Exploring the Role of Consumer's State Dependence Behavior and Strategic Interaction Between a Retailer and Manufacturers

March 2017

Tomohito Kamai

Exploring the Role of Consumer's State Dependence Behavior and Strategic Interaction Between a Retailer and Manufacturers

Graduate School of Systems and Information Engineering

University of Tsukuba

March 2017

Tomohito Kamai

Acknowledgements

With great appreciation, I would like to express my sincere admiration for my academic advisor Professor Yuichiro Kanazawa. This dissertation would never have been completed without his insightful guidance and advice during all stages of my study. I would like to thank all staff of Tsukuba Asotk Foods Company for their understanding, and my special thanks go to Ryotaro Kanzaki for spiritual and material supports throughout the course of this research. I warmly thank and appreciate my parents for their supports in all aspects of my life. Finally, I thank my wife, Naho, and my children, Akihito, Chika, Yukihito, and Kiyoka for their patience, supports, and encouragements.

Contents

List of Tables

Chapter 1

Introduction

Understanding consumer's brand choice behavior is critical in formulating marketing strategies. One of the most significant factors of consumer brand choice behavior is a dynamic behavior whereby the consumer's past choices affect its current choice behavior. This phenomenon is widely known as state dependence.

"The empirical existence of state dependence is universally agreed upon (Freimer & Horsky, 2012, page 647)" and widely applied to the context of brand choice behavior. There are mainly two types of state dependence; positive and negative. They are sometimes referred to as inertia and variety-seeking respectively. In addition to these two state dependence, Bawa (1990) suggests another type of state dependence which he calls "hybrid behavior" whereby a consumer exhibits inertial behavior at first but becomes variety-seeking after a certain period of time.

Despite his unique attempt, there are several limitations in Bawa's model. Most importantly, his model does not account for consumer heterogeneity. Therefore in Chapter 2, we develop the hybrid model to accommodate consumer heterogeneity while increasing the number of brands treated by the model.

In an analysis of consumer good market, one must take into account for the fact that marketing activities of a firm would trigger the reactions from the other players in the market. Some researchers such as Sudhir (2001) and Villas-Boas & Zhao (2005) formulate such interactions among manufacturers and retailers as well as among manufacturers where each party behaves strategically based on its predictions of the other players' behavior and market conditions.

What has been missing in the previous studies is testing retailer Stackelberg formulation, whereby the retailer acts as market leader and manufacturers follow. Therefore in Chapter 3, we accommodate the purported power increase of retailers with respect to manufacturers[1] by mathematically deriving retailer Stackelberg model, extending the approach of Che et al. (2007). We empirically investigate whether manufacturers' effort to develop special featured brands is still rewarding in terms of profit margins using Japanese yogurt as an example. Formulating a new game theoretic framework to describe this phenomenon and testing it with the real data, albeit a small one, would be of great interest to researchers in the field as well as of practical importance to those working for innovative manufacturers facing similar circumstances.

In Chapter 4, we incorporate Nash bargaining framework into the framework employed in Chapter 3 to analyze the relationship between manufacturers and retailers more closely. Though papers in this field such as Misra & Mohanty (2008) and Draganska et al. (2010) well describe the behavior of

[1]The increased power of retailers is attributed to the emergence of giant retailers that exert strong buying power and enjoy economy of scale, their sophisticated information systems regarding consumers, and increased retailer concentration (Kim, 2010)

the market players, the retailer Stackelberg formulation is missing. Thus we accommodate the retailer Stackelberg game in the Nash bargaining model framework given power increase of retailer and test it using Japanese canned tuna as an example. Though the result is limited to a specific product category in a specific market, we believe that our result has a broader significance to the literature.

In summary, the organization of this dissertation is as follows: Chapter 2 exclusively focuses on consumer's brand choice behavior with hybrid state dependence. Chapter 3 extends the traditional market-wide framework in describing the relationship among firms to accommodate an important game reflecting the recent trend of increased power of retailers, namely retailer Stackelberg game. We bring this formulation to the Japanese yogurt market to examine if premium brands are still able to command commensurate a profit given the power shift from manufacturers to retailers. Chapter 4 extends the framework of Chapter 3 further and incorporates Nash bargaining theory, which enables researchers to calibrate the channel-wise bargaining power of manufacturers and retailers in more flexible manner.

Chapter 2

Understanding Consumer's Complex Brand Choice Behaviors with State Dependence

2.1 Introduction

Two behavioral patterns seemingly persisting across more than one purchase occasion have been extensively studied in marketing literature, namely inertia and variety-seeking. Such inter-temporal behavioral phenomena are often jointly referred to as state dependence.

Accounting for state dependence is important in formulating marketing strategy. Chintagunta (1998) states "knowing if consumers in a market are inertial or variety prone is very important in formulating marketing strategies (Chintagunta, 1998, 254)." He suggests that the existence of variety prone consumers motivates managers to expand a product line so that consumers switch to their own other product, and brand retention should be emphasized in an inertial market. Lattin & McAlister (1985) and Gupta et al. (1997) also implicitly suggest to have a variety of products as variety-seeking consumers would not be satisfied by a single product due to their desire to experience a wide range of features. On the other hand, Gupta et al. (1997) recommends encouraging the current choice of the brand through coupon program, a temporal price cut, and manufacturer advertising if consumers are inertial. Roy et al. (1996) claims that inducing trial of a brand should be encouraged if the strong inertia exists.

Knowing the joint effects of marketing mix variables and state dependence is also important in the following sense: The effects of marketing mix variables could be wrongly measured if state dependence is not accounted for. If a repeated purchase of a certain brand is due to inertia rather than promotional activities, the effect of promotion would be overestimated. If variety-seeking effect is ignored, on the other hand, heterogeneity across consumers could be exaggerated or undermined; if increased sales of a promoted brand in current period is partly due to variety-seeking consumers who purchased competing brand on the previous occasion, the effect of current promotion would be overestimated if variety-seeking effect is ignored when it exists. Researchers such as Guadagni & Little (1983), Gupta et al. (1997), and Seetharaman (2003) argue that an effect of promotion must account for the multi-period impact due to inertia; else the effect of promotion would be underestimated by predicting only a single period incremental sales.

2.2 Literature Review

There are many possible rationales of inertia. Jeuland (1979) and Guadagni & Little (1983) call it brand loyalty; Givon (1984) explains it as a risk avoidance behavior; Erdem (1996) calls it habit persistence, that is, a reinforcement of tastes or preferences by past behavior; and Keane (1997) counts habit persistence (use of a brand causes one to acquire a taste for that brand) and learning combined with risk aversion (use of a brand gives one knowledge about its attributes, making it a safe choice for a subsequent purchase occasion) as the explanation for inertia. Some researchers consider it as the consequence of monetary switching cost (e.g. the existence of repeat-purchase coupons and "frequent flyer" programs), transaction cost, and psychological switching cost such as learning cost associated with switching behavior or non-economic brand loyalty (Klemperer, 1987a,b; Farrell & Klemperer, 2007; Dubé et al., 2009).

To account for inertia, the loyalty variable suggested by Guadagni & Little (1983) is sometimes used as the effect of this variable has been shown to remain significant even after controlling the effects of other variables (Lattin, 1987; Keane, 1997). The other specifications include use of the last purchase indicator variable or the number of purchases of a brand by a consumer.

On the other hand, rationales of variety-seeking behavior include a satiation to brand attributes (McAlister, 1982; Lattin & McAlister, 1985), an intrinsic desire for a change (Givon, 1984), and the existence of a composite need, where the consumers' needs cannot be filled best by a single product (Lattin & McAlister, 1985).

The early research on this topic modeled variety-seeking behavior deterministically (e.g., McAlister (1982), Givon (1984), and Lattin & McAlister (1985)), whereas Trivedi et al. (1994) stochastically modeled the intensity of variety-seeking behavior as a random variable. Lattin & McAlister (1985) employed a perceptional difference on features shared between two brands (a previously purchased brand and a brand the consumer faces on the current purchase occasion) to model a satiation from brand attributes. Trivedi et al. (1994) constructed the similarity index for attributes of brands by a questionnaire asking perceived similarity among brands.

Meanwhile, Bawa (1990) suggested the "hybrid behavior," which assumes both inertia and variety-seeking behavior for the same consumer. The hybrid behavior hypothesized in his model was characterized by a consumer who exhibits an inertial behavior for a certain period of time and then switches to exhibit a variety-seeking behavior once a certain period of time passes. In other words, at least for some consumers, the model hypothesizes that the marginal utility of the same brand increases first but starts to decrease after the repeated consumptions of that brand, and brand switching occurs once the utility of that brand becomes lower than those of the other brands.[2] He justifies the hybrid behavior based on the psychological paper of Berlyne (1970) which finds that a hedonistic value, such as pleasantness, increases at first as a stimulus becomes more familiar but starts to decrease once the stimulus loses its novelty due to the repeated exposure.

Despite its unique attempt, there are several limitations in Bawa (1990). Besides he analyzed the only two-brand case, the model did not incorporate the effects of marketing variables and heterogeneity across consumers, which had been empirically found to affect consumers brand choice behavior. However, as Bawa (1990) attempted, a variety-seeking tendency may

[2]There are some papers which assume both inertia and variety-seeking behaviors, such as Lattin (1987), but these papers usually assume inertia for some consumers and variety-seeking behaviors for the other, and these behavioral tendencies are not allowed to change over time.

emerge in the course of repeated consumptions of the same brand and worth more detailed research. Although several papers (e.g., Givon (1984), Lattin (1987), and Seetharaman & Chintagunta (1998)) accommodate both inertia and variety-seeking, they did not allow the same consumer to switch its tendencies to seek or avoid variety over time. Therefore in this chapter, we try to capture such behavior while accounting for heterogeneity across consumers and the effect of marketing variables.

2.3 The Specification of State Dependence

We use the brand loyalty variable of Guadagni & Little (1983), which we will refer to it as "GL variable" henceforth, to express inertial part of the hybrid behavior in constructing the model. We will denote the GL variable at consumer i's purchase occasion t_i for brand $j = 1, \ldots, J$ by GL_{jt_i}. The initial value of the GL variable at $t_i = 1$ is given by

$$GL_{j1} = \begin{cases} \alpha & \text{if brand } j \text{ is the first purchase of consumer } i \\ (1-\alpha)/(J-1) & \text{otherwise} \end{cases} \tag{2.3.1}$$

and

$$GL_{jt_i} = \alpha \cdot GL_{j(t_i-1)} \cdot I_{ij(t_i-1)} \tag{2.3.2}$$

for $t_i = 2, \ldots, T_i$, where α is a parameter between 0 and 1, and $I_{ij(t_i-1)}$ is an indicator function taking unity if consumer i selects brand j on occasion $t_i - 1$. It should be noted that the GL variable is scaled so that their sum across brands is unity on each purchasing occasion; the GL variable is designed not to overwhelm the effect of other components of utility.[3]

[3] To show that the GL variable sums to unity, let us take an example of the GL variables on occasions $t_i = 1$ and $t_i = 2$. On occasion $t_i = 1$, it is obvious that they sum to unity

To capture the effect of variety-seeking, we include "run," the number of consecutive purchases of the same brand, defined in Bawa (1990) to the utility function. The purpose of including run is to "put a brake" on the GL variable, which keeps increasing as long as the same brand is kept being purchased. By including run, the utility for the same brand would start to decline as a result of the repeated consumptions of the same brand if run negatively affects utility. If a consumer has an inertial tendency, the coefficients of both the GL variable and run would significantly be non-negative and the variety-seeking behavior could be detected by the non-positive coefficient of the GL variable and/or run. Meanwhile, the hybrid behavior could be detected by the relative magnitudes of the positive coefficient of the GL variable and negative coefficient of the run. In the next section, we explain the model.

2.4 The Latent Class Model

In this study, we allow for the co-existence of consumers with different behavioral patterns by employing the latent class model. The latent class model is one of the general models to incorporate heterogeneity across consumers assuming a finite number of segments. The segment is a subset to which consumers belong, where members in the same segment are assumed to be homogeneous in intrinsic preferences to brands and responsiveness to the marketing variables. The idea behind this type of model is that there is an underlying multi-dimensional distribution of consumers' heterogeneity (i.e.,

---

because of the division by $J-1$ in (2.3.1). At $t_i = 2$, α is multiplied to all the GL variables consumer i has on occasion $t_i = 1$ by the term $\alpha \cdot GL_{jt_i-1}$ in (2.3.2) regardless of the brand chosen in the period and as a result they sum up to α. In addition, $(1-\alpha)$ is added to one of the GL variables by the second term on the right-hand side of the equation (2.3.2), and hence they together sum up to unity. The same is true for $t_i = 3, 4, \ldots$.

intrinsic preferences for brands and relative responsiveness to the marketing variables) which characterizes their behavior, and the latent class model assumes discrete underlying distribution. The overall choice probabilities of brands are given by the weighted sum of segment-level choice probabilities in this model (Bucklin et al., 1998). In other words, each of the unconditional choice probabilities for brands "can be decomposed into a weighted average of underlying (or "latent") choice probabilities (Kamakura & Russell, 1989, 380)." Because the finite representation of consumer's characteristics of the latent class model coincides well with the concept of a segment, the model is widely employed in the marketing literature.

2.4.1 Specification of the Model

We define the utility of consumer $i = 1, \ldots, N$ for brand $j = 1, \ldots, J$ on occasion $t_i = 1, \ldots, T_i$ as

$$U_{ijt_i} = \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s + \epsilon_{ijt_i}$$

where $\boldsymbol{x}_{ijt_i}$ is a $1 \times R$ vector of the explanatory variables a consumer i faces on occasion t_i, which consists of a set of the dummy variables for brands except for a base brand, the shelf price of brand j, a dummy variable for coupon usage times a coupon face value, the dummy variables for feature and display, the GL variable, and run. The $\boldsymbol{\beta}_s$ is a corresponding $R \times 1$ vector of parameters for segment $s = 1, \ldots, S$. The random error term ϵ_{ijt_i} captures the unobserved part of the utility which is assumed to follow independently, identically distributed ("i.i.d." henceforth) Gumbel distribution.

The relative sizes of segment s is defined as λ_s such that

$$0 < \lambda_s \leq 1$$

for all s and

$$\sum_{s=1}^{S} \lambda_s = 1. \tag{2.4.1}$$

Each consumer has different membership probabilities for each segment because membership probabilities are estimated from each consumer's purchase history. Accordingly, the term λ_s can be viewed as the "likelihood of finding a household in segment s (Kamakura & Russell, 1989, 380)" in the sample.

For consumer i, let y_{ijt_i} be entries of $T_i \times J$ matrix $\boldsymbol{Y}_i$

$$\boldsymbol{Y}_i = \begin{pmatrix} y_{i11} & \cdots & y_{iJ1} \\ \vdots & \ddots & \vdots \\ y_{i1T_i} & \cdots & y_{iJT_i} \end{pmatrix} \tag{2.4.2}$$

and denote each row as $\boldsymbol{y}_{it_i}$. Since we assume ϵ_{ijt_i} follow i.i.d. Gumbel distribution, we can express the probability that consumer i in segment s chooses brand j at the occasion t_i in the standard logit form as

$$\Pr\{(y_{i1t_i}, \ldots, y_{iJt_i}) = (\underbrace{0, \ldots, 0}_{j-1}, 1, \underbrace{0, \ldots, 0}_{J-j}) | S_i = s; \boldsymbol{\beta}_s\} = \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}_s)} \tag{2.4.3}$$

where the random variable S_i indicates the segment consumer i belongs to, assuming we could observe the segment membership of consumer i. We abbreviate (2.4.3) as

$$\Pr(Y_{it_i} = j | S_i = s; \boldsymbol{\beta}_s) = \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}_s)} \tag{2.4.4}$$

henceforth for notational convenience.

The unconditional choice probability of a randomly selected consumer i for brand j can be obtained by integrating out the equation (2.4.3) by the density in the population λ_s as[4]

$$\Pr(Y_{it_i} = j) = \int \Pr(Y_{it_i} = j | S_i = s; \boldsymbol{\beta}_s)\lambda_s ds. \tag{2.4.5}$$

Since the relative size of the segment λ_s is discrete, (2.4.5) is written as

$$\Pr(Y_{it_i} = j) = \sum_{s=1}^{S} \lambda_s \cdot \Pr(Y_{it_i} = j | S_i = s; \boldsymbol{\beta}_s).$$

This is a weighted average of logit formula evaluated at each mass point (segment), as pointed out by Kamakura & Russell (1989).

Suppose that consumer i has the choice history defined as $H_i = (Y_{i1}, \ldots, Y_{iT_i})$, where element Y_{it_i} indicates the brand purchased at occasion t_i. Then the conditional choice probability that consumer i has the choice history H_i given that the consumer belongs to segment s is written as

$$\Pr(H_i | S_i = s; \boldsymbol{\beta}_s) = \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} \left\{\Pr(Y_{it_i} = j | S_i = s; \boldsymbol{\beta}_s)\right\}^{y_{ijt_i}}. \tag{2.4.6}$$

The unconditional probability of randomly selected consumer i having the choice history H_i can be written as

$$\Pr(H_i; \boldsymbol{\beta}) = \sum_{s=1}^{S} \lambda_s \cdot \Pr(H_i | S_i = s; \boldsymbol{\beta}_s) \tag{2.4.7}$$

where $\boldsymbol{\beta}$ is defined as $R \times S$ parameter matrix

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_S) = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1s} & \cdots & \beta_{1S} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \beta_{r1} & \cdots & \beta_{rs} & \cdots & \beta_{rS} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \beta_{R1} & \cdots & \beta_{Rs} & \cdots & \beta_{RS} \end{pmatrix}.$$

[4]The model of the form (2.4.5) is sometimes called mixed logit model and λ_s is called mixing distribution. The latent class model can be regarded as the special case of mixed logit model where mixing distribution is discrete (Train, 2003). See Appendix A.3 for detail.

Let us further define, for each consumer i, the multinomial indicator random variable $z_i(s)$ which takes one if consumer i belongs to segment s and 0 otherwise, assuming we know the membership probability of consumer i belonging to segment s given its purchase history H_i denoted as $\Pr(S_i = s|H_i; \boldsymbol{\beta}_s)$. Then this membership indicator random variables $z_i(s)$'s are entries of $N \times S$ matrix $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_N \end{pmatrix} = \begin{pmatrix} z_1(1) & \dots & z_1(S) \\ \vdots & \ddots & \vdots \\ z_N(1) & \dots & z_N(S) \end{pmatrix}.$$

The row sums of the matrix $\boldsymbol{Z}$ above are all unity. Assuming we were able to observe $\boldsymbol{Z}$, the likelihood given the choice histories of all consumers is written as[5]

$$L(\boldsymbol{\lambda}, \boldsymbol{\beta}|\boldsymbol{H}, \boldsymbol{Z}) = \prod_{i=1}^{N}\prod_{s=1}^{S}\{\lambda_s \cdot \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)\}^{z_i(s)}$$

where $\boldsymbol{H} = (H_1, \dots, H_i, \dots, H_N)$ is the choice history of all consumers in the sample and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_S)$ is $1 \times S$ vector of relative sizes of segments. Accordingly, the log likelihood is

$$l(\boldsymbol{\lambda}, \boldsymbol{\beta}|\boldsymbol{H}, \boldsymbol{Z}) = \sum_{i=1}^{N}\sum_{s=1}^{S} z_i(s) \cdot \ln \Pr(H_i|S_i = s; \boldsymbol{\beta}_s) + \sum_{i=1}^{N}\sum_{s=1}^{S} z_i(s) \cdot \ln \lambda_s. \tag{2.4.8}$$

If we were able to observe $\boldsymbol{Z}$, the algorithm to estimate parameters $(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is as follows:

Step 0.1: Set $t = 0$. Set the initial values $\widehat{\boldsymbol{\beta}}_s^{(0)}$ for $s = 1, \dots, S$ and set

[5] The term $\lambda_s \cdot \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)$ is the joint probability that consumer i belongs to segment s and has choice history H_i. Note, however, that the relative size of segment λ_s is unknown and has to be estimated.

$\lambda_s^{(0)} = 1/S$ for $s = 1, \ldots, S$.

Step 0.2: Calculate $l^{(t)}\left(\boldsymbol{\lambda}^{(t)}, \widehat{\boldsymbol{\beta}}^{(t)}|\mathbf{H},\mathbf{Z}\right)$ using (2.4.8).

Step 1: Calculate $\lambda_s^{(t+1)}$ for $s = 1, \ldots, S$ from the method which will be explained below.

Step 2: Estimate $\widehat{\boldsymbol{\beta}}_s^{(t+1)}$ for $s = 1, \ldots, S$ using the scoring or Newton-Raphson method.[6]

Step 3: Calculate $l^{(t+1)}\left(\boldsymbol{\lambda}^{(t+1)}, \widehat{\boldsymbol{\beta}}^{(t+1)}|\mathbf{H},\mathbf{Z}\right)$ using (2.4.8). If $l^{(t+1)}\left(\boldsymbol{\lambda}^{(t+1)}, \widehat{\boldsymbol{\beta}}_s^{(t+1)}|\mathbf{H},\mathbf{Z}\right)$ and $l^{(t)}\left(\boldsymbol{\lambda}^{(t)}, \widehat{\boldsymbol{\beta}}_s^{(t)}|\mathbf{H},\mathbf{Z}\right)$ are close enough, stop the iteration as the likelihood is maximized. Else set $t = t + 1$ and goto **Step 1**.

Because we cannot possibly obtain $\boldsymbol{Z}$, we employ EM algorithm to estimate $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ as explained in the following subsection.

2.4.2 EM algorithm

If the segment memberships of consumers $\mathbf{Z}$ were completely known, the vector of parameters $\boldsymbol{\beta}_s$ can be estimated by the algorithm described above using well-known methods such as Newton-Raphson method. EM algorithm takes advantage of this fact and in the algorithm, consumer's membership to the segment $z_i(s)$ is first assumed to be missing and is imputed by its "expectation." Then the conditional likelihood is "maximized" based on the expected values of membership to segments. The consumer's expected membership is then updated using the updated likelihood. This cycle of "expectation"

[6]See Appendix A.6 for detail.

of membership to the segment and "maximization" of likelihood is repeated until the likelihood converges.

Taking the expectation with respect to $z_i(s)$ for the log likelihood (2.4.8), we have

$$E[l(\boldsymbol{\lambda}, \boldsymbol{\beta}|\boldsymbol{H}, \boldsymbol{Z})] = \sum_{i=1}^{N}\sum_{s=1}^{S} h_i(s) \cdot \ln \Pr(H_i|S_i = s; \boldsymbol{\beta}_s) + \sum_{i=1}^{N}\sum_{s=1}^{S} h_i(s) \cdot \ln \lambda_s \tag{2.4.9}$$

where

$$h_i(s) = E[z_i(s)] = \sum_{l=1}^{S} z_i(l) \cdot \Pr(S_i = l|H_i; \boldsymbol{\beta}_l) = \Pr(S_i = s|H_i; \boldsymbol{\beta}_s) \tag{2.4.10}$$

is the expected values of the indicator random variable $z_i(s)$. Since parameter $\boldsymbol{\beta}$ only appears in the first term and $\boldsymbol{\lambda}$ only appears in the second term on the right-hand side of the equation (2.4.9), they can be estimated by maximizing $E[l(\boldsymbol{\lambda}, \boldsymbol{\beta}|\boldsymbol{H}, \boldsymbol{Z})]$ alternately.

Let us first look at the second term on the right-hand side of the equation (2.4.9). Since we have the condition $\sum_{s=1}^{S} \lambda_s = 1$ from (2.4.1), the second term can be maximized by the method of Lagrange multipliers given $\boldsymbol{\beta}_s$. Let us define

$$L = \sum_{i=1}^{N}\sum_{s=1}^{S} h_i(s) \cdot \ln \lambda_s - \lambda \left\{ \sum_{s=1}^{S} \lambda_s - 1 \right\}.$$

Then we have $(S + 1)$ set of equations by partially differentiating L with respect to λ_s for $s = 1, \ldots, S$ and λ. Setting resulting formulas zero as

$$\left\{ \begin{array}{rclcl} \frac{\partial L}{\partial \lambda_1} & = & \frac{\sum_{i=1}^{N} h_i(1)}{\lambda_1} - \lambda & = & 0 \\ & \vdots & & & \\ \frac{\partial L}{\partial \lambda_S} & = & \frac{\sum_{i=1}^{N} h_i(S)}{\lambda_S} - \lambda & = & 0 \\ \frac{\partial L}{\partial \lambda} & = & -\sum_{s=1}^{S} \lambda_s + 1 & = & 0, \end{array} \right. \tag{2.4.11}$$

we have

$$\lambda_s = \frac{1}{\lambda}\sum_{i=1}^{N} h_i(s) \tag{2.4.12}$$

for $s = 1, \ldots, S$ from the first S equations in (2.4.11). Substitute these equations to the last equation in (2.4.11), we obtain

$$\frac{1}{\lambda}\sum_{i=1}^{N} h_i(1) + \cdots + \frac{1}{\lambda}\sum_{i=1}^{N} h_i(S) = 1$$

or

$$N = \lambda,$$

since $h_i(1) + \cdots + h_i(S) = 1$. Then we have

$$\lambda_s = \frac{\sum_{i=1}^{N} h_i(s)}{N} \tag{2.4.13}$$

for $s = 1, \ldots, S$ from (2.4.12). The solution (2.4.13) means that the relative size of segment s is the average of segment membership for s across all consumers in the sample. The term $h_i(s) = \Pr(S_i = s|H_i; \boldsymbol{\beta}_s)$ in (2.4.10) can be calculated using the definition of conditional probability as[7]

$$h_i(s) = \frac{\Pr(S_i = s, H_i; \boldsymbol{\beta}_s)}{\Pr(H_i; \boldsymbol{\beta})} = \frac{\lambda_s \cdot \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)}{\sum_{s=1}^{S} \lambda_s \cdot \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)}. \tag{2.4.14}$$

We obtain λ_s by substituting (2.4.14) to (2.4.13).

The parameters of the right-hand side of the equation (2.4.9) for segment s can be estimated independently for each segment since the vectors of parameters $\boldsymbol{\beta}_s$ are independent across segments. The first term on the right-hand side of the equation (2.4.9) for segment s is written with the notation

[7] Note that $h_i(s)$ in (2.4.14) can be interpreted as the posterior distribution of consumer i's membership probability for segment s with prior distribution λ_s and likelihood H_i given segment membership $S_i = s$ as we mentioned earlier.

similar to (2.4.6) as

$$\sum_{i=1}^{N} h_i(s) \cdot \ln \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)$$
$$= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \{h_i(s) \cdot y_{ijt_i} \cdot \ln \Pr(Y_{it_i} = j|S_i = s; \boldsymbol{\beta}_s)\}. \quad (2.4.15)$$

EM algorithm

Step 0.1: Set $t = 0$. Set the initial values $\widehat{\boldsymbol{\beta}}_s^{(0)}$ for $s = 1, \ldots, S$ and set $\lambda_s^{(0)} = 1/S$ for $s = 1, \ldots, S$.

Step 0.2: Set $s = 1$. Obtain $h_i^{(t)}(s)$ for $i = 1, \ldots, N$ by calculating $\Pr(Y_{it_i} = j|S_i = s; \boldsymbol{\beta}_s)$ using (2.4.3) first then (2.4.6) and (2.4.7) successively with $\widehat{\boldsymbol{\beta}}_s^{(t)}$ and $\lambda_s^{(t)}$ and substitute these interim results to (2.4.14). Set $s = s + 1$ and repeat **Step 0.2** until $s = S$.

Step 0.3: Calculate $E\left[l^{(t)}\left(\boldsymbol{\lambda}^{(t)}, \widehat{\boldsymbol{\beta}}^{(t)}|\mathbf{H}, \mathbf{Z}\right)\right]$ using (2.4.9).

Step 1: Update $\lambda_s^{(t+1)}$ from (2.4.13) using $h_i^{(t)}(s)$.

Step 2: Estimate $\widehat{\boldsymbol{\beta}}_s^{(t+1)}$ by maximizing (2.4.15) with (2.4.3) and $h_i^{(t)}(s)$ obtained previously. The actual maximization is done by the Newton-Raphson method or its variant.

Step 3: Update $\Pr(Y_{it_i} = j|S_i = s; \boldsymbol{\beta}_s)^{(t+1)}$ by substituting $\widehat{\boldsymbol{\beta}}_s^{(t+1)}$ to (2.4.4) obtained in **Step 2**.

Step 4: Calculate $h_i^{(t+1)}(s)$ from (2.4.14) with updated $\widehat{\boldsymbol{\beta}}_s^{(t+1)}$ and $\lambda_s^{(t+1)}$

for $i = 1, \ldots, N$. Set $s = s + 1$ and goto **Step 1**. If $s = S$, goto **Step 5**.

Step 5: Calculate $E\left[l^{(t+1)}\left(\boldsymbol{\lambda}^{(t+1)}, \widehat{\boldsymbol{\beta}}^{(t+1)}|\mathbf{H}, \mathbf{Z}\right)\right]$ using (2.4.9). If $E\left[l^{(t+1)}\left(\boldsymbol{\lambda}^{(t+1)}, \widehat{\boldsymbol{\beta}}_s^{(t+1)}|\mathbf{H}, \mathbf{Z}\right)\right]$ and $E\left[l^{(t)}\left(\boldsymbol{\lambda}^{(t)}, \widehat{\boldsymbol{\beta}}_s^{(t)}|\mathbf{H}, \mathbf{Z}\right)\right]$ are close enough, stop the iteration as the expected log likelihood is maximized. Else set $s = 1$ and $t = t + 1$, and return to **Step 1**.

2.5 Data

We use ERIM database, the panel data of U.S. consumers in Sioux Falls, SD which was collected from 1st week of 1986 to 34th week of 1988. ERIM database is the data collected by the now-defunct ERIM division of A.C. Nielsen on panels of consumers in Sioux Falls and Springfield for academic research.[8]

We chose a ketchup category for our empirical analysis for the following reasons. First, because we were interested in consumer's brand choice behavior with the possible presence of state dependence, product categories in which a consumer exhibited a strong genuine preference to a specific brand were not suitable because a consumer would choose the specific brand anyway. Secondly, the products that were purchased with relatively high frequency were preferable, since we would incorporate the effect of past brand purchases on the current purchasing occasion.

Because there were more than forty Stock Keeping Units (SKUs) in the original panel data, we used the following criteria to select SKUs for our analysis. First, we dropped the SKUs whose market shares were less than

[8]We acknowledge the James M. Kilts Center, University of Chicago Booth School of Business for letting us use the data.

1% because some consumers in these data sets were not able to choose them. This left fourteen SKUs. Next, we chose the SKUs whose sizes were either 32 or 28 ounces, which seemed to be the standard sizes of ketchup judged by their market shares; they accounted for 81.6% of the market shares. The other sizes included 14, 40, 44 and 64 ounce, but those who bought the ketchup of these sizes might have different demographic characteristics and thus may have different purchasing patterns from those who bought the standard sized ketchup. This left eight SKUs and there were 516 consumers who chose ketchup from these eight SKUs with 3,933 purchase records.

Next, we checked how many stores carried all these SKUs, because if consumers bought ketchup from the other stores than those carrying all SKUs, their SKU selections could have been influenced by the lack of selection. There were fifteen stores in Sioux Falls but only five of them carried all eight SKUs.[9] If we removed the consumers that bought at least one ketchup in stores other than these five, only 120 consumers with 497 purchase records would be left for the analysis. The large reduction of data was because the sixth, seventh, and eighth selling SKUs were simultaneously available only in few stores in Sioux Falls. Hence we chose to retain only top-selling five SKUs. Among fifteen stores, twelve of them carried all top-selling five SKUs. After eliminating the consumers who purchased ketchup in stores other than these twelve, 255 consumers with 1,791 purchase records remained.

Finally, since we were interested in the consumer's brand choice behavior across time, we chose to retain the consumers who made more than or equal to five purchases of ketchup during the period, which left 137 consumers with 1,504 purchase records. After screening data, we collected consumer ID,

[9] We assumed that the store carried the SKU if at least one purchase record of the SKU was found in that store during the data collection period.

SKU purchased, its shelf price, coupon values (when used), store ID, date of purchase, an indicator variable whether it was displayed, and an indicator variable whether it was featured on each purchasing occasion. Table 2.5.1 presents the summary statistics of the five SKUs analyzed in this study.

Table 2.5.1: Summary statistics of the SKUs

SKU	Market Share	Mean Price per oz.	Mean Value of Coupons	Coupon Usage	Display	Feature
Heinz 32 oz.	31.7%	3.37	1.24	37.9%	11.5%	43.8%
Heinz PLS 28 oz.	15.8%	4.38	2.41	33.1%	16.7%	34.6%
Hunt's PLS & GLS 32 oz.	14.3%	3.22	1.30	32.6%	11.9%	36.7%
Del Monte 32 oz.	6.4%	2.87	1.00	7.2%	11.2%	36.0%
Control 32 oz.	5.0%	2.65	1.64	3.8%	5.7%	24.5%

2.6 Empirical Results

We constructed and tested two other models to calibrate the validity of our proposal model. Model 1 only used marketing variables as explanatory variables, Model 2 incorporated GL variable along with the marketing variables, and Model 3 is our proposal model which incorporated GL variable and run in addition to marketing variables.

We determined the number of segments based on Akaike Information Criteria (AIC).[10] The number of segments was chosen to be four because no significant reduction in AIC was observed for Model 3 when the number of segments was increased from four to five as presented in Table 2.6.1. The estimated parameters of Model 3 are presented in Table 2.6.2. The coefficients

[10]We also used BIC criteria but it yielded the similar result.

Table 2.6.1: Akaike Information Criteria of the three models

	Model 1	Model 2	Model 3
2 segments	1310.6	1058.2	1041.8
3 segments	948.5	861.3	835.6
4 segments	842.6	805.6	796.7
5 segments	819.1	796.5	794.2
6 segments	809.5	803.8	813.0

of SKUs indicate intrinsic preferences for them with respect to "Control 32 oz." which we chose as the base SKU.

All coefficients of Model 3 are consistent with the expected economic behavior; coefficients of prices are negative; those of coupon, display, and feature are all positive in all segments; and the intrinsic preferences to SKUs and responsiveness to marketing variables differ significantly across segments.

To reproduce the dynamic behavioral patterns regulated by the model, we calculated the purchasing probabilities for each SKU and segment, assuming a consumer makes five consecutive purchases of the same SKU. Table 2.6.3 presents the results. For example, the number at rows $t = 3$ is the purchasing probability of SKU given two consecutive purchases of that SKU. In the calculation, we used the average prices and assumed no promotion took place during the period.

Table 2.6.3 shows that, consumers in segments 1 and 4 exhibit strong inertia while those in segments 2 and 3 exhibit weak and modest variety-seeking tendencies respectively. While consumers in segment 2 can be characterized by its strong preferences to Heinz products, those in segment 3 are least price sensitive and have relatively low coefficients for coupons and features. Consumers in segment 4 are most price sensitive and they respond to promotions most.

Table 2.6.2: Parameter estimates of the proposed model

	Segment 1	Segment 2	Segment 3	Segment 4
Heinz 32 oz.	-0.08	4.28	1.89	1.44
	(0.015)	(0.018)	(0.012)	(0.016)
Heinz PLS 28 oz.	1.31	2.66	1.73	2.40
	(0.007)	(0.007)	(0.007)	(0.007)
Hunt's PLS & GLS 32 oz.	0.03	0.30	2.71	-2.53
	(0.007)	(0.005)	(0.010)	(0.006)
Del Monte 32 oz.	1.18	-2.14	1.07	-1.15
	(0.007)	(0.001)	(0.004)	(0.006)
Price	-0.85	-0.74	-0.67	-2.51
	(0.070)	(0.077)	(0.068)	(0.071)
Coupon	2.79	5.02	3.29	5.47
	(0.021)	(0.021)	(0.018)	(0.020)
Display	3.42	3.90	3.86	4.75
	(0.007)	(0.007)	(0.007)	(0.007)
Feature	5.62	2.50	2.94	5.89
	(0.013)	(0.012)	(0.012)	(0.013)
GL variable	4.30	0.74	1.74	5.61
	(0.014)	(0.016)	(0.010)	(0.013)
Run	0.51	*-0.13	-0.22	-0.12
	(0.089)	(0.066)	(0.052)	(0.053)
Size of Segments	26.1%	23.4%	28.9%	21.7%
Total Log Likelihood	-365.8			

*90% level significance. All the other coefficients were significant at 95% level. The numbers in parentheses are standard errors.

