Demand Models with Price Endogeneity and Advertising

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von
Christoph Nagel
Juli 2010

Abteilungssprecher:	Prof. Dr. Martin Peitz
Referent:	Prof. Dr. Carsten Trenkler
Korreferent:	Prof. Stefan Hoderlein, Ph.D.
Tag der mündlichen Prüfung:	21. Oktober 2010

Acknowledgments

I would like to thank my supervisor Stefan Hoderlein for letting me follow my own research ideas and the encouraging and fruitful feedback he gave me. Despite busy times in Mannheim and abroad, he was very much supporting me, widened my perspective towards econometrics and introduced me to many enlightening scholars.

I would also like to thank Carsten Trenkler for his insightful comments on my work and his immediate willingness to join the thesis committee. Further, I would like to thank Enno Mammen for his support and important comments.

I am especially grateful to Elu von Thadden for giving me the opportunity to spend valuable time at the Economics Department of the University College London, CEMMAP and IFS, London.

I am indebted to Kai Kopperschmidt, Martin Schniedermeier and Marc Rossbach of A.C. Nielsen for supplying the dataset used in this work.

I very much enjoyed interacting with my colleagues at the Chair for Statistics and the graduate program of CDSE. I would like to thank Steffen Reinhold and Hannes Ullrich for instructive comments on the third and fourth chapter of this dissertation, Christoph Rothe for inspiring discussions about econometrics. I also thank Renate Bent for always solving my administrative problems.

Many other colleagues made my stay in Mannheim a memorable time, especially Gonzague Vannoorenberghe, Daniel Harenberg, Christoph Rothe, Melanie Schienle, Michal Kowalik, Björn Sass, Heiner Schumacher and Malte Hübner.

My warmest thanks go to my family, in particular to my wife and my children for supporting me during the past years to make my research ideas come true.

Christoph Nagel

Contents

1 Introduction 1

2 Introducing a Household Panel Dataset with Consumption and Advertising 5

2.1 Introduction 5

2.2 Data Structure, Collection and Use for Research 6

2.2.1 Collection Process 9

2.2.2 Using the Dataset for Research 11

2.2.3 Relation to other Datasets 11

2.3 Data Description and Details by File 12

2.3.1 General Purchase Data 13

2.3.1.1 Data Modifications 13

2.3.1.2 Remarks on created Variables 14

2.3.1.3 Summary Statistics 15

2.3.2 Category Purchase Data 16

2.3.2.1 Data Modifications 18

2.3.2.2 Remarks on created Variables 19

2.3.2.3 Summary Statistics 20

2.3.3 Demographic Data 23

2.3.3.1 Data Modifications 25

2.3.3.2 Remarks on created Variables 25

2.3.3.3 Summary Statistics 26

2.3.4 Advertising Data 29

2.3.4.1 Data Modifications 29

2.3.4.2 Remarks on created Variables 30
2.3.4.3 Summary Statistics 30
2.4 Combined Dataset 33
2.4.1 Construction 33
2.4.2 Inferring Prices in Step 5 35
2.4.3 Remarks on created Variables 37
2.4.4 Basic Analysis 38
2.5 Conclusion 42
Appendix 43
Appendix A: Tables and Figures 43
Appendix B: Value Labels of Coded Variables 46

3 Estimating the Magnitude of Causes for Consumer Price Endogeneity 49
3.1 Introduction 49
3.1.1 Related Literature 52
3.2 Model 54
3.2.1 The Empirical Model of Choice 55
3.2.2 Endogeneity Correction Model 60
3.2.3 Causes of Price Endogeneity 62
3.2.3.1 Industry Structure 62
3.2.3.2 Cause 1 - Product Characteristics and Variety Information 63
3.2.3.3 Cause 2 - Retail Activity 65
3.2.3.4 Cause 3 - TV Advertising 66
3.2.3.5 Cause 4 - Individual Inventories 68
3.2.3.6 Cause 5 - State Dependence 69
3.2.3.7 Point 6 - Consumer Heterogeneity 70
3.2.3.8 Point 7 - Data Aggregation 71
3.2.4 Choice and Discussion of Instrumental Variables 71
3.2.5 Method and Interpretation 73
3.3 Data 74
3.3.1 Data Overview 74

3.3.1.1 Relation to other Datasets 79
3.3.2 Variables of Interest in the Analysis 79
3.3.3 Construction of Instrumental Variables 82
3.3.4 Definition of Estimation Samples 83
3.4 Results 84
3.4.1 Price Elasticity Results 84
3.4.2 Parameter Estimates and Estimation Details 88
3.4.3 Specifications with nominal Prices 90
3.4.3.1 Robustness Checks 91
3.4.3.2 Comparison to PT Setup 92
3.4.4 Specifications with efficiency Prices 92
3.4.5 Relation to Literature 93
3.5 Conclusion 94
Appendix 96
Appendix A: Model Details - Estimation of Pooled Case 96
Appendix B: Tables 96
Elasticity Estimates 99
Parameter Estimates 102

4 Local Advertising in a Discrete Choice Demand Model 117

4.1 Introduction 117
4.2 Model 119
4.2.1 Product dummies 123
4.2.2 Brand dummies 125
4.2.3 Instruments 125
4.2.4 The Simple Logit Model 127
4.3 Data 127
4.3.1 Data Overview 127
4.3.2 Advertising 129
4.3.3 Construction of a Market Level Dataset 130
4.4 Results 134

