure 1 that average monthly premiums ranged from \$20 to \$90. So the models without fixed effect generate switching costs which are four to five times the average monthly premium for the lowest cost plan and about equal to the average monthly premium for the highest cost plan. These estimates are noticeably larger than the non-parametric upper bound obtained from the P&P analysis. One question that remains is how much of the difference can be attributed to the logit assumption and how much to the absence of fixed effects.

4.2 Parametric Inequality Estimators: State Dependence with Fixed Effects.

This subsection uses the inequalities from Corollary 3.2 and Theorem 3.4, and the sample described in Table 3 (also used in columns (1) to (5) of Table 4), to estimate bounds on κ_0. This generates over 1,200 inequalities from about 330 groups with an average size of over 400 individuals.

When we used the estimation algorithm described in section 2.2 for the current specification, the simulated value of the test statistic obtained from the normal approximation to the distribution of the moments often implied probabilities that were negative, rendering the assumptions underlying that asymptotic approximation inappropriate. We present the point estimate from that estimation algorithm but do not want to rely on its simulated test statistics for inference. Instead we use the Bayesian approach proposed by Kline and Tamer (2020) with the implementation in Chamberlain

[13] We also tried to use the estimator suggested by Honoré and Kyriazidou (2000), an estimator which does allow for both fixed effects and switching costs. However, their method requires observing four-period sequences in which a person switches plans, followed by two periods in which both choice sets and plan attributes (here, prices) are unchanged. Even though our data contains observation on 623,000 individuals making 1,877,000 individual choices, these restrictions are quite limiting because the data include frequent price and choice set changes. Imposing the restrictions the Honore and Kyriazadou estimator requires leaves us with data on 36 individuals and 144 choices. This is simply not enough data to obtain estimates with reasonable precision.

[14] For a model with fixed effects that explicitly allows for Bayesian learning see Aguirregabiria et al. (2021).

and Imbens (2003). This combines an uninformative prior with the data to generate a posterior distribution for the probabilities.[15] We then take draws from this posterior, calculate the (possibly set-valued) estimate of the parameters that minimizes the sup-norm of the negative part of the inequalities for each draw, and then find a conservative 95 percent confidence set for γ_0, δ_0, and separately for $\kappa_0 \equiv \delta_0/\gamma_0$.

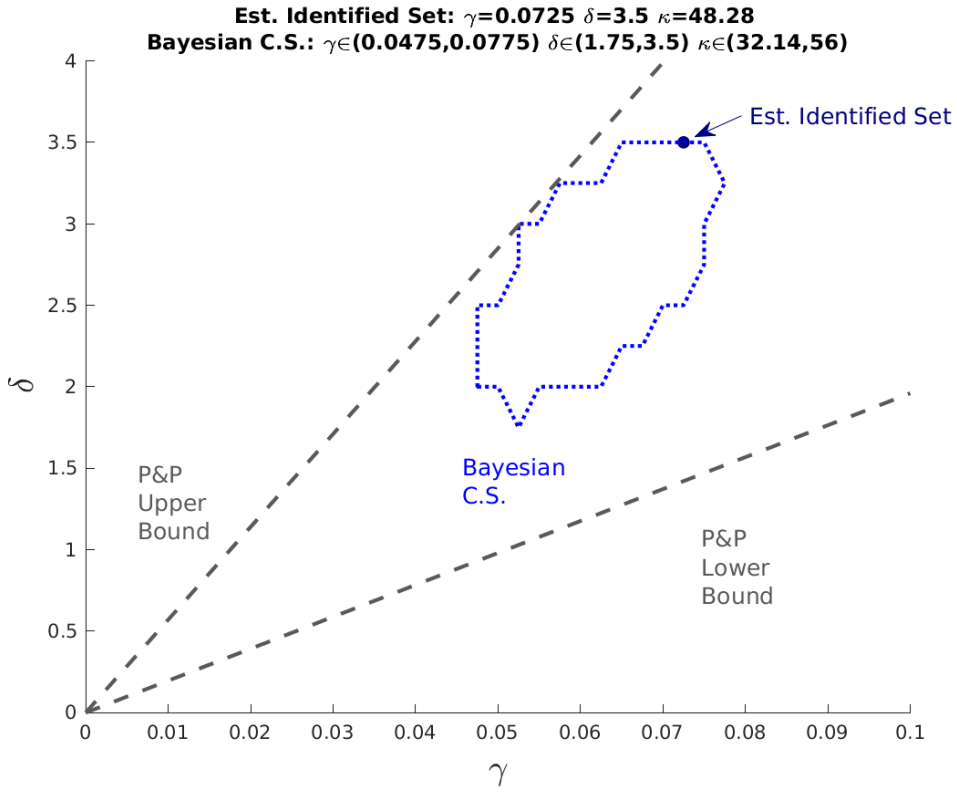


Figure 5: Identified set and Bayesian confidence set for γ_0 and δ_0.

Notes: Cell level moments based on revealed preference and magic of logit inequalities are constructed, and then aggregated into groups.

The results are plotted in (γ, δ) space in Figure 5. The point estimate from minimizing the largest of the negative parts of the moments is given by the dark blue dot. The 95% confidence sets for γ_0, δ_0 and linear combinations of the two are obtained from the 2.5% and 97.5% quantiles of the distribution of their lower and upper bounds found from the posterior draws. The accepted (γ_0, δ_0) combinations are given by the area interior to the shape produced by the blue dots in the figure. The (γ_0, δ_0) combinations that generated the lower and upper P&P bounds for κ_0 are given by the dashed grey lines.

The "point estimate" of κ_0 from the moment minimization problem was $\hat{\kappa}$=\$48. The Bayesian

[15]Treating the choice probabilities for each cell as a multinomial distribution, Chamberlain and Imbens (2003) show that the Dirichlet distribution with parameters set to the observed frequencies is the posterior distribution for the multinomial distribution with uninformative Dirichlet prior. Since the parameter identified set is a simple transformation of the cell probabilities (reduced form parameters), we form a credible set for the identified set by straightforward simulation from the Dirichlet posterior, following Kline and Tamer (2020). Given the large number of inequalities, when a simulated draw of probabilities generates an empty identified set, we conservatively include the parameter value(s) that minimize the criterion based on the worst violation of the inequalities. Kline and Tamer (2020) show the asymptotic connection to a frequentist confidence interpretation of the resulting intervals.

bootstrap produces a 95% credible interval of ($\hat{\underline{\kappa}} = \$32, \hat{\overline{\kappa}} = \$56$). Recall that the lower bound from the P&P estimates was \$20, so the parametric assumptions lead to a considerably tighter lower bound. The upper bound, on the other hand, is almost identical to the upper bound of \$57 we obtained from the P&P estimates that used the sample selected on the basis of exogenous price movements.

The lower bound is 30 to 40% of the estimates of κ_0 obtained from the comparison models in section 4.1 that do not allow for flexible fixed effects and the upper bound is 50 to 70% of those estimates. So the comparison models seem to overestimate the switching cost by a considerable amount. We turn next to an investigation of whether this difference is likely to influence the economic implications of the estimated models.

Counterfactual Comparisons. We now explore whether the difference between the κ_0 bounds obtained from the inequality estimator, and the κ_0 estimates obtained from the comparison models that allow for state dependence but not individual-by-product specific fixed effects, is likely to have economically important implications for a counterfactual of interest. BMC, the largest plan with over a third of the market in 2011 (see Table 2), increased its relative price dramatically in 2012 and then decreased it by an even greater amount in 2013. We consider predictions for what would have happened had they instead kept their price constant at the average of the 2012 and 2013 prices in those two years.

