Discrete Choice Models with Consideration Sets: Identification from Asymmetric Cross-Derivatives*

Jason Abaluck and Abi Adams
*

October 21, 2016

Abstract

The applied literature on consideration sets relies for identification on either auxiliary data on what options were considered or on instruments assumed to impact consideration probabilities or utility but not both. We show that a broad class of consideration set models - including all we are aware of in the applied literature - are identified without these assumptions from asymmetries in the cross-derivatives of choice probabilities with respect to product characteristics. Our identification proof constructively recovers how consideration probabilities vary with observables from these asymmetries. We show that the model recovers the underlying parameters in simulations where traditional models are badly misspecified; in an application to hotel choice data, the model determines that the randomly assigned ordering of hotels impacts attention but not utility.

1 Introduction

Discrete choice models typically assume that consumers are aware of all available options (hereafter, a "full information" model). The literature on "consideration sets" relaxes this assumption and allows for the possibility that consumers consider only a subset of possible options when making a choice. Empirical models in this literature have thus far relied on strong identifying assumptions or additional data to separate the impact of observables on utility and consideration set probabilities. For example, Conlon and Mortimer (2013) assume that consideration sets are known in some periods, Honka (2014) and Honka, Hortaçsu, and Vitorino (2015) use auxiliary information detailing which options consumers are aware of, and Goeree (2008) and Gaynor, Propper, and Seiler (2016) assume that some observables impact either attention or utility but not both.

We show that a broad class of discrete choice models with consideration sets - including all we are aware of estimated in the literature to date - are identified without such assumptions from asymmetries in the responsiveness of choice probabilities to characteristics of rival goods. In addition to (constructively) proving identification, we show in several applications that these models imply very different price elasticities and normative conclusions from the full information models they nest. These results suggest that such consideration set models could be used in a wide variety of settings where full-information models are currently estimated.

*Thanks to Joe Altonji, Steve Berry, Judy Chevalier, Jonathan Gruber, Phil Haile, Costas Meghir, Fiona Scott Morton, Joe Shapiro, K. Sudhir and participants in the Yale labor economics and econometrics workshops for helpful discussions. Thanks to Raluca Ursu for help in replicating her Expedia analysis.

The intuition for identification is as follows: conditional on being attentive to any two goods,[1] discrete choice models typically assume that the choice of good j responds to changes in the characteristics of good k and vice-versa in a way that respects the symmetry of the matrix of cross-derivatives (the discrete choice analogue of Slutsky symmetry). Consider a world with just 2 goods, no outside option and no income effects. Conventional discrete choice models assume that you should be equally responsive if the price of good A changes by \$100 or the price of good B changes in the opposite direction by \$100. If you respond more when the price of good A changes than when the price of good B changes in the opposite direction, this suggests that you were not fully attentive to both goods.[2] Asymmetries in the matrix of cross-derivatives are the J-dimensional generalization of this example. Our identification proof shows how the impact of product characteristics on consideration set probabilities can be constructively recovered from these asymmetries. As long as the data is rich enough to consistently estimate both own and cross-price elasticities, the impact of characteristics on utility can be separately identified from the impact of characteristics on consideration sets. The relationship between inattention and cross-derivative symmetry was (as far as we know) first observed by Gabaix (2011) in the context of his sparsity-based model of attention - in our model, attention is at the product-level rather than the characteristic level and we link the asymmetry of the cross-derivatives to commonly used econometric methods.

Consideration set models have numerous advantages over full information models for a wide variety of questions. Most fundamentally from a positive modeling perspective, full-information random coefficients models misstate how price elasticities vary across goods. Price elasticities will tend to be much larger for goods to which consumers are likely to pay attention and smaller for goods to which consumers are likely to be inattentive.[3] Goeree (2008) shows that this misspecification has important consequences for pricing - in the PC industry, she estimates mark-ups of 19% in the consideration set model compared to mark-ups of 5% in the traditional full-information model. Consideration set models have other positive advantages as well. Consideration set models also permit consideration of counterfactuals where the degree of attentiveness varies.

Normatively, full-information models are seriously deficient relative to consideration set models. The "attentive" utility parameters typically differ substantially from those in the traditional model both in terms of magnitude and relative value (the latter biases any willingness to pay estimates). The choices consumers would make were they fully aware of the alternatives available to them provide a natural normative benchmark which can be used to evaluate the scope for potential benefits from advertising or policies that would make consumers more aware of the options available to them. In an inertial model, assuming that all inertia represents adjustment costs or preferences can vastly overstate the value of making the same choice. Alternatively, assuming that all inertia represents inattention can dramatically understate this value. Both possibilities must be allowed for to permit a proper revealed preference analysis of choices.

[1] In this paper, we use "attentive" as synonymous with "a good is in the consumer's consideration set."

[2] As we make clear in Section 3.1.1, whether this implies inattentiveness to A or B depends on whether one thinks that attentiveness to good B depends on characteristics of good B, good A or both.

[3] One could in principle attempt to remedy this problem by interacting every included observable with product fixed effects or by interacting all observables with each other. But such models would require either $J \times k$ or k^2 additional variables with J alternatives and k observables. Consideration set models allow for this behavior using only $2k$ variables, putting each observable characteristic in both the attentive and utility equation.

Several papers theoretically investigate the revealed preference identification of consideration sets (Manzini and Mariotti (2014), Masatlioglu, Nakajima, and Ozbay (2016)). However, the identification results in these papers require that each individual is observed choosing from a large number of different choice sets, something which is very rarely observed in the field. Perhaps the closest paper to ours is Crawford, Griffith, and Iaria (2016), which presents identification results in discrete choice models with consideration sets in panel data. Their main result is that, given logit errors, preferences can be recovered provided choice sets do not change over time; their model of consideration sets is nonparametric, but they require panel data, the assumption that consideration sets do not change is strong, and their model does not recover how consideration set probabilities vary as a function of observable characteristics.

We prove our result in a general parametric model which nests the models that have been used in the applied literature. The symmetry of the cross-derivatives in a fully attentive discrete choice model follows whenever utility can be written as $u_{ij} = v_{ij} + \epsilon_{ij}$, where $\epsilon \perp x$, $\frac{\partial v_{ij}}{\partial x_{ij}}|_{x_{ij}=\bar{x}} = \frac{\partial v_{ij'}}{\partial x_{ij'}}|_{x_{ij'}=\bar{x}}$ and $\frac{\partial v_{ij}}{\partial x_{ij'}} = 0$. This includes the commonly used class of models where $u_{ij} = x_{ij}\beta_i + \xi_j + \epsilon_{ij}$, but it would exclude models with alternative-specific coefficients on the observables. Likewise, our consideration set model nests the cases considered so far in the applied literature, but it is not without loss of generality - we assume, for example, that unobservable determinants of consideration sets are uncorrelated with the unobservable components of utility. An important question beyond the scope of our work here is to characterize the testable restrictions imposed by the most general possible consideration set model in a fully non-parametric random utility framework - this would generalize the work of Falmagne (1978), who addresses the testable restrictions imposed by optimization in random utility models without consideration sets. Such a result would clarify the degree to which identification in our model comes from underlying features of the data vs. (unavoidable in practice) functional form assumptions. Our result shows that the commonly used set of functional form and distributional assumptions imposes a strong restriction on substitution patterns which rules out patterns typically observed in the data.

We consider two special cases of our model. The first case is the model developed in Goeree (2008) where each good has a probability of being considered which depends on characteristics of that good. We apply this model to online choice of hotels using data from Expedia and show that the model implies that the order of hotels impacts attention and not utility and that price elasticties are substantially higher for higher ranked items. The second case is the model developed in Ho, Hogan, and Scott-Morton (2015) in which there is a clear default good and the probability of considering all other options varies as a function of the characteristics of that default good (the opposite of the assumption in Goeree (2008)). In this model, we apply our results to Medicare Part D and show that the model can be used to disentangle the degree to which observed inertia in health plan choices is due to adjustment costs or inattention.