Table 2.6.3: The dynamic purchase probabilities of the SKUs

Heinz 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	8.8%	87.6%	20.6%	33.5%
t=1	56.1%	91.7%	42.4%	93.6%
t=2	74.9%	91.1%	40.4%	95.3%
t=3	86.7%	90.5%	37.7%	96.2%
t=4	93.1%	89.7%	34.7%	96.8%
t=5	96.4%	88.8%	31.3%	97.1%

Heinz PLS 28 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	15.0%	8.3%	9.0%	7.1%
t=1	70.1%	12.4%	21.9%	68.7%
t=2	84.5%	11.7%	20.5%	75.3%
t=3	92.3%	10.9%	18.8%	79.4%
t=4	96.1%	10.0%	16.8%	81.9%
t=5	98.0%	9.2%	14.8%	83.5%

Hunt's PLS & GLS 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	11.2%	1.8%	52.5%	0.9%
t=1	62.4%	2.8%	75.8%	21.6%
t=2	79.5%	2.7%	74.2%	27.6%
t=3	89.4%	2.5%	72.0%	32.6%
t=4	94.6%	2.3%	69.3%	36.3%
t=5	97.2%	2.0%	66.0%	38.8%

Del Monte 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	47.4%	0.2%	12.8%	9.0%
t=1	92.2%	0.3%	29.4%	74.0%
t=2	96.5%	0.3%	27.7%	79.8%
t=3	98.4%	0.3%	25.5%	83.3%
t=4	99.2%	0.3%	23.1%	85.5%
t=5	99.6%	0.2%	20.5%	86.7%

Control 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	17.6%	2.1%	5.1%	49.5%
t=1	73.9%	3.2%	13.2%	96.6%
t=2	86.9%	3.0%	12.3%	97.5%
t=3	93.5%	2.8%	11.1%	98.0%
t=4	96.8%	2.5%	9.9%	98.3%
t=5	98.3%	2.3%	8.6%	98.5%

2.7 Discussions of the Results

Our proposal model fits the data best compared to the competing models with a fair number of significant variables, indicating that segments are heterogeneous in behavioral patterns over time. These results suggest important implications for marketers. Specifically, the information in Table 2.6.3 can be used as a reference for brand managers to plan their marketing strategies and promotion. For example, since consumers in segment 1 exhibit strong inertia, Del Monte may not need much promotion. When competitor promotes, however, it also may need to promote the brand to retain customers. Since consumers in this segment have low coefficients for coupon and display but have a high coefficient for a feature, Del Monte should use features when it promotes the brand. Consumers in segment 2 exhibit a strong preference for Heinz brands, and there seems to be little chance for the other brands to be selected. From Heinz perspective, promotion for this segment is not necessary since consumers in this segment would purchase its brands anyway. Consumers in segment 3 are the main target for Hunt's. However, because consumers in this segment are the least price sensitive, it may restructure its current promotional planning especially because it uses a lot of coupons. Consumers in segment 4 also exhibit strong inertia but they are most price sensitive and respond to promotions most. The rigorous price competition between Heinz and Control can be expected for this segment. While Heinz should promote its Heinz 32 ounce to segment 4, Hunt's and Del Monte are better off to spend their promotional budgets on segment 1 or 3, since they have little chance to attract consumers in segment 4.

2.8 Conclusion

In this study, we developed the comprehensive model which can accommodate inertia, variety-seeking, and hybrid behavior along with the heterogeneous preferences to brands and sensitivities to marketing variables and empirically tested the model using the panel data of ketchup. Though the hybrid behavior was minimal in our data, this study shed a light on the possible presence of the hybrid behavior and provides plenty of marketing insights of practical importance.

For future studies, the proposed model can be tested using different data sets for the validity of the model. Moreover, analyzing the competitive actions/reactions to promotion and pricing strategies incorporating state dependence would be an opportunity for future research.

Chapter 3

Inference on Strategic Interactions among Manufacturers, Retailers, and Consumers

3.1 Introduction

In consumer packaged good market, manufacturers and retailers strategically interact. The approach called a structural market equilibrium model is sometimes employed to model such interaction. This approach describes the interaction of manufacturers, retailers, and consumers imposing their optimizing behavioral assumptions; manufacturers and retailer are assumed to maximize their own profits and consumers are assumed to maximize their utilities. Examples of such papers are Sudhir (2001), Yang et al. (2003), Villas-Boas & Zhao (2005), and Che et al. (2007) to name a few.

A structural market equilibrium model allows the variety of competitive structure of the market via different combinations of inter-firms interactions. The two main descriptive factors of the market structure in previous studies are horizontal strategic interactions among manufacturers and vertical strategic interaction among manufacturers and retailers; they are Bertrand competition/tacit collusion and manufacturer Stackelberg/vertical Nash respectively. Bertrand competition refers to own-brands profits maximizing behavior of manufacturers and tacit collusion refers to the behavior of manufacturers which collectively maximizes total profits from all brands in the market. Manufacturer Stackelberg game assumes that manufacturers act as Stackelberg leaders with respect to retailers and choose their wholesale prices anticipating a reaction of retailers, conditional on the wholesale prices of competing brands. Vertical Nash game, on the other hand, assumes that manufacturers and retailers move simultaneously; manufacturers choose prices taking retail prices of competing brands and retail margin of their brands as given. In either case, the retailer chooses retail prices to maximize profits taking wholesale prices as given (Choi, 1991; Sudhir, 2001).

The model is widely used in the literature as it offers rich insights of the market and provides an empirical method to test theories. Choi (1991) argues that whether market structure is characterized by manufacturer Stackelberg or vertical Nash depends on the concentration of the market (i.e., whether a market is governed by a few large firms or bunch of small firms). Similarly, Sudhir (2001) argues that firms in long-term competition can achieve tacit collusion partly because it is easier to employ a punishment strategy[11] in the concentrated market where a small number of manufacturers have majority of market share.

[11]It is the strategy in the game theory where a firm penalizes a competing firm by initiating a rigorous price competition once a competing firm deviates from collusive behavior.

The findings in this area are mixed. On the one hand, Nevo (2001) and Che et al. (2007), which analyze cereal market and Villas-Boas & Zhao (2005), which analyzes ketchup market support Bertrand competition in the respective markets. On the other hand, Sudhir (2001) finds cooperative behavior in yogurt and peanut butter market where two leading brands have a majority of market shares (82% and 66% respectively). On the vertical relationship between a retailer and manufacturers, Sudhir (2001) and Che et al. (2007) compare manufacturer Stackelberg and vertical Nash game to find manufacturer Stackelberg outperforms vertical Nash game.[12]

What has been missing in the literature is the retailer Stackelberg formulation whereby the retailer has control over pricing with respect to manufacturers. Given the purported power shift from manufacturers to a retailer, this game has to be considered along with manufacturer Stackelberg and vertical Nash. Therefore in this research, we extend Che et al. (2007) by mathematically formulating retailer Stackelberg and conduct an economic analysis taking Japanese yogurt market as an example to investigate whether manufacturers' effort to develop special featured brands is still rewarding in terms of margins. This is of interest of manufacturers as it is the conventional wisdom that the power in the distribution channel has shifted from manufacturers to retailers and manufacturer's effort may only benefit retailers. To the best of our knowledge, this is the first instance of a retailer Stackelberg formulation in the context of discrete choice model.[13] This study is hence unique in that it successfully portrays the symmetrical relationship

[12]The other empirical papers presume one game as vertical interaction. For example, Nevo (2001) presumes only vertical Nash game whereas Yang et al. (2003) and Villas-Boas & Zhao (2005) only presume manufacturer Stackelberg game.

[13]Choi (1991) introduces retailer Stackelberg formulation but the model in that paper assumes linear demand function.

between manufacturer Stackelberg and retailer Stackelberg games, whereby the vertical Nash game is located in the midpoint of those games.

The rest of this chapter is organized as follows. The next section describes the model. We briefly explain our data in section 3.4. We will present results for data analysis in section 3.5. We discuss the results and concludes this chapter in section 3.6.

3.2 The model

In this section, we explain the model. Our model specification largely follows that of Sudhir (2001), Villas-Boas & Zhao (2005), and Che et al. (2007).

3.2.1 Demand-Side Specification

We employ the multinomial logit model for consumer brand choice behavior. Specifically, the utility of consumer i choosing brand j at time t_i is defined as v_{ijt_i} and written as[14]

$$v_{ijt_i} = \boldsymbol{x}_{jt_i} \cdot \boldsymbol{\beta}_s - \alpha_s \cdot p_{jt_i} + \mathrm{sim}_{kj} \cdot SD_s + \xi_{jt_i}$$

where $\boldsymbol{x}_{jt_i}$ (a subset of $\boldsymbol{x}_{jt}$ defined for all t and j) is the set of explanatory variables including brand dummy variables; p_{jt_i} is the price; sim_{kj} is the attribute similarity index of brand j with respect to the previously purchased brand k; and ξ_{jt_i} is the unobserved demand characteristics which can be observed by firms and consumers but not by a researcher. Then the choice probability of consumer i for brand j at occasion t_i is written as

$$\mathrm{Pr}_{ijt_i} = \frac{\exp(v_{ijt_i})}{1 + \sum_{k=1}^{J} \exp(v_{ikt_i})} \tag{3.2.1}$$

[14] The term ξ_{jt_i} is subset of ξ_{jt} where the latter is defined for all calendar dates and brands in the panel, and the former is retrieved from ξ_{jt}.

where the addition of 1 in the denominator in (3.2.1) stands for the outside option which is a consequence of the specification $v_{i0t_i} = \epsilon_{i0t_i}$.

We assume following properties for ξ_{jt}:

$$E[\xi_{jt}] = 0 \tag{3.2.2}$$

$$\text{Cov}[\xi_{jt}, \mathbf{X}_{jkt}] = 0 \tag{3.2.3}$$

$$E[\xi_{jt}^2 | \mathbf{X}_{jkt}] < \infty \tag{3.2.4}$$

where $\mathbf{X}_{jkt} \equiv (\boldsymbol{x}_{jt}, \text{sim}_{kj})$. The coefficients $\boldsymbol{\beta}_s$, α_s, and SD_s are parameters to be estimated, where subscript s corresponds to segment s in the latent class model which we employed in Chapter 2.

The attribute similarity index

We use the attribute similarity index to capture the state dependence in consumer brand choices following Che et al. (2007).[15] The similarity between the brand purchased on the previous occasion (brand k) and the brand a consumer faces on the current purchase occasion (brand j) is specified as

$$\text{sim}_{kj} = \frac{I_{kj} + \sum_{l=1}^{L} I_{kjl} \cdot r_l}{1 + \sum_{l=1}^{L} r_l}$$

where I_{kj} is an indicator variable taking unity if $k = j$, I_{kjl} is an indicator variable taking unity if two brands share the same level of attribute $l = 1, \cdots, L$, and $r_l > 0$ is importance weight to be estimated. The similarity index is designed to take value between 0 (brands are totally dissimilar)

[15] The idea of the attribute similarity index can be found in previous papers (e.g., Lattin (1987)), but the specification in previous literature requires a questionnaire which explicitly asks subjects for the perceived similarity between brands. The advantage of the specification of Che et al. (2007) is that it does not require such information and similarities between brands can be calibrated from the data, although the level of attributes shared by brands must be arbitrary set by researchers.

and 1 (brands are identical). The coefficient of the similarity index, SD_s, can either be positive or negative which corresponds to inertial and variety-seeking behavior of consumers respectively. Following Che et al. (2007), we parametrize SD_s by the demographic variables as

$$SD_s = \gamma_{s0} + \mathbf{DEMO}_i \cdot \boldsymbol{\gamma}_s$$

where $\mathbf{DEMO}_i = (D_{i1}, \ldots, D_{iQ})$ is vector of demographic characteristics of consumer i, and γ_{s0} and $\boldsymbol{\gamma}_s = (\gamma_{s1}, \ldots, \gamma_{sQ})^T$ are corresponding parameters. Let R be $R \equiv 1 + \sum_{l=1}^{L} r_l$. Then the term $\mathrm{sim}_{kj} \cdot SD_s$ can be written out as

$$\begin{aligned}
\mathrm{sim}_{kj} \cdot SD_s \quad = \quad & \frac{\gamma_{s0}}{R} \cdot I_{kj} + \frac{\gamma_{s0} \cdot r_1}{R} \cdot I_{kj1} + \cdots + \frac{\gamma_{s0} \cdot r_l}{R} \cdot I_{kjL} \\
& + \frac{\gamma_{s1}}{R} \cdot I_{kj} \cdot D_{i1} + \frac{\gamma_{s1} \cdot r_1}{R} \cdot I_{kj1} \cdot D_{i1} + \cdots + \frac{\gamma_{s1} \cdot r_l}{R} \cdot I_{kjL} \cdot D_{i1} + \ldots \\
& \quad + \frac{\gamma_{sQ}}{R} \cdot I_{kj} \cdot D_{iQ} + \frac{\gamma_{sQ} \cdot r_1}{R} \cdot I_{kj1} \cdot D_{iQ} + \cdots + \frac{\gamma_{sQ} \cdot r_l}{R} \cdot I_{kjL} \cdot D_{iQ}.
\end{aligned}$$

In estimation, we treat multiplicative terms of unknown parameters such as γ_{s0}/R as a single parameter and estimate γ_{s0}, $\boldsymbol{\gamma}_s$ and r_l for $l = 1, \ldots, L$ by least squares. We illustrate an example in Appendix A.9 to show how these parameters can be estimated.

The price endogeneity problem

Since ξ_{jt} could be correlated with price and might result in biased estimation (Berry, 1994; Besanko et al., 2003; Nevo, 2001; Villas-Boas & Winer, 1999; Villas-Boas & Zhao, 2005), we employ the two-stage least squares (2SLS) method.[16] In the method, we replace prices with $\kappa_0 + z_{jt} \cdot \kappa_1$ in the assumed pricing equation below

$$p_{jt} = \kappa_0 + z_{jt} \cdot \kappa_1 + \eta_{jt} \tag{3.2.5}$$

[16] See Appendix A.5 for detail.

where z_{jt} is called an instrument which is correlated with p_{jt} but not with ξ_{jt}, κ_0 and κ_1 are parameters to be estimated, and η_{jt} is a random error term. We additionally assume the following properties for η_{jt}:

$$E[\eta_{jt}] = 0 \tag{3.2.6}$$

$$\text{Cov}[z_{jt}, \eta_{jt}] = 0 \tag{3.2.7}$$

$$E[\eta_{jt}^2 | z_{jt}] < \infty. \tag{3.2.8}$$

If prices are endogenously determined, the terms ξ_{jt} and η_{jt} will be correlated since $\kappa_0 + z_{jt} \cdot \kappa_1$ is uncorrelated with ξ_{jt} by construction and thus η_{jt} represents a correlated (with ξ_{jt}) part of p_{jt}. This correlation should arise from the principle of the system where η_{jt} can represent both the cost shock and demand shock (i.e., if the demand for the particular brand is high, a firm can charge a premium price for the brand). In order to check the existence of price endogeneity, we further assume that ξ_{jt} and η_{jt} jointly follow the bi-variate normal distribution as the correlation in that distribution equates dependence between them.

The choice of instruments

The choice of instruments is not a trivial issue. Villas-Boas & Zhao (2005) used the lagged prices as instruments since they are readily available to researchers. Che et al. (2007) used the prices of brands in the other market avoiding the use of the lagged price because they hypothesized that firms might incorporate the effect of current price on the next period. In some studies, average prices of brands produced by the other firms are used as suggested in Berry et al. (1995). In this research, we used the average retail prices of yogurt in five stores we excluded from the analysis owing to lack of price information because those prices in other stores would reflect the general economic condition that would have affected retail prices in the target store as well and they would not be correlated with the unobserved demand shock ξ_{jt} which would include the effect of store-level promotions such as in-store display.

The market share

We denote the observed market share from the panel data and the market share calculated from the estimated demand parameters as $\tilde{S}_{jt}(\boldsymbol{p}_t)$ and $S_{jt}(\boldsymbol{p}_t)$ respectively for brand j at time t. The market share depends on all the explanatory variables in (3.2.1) but we only use an argument $\boldsymbol{p}_t$ to emphasize that market share endogenously depends on prices $\boldsymbol{p}_t = (p_{1t}, \cdots, p_{Jt})^T$. Given the estimated demand parameters, the market share of brand j at t, denoted as $S_{jt}(\boldsymbol{p}_t)$, is calculated as

$$S_{jt}(\boldsymbol{p}_t) = \sum_{s=1}^{S} \left[\sum_{i=1}^{I} \widehat{\Pr_{jts}} \cdot \lambda_s \right]$$

where $s = 1, \ldots, S$ are segments; λ_s is the fraction of segments; the term

$$\widehat{\Pr}_{jts} = \frac{\exp(\widehat{v_{jts}})}{1 + \sum_{k=1}^{J} \exp(\widehat{v_{kts}})}$$

is the estimated probability of brand j being chosen by the consumer belonging to segment s at time t; and $\widehat{v_{jts}}$ is the estimated utility defined as $\widehat{v_{jts}} \equiv \boldsymbol{x}_{jt} \cdot \widehat{\boldsymbol{\beta}_s} - \widehat{\alpha}_s \cdot \widehat{p_{jt}} + \mathrm{sim}_{kj} \cdot \widehat{SD}_s + \widehat{\xi}_{jt}$. Note that we replace the price and ξ_{jt} with the expected price $\widehat{p_{jt}}$ and $\widehat{\xi_{jt}}$ respectively to avoid endogeneity problem in constructing $\widehat{v_{jts}}$ from v_{jts}.

3.2.2 Supply-Side Specification

We follow Che et al. (2007) and estimate margins with a forward-looking model, whereby firms account for the effect of current prices on future demand. In the following subsection, we start with a myopic model and present how to derive manufacturers and retailers margins. Next, we will present how to derive margins in a forward-looking model. Henceforth, we denote manufacturer Stackelberg, vertical Nash, and retailer Stackelberg as MS, VN, and RS respectively.

Following the preceding research, we assume that the retailer is a local monopolist which maximizes joint category profit.[17] The assumption of a local monopolist is often justified by empirical reports which find that there is little evidence of intra-store competitions (Sudhir, 2001). However, we note that the effect of store competition is partly captured by the unobserved demand term ξ as promotion in the other retail store would affects demand or utility of brands in the store we analyze.

Profit functions

We will explain Bertrand competition game in the following, as collusion game is the special case of Bertrand competition game. The profit function of the monopolistic retailer and manufacturers are respectively defined as

$$\pi_{Rt} = \sum_{j=1}^{J} (p_{jt} - w_{jt}) S_{jt} M \tag{3.2.9}$$

and

$$\pi_{ft} = \sum_{j \in J_f} (w_{jt} - mc_{jt}) S_{jt} M \tag{3.2.10}$$

where J_f is a subset of brands produced by manufacturer $f = 1, \ldots, F$; S_{jt}, w_{jt}, and mc_{jt} are the market share, the wholesale price, and the marginal

[17] A retailer could use other pricing rules such as brand profit maximization rule where a retailer sets up a profit function for each brand instead of total profit maximization. However, Sudhir (2001) empirically shows that a retailer attains a maximum profit when it engages in category profit maximization, which supports the assumption widely adopted in the literature.

cost of brand j at time t respectively; and M is the market size. Then the first-order condition (FOC) of the profit functions[18] are

$$S_{jt} + \sum_{k=1}^{J} \left[(p_{kt} - w_{kt}) \frac{\partial S_{kt}}{\partial p_{jt}} \right] - \sum_{k=1}^{J} \left[\frac{\partial w_{kt}}{\partial p_{jt}} S_{kt} \right] = 0 \tag{3.2.11}$$

and

$$S_{lt} + \sum_{k \in J_f} \left[(w_{kt} - mc_{kt}) \sum_{h \in J_f} \frac{\partial S_{kt}}{\partial p_{ht}} \cdot \frac{\partial p_{ht}}{\partial w_{lt}} \right] = 0 \tag{3.2.12}$$

respectively[19] with the fixed M removed.

Stacking (3.2.11) vertically for $j = 1, \ldots, J$ and rearranging them in a matrix form, the retail margins in the general form are obtained as

$$\begin{pmatrix} p_{1t} - w_{1t} \\ \vdots \\ p_{Jt} - w_{Jt} \end{pmatrix} = - \begin{bmatrix} \frac{\partial S_{1t}}{\partial p_{1t}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{1t}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{1t}}{\partial p_{Jt}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{Jt}} \end{bmatrix}^{-1} \left[\boldsymbol{I} - \begin{bmatrix} \frac{\partial w_{1t}}{\partial p_{1t}} & \cdots & \frac{\partial w_{Jt}}{\partial p_{1t}} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_{1t}}{\partial p_{Jt}} & \cdots & \frac{\partial w_{Jt}}{\partial p_{Jt}} \end{bmatrix} \right] \begin{pmatrix} S_{1t} \\ \vdots \\ S_{Jt} \end{pmatrix} \tag{3.2.13}$$

assuming the inverse of the first matrix on the right-hand side of equation (3.2.13) exists. Similarly, by stacking (3.2.12) vertically for $l = 1, \ldots, J$ and rearranging them, the optimal manufacturer margins in the general form can be obtained as

$$\begin{pmatrix} w_{1t} - mc_{1t} \\ \vdots \\ w_{Jt} - mc_{Jt} \end{pmatrix} = - \left[\begin{bmatrix} \frac{\partial p_{1t}}{\partial w_{1t}} & \cdots & \frac{\partial p_{Jt}}{\partial w_{1t}} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{1t}}{\partial w_{Jt}} & \cdots & \frac{\partial p_{Jt}}{\partial w_{Jt}} \end{bmatrix} \begin{bmatrix} \frac{\partial S_{1t}}{\partial p_{1t}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{1t}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{1t}}{\partial p_{Jt}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{Jt}} \end{bmatrix} \cdot * \boldsymbol{\Omega} \right]^{-1} \begin{pmatrix} S_{1t} \\ \vdots \\ S_{Jt} \end{pmatrix} \tag{3.2.14}$$

[18] The optimal retail price p_{kt} should not be affected by the price of the other brands; else, p_{kt} will no longer be optimal. Thus, $\partial p_{kt} / \partial p_{jt}$ becomes 0 if p_{kt} is assessed at its optimal level.

[19] Note that it is assumed $\partial mc_{kt} / \partial w_{lt} = 0$ for all $k, l = 1, \ldots, J$, as wholesale price would not affect marginal cost in general.

where "$\cdot*$" denotes element-by-element multiplication and $\mathbf{\Omega}$ is a $J \times J$ ownership matrix whose (j,k) element, denoted as Ω_{jk}, is an indicator variable taking 1 if brands j and k are made by the same manufacturer and 0 otherwise.[20] The response curves $\partial w_{kt}/\partial p_{jt}$ in (3.2.13) and $\partial p_{ht}/\partial w_{lt}$ in (3.2.14) will be determined in MS, RS, and VN games respectively below.

Retailer margins in the MS game

We briefly review how retailer and manufacturer margins are derived in the MS game. The game is solved backward and retail margins are derived first. In the second stage of the game, since wholesale prices are already determined before retail prices are, we have

$$\frac{\partial w_{kt}}{\partial p_{jt}} = 0 \tag{3.2.15}$$

for all $k, j = 1, \ldots, J$. Substituting (3.2.15) to (3.2.13) yields the optimal retailer margin as

$$(\boldsymbol{p}_t - \boldsymbol{w}_t) = \mathbf{\Phi}_t^{-1} \boldsymbol{S}_t \tag{3.2.16}$$

where $(\boldsymbol{p}_t - \boldsymbol{w}_t) = (p_{1t} - w_{1t}, \ldots, p_{Jt} - w_{Jt})^T$, $\mathbf{\Phi}_t$ is the matrix whose (j,k) element is $-\partial S_{kt}/\partial p_{jt}$, and $\boldsymbol{S}_t = (S_{1t}, \ldots, S_{Jt})^T$.[21] Note that S_{lt} and $\partial S_{kt}/\partial p_{ht}$ in (3.2.16) can be directly observed and calculated.

Manufacturer margins in the MS game

On the other hand, in deriving manufacturer margins, the matrix of how a retailer optimally reacts to wholesale price change, $\partial p_{ht}/\partial w_{lt}$ in (3.2.14), must be indirectly inferred. Since the change in wholesale price of a brand would

[20] The derivation of FOCs is presented in Appendix A.8.

[21] For convenience in comparison to Che et al. (2007), notations and most definitions are the same as those in that paper.

affect retail prices of all brands, the term $\partial p_{ht}/\partial w_{lt}$ needs to be estimated by totally differentiating the FOC of the retail profit function with respect to the wholesale price as

$$\sum_{j=1}^{J}\left[\frac{\partial S_{gt}}{\partial p_{jt}}+\frac{\partial S_{jt}}{\partial p_{gt}}+\sum_{k=1}^{J}(p_{kt}-w_{kt})\frac{\partial^2 S_{kt}}{\partial p_{jt}\partial p_{gt}}\right]dp_{jt}-\frac{\partial S_{lt}}{\partial p_{gt}}dw_{lt}=0 \tag{3.2.17}$$

for some g. Denoting the terms inside the bracket on the left hand side of equation (3.2.17) as $\nu(g,j)$, we have the set of J equations for some l as

$$\begin{cases} \nu(1,1)dp_{1t}+\nu(1,2)dp_{2t}+\cdots+\nu(1,J)dp_{Jt}=\frac{\partial S_{lt}}{\partial p_{1t}}dw_{lt}, \\ \vdots \\ \nu(J,1)dp_{1t}+\nu(J,2)dp_{2t}+\cdots+\nu(J,J)dp_{Jt}=\frac{\partial S_{lt}}{\partial p_{Jt}}dw_{lt}. \end{cases} \tag{3.2.18}$$

Defining $G_{gt}\equiv(\nu(g,1),\ldots,\nu(g,J))$, we rewrite the expression in (3.2.18) in matrix form as

$$\begin{pmatrix} G_{1t} \\ \vdots \\ G_{Jt} \end{pmatrix}\begin{pmatrix} dp_{1t} \\ \vdots \\ dp_{Jt} \end{pmatrix}=\begin{pmatrix} \frac{\partial S_{lt}}{\partial p_{1t}} \\ \vdots \\ \frac{\partial S_{lt}}{\partial p_{Jt}} \end{pmatrix}dw_{lt}$$

or

$$\begin{pmatrix} \partial p_{1t}/\partial w_{lt} \\ \vdots \\ \partial p_{Jt}/\partial w_{lt} \end{pmatrix}=\begin{pmatrix} G_{1t} \\ \vdots \\ G_{Jt} \end{pmatrix}^{-1}\begin{pmatrix} \frac{\partial S_{lt}}{\partial p_{1t}} \\ \vdots \\ \frac{\partial S_{lt}}{\partial p_{Jt}} \end{pmatrix}, \tag{3.2.19}$$

assuming the inverse of the $J\times J$ matrix $(G_{1t}\cdots G_{Jt})^T$ exists. Transposing the both sides of (3.2.19), we have

$$\left(\frac{\partial p_{1t}}{\partial w_{lt}}\cdots\frac{\partial p_{Jt}}{\partial w_{lt}}\right)=\left(\frac{\partial S_{lt}}{\partial p_{1t}}\cdots\frac{\partial S_{lt}}{\partial p_{Jt}}\right)\left(G_{1t}^T\cdots G_{Jt}^T\right)^{-1}$$

for some l. Stacking this vector vertically for $l=1\cdots J$, we have

$$\begin{bmatrix} \frac{\partial p_{1t}}{\partial w_{1t}} & \cdots & \frac{\partial p_{Jt}}{\partial w_{1t}} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{1t}}{\partial w_{Jt}} & \cdots & \frac{\partial p_{Jt}}{\partial w_{Jt}} \end{bmatrix}=\begin{bmatrix} \frac{\partial S_{1t}}{\partial p_{1t}} & \cdots & \frac{\partial S_{1t}}{\partial p_{Jt}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{Jt}}{\partial p_{1t}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{Jt}} \end{bmatrix}\left(G_{1t}^T\cdots G_{Jt}^T\right)^{-1}. \tag{3.2.20}$$

Substituting (3.2.20) to (3.2.14), we have the manufacturers' margins as

$$(\boldsymbol{w}_t - \boldsymbol{mc}_t) = -[[\boldsymbol{\Phi}_t]^T \boldsymbol{G}_t^{-1} \boldsymbol{\Phi}_t \cdot * \boldsymbol{\Omega}]^{-1} \boldsymbol{S}_t \tag{3.2.21}$$

where $(\boldsymbol{w}_t - \boldsymbol{mc}_t) = (w_{1t} - mc_{1t}, \ldots, w_{Jt} - mc_{Jt})^T$. We note that $\boldsymbol{G}_t$ is the matrix whose (j, h) element is

$$\frac{\partial S_{jt}}{\partial p_{ht}} + \frac{\partial S_{ht}}{\partial p_{jt}} + \sum_{k=1}^{J} (p_{kt} - w_{kt}) \frac{\partial^2 S_{kt}}{\partial p_{jt} \partial p_{ht}}.$$

Manufacturer margins in the RS game

In the RS game, we have $\partial(p_{ht} - w_{ht})/\partial w_{lt} = 0$ for all $h, l = 1, \ldots, J$ in the second stage since the retail margin on brand h or $(p_{ht} - w_{ht})$ is set prior to wholesale prices being set. Equivalently, we have

$$\begin{cases} \partial p_{lt}/\partial w_{lt} = 1 \\ \partial p_{ht}/\partial w_{lt} = 0 \end{cases} \tag{3.2.22}$$

since $\partial w_{lt}/\partial w_{lt} = 1$ and $\partial w_{ht}/\partial w_{lt} = 0$. Then, from (3.2.22) and (3.2.12), we have

$$S_{lt} + \sum_{k=1}^{J} \Omega_{lk} \left[(w_{kt} - mc_{kt}) \frac{\partial S_{kt}}{\partial p_{lt}} \right] = 0. \tag{3.2.23}$$

Stacking (3.2.23) vertically for $l = 1, \ldots, J$ and rearranging them, we derive the optimal manufacturer margins in the RS game as

$$(\boldsymbol{w}_t - \boldsymbol{mc}_t) = [\boldsymbol{\Phi}_t \cdot * \boldsymbol{\Omega}]^{-1} \boldsymbol{S}_t. \tag{3.2.24}$$

Retailer margins in the RS game

To derive retail margins in the RS game, the matrix of how manufacturers optimally react to retail price change $\partial w_{kt}/\partial p_{jt}$ in (3.2.13) must be inferred. Similar to the MS case, we totally differentiate the FOC of the manufacturers' profit function in (3.2.23) with respect to p_{jt} and solve the resulting equations for $\partial w_{kt}/\partial p_{jt}$, the optimal reaction curve of the manufacturer.

Stacking the total derivatives of FOC of the manufacturer profit function in (3.2.23) with respect to p_{jt} vertically for $l = 1, \ldots, J$, we have

$$\begin{cases} \frac{dS_{1t}}{dp_{jt}} + \sum_{k=1}^{J} \left[\Omega_{1k} \frac{dw_{kt}}{dp_{jt}} \cdot \frac{\partial S_{kt}}{\partial p_{1t}} + \Omega_{1k}(w_{kt} - mc_{kt}) \frac{d}{dp_{jt}} \left(\frac{\partial S_{kt}}{\partial p_{1t}} \right) \right] = 0 \\ \vdots \\ \frac{dS_{Jt}}{dp_{jt}} + \sum_{k=1}^{J} \left[\Omega_{Jk} \frac{dw_{kt}}{dp_{jt}} \cdot \frac{\partial S_{kt}}{\partial p_{Jt}} + \Omega_{Jk}(w_{kt} - mc_{kt}) \frac{d}{dp_{jt}} \left(\frac{\partial S_{kt}}{\partial p_{Jt}} \right) \right] = 0 \end{cases} \tag{3.2.25}$$

since the marginal cost is not affected by the retail price (i.e., $\partial mc_{kt}/\partial p_{jt} = 0$ for all $k, j = 1, \ldots, J$). Further we have

$$\frac{dS_{lt}}{dp_{jt}} = \frac{\partial S_{lt}}{\partial p_{1t}} \cdot \frac{\partial p_{1t}}{\partial p_{jt}} + \cdots + \frac{\partial S_{lt}}{\partial p_{Jt}} \cdot \frac{\partial p_{Jt}}{\partial p_{jt}} = \frac{\partial S_{lt}}{\partial p_{jt}}$$

since $\partial S_{lt}/\partial p_{jt} \equiv \partial S_{lt}/\partial p_{jt}|_{\boldsymbol{p}=\boldsymbol{p}}$ and

$$\frac{dw_{kt}}{dp_{jt}} = \frac{\partial w_{kt}}{\partial p_{1t}} \cdot \frac{\partial p_{1t}}{\partial p_{jt}} + \cdots + \frac{\partial w_{kt}}{\partial p_{Jt}} \cdot \frac{\partial p_{Jt}}{\partial p_{jt}} = \frac{\partial w_{kt}}{\partial p_{jt}}$$

since $\partial p_{ht}/\partial p_{jt} = 0$ for all $h, j = 1, \ldots, J$ and $\partial p_{jt}/\partial p_{jt} = 1$ for all $j = 1, \ldots, J$. Rearranging (3.2.25) as a matrix, we have

$$\left[\begin{bmatrix} \frac{\partial S_{1t}}{\partial p_{1t}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{1t}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{1t}}{\partial p_{Jt}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{Jt}} \end{bmatrix} \cdot * \boldsymbol{\Omega} \right] \begin{pmatrix} \frac{\partial w_{1t}}{\partial p_{jt}} \\ \vdots \\ \frac{\partial w_{Jt}}{\partial p_{jt}} \end{pmatrix} = - \begin{bmatrix} \frac{\partial S_{1t}}{\partial p_{jt}} + \sum_{k=1}^{J} \Omega_{1k}(w_{kt} - mc_{kt}) \frac{\partial^2 S_{kt}}{\partial p_{1t} \partial p_{jt}} \\ \vdots \\ \frac{\partial S_{Jt}}{\partial p_{jt}} + \sum_{k=1}^{J} \Omega_{Jk}(w_{kt} - mc_{kt}) \frac{\partial^2 S_{kt}}{\partial p_{Jt} \partial p_{jt}} \end{bmatrix}. \tag{3.2.26}$$

Stacking (3.2.26) horizontally for $j = 1, \ldots, J$ and rearranging them, we have

$$\begin{bmatrix} \frac{\partial w_{1t}}{\partial p_{1t}} & \cdots & \frac{\partial w_{1t}}{\partial p_{Jt}} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_{Jt}}{\partial p_{1t}} & \cdots & \frac{\partial w_{Jt}}{\partial p_{Jt}} \end{bmatrix} = \left[\boldsymbol{\Phi}_t \cdot * \boldsymbol{\Omega} \right]^{-1} \cdot \boldsymbol{H}_t \tag{3.2.27}$$

where $\boldsymbol{H}_t$ is a $J \times J$ matrix whose (l, j) element is

$$\frac{\partial S_{lt}}{\partial p_{jt}} + \sum_{k=1}^{J} \Omega_{lk}(w_{kt} - mc_{kt}) \frac{\partial^2 S_{kt}}{\partial p_{lt} \partial p_{jt}}.$$

We obtain retailer margins in the RS game by transposing both sides of (3.2.27) and substituting it to (3.2.13) as

$$(\boldsymbol{p}_t - \boldsymbol{w}_t) = \boldsymbol{\Phi}_t^{-1}[\boldsymbol{I} - \boldsymbol{H}_t^T[[\boldsymbol{\Phi}_t]^T \cdot *\boldsymbol{\Omega}]^{-1}]\boldsymbol{S}_t. \tag{3.2.28}$$

Margins in the VN game

In the VN game, manufacturers and a retailer move simultaneously based on their predictions of other players' behavior. More specifically, manufacturers set wholesale price assuming a certain level of retail margin for the brand; a retailer sets retail margins irrespective of wholesale prices. This structure is likely to emerge in the market where manufacturers and the retailer have approximately equal power (Choi, 1991).