The calculation conditions on the 2011 choices of enrolled individuals. We then predict BMC's market share in 2012 twice; once using the actual and once the counterfactual prices. Finally, we use these predictions and the actual and counterfactual prices in 2013 to obtain the predicted shares from the counterfactual policy for the two year period from 2011-2013. The predictions for these sequences are done in pairs, one of which uses the (γ_0, δ_0) estimates from a comparison model in Table 4, the other uses the γ_0 estimate from the relevant comparison model but restricts δ to equal $\gamma\hat{\kappa}$ where $\hat{\kappa} = \$48.28$, as in Figure 5. The latter need not equal what our model would predict, as that would require either a model or bounds for the $\{\lambda_{i,d}\}_{i,d}$. Still the difference between the two predictions should provide an indication of whether the implications of a model that allowed for fixed effects are likely to be different than a model which does not.

Table 5 provides the results. The bottom row shows that the average BMC premiums, averaged over all incumbent enrollees who were not in the below-poverty group (and hence paid premiums), was \$58.4 per month in 2011. In 2012 that average increased to \$91.1, and in 2013 it fell to \$41.5; the changes that generated the sharp spike in the price plot in Figure 1. We consider counterfactual prices that equal the average of the prices in 2012 and 2013 in each income group, and then maintains that price in both years. That results in an average price of \$63 in 2012 and \$65 in 2013 (with the slight difference coming from changes in the relative size of different income groups in the two years).

The actual predictions differ somewhat between the pairs defined by the comparison models but their qualitative nature does not. The fall in price in 2012 from the \$91.1 to \$63 leads to a prediction of an 11% to 15% increase in share when we use the parameters estimated by the comparison models, but a prediction of a dramatic 53% to 67% increase in share when we constrain $\hat{\kappa} = 48.28$. In 2013 when the counterfactual average price was \$65 compared to the actual average price of \$41.5 , the estimates from the comparison models predict an 8 to 14% *higher* share from the *higher* counterfactual price. In contrast when we use $\hat{\kappa} = 48.28$ the higher counterfactual price in 2013 generates a two period prediction of a 17 to 20% *lower* share than the prediction from the status quo prices.

Recall that this is the prediction for 2013 which conditions on 2011 shares and the counterfactual

prices in both 2012 and 2013. The comparison models do predict the shares fall from 2012 to 2013 (by 1 to 2%). However because the comparison models' estimates of κ_0 are so high, this decrease is more than offset by the comparison model's increased share in 2012. That is, the impact of the higher κ_0 estimates in the comparison models' prediction in any one year spills over to the following years, making longer term predictions particularly problematic.

Table 5: Counterfactual Comparisons

Specification	2011 market shares	2012 status-quo	2012 counterfactual	2012 *% diff*	2013 status-quo	2013 counterfactual	2013 *% diff*
Market shares without imposing κ							
Plan FE	0.357	0.289	0.321	*11.0*	0.266	0.304	*14.2*
Plan × Region FE	0.357	0.289	0.320	*10.6*	0.266	0.298	*12.2*
Plan FE + RE	0.357	0.282	0.324	*15.1*	0.289	0.311	*7.6*
Market shares imposing κ							
Plan FE	0.357	0.186	0.306	*64.1*	0.399	0.326	*-18.3*
Plan × Region FE	0.357	0.205	0.313	*52.8*	0.381	0.318	*-16.6*
Plan FE + RE	0.357	0.183	0.305	*66.8*	0.410	0.329	*-19.7*
Premium	58.4	91.1	62.9		41.5	65.3	

Note: Table shows a counterfactual comparison of BMC market shares among current enrollees above 100% FPL. The top panel shows observed market shares in 2011, and then predicted market shares under status-quo premium and counterfactual premium, as well as their percentage difference in 2012 and 2013. We include results based on two FE specifications and one random coefficient specification. "Imposing κ" indicates whether we restrict the switching cost coefficient. The bottom panel shows the average BMC premium under status-quo and counterfactual in 2011-2013.

5 What Have We Learned?

We have provided both empirical results on switching costs in health insurance choices and methodological results on estimating models with individual by choice specific fixed effects and state dependence.

Our empirical results indicate that health insurance estimates of state dependence that do not allow for very flexible unobserved heterogeneity seem to seriously bias estimates of switching costs upwards; in our data by a factor of 50-130 percent. We found this regardless of whether the model without individual by product specific fixed effects allows for a rich set of plan interactions, random effects conditional on the initial choice, or random effects known prior to the initial choice. Rather, it appears important to allow for flexible individual-level preferences, likely because of the very heterogeneous way that similar consumers value plan provider networks (the key plan attribute in our context). For instance, people may care very strongly about whether *their* current doctor or hospital is covered in a given plan (Shepard, 2020; Tilipman, 2020), an individual-by-plan specific match factor that is not likely to be captured with coarse plan interactions. Our counterfactual, the reversal of what seems to be a failed pricing experiment by the largest insurer, illustrated that the difference in estimated switching costs matters. The comparison models' predicted a one year share change of 10-15% while when we use our estimate of κ we find a share change of 55-65%. Moreover the analogous predicted differences for the share change over the two years that includes

the insurer's policy reversal actually differ in sign; so longer-term predictions using the comparison models' κ estimates can be particularly problematic.

Our methodological results on estimators that allow for both state dependence and fixed effects depend on what the researcher is willing to assume on the distribution of the disturbances. If one does not want to assume a parametric distribution for ϵ and switching probabilities are less than a half, then finite positive bounds for κ_0, the ratio of price sensitivity to switching costs, are obtained employing the stationarity assumption in P&P. If switching probabilities are greater than a half, then revealed preference inequalities can be used to bound κ_0 either without, or if one is willing to make the stationarity assumption, in concert with, the P&P inequalities. If we are willing to make a *parametric* assumption on the ϵ distribution, then it is possible to get finite positive bounds without switching probabilities that are greater than a half from revealed preference. And, if the specified distribution is the logistic (or Gumbel) distribution then there will be positive finite upper and lower bounds. Moreover the estimators obtained from the logisitic distribution are exceptionally easy to compute.

References

Abbring, J. H., J. J. Heckman, P.-A. Chiappori, and J. Pinquet (2003): "Adverse Selection and Moral Hazard in Insurance: Can Dynamic Data Help to Distinguish?" *Journal of the European Economic Association*, 1, 512–521.

Aguirregabiria, V., J. Gu, and Y. Luo (2021): "Sufficient statistics for unobserved heterogeneity in structural dynamic logit models," *Journal of Econometrics*.

Armstrong, T. B. (2014): "Weighted KS statistics for Inference on Conditional Moment Inequalities," *Journal of Econometrics*, 181, 92–116.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *The Review of Economic Studies*, 47, 225–238.

Chamberlain, G. and G. W. Imbens (2003): "Nonparametric Applications of Bayesian Inference," *Journal of Business & Economic Statistics*, 21, 12–18.

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

Conley, T. G. and C. R. Udry (2010): "Learning About a New Technology: Pineapple in Ghana," *American Economic Review*, 100, 35–69.

Dafny, L., K. Ho, and M. Varela (2013): "Let Them Have Choice: Gains from Shifting Away from Employer-Sponsored Health Insurance and Toward an Individual Exchange," *American Economic Journal: Economic Policy*, 5, 32–58.

Ericson, K. M. M. (2014): "Consumer Inertia and Firm Pricing in the Medicare Part D Prescription Drug Insurance Exchange," *American Economic Journal: Economic Policy*, 6, 38–64.