2 Model and Identification Proof

2.1 Full-information Model

Consider first a full-information model. Individual i makes a discrete choice among $J+1$ goods, $\{0, 1, ..., J\}$. Products are characterised as a bundle of $2 < K < \infty$ characteristics, $\mathbf{x}$, with support $\chi \subseteq \mathbb{R}^K$. Let Y_{ij} denote an indicator for whether individual i chooses option j. We assume that $P(Y_{ij} = 1) = P(u_{ij} = \max'_j u_{ij'})$ where utility can be written as:

$$u_{ij} = v_{ij}(x_{ij}) + \epsilon_{ij} \tag{1}$$

where we make the following assumptions:

Assumption 1. *Exogenous Characteristics:* $\epsilon_{ij} \perp\!\!\!\perp \mathbf{x}_{ij'}$ for $\forall i$ and $\forall j, j'$.

Assumption 2. All elements of $\mathbf{x}_{ij}$ for $j = 0, ..., J$ exhibit continuous variation and v_{ij} is differentiable with respect to x_{ij}.

This is required to permit identification of own- and cross-partials of choice probabilities with respect to characteristics. This assumption can be relaxed to require continuous variation only in some neighbourhood if χ if we impose $u_{ij} = x_{ij}\beta + \epsilon_{ij}$. Loosely speaking, we require that we observe sufficient variation in characteristics across either a sufficiently large number of markets (or consumers if characteristics vary at the individual level) in some neighbourhood of the support of independent variables to permit identification of partial derivatives of choice probabilities with respect to characteristics.

Assumption 3. $\frac{\partial v_{ij}}{\partial x_{ij'}} = 0$ for $j \neq j'$. This assumption rules out the case where the utility of product j depends directly on characteristics of product j'.

Assumption 4. $\frac{\partial v_{ij}}{\partial x_{ij}}|_{x_{ij}=\bar{x}} = \frac{\partial v_{ij'}}{\partial x_{ij'}}|_{x_{ij'}=\bar{x}}$.

Evaluated at the same vector of observables $x_{ij} = x_{ij'} = \bar{x}$, the partial derivatives of the deterministic component of utility for good j are symmetric across goods. This rules out - for example - alternative-specific coefficients on observables.

Assumption 5. $F(\epsilon_{i0}, ..., \epsilon_{iJ})$ is absolutely continuous with respect to the Lebesgue measure and gives rise to a density function that is everywhere positive on $\mathbb{R}$.

These assumptions are not without loss of generality, but they nest numerous cases of interest. For example, these assumptions are satisfied by any linear random coefficients model where $u_{ij} = x_{ij}\beta_i + \epsilon_{ij}$ as well as panel data models which include inertial effects and arbitrary correlations of the error term across products and time.

With [] denoting exclusion, the probability that individual i chooses option j having considered all options 0,..., J is given by:

$$s^{\star}_{ij}(\mathbf{x}_{i0},...,\mathbf{x}_{iJ}) = Pr\left(v_{ij}+\epsilon_{ij} = \max_{k=0,...,J} v_{ik}+\epsilon_{ik}\right)$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{v_{ij}+e-v_{i0}} ... \left[\int_{-\infty}^{v_{ij}+e-v_{ij}}\right] ... \int_{-\infty}^{v_{ij}+e-v_{iJ}} f(z_0,...,e,...,z_J)\mathrm{d}z_J...[\mathrm{d}z_j]...\mathrm{d}z_0\mathrm{d}e \quad (2)$$

Maximisation given the structure that we have placed on the utility function implies some specific restrictions on choice probabilities, in addition to those imposed by probability theory. Specifically, these assumptions suffice to demonstrate the discrete choice analogue to Slutsky Symmetry:

Corollary 1. *Symmetry of Cross Derivatives*:

$$\frac{\partial s^{\star}_{ij}}{\partial x^{a}_{ij'}}|_{x_{ij}=x_{ij'}=\bar{x}} = \frac{\partial s^{\star}_{ij'}}{\partial x^{a}_{ij}}|_{x_{ij}=x_{ij'}=\bar{x}} \quad (3)$$

Proof: Without loss of generality, assume $j < j'$:

$$\frac{\partial s^{\star}_{ij}(x_{i0},...,x_{iJ})}{\partial x_{ij'}} = \frac{\partial v_{ij'}}{\partial x_{ij'}}\int_{-\infty}^{\infty}\int_{-\infty}^{v_{ij}+e-v_{i0}} ... \left[\int_{-\infty}^{v_{ij}+e-v_{ij}}\right] ... \left[\int_{-\infty}^{v_{ij}+e-v_{ij'}}\right] ... \int_{-\infty}^{v_{ij}+e-v_{iJ}}$$
$$f(z_0,...,e,...,v_{ij}+e-v_{ij'},...,z_J)\mathrm{d}z_J...[\mathrm{d}z'_j]...[\mathrm{d}z_j]...\mathrm{d}z_0\mathrm{d}e$$

Using the change of variables $t = v_{ij}+e-v_{ij'}$ and the fact that $\frac{\partial v_{ij}}{\partial x_{ij'}} = 0$ for $j \neq j'$, one obtains:

$$\frac{\partial s^{\star}_{ij}(x_{i0},...,x_{iJ})}{\partial x_{ij'}} = \frac{\partial v_{ij'}}{\partial x_{ij'}}\int_{-\infty}^{\infty}\int_{-\infty}^{v_{ij'}+t-v_{i0}} ... \left[\int_{-\infty}^{v_{ij'}+t-v_{ij'}}\right] ... \left[\int_{-\infty}^{v_{ij'}+t-v_{ij}}\right] ... \int_{-\infty}^{v_{ij'}+t-v_{iJ}}$$
$$f(z_0,...,v_{ij'}+t-v_{ij},...,t,...,z_J)\mathrm{d}z_J...[\mathrm{d}z'_j]...[\mathrm{d}z_j]...\mathrm{d}z_0\mathrm{d}e \quad (4)$$
$$= \frac{\partial v_{ij'}}{\partial x_{ij'}}\frac{1}{\frac{\partial v_{ij}(x_{ij})}{\partial x_{ij}}}\frac{\partial s^{\star}_{ij'}}{\partial x_{ij}}$$

Evaluated at $x_{ij} = x_{ij'} = \bar{x}$, we have:

$$\frac{\partial s^{\star}_{ij}(x_{i0},...,\bar{x},...,\bar{x},...x_{iJ})}{\partial x_{ij'}} = \frac{\partial v_{ij'}(\bar{x})}{\partial x_{ij'}}\frac{1}{\frac{\partial v_{ij}(\bar{x})}{\partial x_{ij}}}\frac{\partial s^{\star}_{ij'}}{\partial x_{ij}}$$
$$= \frac{\partial s^{\star}_{ij'}(x_{i0},...,\bar{x},...,\bar{x},...x_{iJ})}{\partial x_{ij}} \quad (5)$$

This result is the lynchpin of our identification strategy. Any full-information discrete choice model satisfying the above assumptions satisfies cross-derivative symmetry. This symmetry need not hold with consideration sets - and we can use the resulting asymmetries to recover consideration set probabilities.

2.2 Consideration Sets

We generalize the full-information model to allow for the possibility that consumers are aware of only a subset of goods. We assume that option 0 is chosen by default if consumers are aware of no other goods.

We allow the consideration set probabilities for each alternative to depend on the characteristics of all alternatives via a linear index $y_{ij} = x_{ij}\gamma$. We assume that option j is considered if:

$$h_j(y_{i0}, ..., y_{iJ}) + \eta_{ij} \geq 0 \tag{6}$$

where we further assume:

ASSUMPTION 6. The joint distribution function $G_\eta(\eta_{i1}, ..., \eta_{iJ})$ is absolutely continuous with respect to the Lebesgue measure and gives rise to a density function that is everywhere positive on $\mathbb{R}$.