4.4.1 Simple Logit Model 134

4.4.2 Random Coefficients Model 140

4.4.2.1 Results without Demographics 142

4.4.2.2 Results with Demographics 146

4.4.3 Discussion of Details for other Model Variants 149

4.5 Conclusion 151

Appendix 153

Appendix A: Berry's (1994) Model 153

Appendix B: Tables and Figures 154

Model Setup Information 154

Simple Logit Model Results 160

Random Coefficient Model Results without Demographics 166

Random Coefficient Model Results with Demographics 169

Bibliography **175**

List of Figures

2.1 Geographic Distribution of Households according to the Zip Code of Residence . . . 8

2.2 Purchases per Month per Household . . . 21

2.3 Purchases in Sample per Household . . . 21

2.4 Sample Membership Duration per Household . . . 22

2.5 Net Monthly Household Income . . . 27

2.6 Occupational Status of Households . . . 27

2.7 Education of Households . . . 28

2.8 Advertisement Contacts per Household . . . 31

2.9 Advertisement Contacts broken by TV Stations . . . 32

2.10 Advertisements by Day time Windows . . . 33

2.11 Geographic Distribution of Purchase Trips according to the Zip Code of Trip Destination . . . 43

4.1 Overview of 30 Runs for Setup 116 without Demographics with logarithmized Advertising Variable . . . 145

4.2 Overview of 9 Runs for Setup 116 with Demographics with logarithmized Advertising Variable . . . 148

4.3 Overview of 30 Runs for Setup 49 without Demographics with logarithmized Advertising Variable . . . 166

4.4 Overview of 40 Runs for Setup 47 without Demographics with logarithmized Advertising Variable . . . 167

4.5 Overview of 40 Runs for Setup 111 without Demographics with logarithmized Advertising Variable . . . 168

4.6 Overview of 15 Runs for Setup 49 with Demographics with logarithmized Advertising Variable . . . 172

4.7 Overview of 30 Runs for Setup 47 with Demographics with logarithmized Advertising Variable 173

4.8 Overview of 30 Runs for Setup 111 with Demographics with logarithmized Advertising Variable 174

List of Tables

2.1 Overview of Consumer Data Files 7

2.2 Number of Households by Required Information 7

2.3 Overview of Variables in General Purchase Data 13

2.4 Summary Statistics for General Purchase Data 15

2.5 Overview of Variables in Category Purchase Data 16

2.6 Summary Statistics for Category Purchase Data 20

2.7 Summary Statistics for Category Purchase Data 22

2.8 Overview of Variables in Sociogemographic Data 24

2.9 Summary Statistics for Sociodemographic Data 26

2.10 Overview of Variables in Advertisement Contact Data 29

2.11 Campaign Information Data 30

2.12 Summary Statistics for Advertising Data 31

2.13 Detergent Market - Market Shares, Advertising and Characteristics for all Product Purchases broken up by Brand 38

2.14 Linear Probability Model and Discrete Choice Logit Model with all Product Alternatives 40

2.15 Linear Probability Model and Discrete Choice Logit Model without no-purchase alternatives 44

2.16 Linear Probability Model and Discrete Choice Logit Model without no-purchase alternatives and private labels 45

2.17 Overview of Value Label for variables `hhfpsex, psex1, ..., psex8` and `stand`: gender & famstatus 46

2.18 Overview of Value Label for variable `bstel`: occupation 46

2.19 Overview of Value Label for variable `schulab`: degree 47

2.20 Overview of Value Label for variables `ng` and `bik`: nielsenarea & community . 47

2.21 Overview of Value Label for variables `duft`, `kons` and `uwg`: duftcode, konscode & uwgcode . . . 48

3.1 Overview of Consumer Data Files . . . 75
3.2 Number of Households by Required Information . . . 75
3.3 Summary Statistics of Household level Raw Data . . . 76
3.4 Market Shares and Means per Brand for the full Estimation Sample . . . 77
3.5 Characteristics for all Product Purchases broken up by Brand . . . 78
3.6 Empirical Hypothesis and Summary of Results . . . 84
3.7 Mean own price Elasticities - Full sample . . . 85
3.8 Alternative mean own price Elasticities - Full sample . . . 86
3.9 Mean own price Elasticities - Advertising Sample . . . 87
3.10 Mean own price Elasticities - Advertising Sample - Efficiency Prices . . . 88
3.11 Overview of Variables . . . 96
3.12 Own price Elasticities per brand - Full Sample . . . 99
3.13 Own price Elasticities per brand - Advertising Sample . . . 99
3.14 Own price Elasticities per brand - Advertising Sample . . . 100
3.15 Own price Elasticities per brand - Full Sample - Efficiency Prices . . . 100
3.16 Own price Elasticities per brand - Advertising Sample - Efficiency Prices 101
3.17 Own price Elasticities per brand - Advertising Sample - Efficiency Prices 101
3.18 Multiple RC IV Mixed Logit - Basic Models . . . 102
3.19 Multiple RC IV Mixed Logit - Inventory and Varieties . . . 103
3.20 Multiple RC IV Mixed Logit - Habit Formation . . . 104
3.21 Multiple RC IV Mixed Logit - Advertising . . . 105
3.22 Multiple RC IV Mixed Logit - Combined Models . . . 106
3.23 Multiple RC IV Mixed Logit - Combined Models continued . . . 107
3.24 Multiple RC IV Mixed Logit - Stability across Samples . . . 108
3.25 Multiple RC IV Mixed Logit - Varying Number of Choice Alternatives . 109
3.26 Petrin and Train Setup - Liquid Detergents Weekly . . . 110
3.27 Multiple RC IV Mixed Logit - Efficiency Prices - Basic Models . . . 111
3.28 Multiple RC IV Mixed Logit - Efficiency Prices - Inventory and Varieties 112
3.29 Multiple RC IV Mixed Logit - Efficiency Prices - Habit Formation . . . 113

3.30 Multiple RC IV Mixed Logit - Efficiency Prices - Advertising 114

3.31 Multiple RC IV Mixed Logit - Efficiency Prices - Combined Models . . . 115

3.32 Multiple RC IV Mixed Logit - Efficiency Prices - Combined Models continued . . . 116

4.1 Overview - Market Shares / Means broken up by Brand . . . 128

4.2 Characteristics of all Purchases broken up by Brand . . . 129

4.3 Variation of Prices and Marketing . . . 130

4.4 Overview of Product Definition Setups . . . 133

4.5 Logit Market Models for specific Setup . . . 135

4.6 Frequency of Variable Significance for all 120 Setups . . . 136

4.7 Frequency of Variable Significance without "small" Market Shares for all 120 Setups . . . 137

4.8 Frequency of Advertising Significance for all 120 Setups . . . 138

4.9 Frequency of Advertising Significance for all 120 Setups . . . 139

4.10 Overview of Market Data Setups . . . 141

4.11 Random Coefficients Model with Setup 116 and 49 without Demographics and logarithmized Advertising Totals (ad20) . . . 142

4.12 Random Coefficients Model with Setup 47 and 111 without Demographics and logarithmized Advertising Totals (ad20) . . . 143