Finkelstein, A., N. Hendren, and M. Shepard (2019): "Subsidizing Health Insurance for Low-income Adults: Evidence from Massachusetts," *American Economic Review*, 109, 1530–67.

Fox, J. (2007): "Semiparametric Estimation of Multinomial Discrete-Choice Models using a Subset of Choices," *The RAND Journal of Economics*, 38, 1002–1019.

Handel, B. R. (2013): "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, 103, 2643–82.

Heckman, J. (1978): "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence," *Annales de l'inséé*, Apr–Sep, 227–269.

Heckman, J. J. (1981): "Heterogeneity and State Dependence," in *Studies in Labor Markets*, University of Chicago Press, 91–140.

Ho, K., J. Hogan, and F. Scott Morton (2017): "The Impact of Consumer Inattention on Insurer Pricing in the Medicare Part D Program," *The RAND Journal of Economics*, 48, 877–905.

Honoré, B. E. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

Honoré, B. E. and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629.

Honoré, B. E. and M. Weidner (2020): "Moment Conditions for Dynamic Panel Logit Models with Fixed Effects," Tech. rep.

Keane, M. P. (1997): "Modeling Heterogeneity and State Dependence in Consumer Choice Behavior," *Journal of Business & Economic Statistics*, 15, 310–327.

Khan, S., F. Ouyang, and E. Tamer (forthcoming): "Inference on Semiparametric Multinomial Response Models," *Quantitative Economics*.

Khan, S., M. Ponomareva, and E. Tamer (2020): "Identification of Dynamic Binary Response Models," Tech. rep., Working Paper.

Kline, B. and E. Tamer (2020): "Bayesian Inference in a Class of Partially Identified Models," *Quantitative Economics*, 7, 329–366.

Kroft, K., F. Lange, and M. J. Notowidigdo (2013): "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment," *The Quarterly Journal of Economics*, 128, 1123–1167.

Manski, C. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

Manski, C. F. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 357–362.

McIntyre, A., M. Shepard, and M. Wagner (2021): "Can Automatic Retention Improve Health Insurance Market Outcomes?" Tech. rep., Harvard University Working Paper.

Pakes, A. and J. Porter (2016): "Moment Inequalities for Multinomial Choice with Fixed Effects," Tech. rep., National Bureau of Economic Research.

Polyakova, M. (2016): "Regulation of Insurance with Adverse Selection and Switching Costs: Evidence from Medicare Part D," *American Economic Journal: Applied Economics*, 8, 165–95.

Romano, J. P., A. M. Shaikh, and M. Wolf (2014): "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 82, 1979–2002.

Samuelson, P. A. (1938): "A Note on the Pure Theory of Consumer's Behaviour," *Economica*, 5, 61–71.

Shepard, M. (2020): "Hospital Network Competition and Adverse Selection: Evidence from the Massachusetts Health Insurance Exchange," Tech. rep., Harvard University Working Paper.

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

Shin, S., S. Misra, and D. Horsky (2012): "Disentangling Preferences and Learning in Brand Choice Models," *Marketing Science*, 31, 115–137.

Sorensen, A. T. (2006): "Social Learning and Health Plan Choice," *The Rand Journal of Economics*, 37, 929–945.

Tebaldi, P., A. Torgovitsky, and H. Yang (2019): "Nonparametric estimates of demand in the california health insurance exchange," Tech. rep., National Bureau of Economic Research.

Tilipman, N. (2020): "Employer Incentives and Distortions in Health Insurance Design: Implications for Welfare and Costs," Tech. rep., Working Paper.

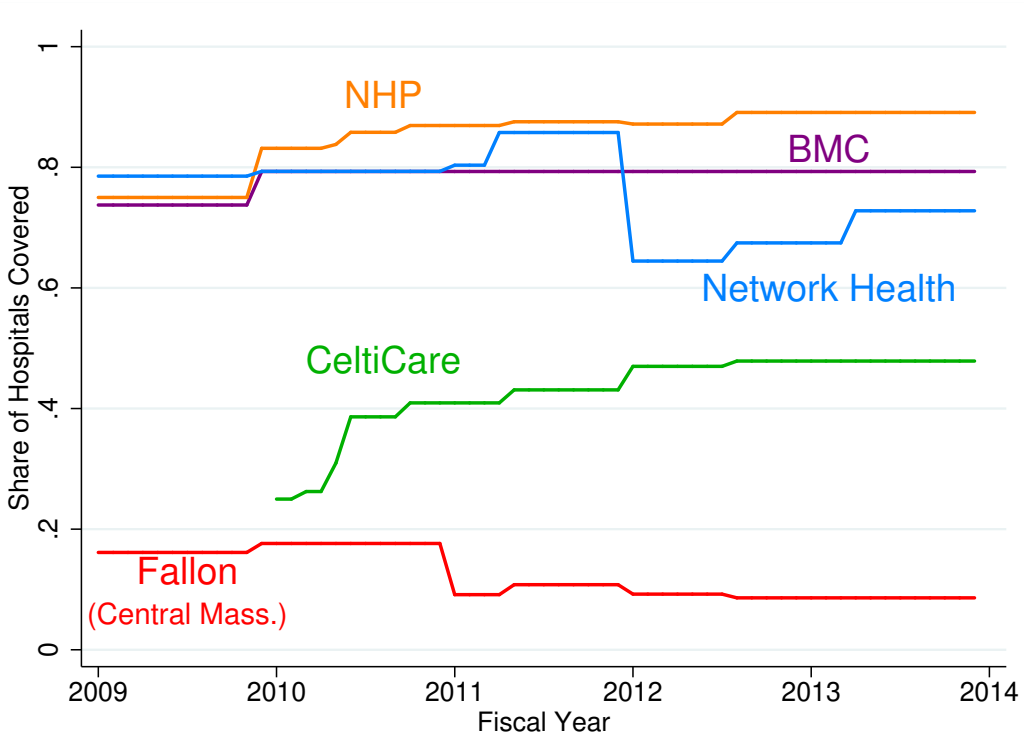
Torgovitsky, A. (2019): "Nonparametric inference on state dependence in unemployment," *Econometrica*, 87, 1475–1505.

Yan, J. (2013): "A Smoothed Maximum Score Estimator for Multinomial Discrete Choice Models," University of Wisconsin Working Paper.

Appendix

A Plan Hospital Networks

Figure 6: Hospital Coverage in Massachusetts Exchange Plans



NOTE: The graph shows the shares of Massachusetts hospitals covered by each CommCare plan, where shares are weighted by hospital bed size in 2011. Fallon's hospital coverage share is much lower than other plans largely because it mainly operates in central Massachusetts and therefore does not have a statewide network. The large decline in Network Health's network size in 2012 reflects its dropping of the Partners Healthcare System and several other providers from its network.