ASSUMPTION 7. *Exogeneity:* $\eta_{ij} \perp\!\!\!\perp \mathbf{x}_{ik}$ and $\eta_{ij} \perp\!\!\!\perp \epsilon_{ik}$ for $\forall i$ and $\forall j, k$.

This functional form is again not without loss although it nests several cases of interest, including all of the discrete choice models with consideration sets we have seen estimated in practice. For example, Goeree (2008) assumes that attention probabilities for good j depend only on the characteristics of good j so that $h_j(y_{i0}, ..., y_{iJ})) = y_{ij}$ and η_{ij} is i.i.d. logit. Ho, Hogan, and Scott-Morton (2015) assumes instead that attention probabilities for all goods are perfectly correlated and depend only on the default good so that $h_j(y_{i0}, ..., y_{iJ})) = y_{i0}$ and $\eta_{ij} = \eta_{i0}$, assumed to be logit. The single index assumption implies that for any two variables in x_{ij} denoted by k and k', $\frac{\partial P(c|x)}{\partial x_{ij}^k} / \frac{\partial P(c|x)}{\partial x_{ij}^{k'}}$ is equal for all j. In the health plan example considered below, if a \$100 increase in premiums for plan j impacts the probability that an alternative is considered twice as much as adding a \$100 deductible to plan j, then the same must be true of plan k. This is the consideration-set analogue of the assumption that the coefficients on product characteristics in utility are not alternative specific.

The exogeneity assumption is also worth highlighting. If for example consideration sets arise due to the endogenous actions of suppliers who are choosing which items to stock in response to shocks to demand, this assumption might be unlikely to hold. In such circumstances, one would need to generalize the model considered here to allow for such correlations.

The set of goods that a consumer pays attention to is referred to as the 'consideration set'. Without further restrictions, there exist $\mathcal{C} = 2^{J+1} + 1$ potential consideration sets — all possible subsets of the $J + 1$ goods. With $I(\cdot)$ as the indicator function, the probability of a consideration set $c \in \mathcal{C}$ is given as:

$$Pr(c|\mathbf{x}_{i0},...\mathbf{x}_{iJ}) = \int_{-(I(1\in c)q_{i0}+I(1\notin c)\infty)}^{(I(1\in c)\infty-I(1\notin c)q_{i0})} ... \int_{-(I(J\in c)q_{iJ}+I(J\notin c)\infty)}^{(I(J\in c)\infty-I(J\notin c)q_{iJ})} g(z_1,...,z_J)\ \mathrm{d}z_J...\mathrm{d}z_0$$
$$= \pi_c(q_{i0},...,q_{iJ}) \tag{7}$$

Let $\pi_\emptyset$ denote the probability of an empty consideration set (in which case the default good is chosen).

The probability of selecting option j then becomes:

$$s_{ij}(\mathbf{x}_{i0},...,\mathbf{x}_{iJ}) = I(j=d)\pi_\emptyset + \sum_{c:j\in c} \pi_c s^\star_{ij}(c) \tag{8}$$

where $s^\star_{ij}(c)$ denotes the probability of choosing j conditional on the consideration set.

2.3 Identification of Consideration Probabilities

In what follows, we assume that $F(\cdot)$ and $G(\cdot)$ are known but can be any distribution (subject to the independence assumptions stated above). Without any restrictions on these structures, a scale normalisation is required.

Normalisation 1. *Attention Normalisation:* The γ coefficients are only defined up to a scale normalization. Thus we normalize the coefficient on the first characteristic x^1_{ij} in the attentive equation to be 1:

$$\gamma^1 = 1 \tag{9}$$

This leaves us with a constant and $K-1$ other coefficients to estimate. We make use of an "envelope theorem" like simplification where - because the cross-derivatives of the market shares conditional on a consideration set are symmetric - the difference of unconditional cross-derivatives depends only on how consideration set probabilities shift with characteristics. First note that we can write the cross-derivatives with respect to characteristic k as:

$$\frac{\partial s_{ij}}{\partial x^k_{ij'}} = \sum_{c=1}^{\mathcal{C}} \left[\frac{\partial \pi_c}{\partial x^k_{ij'}} s^\star_{ij}(c) + \pi_c \frac{\partial s^\star_{ij}(c)}{\partial x^k_{ij'}} \right] \tag{10}$$

Given Corollary 1,

$$
\begin{aligned}
\frac{\partial s_{ij}}{\partial x^k_{ij'}} - \frac{\partial s_{ij'}}{\partial x^k_{ij}} &= \sum_{c=1}^{\mathcal{C}} \left[\frac{\partial \pi_c}{\partial x^k_{ij'}} s^\star_{ij}(c) - \frac{\partial \pi_c}{\partial x^k_{ij}} s^\star_{ij'}(c) \right] + \sum_{c=1}^{\mathcal{C}} \pi_c \left[\frac{\partial s^\star_{ij}(c)}{\partial x^k_{ij'}} - \frac{\partial s^\star_{ij'}(c)}{\partial x^k_{ij}} \right] \\
&= \sum_{c=1}^{\mathcal{C}} \left[\frac{\partial \pi_c}{\partial x^k_{ij'}} s^\star_{ij}(c) - \frac{\partial \pi_c}{\partial x^k_{ij}} s^\star_{ij'}(c) \right] \\
&= \gamma^k \sum_{c=1}^{\mathcal{C}} \left[\frac{\partial \pi_c}{\partial q_{ij'}} s^\star_{ij}(c) - \frac{\partial \pi_c}{\partial q_{ij}} s^\star_{ij'}(c) \right]
\end{aligned} \tag{11}
$$

Thus,

$$
\boxed{\frac{\partial s_{ij}/\partial x^k_{ij'} - \partial s_{ij'}/\partial x^k_{ij}}{\partial s_{ij}/\partial x^1_{ij'} - \partial s_{ij'}/\partial x^1_{ij}} = \gamma^k} \tag{12}
$$

[TO ADD:

1) Invoke mixture model results to show identification of conditional distribution of v_{ij} given γ

2) Conditions for identification of constant γ_0. Functional form identification in logit. Semi-parametrically, have identification as long as choice probability varies with x_{ij}. Need restrictions on utility?

3) Identification of fixed effects - similar to identification of constant, but need variation in attentive probability for each product]

2.4 Elasticities

Thus far, we have shown that - given a broad class of parametric restrictions commonly imposed in discrete choice models - consideration set models relax an important restriction on how choice probabilities vary with observable characteristics. We now explore more deeply how price elasticities vary across goods in consideration set models and how this differs from the usual full information models.

Let $\varepsilon_{s_{ij},x^k_{ij'}} = \frac{\partial s_{ij}}{\partial x^k_{ij'}} \frac{x^k_{ij'}}{s_{ij}}$, the elasticity of the choice probability of good j with respect to characteristic k of good j'. Combining equations 8 and 10 and rearranging terms, we obtain:

$$
\varepsilon_{s_{ij},x^k_{ij'}} = \sum_c p_{ij}(c) \left(\varepsilon_{\pi_{ij}(c),x^k_{ij'}} + \varepsilon_{s^*_{ij}(c),x^k_{ij'}} \right) \tag{13}
$$

where $p_{ij} = \frac{\pi(c) s^*_{ij}(c)}{1(j=d)\pi_\emptyset + \sum_{c'} \pi(c') s^*_{ij}(c')}$, the conditional probability that good j is chosen from choice set c given that good j is chosen in a consideration set model.