4.13 Random Coefficients Model with Setup 116 with Demography and logarithmized Advertising Totals (ad20) . . . 147

4.14 Overview of Advertising Variable Definitions . . . 154

4.15 Overview of Market Data Setups . . . 155

4.16 Overview of t-statistics for Price, Retail and Advertising for all Market Setups given two different Minimal Sales Requirements per Product . . . 160

4.17 Frequency of Advertising Significance for all 120 Setups . . . 165

4.18 Frequency of Advertising Significance for all 120 Setups . . . 165

4.19 Random Coefficients Model with Setup 49 with Demographics and logarithmized Advertising Totals (ad20) . . . 169

4.20 Random Coefficients Model with Setup 47 with Demographics and logarithmized Advertising Totals (ad20) . . . 170

4.21 Random Coefficients Model with Setup 111 with Demographics and logarithmized Advertising Totals (ad20) . . . 171

Chapter 1

Introduction

In this dissertation, I estimate two differentiated products demand models with price endogeneity and advertising using a novel individual level panel dataset. Trying to understand how consumers do their consumption decisions has a long history in the economics and business literature. There exist several modeling approaches for this elementary economic event. The classical demand model for homogeneous goods is what economists have used to explain and study demand. Its well developed mathematical architecture does not allow for product characteristics, but it is well suited to model demand for many different products and understand household consumption as a whole. Typically, in applications demand functions are directly modeled without building demand from the individual, potentially heterogeneous, consumer.

Starting with Lancaster (1966), economists have emphasized the role of product characteristics for the demand of goods. This has led to models that describe demand for differentiated products, for substitutes with different product dimensions.[1] The classical and differentiated model approaches are summarized in Blundell (1988) and Anderson, de Palma, and Thisse (1992), respectively.[2] With the work of Berry (1994) a differentiated products demand model based on the probabilistic choice model of McFadden (1974) has become popular in research and among practitioners. In that model demand is built bottom up from individual heterogeneous consumers that maximize their utility of consumption, precisely speaking of their consumption of the product characteristics. For the estimation only market level data is required. When individual level data are available, the model of McFadden and Train (2000) that has been extended by Chintagunta, Dubé, and Goh (2005) and Petrin and Train (2006) describes individual consumers. In order to derive market demand individuals can be aggregated without re-

[1] In these models solely vertical differentiation is considered, see Shaked and Sutton (1983).

[2] There are of course other types of models to tackle demand. Among these are the hedonic model of Heckman, Matzkin, and Nesheim (2005), the nonprobabilistic characteristics approach of Berry and Pakes (2007), a nonparametric approach of Hoderlein (2007) or count processes of Bijwaard, Franses, and Paap (2003) to name a few examples.

strictions. Both models, estimated with market or individual level data, require solving high dimensional integrals which makes estimation difficult. Classical demand models ignore product characteristics and heterogeneous consumers and traditionally analytical closed form solutions are preferred.

Upcoming simulation methods, which became feasible with increasing computing power, fostered to break with the tradition of analytical closed form solutions. This allowed to consider a new class of models that previously could not be estimated without restrictions, such as the differentiated products demand models estimated in this work.

Coming back to the foundation of these models, the differentiated products demand model explicitly postulates the existence of an unobserved product characteristic. The existence of the unobserved product characteristic, if not controlled for in the econometric model, causes an endogeneity problem following conventional omitted variable bias arguments. Advertising has been frequently mentioned as example for an unobserved product characteristic.[3] Especially, economists have perceived endogeneity of prices as a cause for concern, for price is the central economic variable controlling demand. For example, competition and merger analyses rely on estimating the impact of price correctly to give the right competition policy recommendations. Yet, the focus has been on alleviating the price endogeneity by using econometric correction routines based on instrumental variables or control function approaches. Recall that the type of endogeneity tackled with these routines is different from the classical endogeneity problem in traditional demand analysis, where the simultaneity of demand and supply requires instrumental variables that solely affect the supply side to identify the demand curve.

Parallel to the development of the above mentioned models, the availability of datasets and especially their informational richness has increased. While in the earlier days only aggregate figures were available for certain products or markets, now individuals are tracked to record detailed consumption profiles. The profiles contain variables that have previously neither been recorded at this level of detail nor been collected simultaneously in this combination. This has of course urged the development of suitable models to process this type of data. Moreover, as the level of detail has become high, previously unobserved variables in existing models that could have caused an endogeneity are now actually observed by the researcher.

These two developments, on the one hand the evolution of econometric methods to cope with endogeneity and on the other hand the increasing richness of data, are both beneficial. Still, it is necessary to analyze how both benefits will impact the results in existing econometric models. For example, does richer data used in an existing model change the previous results? Is part of the estimation method used in the model obsolete if the better data is available? Does the better data matter in an existing model? This

[3] See the work of Berry (1994), Berry, Levinsohn, and Pakes (1995), Nevo (2001) and Petrin and Train (2006). Villas-Boas and Winer (1999) were one of the first to consider price endogeneity on the individual level.

selective list of questions serves the purpose of emphasizing the simple questions outside of a specific application that arise once the research complexity increases due to superior methods and richer data.

In this spirit, I contribute to the literature by using existing differentiated products demand models with a novel dataset. I augment the existing models by using information on advertising exposure of consumers that is closely linked to price endogeneity. Therefore, I contribute to the literature on price endogeneity and advertising in differentiated products demand models. I ask new questions in existing models that can be solely answered by using the superior data at hand. I use the identical data for two models that operate on different data levels, but both models suffer from the same type of price endogeneity. They share the same core element ruling the behavior of heterogeneous consumers, McFadden's (1974) probabilistic choice model augmented by random coefficients. In both models, advertising is one of the frequently mentioned candidates to cause price endogeneity.

In detail, this dissertation consists of three main chapters following this introduction. The focus lies on two applications for a novel individual level dataset. Chapter 2 contains detailed information on the dataset that is used to estimate two differentiated product demand models in the chapters 3 and 4. The model in chapter 3 is an individual level model, whereas in chapter 4 the data is modified to study the impact for a market level model. In the following I outline the three chapters.