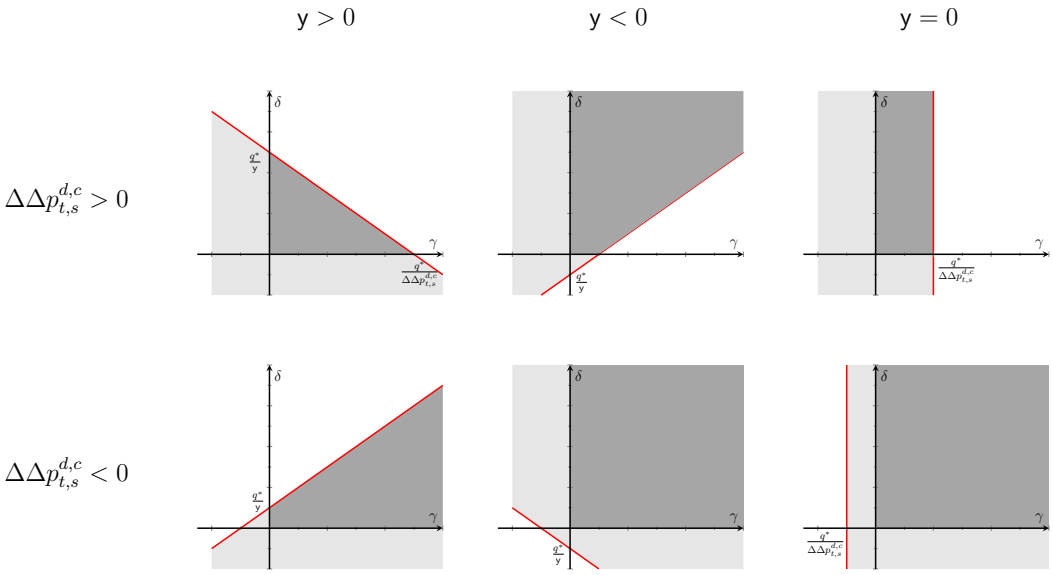
B Revealed Preferences Cases

Below we graphically display the information on the parameters (γ_0, δ_0) contained in Corollary 3.2. Given y, define

$$q^* = \frac{(F_{t,s}^{d,c})^{-1}\left(1 - \sum_{r:\Delta\Delta y_{t,s}^{d,c}(r,y_{i,s-1})\geq \mathsf{y}} \Pr(y_{i,t}=d, y_{i,t-1}=r, y_{i,s}=c|z, y_{i,s-1}); \sigma\right)}{\beta(x_i)}$$

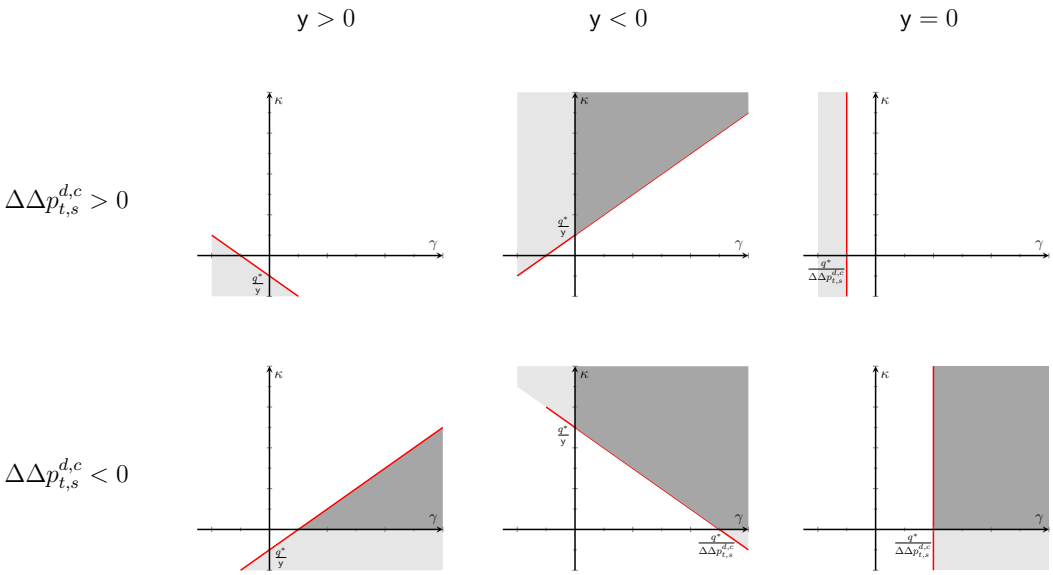
Figure 7 [8] considers the case where $q^* > 0$ [< 0]. If $F_{t,s}^{d,c}(0) = 0.5$, then the case $q^* > 0$ [< 0] corresponds to $\sum_{r:\Delta\Delta y_{t,s}^{d,c}(r,y_{i,s-1})\geq \mathsf{y}} \Pr(y_{i,t}=d, y_{i,t-1}=r, y_{i,s}=c|z, y_{i,s-1}) < 0.5$ [> 0.5].

Figure 7: Corollary 3.2 Inequalities for $q^* > 0$



Two cases in Figure 7 corresponding to $\Delta\Delta p^{d,c}_{t,s} < 0$ and $\mathsf{y} \leq 0$ are uninformative. For these cases, the whole first quadrant satisfies the inequality. In our empirical work, we use the remaining four cases to inform bound on κ_0. As noted previously, the case $q^* < 0$ does not occur in the empirical work, so we do not make use of the cases in Figure 8.

Figure 8: Corollary 3.2 Inequalities for $q^* < 0$



C Proofs

Proof of Lemma 2.2:

For all $c \notin D_0$, $d \in D_0$,

$$\begin{aligned}&(-p_{c,i,s} - \mathbf{1}\{y_{i,s-1} \neq c\}\kappa_0)\,\beta_i - (-p_{d,i,s} - \mathbf{1}\{y_{i,s-1} \neq d\}\kappa_0)\,\beta_i + (\lambda_{c,i} - \lambda_{d,i}) \\ &\quad \geq\ (-p_{c,i,t} - \mathbf{1}\{y_{i,t-1} \neq c\}\kappa_0)\,\beta_i - (-p_{d,i,t} - \mathbf{1}\{y_{i,t-1} \neq d\}\kappa_0)\,\beta_i + (\lambda_{c,i} - \lambda_{d,i})\end{aligned}$$

Hence,

$$\begin{aligned}&\left\{\varepsilon_{i,s} \,\middle|\, \varepsilon_{d,i,s} \geq \max_{c\notin D_0}\ [(-p_{c,i,s} - \mathbf{1}\{y_{i,s-1} \neq c\}\kappa_0)\,\beta_i - (-p_{d,i,s} - \mathbf{1}\{y_{i,s-1} \neq d\}\kappa_0)\,\beta_i + (\lambda_{c,i} - \lambda_{d,i}) + \varepsilon_{c,i,s}]\right\} \\ \subseteq\ &\left\{\varepsilon_{i,t} \,\middle|\, \varepsilon_{d,i,t}\ \geq\ \max_{c\notin D_0}\ [(-p_{c,i,t} - \mathbf{1}\{y_{i,t-1} \neq c\}\kappa_0)\,\beta_i - (-p_{d,i,t} - \mathbf{1}\{y_{i,t-1} \neq d\}\kappa_0)\,\beta_i + (\lambda_{c,i} - \lambda_{d,i}) + \varepsilon_{c,i,t}]\right\} \\ =\ &\left\{\varepsilon_{i,t} \,\middle|\, (-p_{d,i,t} - \mathbf{1}\{y_{i,t-1} \neq d\}\kappa_0)\,\beta_i + \lambda_{d,i} + \varepsilon_{d,i,t}\ \geq\ \max_{c\notin D_0}\ [(-p_{c,i,t} - \mathbf{1}\{y_{i,t-1} \neq c\}\kappa_0)\,\beta_i + \lambda_{c,i} + \varepsilon_{c,i,t}]\right\}\end{aligned}$$