The symmetry argument in Section 2 suggests that, for two goods with equal market share, if $\varepsilon_{\pi_{ij}(c),x^k_{ij'}} = \varepsilon_{\pi_{ij'}(c),x^k_{ij}} = 0$, and we evaluate the elasticity at $x^k_{ij} = x^k_{ij'} = \bar{x}$ then we will have: $\varepsilon_{s_{ij},x^k_{ij'}} = \varepsilon_{s_{ij'},x^k_{ij}}$. In other words, the asymmetries in the model arise because of differences in how consideration set probabilities respond to product characteristics.

Models applied in practice will typically impose further constraints on how elasticities vary across products and allowing for consideration sets will relax these constraints. Specifically, ignoring for the moment the $\varepsilon_{\pi_{ij}(c),x^k_{ij'}}$ term, equation 13 will imply that elasticities for a given product are a weighted average of the elasticities conditional on each consideration set in which the product appears. As we discuss below, commonly estimated parametric models impose strong restrictions on how - for example - own price elasticities vary across products and consideration set models will allow for additional flexibility in this regard; specifically, they will tend to predict that own-price elasticities vary in a systematic way with the factors that predict attention. We discuss this further in sections 3.1.2 and 3.2.2.

3 Special Cases of Interest

In this section, we consider two special cases of the model that have been estimated in the applied literature before, but using stronger identifying assumptions.

Roughly, we can think of the first model - from Goeree (2008) - as being more appropriate in cross-sectional data without a clear default while the second model, from Ho, Hogan, and Scott-Morton (2015), is more appropriate in settings where inertia or choice of defaults explains a large fraction of choices.

In both cases, we show that the model with consideration sets is equivalent to a random utility model in which consideration sets are absent but the utility of each good depends in a particular way on the characteristics of rival goods. We use this representation to clarify the patterns in the data that identify inattention according to our general proof. We also highlight how both models relax restrictions on how price elasticities vary across goods which are often imposed in practice but are not imposed by the general framework considered in Section 2.

3.1 Alternative Specific Inattention

A natural special case of the model described above is the case where there is a probability of paying attention to each good which is a function of the utility-relevant characteristics of that good and potentially other non-utility relevant characteristics of the good or individual. We further assume that these probabilities are independent across goods - this is the case studied by Goeree (2008), although unlike us, she relies on the presence of non-utility relevant characteristics for identification.

Let A_{ij} be an indicator for individual i paying attention to product j. Then for each good we have a probability of attention:

$$\phi_{ij} = P(A_{ij} = 1 | x_{ij}) \tag{14}$$

which we model as a binary choice model. In other words:

$$\begin{aligned} P(A_{ij} = 1 | x_{ij}) &= P(A^*_{ij} > 0) \text{ and} \\ A^*_{ij} &= x_{ij}\gamma + \eta_{ij} \end{aligned} \tag{15}$$

Like Goeree (2008), we can of course impose that $\beta_k = 0$ for some characteristics but the above proof demonstrates that this is not necessary for identification.

One interpretation of this model is that consumers choose whether or not to pay attention to each good as a function of that good's characteristics. But this need not be the case. If for example consumers happen to be more likely to consider goods which advertise more but they are not consciously doing so, the model could still represent their choices. The substantive point is that consumers only consider a subset of the available choices and their probability of considering each subset of choices depends on the characteristics of that good and potentially other individual-level characteristics.

If, following Goeree (2008), we assume further that the η_{ij} are independent across goods, then we can compute the probability of a choice set c as:

$$P(c|x) = \prod_{l \in c} \phi_{il} \prod_{k \notin c} (1 - \phi_{ik}) \tag{16}$$

This is a special case of the model defined in Section 2 with $h_j(y_{i0}, ..., y_{iJ})) = y_{ij}$ and η_{ij} is i.i.d. across goods.

3.1.1 Identification

To gain some intuition for how this model differs from a standard discrete choice model, note that if the conditional choice probabilities are random coefficients logit (the ϵ_{id} terms are i.i.d. extreme value), then we can write this model as a model where the utility of each alternative is given by:

$$u_{ij} = x_{ij}\beta_i + \psi_{ij} + \epsilon_{ij} \tag{17}$$

where:

$$\begin{aligned} \psi_{ij} &= \ln\left(\frac{P_{ij}}{1 - P_{ij}}\right) - \ln\left(\frac{P_{ij}^*}{1 - P_{ij}^*}\right) \\ &= \ln\left(\frac{P(A_j|x_{ij}) \sum_{k \neq j} \exp(x_{ik}\beta_i + \psi_k)}{(1 - P(A_j|x_{ij})) \exp(x_{ij}\beta_i) + \sum_{k \neq j} \exp(x_{ik}\beta_i + \psi_k)}\right) \end{aligned} \tag{18}$$

where P_{ij} is the probability that option j is chosen and P_{ij}^* denotes the probability of choosing option j conditional on paying attention to that option. We show this formally in Appendix A. In other words, the inattention term is the difference between (a monotonic function of) the observed probability of choosing the default option and what that probability would be if consumers were paying attention.[4]

Unlike a traditional logit model, the utility of good j in this representation depends directly on the characteristics of other goods via the ψ_{ij} term if the consumer has some probability of being inattentive. While these characteristics do not directly impact the probability of inattention, they

[4]Note that if we observed some subset of consumers that we knew were paying attention *and* we knew had exactly the same preferences and choice set as inattentive consumers, then we could estimate $P_{id}^{\star}$ and directly compute ψ_{ij}. In practice however, this condition is unlikely to be met. Consider the context of health insurance plan choice. One might consider using the choices of new enrollees making a de novo choice to estimate $P_{id}^{\star}$ and then compare those choices to P_{id} estimated among returning enrollees. This method would incorrectly assume returning enrollees have no true adjustment costs or persistent unobserved preferences. The proof in section 2.3 shows that this model is identified without observing any such consumers.

do impact the *difference* between the probability that plan is chosen with full attention and the actual choice probability with inattention. Suppose plan j is the only plan to which one might be inattentive (so $\psi_{ik} = 0$ for $k \neq j$). With full attention, a change in the characteristics of rival goods would impact the choice of good j via substitution. Allowing utility to depend directly on the characteristics of rival goods effectively undoes this dependence in a way proportional to the degree of inattention. If the observable characteristics of good k become less desirable, then the gap between actual and fully-attentive choice probabilities will become larger thus directly shrinking the utility from good k so that substitution does not in fact occur. While this representation is a positive reflection of choices, we do not advocate including ψ_{ij} in the normative utility function used to evaluate welfare.

Identification comes from the fact that changes in the characteristics of good j have an asymmetric impact on the choice of good k. If the consumer is inattentive to good j, changes in the characteristics of good j impact good k through changes in both $x_{ij}\beta_i$ and ψ_{ij}. The impact via the first channel is smaller than we would expect with full attention (and symmetric with the impact of characteristics of good k on the choice of j), but the impact via the second channel does not exist with full attention. Thus, if consumers are inattentive to good j and if characteristics impact attention with the same sign as their impact on utility, then we expect the impact of a change in the characteristics of good j on the choice of good k to exceed in magnitude the impact of a change in the characteristics of good k on the choice of good j.

This pattern makes sense in some contexts but not others. Suppose that price reduces both attention and utility so that the Goeree (2008) model predicts that the market share of an attentive good responds more to a change in the characteristics of an inattentive good than vice-versa. For a consumer choosing from an ordered list of items, we might observe that when the price of one of the top 3 items increases it has little effect on an item ranked 20th but when the price of an item ranked 20th decreases we do see an effect on the top 3 items because the 20th ranked item moves up in the list and is brought to consumers' attention. This would be consistent with the model. Alternatively, we might observe that a consumer choosing amongst health plans is sensitive to changes in the plan they chose last year that they are paying attention to but insensitive to changes in rival plans. This would contradict the pattern implied by Goeree (2008) and motivates our investigation of the model in Section 3.2.