In chapter 2, I give an overview of a German dataset collected by the marketing research company A.C. Nielsen. The chapter is intended for potential users of the dataset and contains numerous details. Therefore, the focus is on describing and not on analyzing the data. The data highlight the recent tendency to have very detailed profiles of tracked households. The data comprise a panel of consumer households over a period of two years. Product purchases and the associated exposure to TV advertisements are recorded. The main feature of the dataset is a nationwide collection of this information for the same households with a matured technology. The purchases are available for two product categories: detergents and chocolate. The data comprise a record of the advertisement contacts of each household for all advertisements that were nationally broadcast on TV. Moreover, the dataset includes common sociodemographic information on the households. I give file descriptions, detailed variable discussions and descriptive statistics. I outline the steps necessary to prepare the data for discrete choice analysis. Finally, a basic analysis with binary choice and discrete choice models demonstrates the possibilities and usefulness of the dataset. Data of the same type has been used recently by Erdem and Keane (1996), Ackerberg (2001, 2003) and Shum (2004).

In chapter 3, I study five causes for the price endogeneity bias in differentiated products demand models. Previous work by Chintagunta, Dubé, and Goh (2005) and Petrin and Train (2006) tackled the bias by suggesting a price endogeneity correction, but never accounted for all causes simultaneously. The five causes are (i) variety characteristics (ii) unobserved retail activity (iii) TV advertising (iv) household inventories and (v) state dependence/habits. I discuss each cause in detail and how it relates to the endogeneity correction. In particular, I discuss the causes that the endogeneity correction cannot capture. The model estimated with the data from chapter 2 is an individual level discrete choice mixed logit model. The data enable a treatment of the five causes simultaneously under two regimes where prices enter nominally or in efficiency units. Besides, I propose a new approach to construct instruments in the absence of wholesale prices. The results, measured in terms of price elasticities, shed light on the relative importance of the five causes under both price regimes. The importance of the endogeneity correction and a favorite candidate cause for price endogeneity, TV advertising, are studied as well. Previous literature put an emphasis on studying causes in detail, without a focus on price endogeneity. For example, Keane (1997) studies consumer habits and heterogeneity, Allenby and Rossi (1998) focus on consumer heterogeneity and Shum (2004) looks at advertising and habits. Recent structural work has looked at inventory holdings with forward looking consumers, e.g. Hendel and Nevo (2006), Melnikov (2001) and Gowrisankaran and Rysman (2007). Following an alternative avenue Ching, Erdem, and Keane (2009) extend the classical discrete choice model and modify the consumers' perception of prices.

In chapter 4, I assess the role of local TV advertising and retail activity information for aggregate demand outcomes. Commonly, due to nationwide synchronicity of the TV program, it is not possible to measure local advertising exposure of geographically distinct markets within a country. Therefore, only national advertising has been accounted for in previous approaches of Goeree (2008) or Barroso (2009). I construct a market level dataset with local advertising information from the rich individual panel dataset from chapter 2 and use a state of the art standard market level discrete choice model with random coefficients of Berry, Levinsohn, and Pakes (1995). The market setup is as in Nevo (2001). I estimate the model on the constructed data to evaluate the importance of advertising. Just as the model in chapter 3, the model allows for price endogeneity caused by an unobserved product characteristic and again advertising is one of the candidate explanations. As the construction of the dataset is not unique, as first step I compare 120 possible setups with the simple discrete choice logit model to assess the role of advertising. Advertising is analyzed for twenty differently constructed advertising measures. Commonly, this first necessary step has not been discussed in previous applications. For four candidate setups that give clear results in the first step and are economically plausible, I estimate the random coefficients version, considering the recent advances and caveats of this model class put forward by Dubé, Fox, and Su (2009) and Knittel and Metaxoglou (2008). I implement versions with and without local sociodemographic market information to find out whether this impacts the findings for local advertising.

Chapter 2

Introducing a Household Panel Dataset with Consumption and Advertising

2.1 Introduction

In modern applied work the availability of detailed micro data is becoming a common paradigm. Due to informational richness of the data, detailed research questions can be addressed in the hope of finding satisfactory answers. Elaborate methods have been developed especially in the context of micro data. To be sure that the models estimated on the specific dataset are not driven merely by data artifacts, the researcher must have a sound knowledge of the dataset details. Especially, demand models aim at understanding fundamental consumer transactions and describe the economic behavior of many consumers. Traditionally, this has been done with market level data. Since the availability of individual level data increases, it is possible to study consumer demand on this more detailed level, coming at the cost of having more complex and large data. Therefore, this chapter is intended for potential users of the dataset and contains numerous details to permit profound research usage of it.

In this chapter I present the details of a dataset collected by A.C. Nielsen from 2004 through 2006. It comprises detailed household data on purchases, sociodemographics and exposure to TV advertisement. The interplay of purchase incidents and TV advertisement contacts measured for each household allows to study the effects of advertising on meaningful economic quantities of interest, e.g. marginal effects or elasticities of prices on demand. The advertising data are recorded daily for each advertisement by household distinguished by TV channel. Since advertising and the economic result (the purchase action) is known, advertising itself can be analyzed more thoroughly than was possible with previous datasets.

The rise of hedonic models, see Lancaster (1966) as one of the early references, has emphasized the role of product characteristics for economic demand models. This is vital for differentiated product markets because it can explain the existence of abundant varieties. Here, advertising has a natural role to either inform consumers about product characteristics, build an image/prestige for a product or influence consumers' perception in some other way. Ackerberg (2001, 2003) has addressed this question. To do studies in this fashion, the researcher must have profound knowledge of the data at hand to ensure that they are not driving an empirical model result due to an unnoticed artifact.

The work proceeds as follows: In section 2.2, the data files are described and details on the collection process are given. I outline the conditions under that the data can be used for further research. Section 2.3 describes each data file in detail, accompanied by descriptive statistics. I restrict myself to simple statistics and interesting graphics. In section 2.4, the dataset is combined to a joint file and prepared for estimation of product choice and demand models. The section is completed with an exemplary analysis. The final section 2.5 concludes.