So,

$$\begin{aligned}&\Pr(y_{i,t} \in D_0 \,|\, p_i, y_{i,t-1}, \beta_i, \lambda_i) \\ =\ &\Pr\left(\bigcup_{d\in D_0}\left\{\varepsilon_{i,t} \,\middle|\, (-p_{d,i,t} - \mathbf{1}\{y_{i,t-1} \neq d\}\kappa_0)\,\beta_i + \lambda_{d,i} + \varepsilon_{d,i,t}\right.\right. \\ &\qquad\qquad \left.\left. \geq\ \max_{c\notin D_0}\ [(-p_{c,i,t} - \mathbf{1}\{y_{i,t-1} \neq c\}\kappa_0)\,\beta_i + \lambda_{c,i} + \varepsilon_{c,i,t}]\right\} \,\middle|\, \beta_i, \lambda_i\right) \\ \geq\ &\Pr\left(\bigcup_{d\in D_0}\left\{\varepsilon_{i,s} \,\middle|\, (-p_{d,i,s} - \mathbf{1}\{y_{i,s-1} \neq d\}\kappa_0)\,\beta_i + \lambda_{d,i} + \varepsilon_{d,i,s}\right.\right. \\ &\qquad\qquad \left.\left. \geq\ \max_{c\notin D_0}\ [(-p_{c,i,s} - \mathbf{1}\{y_{i,s-1} \neq c\}\kappa_0)\,\beta_i + \lambda_{c,i} + \varepsilon_{c,i,s}]\right\} \,\middle|\, \beta_i, \lambda_i\right) \\ =\ &\Pr(y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i)\end{aligned}$$

In the second and third probabilities, the terms $p_{i,t}$, $p_{i,s}$, $y_{i,t-1}$, and $y_{i,s-1}$ denote the realized value of the price variable and lagged dependent variable from the conditioning statement.

⊠

Proof of Theorem 2.4:

(a) The supposition of Lemma 2.2 is satisfied for D_0 and $y_{i,t-1} = d'$ for any $d' \in D_0$. Hence,

$$\begin{aligned}&\Pr(y_{i,t} \in D_0, y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \\ &\quad = \sum_{d'\in D_0} \Pr(y_{i,t} \in D_0, y_{i,t-1} = d' \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \\ &\quad = \sum_{d'\in D_0} \Pr(y_{i,t} \in D_0 \,|\, p_i, y_{i,t-1} = d', \beta_i, \lambda_i) \cdot \Pr(y_{i,t-1} = d' \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \\ &\quad \geq \sum_{d'\in D_0} \Pr(y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \cdot \Pr(y_{i,t-1} = d' \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \\ &\quad = [\Pr(y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i)]^2\end{aligned}$$

Next, apply Jensen's Inequality to integrate out (β_i, λ_i).

$$\begin{aligned}
&\Pr(y_{i,t} \in D_0, y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2}) \\
&= E\left[\Pr(y_{i,t} \in D_0, y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \,|\, p_i, y_{i,t-2}\right] \\
&\geq E\left[\left[\Pr(y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i)\right]^2 \,|\, p_i, y_{i,t-2}\right] \\
&\geq \left[E\left[\Pr(y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \,|\, p_i, y_{i,t-2}\right]\right]^2 \\
&= \left[\Pr(y_{i,t-1} \in D_0 \,|\, p_i, y_{i,t-2})\right]^2
\end{aligned}$$

(b) $s < t-1$.
The supposition of Lemma 2.2 is satisfied for D_0 and $y_{i,t-1} = d'$ for any $d' \in D_1$. Hence,

$$\begin{aligned}
&\Pr(y_{i,t} \in D_0, y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \\
&= \sum_{d' \in D_1} \Pr(y_{i,t} \in D_0, y_{i,t-1} = d', y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \\
&= \sum_{d' \in D_1} \Pr(y_{i,t} \in D_0 \,|\, p_i, y_{i,t-1} = d', \beta_i, \lambda_i) \cdot \Pr(y_{i,t-1} = d', y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \\
&\geq \sum_{d' \in D_1} \Pr(y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \cdot \Pr(y_{i,t-1} = d', y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \\
&= \Pr(y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \cdot \Pr(y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \\
&\geq \left[\Pr(y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i)\right]^2
\end{aligned}$$

Next, apply Jensen's Inequality to integrate out (β_i, λ_i).

$$\begin{aligned}
&\Pr(y_{i,t} \in D_0, y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}) \\
&= E\left[\Pr(y_{i,t} \in D_0, y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \,|\, p_i, y_{i,s-1}\right] \\
&\geq E\left[\Pr(y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i)^2 \,|\, p_i, y_{i,s-1}\right] \\
&\geq \left(E\left[\Pr(y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \,|\, p_i, y_{i,s-1}\right]\right)^2 \\
&= \left[\Pr(y_{i,t-1} \in D_1, y_{i,s} \in D_0 \,|\, p_i, y_{i,s-1})\right]^2
\end{aligned}$$

⊠

Proof of Theorem 2.6:

(a)

$$\begin{aligned}
A \;\equiv\; &\Big\{(\varepsilon_{i,t}, \varepsilon_{i,t-1}) \,|\, (-p_{d,i,t} - \mathbf{1}\{c \neq d\}\kappa_0)\,\beta_i + \lambda_{d,i} + \varepsilon_{d,i,t} \geq \max_{d' \neq d} (-p_{d',i,t} - \mathbf{1}\{c \neq d'\}\kappa_0)\,\beta_i + \lambda_{d',i} + \varepsilon_{d',i,t}, \\
&\quad (-p_{c,i,t-1} - \mathbf{1}\{y_{i,t-2} \neq c\}\kappa_0)\,\beta_i + \lambda_{c,i} + \varepsilon_{c,i,t-1} \geq \max_{c' \neq c} (-p_{c',i,t-1} - \mathbf{1}\{y_{i,t-2} \neq c'\}\kappa_0)\,\beta_i + \lambda_{c',i} + \varepsilon_{c',i,t-1}\Big\} \\
\subset\; &\Big\{(\varepsilon_{i,t}, \varepsilon_{i,t-1}) \,|\, (-p_{d,i,t} - \kappa_0)\,\beta_i + \lambda_{d,i} + \varepsilon_{d,i,t} \geq (-p_{c,i,t})\,\beta_i + \lambda_{c,i} + \varepsilon_{c,i,t}, \\
&\quad (-p_{c,i,t-1} - \mathbf{1}\{y_{i,t-2} \neq c\}\kappa_0)\,\beta_i + \lambda_{c,i} + \varepsilon_{c,i,t-1} \geq (-p_{d,i,t-1} - \mathbf{1}\{y_{i,t-2} \neq d\}\kappa_0)\,\beta_i + \lambda_{d,i} + \varepsilon_{d,i,t-1}\Big\} \\
\subset\; &\Big\{(\varepsilon_{i,t}, \varepsilon_{i,t-1}) \,|\, \Delta\Delta\varepsilon^{d,c}_{i,t,t-1} \geq \Big(\Delta\Delta p^{d,c}_{i,t,t-1} + \kappa_0 \Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2}))\Big)\,\beta_i\Big\},
\end{aligned}$$

which implies

$$\begin{aligned}
&\Pr(y_{i,t}=d, y_{i,t-1}=c \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \le \Pr((\varepsilon_{i,t}, \varepsilon_{i,t-1}) \in A \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i) \\
&\quad \le \Pr(\Delta\Delta\varepsilon^{d,c}_{i,t,t-1} \ge \left(\Delta\Delta p^{d,c}_{i,t,t-1} + \kappa_0 \Delta\Delta \mathbf{1}^{d,c}(c, y_{i,t-2})\right) \beta_i \,|\, p_i, y_{i,t-2}, \beta_i, \lambda_i)
\end{aligned}$$

Integrate both sides with respect to the conditional distribution of λ_i, and the result follows.