3.1.2 Elasticities

In the conditional logit version of this model, we can obtain a simple analytical expression to help us understand how elasticities vary across products. Consider own-good elasticities and to illustrate the effects, consider a variable that impacts utility but not attention (i.e. imagine we are in a world where attentiveness varies across goods, but not because of price). Then:

$$\begin{aligned}
\varepsilon_{s_{ij},x_{ij}^k} &= \sum_c p_{ij}(c)\left(\varepsilon_{s_{ij}^*(c),x_{ij}^k}\right) \\
&= \sum_c p_{ij}(c)\left(\beta_k x_{ij}(1-s_{ij}^*(c))\right) \\
&= \beta_k x_{ij}\left(1-\sum_c p_{ij}(c)s_{ij}^*(c)\right) \qquad (19)
\end{aligned}$$

Thus, in this model, elasticities are inversely proportional to a weighted average of the conditional market share. For a given observed market share, if that market share arises because a product has low utility but appears in many choice sets, $s_{ij}^*(c)$ will be small, meaning that the observed elasticities will be larger in magnitude. Conversely, if a product has high utility but appears in few choice sets, $s_{ij}^*(c)$ will be large, meaning that elasticities will be small. Thus, for given market share, we will tend to observe large elasticities for products to which consumers are attentive and small elasticities for products to which consumers are inattentive.

To be clear, this is *not* a semiparametric identification result. As our proof makes clear, semiparametric identification of the attention probabilities comes from the $\varepsilon_{\pi_{ij}(c),x_{ij'}^k}$ terms - we are specifically ignoring those terms here and instead asking what happens to own-price elasticities across goods as attention exogenously shifts. This is not generally a source of identification because the model written in section 2 can allow own-price elasticities to vary across goods very flexibly.

Our point is that, subject to the additional strong parametric assumptions of a conditional logit model, the Goeree (2008) model predicts that own price elasticities vary across goods in a systematic way with the degree of attention. The same tendency will also be present in nested logit and random coefficients models if we exogenously shift attention and utility for a given market share. In other words, the Goeree (2008) model provides a particularly parsimonious way of allowing for substitution behavior in which: a) elasticities are asymmetric (identifying the degree of inattention) and b) elasticities vary systematically across goods with that degree of inattention. We can test whether in practice elasticities vary in this way.

3.1.3 Estimation

Goeree (2008) provides details of the estimation process. We sketch the main ideas here. With a small number of available alternatives, estimation in the alternative-specific inattention model is straightforward. The probability of choosing any specific alternative as a function of the parameters $\theta = (\beta, \gamma)$ is given by:

$$P(Y_{ij}=1|\theta) = P(c=\emptyset|\theta)\cdot \mathbb{1}_{j=d} + \sum_{c\in C}\prod_{l\in c}\phi_{il}(\theta)\prod_{k\notin c}(1-\phi_{ik}(\theta))P(Y_{ij}=1|c,\theta) \qquad (20)$$

We can use this to construct the likelihood function and then estimate the parameters β and γ by maximum likelihood.

In larger choice sets, a major computational issue arises - there are 2^J possible consideration sets to sum over. To deal with this problem, we follow Goeree (2008) in using a simulated likelihood approach. The basic idea is to estimate the term $\sum_{c\in C}\prod_{l\in c}\phi_{il}(\theta)\prod_{k\notin c}(1-\phi_{ik}(\theta))P(Y_{ij}=1|c,\theta)$ by simulating R consideration sets per individual where, for each r, each option is added to the consideration set with probability ϕ_{ij} so that the probability a given consideration set is simulated is given by: $\prod_{l\in c}\phi_{il}\prod_{k\notin c}(1-\phi_{ik})$. We then compute:

$$\hat{P}_{ij} = \frac{1}{R}\sum_{r} P(Y_{ij}=1|c_r,\theta) \tag{21}$$

Since each c_r is chosen with probability $\prod_{l\in c}\phi_{il}\prod_{k\notin c}(1-\phi_{ik})$, we have that:

$$\begin{aligned}\hat{P}_{ij} \quad &\to_p \quad \frac{1}{R}\sum_{r}\sum_{c\in C}\prod_{l\in c}\phi_{il}(\theta)\prod_{k\notin c}(1-\phi_{ik}(\theta))P(Y_{ij}=1|c,\theta)\\ &= \quad \sum_{c\in C}\prod_{l\in c}\phi_{il}(\theta)\prod_{k\notin c}(1-\phi_{ik}(\theta))P(Y_{ij}=1|c,\theta) \end{aligned}\tag{22}$$

This procedure would still be computationally burdensome because it would require computing $P(Y_{ij}=1|c_r)$ for every simulation r for all individuals at each candidate set of parameter values (since as the underlying parameters shift, the ϕ, and thus the choice sets would shift).

Following Goeree (2008), two additional tricks are used so that the choice probabilities need to be evaluated only once per person for each simulation r. First, we use the same uniform draws to simulate choice sets at each set of parameter values. Second, we use an importance sampler so that the choice probabilities need only be evaluated at the consideration sets implied by the parameters at their initial values. Specifically, we can compute equation 21 using:

$$\hat{P}_{ij} = \frac{1}{R}\sum_{r}\prod_{l\in c}\phi_{il}\prod_{k\notin c}(1-\phi_{ik})\frac{P(Y_{ij}=1|c_0,\theta)}{\phi^0_{ir}(\theta_0)} \tag{23}$$

where $\phi^0_{ir}(\theta_0) = \prod_{l\in S_0}\phi_{il}\prod_{k\notin S_0}(1-\phi_{ik})$ and each consideration set is sampled with probability $\phi^0_{ir}(\theta_0)$.[5]

3.2 Inattention and Adjustment Costs

A second special case is one in which consumers have correlated inattention probabilities across all goods and where there is a clear default amongst the "inside goods". This model can be used to identify whether inertia observed in panel data arises because consumers do not consider other options (and thus might be better off if they switched) or because consumers are actively choosing not to switch due to adjustment costs or persistent unobserved heterogeneity.

One such model is considered in Ho, Hogan, and Scott-Morton (2015). In their interpretation, consumers choose in two stages. First, they decide whether to be attentive or not as a function of unobservables and the characteristics of the default good and second, if they are attentive, they make an active choice among all goods. Following their notation, we write the choice probabilities as a function of the probability of *inattention*, defined as one minus the attentive probability. This

[5]Recall that an importance sampler estimates a density $f(x)$ by drawing from a density $g(x)$, labeling the resulting value as x_1 and then weighting each draw by $f(x_1)/g(x_1)$. The resulting density is equivalent to drawing directly from $f(x)$.

implies that choice probabilities are given by:

$$
\begin{aligned}
P(Y_{id}=1) &= P(I_i|x_{id},z_i)+(1-P(I_i|x_{id},z_i))P_a(Y_{id}=1) \\
P(Y_{ij}=1) &= (1-P(I_i|x_{id},z_i))P_a(Y_{ij}=1) \text{ for } j\neq d
\end{aligned} \tag{24}
$$

where $P_a(Y_{ij}=1)$ is the probability of choosing plan j conditional on paying attention.

We assume that consumers have utility given by:

$$u_{ij} = x_{ij}\beta_i + \xi_{i,j=d}(x_{id}) + \epsilon_{ij} \tag{25}$$

where $\xi_{i,j=d}$ is the traditional adjustment cost term which takes the value $\xi_i(x_{id})$ for plan d and is 0 otherwise and ϵ_{ij} is i.i.d. extreme value. Note that ξ is allowed to vary with the characteristics of the default plan but cannot generally vary with the characteristics of other alternatives. This is a crucial identifying assumption (and is in our view quite natural in most cases).

This is a special case of the model defined in Section 2 with $h_j(y_{i0},...,y_{iJ})) = y_{i0}$ and η_{ij} perfectly correlated across all goods.