Che et al. (2007) substitutes (3.2.15) to (3.2.13) and derives retail margin in the VN game as (3.2.16), and substitutes (3.2.22) to (3.2.14) and derives the manufacturer margin in the VN game as (3.2.24) because conditions (3.2.15) and (3.2.22) simultaneously hold in VN game since the retailer and manufacturers move simultaneously. We note that the margins of the retailer and manufacturers become identical if manufacturers collude in this game. This makes sense as the VN game assumes approximately equal power between manufacturers and retailer.

Arriving at VN from two extreme directions

The term $\boldsymbol{H}_t^T[[\boldsymbol{\Phi}_t]^T \cdot *\boldsymbol{\Omega}]^{-1}$ in retail profit in the RS game is the matrix whose (l, j) element is $\partial w_{jt}/\partial p_{lt}$. Notice that these terms are 0 for $l, j = 1, \ldots, J$ when we employ the behavior (3.2.15) of manufacturers in the MS game. In other words, retailer profit in the VN game can be obtained by applying the manufacturer behavior in the MS game to the retail margin. Similarly, the term $[\boldsymbol{\Phi}_t]^T \boldsymbol{G}_t^{-1}$ in manufacturer profit in the MS game is (3.2.20) whose

Table 3.2.1: Margins under each game

	Manufacturer Stackelberg	Vertical Nash	Retailer Stackelberg
Retailer Margin	$\boldsymbol{\Phi}_t^{-1}\boldsymbol{S}_t$	$\boldsymbol{\Phi}_t^{-1}\boldsymbol{S}_t$	$\boldsymbol{\Phi}_t^{-1}[\boldsymbol{I}-\boldsymbol{H}_t^T[[\boldsymbol{\Phi}_t]^T\cdot*\boldsymbol{\Omega}]^{-1}]\boldsymbol{S}_t$
Manufacturer Margin	$-[[\boldsymbol{\Phi}_t]^T\boldsymbol{G}_t^{-1}\boldsymbol{\Phi}_t\cdot*\boldsymbol{\Omega}]^{-1}\boldsymbol{S}_t$	$[\boldsymbol{\Phi}_t\cdot*\boldsymbol{\Omega}]^{-1}\boldsymbol{S}_t$	$[\boldsymbol{\Phi}_t\cdot*\boldsymbol{\Omega}]^{-1}\boldsymbol{S}_t$

(l,h) element is $\partial p_{ht}/\partial w_{lt}$. Note that the matrix of these terms becomes an identity matrix when we employ the behavior (3.2.22) of the retailer in the RS game. This is the symmetrical relationship of MS and RS games we refer to in section 3.1. In summary, we present the formulation of margins under each game in Table 3.2.1. We note that margins under each game with collusive manufacturers can be derived by making $\boldsymbol{\Omega}$ an matrix with all elements being unity.

The forward-looking model

We briefly review how to derive margins in the forward-looking model in this section. Though we consider firms only look one-period ahead as in Che et al. (2007), the following derivations can be generalized to more than one-period ahead behavior.

Retailer's margin (forward-looking model)

The objective function of forward-looking retailer is $V_R = \pi_{R1} + \delta\pi_{R2}$, where π_{Rt} is a profit function defined in (3.2.9) for period $t = 1, 2$, and the term δ is some exogenously given discount rate. Then FOCs for some j are

$$\begin{cases} \frac{\partial \pi_{R1}}{\partial p_{j1}} + \delta \sum_{k=1}^{J} \frac{\partial \pi_{R2}}{\partial S_{k2}} \cdot \frac{\partial S_{k2}}{\partial S_{k1}} \cdot \frac{\partial S_{k1}}{\partial p_{j1}} & = & 0 \\ \frac{\partial \pi_{R2}}{\partial p_{j2}} & = & 0. \end{cases} \tag{3.2.29}$$

We have the set of equations as in (3.2.29) for $j = 1, \ldots, J$. The first equation corresponds to the objective function in period 1 and the second equation corresponds to that in period 2. Note here that firms become myopic in period 2 in this setting.

As the margins in period 2, which is identical to the myopic case, is already derived, we only concern for the profit function in period 1 in the following derivation. Our strategy is to decompose and translate each component of the second term on the left-hand side of the first equation above to the expression we can calculate. Clearly, the first component $\partial\pi_{R2}/\partial S_{k2}$ is $(p_{k2} - w_{k2})$. To calculate the second component $\partial S_{k2}/\partial S_{k1}$, we exploit the following relationship:

$$S_{k2} = \theta_{k2|k1} \times S_{k1} + \sum_{l=1, l\neq k}^{J} \theta_{k2|l1} \times S_{l1} \tag{3.2.30}$$

where $\theta_{k2|k1}$ is the probability of purchasing brand k in period 2 given the purchase of the brand in period 1. The term $\theta_{k2|l1}$ is defined likewise for brand l. The second term on the right-hand side of equation (3.2.30) can be rewritten as

$$\begin{aligned}\sum_{l=1, l\neq k}^{J} \theta_{k2|l1} \times S_{l1} &= \theta_{k2|11} \times S_{11} + \cdots + \theta_{k2|k-1,1} \times S_{k-1,1} + \theta_{k2|k+1,1} \times S_{k+1,1} + \\ &\quad \cdots + \theta_{k2|J1} \times S_{J1}.\end{aligned}$$

Since the terms S_{l1} on the right hand side of equation can be rewritten as $S_{l1} = (1 - S_{11} - \cdots - S_{l-1,1} - S_{l+1,1} - \cdots - S_{J1})$ for all $l = 1, \ldots, J$, $l \neq k$, all S_{l1} include the term $-S_{k1}$ on these relationships. Thus, the partial derivative of the second term on the right-hand side of equation (3.2.30) with respect to S_{k1} is

$$\frac{\partial \left[\sum_{l=1, l\neq k}^{J} \theta_{k2|l1} \times S_{l1}\right]}{\partial S_{k1}} = -\sum_{l=1, l\neq k}^{J} \theta_{k2|l1}$$

as $\partial S_{l1}/\partial S_{k1} = -1$ for $l = 1, \ldots, J$, $l \neq k$. Thus taking partial derivative of both sides of (3.2.30) with respect to S_{k1}, we have

$$\frac{\partial S_{k2}}{\partial S_{k1}} = \theta_{k2|k1} - \sum_{l=1,l\neq k}^{J} \theta_{k2|l1}.$$

This is the second component and let us define it as Δ_k. This term is the repurchasing probability of the brand minus the switching probabilities from the other brands, and can be calculated from the predicted market share estimates.

In the same manner as we showed in the derivation of vector $(\boldsymbol{p}_t - \boldsymbol{w}_t)$, the second term on the left-hand side of the first equation in (3.2.29) can be stacked vertically for $j = 1, \ldots, J$ and expressed by matrix form as

$$\delta \begin{bmatrix} \frac{\partial S_{11}}{\partial p_{11}} & \cdots & \frac{\partial S_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{11}}{\partial p_{J1}} & \cdots & \frac{\partial S_{J1}}{\partial p_{J1}} \end{bmatrix} \begin{bmatrix} \Delta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta_J \end{bmatrix} \begin{pmatrix} p_{12} - w_{12} \\ \vdots \\ p_{J2} - w_{J2} \end{pmatrix}$$

where the second matrix is diagonal matrix with elements Δ_k which we will express as $\boldsymbol{\Delta}$. We also know from (3.2.9) that the first term on the left-hand side of the first equation in (3.2.29) can be stacked vertically for $j = 1, \ldots, J$ and expressed by

$$\begin{pmatrix} \frac{\partial \pi_{R1}}{\partial p_{11}} \\ \vdots \\ \frac{\partial \pi_{R1}}{\partial p_{J1}} \end{pmatrix} = \left[\boldsymbol{I} - \begin{bmatrix} \frac{\partial w_{11}}{\partial p_{11}} & \cdots & \frac{\partial w_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_{11}}{\partial p_{J1}} & \cdots & \frac{\partial w_{J1}}{\partial p_{J1}} \end{bmatrix} \right] \begin{pmatrix} S_{11} \\ \vdots \\ S_{J1} \end{pmatrix} + \begin{bmatrix} \frac{\partial S_{11}}{\partial p_{11}} & \cdots & \frac{\partial S_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{11}}{\partial p_{J1}} & \cdots & \frac{\partial S_{J1}}{\partial p_{J1}} \end{bmatrix} \begin{pmatrix} p_{11} - w_{11} \\ \vdots \\ p_{J1} - w_{J1} \end{pmatrix}.$$

Thus we have

$$\left[\boldsymbol{I} - \begin{bmatrix} \frac{\partial w_{11}}{\partial p_{11}} & \cdots & \frac{\partial w_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_{11}}{\partial p_{J1}} & \cdots & \frac{\partial w_{J1}}{\partial p_{J1}} \end{bmatrix} \right] \begin{pmatrix} S_{11} \\ \vdots \\ S_{J1} \end{pmatrix} + \begin{bmatrix} \frac{\partial S_{11}}{\partial p_{11}} & \cdots & \frac{\partial S_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{11}}{\partial p_{J1}} & \cdots & \frac{\partial S_{J1}}{\partial p_{J1}} \end{bmatrix} \begin{pmatrix} p_{11} - w_{11} \\ \vdots \\ p_{J1} - w_{J1} \end{pmatrix}$$

$$+\delta \begin{bmatrix} \frac{\partial S_{11}}{\partial p_{11}} & \cdots & \frac{\partial S_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{11}}{\partial p_{J1}} & \cdots & \frac{\partial S_{J1}}{\partial p_{J1}} \end{bmatrix} \begin{bmatrix} \Delta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta_J \end{bmatrix} \begin{pmatrix} p_{12} - w_{12} \\ \vdots \\ p_{J2} - w_{J2} \end{pmatrix} = 0$$

or

$$
\begin{pmatrix} p_{11} - w_{11} \\ \vdots \\ p_{J1} - w_{J1} \end{pmatrix} = - \begin{bmatrix} \frac{\partial S_{11}}{\partial p_{11}} & \cdots & \frac{\partial S_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{11}}{\partial p_{J1}} & \cdots & \frac{\partial S_{J1}}{\partial p_{J1}} \end{bmatrix}^{-1} \left[\boldsymbol{I} - \begin{bmatrix} \frac{\partial w_{11}}{\partial p_{11}} & \cdots & \frac{\partial w_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_{11}}{\partial p_{J1}} & \cdots & \frac{\partial w_{J1}}{\partial p_{J1}} \end{bmatrix} \right] \begin{pmatrix} S_{11} \\ \vdots \\ S_{J1} \end{pmatrix}
$$

$$
-\delta \begin{bmatrix} \Delta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta_J \end{bmatrix} \begin{pmatrix} p_{12} - w_{12} \\ \vdots \\ p_{J2} - w_{J2} \end{pmatrix} \tag{3.2.31}
$$

assuming the inverse of $\mathbf{\Phi}_1$ exists. The second equation of (3.2.29) can be obtained as the myopic case. In estimation, we first estimate margins of myopic model for $t = 2, \ldots, T$ and use these margins in computing the first-period margins. This derivation is that of RS game; those of MS and VN game can be obtained by applying (3.2.15) to (3.2.31), where (3.2.31) reduces to

$$
\begin{pmatrix} p_{11} - w_{11} \\ \vdots \\ p_{J1} - w_{J1} \end{pmatrix} = - \begin{bmatrix} \frac{\partial S_{11}}{\partial p_{11}} & \cdots & \frac{\partial S_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{11}}{\partial p_{J1}} & \cdots & \frac{\partial S_{J1}}{\partial p_{J1}} \end{bmatrix}^{-1} \begin{pmatrix} S_{11} \\ \vdots \\ S_{J1} \end{pmatrix}
$$

$$
-\delta \begin{bmatrix} \Delta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta_J \end{bmatrix} \begin{pmatrix} p_{12} - w_{12} \\ \vdots \\ p_{J2} - w_{J2} \end{pmatrix}.
$$

Manufacturers' margins (forward-looking)

The derivation of manufacturers' margins in the one-period forward-looking model follows much of the retail counterpart. The objective function is $V_M = \pi_{f1} + \delta \pi_{f2}$ and the FOCs are

$$
\begin{cases} \frac{\partial \pi_{f1}}{\partial w_{j1}} + \delta \sum_{j \in J_f} \frac{\partial \pi_{f2}}{\partial S_{k2}} \cdot \frac{\partial S_{k2}}{\partial S_{k1}} \cdot \frac{\partial S_{k1}}{\partial w_{j1}} & = & 0 \\ \frac{\partial \pi_{f2}}{\partial w_{j2}} & = & 0. \end{cases} \tag{3.2.32}
$$

Clearly, $\partial\pi_{f2}/\partial S_{k2} = (w_{k2} - mc_{k2})$. The term $\partial S_{k2}/\partial S_{k1}$ is the same as in the case of retailer's margin. Note that the set of equations consist of $\sum_{j\in J_f} \partial\pi_{f2}/\partial S_{k2} \cdot \partial S_{k1}/\partial w_{j1}$ or $\sum_{j\in J_f}(w_{k2} - mc_{k2})\partial S_{k1}/\partial w_{j1}$ for $j = 1, \ldots, J$ is equal to the second terms of the left-hand side of equations in (3.2.12) and thus can be written as

$$\left[\begin{bmatrix} \frac{\partial p_{11}}{\partial w_{11}} & \cdots & \frac{\partial p_{J1}}{\partial w_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{11}}{\partial w_{J1}} & \cdots & \frac{\partial p_{J1}}{\partial w_{J1}} \end{bmatrix} \begin{bmatrix} \frac{\partial S_{11}}{\partial p_{11}} & \cdots & \frac{\partial S_{J1}}{\partial p_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{11}}{\partial p_{J1}} & \cdots & \frac{\partial S_{J1}}{\partial p_{J1}} \end{bmatrix} \cdot *\boldsymbol{\Omega}\right] \begin{pmatrix} w_{12} - mc_{12} \\ \vdots \\ w_{J2} - mc_{J2} \end{pmatrix}$$

or simply $[[\boldsymbol{\Phi}_1]^T \boldsymbol{G}_1^{-1} \boldsymbol{\Phi}_1 \cdot *\boldsymbol{\Omega}](\boldsymbol{w}_2 - \boldsymbol{mc}_2)$. Thus the second term on the left-hand side of the first equation of (3.2.32) becomes $\delta[[\boldsymbol{\Phi}_1]^T \boldsymbol{G}_1^{-1} \boldsymbol{\Phi}_1 \cdot *\boldsymbol{\Omega}]\boldsymbol{\Delta}(\boldsymbol{w}_2 - \boldsymbol{mc}_2)$. Since the first term of the first equation in (3.2.32), if stacked vertically for $j = 1, \ldots, J$, is $\boldsymbol{S}_1 + [[\boldsymbol{\Phi}_1]^T \boldsymbol{G}_1^{-1} \boldsymbol{\Phi}_1 \cdot *\boldsymbol{\Omega}](\boldsymbol{w}_1 - \boldsymbol{mc}_1)$, we have $\boldsymbol{S}_1 + [[\boldsymbol{\Phi}_1]^T \boldsymbol{G}_1^{-1} \boldsymbol{\Phi}_1 \cdot *\boldsymbol{\Omega}](\boldsymbol{w}_1 - \boldsymbol{mc}_1) + \delta[[\boldsymbol{\Phi}_1]^T \boldsymbol{G}_1^{-1} \boldsymbol{\Phi}_1 \cdot *\boldsymbol{\Omega}]\boldsymbol{\Delta}(\boldsymbol{w}_2 - \boldsymbol{mc}_2) = 0$ or $(\boldsymbol{w}_1 - \boldsymbol{mc}_1) = -[[\boldsymbol{\Phi}_1]^T \boldsymbol{G}_1^{-1} \boldsymbol{\Phi}_1 \cdot *\boldsymbol{\Omega}]^{-1}\boldsymbol{S}_1 - \delta \cdot \boldsymbol{\Delta}(\boldsymbol{w}_2 - \boldsymbol{mc}_2)$ assuming the inverse of $[[\boldsymbol{\Phi}_1]^T \boldsymbol{G}_1^{-1} \boldsymbol{\Phi}_1 \cdot *\boldsymbol{\Omega}]$ exists. This derivation is that of MS game; those of VN and RS game can be obtained by assuming (3.2.22) in which case $\boldsymbol{G}_1^{-1}\boldsymbol{\Phi}_1$ becomes an identity matrix. The collusion case is obtained by making $\boldsymbol{\Omega}$ an matrix with all elements being unity.

3.3 Estimation

3.3.1 Demand-Side Estimation

On the demand-side estimation, we first obtain $\widehat{p_{jt}}$, $\widehat{\eta_{jt}}$ and $\widehat{\xi_{jt}}$ by preliminary analysis by 2SLS estimation. Then we maximize simulated likelihood function using $\widehat{p_{jt}}$ and $\widehat{\xi_{jt}}$.

Estimating ξ_{jt} and η_{jt}

The joint distributions of ξ_{jt} and η_{jt} are estimated as residuals in the following equations if the logit specification in (3.2.1) are justified:

$$\ln \tilde{S}_{jt} - \ln \tilde{S}_{0t} = \boldsymbol{x}_{jt} \cdot \boldsymbol{\beta} - \alpha \cdot p_{jt} + \mathrm{sim}_{kj} \cdot SD + \xi_{jt} \tag{3.3.1}$$

$$p_{jt} = \kappa_0 + \boldsymbol{z}_{jt} \cdot \kappa_1 + \eta_{jt} \tag{3.3.2}$$

for all i, j and t respectively. Note that these equations follow from (3.2.1) and (3.2.5) respectively. We first estimate the equation (3.3.2) by ordinary least squares (OLS) and obtain $\widehat{p_{jt}}$ and $\widehat{\eta_{jt}}$. Then we substitute $\widehat{p_{jt}}$ to p_{jt} in the equation (3.3.1) to correct possible endogeneity between price and the unobserved product characteristics.[22] The structure of (3.3.1) and (3.3.2) guarantees that all $\widehat{\xi_{jt}}$ is paired with $\widehat{\eta_{jt}}$ and vice versa for all j and t.

Following Draganska & Jain (2004), we assume that ξ_{jt} and η_{jt} are independent across t and their marginal densities are both normally distributed. Let the sum of the purchasing occasions for all consumers $T_1 + \cdots + T_N$ be $\boldsymbol{T}$. We estimate ξ_{jt} and η_{jt} as follows:

[22] Since we assume conditions (3.2.2) and (3.2.3), the least-square estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\alpha}$, and $\widehat{SD}$ are BLUE by Gauss-Markov theorem if $\mathrm{Cov}(\xi_{jt}, \xi_{jt'}) = 0$ for all $t \neq t'$ and $V(\xi_{jt}) = \sigma^2_{\xi_j}$ for all t. Moreover, under rank condition where the outer product of the vector defined as $\boldsymbol{Z} \equiv (\boldsymbol{x}_{jt}, \mathrm{sim}_{kj}, 1, z_{jt})$ being full-rank, the assumption of $E[\boldsymbol{Z}^T, \xi_{jt}] = \boldsymbol{0}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\alpha}$, and $\widehat{SD}$ are consistent. Since $V(\xi_{jt}) = \sigma^2_{\xi_j}$ for all t, $\widehat{\boldsymbol{\beta}}$, $\widehat{\alpha}$, and $\widehat{SD}$ are asymptotically normal. Similarly, since we assume conditions (3.2.6) and (3.2.7), $\widehat{\kappa_0}$ and $\widehat{\kappa_1}$ are BLUE by Gauss-Markov theorem if $\mathrm{Cov}(\eta_{jt}, \eta_{jt'}) = 0$ for all $t \neq t'$ and $V(\eta_{jt}) = \sigma^2_{\eta_j}$ for all t. Under the rank condition of the outer product of $(\boldsymbol{1}, z^T_{jt})$ being full-rank and the condition $E[z^T_{jt}, \eta_{jt}] = \boldsymbol{0}$ which follows from conditions (3.2.6) and (3.2.7), $\widehat{\kappa_0}$ and $\widehat{\kappa_1}$ are consistent. Since $V(\eta_{jt}) = \sigma^2_{\eta_j}$ for all t, $\widehat{\kappa_0}$ and $\widehat{\kappa_1}$ are asymptotically normal.

Step 1: Run OLS regression to estimate $\widehat{\kappa}_0$ and $\widehat{\kappa_1}$ as follows:

$$\begin{pmatrix} p_{j,t_1=1} \\ p_{j,t_1=2} \\ \vdots \\ p_{j,t_1=T_1} \\ \vdots \\ p_{j,t_i=1} \\ p_{j,t_i=2} \\ \vdots \\ p_{j,t_i=T_i} \\ \vdots \\ p_{j,t_N=1} \\ p_{j,t_N=2} \\ \vdots \\ p_{j,t_N=T_N} \end{pmatrix} = \begin{bmatrix} 1, z_{j,t_1=1} \\ 1, z_{j,t_1=2} \\ \vdots \\ 1, z_{j,t_1=T_1} \\ \vdots \\ 1, z_{j,t_i=1} \\ 1, z_{j,t_i=2} \\ \vdots \\ 1, z_{j,t_i=T_i} \\ \vdots \\ 1, z_{j,t_N=1} \\ 1, z_{j,t_N=2} \\ \vdots \\ 1, z_{j,t_N=T_N} \end{bmatrix} \begin{pmatrix} \kappa_0 \\ \kappa_1 \end{pmatrix} + \begin{pmatrix} \eta_{j,t_1=1} \\ \eta_{j,t_1=2} \\ \vdots \\ \eta_{j,t_1=T_1} \\ \vdots \\ \eta_{j,t_i=1} \\ \eta_{j,t_i=2} \\ \vdots \\ \eta_{j,t_i=T_i} \\ \vdots \\ \eta_{j,t_N=1} \\ \eta_{j,t_N=2} \\ \vdots \\ \eta_{j,t_N=T_N} \end{pmatrix}.$$

The left-hand side is the $\boldsymbol{T} \times 1$ vector with each a vector with each row being p_{jt_i} with inner loop of t_i and outer loop for $i = 1, \cdots, N$ and the first term on the right-hand side if the $\boldsymbol{T} \times 2$ matrix with each row being $(1, z_{jt_i})$ with inner loop of t_i and outer loop for $i = 1, \cdots, N$. Obtain $\widehat{p_{jt_i}}$ and $\widehat{\eta_{jt_i}}$.

Step 2: Run OLS regression as follows:

$$\begin{pmatrix} \ln \tilde{S}_{j,t_1=1} - \ln \tilde{S}_{0,t_1=1} \\ \ln \tilde{S}_{j,t_1=2} - \ln \tilde{S}_{0,t_1=2} \\ \vdots \\ \ln \tilde{S}_{jt_1=T_1} - \ln \tilde{S}_{0,t_1=T_1} \\ \vdots \\ \ln \tilde{S}_{j,t_i=1} - \ln \tilde{S}_{0,t_i=1} \\ \ln \tilde{S}_{j,t_i=2} - \ln \tilde{S}_{0,t_i=2} \\ \vdots \\ \ln \tilde{S}_{j,t_i=T_i} - \ln \tilde{S}_{0,t_i=T_i} \\ \vdots \\ \ln \tilde{S}_{j,t_N=1} - \ln \tilde{S}_{0,t_N=1} \\ \ln \tilde{S}_{j,t_N=2} - \ln \tilde{S}_{0,t_N=2} \\ \vdots \\ \ln \tilde{S}_{j,t_N=T_N} - \ln \tilde{S}_{0,t_N=T_N} \end{pmatrix} = \begin{bmatrix} \boldsymbol{x}_{j,t_1=1}, \widehat{p_{j,t_1=1}}, \mathbf{I}_{kj} \\ \boldsymbol{x}_{j,t_1=2}, \widehat{p_{j,t_1=2}}, \mathbf{I}_{kj} \\ \vdots \\ \boldsymbol{x}_{j,t_1=T_1}, \widehat{p_{j,t_1=T_1}}, \mathbf{I}_{kj} \\ \vdots \\ \boldsymbol{x}_{j,t_i=1}, \widehat{p_{j,t_i=1}}, \mathbf{I}_{kj} \\ \boldsymbol{x}_{j,t_i=2}, \widehat{p_{j,t_i=2}}, \mathbf{I}_{kj} \\ \vdots \\ \boldsymbol{x}_{j,t_i=T_i}, \widehat{p_{j,t_i=T_i}}, \mathbf{I}_{kj} \\ \vdots \\ \boldsymbol{x}_{j,t_N=1}, \widehat{p_{j,t_N=1}}, \mathbf{I}_{kj} \\ \boldsymbol{x}_{j,t_N=2}, \widehat{p_{j,t_N=2}}, \mathbf{I}_{kj} \\ \vdots \\ \boldsymbol{x}_{j,t_N=T_N}, \widehat{p_{j,t_N=T_N}}, \mathbf{I}_{kj} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ -\alpha \\ \boldsymbol{\psi}^T \end{pmatrix} + \begin{pmatrix} \xi_{j,t_1=1} \\ \xi_{j,t_1=2} \\ \vdots \\ \xi_{j,t_1=T_1} \\ \vdots \\ \xi_{j,t_i=1} \\ \xi_{j,t_i=2} \\ \vdots \\ \xi_{j,t_i=T_i} \\ \vdots \\ \xi_{j,t_N=1} \\ \xi_{j,t_N=2} \\ \vdots \\ \xi_{j,t_N=T_N} \end{pmatrix}.$$

The vector on the left-hand side is the $\boldsymbol{T} \times 1$ vector with each row being $\ln \tilde{S}_{jt_i} - \ln \tilde{S}_{0t_i}$ with inner loop of t_i and outer loop for $i = 1, \cdots, N$. The first term on the right-hand side is the $\boldsymbol{T} \times (7 + 1 + 12)$ matrix with each row being $(\boldsymbol{x}_{jt_i}, \widehat{p_{jt_i}}, \mathbf{I}_{kj})$ with inner loop of t_i and outer loop for $i = 1, \cdots, N$ where $\mathbf{I}_{kj} = (I_{kj}, I_{kj1}, \ldots, I_{kj3}, I_{kj} \cdot gen, \ldots, I_{kj3} \cdot gen, I_{kj} \cdot age, \ldots, I_{kj3} \cdot age)$ in our case and $\boldsymbol{\psi}^T$ is corresponding vector of parameters. Obtain $\widehat{\xi_{jt_i}}$.

Likelihood function

The likelihood function of purchase history of consumer i is written as

$$L_i = \prod_{t_i=1}^{T_i} \int \left\{ \prod_{j=0}^{J} [\Pr_{ijt_i}]^{y_{ijt_i}} \times f(\xi_{jt_i} | \eta_{jt_i}) \times f(\eta_{jt_i}) \right\} d\xi_{jt_i} \tag{3.3.3}$$

where y_{ijt_i} is an indicator function taking unity if consumer i chooses brand j at time t_i and 0 otherwise, $f(\eta_{jt_i})$ is the density function of η_{jt_i}, and $f(\xi_{jt_i}|\eta_{jt_i})$ is the conditional density of ξ_{jt_i}.

We use the latent class model of Kamakura & Russell (1989) under which the likelihood (3.3.3) for consumer i is replaced with $L_i(S_i = s)$, the likelihood of consumer i belonging to the segment s or $S_i = s$, as

$$L = \prod_{i=1}^{I} \left\{ \prod_{s=1}^{S} L_i(S_i = s) \times \mathrm{Pr}_i(s) \right\} \tag{3.3.4}$$

where S is the number of supports of the discrete mass points (i.e., segments) and $\mathrm{Pr}_i(s)$ is the membership probability of consumer i belonging to segment s. The parameters $\boldsymbol{\beta}_s$, α_s, and SD_s are estimated by maximizing this likelihood function.

3.3.2 Supply-Side Estimation

On the supply-side, we first compute the margins under different competitive assumptions conditional on the estimated demand-side parameters. Then we estimate the parameters in marginal cost equation which is written as

$$mc_{jt} = w_{j0} + \mathbf{input}_{jt} \cdot \boldsymbol{w}_r$$

where w_{j0} is brand-specific intercept term, $\mathbf{input}_{jt}$ is vector of observable cost shifters with R elements, and $\boldsymbol{w}_r = (w_1, \cdots, w_R)^T$ is vector of coefficients. We utilize the following equation

$$p_{jt} - \widehat{CMR}_{jt} - \widehat{CMM}_{jt} = mc_{jt} + \varepsilon_{jt} \tag{3.3.5}$$

to estimate w_{j0} and $\boldsymbol{w}_r$, where $\widehat{CMR}_{jt}$ and $\widehat{CMM}_{jt}$ are computed margin for retailer and computed margin for manufacturer for brand j at time t respectively. Assuming error terms ε_{jt} follow the normal distribution with mean zero and finite variance (which we will estimate), the right-hand side of the following equation

$$\varepsilon_{jt} = p_{jt} - \widehat{CMR}_{jt} - \widehat{CMM}_{jt} - w_{j0} - \mathbf{input}_{jt} \cdot \boldsymbol{w}_r \tag{3.3.6}$$

also follows the normal distribution. Then we set up the likelihood function on the supply-side as

$$\prod_{t=1}^{T} \prod_{j=1}^{J} g(\varepsilon_{jt}) \tag{3.3.7}$$

where $g(\cdot)$ is the marginal density of ε_{jt}. Given those specifications, the algorithm to estimate w_{j0} and $\boldsymbol{w}_r$ is as follows:

Step 0: Set $ite = 0$. Set the initial values of demand parameters $\boldsymbol{\beta}_s^{(ite)}$, $\alpha_s^{(ite)}$, and $\boldsymbol{\psi}^{(ite)}$ for $s = 1, \cdots, S$ where superscript stands for iteration and subscript stands for segment. Set $\mathrm{Pr}_i(s)^{(ite)}$, the initial value of membership probability of consumer i for segment s, to be $1/S$ for $i = 1, \cdots, N$ and $s = 1, \cdots, S$.

Step 1: Set $i = 1$, $s = 1$, and $t_i = 1$.

Step 2: Calculate utilities $\widehat{v_{ijt_i}}$ for consumer i for brands $j = 1, \cdots, J$ using $\widehat{p_{jt}}$, $\boldsymbol{\beta}_s^{(ite)}$, $\alpha_s^{(ite)}$, $\boldsymbol{\psi}^{(ite)}$, and $\widehat{\xi_{jt}}$.

Step 3: Calculate the logit probabilities as in (3.2.1) using $\widehat{v_{ijt_i}}$ obtained in **Step 2** for $j = 1, \cdots, J$. Denote it as $\widehat{P_{ijt_i}}$.

Step 4: Calculate L_{it_i} as

$$L_{it_i} = \prod_{j=1}^{J} \left[\widehat{P_{ijt_i}} \right]^{y_{ijt_i}}.$$

Step 5: Increase t_i by 1. If $t_i < T_i$, go back to **Step 2**. Else calculate

$$L_i = \prod_{t_i=1}^{T_i} L_{it_i}.$$

This is the contribution of consumer i to the likelihood function.

Step 6: Increase i by 1. If $i < N$, go back to **Step 2**.

Step 7: Increase s by 1. If $s < S$, go back to **Step 2**.

Step 8: Calculate the likelihood L as in (3.3.4).

Step 9: Update demand parameters by maximizing the likelihood function L in equation (3.3.4). Denote them as $\boldsymbol{\beta}_s^{(ite+1)}$, $\alpha_s^{(ite+1)}$, $\boldsymbol{\psi}^{(ite+1)}$, and $\Pr_i(s)^{(ite+1)}$. Set $\Pr_i(s)^{(ite+1)} = \sum_{i=1}^{N} \Pr_i(s)^{(ite)}/N$ for $i = 1, \cdots, N$ and $s = 1, \cdots, S$. Let $ite = ite + 1$.

Step 10: Repeat **Step 2** through **Step 9** until the likelihood function L converges.

Step 11: Calculate the retailer and manufacturers margins given the estimated demand parameters. Note that margins are different depending on the type of games employed.

Step 12: Estimate marginal cost parameters by maximizing the likelihood in (3.3.7) using (3.3.6) and calculate residual η_{jt}.

Step 13: Calculate the supply-side likelihood as

$$\prod_{t=1}^{T}\prod_{j=1}^{J} g(\eta_{jt}).$$

3.4 Data

We use daily scanner-panel data on the yogurt category sales between January 2007 and December 2008 in an anonymous retail chain located in western Tokyo, Japan.[23] This market is suitable for our analysis because it already had two well-established brands with a special feature using newly found bacilli[24] and a power shift from manufacturers to retailers was said to already have been observed in the Japanese food industry (Kim, 2010). Between two types of yogurt−box type and snack type−we choose the latter for our empirical analysis as the former did not have a brand with a special feature.

We choose the seven top-selling brands for our empirical analysis.[25] Table 3.4.1 summarizes the data of the brands. Brand 5 and its low-fat version, brand 6 are the brands with a special feature which had existed during the observation period.

After choosing consumers that only purchased the selected brands at least twice during the period, 183 consumers who made 15,194 shopping trips

---

[23]This work is supported by Grant-in-Aid for Scientific Research (A)21243030. The original CCL-CAFE data are provided by CUSTOMER COMMUNICATIONS, Ltd, through the introduction from Prof. Tadahiko Sato of the Graduate School of Business Sciences of the University of Tsukuba, Tokyo. We thank them for the generous offer.

[24]We cannot disclose the name of the bacilli as it would identify the product.

[25]The combined market share of the seven selected brands is 44.5%, excluding box-type yogurt. The number is relatively small because there existed 300 brands during the study period and market share of each brand was small. We chose the top-selling seven brands because the minor brands had many missing daily price information.

Table 3.4.1: Summary statistic of the brands

	Average Price (yen per gram)	Manufacturer ID	Market Share	Raw Milk Usage	Fat Level	Agar Usage	Fat (g/100g)	Sugar (g/100g)
Brand 1	0.45	1	4.0%	No	Middle	Yes	2.5	7.8
Brand 2	0.50	2	10.4%	Partial	Middle	Yes	2.1	14.6
Brand 3	0.51	3	2.5%	All	High	Yes	4.1	14.9
Brand 4	0.48	4	3.7%	Partial	Low	No	1.8	15.2
Brand 5	1.13	4	10.9%	Partial	Middle	No	3.0	9.7
Brand 6	1.13	4	4.8%	Partial	Low	No	1.4	9.2
Brand 7	0.86	5	8.2%	Partial	low	No	1.9	13.4

and purchased 2,550 units of yogurt remained. The available demographic variables in our data are age and gender. The average age of the consumers is 59.4 (with the standard deviation of 19.6) and 76.5% of them are female.

In addition, we collected weekly data on ingredients (raw milk prices, cream price indexes, and international sugar prices), labor wages for the four prefectures where the seven selected brands had been produced, and international oil price during the study period. Because all data are only available on monthly basis, we transformed them into weekly data by applying the linear filtering process employed by Slade (1995) as

$$W_t = 0.25W_{t-1} + 0.50W_t + 0.25W_{t+1}$$

where W_t in week t is the input price in the corresponding month (Besanko et al., 1998). As for the international price of sugar, we multiplied it to the amount of sugar each brand contains. Also, since cream is mixed in yogurt to increase fat content, we multiplied price index of cream to the amount of fat each brand contains. We used raw milk prices as they were, and we took log for labor wage cost and international oil prices. We also included manufacturer dummy variables to infer the firm-specific cost structure with manufacturer 1 as the basis.

The attributes we used for similarity index are: "Raw Milk Usage" (the amount of raw milk contained, 3 levels), "Fat Level" (the amount of fat, 3 levels), and "Agar Usage" (whether it uses agar or not, 2 levels).[26]

3.5 Empirical Results

3.5.1 Demand-Side Results

We find that the latent class model with four segments is optimal.[27] Table 3.5.1 presents parameter estimates of the demand model (with standard errors in parentheses); "Brand" entries represent the brand-specific intercepts relative to the outside option; presented under "Demographics" are estimated parameters for SD_s[28]; "Agar Usage" entry is the estimate of importance weight for this attribute in calculating the attribute similarity index[29]; estimated segment sizes are reported below price coefficients.