2.2 Data Structure, Collection and Use for Research

The "Single Source" data are an extensive household level panel supplied by A.C. Nielsen, Germany.[1] It provides household, daily purchase and real-time media information over a period of 2 years from June 30^{th} 2004 through June 30^{th} 2006.[2] The name Single Source highlights the fact that daily purchase and high frequency TV advertisement history are each recorded for the same household. The A.C. Nielsen competitor GfK (Gesellschaft für Konsumforschung AG, Nürnberg, Germany) does not supply these data based on the same households, but tries to combine the information from two separate panels using matching procedures. Thus, it is a unique feature of the A.C. Nielsen dataset to observe the households' purchases and advertising exposure simultaneously.

The dataset is collected nationwide throughout Germany and consists of two components: a household panel where purchases are followed and a subsample of the former where additionally all TV advertisement contacts are recorded automatically. As the data consist of several collected data files from A.C. Nielsen, the combined sample size differs from the sample size per file.

[1] "Single Source" is a registered trademark of A.C. Nielsen.

[2] Precisely speaking the purchase information "Homescan" is collected by A.C. Nielsen, and the media information for the same households is collected by Nielsen Media Research, both companies belonging to the Nielsen group. A.C. Nielsen supplies the combined data. "Homescan" is a registered trademark of A.C. Nielsen.

Table 2.1. Overview of Consumer Data Files

File	Household appears if ...	Description
Cash	purchased anything	total value of purchases with time, store, zip
Wash	purchased detergent	detergent purchases with time, store, zip and product details (price, quantity, characteristics)
Demo	sampled	time constant sociodemographic variables
Contact	TV telemeter equipped	TV advertisement, TV representation factors

See table 2.1 for details on the four data files that contain purchase data, category purchase data, sociodemographics and advertising contacts. Table 2.2 shows the number of households for whom relevant information is available. Sociodemographics are available for all households. 80% of the sampled and purchasing households buy detergents. 23% of purchasing households have participated in recording advertising exposure.

Table 2.2. Number of Households by Required Information

Dataset	Criterion	No. of Households
Demo	sociodemographics known	17,978
Cash	any purchase	16,757
Cash	any purchase in "detergent" store	16,737
Cash	above plus demographics	16,737
Wash	any purchase of detergent	13,455
Wash	above plus demographics	13,455
Wash	TV coverage in any year	3,783
Wash	TV coverage 2004	2,953
Wash	TV coverage 2005	2,630
Wash	TV coverage 2006	2,571
Wash	TV coverage 2004 and 2005	2,250
Wash	TV coverage 2005 and 2006	1,993
Wash	TV coverage 2004 to 2006	1,735

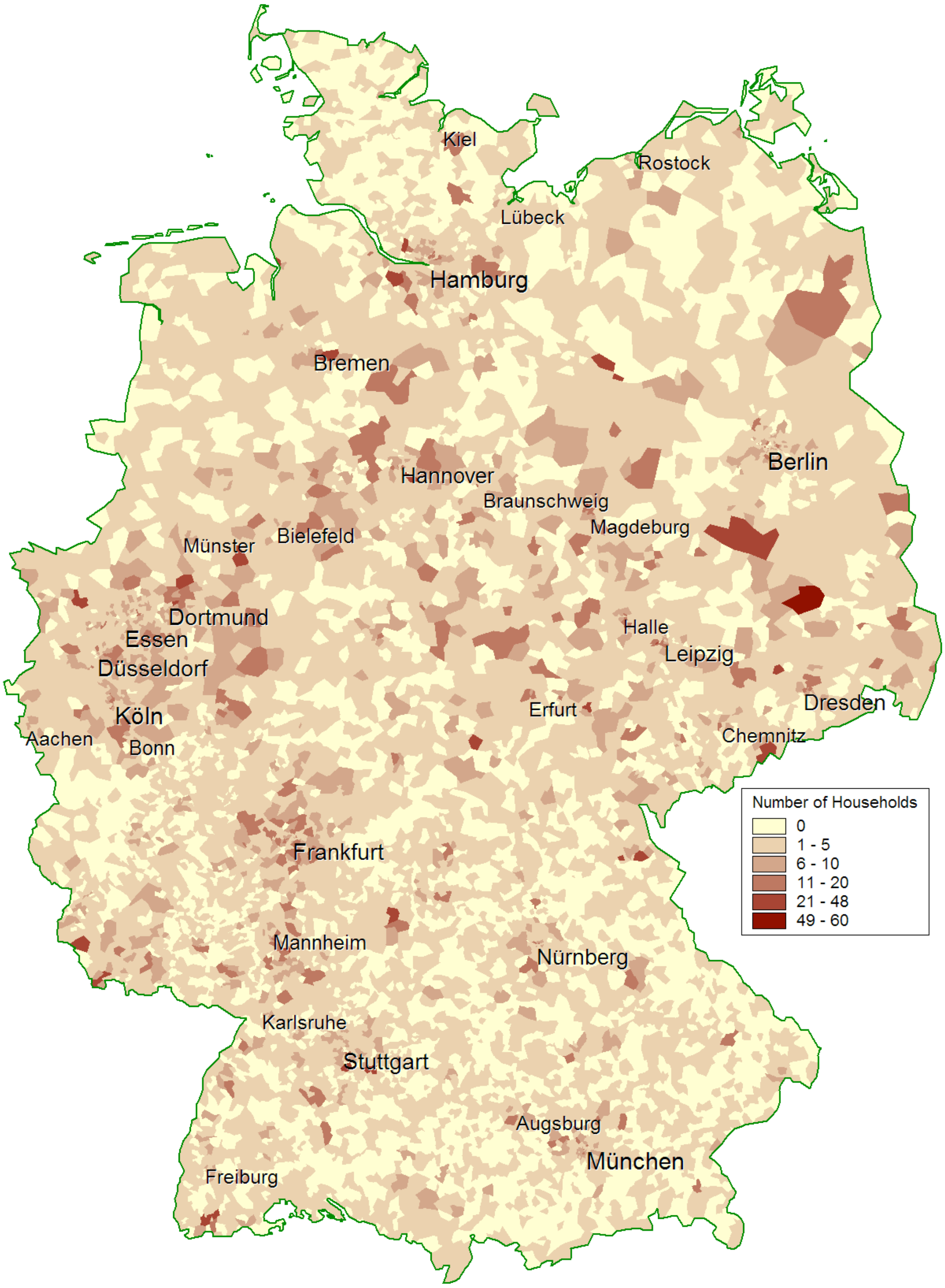
Notes: A “detergent” store is defined as store where it is possible to buy detergents.