(b) $s < t-1$.
Define

$$B^{d,c}_{t,s}(\mathsf{p}, \mathsf{y}, \beta_i) = \left\{ (\varepsilon_{i,t}, \varepsilon_{i,s}) \,|\, \Delta\Delta\varepsilon^{d,c}_{i,t,s} \ge (\mathsf{p} + \kappa_0 \mathsf{y})\beta_i \right\}$$

and note that since $\mathcal{F}^{d,c}_{t,s}$ denotes the c.d.f. of the conditional distribution of $\Delta\Delta\varepsilon^{d,c}_{i,t,s}$

$$\Pr((\varepsilon_{i,t}, \varepsilon_{i,s}) \in B^{d,c}_{t,s}(\mathsf{p}, \mathsf{y}) \,|\, p_i, y_{i,s-1}, \beta_i) = 1 - \mathcal{F}^{d,c}_{t,s}((\mathsf{p} + \kappa_0 \mathsf{y})\beta_i)$$

where p, y are constant values or functions of the conditioning set.

Using the same argument as in (a),

$$\begin{aligned}
A(r) &= \Big\{ (\varepsilon_{i,t}, \varepsilon_{i,s}) \,|\, (-p_{d,i,t} - \mathbf{1}\{r \ne d\}\kappa_0)\beta_i + \lambda_{d,i} + \varepsilon_{d,i,t} \ge \max_{d' \ne d} (-p_{d',i,t} - \mathbf{1}\{r \ne d'\}\kappa_0)\beta_i + \lambda_{d',i} + \varepsilon_{d',i,t}, \\
&\qquad (-p_{c,i,s} - \mathbf{1}\{y_{i,s-1} \ne c\}\kappa_0)\beta_i + \lambda_{c,i} + \varepsilon_{c,i,s} \ge \max_{c' \ne c} (-p_{c',i,s} - \mathbf{1}\{y_{i,s-1} \ne c'\}\kappa_0)\beta_i + \lambda_{c',i} + \varepsilon_{c',i,s} \Big\} \\
&\subset \left\{ (\varepsilon_{i,t}, \varepsilon_{i,s}) \,|\, \Delta\Delta\varepsilon^{d,c}_{i,t,s} \ge \left(\Delta\Delta p^{d,c}_{i,t,s} + \kappa_0 \Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1})\right) \beta_i \right\} \\
&= B^{d,c}_{t,s}(\Delta\Delta p^{d,c}_{i,t,s}, \Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}), \beta_i)
\end{aligned}$$

Note that if $\kappa_0 \ge 0$ and $\mathsf{y}' \ge \mathsf{y}$, then $B^{d,c}_{t,s}(\mathsf{p}, \mathsf{y}', \beta_i) \subset B^{d,c}_{t,s}(\mathsf{p}, \mathsf{y}, \beta_i)$. It follows that for any r such that $\Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \ge \mathsf{y}$, $A(r) \subset B^{d,c}_{t,s}(\Delta\Delta p^{d,c}_{i,t,s}, \mathsf{y}, \beta_i)$. And so,

$$\bigcup_{r:\Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \ge \mathsf{y}} A(r) \subset B^{d,c}_{t,s}(\Delta\Delta p^{d,c}_{i,t,s}, \mathsf{y}, \beta_i)$$

$$\begin{aligned}
&\sum_{r:\Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \ge \mathsf{y}} \Pr(y_t = d, y_{t-1} = r, y_s = c \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \\
&\quad \le \sum_{r:\Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \ge \mathsf{y}} \Pr(\{(\varepsilon_{i,t}, \varepsilon_{i,s}) \in A(r)\} \cap \{y_{i,t-1} = r\} \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \\
&\quad \le \Pr\left((\varepsilon_{i,t}, \varepsilon_{i,s}) \in \bigcup_{r:\Delta\Delta \mathbf{1}^{d,c}(r, y_{i,s-1}) \ge \mathsf{y}} A(r) \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i \right) \\
&\quad \le \Pr((\varepsilon_{i,t}, \varepsilon_{i,s}) \in B^{d,c}_{t,s}(\Delta\Delta p^{d,c}_{i,t,s}, \mathsf{y}, \beta_i) \,|\, p_i, y_{i,s-1}, \beta_i, \lambda_i) \qquad \text{(C.1)}
\end{aligned}$$

The conclusion of part (b) follows by integrating out both sides of the inequality with respect to the conditional distribution of λ_i. ⊠

Proof of Corollary 2.7:
The supposition in part (a) implies that $\Pr(y_{i,t} = d, y_{i,t-1} = c \,|\, p_i, y_{i,t-2}, \beta_i) \geq 0.5$ for some β_i. The assumptions of the corollary then lead immediately to the conclusion by Theorem 2.6. Part (b) follows similarly. ⊠

Lemma C.1. *(a) Let* $\mathcal{M}_t(c, \lambda_i) = \sum_{r \in \mathcal{D}_t} \exp[(-\gamma_0 p_{r,t} - \delta_0 \mathbf{1}\{c \neq r\})\beta(x) + \lambda_{r,i}]$. *For any* c, d,

$$e^{-\delta_0 \beta(x)} \;\leq\; \frac{\mathcal{M}_t(d, \lambda_i)}{\mathcal{M}_t(c, \lambda_i)} \;\leq\; e^{\delta_0 \beta(x)}$$

(b) Let $\mathcal{S}_{t,s}(d, c, \lambda_i) = \sum_{r \in \mathcal{D}_{t-1}, r' \in \mathcal{D}_{s+1}} \frac{\exp[-\delta_0(\mathbf{1}\{r \neq d\} + \mathbf{1}\{c \neq r'\})\beta(x) - \gamma_0 p_{r',i,s+1}\beta(x) + \lambda_{r',i})}{\mathcal{M}_t(r, \lambda_i)} \Pr(y_{i,t-1} = r \,|\, p_i, y_{i,s+1} = r', x_i = x, \lambda_i)$. *For any* c, d,

$$e^{-2\delta_0 \beta(x)} \;\leq\; \frac{\mathcal{S}_{t,s}(d, c, \lambda_i)}{\mathcal{S}_{t,s}(c, d, \lambda_i)} \;\leq\; e^{2\delta_0 \beta(x)}$$

Proof of Lemma C.1:
(a) For any c, d, r, $-1 - \mathbf{1}\{c \neq r\} \;\leq\; -\mathbf{1}\{d \neq r\} \;\leq\; 1 - \mathbf{1}\{c \neq r\}$, and so

$$\begin{aligned}
e^{-\delta_0 \beta(x)} \mathcal{M}_t(c, \lambda_i) &= \sum_{r \in \mathcal{D}_t} e^{\delta_0 \beta(x)(-1 - \mathbf{1}\{c \neq r\})} \exp[-\gamma_0 p_{r,t} \beta(x) + \lambda_{r,i}] \\
&\leq \sum_{r \in \mathcal{D}_t} e^{\delta_0 \beta(x)(-\mathbf{1}\{d \neq r\})} \exp[-\gamma_0 p_{r,t} \beta(x) + \lambda_{r,i}] \\
&= \mathcal{M}_t(d, \lambda_i) \;\leq\; \sum_{r \in \mathcal{D}_t} e^{\delta_0 \beta(x)(1 - \mathbf{1}\{c \neq r\})} \exp[-\gamma_0 p_{r,t} \beta(x) + \lambda_{r,i}] \;=\; e^{\delta_0 \beta(x)} \mathcal{M}_t(c, \lambda_i)
\end{aligned}$$

Result (a) follows.