Overall, our findings are economically consistent. Segments are heterogeneous in responsiveness to marketing variables and coefficient of prices are negative. Although the estimates of demographics are generally non-significant, we find some patterns for each segment. For example, segment 2 is characterized by variety-seeking behavior regardless of the age and gender. Specifically in segment 2, a male aged 94 (the maximum age in the

[26] Agar is used to produce so-called "hard-type" yogurt.

[27] We increased the number of segments to minimize AIC. Although AIC was lower for the five-segments model than for the four-segments model, we chose the latter because the size of fifth segment became 0.7% in the five-segments model, as targeting a segment size less than 0.7% out of a sample size of 183 does not make much sense.

[28] We only presented importance weight estimates for segment 1 because we used these estimates for the models with greater number of segments.

[29] We found "Raw Milk Usage" and "Fat Level" to be non-significant.

Table 3.5.1: Parameter estimates of the proposed model

Variables	Segment 1	Segment 2	Segment 3	Segment 4
Brand 1	-0.64	-1.30**	-3.97**	0.39
	(0.591)	(0.331)	(0.472)	(0.595)
Brand 2	0.67	4.00	1.92*	-1.85**
	(1.029)	(3.224)	(0.868)	(0.366)
Brand 3	-0.76	-0.41	1.01	2.68*
	(0.505)	(0.381)	(0.551)	(1.200)
Brand 4	0.40	2.03	-8.14**	5.57
	(0.937)	(2.017)	(1.102)	(3.754)
Brand 5	8.79**	2.45**	6.84**	0.94**
	(1.683)	(0.060)	(2.143)	(0.011)
Brand 6	7.18**	8.98**	-5.63**	-0.40**
	(0.785)	(1.400)	(1.493)	(0.006)
Brand 7	4.04**	1.92**	-1.49**	7.60**
	(0.986)	(0.327)	(0.335)	(1.590)
Price Coefficient	-11.27**	-12.59**	-10.03**	-14.33**
	(2.227)	(2.391)	(3.016)	(2.250)
Segment Size	76.5%	12.4%	3.7%	7.5%
Demographics				
Intercept	-0.32	-6.84*	-9.43**	-3.77
	(1.333)	(2.887)	(2.015)	(1.999)
Male Dummy	0.33	0.02	8.81**	1.35
	(0.769)	(1.395)	(0.590)	(0.994)
Age (logged)	0.19	2.26	5.91	1.59
	(5.537)	(12.120)	(8.285)	(8.183)
The Attribute Similarity Index				
Agar Usage	0.36**			
	(0.078)			
Number of Parameters	47			
Number of Observations	15,194			
Log-likelihood	-7,324.7			

** Significant at 1% level.

* Significant at 5% level.

Table 3.5.2: Fits across games in the forward-looking model

Manufacturer-Retailer Interaction	Manufacturer Interaction	Log-likelihood
Manufacturer Stackelberg	Bertrand competition	5.4
	Collusion	36.3
Vertical Nash	Bertrand competition	23.7
	Collusion	49.0
Retailer Stackelberg	Bertrand competition	-174.6
	Collusion	-173.4

sample) would have SD of $-6.84 + 0.02 + 2.26 * \log(94) = -2.36$. All the other consumers in this segment would have SD lower than -2.36 and thus would exhibit variety-seeking behavior. In segment 1, males of all ages and females aged more than 48 years have a tendency toward inertia. In segment 3, males of all ages and females aged more than 39 years have a tendency toward inertia. In segment 4, females of all ages and males aged less than 34 years would exhibit variety-seeking behavior.

3.5.2 Supply-Side Results

The fits across games

We calculated the log-likelihood for six games (i.e., Bertrand/Collusion and MS/VN/RS games) to compare the fits. Table 3.5.2 presents the result. We find that VN-Collusion game fits the data best. Thus we report the result of this game in the following.

Margins

Table 3.5.3 reports the estimated margins in the best-fitting model and their standard errors (in parentheses). The standard errors turn out to be very small because the prices of those brands stay fairly constant during the study

Table 3.5.3: Retailer and manufacturer margins under each game

Retail Margin	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6	Brand 7
MS and VN	0.053	0.061	0.048	0.056	0.188	0.219	0.135
	(0.0014)	(0.0009)	(0.0009)	(0.0016)	(0.0013)	(0.0013)	(0.0015)
RS							
Bertrand	0.083	0.103	0.079	0.097	0.329	0.375	0.222
	(0.0016)	(0.0010)	(0.0008)	(0.0025)	(0.0016)	(0.0013)	(0.0018)
Collusion	0.053	0.077	0.064	0.081	0.178	0.243	0.155
	(0.0015)	(0.0019)	(0.0015)	(0.0033)	(0.0010)	(0.0022)	(0.0015)
Manufacturer Margin							
MS							
Bertrand	0.029	0.061	0.045	0.071	0.141	0.181	0.101
	(0.0010)	(0.0014)	(0.0009)	(0.0030)	(0.0011)	(0.0015)	(0.0010)
Collusion	0.053	0.077	0.065	0.083	0.177	0.242	0.155
	(0.0016)	(0.0020)	(0.0018)	(0.0035)	(0.0011)	(0.0023)	(0.0015)
VN and RS							
Bertrand	0.033	0.047	0.034	0.046	0.157	0.170	0.093
	(0.0006)	(0.0007)	(0.0004)	(0.0013)	(0.0008)	(0.0008)	(0.0003)
Collusion	0.053	0.061	0.048	0.056	0.188	0.219	0.135
	(0.0014)	(0.0009)	(0.0009)	(0.0016)	(0.0013)	(0.0013)	(0.0015)

period. We note that brand 5 and brand 6 yield the two largest margins, and brand 6, in particular, yields the highest margins relative to the average retail prices, implying that brands with the aforementioned special feature could indeed earn a large amount of margins. The implications of these results are discussed in section 3.6.

Marginal cost

Table 3.5.4 reports the result for the marginal cost estimation of the best-fitting model. We find that after including the manufacturer dummy variables, all cost variables except for international oil price become non-significant in the marginal cost estimation in the best-fitting model. International oil price affects marginal cost as oil is required for a yogurt-making machine,

Table 3.5.4: Marginal cost estimation of the vertical Nash-collusion forward-looking model

	Estimate	Std.Err	t-value	Pr(> $\|t\|$)
Intercept	0.196	0.082	2.401	0.017
Manufacturer Dummy 2	-0.048	0.021	-2.280	0.023
Manufacturer Dummy 3	-0.037	0.021	-1.778	0.076
Manufacturer Dummy 4	0.163	0.017	9.563	0.000
Manufacturer Dummy 5	0.315	0.021	15.108	0.000
International Oil Price	0.037	0.018	2.015	0.044

refrigeration, air conditioning in yogurt factories, and transportation by refrigerator trucks.

3.6 Conclusion

In this chapter, we derived an RS game extending Che et al. (2007) and showed that MS and RS games stand at opposite extremes whereas the VN game lies in between these two games. We then empirically analyzed Japanese yogurt market under the extended formulation, incorporating heterogeneity among consumers, state dependence in brand choice by the similarity index variable, and firms' forward-looking behavior while correcting for price endogeneity.

We find that the brands with the differentiating feature (i.e., enhancing the health effect of yogurt by newly found bacilli) enable the manufacturer to command larger margins than the other brands, showing that the manufacturer's effort in this direction could be interpreted as rewarding. However, we also find that the power to charge larger margins does not spill over to the other brands of a manufacturer, as the manufacturer's margin on brand 4 is in line with those of the others even though brand 4 is owned by the manufacturer producing brand 5 and brand 6. The fact that the retailer earns the same amount of margins as manufacturers is somewhat counter-intuitive given the conventional wisdom of the power shift from manufacturers toward retailers; yet the fact that the retailer still earns large margins on these brands might indicate such a power shift.

One of the limitations of this study is the assumption of the monopolistic retailer, as retail competition is shown to affect the relationship between a retailer and manufacturers (Raju & Zhang, 2005). We leave this issue for future research.

Chapter 4

Incorporating the Retailer Stackelberg Game in Nash Bargaining Model

4.1 Introduction

The relationship between manufacturers and retailers has been attracting great attention from marketing literature. One of the main interests of the literature has been to gain insight on the strategic interaction of these firms such as a degree of coordination or split of profit in a channel (Besanko et al., 1998; Nevo, 2001; Sudhir, 2001; Villas-Boas & Zhao, 2005). Some studies step further to examine the effectiveness of specific tactics while incorporating such interaction. For example, Che et al. (2007) analyzes a breakfast cereal market in U.S. and shows that suppliers in midst of strategic interaction also look forward in demand and make pricing decisions accordingly. Given the purported increase in power of some of the retailers due to their increased size and willingness to introduce private labels aggressively as well as willingness to invest in sophisticated information systems regarding consumers, we extended the framework of Che et al. (2007) and derives a mathematical formulation of market-level RS game in analyzing a Japanese yogurt market in Chapter 3.

The research mentioned above mainly portrays the overall picture of the market. Recent research instead formulates wholesale prices as the outcome of the bargaining in each manufacturer-retailer pair to investigate the power structure in the market more closely and obtain implications via Nash bargaining model. For example, a theoretical study of Dukes et al. (2006) employing Nash bargaining model in two manufacturers−two retailers setting (one powerful retailer and the other less powerful retailer) finds that the presence of a dominant retailer, in fact, benefits the manufacturers by increasing channel margins via operational efficiency. The empirical work of Misra & Mohanty (2008), which applies the generalized Nash bargaining model to the real data in the context of logit demand model with a single retailer and multiple manufacturers setup, shows that demand-supply structural model incorporating Nash bargaining game outperforms the market-level game-theoretic models in terms of data fit, providing the empirical support to the model. Draganska et al. (2010) extends their work by allowing multiple retailers[30] and empirically shows that, in a German coffee market, the relative bargaining power between manufacturers and retailers is different across each manufacturer-retailer pair rather than uniform across pairs as expected. These studies contrast to previous models in that they enable researchers to calibrate relative bargaining power between each manufacturer-retailer pair,

[30]Though Dukes et al. (2006) considers multiple retailers setup as well, their model employs simple linear demand where demand decreases linearly as price of the goods increases.

rather than overall power structure in the market thus providing more detailed insight into the market.

While models of Dukes et al. (2006), Misra & Mohanty (2008), and Draganska et al. (2010) successfully portray the behavior of market players to a certain extent, what has been missing in their papers is the RS formulation. Given aforementioned power increase of retailers, it is possible that some of the retailers bargain aggressively with manufacturers over wholesale prices and thus enjoy a larger profit. The limitation of the preceding papers is that such behavior cannot be modeled because retailers are implicitly assumed to passively react to whatever manufacturers offer. Therefore in this chapter, we derive how to accommodate the RS game in the Nash bargaining model framework so as to model powerful retailers and illustrate it empirically using Japanese canned tuna as an example.

The rest of this chapter is organized as follows. Section 4.2 describes our data. Section 4.3 reviews the previous approach in this literature and presents our model and estimation procedure. Section 4.4 presents empirical results and discuss their implications. Section 4.5 concludes our results.

4.2 Data

We use daily scanner-panel data on canned tuna between October 2008 and December 2009 in an anonymous retail chain located in western Tokyo, Japan. This market is suitable for our analysis because there existed a private brand in this category during the research period. By using the category with a private brand, we can illustrate how to deal with it since a private brand may not be the subject of bargaining. We choose the six top-selling brands

Table 4.2.1: Summary statistic of the brands

	Average Price (yen per gram)	Manufacturer ID	Market Share
Brand 1	1.08	1	41.8%
Brand 2	1.92	2	11.6%
Brand 3	1.29	3	21.6%
Brand 4	1.50	3	7.4%
Brand 5	1.52	4	8.2%
Brand 6	1.36	4	8.5%

for our empirical analysis.[31] After choosing consumers who purchased out of six brands more than twice during the period, 281 consumers who made 8,479 purchases remained. Table 4.2.1 summarizes the data on the brands. The unit of price is Japanese yen per one gram. We note that brand 1 is the private brand. To estimate marginal cost, we collected weekly data on ingredients (wholesale prices of frozen tuna, big-eye, yellow-fin tuna, blue-fin tuna, and southern tuna in Metropolitan Central Wholesale Market), international oil prices, and heavy-oil prices during the study period.

4.3 The Model

In this section, we present the demand and supply models as well as the estimation procedure.

4.3.1 Demand-Side Specification

We use a multinomial logit model to estimate consumer's brand choice behavior employing the latent class model. The indirect utility of consumer

[31]The combined market share of the six selected brands is 99.1%.

i $(i = 1, \ldots, N)$ for brand j $(j = 1, \ldots, J)$ on shopping occasion t_i $(t_i = 1, \ldots, T_i)$ is defined as

$$v_{ijt_i} = \boldsymbol{x}_{jt_i}\boldsymbol{\beta}_s + I_{kj}\gamma_s + \xi_{jt_i} + \epsilon_{ijt_i} \tag{4.3.1}$$

where $\boldsymbol{x}_{jt_i}$ is vector of brand dummy variables and the retail price of brand j that consumer i faces on shopping occasion t_i, and I_{kj} is the last choice variable of brand j relative to the previously purchased brand k which takes unity if $j = k$ and zero else. The term $\boldsymbol{\beta}_s$ is the corresponding vector of parameters for consumers in segment s and γ_s is the parameter for the last choice variable. A positive (negative) value of γ_s reveals inertial (variety-seeking) behavior, that is, a brand consumption experience increases (decreases) the probability of repurchasing the brand on the consecutive purchasing occasion. The term ξ_{jt_i} is a composite measure of unobserved (to the researcher) demand characteristics that affect all consumers commonly and ϵ_{ijt_i} are errors distributed i.i.d. Gumbel. The outside option $(j = 0)$ is specified as determinant part of the utility being zero.

Demand-Side Estimation

To avoid the endogeneity problem between p_{jt} and ξ_{jt}, we employ 2SLS for price as follows:

$$p_{jt} = \kappa_0 + z_{jt} \cdot \kappa_1 + \eta_{jt}$$

where z_{jt} is the instrument for retail price p_{jt}, κ_0 and κ_1 are parameters to be estimated, and η_{jt} is an error term. They are defined for all brands $j = 1, \ldots, J$ and dates $t = 1, \ldots, T$ in the study period. For the instrument, we use the average retail prices of canned tuna in five stores we excluded from the analysis owing to lack of price information.

The likelihood function of the purchase history of consumer i belonging to segment s ($L_{i\in s}$) is given by

$$L_{i\in s} = \prod_{t_i=1}^{T_i} \int \left\{ \prod_{j=0}^{J} \left(\mathrm{Pr}_{ijt_i}^{s}\right)^{y_{ijt_i}} \times f(\xi_{jt_i}|\eta_{jt_i}) \times h(\eta_{jt_i}) \right\} d\xi_{jt_i}$$

where $\mathrm{Pr}_{ijt_i}^{s}$ is the logit purchase probability of consumer i who belongs to segment s choosing brand j on shopping occasion t_i, y_{ijt_i} is the indicator function taking 1 if consumer i chooses brand j at time t_i and 0 otherwise, $f(\xi_{jt_i}|\eta_{jt_i})$ is the conditional density function of ξ_{jt_i} given η_{jt_i}, and $h(\eta_{jt_i})$ is the density function of η_{jt_i}. Then, the demand-side likelihood function is

$$L = \prod_{i=1}^{I} \left\{ \prod_{s=1}^{S} L_{i\in s} \times \mathrm{Pr}_i(s) \right\}$$

where $\mathrm{Pr}_i(s)$ is consumer i's probability of membership in segment s.

4.3.2 Supply-Side Specification

In modeling supply-side behavior, we make the following assumptions. First, we assume that a retailer is a local monopolist.[32] The second assumption is that, though some brands are produced by the same manufacturer, each brand has its own bargaining power with respect to the retailer. However, in setting the wholesale price of brands, the manufacturers account for their impact on the other brands they own. The third assumption is that all manufacturers and a retailer have rational expectations and can anticipate ultimate equilibrium outcomes.

In this section, we first review the general framework of the Nash bargaining game. Then we review the specification of Draganska et al. (2010) and describe how an RS game can be accommodated to that framework.

[32] Though we analyze one-store data, retail competition is in fact reflected in that data as the retailers compete in reality. If the RS behavior is found in our data, it would empirically support our claim that the retailer acts aggressively despite the retail competition.

The Nash bargaining model

Let us suppose the manufacturer and the retailer bargain over brand j. Then the Nash bargaining solution can be obtained by maximizing

$$f(\pi_{jt}^r, \pi_{jt}^w) = (\pi_{jt}^r - d_{jt}^r)^{\lambda_j}(\pi_{jt}^w - d_{jt}^w)^{1-\lambda_j} \tag{4.3.2}$$

where π_{jt}^r and π_{jt}^w are the retailer and the manufacturer profits from brand j respectively, d_{jt}^r and d_{jt}^w are the retailer and the manufacturer disagreement payoffs that are obtained if the negotiation fails, and λ is the parameter representing the retailer's bargaining power relative to the manufacturer which takes between 0 and 1 inclusive.[33] Sometimes the term λ is called "bargaining power" and it depends on numerous factors such as negotiation skill of the agent, risk tolerance level, and patience of a party that affects the outcome of the bargaining (Dukes et al., 2006; Draganska et al., 2010). In contrast, "bargaining position" refers to the relative strength of the party which is already realized before a negotiation starts (Dukes et al., 2006). Specifically, it is the value of a party's outside option; the more it loses if the negotiation fails, the weaker the party's bargaining position. Both bargaining power and bargaining position will affect the total channel profit and their split.

How the wholesale price is determined in the Nash bargaining model

We review how the wholesale price is determined in the generalized Nash bargaining model as proposed in Draganska et al. (2010) in the following. One notable assumption of their paper is fixed retail price assumption; their justification of that assumption is that retail prices cannot be contracted upon (known as retail price unobservability). Notice that, if this formulation

[33] See Appendix A.10 for its derivation.

is adopted, manufacturers can exercise price control on a product without worrying about the decrease in its market share. In this sense, this formulation presupposes MS game or retail prices would not be affected by the change in wholesale prices.[34]

Taking partial derivative of (4.3.2) by wholesale price of brand j and setting it zero, they obtain the FOC of (4.3.2) as

$$\lambda_j(\pi^r_{jt} - d^r_{jt})^{\lambda_j - 1}\frac{\partial \pi^r_{jt}}{\partial w_{jt}}(\pi^w_{jt} - d^w_{jt})^{1-\lambda_j} + (\pi^r_{jt} - d^r_{jt})^{\lambda_j}(1-\lambda_j)(\pi^w_{jt} - d^w_{jt})^{-\lambda_j}\frac{\partial \pi^w_{jt}}{\partial w_{jt}} = 0$$

or

$$\lambda_j(\pi^w_{jt} - d^w_{jt})\frac{\partial \pi^r_{jt}}{\partial w_{jt}} + (1-\lambda_j)(\pi^r_{jt} - d^r_{jt})\frac{\partial \pi^w_{jt}}{\partial w_{jt}} = 0. \tag{4.3.3}$$

Note that the wholesale margin is a function of wholesale price, so $\pi^w_{jt} = \pi^w_{jt}(w_{jt})$. Although the retail prices are not affected by the wholesale price, the retail margins are $\pi^r_{jt} = \pi^r_{jt}(w_{jt})$, so π^r_{jt} is also a function of w_{jt}. The retailer and the manufacturer profits of brand j are respectively defined as

$$\pi^r_{jt} = (p_{jt} - w_{jt})S_{jt}M \tag{4.3.4}$$

and

$$\pi^w_{jt} = (w_{jt} - mc_{jt})S_{jt}M \tag{4.3.5}$$

where S_{jt}, w_{jt}, and mc_{jt} are the market share, the wholesale price, and the marginal cost of brand j at time t respectively and M is the market size. Abbreviate $S_{jt}(p_{jt}) = S_{jt}(p_{jt}(w_{jt})) = S_{jt}(w_{jt})$ as S_{jt}. Then FOCs of the profit functions (4.3.4) and (4.3.5) with respect to wholesale price of brand j are

$$\frac{\partial \pi^r_{jt}}{\partial w_{jt}} = (p_{jt} - w_{jt})\frac{\partial S_{jt}}{\partial w_{jt}}M + \left(\frac{\partial p_{jt}}{\partial w_{jt}} - 1\right)S_{jt}M \tag{4.3.6}$$

[34] This assumption will be replaced in our formulation as will be explained.

and

$$\frac{\partial \pi_{jt}^{w}}{\partial w_{jt}} = (w_{jt} - mc_{jt})\frac{\partial S_{jt}}{\partial w_{jt}}M + S_{jt}M \tag{4.3.7}$$

respectively.[35] However, since

$$\frac{\partial S_{jt}}{\partial w_{jt}} = \sum_{h=1}^{J} \frac{\partial S_{jt}}{\partial p_{ht}} \cdot \frac{\partial p_{ht}}{\partial w_{jt}} \tag{4.3.8}$$

and $\partial p_{ht}/\partial w_{jt} = 0$ for all h and j from the fixed retail price assumption, FOCs in (4.3.6) and (4.3.7) reduce to

$$\frac{\partial \pi_{jt}^{r}}{\partial w_{jt}} = -S_{jt}M \tag{4.3.9}$$

and

$$\frac{\partial \pi_{jt}^{w}}{\partial w_{jt}} = S_{jt}M. \tag{4.3.10}$$

The disagreement payoffs of the retailer and the manufacturer are respectively defined as

$$d_{jt}^{r} = \sum_{k \in J\setminus\{j\}} (p_{kt} - w_{kt})\Delta s_{kt}^{-j}(p)M \tag{4.3.11}$$

$$d_{jt}^{w} = \sum_{k \in J_f\setminus\{j\}} (p_{kt} - w_{kt})\Delta s_{kt}^{-j}(p)M \tag{4.3.12}$$

where J_f is the set of products sold by manufacturer. The term $\Delta s_{kt}^{-j}(p)$ is the difference in market share of brand k when brand j is available and when not, which is defined as

$$\Delta s_{kt}^{-j}(p) = \int \left[\frac{\exp(v_{kt}^{s})}{1 + \sum_{l \in \Omega\setminus\{j\}} \exp(v_{lt}^{s})} - \frac{\exp(v_{kt}^{s})}{1 + \sum_{l \in \Omega} \exp(v_{lt}^{s})} \right] dF(s)$$

where v_{jt}^{s} is the deterministic part of the utility of brand j for consumers in segment s and $F(s)$ is the distribution function of segment s, which is discrete in our specification.

[35] Note that it is assumed $\partial mc_{kt}/\partial w_{lt} = 0$ for all $k, l = 1, \ldots, J$, as wholesale price would not affect marginal cost in general. Similarly, $\partial w_{jt}/\partial w_{kt} = 0$.

Substituting (4.3.9), (4.3.10), (4.3.11), and (4.3.12) to (4.3.3) and rearranging, they have the optimal manufacturer's margin in relation to the retailer margin as

$$\pi_{jt}^{w} - d_{jt}^{w} = \frac{1-\lambda_j}{\lambda_j}(\pi_{jt}^{r} - d_{jt}^{r}). \tag{4.3.13}$$

Denote $(p_{kt} - w_{kt})$ and $(w_{kt} - mc_{kt})$ as m_{kt}^{r} and m_{kt}^{w} respectively. From (4.3.4), (4.3.5), (4.3.11), and (4.3.12), (4.3.13) can be rewritten as

$$m_{jt}^{w} S_{jt} - \sum_{k \in J_f \backslash \{j\}} m_{kt}^{r} \Delta s_{kt}^{-j}(p) = \frac{\lambda}{1-\lambda}\left[m_{jt}^{r} S_{jt} - \sum_{k \in J \backslash \{j\}} m_{kt}^{r} \Delta s_{kt}^{-j}(p) \right]. \tag{4.3.14}$$

Stacking (4.3.14) for all brands and rearranging yields

$$\begin{aligned}
&\mathbf{\Omega} \cdot * \begin{bmatrix} s_1 & -\Delta s_2^{-1} & \cdots & -\Delta s_J^{-1} \\ -\Delta s_1^{-2} & s_2 & \cdots & -\Delta s_J^{-2} \\ \vdots & \vdots & \ddots & \vdots \\ -\Delta s_1^{-J} & -\Delta s_2^{-J} & \cdots & s_J \end{bmatrix} \begin{pmatrix} m_1^w \\ m_2^w \\ \vdots \\ m_J^w \end{pmatrix} \\
&= \begin{bmatrix} \frac{1-\lambda_1}{\lambda_1} s_1 & -\frac{1-\lambda_1}{\lambda_1}\Delta s_2^{-1} & \cdots & -\frac{1-\lambda_1}{\lambda_1}\Delta s_J^{-1} \\ -\frac{1-\lambda_2}{\lambda_2}\Delta s_1^{-2} & \frac{1-\lambda_2}{\lambda_2} s_2 & \cdots & -\frac{1-\lambda_2}{\lambda_2}\Delta s_J^{-2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1-\lambda_J}{\lambda_J}\Delta s_1^{-J} & -\frac{1-\lambda_J}{\lambda_J}\Delta s_2^{-J} & \cdots & \frac{1-\lambda_J}{\lambda_J} s_J \end{bmatrix} \begin{pmatrix} m_1^r \\ m_2^r \\ \vdots \\ m_J^r \end{pmatrix}
\end{aligned} \tag{4.3.15}$$

dropping t for notational convenience.

Incorporating the RS game to the generalized Nash bargaining model

We explain how the RS game can be incorporated into the generalized Nash bargaining model in the following. In the RS game, a retailer sets its margin anticipating the behavior of manufacturers and then manufacturers follow. Accordingly, we would not treat retail price being fixed unlike in Draganska et al. (2010). Even under such circumstances, manufacturers would be able to know the retail prices of their brands sooner or later and thus would take them into account in the next negotiation. In the RS formulation, therefore, a manufacturer cannot demand as much as it would under the fixed retail price assumption since the increase in wholesale price would, in turn, increase its retail price and is likely to lower its market share.

Specifically in the RS formulation, the retail margin is determined beforehand or $\partial(p_{jt} - w_{jt})/\partial w_{kt} = 0$ for all j and k, or $\partial p_{jt}/\partial w_{kt}$ is 1 if $j = k$ and 0 else. As a result, (4.3.8) reduces to

$$\frac{\partial S_{jt}}{\partial w_{jt}} = \frac{\partial S_{jt}}{\partial p_{jt}}. \tag{4.3.16}$$

By $\partial p_{jt}/\partial w_{jt} = 1$ and (4.3.16), equations (4.3.6) and (4.3.7) become

$$\frac{\partial \pi^r_{jt}}{\partial w_{jt}} = (p_{jt} - w_{jt})\frac{\partial S_{jt}}{\partial p_{jt}}M, \tag{4.3.17}$$

and

$$\frac{\partial \pi^w_{jt}}{\partial w_{jt}} = (w_{jt} - mc_{jt})\frac{\partial S_{jt}}{\partial p_{jt}}M + S_{jt}M, \tag{4.3.18}$$

respectively. Notice, in (4.3.18), manufacturers would not be as well-off as they would be under (4.3.10) in increasing wholesale prices because the first term of (4.3.18) on the right-hand side is negative. Substituting (4.3.17) and (4.3.18) to (4.3.3) yields

$$\lambda_j(\pi^w_{jt} - d^w_{jt})m^r_{kt}\Phi_{jt} + (1 - \lambda_j)(\pi^r_{jt} - d^r_{jt})\left(m^w_{kt}\Phi_{jt} + S_{jt}\right) = 0, \tag{4.3.19}$$

where $\Phi_{jt} = \partial S_{jt}/\partial p_{jt}$. Moving the second term on the left of (4.3.19) to the right-hand side, substituting (4.3.11) and (4.3.12), dividing both sides by λ_j

and stacking for $j = 1, \ldots, J$, we have the set of equations in matrix form as

$$\mathbf{\Omega} \cdot * \begin{bmatrix} s_1 m_1^r \Phi_1 & -\Delta s_2^{-1} m_1^r \Phi_1 & \ldots & -\Delta s_J^{-1} m_1^r \Phi_1 \\ -\Delta s_1^{-2} m_2^r \Phi_2 & s_2 m_2^r \Phi_2 & \ldots & -\Delta s_J^{-2} m_2^r \Phi_2 \\ -\Delta s_1^{-3} m_3^r \Phi_3 & -\Delta s_2^{-3} m_3^r \Phi_3 & \ldots & -\Delta s_2^{-3} m_3^r \Phi_3 \\ \vdots & \vdots & \ddots & \vdots \\ -\Delta s_1^{-J} m_J^r \Phi_J & -\Delta s_2^{-J} m_J^r \Phi_J & \ldots & s_J m_J^r \Phi_J \end{bmatrix} \begin{pmatrix} m_1^w \\ m_2^w \\ \vdots \\ m_J^w \end{pmatrix} =$$

$$\begin{bmatrix} \frac{\lambda_1 - 1}{\lambda_1} s_1 (m_1^w \Phi_1 + S_1) & -\frac{\lambda_1 - 1}{\lambda_1} \Delta s_2^{-1} (m_1^w \Phi_1 + S_1) & \cdots & -\frac{\lambda_1 - 1}{\lambda_1} \Delta s_J^{-1} (m_1^w \Phi_1 + S_1) \\ -\frac{\lambda_2 - 1}{\lambda_2} \Delta s_1^{-2} (m_2^w \Phi_2 + S_2) & \frac{\lambda_2 - 1}{\lambda_2} s_2 (m_2^w \Phi_2 + S_2) & \cdots & -\frac{\lambda_2 - 1}{\lambda_2} \Delta s_J^{-2} (m_2^w \Phi_2 + S_2) \\ -\frac{\lambda_3 - 1}{\lambda_3} \Delta s_1^{-3} (m_3^w \Phi_3 + S_3) & -\frac{\lambda_3 - 1}{\lambda_3} \Delta s_2^{-3} (m_3^w \Phi_3 + S_3) & \cdots & -\frac{\lambda_3 - 1}{\lambda_3} \Delta s_J^{-3} (m_3^w \Phi_3 + S_3) \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\lambda_J - 1}{\lambda_J} \Delta s_1^{-J} (m_J^w \Phi_J + S_J) & -\frac{\lambda_J - 1}{\lambda_J} \Delta s_2^{-J} (m_J^w \Phi_J + S_J) & \cdots & \frac{\lambda_J - 1}{\lambda_J} s_J (m_J^w \Phi_J + S_J) \end{bmatrix}$$

$$\cdot \begin{pmatrix} m_1^r \\ m_2^r \\ \vdots \\ m_J^r \end{pmatrix} \tag{4.3.20}$$

dropping t for convenience where $\mathbf{\Omega}$ is a $J \times J$ ownership matrix whose (j, k) element, denoted as Ω_{jk}, is an indicator variable taking unity if brands j and k are made by the same manufacturer and 0 otherwise. Notice that in (4.3.15), the manufacturers' and the retailer's margins are clearly separated, while (4.3.20) contains retailer margin on the left-hand side and manufacturer margin on the right-hand side as well. This is because only the total profit of each party (i.e., $(\pi_{jt}^r - d_{jt}^r)$ and $(\pi_{jt}^w - d_{jt}^w)$) affects margins in (4.3.15), while the rate of change in market shares with respect to retail prices (Φ_{jt}), market shares (S_{jt}), and retailer margin affect manufacturer margin in (4.3.20).

The retailer margin

As the equation (4.3.20) only relates manufacturer margins to retailer margin, we still need to derive the retailer margins. There are two distinct approaches doing so. One is to derive from the model by Draganska et al. (2010); the other is to derive from the RS game as in Chapter 3. These two specifications represent two different behavior of the retailer toward manufacturers. In the model of Draganska et al. (2010), the retailer acts naively and does not attempt to control wholesale prices in its favor but just accepts whatever manufacturers offer. On the other hand, in the RS game, the retailer acts aggressively and sets retail prices possibly in its favor at the expense of manufacturers' margins.

Retailer margins

We know from Chapter 3 that the retailer margin in the MS and VN games is

$$(\boldsymbol{p}_t - \boldsymbol{w}_t) = [\boldsymbol{\Phi}_t]^{-1}\boldsymbol{S}_t$$

where

$$\boldsymbol{\Phi}_t = \begin{bmatrix} \frac{\partial S_{1t}}{\partial p_{1t}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{1t}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{1t}}{\partial p_{Jt}} & \cdots & \frac{\partial S_{Jt}}{\partial p_{Jt}} \end{bmatrix}.$$

We also know that retailer margin in the RS game when manufacturers compete in Bertrand manner is

$$(\boldsymbol{p}_t - \boldsymbol{w}_t) = [\boldsymbol{\Phi}_t]^{-1}[\boldsymbol{I} - \boldsymbol{H}_t^T[[\boldsymbol{\Phi}_t]^T \cdot *\boldsymbol{\Omega}]^{-1}]\boldsymbol{S}_t$$

where $\boldsymbol{H}_t$ is a $J \times J$ matrix whose (l, j) element is

$$\frac{\partial S_{lt}}{\partial p_{jt}} + \sum_{k=1}^{J} \Omega_{lk}(w_{kt} - mc_{kt})\frac{\partial^2 S_{kt}}{\partial p_{lt}\partial p_{jt}}$$

as in (3.2.28). When manufacturers collude, we make $\mathbf{\Omega}$ an matrix with all elements being unity.

Supply-Side Estimation

To obtain the likelihood of the supply-model, we exploit the relationship

$$p_{jt} - \widehat{MR}_{jt} - \widehat{MM}_{jt} = mc_{jt} + \varepsilon_{jt}$$

where $\widehat{MR}_{jt}$ and $\widehat{MM}_{jt}$ are estimated margins of the retailer and the manufacturer on brand j at time t respectively, and ε_{jt} is random error term. If we assume that errors ε_{jt} follow a normal distribution with mean zero and finite variance, the right-hand side of the equation

$$\varepsilon_{jt} = p_{jt} - \widehat{MR}_{jt} - \widehat{MM}_{jt} - mc_{jt} \tag{4.3.21}$$

would also follow the normal distribution. Furthermore, we parametrize the marginal cost as

$$mc_{jt} = \psi_j + \mathbf{input}_{jt}\boldsymbol{\psi} \tag{4.3.22}$$

where ψ_j is the brand-specific intercept term, $\mathbf{input}_{jt}$ is the vector of observable cost shifters, and $\boldsymbol{\psi}$ is the corresponding vector of parameters.[36] Then by substituting (4.3.22) to (4.3.21), we have

$$\varepsilon_{jt} = p_{jt} - \widehat{MR}_{jt} - \widehat{MM}_{jt} - \psi_j - \mathbf{input}_{jt}\boldsymbol{\psi}.$$

The supply-side likelihood function is

$$\prod_{t=1}^{T}\prod_{j=1}^{J} g(\varepsilon_{jt}) \tag{4.3.23}$$

where $g(\cdot)$ is the density function of ε_{jt}.

[36]The cost shifters used in this analysis are listed in section 4.2.

4.4 Empirical Results

We present our empirical results in this section. Remember that brand 1 is a private brand totally under control of this retailer ($\lambda_1 = 1$) and we assume that there does not exist manufacturer margin on the private brand ($m_1^w = 0$). With this assumption, the equation (4.3.20) reduces to

$$\mathbf{\Omega} \cdot * \begin{bmatrix} s_2 m_2^r \Phi_2 & \dots & -\Delta s_J^{-2} m_2^r \Phi_2 \\ \vdots & \ddots & \vdots \\ -\Delta s_2^{-J} m_J^r \Phi_J & \dots & s_J m_J^r \Phi_J \end{bmatrix} \begin{pmatrix} m_2^w \\ \vdots \\ m_J^w \end{pmatrix} =$$

$$\begin{bmatrix} \frac{\lambda_2 - 1}{\lambda_2} s_2 (m_2^w \Phi_2 + S_2) & \cdots & -\frac{\lambda_2 - 1}{\lambda_2} \Delta s_J^{-2} (m_2^w \Phi_2 + S_2) \\ \vdots & \ddots & \vdots \\ -\frac{\lambda_J - 1}{\lambda_J} \Delta s_2^{-J} (m_J^w \Phi_J + S_J) & \cdots & \frac{\lambda_J - 1}{\lambda_J} s_J (m_J^w \Phi_J + S_J) \end{bmatrix} \begin{pmatrix} m_2^r \\ \vdots \\ m_J^r \end{pmatrix}.$$

4.4.1 Demand-Side Results

We find that the latent class model with five segments is optimal.[37] Table 4.4.1 presents parameter estimates of the demand model (with standard errors in parentheses). In Table 4.4.1, "Brand" entries represent the brand-specific intercepts relative to the outside option. The values of these intercepts vary across segments. The coefficients of "Last Choice" variable are all positive and significant, suggesting that consumers all have inertial tendency in this product category in our data.