3.2.1 Identification

As in the alternative-specific model, the above model is equivalent to a standard logit model with an additional inertial term:

$$u_{ij} = x_{ij}\beta + \xi_{i,j=d}(x_{id}) + \psi_{i,j=d} + \epsilon_{ij} \tag{26}$$

where $\psi_{i,j=d}$ takes the value ψ_i for plan d and is 0 otherwise. We show in Appendix A that ψ_i is given by:

$$\psi_i = \ln\left(\frac{1+P(I_i|x_{id},z_i)\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i - \xi_i(x_{id}))}{1-P(I_i|x_{id},z_i)}\right) \tag{27}$$

The term $\xi_{i,j=d}(x_{id})$ in this model can be thought of as all of the reasons why an attentive consumer might nonetheless prefer to choose the same plan - for example, because there are adjustment costs to switching or persistent unobserved heterogeneity.[6] The $\psi_{i,j=d}$ term by contrast captures the possibility that the consumer chose the default plan not because it had higher utility, but simply because they were inattentive to the available options.

Identification in this model parallels identification in the alternative specific case. We observe that beneficiaries are inertial and we want to know - is this because of adjustment costs or inattention? If it is because of adjustment costs ($\psi_i = 0$) then consumers will readily switch to alternative options if those options become more desirable. If it is because of inattention, consumers will be insensitive to characteristics of alternative plans. They may however still switch in response to changes

[6]The inertia term ξ is not formally identical to a model where error terms are correlated over time. Abaluck and Gruber (2016) gives one example of a model which allows for both possibilities in the empirical setting of health plan choice we consider below. Nonetheless, in a model that does not allow for such a correlation, the ξ term may proxy for it.

in the characteristics of the default plan because these changes are allowed to impact the degree of inattention. Thus, in the health plan context, we may see the characteristic pattern documented in Ho, Hogan, and Scott-Morton (2015) wherein consumers readily switch when the premiums of their prior year plan increase but do not switch when the premiums of alternative plans fall. Provided we observe enough determinants of attentive behavior (which again, need not include any characteristics beyond those which enter utility), we can separately identify ξ and ψ based on the asymmetry in how the share of rival plans responds to changes in the characteristics of the default good relative to how the share of the default good responds to changes in the characteristics of rival plans.

3.2.2 Elasticities

In this model, the identifying asymmetry is between how choice probabilities respond to characteristics of the non-default good vs. characteristics of the default good. The elasticity of a good j with respect to characteristics of the non-default good $j' \neq d$ is given by:

$$\varepsilon_{s_{ij},x^k_{ij'}} \quad = \quad p_{i,j=d}\varepsilon_{s^*_{ij},x^k_{ij'}} \tag{28}$$

where $p_{i,j=d}$ is the conditional probability that you are paying attention given that you choose the good. For a non-default good, this is 1; for the default good, this is $p_i = \frac{P(A_i)s^*_{id}}{(1-P(A_i))+P(A_i)s^*_{id}}$.

The elasticity with respect to characteristics of the default good is given by:

$$\varepsilon_{s_{ij},x^k_{id}} = \alpha_{i,j=d}\varepsilon_{P(A_i),x^k_{id}} + p_{i,j=d}\varepsilon_{s^*_{ij},x^k_{id}} \tag{29}$$

where $\alpha_{i,j=d}$ is 1 for $j \neq d$ and equals $\alpha_i = \frac{s_{id}-1}{s_{id}}$ for the default good.

In this model then, the elasticity of the default good with respect to characteristics of the non-default good j is given by: $p_i\varepsilon_{s^*_{id},x^k_{ij}}$, while the elasticity of j to characteristics of the default good is given by: $\varepsilon_{P(A_i),x^k_{id}} + \varepsilon_{s^*_{ij},x^k_{id}}$. Note that with equal observed market shares and characteristics, we will have: $p_i\varepsilon_{s^*_{id},x^k_{ij}} = \varepsilon_{s^*_{ij},x^k_{id}}$. The asymmetry arises from the fact that default good characteristics also impact the probability of attention. If they do so in the same direction as they impact utility, then the elasticity of characteristics of good j with respect to the default good will be larger than the elasticity of the default good with respect to characteristics of good j.

3.2.3 Estimation

To estimate this model by the usual methods, one must assume a functional form for $P(I_i|x_{id})$. If we assume that $P(I_i|x_{id})$ is itself given by a logit, we obtain a particularly simple expression. This is the same assumption used in HHS. Suppose consumers are inattentive whenever:

$$x_{id}\beta + \epsilon_{id} > f(z_i) + v_i \tag{30}$$

where x_{id} are characteristics of the default good and z_i is a vector of other individual characteristics and ϵ_{id} and v_i are both type 1 extreme value. Then the probability of being inattentive is:

$$P(I|x_{id}) = \frac{\exp(x_{id}\beta)}{\exp(f(z_i)) + \exp(x_{id}\beta)} \tag{31}$$

and we can simplify the inattention term to:

$$\psi_{i,j=d} \quad = \quad \ln\left[1+\frac{\sum_j \exp(x_{ij}\beta)}{\exp(f(z_i))}\right] \tag{32}$$

Note first that we do not need to observe any additional individual characteristics in order to estimate this model. We can assume that $f(z_i) = 0$ and the model is still identified. Including individual characteristics just produces a more flexible model of inattention and thus reduces the likelihood that the error term is misspecified due to heteroscedasticity.

4 Estimation Results

In this section, we present estimation results from the models described above. First, we show in a simulation that - when consideration sets are present - the Goeree (2008) model without instruments recovers the true parameters while conventional models are badly misspecified. Second, we consider an application to Expedia Hotel data. We show that the Goeree model implies that consumers pay far more attention to hotels which rank higher in search results and that price elasticities are much larger for these hotels. Third, we consider an application to Medicare Part D, and we show that most of the observed inertia in plan choices appears to result from inattention rather than adjustment costs.

4.1 Simulation Results

To validate that the maximum likelihood estimation method proposed by Goeree (2008) does in fact recover the true parameters we conduct a simulation exercise.

We simulate 5,000 beneficiaries with 10 choices and prices uniformly distributed between 1 and 5. Utility is given by $u_{ij} = \beta p_{ij} + \epsilon_{ij}$ and the attention probability is given by $P(A_{ij}) = \Delta(\gamma_0 + \gamma_1 p_{ij})$ where $\Delta(\cdot)$ is the logistic function. The table shows the results for alternative values of β_1 and γ_1. In each case, γ_0 is chosen so that the average attention probability is 0.25.

Table 1 shows the results from estimating this model using both a conventional conditional logit command and the "attentive logit" model, implemented using the method described in Section 3.1.3. The goodness of fit measure reported is the average absolute deviation between the choice probability given the true parameters and the choice predicted by the model given the estimated parameters.

The attentive logit model recovers the true parameters. The average absolute deviation between the attentive logit market share and the observed market share is in all cases less than 0.4 percentage points. By contrast, the logit model shows systematic biases. There are two offsetting effects - β tends to be biased towards 0 because choices are not sensitive to variation in the prices of products to which consumers are inattentive but β is biased upwards because consumers are less likely to be attentive to goods with higher prices which looks like they dislike those prices. The net effect is ambiguous, but the fit of the model to choice probabilities is poor because the model fails to account for the fact that choices are more elastic for the goods which consumers are more likely to pay attention to. In some cases, the fit can be quite poor. The average market share in this case

Table 1: Simulation Results: Conditional Logit Model vs. Attentive Logit

beta_1	gamma_1	logit beta_1	alogit beta_1	alogit gamma_1	logit abs dev	alogit abs dev
-0.5	-0.5	-0.6202	-0.5334	-0.4489	0.032	0.0031
-1	-0.5	-0.9188	-1.0572	-0.4187	0.0185	0.0036
-2	-0.5	-1.3003	-2.0432	-0.4409	0.0105	0.0028
-0.5	-1	-0.8597	-0.5447	-0.9507	0.0419	0.0023
-1	-1	-1.1774	-1.0631	-0.9314	0.0266	0.0011
-2	-1	-1.6186	-2.0329	-0.9585	0.0173	0.001
-0.5	-2	-1.137	-0.4904	-2.0311	0.0758	0.0039
-1	-2	-1.4769	-0.9276	-2.0578	0.0594	0.0026
-2	-2	-2.0969	-2.0274	-2.0776	0.049	0.0038

is 0.1. When the true $\beta_1 = -1$ and $\gamma_1 = -2$, we find that the average absolute deviation of the market share is 7.6 percentage points given an average market share of 10 percentage points.