(b) For any c, d, r, r',
$-2 - \mathbf{1}\{r \neq c\} - \mathbf{1}\{d \neq r'\} \;\leq\; -\mathbf{1}\{r \neq d\} - \mathbf{1}\{c \neq r'\} \;\leq\; 2 - \mathbf{1}\{r \neq c\} - \mathbf{1}\{d \neq r'\}$.
Hence,

$$\begin{aligned}
&e^{-2\delta_0 \beta(x)} \mathcal{S}_{t,s}(c, d, \lambda_i) \\
&= \sum_{r \in \mathcal{D}_{t-1}, r' \in \mathcal{D}_{s+1}} \frac{e^{\delta_0(-2 - \mathbf{1}\{r \neq c\} - \mathbf{1}\{d \neq r'\})\beta(x)} e^{-\gamma_0 p_{r',i,s+1}\beta(x) + \lambda_{r',i}}}{\mathcal{M}_t(r, \lambda_i)} \Pr(y_{i,t-1} = r \,|\, p_i, y_{i,s+1} = r', x_i = x, \lambda_i) \\
&\leq \sum_{r \in \mathcal{D}_{t-1}, r' \in \mathcal{D}_{s+1}} \frac{e^{\delta_0(-\mathbf{1}\{r \neq d\} - \mathbf{1}\{c \neq r'\})\beta(x)} e^{-\gamma_0 p_{r',i,s+1}\beta(x) + \lambda_{r',i}}}{\mathcal{M}_t(r, \lambda_i)} \Pr(y_{i,t-1} = r \,|\, p_i, y_{i,s+1} = r', x_i = x, \lambda_i) \\
&= \mathcal{S}_{t,s}(d, c, \lambda_i) \\
&\leq \sum_{r \in \mathcal{D}_{t-1}, r' \in \mathcal{D}_{s+1}} \frac{e^{\delta_0(2 - \mathbf{1}\{r \neq c\} - \mathbf{1}\{d \neq r'\})\beta(x)} e^{-\gamma_0 p_{r',i,s+1}\beta(x) + \lambda_{r',i}}}{\mathcal{M}_t(r, \lambda_i)} \Pr(y_{i,t-1} = r \,|\, p_i, y_{i,s+1} = r', x_i = x, \lambda_i) \\
&\leq e^{2\delta_0 \beta(x)} \mathcal{S}_{t,s}(c, d, \lambda_i)
\end{aligned}$$

and (b) follows. ⊠

The following theorem extends the results in Theorem 3.4.

Theorem C.2. *Suppose Assumption 3.3 holds, and* $(d,c) \in \mathcal{D}_t \cap \mathcal{D}_s$. *Let* $\tau = \mathbf{1}\{r = c\}$, *and* $\Lambda = \begin{cases} 1 & \text{if } s = t-1 \text{ or } t-2 \\ 3 & \text{if } s < t-2 \end{cases}$
Then, for $s \leq t-1$ *and* $r \neq d$,

$$\begin{aligned}
&\Pr(y_{i,t} = c, y_{i,s} = d \,|\, p_i, y_{i,s-1} = r, x_i = x)\ e^{(\tau-\Lambda)\delta_0\beta(x)} e^{\gamma_0\left(\Delta\Delta p^{c,d}_{t,s}\right)\beta(x)} \\
&\qquad \leq \Pr(y_{i,t} = d, y_{i,s} = c \,|\, p_i, y_{i,s-1} = r, x_i = x) \\
&\qquad \leq \Pr(y_{i,t} = c, y_{i,s} = d \,|\, p_i, y_{i,s-1} = r, x_i = x)\ e^{(\tau+\Lambda)\delta_0\beta(x)} e^{\gamma_0\left(\Delta\Delta p^{c,d}_{t,s}\right)\beta(x)}
\end{aligned}$$

Remark C.3. The case $r = d$ is implied by the case $r = c$.

Proof of Theorem C.2:
(i) $s = t-1$.

$$\begin{aligned}
&\Pr(y_{i,t} = d, y_{i,t-1} = c \,|\, p_i, y_{i,t-2}, x_i = x, \lambda_i) \\
&= \Pr(y_{i,t} = d \,|\, p_i, y_{i,t-1} = c, x_i = x, \lambda_i) \Pr(y_{i,t-1} = c \,|\, p_i, y_{i,t-2}, x_i = x, \lambda_i) \\
&= \frac{e^{(-\gamma_0 p_{d,i,t} - \delta_0 \mathbf{1}\{c \neq d\})\beta(x) + \lambda_{d,i}}}{\mathcal{M}_t(c, \lambda_i)} \frac{e^{(-\gamma_0 p_{c,i,t-1} - \delta_0 \mathbf{1}\{y_{i,t-2} \neq c\})\beta(x) + \lambda_{c,i}}}{\mathcal{M}_{t-1}(y_{i,t-2}, \lambda_i)} \\
&= \frac{e^{-\gamma_0 p_{d,i,t}\beta(x)} e^{(-\gamma_0 p_{c,i,t-1} - \delta_0 \mathbf{1}\{y_{i,t-2} \neq c\})\beta(x)} e^{-\delta_0 \mathbf{1}\{c \neq d\}\beta(x)} e^{\lambda_{d,i} + \lambda_{c,i}}}{\mathcal{M}_t(c, \lambda_i)\mathcal{M}_{t-1}(y_{i,t-2}, \lambda_i)}
\end{aligned}$$

Similarly, for $\Pr(y_{i,t} = c, y_{i,t-1} = d \,|\, p_i, y_{i,t-2}, x_i = x, \lambda_i)$.
So,

$$\begin{aligned}
\frac{\Pr(y_{i,t} = d, y_{i,t-1} = c \,|\, p_i, y_{i,t-2}, x_i = x, \lambda_i)}{\Pr(y_{i,t} = c, y_{i,t-1} = d \,|\, p_i, y_{i,t-2}, x_i = x, \lambda_i)} &= \frac{e^{-\gamma_0 p_{d,i,t}\beta(x)} e^{(-\gamma_0 p_{c,i,t-1} - \delta_0 \mathbf{1}\{y_{i,t-2} \neq c\})\beta(x)} \mathcal{M}_t(d, \lambda_i)}{e^{-\gamma_0 p_{c,i,t}\beta(x)} e^{(-\gamma_0 p_{d,i,t-1} - \delta_0 \mathbf{1}\{y_{i,t-2} \neq d\})\beta(x)} \mathcal{M}_t(c, \lambda_i)} \\
&= \exp\left[-\gamma_0 \left(\Delta\Delta p^{d,c}_{i,t,t-1}\right)\beta(x)\right] \exp\left[-\delta_0 \left(\mathbf{1}\{y_{i,t-2} \neq c\} - \mathbf{1}\{y_{i,t-2} \neq d\}\right)\beta(x)\right] \frac{\mathcal{M}_t(d, \lambda_i)}{\mathcal{M}_t(c, \lambda_i)}
\end{aligned}$$

By Lemma C.1,

$$\begin{aligned}
&\Pr(y_{i,t} = c, y_{i,t-1} = d \,|\, p_i, y_{i,t-2}, x_i = x, \lambda_i)\ e^{-\delta_0\beta(x)} e^{-\gamma_0\left(\Delta\Delta p^{d,c}_{i,t,t-1}\right)\beta(x)} e^{-\delta_0(\mathbf{1}\{y_{i,t-2} \neq c\} - \mathbf{1}\{y_{i,t-2} \neq d\})\beta(x)} \\
&\leq \Pr(y_{i,t} = d, y_{i,t-1} = c \,|\, p_i, y_{i,t-2}, x_i = x, \lambda_i) \\
&\leq \Pr(y_{i,t} = c, y_{i,t-1} = d \,|\, p_i, y_{i,t-2}, x_i = x, \lambda_i)\ e^{\delta_0\beta(x)} e^{-\gamma_0\left(\Delta\Delta p^{d,c}_{i,t,t-1}\right)\beta(x)} e^{-\delta_0(\mathbf{1}\{y_{i,t-2} \neq c\} - \mathbf{1}\{y_{i,t-2} \neq d\})\beta(x)}
\end{aligned}$$

The result for the case $s = t-1$ follows by integrating out λ_i.