Preference to a private brand is mixed among segments; segments 3 and segment 5 with the respective estimated segment sizes 40.2% and 24.1% prefer the private brand while segment 4 with the estimated segment size 5.8% does not.

[37]We chose the five-segments model because the size of a segment becomes 0.01% in the six-segments model though AIC supported six-segments model.

Table 4.4.1: Parameter estimates of the demand model

Variables	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Brand 1	2.54	-0.28	4.66**	-5.27**	3.42**
	(1.471)	(0.581)	(1.270)	(0.063)	(0.869)
Brand 2	2.66**	1.30**	7.78**	8.72**	7.28**
	(0.421)	(0.269)	(0.198)	(1.240)	(0.376)
Brand 3	-3.25**	2.16	3.11**	-1.82**	-6.52**
	(0.264)	(1.219)	(0.681)	(0.256)	(0.103)
Brand 4	1.22*	-4.97**	3.10**	3.51**	1.97**
	(0.503)	(0.054)	(0.275)	(0.876)	(0.231)
Brand 5	3.42**	-2.86**	6.99**	1.19**	2.22**
	(1.243)	(0.173)	(1.179)	(0.416)	(0.285)
Brand 6	-5.99**	-6.72**	-5.49**	-4.21**	-5.71**
	(0.605)	(0.514)	(0.785)	(1.349)	(0.807)
Price Coefficient	-5.54*	-5.73**	-8.16**	-6.33*	-7.85**
	(2.659)	(1.648)	(2.367)	(2.698)	(1.268)
Last Choice	8.07**	6.94**	3.50**	4.69**	8.88**
	(0.560)	(0.544)	(0.566)	(0.449)	(0.460)
Segment Sizes	12.4%	18.7%	40.2%	5.8%	24.1%
Number of Parameters	44				
Number of Observations	8,479				
Log-likelihood	-5,180.3				

** Significant at 1% level.

* Significant at 5% level.

Table 4.4.2: Log-likelihood under each game

Manufacturer - Retailer game	Manufacturers' game	Log-likelihood
Retailer Stackelberg	Bertrand	-192.96
	Collusion	-192.94
Manufacturer Stackelberg	Bertrand	-192.86
	Collusion	-192.86
Vertical Nash	Bertrand	-192.89
	Collusion	-192.88
Nash Bargaining Solution (Draganska et al., 2010)	Bertrand	-192.81
	Collusion	-192.84
Nash Bargaining Solution (Proposed)	Bertrand	-192.69
	Collusion	-193.25

Table 4.4.3: The relative bargaining power of the retailer (λ_j)

	Draganska et al. (2010) Model	Proposed Model
Brand 2	0.36	0.98
Brand 3	0.32	0.69
Brand 4	0.49	0.57
Brand 5	0.42	0.64
Brand 6	0.27	0.23

4.4.2 Supply-Side Results

Table 4.4.2 presents supply-side log-likelihood of each model.[38] Though fits across these games are very close, the proposed retailer Stackelberg model with Bertrand competition fits the data best. We note that brand preferences presented in Table 4.4.1 do not seem to be correlated with bargaining power reported on the right-hand side column of Table 4.4.3. Hence bargaining power is not an inherent characteristic of brand, and this finding is consistent

[38]We present log-likelihood because it turns out that all models have the same number of explanatory variables.

Table 4.4.4: Retailer and manufacturer margins under each game

		Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6
Retail Margin							
Manufacturer Stackelberg		0.152	0.158	0.172	0.167	0.150	0.149
		(0.00053)	(0.00052)	(0.00086)	(0.00076)	(0.00088)	(0.00073)
Retailer Stackelberg	Bertrand	0.152	0.285	0.311	0.298	0.271	0.270
		(0.00053)	(0.00066)	(0.00131)	(0.00117)	(0.00123)	(0.00098)
	Collusion	0.152	0.293	0.316	0.309	0.279	0.278
		(0.00053)	(0.00098)	(0.00152)	(0.00133)	(0.00158)	(0.00135)
Manufacturer Margin							
Nash Bargaining Model	Bertrand	0 (assumed)	0.070	0.065	0.041	0.033	0.013
by Draganska et al. (2010)		-	(0.00137)	(0.00149)	(0.00099)	(0.00142)	(0.00088)
	Collusion	0 (assumed)	0.152	0.167	0.161	0.143	0.143
		-	(0.00039)	(0.00065)	(0.00078)	(0.00065)	(0.00047)
Nash Bargaining Model	Bertrand	0 (assumed)	0.077	0.073	0.016	0.011	0.011
by the proposed model		-	(0.00020)	(0.00027)	(0.00014)	(0.00002)	(0.00040)
	Collusion	0 (assumed)	0.056	0.064	0.140	0.214	0.047
		-	(0.00031)	(0.00033)	(0.00037)	(0.00115)	(0.00008)

with that of Draganska et al. (2010). The bargaining power does not seem to be correlated with market share either.

Margins

Table 4.4.4 reports the estimated margins and their standard errors (in parentheses). Since brand 1 is a private brand, we assume that there is no manufacturer's margin on it and its retailer margin is common across games. The standard errors turn out to be very small because the prices of those brands stay fairly constant during the study period. The interpretation of the result is discussed in the next section.

Table 4.4.5: Marginal cost estimation of the proposed model

	Estimate	Std.Err	t-value
Intercept	-0.03	0.24	-0.10
Brand Dummy 1	-0.15	0.03	-5.64
Brand Dummy 2	0.47	0.03	17.88
Brand Dummy 3	-0.14	0.03	-5.21
Brand Dummy 4	0.14	0.03	5.49
Brand Dummy 5	0.15	0.03	5.55
Lagged Heavy Oil Price	0.09	0.03	2.89
Lagged International Oil Price	0.16	0.07	2.40

Marginal cost

Table 4.4.5 reports the marginal cost estimation result for the proposed model. We find that each of price of yellow-fin tuna, one-year-lagged international oil price, and a one-year-lagged heavy-oil price is significant. But when we included all of them, the price of yellow-fin tuna became insignificant, so we dropped it from our marginal cost estimation.

4.5 Conclusion and Discussion

In this chapter, we propose how to incorporate the RS formulation into the Nash bargaining model framework and empirically analyze canned tuna market in Japan. The main finding is that, according to the result of our RS model, the data do support retailer's relative dominance in that retailer margins are much larger than those of manufacturers as shown in Table 4.4.4, though our fit relative to the model proposed by Draganska et al. (2010) is only marginally better as shown in Table 4.4.2. We discuss our empirical results below.

Compared to Draganska et al. (2010), the values of λ_j in our model are higher on average as expected (Table 4.4.3). Moreover, while all brands have more bargaining power with respect to the retailer (i.e., $\lambda_j > 1/2$) in the model of Draganska et al. (2010), the bargaining power lies in the retailer for four out of five brands in our model. The explanation of this difference is that the model of Draganska et al. (2010) presupposes an MS game in deriving retail margins while we assume an RS formulation. If our analysis is correct, the previous models significantly underestimate the power of retailers by presupposing an MS game. The accurate description of the relationship between a retailer and manufacturers is indispensable if firms make decisions accounting for the consequences these models suggest.

Interestingly, brand 2 and brand 3 are actually better-off in terms of margin in our model than in the model of Draganska et al. (2010) (Table 4.4.4). This is counter-intuitive because their respective bargaining power (i.e., $1 - \lambda_j$) are 0.02 and 0.31, which are much smaller than corresponding estimates of 0.64 and 0.68 in the model by Draganska et al. (2010). This could imply that the retailer could act as if it vertically integrates the brands, eliminating or mitigating the double marginalization problem. This finding is consistent with that of Dukes et al. (2006) which finds that the presence of a dominant retailer actually benefits manufacturers. This result could encourage both retailers and manufacturers to cooperate, not only on developing private brands but also on selling national brands.

In summary, we conclude that the power lies to the retailer in this market from the facts that retailer has more bargaining power (Table 4.4.3) and larger margins (Table 4.4.4). To our knowledge, this is the first study which empirically shows that the retailer has more bargaining power than manufacturers.

For future research, we need to develop a framework that embeds both a model by Draganska (2010) and our model within so that employing this comprehensive model would enable us to measure relative power of a retailer to multiple manufacturers channel by channel.

Appendix A

Appendix

A.1 Discrete Choice Model

In this section, we review widely used models for predicting consumers' choice behaviors in the differentiated products markets within the framework of discrete choice models. In the framework, the consumer chooses only one product among the choice set of the differentiated products.[39] Importantly, we incorporate the unobserved product characteristic in the models. This is because some product characteristics such as style, durability, status, and service at a point-of-sale are difficult to quantify but are likely to be correlated with prices. It is well known that ignoring such correlation causes the systematic error in estimation.

The models discussed in this section include logit model and mixed logit model. The properties of these models are discussed as well.

[39]The usual assumption of the choice model is that it is exclusive, in the sense that choice set includes all possible products or alternatives, and products are mutually exclusive and the number of which is finite.

A.1.1 A General Choice Probability Function

Consumers make decisions "under an assumption of (their) utility-maximizing behavior (Train, 2003, 18)." The models under this assumption are referred to as random utility models, and they are derived as follows. First, let us assume that there are N consumers facing $J+1$ ($j = 0, \cdots, J$) products in a market, and the consumer i chooses product j. An alternative of not choosing any of the $j = 1, \cdots, J$ products is sometimes characterized as choosing "outside goods" and denoted as $j = 0$. The utility of consumer i for choosing product j is denoted as U_{ij}, $j = 0, \ldots, J$ and it is assumed to be known to the consumer but not to the researcher. In the framework, a consumer chooses product m if and only if the utility U_{im} is greater than the utilities U_{ij} for all product $j \neq m$, which is expressed as

$$\mathrm{Pr}_{im} \quad = \quad \Pr\left\{U_{ij} < U_{im}, \forall j \neq m\right\}. \tag{A.1.1}$$

where Pr_{im} stands for consumer i's choice probability for product m, and $\{U_{ij} < U_{im}, \forall j \neq m\}$ is an indicator function. The researcher usually observes the vector of some attributes $\boldsymbol{x}$ and specifies a function that relates these observed attributes to consumer's utility. We write this function as V_{ij}, which we call the "representative utility." The utility is decomposed into V_{ij} and ϵ_{ij}; the term ϵ_{ij} captures the factors that affect utility but are not included in V_{ij}. Now the utility of consumer i for product j is rewritten as

$$U_{ij} = V_{ij} + \epsilon_{ij}. \tag{A.1.2}$$

The term ϵ_{ij} is assumed to have some density $f(\epsilon_{ij})$. The behavior of the models largely depends on the specification of $f(\epsilon_{ij})$, as we will show in the following subsections.

With these assumptions, the choice probability of consumer i for product

m in (A.1.1) is expressed further as

$$\begin{aligned}
\Pr_{im} &= \Pr\{U_{ij} < U_{im}, \forall j \neq m\} \\
&= \Pr\{V_{ij} + \epsilon_{ij} < V_{im} + \epsilon_{im}, \forall j \neq m\} \\
&= \Pr\{\epsilon_{ij} < \epsilon_{im} + V_{im} - V_{ij}, \forall j \neq m\} \\
&= \int_{\boldsymbol{\epsilon}_{ij}} \{\epsilon_{ij} < \epsilon_{im} + V_{im} - V_{ij}, \forall j \neq m\} f(\boldsymbol{\epsilon}_{ij}) d\boldsymbol{\epsilon}_{ij}, \qquad \text{(A.1.3)}
\end{aligned}$$

where $f(\boldsymbol{\epsilon}_{ij})$ is the joint density function of random error vector $\boldsymbol{\epsilon}_{ij} = (\epsilon_{i0}, \cdots, \epsilon_{iJ})$. In words, the choice probability is integral of the indicator function over all values of ϵ_{ij} weighted by its density. We calculate (A.1.3) as follows:

$$\begin{aligned}
\Pr_{im} &= \Pr\{\epsilon_{ij} < \epsilon_{im} + V_{im} - V_{ij}, \forall j \neq m\} \\
&= \int_{-\infty}^{\infty} \Big[\int_{-\infty}^{\epsilon_{im}+V_{im}-V_{i0}} \cdots \int_{-\infty}^{\epsilon_{im}+V_{im}-V_{i(m-1)}} \int_{-\infty}^{\epsilon_{im}+V_{im}-V_{i(m+1)}} \cdots \\
&\qquad \int_{-\infty}^{\epsilon_{im}+V_{im}-V_{iJ}} f(\epsilon_{i0}, \cdots, \epsilon_{iJ}) d\epsilon_{i0} \cdots d\epsilon_{i(m-1)} d\epsilon_{i(m+1)} \cdots d\epsilon_{iJ} \Big] d\epsilon_{im} \\
&= \int_{-\infty}^{\infty} \Big[\frac{\partial}{\partial \epsilon_{im}} \int_{-\infty}^{\epsilon_{i0}} \cdots \int_{-\infty}^{\epsilon_{iJ}} f(\epsilon_{i0}, \cdots, \epsilon_{iJ}) \\
&\qquad d\epsilon_{i0} \cdots d\epsilon_{iJ} \Big] \Big|_{\epsilon_{i0}=\epsilon_{im}+V_{im}-V_{i0}, \cdots, \epsilon_{iJ}=\epsilon_{im}+V_{im}-V_{iJ}} d\epsilon_{im} \\
&= \int_{-\infty}^{\infty} \Big[\frac{\partial}{\partial \epsilon_{im}} f(\epsilon_{i0}, \cdots, \epsilon_{iJ}) \Big] \Big|_{\epsilon_{i0}=\epsilon_{ij}+V_{ij}-V_{i0}, \cdots, \epsilon_{iJ}=\epsilon_{ij}+V_{ij}-V_{iJ}} d\epsilon_{im} \\
&= \int_{-\infty}^{\infty} P_m(\epsilon_{im} + V_{im} - V_{i0}, \cdots, \epsilon_{im} + V_{im} - V_{iJ}) d\epsilon_{im}, \qquad \text{(A.1.4)}
\end{aligned}$$

where P_m is the partial derivative of $f(\boldsymbol{\epsilon}_{ij})$ with respect to its mth argument. In the subsections immediately follows, we derive so-called the logit model and discuss some properties of the model.

A.2 Logit Model

The logit model is "the easiest and most widely used discrete model (Train, 2003, 38)." Originally, Luce (1959) derived this model from the assumption of independence from irrelevant alternatives (IIA); on the other hand, McFadden (1974) derived the model from the assumption of utility-maximizing behavior.

The representative utility used in the more modern logit model consists of the observed characteristics and the unobserved characteristics as proposed by Berry (1994). For product j, the observed characteristics are denoted by an $1 \times R$ row vector $\boldsymbol{x}_{j\cdot} = (x_{j1}, \dots, x_{jr}, \dots, x_{jR})$, where $r = 1, \cdots, R$ is product characteristics which affect demands. Note that p_j, the price of product j, is an important element of $\boldsymbol{x}_{j\cdot}$.

The unobserved product characteristic random variable is denoted as ξ_j. The term ξ_j might be thought of the mean preferences of consumers i for the unobserved product characteristics. The representative utility of choosing product j for any consumer is written as

$$V_j = \boldsymbol{x}_{j\cdot}\boldsymbol{\beta} + \xi_j$$

where $\boldsymbol{\beta}$ is an $R \times 1$ column vector of coefficient of $\boldsymbol{x}_{j\cdot}$. The logit choice probability is derived by assuming each ϵ_{ij} is i.i.d. extreme value (Gumbel or type I extreme value) across consumer i and products j, whose cumulative distribution function is in the form of

$$P(\epsilon_{ij}) = \exp\{-\exp(-\epsilon_{ij})\}$$

where $P(\cdot)$ is cumulative distribution function of ϵ_{ij}. Accordingly, the joint cumulative distribution function of $\boldsymbol{\epsilon}_{ij}$ is

$$P(\boldsymbol{\epsilon}_{ij}) = \prod_{j=0}^{J} \exp\{-\exp(-\epsilon_{ij})\}, \tag{A.2.1}$$

since ϵ_{ij} are independent across i and j. Taking partial derivative of (A.2.1) with respect to ϵ_{im} yields

$$\begin{aligned}
& \frac{\partial P(\boldsymbol{\epsilon}_{ij})}{\partial \epsilon_{im}}(\epsilon_{im} + V_m - V_0, \cdots, \epsilon_{im} + V_m - V_J) \\
= \;& (-1) \times \{-\exp(-\epsilon_{im})\} \prod_{j=0}^{J} \exp\{-\exp(-\epsilon_{im} - V_m + V_j)\} \\
= \;& \exp(-\epsilon_{im}) \prod_{j=0}^{J} \exp\{-\exp(-\epsilon_{im} - V_m + V_j)\}. \qquad \text{(A.2.2)}
\end{aligned}$$

The logit model can be derived by substituting equation (A.2.2) to (A.1.4) as

$$\begin{aligned}
\Pr_{im} &= \int_{-\infty}^{\infty} \left[\frac{\partial P(\boldsymbol{\epsilon}_{ij})}{\partial \epsilon_{im}}(\epsilon_{im} + V_m - V_0, \cdots, \epsilon_{im} + V_m - V_J) \right] d\epsilon_{im} \\
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{im}) \prod_{j=0}^{J} \exp\{-\exp(-\epsilon_{im} - V_m + V_j)\}\, d\epsilon_{im} \\
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{im}) \exp\left\{ -\sum_{j=0}^{J} \exp(-\epsilon_{im} - V_m + V_j) \right\} d\epsilon_{im} \\
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{im}) \exp\left\{ -\exp(-\epsilon_{im}) \sum_{j=0}^{J} \exp(V_j - V_m) \right\} d\epsilon_{im}.
\end{aligned} \qquad \text{(A.2.3)}$$

If we let $x = -\exp(-\epsilon_{im})$, x is in interval $(-\infty, 0)$ and $d\epsilon_{im} = dx/\exp(-\epsilon_{ij})$, so that (A.2.3) is rewritten as

$$
\begin{aligned}
\Pr_m &= \int_{-\infty}^{0} \exp(-\epsilon_{im}) \exp\left\{ x \sum_{J=0}^{J} \exp(V_j - V_m) \right\} \frac{dx}{\exp(-\epsilon_{ij})} \\
&= \int_{-\infty}^{0} \exp\left\{ x \sum_{J=0}^{J} \exp(V_j - V_m) \right\} dx \\
&= \left[\frac{\exp\left\{ x \sum_{J=0}^{J} \exp(V_j - V_m) \right\}}{\sum_{J=0}^{J} \exp(V_j - V_m)} \right]_{-\infty}^{0} \\
&= \frac{\exp(0) - \exp(-\infty)}{\sum_{J=0}^{J} \exp(V_j - V_m)} = \frac{1}{\sum_{J=0}^{J} \exp(V_j - V_m)} \\
&= \frac{\exp(V_m)}{\sum_{J=0}^{J} \exp(V_j)}. \qquad \text{(A.2.4)}
\end{aligned}
$$

This is the standard formula for the logit choice probability. Note that market share of product j is the choice probability itself in the standard logit model framework as

$$s_m^e(\boldsymbol{V}) = \Pr_m$$

where $\boldsymbol{V} \equiv (V_0, \ldots, V_J)$, since the estimated market share of m denoted as s_m^e, where superscript e stands for "estimated", depends on the representative utilities of all the products in the market $j = 0, \ldots, J$.

A.2.1 Estimation of the Logit Model

In this subsection, we follow Berry (1994) and review a method to estimate parameter of logit model assuming we have the market share data including the share of outside goods. If we take log of the right-hand side of the equation (A.2.4), we have

$$\log(s_m^e(\boldsymbol{V})) = \log\left\{ \frac{\exp(V_m)}{\sum_{j=0}^{J} \exp(V_j)} \right\}. \qquad \text{(A.2.5)}$$

Setting the utilities of the observed product characteristics and unobserved product characteristics of a outside goods as $\boldsymbol{x}_{0\cdot} = (0, \ldots, 0)$ and $\xi_0 = 0$ respectively, we derive the utility of the outside goods as

$$U_{i0} = \epsilon_{i0} \tag{A.2.6}$$

since $V_0 = 0$. Applying (A.2.6) to (A.2.4), the choice probability of outside goods $\Pr_0$ under the logit model is

$$\Pr_0 = \frac{\exp(V_0)}{\sum_{j=0}^{J} \exp(V_j)} = \frac{\exp(0)}{\sum_{j=0}^{J} \exp(V_j)} = \frac{1}{\sum_{j=0}^{J} \exp(V_j)}.$$

Notice that this choice probability of outside goods coincides with the market share of outside goods, and we have the market share of outside goods as

$$s_0^e(\boldsymbol{V}) = \frac{1}{\sum_{j=0}^{J} \exp(V_j)}. \tag{A.2.7}$$

Substitute (A.2.7) to the denominator of (A.2.5) and obtain

$$\log(s_m^e(\boldsymbol{V})) = \log(\exp(V_m) \cdot s_0^e(\boldsymbol{V})) = V_m + \log(s_0^e(\boldsymbol{V})).$$

Therefore, we have

$$\log(s_m^e(\boldsymbol{V})) - \log(s_0^e(\boldsymbol{V})) = V_m = \boldsymbol{x}_{m\cdot}\boldsymbol{\beta} + \xi_m.$$

Assuming $s_m^e(\boldsymbol{V})$ and $s_0^e(\boldsymbol{V})$ equal to their observed counterparts s_m^o and s_0^o respectively at the true value of $\boldsymbol{\beta}$ and ξ_m, we can replace $s_m^e(\boldsymbol{V})$ and $s_0^e(\boldsymbol{V})$ with the observed market share s_m^o and s_0^o respectively to obtain

$$\log(s_m^o) - \log(s_0^o) = \boldsymbol{x}_{m\cdot}\boldsymbol{\beta} + \xi_m. \tag{A.2.8}$$

If OLS is applied to estimate $\boldsymbol{\beta}$ and ξ_m using the left-hand side of the equation (A.2.8) as a dependent variable and regarding ξ_m as the residuals, then the OLS estimator would be inconsistent as explained in section A.4 if the element of $\boldsymbol{x}_{m\cdot}$, especially p_m, is correlated with ξ_m.

In order to avoid this problem, the 2SLS method is implemented. An instrumental variable is defined as the variable that is correlated with the endogenous variables $\boldsymbol{x}_{m\cdot}$ but is not with the error term (ξ_m in this case). The 2SLS method allows the correlation between the endogenous variables and error terms, but still estimates $\boldsymbol{\beta}$ consistently. The formal definition of an instrumental variable and detailed discussion of 2SLS are given in section A.5.

A.3 Mixed Logit Model

The mixed logit model is a highly flexible model which can approximate any random utility model by allowing the parameters associated with each observed variable to vary across consumers (Revelt & Train, 1998; McFadden & Train, 2000). The mixed logit model thus can represent heterogeneity across consumers. Mixed logit probabilities are obtained as the integral of the standard logit probabilities over a density of parameter,

$$\Pr{}_{ij} = \int L_{ij}(\boldsymbol{\beta}) f(\boldsymbol{\beta}) d\boldsymbol{\beta}, \tag{A.3.1}$$

where $L_{ij}(\boldsymbol{\beta})$ is the logit probability evaluated at parameter $\boldsymbol{\beta}$ as

$$L_{ij}(\boldsymbol{\beta}) = \frac{\exp V_{ij}(\boldsymbol{\beta})}{\sum_{j=0}^{J} \exp V_{ij}(\boldsymbol{\beta})}$$

and $f(\boldsymbol{\beta})$ is a density function of parameter. In words, "the mixed logit probability is a weighted average of the logit formula evaluated at the different value of $\boldsymbol{\beta}$ (Train, 2003, 139)." The word "mixed" comes from the statistical custom to call the weighted average of several functions a mixed function, and the density which provides the weight is called mixing distribution.

"The mixed logit probability can be derived from utility-maximizing behavior in several ways that are formally equivalent but provide different interpretations (ibid, 141)." In the following subsections, two interpretations of the mixed logit model are presented. The first one is the random-coefficients and the other is error-component.

A.3.1 Random-Coefficients

Unlike (A.1.2), the utility that consumer i obtains from the alternative j is written as

$$U_{ij} = V_{ij}(\boldsymbol{\beta}_i) + \epsilon_{ij}$$

where $V_{ij}(\boldsymbol{\beta}_i)$ is the observed component of utility which depends on the parameter $\boldsymbol{\beta}_i$, which varies across consumers and has the density function $f(\boldsymbol{\beta}_i|\Omega)$, where Ω is the set of the true parameters of this density function, and ϵ_{ij} is an unobserved error with i.i.d. extreme value.

We can write the probability that consumer i chooses alternative m conditional on given $\boldsymbol{\beta}_i$, denoted as $L_{im}(\boldsymbol{\beta}_i)$, in the standard logit form as

$$L_{im}(\boldsymbol{\beta}_i) = \frac{\exp(V_{im}(\boldsymbol{\beta}_i))}{\sum_{j=0}^{J} \exp(V_{ij}(\boldsymbol{\beta}_i))}.$$

Since the researchers do not know the values of $\boldsymbol{\beta}_i$, the unconditional probability is obtained by integrating the conditional probability over all possible value of $\boldsymbol{\beta}_i$ as

$$\Pr{}_{im} = \int L_{im}(\boldsymbol{\beta}_i) f(\boldsymbol{\beta}_i|\Omega) d\boldsymbol{\beta}_i,$$

which is mixed logit probability in (A.3.1). If the utility is linear in $\boldsymbol{\beta}_i$, then $V_{ij}(\boldsymbol{\beta})$ can be written as $\boldsymbol{x}_{ij\cdot}\boldsymbol{\beta}_i$, where $\boldsymbol{x}_{ij\cdot} = (x_{ij0}, \ldots, x_{ijr}, \ldots, x_{ijR})$ is vector of observed variables relating to alternative j and $\boldsymbol{\beta}_i$ is corresponding vector of parameters. If that is the case, $L_{im}(\boldsymbol{\beta}_i)$ becomes

$$L_{im}(\boldsymbol{\beta}_i) = \frac{\exp(\boldsymbol{x}_{ij\cdot}\boldsymbol{\beta}_i)}{\sum_{j=0}^{J} \exp(\boldsymbol{x}_{ij\cdot}\boldsymbol{\beta}_i)},$$

and the mixed logit probability becomes

$$\Pr{}_{im} = \int \frac{\exp(\boldsymbol{x}_{ij\cdot}\boldsymbol{\beta}_i)}{\sum_{j=0}^{J}\exp(\boldsymbol{x}_{ij\cdot}\boldsymbol{\beta}_i)} f(\boldsymbol{\beta}_i|\Omega)d\boldsymbol{\beta}_i.$$

A.3.2 Error-Components

The second interpretation is called error-components. The utility in error-component is specified as

$$U_{ij} = \boldsymbol{x}_{ij\cdot}\boldsymbol{\alpha} + \boldsymbol{z}_{ij\cdot}\boldsymbol{\mu}_{i\cdot} + \epsilon_{ij}, \tag{A.3.2}$$

where $\boldsymbol{x}_{ij\cdot}$ and $\boldsymbol{z}_{ij\cdot} \equiv (z_{ij0}, \ldots, z_{ijl}, \ldots, z_{ijL})$ is vector of observed variables relating to alternative j, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_r, \ldots, \alpha_R)^T$ is vector of parameter which are fixed over consumers and alternatives, and $\boldsymbol{\mu}_{i\cdot} = (\mu_{i1}, \ldots, \mu_{il}, \ldots, \mu_{iL})^T$ is vector of random terms that vary over consumers which has the distribution $g(\boldsymbol{\mu}|\Omega)$ where Ω is the set of fixed parameters of the distribution $g(\cdot)$.

In this setting, $\boldsymbol{x}_{ij\cdot}\boldsymbol{\alpha}$ is the component of the utility which is invariant across consumers and $\boldsymbol{z}_{ij\cdot}\boldsymbol{\mu}_{i\cdot}$ is that varies across consumers. The terms $\boldsymbol{z}_{ij\cdot}\boldsymbol{\mu}_{i\cdot}$ along with ϵ_{ij} define the stochastic portion of utility which we denote as $\eta_{ij} \equiv \boldsymbol{z}_{ij\cdot}\boldsymbol{\mu}_{i\cdot} + \epsilon_{ij}$. In the error-component specification, utilities are correlated with one another if $\boldsymbol{z}_{ij\cdot}$ is non-zero, since the covariance of the two stochastic component of alternatives η_{ij} and η_{ig} is $\mathrm{Cov}(\eta_{ij}, \eta_{ig}) = E[\boldsymbol{z}_{ij\cdot}\boldsymbol{\mu}_{i\cdot} + \epsilon_{ij}]'[\boldsymbol{z}_{ig\cdot}\boldsymbol{\mu}_{g\cdot} + \epsilon_{ig}] = \boldsymbol{z}'_{ij\cdot}W\boldsymbol{z}_{ig\cdot}$, where W is the covariance of $\boldsymbol{\mu}_{i\cdot}$. The terms $\boldsymbol{z}_{ij\cdot}\boldsymbol{\mu}_{i\cdot}$ are interpreted as "error-components" because they can induce heteroskedasticities and the correlations over the alternatives (Brownstone et al., 2000).

The choice probability for each consumer i now depends on $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}_{i\cdot}$. We have the conditional choice probability of consumer i for alternative m as

$$\Pr{}_{im|\boldsymbol{\mu}_i} = \frac{\exp(\boldsymbol{x}_{im\cdot}\boldsymbol{\alpha} + \boldsymbol{z}_{im\cdot}\boldsymbol{\mu}_{i\cdot})}{\sum_{j\in J}\exp(\boldsymbol{x}_{ij\cdot}\boldsymbol{\alpha} + \boldsymbol{z}_{ij\cdot}\boldsymbol{\mu}_{i\cdot})}. \tag{A.3.3}$$

The unconditional choice probability is obtained by integrating (A.3.3) respect to all the value of $\boldsymbol{\mu}_{i\cdot}$ as

$$\Pr{}_{im} = \int \left[\frac{\exp(\boldsymbol{x}_{im\cdot}\boldsymbol{\alpha} + \boldsymbol{z}_{im\cdot}\boldsymbol{\mu}_{i\cdot})}{\sum_{j\in J}\exp(\boldsymbol{x}_{ij\cdot}\boldsymbol{\alpha} + \boldsymbol{z}_{ij\cdot}\boldsymbol{\mu}_{i\cdot})} \right] g(\boldsymbol{\mu}|\Omega) d\boldsymbol{\mu}_{i\cdot}.$$

As stated, random-coefficients and error-components specifications are formally equivalent because if we decompose $\boldsymbol{\beta}_i$ of random-coefficients into its mean α and deviation μ_i, we have the utility $U_{ij} = \boldsymbol{x}_{ij\cdot}\boldsymbol{\alpha} + \boldsymbol{x}_{ij\cdot}\boldsymbol{\mu}_{i\cdot} + \epsilon_{ij}$, which is that of error-components specification, defined by $\boldsymbol{x}_{ij\cdot} = \boldsymbol{z}_{ij\cdot}$ (Train, 2003). The different interpretation or application of these models depends on the purpose of research and the appropriateness of the model depending on the situation. While random-coefficient model allows coefficients to vary, which is more intuitively plausible than error-component, it is unreasonable to apply it when there are many variables (ibid). On the other hand, when substitution patterns are emphasized, error-components is more convenient. While the different substitution patterns can be obtained by a different specification of function $g(\cdot)$, the most widely used mixing distributions are the normal distribution and log-normal distribution (Bhat, 2001). When the mixing distribution is discrete, the mixed logit model becomes the latent class model. Because of the integration in the mixed logit model, the probability cannot be calculated in the closed form, especially if dimensions of the parameter is large.

A.4 The Standard Asymptotic Results of Ordinary Least Squares (OLS) Estimator

OLS method is sometimes used in the regression analysis for parameter estimation. In this section, the properties of OLS estimator are discussed. In subsection A.4.1, OLS estimator is derived. In subsection A.4.2, (weak) consistency of OLS estimator is shown under a set of standard assumptions. In subsection A.4.3, asymptotic normality of OLS estimator is demonstrated with an additional set of standard assumptions.

A.4.1 The OLS Estimator

Let us consider regression model with J observations

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{x}_{1\cdot}\boldsymbol{\beta} + \epsilon_1 \\ \vdots \\ \boldsymbol{x}_{i\cdot}\boldsymbol{\beta} + \epsilon_i \\ \vdots \\ \boldsymbol{x}_{J\cdot}\boldsymbol{\beta} + \epsilon_J \end{bmatrix} = \begin{bmatrix} \sum_{r=1}^{R} x_{1r}\beta_r + \epsilon_1 \\ \vdots \\ \sum_{r=1}^{R} x_{ir}\beta_r + \epsilon_i \\ \vdots \\ \sum_{r=1}^{R} x_{Jr}\beta_r + \epsilon_J \end{bmatrix} \tag{A.4.1}$$

where $\boldsymbol{y} = (y_1, \cdots, y_i, \cdots, y_J)^T$ is a $J \times 1$ vector of response or dependent variables,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{1\cdot} \\ \vdots \\ \boldsymbol{x}_{i\cdot} \\ \vdots \\ \boldsymbol{x}_{J\cdot} \end{bmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1r} & \cdots & x_{1R} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ir} & \cdots & x_{iR} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{J1} & \cdots & x_{Jr} & \cdots & x_{JR} \end{pmatrix} = [\boldsymbol{x}_{\cdot 1}, \cdots, \boldsymbol{x}_{\cdot r}, \cdots, \boldsymbol{x}_{\cdot R}].$$

is an $J \times R$ matrix of R stochastic explanatory variables whose ith row is denoted by $\boldsymbol{x}_{i\cdot} = (x_{i1}, \cdots, x_{ir}, \cdots, x_{iR})$ and whose rth column is denoted by $\boldsymbol{x}_{\cdot r} = (x_{1r}, \cdots, x_{ir}, \cdots, x_{Jr})^T$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_r, \cdots, \beta_R)^T$ is a $R \times 1$ vector of parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_i, \cdots, \epsilon_J)^T$ is a $J \times 1$ vector of errors. The assumption required for the existence of OLS estimator is as follows.

Assumption 1a

The term $\boldsymbol{X}^T\boldsymbol{X}$ has rank R as

$$\boldsymbol{X}^T\boldsymbol{X} = R \tag{A.4.2}$$

where the term $\boldsymbol{X}^T\boldsymbol{X}$ can be written as

$$\begin{aligned}
\boldsymbol{X}^T\boldsymbol{X} &= \begin{bmatrix} \boldsymbol{x}_{1\cdot}{}^T \cdots \boldsymbol{x}_{i\cdot}{}^T \cdots \boldsymbol{x}_{J\cdot}{}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{1\cdot} \\ \vdots \\ \boldsymbol{x}_{i\cdot} \\ \vdots \\ \boldsymbol{x}_{J\cdot} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{\cdot 1}{}^T \\ \vdots \\ \boldsymbol{x}_{\cdot r}{}^T \\ \vdots \\ \boldsymbol{x}_{\cdot R}{}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{\cdot 1} \cdots \boldsymbol{x}_{\cdot r} \cdots \boldsymbol{x}_{\cdot R} \end{bmatrix} \\
&= \begin{pmatrix} x_{11} & \cdots & x_{i1} & \cdots & x_{J1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1r} & \cdots & x_{ir} & \cdots & x_{Jr} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1R} & \cdots & x_{iR} & \cdots & x_{JR} \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1r} & \cdots & x_{1R} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ir} & \cdots & x_{iR} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{J1} & \cdots & x_{Jr} & \cdots & x_{JR} \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^J x_{i1} \cdot x_{i1} & \cdots & \sum_{i=1}^J x_{i1} \cdot x_{ir} & \cdots & \sum_{i=1}^J x_{i1} \cdot x_{iR} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sum_{i=1}^J x_{ir} \cdot x_{i1} & \cdots & \sum_{i=1}^J x_{ir} \cdot x_{ir} & \cdots & \sum_{i=1}^J x_{ir} \cdot x_{iR} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sum_{i=1}^J x_{iR} \cdot x_{i1} & \cdots & \sum_{i=1}^J x_{iR} \cdot x_{ir} & \cdots & \sum_{i=1}^J x_{iR} \cdot x_{iR} \end{pmatrix}.
\end{aligned} \tag{A.4.3}$$

Since $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric $R \times R$ matrix, $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular and invertible with this assumption.