The data provide information on daily visits to supermarkets and the amounts spent at each visit for two product categories: chocolate and detergent. Additionally, I know aggregate amounts spent per visit, the exact brand-size combinations bought, quantity and transaction price. Retail activity is measured by feature and display variables. The feature variable indicates whether the brand of the product was featured in the newspaper circulars for a store. The display variable measures whether the brand was promoted via a display, e.g. lobby, aisle (front, end, back) and specialty/shipper.

See figure 2.1 for the geographic distribution of consumers according to their zip code of residence.[3] The cities in the map have more than 200,000 inhabitants. In appendix part A the distribution of the shopping trips is depicted and it resembles very much the

[3] The graph is created with the software PLZ-Diagramm v3.8.

Figure 2.1. Geographic Distribution of Households according to the Zip Code of Residence



distribution of household residence.[4] Given that there are roughly 18,000 households in potentially about 15,000 zip codes, it is not surprising that large areas of Germany are not represented in the sample, indicated by the light yellow area in the figure. Interestingly, the areas with the most households are not necessarily near or in metropolitan areas, i.e. near the cities in the map that all have at least 200,000 inhabitants. It is also not the case that the zip code areas are too small to be visually recognizable, i.e. in Berlin it is possible to see yellow spots (no household participates) and most areas have less than five households. Considering the representative character of the sampled households for the national German market in the geographic dimension, there are no other obvious concerns since households are in fact sampled nationwide.

2.2.1 Collection Process

In general, it is necessary to distinguish store level and household level data. The first are collected in a store, summarize all transactions and do not identify the household, whereas the latter only record the transactions of the households involved in a panel.

"Classic" scanner data are store level data. Scanner data are collected at the store where the purchase is done. There the cashier will scan or identify the product, usually by EAN Code.[5] Thereby transaction information is perfectly monitored by the IT system of the retailer. The drawback of this approach is that transactions cannot be associated with the consumers, while it is perfectly suited to track all store sales. Of course, customer or loyalty cards/memberships are a way to mitigate this problem.

The Single Source dataset by A.C. Nielsen consists of a "Homescan" panel of consumers. The term "Homescan" is very descriptive and suits well the fact that this is household level data. See the paper by Einav, Leibtag, and Nevo (2008) for a validation study of the US Homescan Panel. Participants are equipped with a hand scanning device, a charger for the latter, a phone box to connect directly to A.C. Nielsen and a bar code handbook. Homescan households can apply to report their advertising exposure to A.C. Nielsen. Those get a telemeter to measure advertising exposure and a special remote control. The actual data collection consists of two components. Collection of purchase data and of advertising data.

First, I outline the procedure for the purchase data. Consumers shop as usual and when arriving at home, they use a hand scanning device to scan all the purchases they have

[4] This is not surprising, since missing values of shopping trip zip codes are partly imputed with the zip code of residence, see the discussion of the general purchase data in section 2.3. However, the relative importance could have varied, as the number of trips is different per household. For example, if residents within cities should had conducted more purchase trips, this would have changed their relative importance compared to residents of rural areas.

[5] EAN is short for European Article Number and identifies consumer products uniquely. This is the number also coded as barcode on most products. Manufacturers can request (for a fee) these numbers from the organization GS1 (Global Standards One) that is in charge of these codes.

done. Products with EAN Code can be easily scanned. For non-coded products the consumer looks up the product in the bar code handbook and scans the right barcode, e.g. for bread, vegetables or meat products. In the scanner display, the consumer selects the store from a drop-down list. The person that scans also selects in the scanner display the household members that participated in the purchase trip. This information is transmitted daily and automatically via phone box to A.C. Nielsen, who can reconstruct the transactions from the transmitted data. In the case of absence due to holidays, a household member types this into the scanning device.

Second, the advertising data are collected in the following way. The TV in the household is augmented by a device ("telemeter") that checks to which TV program the household was tuned in. When using the TV, the household members have to use exclusively the special remote control that also works for the normal TV. Each member has a button on the remote control. By pushing this button, the member logs in and out from watching TV. The information from the special remote control is received from the telemeter and transmitted automatically via phone box to A.C. Nielsen. Thereby, the exposure is measured with high accuracy because all the members must do is to use the right remote control.[6]

Households can enter the panel in several ways. A.C. Nielsen randomly requests participation by mail, advertises the panel to receive applications or existing panel members can recommend new members. Households are interviewed and if the candidate household is suitable, it is admitted to the panel. Then the household reports detailed sociodemographic information that is updated regularly. A.C. Nielsen tries to select households in such a way that all relevant types of German consumers are present in the sample. I have no information on other suitability criteria. Participants commit to record all their shopping trips. The consumers are not paid for their participation, but they get bonus points for which they can choose products from a catalogue. If Homescan households participate in measuring advertising exposure they get additional bonus points. Households can participate in extra programs to acquire more bonus points, e.g. admit to fill out questionnaires with supplementary questions. Households exit from the panel voluntarily or are taken out if they do not comply with the participation rules, e.g. they do not scan their purchases or do not report their absence from the panel due to holidays. The participation is checked regularly. Participating consumers are quite satisfied with the collection process.[7]

[6] In earlier implementations this process was not reliable. This would lead to many missing values, due to technical failures or misuse, see the freely available data of the Kilts Marketing Research Center at the Graduate School of Business at Chicago that suffer from these issues. Nevertheless, the data have been used for research, as discussed later on.

[7] See the reviews about the data collection procedure for the Single Source panel authored from various sampled consumers at the Website of Ciao GmbH in 2007. Viewed January 14th 2007: `http://www.ciao.de/ACNielsen_Werbeforschungsunternehmen__942530`.

2.2.2 Using the Dataset for Research

The dataset has been acquired recently by the author for research and is available for further research that goes beyond this dissertation. The Chair of Statistics of Prof. Dr. Enno Mammen has a contract with A.C. Nielsen that permits research use of the data and is entitled to collaborate with researchers to use the data after signing a confidentiality agreement.[8] Research is not limited in any fashion. Names of any company, product and retail chain have to stay anonymous in the publications but are contained in the data. Obviously the same is required for the sampled households. Part of the contract with A.C. Nielsen requires all publications to be sent in to A.C. Nielsen before publication.