[A lab experiment in which consumers make actual choices given known underlying attention parameters in in progress. This further tests whether the model recovers the true parameters given actual choices where - for example - the model may be misspecified as choices are not guaranteed to be a linear function of prices and unobserved factors are not guaranteed to be i.i.d. logit]

4.2 Expedia Results

We next apply the model to a dataset with information about consumers' online hotel choices from Expedia. The data are publicly available from Kaggle.com. This dataset is particularly attractive because - in a subset of the data - the order in which hotels were listed in search results was randomized. Ursu (2015) describes several sample selection restrictions designed primarily to clean the data (e.g. dropping all hotels with prices of less than \$10 / night or more than \$1,000 / night). We impose the exact same sample selection restrictions as Ursu (2015) with two exceptions: we restrict to the top 10 choices and we do not restrict to the 4 largest hotel destinations.

Given that the order of the hotels was randomized, we might expect the position of the hotels in the search results to impact only attention and not utility. This need not be the case - Expedia did not inform consumers that the order was randomized so they may believe that higher ranked hotels are better in some unobservable respect. The estimation results with a conditional logit model and the Goeree (2008) model, referred to as the "attentive logit" model, are shown in Table 2. We see that while the conditional logit model says that position has a larger impact on utility than any of the observed characteristics, the attentive logit model reveals that position matters for attention and not utility. Intriguingly, the model also suggests that promotions in which some hotels are highlighted impact attention but not utility.

Table 3 shows how choice probabilities and attentive probabilities vary with the ranking. The model suggests that the attentive probability ranges from 0.3 for a hotel in the 10th position to 0.6 for the highest ranked hotel. We also compare the price elasticities estimated in the conditional logit model with the attentive logit model. The logit model seems to modestly attenuate own-price

Table 2: Expedia Data: β and γ

		Conditional Logit		Attentive Logit	
Utility		coef	std error	coef	std error
	Hotel Stars (1-5)	0.63	0.05	0.87	0.18
	Review Score (1-5)	0.40	0.05	0.52	0.13
	Popular Brand Dummy	0.07	0.06	0.37	0.17
	Location	0.16	0.03	-0.13	0.09
	Price	-0.02	0.00	-0.02	0.00
	Ongoing Promotion	0.22	0.06	-0.07	0.17
	Position	-0.12	0.01	-0.03	0.03
Attention					
	Hotel Stars (1-5)			0.01	0.18
	Review Score (1-5)			0.08	0.13
	Popular Brand Dummy			-0.40	0.25
	Location			0.37	0.10
	Price			0.00	0.00
	Ongoing Promotion			0.48	0.26
	Position			-0.15	0.03
	Constant			-0.33	0.83

elasticities, with an average error of about 10%.

Table 3: Expedia Data: Choice Probabilities and Elasticities

Position	Market Share	Attentive Probability	Clogit Elasticity	Alogit Elasticity
1	0.178222	0.60254121	-2.13	-2.3
2	0.151425	0.57203413	-2.14	-2.3
3	0.1208	0.53594103	-2.21	-2.36
4	0.100808	0.50647202	-2.2	-2.37
5	0.089324	0.4638763	-2.24	-2.44
6	0.078265	0.4298252	-2.26	-2.49
7	0.074862	0.39467121	-2.26	-2.49
8	0.061251	0.35690565	-2.32	-2.54
9	0.080817	0.32817872	-2.3	-2.54
10	0.064228	0.29712054	-2.36	-2.24

4.3 Medicare Part D Results

We also apply the model to evaluate whether the observed inertia in Medicare Part D plans is due to inattention, adjustment costs or both. Heiss, McFadden, Winter, Wupperman, and Zhou (2016) perform a similar exercise but rely on assumptions that some variables impact attention and not utility. We instead rely on the asymmetry between how the market share of the default plan responds to prices of alternative plans relative to how the market shares of alternative plans respond to prices of the default plans. We use the sample selection approach described in Abaluck and Gruber (2016) and consider plan choices in 2009.

We estimate the Ho, Hogan, and Scott-Morton (2015) model described in Section 3.2 to separately identify an inertial term $\xi_{j=d}$ and an adjustment cost term. In the conditional logit model, the ratio of the inertial term to premiums implies a willingness to pay of \$1,360 each year to remain in the same plan. This is larger than the total premiums and out of pocket costs paid by the majority of beneficiaries. In the attentive logit model, we find that this reflects almost entirely inattention. The dollar-equivalent $\xi_{j=d}$ term is \$30. The model reaches this conclusion because, as in Ho, Hogan, and Scott-Morton (2015), the elasticity of the switching probability with respect to the prices of plans other than the default plan is statistically indistinguishable from zero.

5 Conclusion

Discrete choice models with consideration sets relax the strong assumption that beneficiaries consider all of the options available to them before making a choice. In the literature to date, such models have been identified by either bringing in auxiliary information on what options consumers consider or assuming that some characteristics impact attention or utility but not both. This paper shows that these assumptions are unnecessary. This paper shows that a broad class of such models - broad enough to include all discrete choice models with consideration sets that have been applied to date - are identified from variation already available in the data. Consideration set probabilities can be constructively recovered from asymmetries in the matrix of cross-derivatives of choice probabilities with respect to characteristics of rival goods.

This identification result highlights one principle motivation for consideration sets - there are patterns of substitution which are not allowed for by conventional models which are permitted in consideration set models. Our empirical applications suggest that these types of substitution may be first-order in practice. There may be large asymmetries in cross-derivatives with respect to some characteristics. Models estimated in practice - such as random coefficients logit models - also often impose strong assumptions on how own-price elasticities vary across goods. While these assumptions could in principle be relaxed by - for example - allowing for heteroscedasticity in the idiosyncratic error term, consideration set models relax these assumptions in a particularly parsimonious way while making systematic and testable restrictions on how asymmetries in the cross-derivatives vary across characteristics for different goods.

Even putting aside their positive implications for how market shares respond to choice characteristics, consideration set models are attractive because they relax the typically implausible assumption that beneficiaries are aware of all of the alternatives which they might choose. Assuming exogeneity, a major reason we estimate random utility models rather than using purely statistical methods to model choices is that we believe the theoretical assumption of consumer optimization generates plausible restrictions on choices. But this theoretical assumption is often suspect, and it may be more plausible to assume that consumers are optimizing given only a subset of the information available to the econometrician. Consideration set models are one way of formalizing this notion.

We hope that the results in this paper will make it possible to adapt consideration set models to a wider range of settings than they have previously been applied. These models also enable us to consider counterfactuals and normative assumptions which are not possible in conventional models. We can ask, how might beneficiaries choose if they considered all available options? Do some demographic or choice set features (such as the number of plans) increase the likelihood that consumers are attentive? How much better off might consumers be if they were fully informed about the relevant choices? We hope that future work will explore these questions in more detail.

References

Abaluck, J. and J. Gruber (2016). Evolving choice inconsistencies in choice of prescription drug insurance. *The American Economic Review, Forthcoming*.

Conlon, C. T. and J. H. Mortimer (2013). Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics 5*(4), 1–30.

Crawford, G. S., R. Griffith, and A. Iaria (2016). Demand estimation with unobserved choice set heterogeneity.

Falmagne, J.-C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology 18*(1), 52–72.