OLS estimator of $\boldsymbol{\beta}$, which we denote $\widehat{\boldsymbol{\beta}}_{OLS}$, can be determined to make the column vectors of $\boldsymbol{X}$ and the regression residuals $\widehat{\boldsymbol{\epsilon}}_i$ be orthogonal, i.e., $\boldsymbol{X}^T\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{0}$. Since the residual $\widehat{\boldsymbol{\epsilon}}_i$ is written as $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{OLS}$, we have $\boldsymbol{X}^T(\boldsymbol{y} -$

$\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{OLS}) = \mathbf{0}$, so that we obtain $\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{X}^T\boldsymbol{y}$. From **Assumption 1a**, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists and we obtain the OLS estimator of $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{A.4.4}$$

Substituting (A.4.1) to (A.4.4) gives

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{OLS} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_i)) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{\epsilon}_i) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}_i \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}_i. \qquad \text{(A.4.5)}
\end{aligned}$$

If the second term of the last equation goes to $\mathbf{0}$ as the sample size increases, the OLS estimator is (weakly) consistent, as we will show in the following subsection.

A.4.2 (Weak) Consistency of the OLS Estimator

Assumptions for (weakly) consistent OLS estimator

Below we list the assumptions for OLS estimator to be (weakly) consistent.

Assumption 1b

The quantity $\boldsymbol{x}_{\cdot r} = (x_{1r}, \cdots, x_{Jr})^T$ is the realization of the r-th stochastic explanatory variables whose distribution has finite mean $\mu_{\boldsymbol{x}_r}$ and finite variance $\sigma^2_{\boldsymbol{x}_r}$.[40]

[40] The mean of $\boldsymbol{x}_{\cdot r}$ is calculated using the joint distribution $f_{X_{\cdot 1},\ldots,X_{\cdot R}}(\boldsymbol{x}_{\cdot 1},\ldots,\boldsymbol{x}_{\cdot R})$ as

$$\begin{aligned}
\mu_{\boldsymbol{x}_r} = \mathrm{E}[\boldsymbol{x}_{\cdot r}] &= \int \boldsymbol{x}_{\cdot r} \int \ldots \int f_{X_{\cdot 1},\ldots,X_{\cdot R}}(\boldsymbol{x}_{\cdot 1},\ldots,\boldsymbol{x}_{\cdot R}) d\boldsymbol{x}_{\cdot 1},\ldots,d\boldsymbol{x}_{\cdot r-1}, d\boldsymbol{x}_{\cdot r+1},\ldots,d\boldsymbol{x}_{\cdot R} \\
&= \int \boldsymbol{x}_{\cdot r} f_{X_{\cdot r}}(\boldsymbol{x}_{\cdot r}) d\boldsymbol{x}_{\cdot r}.
\end{aligned}$$

The variance $\sigma^2_{\boldsymbol{x}_r}$ can be calculated similarly.

Assumption 1c

The R stochastic explanatory variables $\boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot R}$ do not have to be independent, nor identically distributed across r. Therefore the covariance between $\boldsymbol{x}_{\cdot k}$ and $\boldsymbol{x}_{\cdot r} (k \neq r)$ is not necessarily zero and is defined as $\sigma^2_{\boldsymbol{x}_k, \boldsymbol{x}_r}$, assuming their joint distributions have joint moments up to a second order.[41]

Assumption 1d

The set of explanatory variable $\boldsymbol{x}_{i\cdot} = (x_{i1}, \cdots, x_{ir}, \cdots, x_{iR})$ is independent across i. In other words, R explanatory variables are jointly sampled independently over J samples.

On the error term

The error term $\boldsymbol{\epsilon}_i = (\epsilon_1, \ldots, \epsilon_J)^T$ is the collection of measurement errors and the residue effects of the variables not included in the model. We assume that we are able to specify a good set of explanatory variables, and thus we are able to assume the following:

[41] The covariance between $\boldsymbol{x}_{\cdot k}$ and $\boldsymbol{x}_{\cdot r} (k \neq r)$ is written as

$$\sigma^2_{\boldsymbol{x}_k, \boldsymbol{x}_r} = \mathrm{Cov}(\boldsymbol{x}_{\cdot k}, \boldsymbol{x}_{\cdot r}) = \mathrm{E}[\boldsymbol{x}_k, \boldsymbol{x}_r] - \mathrm{E}[\boldsymbol{x}_k]\mathrm{E}[\boldsymbol{x}_r].$$

The term $\mathrm{E}[\boldsymbol{x}_k]$ and $\mathrm{E}[\boldsymbol{x}_r]$ has finite values from **Assumption 1b**. The term $\mathrm{E}[\boldsymbol{x}_k, \boldsymbol{x}_r]$ can be calculated using the joint distribution $f_{X_{\cdot 1}, \ldots, X_{\cdot R}}(\boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot R})$ as

$$\begin{aligned} \mathrm{E}[\boldsymbol{x}_k, \boldsymbol{x}_r] &= \int\int \boldsymbol{x}_{\cdot k}, \boldsymbol{x}_{\cdot r} \int \ldots \int f_{X_{\cdot 1}, \ldots, X_{\cdot R}}(\boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot r}, \ldots, \boldsymbol{x}_{\cdot R}) \\ &\qquad d\boldsymbol{x}_{\cdot 1}, \ldots, d\boldsymbol{x}_{\cdot k-1}, d\boldsymbol{x}_{\cdot k+1}, \ldots, d\boldsymbol{x}_{\cdot r-1}, d\boldsymbol{x}_{\cdot r+1}, \ldots, d\boldsymbol{x}_{\cdot R} \\ &= \int\int \boldsymbol{x}_{\cdot k}, \boldsymbol{x}_{\cdot r} f_{X_{\cdot k}, X_{\cdot r}}(\boldsymbol{x}_{\cdot k}, \boldsymbol{x}_{\cdot r}) d\boldsymbol{x}_{\cdot k} d\boldsymbol{x}_{\cdot r}. \end{aligned}$$

Assumption 2a

$$\mathrm{E}[\boldsymbol{\epsilon}_i] = \mathbf{0}, \tag{A.4.6}$$

where $\mathbf{0}$ is $J \times 1$ vector of zeros.

Assumption 2b

The terms ϵ_i are i.i.d. across i.

On the relationship between explanatory variables and error term

Assumption 3a

Between explanatory variable x_{ir} and the error ϵ_i for the same individual i, we assume

$$\mathrm{Cov}(x_{ir} \cdot \epsilon_i) = 0 \tag{A.4.7}$$

for $i = 1, \ldots, J$ and $r = 1, \ldots, R$. This assumption along with **Assumption 2a** and **Assumption 2a** leads to the condition[42]

$$\mathrm{E}[x_{ir} \cdot \epsilon_i] = 0 \tag{A.4.8}$$

for $i = 1, \ldots, J$ and $r = 1, \ldots, R$. This suggests the orthogonal condition of x_{ir} and ϵ_i, which is standard assumption of OLS.

With these assumptions, the (weak) consistency of OLS estimator is shown as follows: The second term of the right hand side of (A.4.5) can be rewritten as

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}_i = (J^{-1}\boldsymbol{X}^T\boldsymbol{X})^{-1}J^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}_i. \tag{A.4.9}$$

[42]From (A.4.6) and (A.4.7), we have

$$\mathrm{E}[x_{ir} \cdot \epsilon_i] = \mathrm{Cov}(x_{ir}, \epsilon_i) + \mathrm{E}[x_{ir}]\mathrm{E}[\epsilon_i] = 0 + \mathrm{E}[x_{ir}] \cdot 0 = 0.$$

Since elements of $\boldsymbol{x}_{\cdot r}$ are i.i.d. across i for all r from **Assumption 1b**, Etemadi's Strong Law of Large Numbers (SLLN) can be applied. With the SLLN, (r,r)-th and (k,r)-th element of (A.4.3) times J^{-1} respectively converge to their moments as

$$J^{-1}(\boldsymbol{X}^T\boldsymbol{X})_{rr} = J^{-1}\sum_{i=1}^{J} {x_{ir}}^2 \overset{\text{a.s.}}{\to} \mathrm{E}[\boldsymbol{x}_{\cdot r}^T \cdot \boldsymbol{x}_{\cdot r}] = \mu_{\boldsymbol{x}_r}^2 + \sigma_{\boldsymbol{x}_r}^2,$$

from **Assumption 1b** and

$$J^{-1}(\boldsymbol{X}^T\boldsymbol{X})_{kr} = J^{-1}\sum_{i=1}^{J}(x_{ik}\cdot x_{ir}) \overset{\text{a.s.}}{\to} \mathrm{E}[\boldsymbol{x}_{\cdot k}^T \cdot \boldsymbol{x}_{\cdot r}] = \mu_{\boldsymbol{x}_k}\cdot\mu_{\boldsymbol{x}_r} + \sigma_{\boldsymbol{x}_k,\boldsymbol{x}_r}^2,$$

from **Assumption 1b** and **Assumption 1c**. Therefore, $J^{-1}\boldsymbol{X}^T\boldsymbol{X}$ has non-stochastic probability limit $\boldsymbol{Q}$, since almost sure convergence implies convergence in probability. The elements of $\boldsymbol{Q}$ are written as

$$\operatorname*{plim}_{J\to\infty} J^{-1}\boldsymbol{X}^T\boldsymbol{X} \equiv \boldsymbol{Q} = \begin{pmatrix} E[{\boldsymbol{x}_{\cdot 1}}^T\cdot\boldsymbol{x}_{\cdot 1}] & \cdots & E[{\boldsymbol{x}_{\cdot 1}}^T\cdot\boldsymbol{x}_{\cdot r}] & \cdots & E[{\boldsymbol{x}_{\cdot 1}}^T\cdot\boldsymbol{x}_{\cdot R}] \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ E[{\boldsymbol{x}_{\cdot r}}^T\cdot\boldsymbol{x}_{\cdot 1}] & \cdots & E[{\boldsymbol{x}_{\cdot r}}^T\cdot\boldsymbol{x}_{\cdot r}] & \cdots & E[{\boldsymbol{x}_{\cdot r}}^T\cdot\boldsymbol{x}_{\cdot R}] \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ E[{\boldsymbol{x}_{\cdot R}}^T\cdot\boldsymbol{x}_{\cdot 1}] & \cdots & E[{\boldsymbol{x}_{\cdot R}}^T\cdot\boldsymbol{x}_{\cdot r}] & \cdots & E[{\boldsymbol{x}_{\cdot R}}^T\cdot\boldsymbol{x}_{\cdot R}] \end{pmatrix}$$

$$= \begin{pmatrix} \mu_{\boldsymbol{x}_1}^2 + \sigma_{\boldsymbol{x}_1}^2 & \cdots & \mu_{\boldsymbol{x}_1}\cdot\mu_{\boldsymbol{x}_r} + \sigma_{\boldsymbol{x}_1,\boldsymbol{x}_r}^2 & \cdots & \mu_{\boldsymbol{x}_1}\cdot\mu_{\boldsymbol{x}_R} + \sigma_{\boldsymbol{x}_1,\boldsymbol{x}_R}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\boldsymbol{x}_r}\cdot\mu_{\boldsymbol{x}_1} + \sigma_{\boldsymbol{x}_r,\boldsymbol{x}_1}^2 & \cdots & \mu_{\boldsymbol{x}_r}^2 + \sigma_{\boldsymbol{x}_r}^2 & \cdots & \mu_{\boldsymbol{x}_r}\cdot\mu_{\boldsymbol{x}_R} + \sigma_{\boldsymbol{x}_r,\boldsymbol{x}_R}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\boldsymbol{x}_R}\cdot\mu_{\boldsymbol{x}_1} + \sigma_{\boldsymbol{x}_R,\boldsymbol{x}_1}^2 & \cdots & \mu_{\boldsymbol{x}_R}\cdot\mu_{\boldsymbol{x}_r} + \sigma_{\boldsymbol{x}_R,\boldsymbol{x}_r}^2 & \cdots & \mu_{\boldsymbol{x}_R}^2 + \sigma_{\boldsymbol{x}_r}^2 \end{pmatrix}.$$

Thus we have

$$\operatorname{plim}(J^{-1}\boldsymbol{X}^T\boldsymbol{X})^{-1} = \boldsymbol{Q}^{-1}$$

as long as $\boldsymbol{X}^T\boldsymbol{X}$ is invertible because of Continuous Mapping Theorem, since the inverse operator is continuous on the space of invertible matrices. Since

$\boldsymbol{X}^T\boldsymbol{X}$ is invertible from **Assumption 1a**, $(J^{-1}\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{p} \boldsymbol{Q}^{-1}$ holds. Likewise, $\boldsymbol{X}^T\boldsymbol{\epsilon}_i$ can be written as

$$\boldsymbol{X}^T\boldsymbol{\epsilon}_i = \left[\ \boldsymbol{x}_{1\cdot}{}^T \cdots \boldsymbol{x}_{i\cdot}{}^T \cdots \boldsymbol{x}_{J\cdot}{}^T\ \right] \boldsymbol{\epsilon}_i = \begin{bmatrix} \boldsymbol{x}_{\cdot 1}{}^T \\ \vdots \\ \boldsymbol{x}_{\cdot r}{}^T \\ \vdots \\ \boldsymbol{x}_{\cdot R}{}^T \end{bmatrix} \boldsymbol{\epsilon}_i$$

$$= \begin{pmatrix} x_{11} & \cdots & x_{i1} & \cdots & x_{J1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1r} & \cdots & x_{ir} & \cdots & x_{Jr} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1R} & \cdots & x_{iR} & \cdots & x_{JR} \end{pmatrix} \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_J \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^J x_{i1} \cdot \epsilon_i \\ \vdots \\ \sum_{i=1}^J x_{ir} \cdot \epsilon_i \\ \vdots \\ \sum_{i=1}^J x_{iR} \cdot \epsilon_i \end{bmatrix}. \tag{A.4.10}$$

Let us denote r-th row of a vector as $(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)_r$ and multiply it by J^{-1}, and write

$$J^{-1} \cdot (\boldsymbol{X}^T\boldsymbol{\epsilon}_i)_r = J^{-1} \sum_{i=1}^{J} x_{ir} \cdot \epsilon_i. \tag{A.4.11}$$

From the SLLN, the equation (A.4.11) becomes

$$J^{-1} \sum_{i=1}^{J} x_{ir} \cdot \epsilon_i \xrightarrow{\text{a.s.}} \mathrm{E}(x_{ir} \cdot \epsilon_i) = 0, \tag{A.4.12}$$

from the condition (A.4.8). Since (A.4.12) holds for every $r = 1, \ldots, R$ from **Assumption 3a**, $J^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}_i$ have nonstochastic probability limit of $\mathbf{0}$ as

$$\operatorname*{plim}_{J\to\infty} J^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}_i = \mathbf{0}.$$

Thus the plim of the second term of the right side of the equation in (A.4.5) becomes

$$\left(\operatorname*{plim}_{J\to\infty} J^{-1}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \left(\operatorname*{plim}_{J\to\infty} J^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}_i\right) = (\boldsymbol{Q})^{-1} \cdot \mathbf{0} = \mathbf{0}$$

from Slutsky's theorem. Therefore, the OLS estimator is (weakly) consistent since the second term in the last equation of (A.4.5) becomes a vector of zeros.

Remark If, on the other hand, the explanatory variables are correlated with error term as $E[x_{ir} \cdot \epsilon_i] \neq 0$ for some r, the OLS estimator in general is not (weakly) consistent. Such a situation is taken care of in the next section by introducing the 2SLS method.

A.4.3 Asymptotic Normality of the OLS Estimator

Asymptotic normality of OLS estimator will be demonstrated in this subsection. We first need an additional set of assumptions required for asymptotic normality of OLS estimator as follows:

On the error term (continued.)

Assumption 2c

$$\mathrm{E}[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T] = \sigma^2 \mathbf{I}. \tag{A.4.13}$$

On the relationship between explanatory variables and error term (continued.)

Assumption 3b

The variance covariance matrix of $\boldsymbol{X}^T \boldsymbol{\epsilon}_i$ are finite as

$$\mathrm{E}[(\boldsymbol{X}^T \boldsymbol{\epsilon}_i)(\boldsymbol{X}^T \boldsymbol{\epsilon}_i)^T] < \infty.$$

From (A.4.5), we have

$$\sqrt{J}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) = \{J^{-1}(\boldsymbol{X}^T\boldsymbol{X})\}^{-1}J^{-1/2}\boldsymbol{X}^T\boldsymbol{\epsilon}_i. \tag{A.4.14}$$

Let us examine the asymptotic property $J^{-1/2}\boldsymbol{X}^T\boldsymbol{\epsilon}_i$ in the following. We see that $\boldsymbol{X}^T\boldsymbol{\epsilon}_i$ can be written as $\boldsymbol{X}^T\boldsymbol{\epsilon}_i = \sum_{i=1}^J \boldsymbol{x}_{i\cdot}^T\epsilon_i$ from (A.4.10). From **Assumption 3a**, $\boldsymbol{x}_{i\cdot}^T\epsilon_i$ has mean $\mathbf{0}$ and its covariance $E[(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)^T]$ is finite from **Assumption 3b**. Then $\sum_{i=1}^J \boldsymbol{x}_{i\cdot}^T\epsilon_i$ becomes, as we will show soon, as

$$J^{-1/2}\sum_{i=1}^{J}\boldsymbol{x}_{i\cdot}^T\epsilon_i \overset{\text{w}}{\rightsquigarrow} N(\mathbf{0}, \boldsymbol{B}_{\boldsymbol{X}}), \tag{A.4.15}$$

where $\boldsymbol{B}_{\boldsymbol{X}}$ is asymptotic variance of $\boldsymbol{X}^T\boldsymbol{\epsilon}_i$ which is

$$\begin{aligned} V(\boldsymbol{X}^T\boldsymbol{\epsilon}_i) &= \mathrm{E}[(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)^T] - \mathrm{E}[(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)]\mathrm{E}[(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)^T] \\ &= \mathrm{E}[(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)(\boldsymbol{X}^T\boldsymbol{\epsilon}_i)^T] = \mathrm{E}[\boldsymbol{X}^T\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T\boldsymbol{X}]. \end{aligned} \tag{A.4.16}$$

Since the term $\mathrm{E}[\boldsymbol{X}^T\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T\boldsymbol{X}]$ involves two random variables, it can be written as

$$\begin{aligned} \mathrm{E}[\boldsymbol{X}^T\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T\boldsymbol{X}] &= \int\int \left(\boldsymbol{X}^T\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T\boldsymbol{X}\right) f_{\boldsymbol{X},\boldsymbol{\epsilon}_i}(\boldsymbol{X},\boldsymbol{\epsilon}_i)d\boldsymbol{X}d\boldsymbol{\epsilon}_i \\ &= \int \left[\int \left(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T|\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{x}^T\boldsymbol{x}\right) f_{\boldsymbol{\epsilon}_i|\boldsymbol{X}}(\boldsymbol{\epsilon}_i|\boldsymbol{X})d\boldsymbol{\epsilon}_i\right] \boldsymbol{X}^T\boldsymbol{X}f_{\boldsymbol{X}}(\boldsymbol{X})d\boldsymbol{X} \\ &= \mathrm{E}\left[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T|\boldsymbol{x}^T\boldsymbol{x}\right]\int \boldsymbol{X}^T\boldsymbol{X}f_{\boldsymbol{X}}(\boldsymbol{X})d\boldsymbol{X} \\ &= \mathrm{E}\left[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T\right]\mathrm{E}\left[\boldsymbol{X}^T\boldsymbol{X}\right], \end{aligned}$$

since $\boldsymbol{X}$ and $\boldsymbol{\epsilon}_i$ are orthogonal from **Assumption 3a**. From **Assumption 2c**, we have

$$V(\boldsymbol{X}^T\boldsymbol{\epsilon}_i) = \sigma^2\mathrm{E}[\boldsymbol{X}^T\boldsymbol{X}].$$

Iin order to show (A.4.15), we first need the following theorem.

Theorem A.4.3.1 [43]

For random vectors $\boldsymbol{X} = (X_1, \cdots, X_R)$ and $\mathbf{Y} = (Y_1, \cdots, Y_R)$, a necessary and sufficient condition for $\boldsymbol{X} \overset{\text{w}}{\rightsquigarrow} \mathbf{Y}$ is that

$$\sum_{r=1}^{R} t_r X_r \overset{\text{w}}{\rightsquigarrow} \sum_{r=1}^{R} t_r Y_r$$

for each $\mathbf{t} = (t_1, \cdots, t_R)$ in $\mathcal{R}^R$.

Proof The necessity part will be proven first as follows: Define the function $h_t(x) = \mathbf{t} \cdot \boldsymbol{x}$, which maps $\mathcal{R}^R \to R^1$, where $\mathbf{t} \cdot \boldsymbol{x}$ denotes the inner product. Let us assume that this mapping $\mathcal{R}^R \to R^1$ is measurable and the set of D_h of its discontinuities are measurable. If $\Pr\{\boldsymbol{X} \in D_h\} = 0$, $\Pr\{\mathbf{Y} \in D_h\} = 0$ and $\boldsymbol{X} \overset{\text{w}}{\rightsquigarrow} \mathbf{Y}$, then

$$\sum_{r=1}^{R} t_r X_r = h_t(\boldsymbol{X}) \overset{\text{w}}{\rightsquigarrow} h_t(\mathbf{Y}) = \sum_{r=1}^{R} t_r Y_r$$

holds from Continuous Mapping Theorem.

The sufficiency of the proof is as follows: The continuity theorem implies that if for one dimensional characteristic function of $\boldsymbol{X}$, $\mathrm{E}[\exp(is \sum_{r=1}^{R} t_r X_r)]$ converges to that of $\mathbf{Y}$, $\mathrm{E}[\exp(is \sum_{r=1}^{R} t_r Y_r)]$ or

$$\mathrm{E}[\exp(is \sum_{r=1}^{R} t_r X_r)] \overset{\text{w}}{\rightsquigarrow} \mathrm{E}[\exp(is \sum_{r=1}^{R} t_r Y_r)] \quad \text{for all } s,$$

then it follows that $\boldsymbol{X} \overset{\text{w}}{\rightsquigarrow} \mathbf{Y}$. If we let $s = 1$, we immediately know that the characteristic function of $\boldsymbol{X}$ converges pointwisely to that of $\mathbf{Y}$ from the assumption of the Theorem. $\square$

[43] See (Billingsley, 1986, 397).

Let $S_i = \sum_{i=1}^{J} \boldsymbol{x}_{i\cdot}^T \epsilon_i$ for notational convenience and define $\mathbf{Y} = (Y_1, \cdots, Y_R)$ which is normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{B}_{\boldsymbol{Y}}$. For given $\mathbf{t} = (t_1, \cdots, t_R)$, let

$$\begin{aligned} Z_i &= \sum_{r=1}^{R} t_r \cdot \boldsymbol{x}_{i\cdot}^T \epsilon_i, \\ Z &= \sum_{r=1}^{R} t_r \cdot Y_r. \end{aligned}$$

From Theorem A.4.3.1, it suffices to show that

$$J^{-1/2} \sum_{i=1}^{J} Z_i \overset{\mathrm{w}}{\rightsquigarrow} Z \tag{A.4.17}$$

for arbitrary t. We have

$$J^{-1/2} \sum_{i=1}^{J} Z_i = J^{-1/2} \sum_{i=1}^{J} \sum_{r=1}^{R} t_r \cdot \boldsymbol{x}_{i\cdot}^T \epsilon_i = J^{-1/2} \sum_{r=1}^{R} t_r \sum_{i=1}^{J} \boldsymbol{x}_{i\cdot}^T \epsilon_i = J^{-1/2} \sum_{r=1}^{R} t_r S_i.$$

Since variance of S_i is finite from **Assumption 3b**, Lindeberg's condition holds as $\mathrm{E}[(\boldsymbol{X}^T \boldsymbol{\epsilon}_i)(\boldsymbol{X}^T \boldsymbol{\epsilon}_i)^T]\{|\boldsymbol{X}^T \boldsymbol{\epsilon}_i| > \boldsymbol{\epsilon}_i\} \to \mathbf{0}$ as $J \to \infty$ for some constant vector $\boldsymbol{\epsilon}_i > \mathbf{0}$ from **Assumption 3a**. Then Lindeberg-Lévy Central Limit Theorem suggests $J^{-1/2} S_i \overset{\mathrm{w}}{\rightsquigarrow} N(0, \sigma^2)$ for all i. Then condition (A.4.17) holds as

$$J^{-1/2} \sum_{r=1}^{R} t_r S_i \overset{\mathrm{w}}{\rightsquigarrow} \sum_{r=1}^{R} t_r Y_r,$$

for all i. Then from the Theorem A.4.3.1, we have

$$J^{-1/2} \sum_{i=1}^{J} \boldsymbol{x}_{i\cdot}^T \epsilon_i \overset{\mathrm{w}}{\rightsquigarrow} \mathbf{Y}$$

or

$$J^{-1/2} \sum_{i=1}^{J} \boldsymbol{x}_{i\cdot}^T \epsilon_i \overset{\mathrm{w}}{\rightsquigarrow} N(\mathbf{0}, \boldsymbol{B}_{\boldsymbol{X}}).$$

The equation (A.4.14) becomes

$$\sqrt{J}(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \overset{\mathrm{w}}{\rightsquigarrow} N(\mathbf{0}, \boldsymbol{Q}^{-1} \boldsymbol{B}_{\boldsymbol{X}} \boldsymbol{Q}^{-1}).$$

Since $\boldsymbol{B}_{\boldsymbol{X}} = \sigma^2 \boldsymbol{Q}$, we have

$$\sqrt{J}(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \overset{\text{w}}{\rightsquigarrow} N(\mathbf{0}, \sigma^2 \boldsymbol{Q}^{-1}).$$

This tells us that $\widehat{\boldsymbol{\beta}}_{OLS}$ will be normally distributed with mean $\boldsymbol{\beta}$ as sample increases. This property is called asymptotic normality of OLS.

A.5 The Standard Asymptotic Result of Two-Stage Least Squares (2SLS) Estimator

In this section, the property of 2SLS Estimator is presented.

A.5.1 Assumptions of the 2SLS Estimator

Remember a standard linear regression model with J observations in (A.4.1) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with assumptions

$$E(\epsilon_i) = 0,$$
$$\mathrm{Cov}(x_{ir}, \epsilon_i) = 0, \quad i = 1, \ldots, J, \quad r = 1, \ldots, R-1.$$

This time, assume that an ith element of Rth column vector of $\boldsymbol{X}$, x_{iR}, is correlated with the error term ϵ_i, but the other elements of $\boldsymbol{x}_{i\cdot}$, $(x_{i1}, \ldots, x_{i,R-1})$ are uncorrelated with ϵ_i for $i = 1, \ldots, J$. For instance, in equality (A.2.8), it is clear that the unobserved product quality ξ_m is correlated to at least one of $\boldsymbol{x}_{m\cdot}$ such as prices. As we discussed at the end of the previous section, the ordinary least square method would lead to an inconsistent estimator in such a situation because of the correlation between x_{iR} and ϵ_i.

The 2SLS method to be discussed below is one way to find a (weak) consistent estimator even in such case. In this method, what is called "instrumental variable", denoted as $z_{ih}, h = 1, \ldots, M$, is introduced. The instrumental variables are defined as the variables such that

$$\left.\begin{array}{r} \mathrm{Cov}(z_{ih}, x_{iR}) \neq 0 \\ \mathrm{Cov}(z_{ih}, \epsilon_i) = 0 \end{array}\right\} \text{ for } \quad h = 1, \ldots, M, \quad i = 1, \ldots, J.$$

If the variables $(x_{i1}, \ldots, x_{i,R-1})$ satisfy the condition above, they also can serve as the instrumental variable for x_{iR}. Define the matrix

$$\boldsymbol{Z} \equiv (\boldsymbol{z}_{1\cdot}, , \ldots, \boldsymbol{z}_{i\cdot}, , \ldots, \boldsymbol{z}_{J\cdot})^T,$$

where $\boldsymbol{z}_{i\cdot} \equiv (x_{i1}, \ldots, x_{i,R-1}, z_{i1}, \ldots, z_{iM})$ is an $1 \times L$ vector $(L = R-1+M)$ of instrumental variables, whose lth column is denoted as $\boldsymbol{z}_{\cdot l}$. Let us assume, as in the previous section, that for a given l, the elements of $\boldsymbol{z}_{\cdot l}$ are realizations of lth stochastic variables whose moments exist up to second order, and z_{il} is i.i.d. across i. Let us denote its population mean and variance of $\boldsymbol{z}_{\cdot l}$ as $\mu_{\boldsymbol{z}_l}$ and $\sigma_{\boldsymbol{z}_l}$ respectively.

A.5.2 The 2SLS Estimator

In the first step of 2SLS method, the Rth column vector of $\boldsymbol{X}$ is regressed by OLS on the space spanned by the column vector of instrumental variables $\boldsymbol{Z}$ to obtain $\widehat{\boldsymbol{x}_{\cdot R}}$. We can write

$$\widehat{\boldsymbol{x}_{\cdot R}} = \begin{bmatrix} \widehat{x}_{1R} \\ \vdots \\ \widehat{x}_{iR} \\ \vdots \\ \widehat{x}_{JR} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1,R-1} & z_{11} & \cdots & z_{1M} \\ \vdots & \ddots & \vdots & & \ddots & \vdots \\ x_{i1} & \cdots & x_{i,R-1} & z_{i1} & \cdots & z_{iM} \\ \vdots & \ddots & \vdots & & \ddots & \vdots \\ x_{J1} & \cdots & x_{J,R-1} & z_{J1} & \cdots & z_{JM} \end{bmatrix} \boldsymbol{b}^1 = \boldsymbol{Z}\boldsymbol{b}^1,$$

where $\boldsymbol{b}^1 \equiv (b_1{}^1, \ldots, b_L{}^1)^T$ is $L \times 1$ vector of parameters. Define $\widehat{\boldsymbol{X}}$ as $\widehat{\boldsymbol{X}} \equiv (\widehat{\boldsymbol{x}}_{1\cdot}, \ldots, \widehat{\boldsymbol{x}}_{i\cdot}, \ldots, \widehat{\boldsymbol{x}}_{J\cdot})^T$, where $\widehat{\boldsymbol{x}_{i\cdot}} \equiv (x_{i1}, \ldots, x_{i,R-1}, \widehat{x}_{iR})$.

As the second step, $\boldsymbol{y}$ is regressed on $\widehat{\boldsymbol{X}}$ to obtain the 2SLS estimator denoted as $\widehat{\boldsymbol{\beta}}_{2sls}$, as

$$\widehat{\boldsymbol{\beta}}_{2SLS} = (\widehat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \widehat{\boldsymbol{X}}^T \boldsymbol{y}, \tag{A.5.1}$$

assuming that $\widehat{\boldsymbol{X}}^T \boldsymbol{X}$ is invertible. Since $\widehat{\boldsymbol{x}_{\cdot R}}$ is orthogonal projection, $\boldsymbol{b}^1 = (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \widehat{\boldsymbol{x}_{\cdot R}}$, the projection matrix associated with this operation $\boldsymbol{P}_z$ is

$$\boldsymbol{P}_z = \boldsymbol{Z}(\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T.$$

Since $\boldsymbol{P}_z$ is symmetric idempotent, we have

$$\begin{aligned}
\widehat{\boldsymbol{X}}^T \widehat{\boldsymbol{X}} &= (\boldsymbol{P}_z \boldsymbol{X})^T \boldsymbol{P}_z \boldsymbol{X} = \boldsymbol{X}^T \boldsymbol{P}_z{}^T \cdot \boldsymbol{P}_z \boldsymbol{X} \\
&= \boldsymbol{X}^T \boldsymbol{P}_z{}^T \boldsymbol{P}_z{}^T \boldsymbol{X} = \boldsymbol{X}^T \boldsymbol{P}_z{}^T \boldsymbol{X} = (\boldsymbol{P}_z \boldsymbol{X})^T \boldsymbol{X} = \widehat{\boldsymbol{X}}^T \boldsymbol{X}.
\end{aligned}$$

Let us write $\widehat{\boldsymbol{X}}$ using $\boldsymbol{Z}$ as $\widehat{\boldsymbol{X}} = \boldsymbol{Z}\boldsymbol{b}$ as

$$\widehat{\boldsymbol{X}} = \begin{bmatrix} x_{11} & \cdots & x_{1,R-1} & \cdots & \widehat{x}_{1R} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{i,R-1} & \cdots & \widehat{x}_{iR} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{J1} & \cdots & x_{J,R-1} & \cdots & \widehat{x}_{JR} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1,R-1} & z_{11} & \cdots & z_{1M} \\ \vdots & \ddots & \vdots & & \ddots & \vdots \\ x_{i1} & \cdots & x_{i,R-1} & z_{i1} & \cdots & z_{iM} \\ \vdots & \ddots & \vdots & & \ddots & \vdots \\ x_{J1} & \cdots & x_{J,R-1} & z_{J1} & \cdots & z_{JM} \end{bmatrix}.$$

$$\begin{bmatrix} 1 & \ldots & 0 & b_1{}^1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 1 & \vdots \\ 0 & \ldots & 0 & \vdots \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & b_L{}^1 \end{bmatrix} = \boldsymbol{Z}\boldsymbol{b}. \tag{A.5.2}$$

There is $(R-1) \times (R-1)$ identity matrix stacked on the $M \times (R-1)$ zero matrix in the last matrix $\boldsymbol{b}$, which corresponds to $(\boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot R-1})$, whereas $({b_1}^1, \ldots, {b_L}^1)$ in the last matrix corresponds to $\widehat{\boldsymbol{x}_{\cdot R}}$ as the product of $\boldsymbol{Z}$ and $({b_1}^1, \ldots, {b_L}^1)$ produces $\widehat{\boldsymbol{x}_{\cdot R}}$. The quantity $\boldsymbol{b}$ is an $(L \times R)$ matrix

$$\boldsymbol{b} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}, \tag{A.5.3}$$

since $\boldsymbol{Z}\boldsymbol{b} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X} = \boldsymbol{P}_z\boldsymbol{X}=\widehat{\boldsymbol{X}}$. With (A.5.2) and (A.5.3), $\widehat{\boldsymbol{X}}^T\boldsymbol{X}$ can be written as

$$\widehat{\boldsymbol{X}}^T\boldsymbol{X} \;=\; (\boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{X} = \boldsymbol{b}^T\boldsymbol{Z}^T\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}. \tag{A.5.4}$$

In the same manner, $\widehat{\boldsymbol{X}}^T\boldsymbol{y}$ can be written as

$$\widehat{\boldsymbol{X}}^T\boldsymbol{y} \;=\; (\boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{y} = \boldsymbol{b}^T\boldsymbol{Z}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}(\boldsymbol{Z}^T\boldsymbol{y}). \tag{A.5.5}$$

Let us substitute (A.5.4) and (A.5.5) for (A.5.1) to obtain

$$\widehat{\boldsymbol{\beta}}_{2SLS} \;=\; \left(\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)^{-1} \cdot \left(\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}\right). \tag{A.5.6}$$

Substituting (A.4.1) to (A.5.6), we have the 2SLS estimator as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{2SLS} \;=&\; \left(\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)^{-1} \\
&\qquad \cdot \left(\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right) \\
=&\; \left(\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)^{-1} \\
&\qquad \cdot \left(\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)\boldsymbol{\beta} \\
&\qquad\quad + \left(\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)^{-1} \\
&\qquad\qquad \cdot \left(\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon}\right) \\
=&\; \boldsymbol{\beta} + \left(J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)^{-1} \\
&\qquad\quad \cdot \left(J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon}\right).
\end{aligned} \tag{A.5.7}$$

If the second term in the last equation above goes to $\mathbf{0}$ as sample size increases, the 2SLS estimator is consistent, which is the case under certain conditions as we will show in the next subsection.