2.2.3 Relation to other Datasets

There are many datasets used in the literature, where each one has comparative advantages for certain applications. Most important for the current chapter are the differences in the quality and quantity of pricing and advertisement information.

In general, it is necessary to distinguish store level and household level data. The first contains all prices of all products sold at a given time in a store and does not identify the household, whereas the latter only records the transaction prices of the households involved. Hence, store level data contain also price information of products that were not purchased in a household level dataset. Although store level data record price information optimally, they lack the advertising information outside the store totally, whereas household level data can almost optimally measure advertising exposure. Since A.C. Nielsen collects data on both levels, it is theoretically feasible to link the data to deliver the optimal dataset.

Besides, there is commonly a difference in the geographic dimension of both levels of data. Store level data are collected obviously by store and can be geographically concentrated if they are collected for a little number of stores. Household level data are usually collected for households in different locations so that they are geographically more dispersed.

In the data of Hendel and Nevo (2006) households are tracked that purchase in one store and a complete store level dataset is available to deliver all prices during purchase decisions of the consumer in a simple fashion, but it lacks the TV advertisement information of the dataset in this chapter. The data was provided by Information Resources Inc. (IRI) from 1991 to 1993 and is an example of a combination of both store and household level data.

Concerning prices, my data are a household level panel of the same kind as Keane (1997), albeit he has only households in a few regional US markets. He has to impute missing

[8] Address: Chair of Statistics, Economics Department, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany.

price data. I deal with this issue in a similar fashion as explained later in section 2.4.2. Erdem and Keane (1996) use a household level dataset that tracks households daily in two stores from 1986 to 1988 and is also collected by A.C. Nielsen.[9] The product category of interest is laundry detergents. In principle it has the same TV advertising data as the data employed in this chapter, but only for 1800 households during 51 weeks. It lacks the precise advertisement information of my data that enables to identify image and product specific spots. Moreover, the tracking technology was not fully developed at this time, leading to errors and missing values in this data. The same data with a different product was used by Ackerberg (2001); he studies one single product that is only advertised during its introduction to the market and exactly for this time the advertising data are available without obvious errors. Shum (2004) has geographically restricted cereal sales data from IRI for the period from 1991 to 1992 and uses quarterly national advertising expenditures from leading national advertisers (LNA). Griffith, Leibtag, Leicester, and Nevo (2009) use household level data from the TNS Worldpanel for Great Britain with more information on retail activity but no TV advertising information.

2.3 Data Description and Details by File

For each data file provided by A.C. Nielsen I present a modified version that differs from the raw data. The attention is restricted to the detergent category. I specify the modifications, give details on created variables, show an overview of all variables and present descriptive statistics per data file. The modifications have the sole purpose of preparing the data for analysis without changing informational content.

The value labels for the brands, manufacturers and stores are in the data, but they are not displayed due to an confidentiality agreement with A.C. Nielsen.[10] Moreover, it not possible to identify individuals but only households with an id number although the individual household member information is recorded as detailed in section 2.2.1. All monetary values are measured in €. There are four files that will be discussed in detail. The general purchase data contain all purchase trips of households. The category purchase data consist of all purchases in the detergent category for each household. The demographic data comprise the sociodemographic information of all households. The advertising data are composed of the advertising exposure per household.

[9] This is the freely available data of the Kilts Marketing Research Center at the Graduate School of Business at Chicago.

[10] Value Labels are a possibility to code strings as numbers. Then each code represents a string. This saves RAM (Random Access Memory) during statistical analysis for datasets with many observations.

2.3.1 General Purchase Data

The general purchase data contain all purchase incidents (i.e. shopping trips) for each household for the time period from June 30th 2004 through June 30th 2006. The raw data have 4,179,716 observations of 16,757 households. Each observation is a purchase incident and several incidents per day are possible. As all incidents of a household are recorded, some stores exist in which no product of the category of interest, i.e. detergents, was sold. I remove those purchase incidents. In table 2.3 all variables are displayed. There are 3,058,880 observations. The following section explains the steps taken to get from the raw data to the presented data.

Table 2.3. Overview of Variables in General Purchase Data

Variable	Numeric	Position	Description
`calweek`	1(=yes)	14	Calendar Week (STATA) YYYYwWW
`date`	1	4	Date as YYYYddd, where ddd is absolute day in year YYYY
`date2`	1	5	Date as YYYYMMDD
`datebegc`	1	20	Date of first observation in this file
`dateday`	1	12	Date DD
`dateendc`	1	19	Date of last observation in this file
`datemonth`	1	11	Date MM
`dateyear`	1	10	Date YYYY
`durhhmean2c`	1	18	Mean duration per household in days since last purchase, exclude zero durations
`durhhmeanc`	1	17	Mean duration per household in days since last purchase
`durobs2c`	1	16	Duration in days since last purchase, exclude durations of zero
`durobsc`	1	15	Duration in days since last purchase
`dursamplec`	1	21	Duration between first and last observation in this file
`edate`	1	13	Date (STATA) YYYYMMDD
`hhnr`	1	1	Household ID code
`hhobsc`	1	9	No of Purchases in sample for HH
`key_acc`	0(=no)	6	Name of store
`plz`	1	8	German zip code, all five digits
`stadt`	0	7	Name of city
`store`	1	2	Store ID code
`value`	1	3	Total value of purchase in €

Notes: Numeric indicates whether a variable is numeric. Position gives the column position of the variable in the file. STATA in the description means that the variable is in STATA date format.

2.3.1.1 Data Modifications

This section details the modifications applied to the zip codes, store names and their impact on sample size.

Zip codes (variable: `plz`) For several observations, zip codes are missing. To fill up missing values I assume that people will mostly do consumer goods shopping trips in the most frequent occurring zip code of all their shopping trips. There are 2,741,313 missing values, after filling up the number is down to 185,364.

Store names (variable: key_acc) For some purchase trips, there are missing values for the store names that are necessary to identify stores and retail chains. To fill the gaps, I look for identical store codes (variable store) on other purchase trips within the same file. I replace missing key_acc store name values with the key_acc values of the identical store code if it was present in the file for another purchase trip.[11] Of 11,021 missing values 5,621 can be constructed, 5,400 remain unknown.