Gabaix, X. (2011). A sparsity-based model of bounded rationality. Technical report, National Bureau of Economic Research.

Gaynor, M., C. Propper, and S. Seiler (2016). Free to choose? reform, choice, and consideration sets in the english national health service. *American Economic Review*.

Goeree, M. S. (2008). Limited information and advertising in the us personal computer industry. *Econometrica 76*(5), 1017–1074.

Heiss, F., D. McFadden, J. Winter, A. Wupperman, and B. Zhou (2016). Inattention and switching costs as sources of inertia in medicare part d. Working Paper.

Ho, K., J. Hogan, and F. Scott-Morton (2015). The impact of consumer inattention on insurer pricing in the medicare part d program. Technical report, National Bureau of Economic Research.

Honka, E. (2014). Quantifying search and switching costs in the us auto insurance industry. *The RAND Journal of Economics 45*(4), 847–884.

Honka, E., A. Hortaçsu, and M. A. Vitorino (2015). Advertising, consumer awareness, and choice: Evidence from the us banking industry. *Consumer Awareness, and Choice: Evidence from the US Banking Industry (September 16, 2015)*.

Manzini, P. and M. Mariotti (2014). Stochastic choice and consideration sets. *Econometrica 82*(3), 1153–1176.

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2016). Revealed attention. In *Behavioral Economics of Preferences, Choices, and Happiness*, pp. 495–522. Springer.

Ursu, R. M. (2015). The power of rankings: Quantifying the effects of rankings on online consumer search and choice. *Available at SSRN 2729325*.

A Proof of Utility Representations for Consideration Set Models

Consider first the Goeree model. Let's start by assuming there is a single plan to which you might be inattentive, plan 1 and an alternative, plan 0 to which you are attentive. Suppose that the probability of choosing good 1 is given by:

$$P(Y_{i1} = 1) \quad = \quad P(A_{i1}|x_{i1})P(Y^*_{i1} = 1) \tag{33}$$

where $P(Y^*_{i1} = 1)$ is the probability of choosing good 1 conditional on paying attention. In a logit setting, this is equivalent to a model where:

$$u_{i1} = x_{ij}\beta + \psi_{j=1} + \epsilon_{ij} \tag{34}$$

where $\psi_{j=1}$ takes the value ψ_1 for plan 1 and is 0 otherwise and ψ_1 is given by:

$$\psi_1 = \ln\left(\frac{P(A_1)\exp(x_{i0}\beta)}{(1 - P(A_1))\exp(x_{i1}\beta) + \exp(x_{i0}\beta)}\right) \tag{35}$$

I will prove that an analogous result holds in a J good model. Specifically, suppose there are $J - 1$ goods to which you might be inattentive and a default good 0 to which you are attentive with certainty (this is a normalization). For each of the $J - 1$ goods to which you might be inattentive:

$$P(Y_{ij} = 1) \quad = \quad P(A_{ij}|x_{ij})P(Y^a_{ij} = 1) \tag{36}$$

where Y^a_{ij} is the probability of choosing good j conditional on paying attention to that good. In a logit setting, this can be represented by writing:

$$u_{ij} = x_{ij}\beta + \psi_j + \epsilon_{ij} \tag{37}$$

where ψ_j is 0 for the default plan and is given by:

$$\psi_j = \ln\left(\frac{P(A_j)\sum_{k\neq j}\exp(x_{ik}\beta + \psi_k)}{(1 - P(A_j))\exp(x_{ij}\beta) + \sum_{k\neq j}\exp(x_{ik}\beta + \psi_k)}\right) \tag{38}$$

And if you are fully attentive to good J, then the same representation holds with $\psi_J = 0$.

Let's prove this by induction. I showed above that this representation holds for a 2 plan choice set. Next, suppose it holds for a $J - 1$ plan choice set. If we add a Jth plan to which you might be inattentive:

$$P(Y_{iJ} = 1) \quad = \quad P(A_{iJ}|x_{iJ})P(Y^a_{iJ} = 1) \tag{39}$$

By the inductive hypothesis, we have:

$$P(Y^a_{iJ} = 1) = \frac{\exp(x_{iJ}\beta)}{\exp(x_{iJ}\beta) + \sum_{k\neq J}\exp(x_{ik}\beta + \psi_k)} \tag{40}$$

And therefore:

$$P(Y_{iJ}=1) \quad = \quad \frac{P(A_{iJ}|x_{iJ})\exp(x_{iJ}\beta)}{\exp(x_{iJ}\beta)+\sum_{k\neq J}\exp(x_{ik}\beta+\psi_k)} \tag{41}$$

And it is straightforward to confirm that we obtain this representation if we set:

$$\psi_J = \ln\left(\frac{P(A_J)\sum_{k\neq J}\exp(x_{ik}\beta+\psi_k)}{(1-P(A_J))\exp(x_{iJ}\beta)+\sum_{k\neq J}\exp(x_{ik}\beta+\psi_k)}\right) \tag{42}$$

Next, consider the Ho, Hogan and Scott-Morton model.

$$\begin{aligned} P(Y_{id}=1) &= P(I_i|x_{id})+(1-P(I_i|x_{id}))P(Y^*_{id}=1) \\ P(Y_{ij}=1) &= (1-P(I_i|x_{id}))P(Y^*_{ij}=1) \text{ for } j\neq d \end{aligned} \tag{43}$$

where $Y^*_{ij}=1$ are the choices given by maximizing:

$$u^*_{ij} = x_{ij}\beta_i + \xi_{i,j=d} + \epsilon_{ij} \tag{44}$$

We want to show that this is equivalent to a model given by:

$$u_{ij} = x_{ij}\beta_i + \xi_{i,j=d} + \psi_{i,j=d} + \epsilon_{ij} \tag{45}$$

We will derive an expression for $\psi_{i,j=d}$ in the case where the ϵ_{ij} are i.i.d. type I extreme value. Let ψ_i and ξ_i denote the values of $\psi_{i,j=d}$ and $\xi_{i,j=d}$ when $j=d$. In this case, the probability of choosing the default plan is given by:

$$\begin{aligned} P(Y^*_{id}=1) &= \int \frac{\exp(x_{id}\beta_i+\xi_i)}{\sum_k \exp(u_{ik})} f(\beta,\xi)d\beta d\xi \\ &= \int \frac{1}{1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)} f(\beta,\xi)d\beta d\xi \end{aligned} \tag{46}$$

We want to solve for ψ_i satisfying:

$$\begin{aligned} &\int \frac{1}{1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i-\psi_i)} f(\beta,\xi)d\beta d\xi = \\ P(I_i|x_{id})+(1-P(I_i|x_{id}))&\int \frac{1}{1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)} f(\beta,\xi)d\beta d\xi \end{aligned} \tag{47}$$

We can rewrite the 2nd term as:

$$\int \frac{P(I_i|x_{id})(1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)+(1-P(I_i|x_{id}))}{1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)} f(\beta,\xi)d\beta d\xi \tag{48}$$

Thus, our problem reduces to finding ψ_i which satisfies:

$$\frac{1}{1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i-\psi_i)} = \frac{P(I_i|x_{id})(1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)+(1-P(I_i|x_{id}))}{1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)} \tag{49}$$

Taking reciprocals of both sides yields:

$$1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i-\psi_i) = \frac{1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)}{P(I_i|x_{id})(1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)+(1-P(I_i|x_{id}))} \tag{50}$$

Subtracting 1 gives:

$$\exp(-\psi_i)\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i) = \frac{(1-P(I_i|x_{id}))\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)}{P(I_i|x_{id})(1+\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)+(1-P(I_i|x_{id}))} \tag{51}$$

And so:

$$\psi_i = \ln\left(\frac{1+P(I_i|x_{id})\sum_{k\neq d}\exp((x_{ik}-x_{id})\beta_i-\xi_i)}{(1-P(I_i|x_{id}))}\right) \tag{52}$$