A.5.3 (Weak) Consistency of the 2SLS Estimator

The (weak) consistency of 2SLS estimator would be shown in this subsection. The assumptions needed for 2SLS estimator to be (weak) consistent are

$$E[(\boldsymbol{x}_{i\cdot}^T \epsilon_i)^T (\boldsymbol{x}_{i\cdot}^T \epsilon_i)] < \infty, \tag{A.5.8}$$

$$\mathrm{E}(\boldsymbol{z}_{i\cdot}{}^T \epsilon_i) = 0, \quad i = 1, \ldots, J, \tag{A.5.9}$$

$$\mathrm{E}[(\boldsymbol{z}_{i\cdot}^T \epsilon_i)^T (\boldsymbol{z}_{i\cdot}^T \epsilon_i)] < \infty, \quad i = 1, \ldots, J, \tag{A.5.10}$$

$$\text{rank } \mathrm{E}(\boldsymbol{Z}^T \boldsymbol{Z}) = L, \tag{A.5.11}$$

$$\text{rank } \mathrm{E}(\boldsymbol{Z}^T \boldsymbol{X}) = R. \tag{A.5.12}$$

As in the last section, let us assume that $\boldsymbol{x}_{\cdot r}$ has population mean $\mu_{\boldsymbol{x}_r}$ and population variance $\sigma_{\boldsymbol{x}_r}{}^2$ for $r = 1, \ldots, R$, and that the joint distribution of $\boldsymbol{x}_{\cdot R}$ and $\boldsymbol{z}_{\cdot l}$ has a moment up to second order. Then matrices $\boldsymbol{X}^T\boldsymbol{Z}, \boldsymbol{Z}^T\boldsymbol{Z}$, and $\boldsymbol{Z}^T\boldsymbol{X}$ are

$$\boldsymbol{X}^T\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{x}_{\cdot 1}{}^T \cdot \boldsymbol{z}_{\cdot 1} & \cdots & \boldsymbol{x}_{\cdot 1}{}^T \cdot \boldsymbol{z}_{\cdot l} & \cdots & \boldsymbol{x}_{\cdot 1}{}^T \cdot \boldsymbol{z}_{\cdot L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{\cdot r}{}^T \cdot \boldsymbol{z}_{\cdot 1} & \cdots & \boldsymbol{x}_{\cdot r}{}^T \cdot \boldsymbol{z}_{\cdot l} & \cdots & \boldsymbol{x}_{\cdot r}{}^T \cdot \boldsymbol{z}_{\cdot L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{\cdot R}{}^T \cdot \boldsymbol{z}_{\cdot 1} & \cdots & \boldsymbol{x}_{\cdot R}{}^T \cdot \boldsymbol{z}_{\cdot l} & \cdots & \boldsymbol{x}_{\cdot R}{}^T \cdot \boldsymbol{z}_{\cdot L} \end{bmatrix},$$

$$\boldsymbol{Z}^T\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_{\cdot 1}{}^T \cdot \boldsymbol{z}_{\cdot 1} & \cdots & \boldsymbol{z}_{\cdot 1}{}^T \cdot \boldsymbol{z}_{\cdot l} & \cdots & \boldsymbol{z}_{\cdot 1}{}^T \cdot \boldsymbol{z}_{\cdot L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{z}_{\cdot l}{}^T \cdot \boldsymbol{z}_{\cdot 1} & \cdots & \boldsymbol{z}_{\cdot l}{}^T \cdot \boldsymbol{z}_{\cdot l} & \cdots & \boldsymbol{z}_{\cdot l}{}^T \cdot \boldsymbol{z}_{\cdot L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{z}_{\cdot L}{}^T \cdot \boldsymbol{z}_{\cdot 1} & \cdots & \boldsymbol{z}_{\cdot L}{}^T \cdot \boldsymbol{z}_{\cdot l} & \cdots & \boldsymbol{z}_{\cdot L}{}^T \cdot \boldsymbol{z}_{\cdot L} \end{bmatrix},$$

and

$$\boldsymbol{Z}^T\boldsymbol{X} = \begin{bmatrix} \boldsymbol{z}_{\cdot 1}{}^T \cdot \boldsymbol{x}_{\cdot 1} & \cdots & \boldsymbol{z}_{\cdot 1}{}^T \cdot \boldsymbol{x}_{\cdot r} & \cdots & \boldsymbol{z}_{\cdot 1}{}^T \cdot \boldsymbol{x}_{\cdot R} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{z}_{\cdot l}{}^T \cdot \boldsymbol{x}_{\cdot 1} & \cdots & \boldsymbol{z}_{\cdot l}{}^T \cdot \boldsymbol{x}_{\cdot r} & \cdots & \boldsymbol{z}_{\cdot l}{}^T \cdot \boldsymbol{x}_{\cdot R} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{z}_{\cdot L}{}^T \cdot \boldsymbol{x}_{\cdot 1} & \cdots & \boldsymbol{z}_{\cdot L}{}^T \cdot \boldsymbol{x}_{\cdot r} & \cdots & \boldsymbol{z}_{\cdot L}{}^T \cdot \boldsymbol{x}_{\cdot R} \end{bmatrix},$$

respectively.

Since elements of $\boldsymbol{x}_{\cdot r}$ and $\boldsymbol{z}_{\cdot l}$ are assumed to be i.i.d. across i, Etemadi's SLLN can be applied, assuming the mean $\mu_{\boldsymbol{x}_r}$ and $\mu_{\boldsymbol{z}_l}$ are finite for all r and l. We have

$$J^{-1}\mathbf{1} \cdot \boldsymbol{x}_{\cdot r} = J^{-1}\sum_{i=1}^{J} x_{ir} \overset{\text{a.s.}}{\to} \mathrm{E}(\boldsymbol{x}_{\cdot r}) = \mu_{\boldsymbol{x}_r}, \quad r = 1, \ldots, R$$

and

$$J^{-1}\mathbf{1} \cdot \boldsymbol{z}_{\cdot l} = J^{-1}\sum_{i=1}^{J} z_{il} \overset{\text{a.s.}}{\to} \mathrm{E}(\boldsymbol{z}_{\cdot l}) = \mu_{\boldsymbol{z}_l}, \quad l = 1, \ldots, L.$$

As stated in the previous section, SLLN can be applied to the sample average of square of $\boldsymbol{z}_{\cdot l}$, and sample average of the product of $\boldsymbol{x}_{\cdot r} \cdot \boldsymbol{z}_{\cdot l}$ as long as their means are finite. The terms $J^{-1}\boldsymbol{z}_{\cdot l}{}^T\boldsymbol{z}_{\cdot l}$ converge as

$$J^{-1}\sum_{i=1}^{J} z_{il}{}^2 \overset{\text{a.s.}}{\to} \mathrm{E}(\boldsymbol{z}_{\cdot l}{}^2) = \mu_{\boldsymbol{z}_l}{}^2 + \sigma_{\boldsymbol{z}_l}{}^2$$

for $l = 1, \ldots, L$, since $\sigma_{\boldsymbol{z}_l}{}^2 = \mathrm{E}(\boldsymbol{z}_{\cdot l}{}^2) - \mu_{\boldsymbol{z}_l}{}^2$. The terms $J^{-1}\boldsymbol{z}_g{}^T\boldsymbol{z}_h$ $(g \neq h)$ converge as

$$J^{-1}\sum_{i=1}^{J}(z_{ig} \cdot z_{ih}) \overset{\text{a.s.}}{\to} \mathrm{E}(\boldsymbol{z}_{\cdot g} \cdot \boldsymbol{z}_{\cdot h}) = \mu_{\boldsymbol{z}_g} \cdot \mu_{\boldsymbol{z}_h} + \sigma_{\boldsymbol{z}_g,\boldsymbol{z}_h}{}^2$$

since $\sigma_{\boldsymbol{z}_g,\boldsymbol{z}_h}{}^2$, the covariance between $\boldsymbol{z}_{\cdot g}$ and $\boldsymbol{z}_{\cdot h}$, equals to $\mathrm{E}(\boldsymbol{z}_{\cdot g} \cdot \boldsymbol{z}_{\cdot h}) - \mu_{\boldsymbol{z}_g} \cdot \mu_{\boldsymbol{z}_h}$. Also $J^{-1}\boldsymbol{x}_{\cdot r}{}^T \cdot \boldsymbol{z}_{\cdot l}$ converges as

$$J^{-1}\sum_{i=1}^{J}(x_{ir} \cdot z_{il}) \overset{\text{a.s.}}{\to} \mathrm{E}(\boldsymbol{x}_{\cdot r} \cdot \boldsymbol{z}_{\cdot l}) = \mu_{\boldsymbol{x}_r} \cdot \mu_{\boldsymbol{z}_l} + \sigma_{\boldsymbol{x}_r,\boldsymbol{z}_l}{}^2$$

for $l = 1, \ldots, L$ and $r = 1, \ldots, R$, since ${\sigma_{\boldsymbol{x}_r \boldsymbol{z}_l}}^2$, the covariance between $\boldsymbol{x}_{\cdot r}$ and $\boldsymbol{z}_{\cdot l}$, equals to $\mathrm{E}(\boldsymbol{x}_{\cdot r} \cdot \boldsymbol{z}_{\cdot l}) - \mu_{\boldsymbol{x}_r} \cdot \mu_{\boldsymbol{z}_l}$, as J increases. The same is true for $J^{-1} {\boldsymbol{z}_{\cdot l}}^T \cdot \boldsymbol{x}_{\cdot r}$. Therefore, $J^{-1}\boldsymbol{X}^T\boldsymbol{Z}$, $J^{-1}\boldsymbol{Z}^T\boldsymbol{Z}$ and $J^{-1}\boldsymbol{Z}^T\boldsymbol{X}$ have non-stochastic probability limits denoted as $\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}$, $\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}$, and $\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}}$ respectively, since almost sure convergence implies convergence in probability. Therefore we have

$$\operatorname*{plim}_{J\to\infty} J^{-1}\boldsymbol{X}^T\boldsymbol{Z} \equiv \boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}} = \begin{bmatrix} \mathrm{E}({\boldsymbol{x}_{\cdot 1}}^T \cdot \boldsymbol{z}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{x}_{\cdot 1}}^T \cdot \boldsymbol{z}_{\cdot l}) & \cdots & \mathrm{E}({\boldsymbol{x}_{\cdot 1}}^T \cdot \boldsymbol{z}_{\cdot L}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathrm{E}({\boldsymbol{x}_{\cdot r}}^T \cdot \boldsymbol{z}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{x}_{\cdot r}}^T \cdot \boldsymbol{z}_{\cdot l}) & \cdots & \mathrm{E}({\boldsymbol{x}_{\cdot r}}^T \cdot \boldsymbol{z}_{\cdot L}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathrm{E}({\boldsymbol{x}_{\cdot R}}^T \cdot \boldsymbol{z}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{x}_{\cdot R}}^T \cdot \boldsymbol{z}_{\cdot l}) & \cdots & \mathrm{E}({\boldsymbol{x}_{\cdot R}}^T \cdot \boldsymbol{z}_{\cdot L}) \end{bmatrix}$$

$$= \begin{bmatrix} \mu_{\boldsymbol{x}_1} \cdot \mu_{\boldsymbol{z}_1} + {\sigma_{\boldsymbol{x}_1,\boldsymbol{z}_1}}^2 & \cdots & \mu_{\boldsymbol{x}_1} \cdot \mu_{\boldsymbol{z}_l} + {\sigma_{\boldsymbol{x}_1,\boldsymbol{z}_l}}^2 & \cdots & \mu_{\boldsymbol{x}_1} \cdot \mu_{\boldsymbol{z}_L} + {\sigma_{\boldsymbol{x}_1,\boldsymbol{z}_L}}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\boldsymbol{x}_r} \cdot \mu_{\boldsymbol{z}_1} + {\sigma_{\boldsymbol{x}_r,\boldsymbol{z}_1}}^2 & \cdots & \mu_{\boldsymbol{x}_r} \cdot \mu_{\boldsymbol{z}_l} + {\sigma_{\boldsymbol{x}_r,\boldsymbol{z}_l}}^2 & \cdots & \mu_{\boldsymbol{x}_r} \cdot \mu_{\boldsymbol{z}_L} + {\sigma_{\boldsymbol{x}_r,\boldsymbol{z}_L}}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\boldsymbol{x}_R} \cdot \mu_{\boldsymbol{z}_1} + {\sigma_{\boldsymbol{x}_R,\boldsymbol{z}_1}}^2 & \cdots & \mu_{\boldsymbol{x}_R} \cdot \mu_{\boldsymbol{z}_l} + {\sigma_{\boldsymbol{x}_R,\boldsymbol{z}_l}}^2 & \cdots & \mu_{\boldsymbol{x}_R} \cdot \mu_{\boldsymbol{z}_L} + {\sigma_{\boldsymbol{x}_R,\boldsymbol{z}_L}}^2 \end{bmatrix},$$

$$\operatorname*{plim}_{J\to\infty} J^{-1}\boldsymbol{Z}^T\boldsymbol{Z} \equiv \boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}} = \begin{bmatrix} \mathrm{E}({\boldsymbol{z}_{\cdot 1}}^T \cdot \boldsymbol{z}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot 1}}^T \cdot \boldsymbol{z}_{\cdot l}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot 1}}^T \cdot \boldsymbol{z}_{\cdot L}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathrm{E}({\boldsymbol{z}_{\cdot l}}^T \cdot \boldsymbol{z}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot l}}^T \cdot \boldsymbol{z}_{\cdot l}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot l}}^T \cdot \boldsymbol{z}_{\cdot L}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathrm{E}({\boldsymbol{z}_{\cdot L}}^T \cdot \boldsymbol{z}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot L}}^T \cdot \boldsymbol{z}_{\cdot l}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot L}}^T \cdot \boldsymbol{z}_{\cdot L}) \end{bmatrix}$$

$$= \begin{bmatrix} {\mu_{\boldsymbol{z}_1}}^2 + {\sigma_{\boldsymbol{z}_1}}^2 & \cdots & \mu_{\boldsymbol{z}_1} \cdot \mu_{\boldsymbol{z}_l} + {\sigma_{\boldsymbol{z}_1,\boldsymbol{z}_l}}^2 & \cdots & \mu_{\boldsymbol{z}_1} \cdot \mu_{\boldsymbol{z}_L} + {\sigma_{\boldsymbol{z}_1,\boldsymbol{z}_L}}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\boldsymbol{z}_l} \cdot \mu_{\boldsymbol{z}_1} + {\sigma_{\boldsymbol{z}_l,\boldsymbol{z}_1}}^2 & \cdots & {\mu_{\boldsymbol{z}_l}}^2 + {\sigma_{\boldsymbol{z}_l}}^2 & \cdots & \mu_{\boldsymbol{z}_l} \cdot \mu_{\boldsymbol{z}_L} + {\sigma_{\boldsymbol{z}_l,\boldsymbol{z}_L}}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\boldsymbol{z}_L} \cdot \mu_{\boldsymbol{z}_1} + {\sigma_{\boldsymbol{z}_L,\boldsymbol{z}_1}}^2 & \cdots & \mu_{\boldsymbol{z}_L} \cdot \mu_{\boldsymbol{z}_l} + {\sigma_{\boldsymbol{z}_L,\boldsymbol{z}_l}}^2 & \cdots & {\mu_{\boldsymbol{z}_L}}^2 + {\sigma_{\boldsymbol{z}_L}}^2 \end{bmatrix},$$

and

$$\operatorname*{plim}_{J\to\infty} J^{-1}\boldsymbol{Z}^T\boldsymbol{X} \equiv \boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}} = \begin{bmatrix} \mathrm{E}({\boldsymbol{z}_{\cdot 1}}^T \cdot \boldsymbol{x}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot 1}}^T \cdot \boldsymbol{x}_{\cdot r}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot 1}}^T \cdot \boldsymbol{x}_{\cdot R}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathrm{E}({\boldsymbol{z}_{\cdot l}}^T \cdot \boldsymbol{x}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot l}}^T \cdot \boldsymbol{x}_{\cdot r}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot l}}^T \cdot \boldsymbol{x}_{\cdot R}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathrm{E}({\boldsymbol{z}_{\cdot L}}^T \cdot \boldsymbol{x}_{\cdot 1}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot L}}^T \cdot \boldsymbol{x}_{\cdot r}) & \cdots & \mathrm{E}({\boldsymbol{z}_{\cdot L}}^T \cdot \boldsymbol{x}_{\cdot R}) \end{bmatrix}$$

$$= \begin{bmatrix} \mu_{\boldsymbol{z}_1} \cdot \mu_{\boldsymbol{x}_1} + {\sigma_{\boldsymbol{z}_1,\boldsymbol{x}_1}}^2 & \cdots & \mu_{\boldsymbol{z}_1} \cdot \mu_{\boldsymbol{x}_r} + {\sigma_{\boldsymbol{z}_1,\boldsymbol{x}_r}}^2 & \cdots & \mu_{\boldsymbol{z}_1} \cdot \mu_{\boldsymbol{x}_R} + {\sigma_{\boldsymbol{z}_1,\boldsymbol{x}_R}}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\boldsymbol{z}_l} \cdot \mu_{\boldsymbol{x}_1} + {\sigma_{\boldsymbol{z}_l,\boldsymbol{x}_1}}^2 & \cdots & \mu_{\boldsymbol{z}_l} \cdot \mu_{\boldsymbol{x}_r} + {\sigma_{\boldsymbol{z}_l,\boldsymbol{x}_r}}^2 & \cdots & \mu_{\boldsymbol{z}_l} \cdot \mu_{\boldsymbol{x}_R} + {\sigma_{\boldsymbol{z}_l,\boldsymbol{x}_R}}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_{\boldsymbol{z}_L} \cdot \mu_{\boldsymbol{x}_1} + {\sigma_{\boldsymbol{z}_L,\boldsymbol{x}_1}}^2 & \cdots & \mu_{\boldsymbol{z}_L} \cdot \mu_{\boldsymbol{x}_r} + {\sigma_{\boldsymbol{z}_L,\boldsymbol{x}_r}}^2 & \cdots & \mu_{\boldsymbol{z}_L} \cdot \mu_{\boldsymbol{x}_R} + {\sigma_{\boldsymbol{z}_L,\boldsymbol{x}_R}}^2 \end{bmatrix}.$$

Likewise, $\boldsymbol{Z}^T\boldsymbol{\epsilon}$ can be written as

$$\boldsymbol{Z}^T\boldsymbol{\epsilon} = \begin{bmatrix} {\boldsymbol{z}_{1\cdot}}^T \dots {\boldsymbol{z}_{i\cdot}}^T \dots, {\boldsymbol{z}_{J\cdot}}^T \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_J \end{bmatrix} = \begin{bmatrix} {\boldsymbol{z}_{\cdot 1}}^T \\ \vdots \\ {\boldsymbol{z}_{\cdot l}}^T \\ \vdots \\ {\boldsymbol{z}_{\cdot L}}^T \end{bmatrix} \boldsymbol{\epsilon}$$

$$= \begin{bmatrix} z_{11} & \cdots & z_{i1} & \cdots & z_{J1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{1l} & \cdots & z_{il} & \cdots & z_{Jl} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{1L} & \cdots & z_{iL} & \cdots & z_{JL} \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_J \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^J z_{i1} \cdot \epsilon_i \\ \vdots \\ \sum_{i=1}^J z_{il} \cdot \epsilon_i \\ \vdots \\ \sum_{i=1}^J z_{iL} \cdot \epsilon_i \end{bmatrix} = \begin{bmatrix} {\boldsymbol{z}_{\cdot 1}}^T\boldsymbol{\epsilon} \\ \vdots \\ {\boldsymbol{z}_{\cdot l}}^T\boldsymbol{\epsilon} \\ \vdots \\ {\boldsymbol{z}_{\cdot L}}^T\boldsymbol{\epsilon} \end{bmatrix}. \tag{A.5.13}$$

If we multiply resulting vector on the right-hand side of the equation in

(A.5.13) by J^{-1}, we can write lth row of $J^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon}$ as

$$J^{-1}\sum_{i=1}^{J} z_{il}\cdot\epsilon_i.$$

From assumption (A.5.9) and Etemadi's SLLN, we have

$$J^{-1}\sum_{i=1}^{J} z_{il}\cdot\epsilon_i \xrightarrow{\text{a.s.}} \mathrm{E}(z_{il}\cdot\epsilon_i) = 0. \tag{A.5.14}$$

Since the equation (A.5.14) holds for every l, $J^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon}$ have nonstochastic probability limit of $\mathbf{0}$ as

$$\operatorname*{plim}_{J\to\infty} J^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon} = \mathbf{0}.$$

The matrix $\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}$ is invertible from assumption (A.5.11). Also, since $\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}}$ is $(R\times L)\times(L\times L)\times(L\times R) = R\times R$ matrix from (A.5.11) and (A.5.12), this is nonsingular matrix and is invertible. Therefore, from matrix inverse rule, we have

$$(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1} \xrightarrow{p} {\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}$$

and

$$[J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1}\boldsymbol{Z}^T\boldsymbol{X}]^{-1} \xrightarrow{p} (\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}})^{-1}.$$

Thus the probability limit of the second term in (A.5.7) becomes $\mathbf{0}$ from Slutsky's theorem as

$$\begin{aligned}
\boldsymbol{\beta} &+ \big(J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1}\boldsymbol{Z}^T\boldsymbol{b}\big)^{-1} \\
&\qquad\cdot\big(J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon}\big) \\
&\xrightarrow{p} \boldsymbol{\beta} + (\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}})^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}.
\end{aligned}$$

Therefore, the consistency of 2SLS estimator is proven. 2SLS estimator is simply called instrumental variable estimator when the number L equal to that of R.

A.5.4 Asymptotic Normality of the 2SLS Estimator

From the (A.5.7), we see that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta} &= \left(J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)^{-1} \\ &\quad \cdot \left(J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1}\boldsymbol{Z}^T\boldsymbol{\epsilon}\right)\end{aligned}$$

and

$$\begin{aligned}\sqrt{J}(\widehat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}) &= \left(J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)^{-1} \\ &\quad \cdot \left(J^{-1}\boldsymbol{X}^T\boldsymbol{Z}(J^{-1}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}J^{-1/2}\boldsymbol{Z}^T\boldsymbol{\epsilon}\right).\end{aligned} \tag{A.5.15}$$

The term $\boldsymbol{Z}^T\boldsymbol{\epsilon}$ on the right-hand side of the equation (A.5.15) can be written as

$$\boldsymbol{Z}^T\boldsymbol{\epsilon} = \textstyle\sum_{i=1}^J \boldsymbol{z}_{i\cdot}^T\epsilon_i = \begin{pmatrix} z_{11} \\ \vdots \\ z_{1l} \\ \vdots \\ z_{1L}\end{pmatrix}\epsilon_1 + \ldots + \begin{pmatrix} z_{i1} \\ \vdots \\ z_{il} \\ \vdots \\ z_{iL}\end{pmatrix}\epsilon_i + \ldots + \begin{pmatrix} z_{J1} \\ \vdots \\ z_{Jl} \\ \vdots \\ z_{JL}\end{pmatrix}\epsilon_J.$$

Therefore, $\boldsymbol{z}_{i\cdot}^T\epsilon_i$ is a sequence of i.i.d. $L\times 1$ vectors such that $\mathrm{E}[(\boldsymbol{z}_{i\cdot}^T\epsilon_i)^T(\boldsymbol{z}_{i\cdot}^T\epsilon_i)]$ is finite from the assumption, and its mean is $\mathbf{0}$ from assumption as well. Then it follows that $\sum_{i=1}^J \boldsymbol{z}_{i\cdot}^T\epsilon_i$ satisfies the condition for Lindeberg-Lévy Central Limit Theorem which requires

$$\mathrm{E}(\boldsymbol{z}_{i\cdot}^T\epsilon_i)^2\{|\boldsymbol{z}_{i\cdot}^T\epsilon_i| > \epsilon\} \to \mathbf{0}$$

for some $\epsilon > 0$. Then $\sum_{i=1}^J \boldsymbol{z}_{i\cdot}^T\epsilon_i$ satisfies the Lindeberg-Lévy Central Limit Theorem as

$$J^{-1/2}\sum_{i=1}^J \boldsymbol{z}_{i\cdot}^T\epsilon_i \xrightarrow{\text{w}} N(\mathbf{0}, \boldsymbol{B}_Z),$$

where $\boldsymbol{B}_Z$ is a variance of $\boldsymbol{Z}^T\boldsymbol{\epsilon}$ which is

$$\begin{aligned}
V(\boldsymbol{Z}^T\boldsymbol{\epsilon}) &= \mathrm{E}[(\boldsymbol{Z}^T\cdot\boldsymbol{\epsilon})^2]-[\mathrm{E}(\boldsymbol{Z}^T\cdot\boldsymbol{\epsilon})]^2=\mathrm{E}[(\boldsymbol{Z}^T\cdot\boldsymbol{\epsilon})^2]-0\\
&= \mathrm{E}[(\boldsymbol{Z}^T\cdot\boldsymbol{\epsilon})(\boldsymbol{Z}^T\cdot\boldsymbol{\epsilon})^T]=\mathrm{E}[(\boldsymbol{Z}^T\cdot\boldsymbol{\epsilon})(\boldsymbol{\epsilon}^T\cdot\boldsymbol{x}_{i\cdot})]\\
&= \mathrm{E}(\boldsymbol{\epsilon}^2\boldsymbol{Z}^T\boldsymbol{Z})=\mathrm{E}(\boldsymbol{\epsilon}^2)\mathrm{E}(\boldsymbol{Z}^T\boldsymbol{Z}).
\end{aligned}$$

It follows that (A.5.15) is asymptotically normally distributed with mean $\mathbf{0}$ and variance matrix $\mathbf{V_{2SLS}}$ as

$$\begin{aligned}
\mathbf{V_{2SLS}} &= (\boldsymbol{Q})^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\left(\mathrm{E}(\boldsymbol{\epsilon}^2)\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}\right)\cdot\left[(\boldsymbol{Q})^{-1}\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\right]^T\\
&= \mathrm{E}(\boldsymbol{\epsilon}^2)(\boldsymbol{Q})^{-1}\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}\cdot\\
&\qquad {\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}}\cdot\left\{\left[\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}}\right]^T\right\}^{-1}\\
&= \mathrm{E}(\boldsymbol{\epsilon}^2)(\boldsymbol{Q})^{-1}\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}}\cdot\left\{\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot{\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}}^{-1}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}}\right\}^{-1}\\
&= \mathrm{E}(\boldsymbol{\epsilon}^2)(\boldsymbol{Q})^{-1},
\end{aligned}$$

since $\mathrm{E}(\boldsymbol{Z}^T\boldsymbol{Z})=\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{Z}}$, where $(\boldsymbol{Q})^{-1}$ is $(\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot\boldsymbol{Q}^{-1}_{\boldsymbol{Z}^T\boldsymbol{Z}}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}})^{-1}$. Again, we assume that variance of error term to be σ^2. Then we can write

$$\sqrt{J}(\widehat{\boldsymbol{\beta}}_{2SLS}-\boldsymbol{\beta})\xrightarrow{\mathrm{w}}N(\mathbf{0},\sigma^2(\boldsymbol{Q}_{\boldsymbol{X}^T\boldsymbol{Z}}\cdot\boldsymbol{Q}^{-1}_{\boldsymbol{Z}^T\boldsymbol{Z}}\cdot\boldsymbol{Q}_{\boldsymbol{Z}^T\boldsymbol{X}})^{-1}).$$

A.6 Scoring (Newton-Raphson) Method

In order to maximize the (log) likelihood function, the algorithm called scoring method, the variant of Newton-Raphson method, can be employed. First we explain Newton-Raphson method. We denote the $\boldsymbol{\beta}$ at $(t+1)$-th iteration by adding superscript as $\boldsymbol{\beta}^{(t+1)}$. In this method, a second-order Taylor expansion of $LL(\boldsymbol{\beta}^{(t+1)})$ around $LL(\boldsymbol{\beta}^{(t)})$ is taken as

$$LL(\boldsymbol{\beta}^{(t+1)})=\left(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\right)^T g_t+\frac{1}{2}\left(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\right)^T H_t\left(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\right), \tag{A.6.1}$$

where $R \times 1$ vector g_t is the gradient at $\boldsymbol{\beta}^{(t)}$

$$g_t = \left(\frac{\partial LL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}$$

and $R \times R$ matrix H_t is matrix of the second derivatives

$$H_t = \left(\frac{\partial g_t}{\partial \boldsymbol{\beta}} \right) \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \left(\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\beta}^{(t+1)}$. The value of $\boldsymbol{\beta}^{(t+1)}$ maximizing (A.6.1) is obtained by setting its derivative zero as

$$\frac{\partial LL\left(\boldsymbol{\beta}^{(t+1)}\right)}{\partial \boldsymbol{\beta}^{(t+1)}} = g_t + H_t \left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right) = \mathbf{0}.$$

This means

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(-H_t^{-1} g_t\right),$$

assuming H_t is invertible.

The scoring method is version of Newton-Raphson method where by the likelihood function $l_s^{(t)}\left(\boldsymbol{\pi}, \boldsymbol{\beta}_s^{(t)} | \boldsymbol{H}, \boldsymbol{Z}\right)$ is replaced with its expected value $\mathrm{E}\left[l_s^{(t)}\left(\boldsymbol{\pi}, \boldsymbol{\beta}_s^{(t)} | \boldsymbol{H}, \boldsymbol{Z}\right)\right]$ to reduce the average number of iterations that can fluctuate from sample to sample if we employ the random $l_s^{(t)}\left(\boldsymbol{\pi}, \boldsymbol{\beta}_s^{(t)} | \boldsymbol{H}, \boldsymbol{Z}\right)$.

A.6.1 Estimating Parameters

In this subsection, we demonstrate how to calculate gradient and Hessian in the standard logit specification. We assume that a panel data of purchase histories for consumers $i = 1, \ldots, N$ who purchase one of $j = 1, \ldots, J$ products at $t_i = 1, \ldots, T_i$ occasions. We also assume that all the products are available for the group of consumers. Remember that the standard logit model, by modifying (2.4.3), can be written as

$$\Pr(j|\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}$$

where $\boldsymbol{x}_{ijt_i}$ is $1 \times R$ vector and $\boldsymbol{\beta}$ is $R \times 1$ parameter vector. Accordingly the likelihood function of the logit model can be written as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} (\Pr(j|\boldsymbol{\beta}))^{y_{ijt_i}},$$

and the log likelihood is written as

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} y_{ijt_i} \log(\Pr(j|\boldsymbol{\beta})) \\ &= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} y_{ijt_i} \log(\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) - \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})) \\ &= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} y_{ijt_i} \left\{ \boldsymbol{x}_{ijt_i}\boldsymbol{\beta} - \log \left(\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) \right) \right\}. \qquad \text{(A.6.2)} \end{aligned}$$

The gradient

Differentiating (A.6.2) with respect to the vector $\boldsymbol{\beta}$, we have tentatively

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} y_{ijt_i} \left\{ \boldsymbol{x}_{ijt_i}^{T} - \frac{\left(\frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\}, \qquad \text{(A.6.3)}$$

since

$$\frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_1} \\ \vdots \\ \frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_R} \end{bmatrix} = \begin{bmatrix} x_{ijt_i1} \\ \vdots \\ x_{ijt_iR} \end{bmatrix} = \boldsymbol{x}_{ijt_i}^{T}.$$

The last term on the right hand side of (A.6.3) is

$$\frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{l=1}^{J} \begin{bmatrix} \frac{\partial \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_R} \end{bmatrix} = \sum_{l=1}^{J} \begin{bmatrix} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) x_{ilt_i 1} \\ \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) x_{ilt_i 2} \\ \vdots \\ \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) x_{ilt_i R} \end{bmatrix}$$

$$= \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) \begin{bmatrix} x_{ilt_i 1} \\ x_{ilt_i 2} \\ \vdots \\ x_{ilt_i R} \end{bmatrix} = \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) \boldsymbol{x}_{ilt_i}^T. \tag{A.6.4}$$

Substituting (A.6.4) back to (A.6.3) yields

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} y_{ijt_i} \left\{ \boldsymbol{x}_{ijt_i}^T - \frac{\sum_{l=1}^{J} (\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) \boldsymbol{x}_{ilt_i}^T)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \left\{ y_{ijt_i} \boldsymbol{x}_{ijt_i}^T - \frac{y_{ijt_i} \sum_{l=1}^{J} \left(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) \boldsymbol{x}_{ilt_i}^T\right)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \left\{ y_{ijt_i} \boldsymbol{x}_{ijt_i}^T - \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i}^T}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \left\{ y_{ijt_i} \boldsymbol{x}_{ijt_i}^T - \Pr_i(j|\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i}^T \right\} \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \left\{ y_{ijt_i} - \Pr_i(j|\boldsymbol{\beta}) \right\} \boldsymbol{x}_{ijt_i}^T.
\end{aligned} \tag{A.6.5}$$

This is $R \times 1$ vector of gradient.

The Hessian

Differentiate (A.6.5) further with respect to $\boldsymbol{\beta}^T$ to obtain tentatively

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \frac{\partial \left(\sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \{y_{ijt_i} - \Pr_i(j|\boldsymbol{\beta})\} \boldsymbol{x}_{ijt_i}^T \right)}{\partial \boldsymbol{\beta}^T} \\
&= -\sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \frac{\partial \Pr_i(j|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \boldsymbol{x}_{ijt_i}^T, \qquad \text{(A.6.6)}
\end{aligned}$$

where from (A.6.4)

$$\frac{\partial \Pr_i(j|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = \frac{\frac{\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) - \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}}{(\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}))^2}. \qquad \text{(A.6.7)}$$

Since the term $\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})/\partial \boldsymbol{\beta}^T$ becomes

$$\begin{aligned}
\frac{\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} &= \left[\frac{\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\partial \beta_R} \right] \\
&= [\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) x_{ijt_i1}, \ldots, \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) x_{ijt_iR}] \\
&= \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})[x_{ijt_i1}, \ldots, x_{ijt_iR}] = \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i},
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} &= \frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}} \cdot \frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}^T} \\
&= \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \cdot \left[\frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_1}, \ldots, \frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_R} \right] \\
&= \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \cdot [x_{ijt_i1}, \ldots, x_{ijt_iR}] \\
&= \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \cdot \boldsymbol{x}_{ijt_i},
\end{aligned}$$

the equation (A.6.7) becomes

$$\begin{aligned}
\frac{\partial \Pr_i(j|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} &= \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i} \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) - \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i}}{(\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}))^2} \\
&= \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i}}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} - \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \boldsymbol{x}_{ijt_i} \\
&= \Pr_i(j|\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i} - \{\Pr_i(j|\boldsymbol{\beta})\}^2 \boldsymbol{x}_{ijt_i}. \qquad \text{(A.6.8)}
\end{aligned}$$

Substituting (A.6.8) back into (A.6.6) yields

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= -\sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\frac{\partial \mathrm{Pr}_i(j|\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\boldsymbol{x}_{ijt_i}^T \\
&= -\sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left(\mathrm{Pr}_i(j|\boldsymbol{\beta})\boldsymbol{x}_{ijt_i} - \{\mathrm{Pr}_i(j|\boldsymbol{\beta})\}^2\boldsymbol{x}_{ijt_i}\right)\boldsymbol{x}_{ijt_i}^T \\
&= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left(\{\mathrm{Pr}_i(j|\boldsymbol{\beta})\}^2\boldsymbol{x}_{ijt_i}^T\boldsymbol{x}_{ijt_i} - \mathrm{Pr}_i(j|\boldsymbol{\beta})\boldsymbol{x}_{ijt_i}^T\boldsymbol{x}_{ijt_i}\right) \\
&= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\{\mathrm{Pr}_i(j|\boldsymbol{\beta})\}^2\boldsymbol{x}_{ijt_i}^T\boldsymbol{x}_{ijt_i} - \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\mathrm{Pr}_i(j|\boldsymbol{\beta})\boldsymbol{x}_{ijt_i}^T\boldsymbol{x}_{ijt_i},
\end{aligned}$$

which is $R \times R$ Hessian matrix.

A.6.2 BHHH Method

One of the alternative methods to Newton-Raphson method is BHHH method which would be introduced in this subsection. The Newton-Raphson method has two major drawbacks: calculating the Hessian is sometimes computation-intensive and it does not guarantee an increase in the log-likelihood if the log-likelihood is not globally concave.