(ii) $s = t-2$.

$$\begin{aligned}
&\Pr(y_{i,t} = d, y_{i,t-2} = c \,|\, p_i, y_{i,t-3}, x_i = x, \lambda_i) \\
&\quad = \sum_{r \in \mathcal{D}_{t-1}} \Pr(y_{i,t} = d \,|\, p_i, y_{i,t-1} = r, x_i = x, \lambda_i) \Pr(y_{i,t-1} = r \,|\, p_i, y_{i,t-2} = c, x_i = x, \lambda_i) \\
&\qquad \cdot \Pr(y_{i,t-2} = c \,|\, p_i, y_{i,t-3}, x_i = x, \lambda_i) \\
&\quad = \left[\sum_{r \in \mathcal{D}_{t-1}} \frac{e^{-\delta_0(\mathbf{1}\{r \neq d\} + \mathbf{1}\{c \neq r\})\beta(x)} e^{-\gamma_0 p_{r,i,t-1}\beta(x) + \lambda_{r,i}}}{\mathcal{M}_t(r, \lambda_i)} \right] \frac{e^{-\gamma_0 p_{d,i,t}\beta(x)} e^{-\gamma_0 p_{c,i,t-2}\beta(x) - \delta_0 \mathbf{1}\{y_{i,t-3} \neq c\}\beta(x)} e^{\lambda_{d,i} + \lambda_{c,i}}}{\mathcal{M}_{t-1}(c, \lambda_i) \mathcal{M}_{t-2}(y_{i,t-3}, \lambda_i)}
\end{aligned}$$

Similarly, for $\Pr(y_{i,t} = c, y_{i,t-2} = d \,|\, p_i, y_{i,t-3}, x_i = x, \lambda_i)$.

Hence,

$$\begin{aligned}
\frac{\Pr(y_{i,t} = d, y_{i,t-2} = c \,|\, p_i, y_{i,t-3}, x_i = x, \lambda_i)}{\Pr(y_{i,t} = c, y_{i,t-2} = d \,|\, p_i, y_{i,t-3}, x_i = x, \lambda_i)} &= \frac{e^{-\gamma_0(p_{d,i,t} + p_{c,i,t-2})\beta(x)} e^{-\delta_0 \mathbf{1}\{c \neq y_{i,t-3}\}\beta(x)} \mathcal{M}_{t-1}(d, \lambda_i)}{e^{-\gamma_0(p_{c,i,t} + p_{d,i,t-2})\beta(x)} e^{-\delta_0 \mathbf{1}\{d \neq y_{i,t-3}\}\beta(x)} \mathcal{M}_{t-1}(c, \lambda_i)} \\
&= \exp\left[-\gamma_0 \left(\Delta\Delta p^{d,c}_{i,t,t-2} \right) \beta(x) \right] \exp\left[-\delta_0 \left(\mathbf{1}\{y_{i,t-3} \neq c\} - \mathbf{1}\{y_{i,t-3} \neq d\} \right) \beta(x) \right] \frac{\mathcal{M}_{t-1}(d, \lambda_i)}{\mathcal{M}_{t-1}(c, \lambda_i)}
\end{aligned}$$

As in the $s = t-1$ case, the result for $s = t-2$ now follows by application of Lemma C.1(a) and integrating out λ_i.

(iii) $s < t-2$.

$$\begin{aligned}
&\Pr(y_{i,t} = d, y_{i,s} = c \,|\, p_i, y_{i,s-1}, x_i = x, \lambda_i) \\
&\quad = \sum_{r \in \mathcal{D}_{t-1}, r' \in \mathcal{D}_{s+1}} [\Pr(y_{i,t} = d \,|\, p_i, y_{i,t-1} = r, x_i = x, \lambda_i) \Pr(y_{i,t-1} = r \,|\, p_i, y_{i,s+1} = r', x_i = x, \lambda_i) \\
&\qquad \cdot \Pr(y_{i,s+1} = r' \,|\, p_i, y_{i,s} = c, x_i = x, \lambda_i) \Pr(y_{i,s} = c \,|\, p_i, y_{i,s-1}, x_i = x, \lambda_i)] \\
&\quad = \left[\sum_{r \in \mathcal{D}_{t-1}, r' \in \mathcal{D}_{s+1}} \frac{e^{-\delta_0(\mathbf{1}\{r \neq d\} + \mathbf{1}\{c \neq r'\})\beta(x)} e^{-\gamma_0 p_{r',i,s+1}\beta(x) + \lambda_{r',i}}}{\mathcal{M}_t(r, \lambda_i)} \Pr(y_{i,t-1} = r \,|\, p_i, y_{i,s+1} = r', x_i = x, \lambda_i) \right] \\
&\qquad \cdot \frac{e^{-\gamma_0 p_{d,i,t}\beta(x)} e^{(-\gamma_0 p_{c,i,s} - \delta_0 \mathbf{1}\{y_{i,s-1} \neq c\})\beta(x)} e^{\lambda_{d,i} + \lambda_{c,i}}}{\mathcal{M}_{s+1}(c, \lambda_i) \mathcal{M}_s(y_{i,s-1}, \lambda_i)}
\end{aligned}$$

Similarly, for $\Pr(y_{i,t} = c, y_{i,s} = d \,|\, p_i, y_{i,s-1}, x_i = x, \lambda_i)$.

Using the notation from Lemma C.1(b),

$$\begin{aligned}
\frac{\Pr(y_{i,t} = d, y_{i,s} = c \,|\, p_i, y_{i,s-1}, x_i = x, \lambda_i)}{\Pr(y_{i,t} = c, y_{i,s} = d \,|\, p_i, y_{i,s-1}, x_i = x, \lambda_i)} &= \frac{e^{-\gamma_0(p_{d,i,t} + p_{c,i,s})\beta(x)} e^{-\kappa_0 \mathbf{1}\{c \neq y_{i,s-1}\}} \mathcal{M}_{s+1}(d, \lambda_i) \, \mathcal{S}_{t,s}(d, c, \lambda_i)}{e^{-\gamma_0(p_{c,i,t} + p_{d,i,s})\beta(x)} e^{-\kappa_0 \mathbf{1}\{d \neq y_{i,s-1}\}} \mathcal{M}_{s+1}(c, \lambda_i) \, \mathcal{S}_{t,s}(c, d, \lambda_i)} \\
&= \exp\left[-\gamma_0 \left(\Delta\Delta p^{d,c}_{t,s} \right) \beta(x) \right] \exp\left[-\delta_0 \left(\mathbf{1}\{y_{i,s-1} \neq c\} - \mathbf{1}\{y_{i,s-1} \neq d\} \right) \beta(x) \right] \frac{\mathcal{M}_{s+1}(d, \lambda_i) \, \mathcal{S}_{t,s}(d, c, \lambda_i)}{\mathcal{M}_{s+1}(c, \lambda_i) \, \mathcal{S}_{t,s}(c, d, \lambda_i)}
\end{aligned}$$

Now apply both parts of Lemma C.1 and integrate out λ_i. The result for $s < t-2$ follows. ⊠