The BHHH method uses a matrix of the outer products of the score as the alternative to the negative Hessian in determining the next step. The score of an observation for consumer i, indexed by $s_i(\beta_r)$, is defined as the derivative of the observation's log-likelihood with respect to the parameter β_r which is in the form of

$$s_i(\beta_r) = \frac{\partial \ln(\mathrm{Pr}(j|\boldsymbol{\beta}))}{\partial\beta_r}.$$

Since the the log likelihood function for a standard logit model is written as

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{J} y_{ij}\left\{\boldsymbol{x}_{ijt_i}\boldsymbol{\beta} - \ln\left(\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})\right)\right\},$$

differentiating it with respect to vector β_r yields

$$s_i(\beta_r) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{j=1}^{J} y_{ij} \left\{ x_{ijt_ir} - \frac{\left(\frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_r}\right)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\}, \tag{A.6.9}$$

since

$$\frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_r} = x_{ijt_ir}.$$

The last term in (A.6.9) is

$$\frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_r} = \sum_{l=1}^{J} \frac{\partial \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_r} = \sum_{l=1}^{J} \left(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) x_{ilt_ir}\right). \tag{A.6.10}$$

Substituting (A.6.10) back to (A.6.9) yields

$$\begin{aligned}
s_i(\beta_r) &= \sum_{j=1}^{J} y_{ij} \left\{ x_{ijt_ir} - \frac{\sum_{l=1}^{J} \left(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) x_{ilt_ir}\right)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{j=1}^{J} \left\{ y_{ij} x_{ijt_ir} - \frac{y_{ij} \sum_{l=1}^{J} \left(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) x_{ilt_ir}\right)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{j=1}^{J} \left\{ y_{ij} x_{ijt_ir} - \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) x_{ijt_ir}}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{j=1}^{J} \left\{ y_{ij} x_{ijt_ir} - \Pr_i(j|\boldsymbol{\beta}) x_{ijt_ir} \right\} = \sum_{j=1}^{J} \left\{ y_{ij} - \Pr_i(j|\boldsymbol{\beta}) \right\} x_{ijt_ir}.
\end{aligned}$$

We repeat the above procedure for $r = 1, \cdots, R$ and stack them as $R \times 1$ vector which we denote $s_i(\boldsymbol{\beta})$. In the BHHH algorithm, the matrix $s_i(\boldsymbol{\beta}) s_i(\boldsymbol{\beta})^T$ is used instead of negative of Hessian in Newton-Raphson method.

A.6.3 Variance of Estimates

The asymptotic covariance for correctly specified model is calculated as

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N(0, -\boldsymbol{H}^{-1})$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator, $\boldsymbol{\beta}^0$ denotes the true value of the parameter and $\boldsymbol{H}$ is the expected Hessian in the population. The negative of this term $-\boldsymbol{H}$ is often called the information matrix (Train 2003). The asymptotic covariance of is $\hat{\boldsymbol{\beta}}$ is $-\boldsymbol{H}^{-1}/N$. In practice, the asymptotic covariance of $\hat{\boldsymbol{\beta}}$ is calculated as $-H^{-1}/N$ where H is the average Hessian in the sample. In calculating the asymptotic covariance of is $\hat{\boldsymbol{\beta}}$, W^{-1}/N and B^{-1}/N are used other than $-H^{-1}/N$, where W is the sample covariance of the scores and B is the sample average of outer product of the scores because it is known that $W \rightarrow -H$ as $N \rightarrow \infty$ and $B \rightarrow -H$ as $N \rightarrow \infty$ at the maximizing value of $\boldsymbol{\beta}$ by information identity (ibid).

For any model for which the expected score is zero at the true value is calculated as

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N(0, \boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1})$$

where $\boldsymbol{V}$ is the variance of scores in the population (ibid). The asymptotic covariance of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}/N$ in this case, and it is valued whether or not model is correctly specified or not. This matrix is called robust covariance matrix for this reason. In practice, $\boldsymbol{V}$ is substituted by W or B and the matrix is calculated as $H^{-1}WH^{-1}$. If model is correctly specified, $\boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}$ reduces to $-\boldsymbol{H}^{-1}$ since $-\boldsymbol{H}^{-1} = \boldsymbol{V}$ by information identity.

A.7 Asymptotic Efficiency of MLE

In our specification, the log likelihood for segment s can be written from (2.4.15) and (2.4.3) as

$$\begin{aligned} l_s(\boldsymbol{\beta}_s|\boldsymbol{H}) &= \sum_{i=1}^{N} h_i(s) \cdot \ln \Pr(H_i|S_i = s; \boldsymbol{\beta}_s) \\ &= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J} \{h_i(s) \cdot y_{ijt_i} \cdot \ln \Pr(Y_{it_i} = j|S_i = s; \boldsymbol{\beta}_s)\} \\ &= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J} \left\{ h_i(s) \cdot y_{ijt_i} \cdot \ln\left(\frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s)}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}_s)}\right)\right\} \end{aligned}$$

and let us denote the term $\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left\{h_i(s) \cdot y_{ijt_i} \cdot \ln\left(\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s)/\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}_s)\right)\right\}$ on the last equality as $\log f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)$ or

$$l_s(\boldsymbol{\beta}_s|\boldsymbol{H}) = \sum_{i=1}^{N} \log f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s).$$

It is known that one of the estimates which solves condition $\partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$ achieves asymptotic efficiency with the variance covariance matrix being the inverse of $R \times R$ Hessian matrix under some regularity conditions. We list one of the standard regularity conditions below.

Let us assume that $\boldsymbol{y}_{it_i}$ defined in (2.4.2) is i.i.d. sample from the density $f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)$ where $\boldsymbol{\beta}_s$ are real-valued and further assume the followings conditions hold.

(a) The parameter space of $\boldsymbol{\beta}_s$, denoted as $\boldsymbol{\mathcal{B}}$, is an open set.

(b) The set $A = \{\boldsymbol{y}_{it_i} : f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s) > 0\}$ is independent of $\boldsymbol{\beta}_s$.

(c) For every $\boldsymbol{y}_{it_i} \in A$, the density $f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)$ is three times differentiable with respect to $\boldsymbol{\beta}_s$ and third derivative is continuous in $\boldsymbol{\beta}_s$.

(d) The integral $\int f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)d\boldsymbol{y}_{it_i}$ can be twice differentiable under the integral sign.

(e) The Fisher Information $I(\boldsymbol{\beta}_s)$

$$I(\boldsymbol{\beta}_s) = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\beta}_s}\log f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)\right]^2 = \int\left(\frac{f^{(1)}(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)}{f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)}\right)^2 f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)d\boldsymbol{y}_{it_i}$$

satisfies $0 < I(\boldsymbol{\beta}_s) < \infty$.

(f) For any $\boldsymbol{\beta}_s^{(0)} \in \mathcal{B}$, there exists a vector of positive number ϵ and a function $M(\boldsymbol{y}_{it_i})$, both of which may depend on $\boldsymbol{\beta}_s^0$, such that

$$\left|\frac{\partial^3\log f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)}{\partial^3\boldsymbol{\beta}_s}\right| \leq M(\boldsymbol{y}_{it_i})$$

for all $\boldsymbol{\beta}_s^0 \in \mathcal{B}$, $\beta_{rs}^0 - \epsilon < \beta_{rs} < \beta_{rs}^0 + \epsilon$ for all r with $\mathrm{E}_{\boldsymbol{\beta}_s^0}[M(\boldsymbol{y}_{it_i})] < \infty$.

Then any consistent sequence $\hat{\boldsymbol{\beta}}_{s(N)} = \hat{\boldsymbol{\beta}}_{s(N)}(\boldsymbol{y}_{1t_1}, \ldots, \boldsymbol{y}_{Nt_N})$ of roots[44] of the likelihood equation satisfies

$$N^{1/2}\left(\hat{\boldsymbol{\beta}}_{s(N)} - \boldsymbol{\beta}_{s(N)}^0\right) \overset{\mathrm{w}}{\rightsquigarrow} N\left(\mathbf{0}, I^{-1}(\boldsymbol{\beta}_s)\right).$$

In our case, the Fisher information is

$$I(\boldsymbol{\beta}_s) = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\beta}_s}\sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\log f(\boldsymbol{y}_{it_i};\boldsymbol{\beta}_s)\right]^2 = \mathrm{E}\left[\left(\frac{\partial l_s(\boldsymbol{\beta}_s|\mathbf{H})}{\partial\boldsymbol{\beta}_s}\right)\left(\frac{\partial l_s(\boldsymbol{\beta}_s|\mathbf{H})}{\partial\boldsymbol{\beta}_s}\right)^T\right].$$

The rightmost term above is the expected value of outer product of scores, which is known to converge to Hessian as sample size increases at true value of $\boldsymbol{\beta}_s^{(0)}$ by Information Identity. The sequence $\hat{\boldsymbol{\beta}}_{s(N)}$ is called an efficient likelihood estimator of $\hat{\boldsymbol{\beta}}_{s(N)}$, and it is typically provided by MLE.

[44]The root is the vector $\hat{\boldsymbol{\beta}}_{s(N)}(\boldsymbol{y}_{1t_1}, \ldots, \boldsymbol{y}_{Nt_N})$ which tends to the vector of true value $\boldsymbol{\beta}_s^0$ in probability under some assumptions.

A.8 The FOC of Profit Functions

The FOC of retailer profit function

Partially differentiating (3.2.9) with respect to each retail price p_j and setting them zero, we have the set of equations as

$$\begin{cases} (p_1 - w_1)\frac{\partial S_1}{\partial p_1} + \cdots + (p_J - w_J)\frac{\partial S_J}{\partial p_1} = \sum_{k=1}^{J} \frac{\partial w_k}{\partial p_1} S_1 - S_1 \\ \vdots \\ (p_1 - w_1)\frac{\partial S_1}{\partial p_J} + \cdots + (p_J - w_J)\frac{\partial S_J}{\partial p_J} = \sum_{k=1}^{J} \frac{\partial w_k}{\partial p_J} S_J - S_J. \end{cases} \tag{A.8.1}$$

dropping subscript t for convenience. Writing (A.8.1) in the matrix and rearranging, we have (3.2.13).

The FOC of manufacturer profit function

The FOC of the profit function for manufacturers is

$$\frac{\partial \pi_{\forall f}}{\partial w_l} = \left[S_l + \sum_{j=1}^{J} \left[(w_j - mc_j) \sum_{k=1}^{J} \frac{\partial S_j}{\partial p_k} \cdot \frac{\partial p_k}{\partial w_l} \right] \right] M = 0 \tag{A.8.2}$$

for $l = 1, \cdots, J$. Then we have the set of equations with the constant M removed as[45]

$$\begin{cases} S_1 + (w_1 - mc_1)\left(\frac{\partial S_1}{\partial p_1} \cdot \frac{\partial p_1}{\partial w_1} + \cdots + \frac{\partial S_1}{\partial p_J} \cdot \frac{\partial p_J}{\partial w_1}\right) + \cdots \\ \quad +(w_J - mc_J)\left(\frac{\partial S_J}{\partial p_1} \cdot \frac{\partial p_1}{\partial w_1} + \cdots + \frac{\partial S_J}{\partial p_J} \cdot \frac{\partial p_J}{\partial w_1}\right) = 0 \\ \vdots \\ S_J + (w_1 - mc_1)\left(\frac{\partial S_1}{\partial p_1} \cdot \frac{\partial p_1}{\partial w_J} + \cdots + \frac{\partial S_1}{\partial p_J} \cdot \frac{\partial p_J}{\partial w_J}\right) + \cdots \\ \quad +(w_J - mc_J)\left(\frac{\partial S_J}{\partial p_1} \cdot \frac{\partial p_1}{\partial w_J} + \cdots + \frac{\partial S_J}{\partial p_J} \cdot \frac{\partial p_J}{\partial w_J}\right) = 0. \end{cases} \tag{A.8.3}$$

[45]Note that it is assumed $\partial mc_j / \partial w_j = 0$, as wholesale price would not affect the cost structure of manufacturers in general.

We rewrite this system in matrix form first as

$$\begin{pmatrix} S_1 \\ \vdots \\ S_J \end{pmatrix} + (w_1 - mc_1) \begin{bmatrix} \frac{\partial p_1}{\partial w_1} & \cdots & \frac{\partial p_J}{\partial w_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_1}{\partial w_J} & \cdots & \frac{\partial p_J}{\partial w_J} \end{bmatrix} \cdot \begin{pmatrix} \frac{\partial S_1}{\partial p_1} \\ \vdots \\ \frac{\partial S_1}{\partial p_J} \end{pmatrix} + \cdots$$

$$+ (w_J - mc_J) \begin{bmatrix} \frac{\partial p_1}{\partial w_1} & \cdots & \frac{\partial p_J}{\partial w_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_1}{\partial w_J} & \cdots & \frac{\partial p_J}{\partial w_J} \end{bmatrix} \cdot \begin{pmatrix} \frac{\partial S_J}{\partial p_1} \\ \vdots \\ \frac{\partial S_J}{\partial p_J} \end{pmatrix} = \mathbf{0}$$

and we then rearrange this column-wise to have

$$\begin{pmatrix} S_1 \\ \vdots \\ S_J \end{pmatrix} + \begin{bmatrix} \frac{\partial p_1}{\partial w_1} & \cdots & \frac{\partial p_J}{\partial w_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_1}{\partial w_J} & \cdots & \frac{\partial p_J}{\partial w_J} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial S_1}{\partial p_1} & \cdots & \frac{\partial S_J}{\partial p_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_1}{\partial p_J} & \cdots & \frac{\partial S_J}{\partial p_J} \end{bmatrix} \cdot \begin{pmatrix} w_1 - mc_1 \\ \vdots \\ w_J - mc_J \end{pmatrix} = \mathbf{0}.$$

Therefore we have

$$\begin{pmatrix} w_1 - mc_1 \\ \vdots \\ w_J - mc_J \end{pmatrix} = - \left[\begin{bmatrix} \frac{\partial p_1}{\partial w_1} & \cdots & \frac{\partial p_J}{\partial w_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_1}{\partial w_J} & \cdots & \frac{\partial p_J}{\partial w_J} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial S_1}{\partial p_1} & \cdots & \frac{\partial S_J}{\partial p_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_1}{\partial p_J} & \cdots & \frac{\partial S_J}{\partial p_J} \end{bmatrix} \right]^{-1} \begin{pmatrix} S_1 \\ \vdots \\ S_J \end{pmatrix}.$$

A.9 Estimation of Parameters of Similarity Index Variable

To illustrate how to obtain γ_{s0}, $\boldsymbol{\gamma}_s$ and r_l for $l = 1, \cdots, L$, suppose that $L = 3$, $Q = 2$ with D_{i1} being gender dummy variable (abbreviated as *gen*) and D_{i2} being age variable. Then we have

$$SD_s = \gamma_{s0} + gen \cdot \gamma_{s1} + age \cdot \gamma_{s2}$$

and

$$
\begin{aligned}
\text{sim}_{kj} \cdot SD_s \;=\;& \frac{\gamma_{s0}}{R} \cdot I_{kj} + \frac{\gamma_{s0} \cdot r_1}{R} \cdot I_{kj1} + \frac{\gamma_{s0} \cdot r_2}{R} \cdot I_{kj2} + \frac{\gamma_{s0} \cdot r_3}{R} \cdot I_{kj3} \\
& + \frac{\gamma_{s1}}{R} \cdot I_{kj} \cdot gen + \frac{\gamma_{s1} \cdot r_1}{R} \cdot I_{kj1} \cdot gen + \frac{\gamma_{s1} \cdot r_2}{R} \cdot I_{kj2} \cdot gen \\
& + \frac{\gamma_{s1} \cdot r_3}{R} \cdot I_{kj3} \cdot gen + \frac{\gamma_{s2}}{R} \cdot I_{kj} \cdot age + \frac{\gamma_{s2} \cdot r_1}{R} \cdot I_{kj1} \cdot age \\
& + \frac{\gamma_{s2} \cdot r_2}{R} \cdot I_{kj2} \cdot age + \frac{\gamma_{s2} \cdot r_3}{R} \cdot I_{kj3} \cdot age.
\end{aligned}
$$

For the notational convenience, let us rewrite the vector of parameters $(\gamma_{s0}/R, \gamma_{s0} \cdot r_1/R, \ldots, \gamma_{s2} \cdot r_3/R)$ as $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{12})$. Then we have

$$
\begin{aligned}
\text{sim}_{kj} \cdot SD_s \;=\;& \psi_1 \cdot I_{kj} + \psi_2 \cdot I_{kj1} + \psi_3 \cdot I_{kj2} + \psi_4 \cdot I_{kj3} \\
& + \psi_5 \cdot I_{kj} \cdot gen + \psi_6 \cdot I_{kj1} \cdot gen + \psi_7 \cdot I_{kj2} \cdot gen + \psi_8 \cdot I_{kj3} \cdot gen \\
& + \psi_9 \cdot I_{kj} \cdot age + \psi_{10} \cdot I_{kj1} \cdot age + \psi_{11} \cdot I_{kj2} \cdot age + \psi_{12} \cdot I_{kj3} \cdot age.
\end{aligned}
$$

Notice that r_1, r_2 and r_3 can be defined by three relative ratios among the components of $\boldsymbol{\psi}$ and the system is inconsistent. For example, parameters involving r_1 are $\widehat{\psi_2}$, $\widehat{\psi_6}$ and $\widehat{\psi_{10}}$ relative to $\widehat{\psi_1}$, $\widehat{\psi_5}$ and $\widehat{\psi_9}$ respectively, parameters involving r_2 are $\widehat{\psi_3}$, $\widehat{\psi_7}$ and $\widehat{\psi_{11}}$ relative to $\widehat{\psi_1}$, $\widehat{\psi_5}$ and $\widehat{\psi_9}$ respectively, and parameters involving r_3 are $\widehat{\psi_4}$, $\widehat{\psi_5}$ and $\widehat{\psi_{12}}$ relative to $\widehat{\psi_1}$, $\widehat{\psi_5}$ and $\widehat{\psi_9}$ respectively or

$$
\begin{array}{ccccc}
\widehat{\psi_1} \cdot r_1 = \widehat{\psi_2} & & \widehat{\psi_1} \cdot r_2 = \widehat{\psi_3} & & \widehat{\psi_1} \cdot r_3 = \widehat{\psi_4} \\
\widehat{\psi_5} \cdot r_1 = \widehat{\psi_6} & , & \widehat{\psi_5} \cdot r_2 = \widehat{\psi_7} & , & \widehat{\psi_5} \cdot r_3 = \widehat{\psi_8} \\
\widehat{\psi_9} \cdot r_1 = \widehat{\psi_{10}} & & \widehat{\psi_9} \cdot r_2 = \widehat{\psi_{11}} & & \widehat{\psi_9} \cdot r_3 = \widehat{\psi_{12}}
\end{array}
$$

or

$$
\begin{bmatrix} \widehat{\psi_1} \\ \widehat{\psi_5} \\ \widehat{\psi_9} \end{bmatrix} \cdot r_1 = \begin{bmatrix} \widehat{\psi_2} \\ \widehat{\psi_6} \\ \widehat{\psi_{10}} \end{bmatrix}, \begin{bmatrix} \widehat{\psi_1} \\ \widehat{\psi_5} \\ \widehat{\psi_9} \end{bmatrix} \cdot r_2 = \begin{bmatrix} \widehat{\psi_3} \\ \widehat{\psi_7} \\ \widehat{\psi_{11}} \end{bmatrix}, \begin{bmatrix} \widehat{\psi_1} \\ \widehat{\psi_5} \\ \widehat{\psi_9} \end{bmatrix} \cdot r_3 = \begin{bmatrix} \widehat{\psi_4} \\ \widehat{\psi_8} \\ \widehat{\psi_{12}} \end{bmatrix}.
$$

These equations will be solvable only if each of these ratios are the same which is not likely to be the case. However, we can estimate r_1, r_2 and r_3 by the method of least squares as

$$\begin{aligned}
\widehat{r_1} &= \frac{\widehat{\psi_1}\cdot\widehat{\psi_2}+\widehat{\psi_5}\cdot\widehat{\psi_6}+\widehat{\psi_9}\cdot\widehat{\psi_{10}}}{\widehat{\psi_1}^2+\widehat{\psi_5}^2+\widehat{\psi_9}^2},\\
\widehat{r_2} &= \frac{\widehat{\psi_1}\cdot\widehat{\psi_3}+\widehat{\psi_5}\cdot\widehat{\psi_7}+\widehat{\psi_9}\cdot\widehat{\psi_{11}}}{\widehat{\psi_1}^2+\widehat{\psi_5}^2+\widehat{\psi_9}^2},\\
\widehat{r_3} &= \frac{\widehat{\psi_1}\cdot\widehat{\psi_4}+\widehat{\psi_5}\cdot\widehat{\psi_8}+\widehat{\psi_9}\cdot\widehat{\psi_{12}}}{\widehat{\psi_1}^2+\widehat{\psi_5}^2+\widehat{\psi_9}^2}.
\end{aligned}$$

After estimating $\widehat{r_1}$ through $\widehat{r_3}$, $\widehat{R}$ can be obtained. Then γ_{s0} through γ_{sQ} can be estimated accordingly by the same manner in estimating r_l, $l = 1, \ldots, L$ as

$$\begin{bmatrix} 1 \\ r_1 \\ r_2 \\ r_3 \end{bmatrix} \cdot \gamma_{s0} = \begin{bmatrix} \widehat{\psi_1} \\ \widehat{\psi_2} \\ \widehat{\psi_3} \\ \widehat{\psi_4} \end{bmatrix} \hat{R}, \begin{bmatrix} 1 \\ r_1 \\ r_2 \\ r_3 \end{bmatrix} \cdot \gamma_{s1} = \begin{bmatrix} \widehat{\psi_5} \\ \widehat{\psi_6} \\ \widehat{\psi_7} \\ \widehat{\psi_8} \end{bmatrix} \hat{R}, \begin{bmatrix} 1 \\ r_1 \\ r_2 \\ r_3 \end{bmatrix} \cdot \gamma_{s1} = \begin{bmatrix} \widehat{\psi_9} \\ \widehat{\psi_1 0} \\ \widehat{\psi_1 1} \\ \widehat{\psi_1 2} \end{bmatrix} \hat{R}.$$

A.10 Nash Bargaining Solution

Bargaining theory addresses the question of how the surplus generated by cooperation will be divided or distributed among the participants. We consider the case of two participants or agents. For example, one agent has goods to sell and another has the opportunity to buy that good; the potential value of the good to the buyer is more than that of the seller, so the difference in the valuation motivates these agents to trade.

Suppose that there are player 1 and 2 whose utilities are denoted by u_1 and u_2. Let us denote the set of possible agreements in terms of utilities of player 1 and 2 by $\boldsymbol{U} = (u_1, u_2)$, which is a convex set and denote the disagreement point which are the utilities obtained if the negotiation fails as $\boldsymbol{d} = (d_1, d_2)$. In such situation, we review the fact that there exists a unique function which satisfies five axioms (Pareto efficiency, symmetry, independence of utility origins, independence of utility units, and independent of irrelevant alternatives). Then we show that finding such a function is equivalent to maximize so-called "Nash product," $(u_1 - d_1)(u_2 - d_2)$.

For the disagreement point $\boldsymbol{d} = (v, h)$ and a constant $k > 0$, we have the following proposition:

Proposition: If we draw the tangent line to the hyperbola of $(u_1 - v)(u_2 - h) = k$ at any point on the boundary of $\boldsymbol{U}$, the length of the two segments on the tangent, from the tangency point to the vertical asymptote, and from the tangency point to the horizontal asymptote is the same.

Proof Let us consider the plane spanned by u_1 (horizontal axis) and u_2 (vertical axis). Then $(u_1 - v)(u_2 - h) = k$ has the vertical asymptote $u_1 = v$ and the horizontal asymptote $u_2 = h$. Further we consider a tangency point (u_1', u_2') on the boundary of $\boldsymbol{U}$. Let us denote the horizontal location of the intersection point of the tangent and $u_2 = h$ as x. Then to prove above Proposition, it suffices to show that the distance between x and u_1 is equal to the distance between u_1 and v.

The slope of the tangent at this point is obtained by taking the derivative of $(u_1 - v)(u_2 - h) = k$ with respect to u_1. Then by $(u_2 - h) + (u_1 - v)\partial u_2/\partial u_1 = 0$, the slope $\partial u_2/\partial u_1$ is $-(u_2 - h)/(u_1 - v)$. Then the tangent line at (u_1', u_2') is

$$u_2 = -\frac{u_2' - h}{u_1' - v}(u_1 - u_1') + u_2'$$

or

$$u_1 = -\frac{u_1' - v}{u_2' - h}(u_2 - u_2') + u_1'. \tag{A.10.1}$$

Plugging $u_2 = h$ into the linear function (A.10.1), we obtain the corresponding point of horizontal axis as $u_1 = 2u_1' - v$ which is x. Then $x - u_1' = u_1' - v$.

Thus u'_1 divides the distance between x and v, forming the two identical triangles with upper-left to (u'_1, u'_2) and lower-right to (u'_1, u'_2). Thus Proposition is proved. $\square$

As the hyperbola would move further away from the origin (or disagreement point) as k increases, we can find a unique point which is furthest away from the origin (or disagreement point) yet still touches the boundary of $\boldsymbol{U}$ at one point. This point is unique (by the independence of irrelevant alternatives) and satisfies axioms. To find such point is equivalent to a constrained maximization problem

$$\max_{u_1,u_2}(u_1 - d_1)(u_2 - d_2)$$
$$\text{s.t. } (u_1, u_2) \in \boldsymbol{U}.$$

The objective function of $(u_1 - d_1)(u_2 - d_2)$ is called Nash product.

A.10.1 Generalized Nash Bargaining Solution

Axiom

Instead of Axiom of Symmetry, the axiom of Reservation of Proportion is suggested here. It says that we can find a tangent line to a boundary of $\boldsymbol{U}$ which has a property that the length of two segments, from the tangency point to the vertical line through $\boldsymbol{d}$ and from the tangency point to the horizontal line through $\boldsymbol{d}$ have a proportion $\lambda : (1 - \lambda)$, where $0 < \lambda < 1$.

Finding f with the new axiom

Consider a hyperbola $(u_1 - v)^{\lambda}(u_2 - h)^{1-\lambda} = k$ with $k > 0$ constant. If we draw the tangent to the hyperbola at any point, the length of the two segments of the tangent, from the tangency point to the vertical line through $\boldsymbol{d}$ and from the tangency point to the horizontal line through $\boldsymbol{d}$ has a proportion $\lambda : (1-\lambda)$. Then we should find the furthest hyperbola from the origin (or $\boldsymbol{d}$). This will give a hyperbola that is tangent to $\boldsymbol{U}$. It follows that the solution point must be the point of tangency between $\boldsymbol{U}$ and hyperbola. Finding furthest hyperbola still touches $\boldsymbol{U}$ is the constrained maximization problem, namely

$$\max_{u_1,u_2}(u_1-d_1)^{\lambda}(u_2-d_2)^{1-\lambda}$$
$$\text{s.t. } (u_1,u_2)\in\boldsymbol{U}$$

as hyperbola moves away from the origin as k increases.

Proof Let us consider the plane spanned by u_1 (horizontal axis) and u_2 (vertical axis). Then $(u_1-v)^{1-\lambda}(u_2-h)^{\lambda}=k$ has the vertical asymptote $u_1=v$ and the horizontal asymptote $u_2=h$. Further we consider a tangency point (u_1',u_2') on the boundary of $\boldsymbol{U}$. Let us denote the horizontal location of the intersection point of the tangent and $u_2=h$ as x. Then to prove above theory, it suffices to show that the distance between x and u_1 is equal to $(1-\lambda)/\lambda(u_1'-v)$.

The slope of the tangent at this point is obtained by taking the derivative of $(u_1-v)^{1-\lambda}(u_2-h)^{\lambda}=k$ with respect to u_1. Then by $\lambda(u_1-v)^{\lambda-1}(u_2-h)^{1-\lambda}+(u_1-v)^{\lambda}(1-\lambda)(u_2-h)^{1-\lambda-1}\partial u_2/\partial u_1=0$, the slope $\partial u_2/\partial u_1$ is

$$-\frac{\lambda}{1-\lambda}\frac{u_2-h}{u_1-v}.$$

Then the tangent line at (u_1',u_2') is

$$u_2=-\frac{\lambda}{1-\lambda}\frac{u_2'-h}{u_1'-v}(u_1-u_1')+u_2'$$

or

$$u_1=-\frac{1-\lambda}{\lambda}\frac{u_1'-v}{u_2'-h}(u_2-u_2')+u_1'. \tag{A.10.2}$$

To obtain x, we plug $u_2 = h$ into (A.10.2) which yield

$$x = (h - u_2') - \frac{1-\lambda}{\lambda}\frac{u_1' - v}{u_2' - h}(h - u_2') + u_1' = \frac{1-\lambda}{\lambda}(u_1' - v) + u_1'$$

which proves the theorem. $\square$

Bibliography

Bawa, K. (1990). Modeling inertia and variety seeking tendencies in brand choice behavior. *Marketing Science*, *9*(3), 263–278.

Berlyne, D. (1970). Novelty, complexity, and hedonic value. *Journal of Marketing Research*, *8*(5), 279–286.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, *25*(2), 242–262.

Besanko, D., Dubé, J.-P., & Gupta, S. (2003). Competitive price discrimination strategies in a vertical channel using aggregate retail data. *Management Science*, *49*(9), 1121–1138.

Besanko, D., Gupta, S., & Jain, D. (1998). Logit demand estimation under competitive pricing behavior: An equilibrium framework. *Management Science*, *44*(11-part-1), 1533–1547.

Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, *35*(7), 677–693.

Billingsley, P. (1986). *Probability and Measure*. John Willey.

Brownstone, D., Bunch, D. S., & Train, K. (2000). Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research Part B: Methodological*, *34*(5), 315–338.

Bucklin, R. E., Russell, G. J., & Srinivasan, V. (1998). A relationship between market share elasticities and brand switching probabilities. *Journal of Marketing Research*, *35*(1), 99–113.

Che, H., Sudhir, K., & Seetharaman, P. (2007). Bounded rationality in pricing under state-dependent demand: Do firms look ahead, and if so, how far? *Journal of Marketing Research*, *44*(3), 434–449.

Chintagunta, P. K. (1998). Inertia and variety seeking in a model of brand-purchase timing. *Marketing Science*, *17*(3), 253–270.

Choi, S. C. (1991). Price competition in a channel structure with a common retailer. *Marketing Science*, *10*(4), 271–296.

Draganska, M. & Jain, D. (2004). A likelihood approach to estimating market equilibrium models. *Management Science*, *50*(5), 605–616.

Draganska, M., Klapper, D., & Villas-Boas, S. B. (2010). A larger slice or a larger pie? an empirical investigation of bargaining power in the distribution channel. *Marketing Science*, *29*(1), 57–74.

Dubé, J.-P., Hitsch, G., & Rossi, P. (2009). Do switching costs make markets less competitive? *Journal of Marketing Research*, *46*(5), 435–445.

Dukes, A. J., Gal-Or, E., & Srinivasan, K. (2006). Channel bargaining with retailer asymmetry. *Journal of Marketing Research*, *43*(1), 84–97.

Erdem, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing Science*, *15*(4), 359–378.

Farrell, J. & Klemperer, P. (2007). Coordination and lock-in: Competition with switching costs and network effects. *Handbook of industrial organization*, *3*, 1967–2072.

Freimer, M. & Horsky, D. (2012). Periodic advertising pulsing in a competitive market. *Marketing Science*, *31*(4), 637–648.

Givon, M. (1984). Variety seeking through brand switching. *Marketing Science*, *3*(1), 1–22.

Guadagni, P. M. & Little, J. D. C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science*, *2*(3), 203–238.

Gupta, S., Chintagunta, P. K., & Wittink, D. R. (1997). Household heterogeneity and state dependence in a model of purchase strings: Empirical results and managerial implications. *International Journal of Research in Marketing*, *14*(4), 341–357.

Jeuland, A. P. (1979). Brand choice inertia as one aspect of the notion of brand loyalty. *Management Science*, *25*(7), 671–682.

Kamakura, W. A. & Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, *26*(4), 379–390.

Keane, M. P. (1997). Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business and Economic Statistics*, *15*(3), 310–327.

Kim, C. (2010). A change in channel relations in grocery industry and the industrial organization theory approach to distribution channels. *The Business review (Osaka City University)*, *60*(4), 189–208.

Klemperer, P. (1987a). The competitiveness of markets with switching costs. *The RAND Journal of Economics*, 138–150.

Klemperer, P. (1987b). Entry deterrence in markets with consumer switching costs. *The Economic Journal*, *97*, 99–117.

Lattin, J. M. (1987). A model of balanced choice behavior. *Marketing Science*, *6*(1), 48–65.

Lattin, J. M. & McAlister, L. (1985). Using a variety-seeking model to identify substitute and complementary relationships among competing products. *Journal of Marketing Research*, *22*(3), 330–339.

Luce, R. D. (1959). *Individual Choice Behavior a Theoretical Analysis*. john Wiley and Sons.

McAlister, L. (1982). A dynamic attribute satiation model of variety-seeking behavior. *Journal of Consumer Research*, *9*(2), 141–150.

McFadden, D. (1974). *Conditional Logit Analysis of Qualitative Choice Behavior*. New York: Academic Press.

McFadden, D. & Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, *15*(5), 447–470.

Misra, S. & Mohanty, S. K. (2008). Estimating bargaining games in distribution channels. Technical report, Citeseer.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, *69*(2), 307–342.

Raju, J. & Zhang, Z. J. (2005). Channel coordination in the presence of a dominant retailer. *Marketing Science*, *24*(2), 254–262.

Revelt, D. & Train, K. (1998). Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of economics and statistics*, *80*(4), 647–657.

Roy, R., Chintagunta, P. K., & Haldar, S. (1996). A framework for investigating habits, "the hand of the past," and heterogeneity in dynamic brand choice. *Marketing Science*, *15*(3), 280–299.

Seetharaman, P. (2003). Probabilistic versus random-utility models of state dependence: an empirical comparison. *International Journal of Research in Marketing*, *20*(1), 87–96.

Seetharaman, P. & Chintagunta, P. (1998). A model of inertia and variety-seeking with marketing variables. *International Journal of Research in Marketing*, *15*(1), 1–17.

Slade, M. E. (1995). Product rivalry with multiple strategic weapons: An analysis of price and advertising competition. *Journal of Economics & Management Strategy*, *4*(3), 445–476.

Sudhir, K. (2001). Structural analysis of manufacturer pricing in the presence of a strategic retailer. *Marketing Science*, *20*(3), 244–264.

Train, K. (2003). *Discrete choice methods with simulation*. Cambridge Univ Pr.

Trivedi, M., Bass, F. M., & Rao, R. C. (1994). A model of stochastic variety-seeking. *Marketing Science*, *13*(3), 274–297.

Villas-Boas, J. M. & Winer, R. S. (1999). Endogeneity in brand choice models. *Management Science*, *45*(10), 1324–1338.

Villas-Boas, J. M. & Zhao, Y. (2005). Retailer, manufacturers, and individual consumers: Modeling the supply side in the ketchup marketplace. *Journal of Marketing Research*, *42*(1), 83–95.

Yang, S., Chen, Y., & Allenby, G. M. (2003). Bayesian analysis of simultaneous demand and supply. *Quantitative Marketing and Economics*, *1*(3), 251–275.

A List of Papers

Peer Reviewed Papers

Kamai, Tomohito, and Yuichiro Kanazawa. (2016). Is product with a special feature still rewarding? The case of the Japanese yogurt market. *Cogent Economics Finance 4.1 (2016): 1221231.*

Discussion Papers

Kamai, Tomohito, and Yuichiro Kanazawa. (2012). Understanding Households Complex Brand Choice Behaviors: Variety-seeking, Inertia, or Hybrid Behavior?. *Discussion Paper No.1300 (University of Tsukuba).*

Presented Papers

Kamai, Tomohito, and Yuichiro Kanazawa. (2012). The Latent Class Model of Brand Choice Behaviors. *RIMS Kôkyûroku 1804: 144-161.*