Store names, store id for category purchases (variables: key_acc, store) Not all stores offer the possibility to buy a product from the category of interest. Therefore, all incidents are dropped if it is not possible to buy detergents in the store. 1,120,772 observations could be identified not to offer products from the detergent category and were dropped, 3,058,944 observations remain. To identify the stores to be dropped, I compared the store names and store ids in the general purchase data with those that appeared in the category purchase data. If they appeared, these stores were kept while the rest was dropped.

After these adjustments I checked for duplicates in terms of the variables hhnr, edate, key_acc and value that are to identify a purchase incident. Then a given household can be only once a day in a store and spend a specific amount. It is very unlikely that this will identify real observations as wrong duplicates. In fact, this approach delivered 64 duplicates that were dropped. The remaining number of observations is 3,058,880.

2.3.1.2 Remarks on created Variables

In this subsection I want to emphasize details of newly constructed variables.

Duration and Timing variables All of these variables rely on the purchase incident dates that indicate the day of purchase. The duration variables are calculated by measuring the time between two adjacent purchases. If purchases occur on the same day duration is zero. This may underestimate the duration the researcher is interested in if the duration variable is used to construct statistics such as means. That is why for all durations there is a variable whose name is identical but ends with 2, and these ignore zeros when calculating durations and may be the actual duration the researcher is interested in.[12]

[11] The relation of key_acc to store code is 1:n, i.e. there are several store codes that map into the same store name. An natural example for this relation is a retail chain.

[12] For variables durobsc, durhhmeanc there exist versions durobs2c, durhhmean2c. The c indicates the duration is calculated for the general purchase data. The zeros are “ignored” by setting the variable to missing.

2.3.1.3 Summary Statistics

This section provides summary statistics for the general purchase data. Consult table 2.4 for descriptive statistics.[13] Recall that only trips to stores that also sell detergents are considered. For easier navigation in the text discussion of the table, I shall give the relevant variable name in parenthesis to make the results more traceable. In the table, the first column gives the variable name, the second column indicates whether statistics are constructed by purchase incident (PI) or by household (HH). First, I discuss the PI variables. The average basket value in € per shopping trip is 18.91 and gives with the inter quartile range of 18.77 an intuitive range (`value`). As noted before, duration between two purchases is underestimated by counting several purchases on the same day as observation with zero duration (`durbosc, durobs2c`). Therefore, the difference between the duration measures is quite high. Note that the median is two for the second measure so that there are two days between days with any shopping trip. The difference between the duration measures also highlights that there are numerous purchase trips on the same day. Note that there are also durations up to 549, so consumers exist that almost did not shop for over one and a half years. Those are consumers that did not fully participate in the Homescan panel, but they are still in the data. Keeping this in mind, the interested reader will find that several variables in the data indicate similar findings.

Table 2.4. Summary Statistics for General Purchase Data

Variable		Mean	Median	SD	IQR	Min	10%	90%	Max
`value`	PI	18.91	12.67	19.60	18.77	0.10	2.99	42.24	204.51
`durobsc`	PI	2.47	1.00	3.97	3.00	0.00	0.00	6.00	549.00
`durobs2c`	PI	3.56	2.00	4.35	3.00	1.00	1.00	7.00	549.00
`hhobsc`	HH	182.77	133.00	164.26	188.00	1.00	29.00	403.00	1535.00
`dursamplec`	HH	601.28	725.00	200.90	314.00	0.00	242.00	730.00	730.00
`durhhmeanc`	HH	2.47	2.05	1.59	1.55	0.00	1.09	4.39	88.00
`durhhmean2c`	HH	3.47	2.99	1.83	1.90	1.00	1.80	5.77	90.00
`numstore`	HH	12.06	12.00	5.46	8.00	1.00	5.00	19.00	41.00
`storehhi`*	HH	0.34	0.29	0.19	0.21	0.06	0.16	0.61	1.00

Notes: The second column indicates whether statistics are constructed by purchase incident (PI) or by household (HH). * indicates the HHI is calculated with value as weights.

Now I turn to the HH variables. There are many purchases per household (`hhobsc`) and that number of purchases varies a lot. The average time a household spends in the sample is 601 days (`dursamplec`). Taking the mean of the duration between purchases per HH and not per PI increases the median numbers and leaves means almost unchanged. Now

[13] The Herfindahl-Hirschmann Index (=HHI) measures concentration of a variable of interest, and indicates whether consumers stick to their favorite stores or brands. HHI measures are calculated in normalized form so that values range from $\frac{1}{b}$ to 1 where a higher number indicates higher concentration and b is the number of alternatives. For $b > 10$, small values are < 0.3, the medium range is > 0.3 and < 0.7 and high values are > 0.7.

there is a median of 2.99 days between days with any shopping trip (durhhmean2c). Obviously, even on this crude and simple level the right definition of the variable of interest matters a lot for the result. Households buy from about 12 different stores on average (numstore) and consequently the concentration measure displays only low to medium concentration (storehhi).

2.3.2 Category Purchase Data

The category purchase data contain all purchases in the detergent category. The raw data have 94,747 observations. Each line in the file represents a transaction of a specific product. Products are identified by the EAN Code, see footnote 5 for details. For example, if two different products are bought during one shopping trip, there will be two observations for this. Thus, the number of observations is the number of (purchase trip, product) pairs. Consult table 2.5 for an overview of all variables in this file. If variables are numerically coded, the corresponding codes can be found in the tables of appendix part B. The tables represent a modified version of the raw data originally provided and 94,222 observations remain. In the following I outline the steps undertaken.

Table 2.5. Overview of Variables in Category Purchase Data

Variable	Numeric	Position	Description
bigpack	1(=yes)	41	Dummy for detergent is sold in extra big pack
calweek	1	27	Calendar Week (STATA) YYYYwWW
cnanlabl	0(=no)	2	Detailed Product Information String
color	1	36	Dummy for detergent is color
date	1	1	Date as YYYYddd, where ddd is absolute day in year YYYY
date2	1	21	Date as YYYYMMDD
datebegw	1	33	Date of first observation in this file
dateday	1	25	Date DD
dateendw	1	32	Date of last observation in this file
datemonth	1	24	Date MM
dateyear	1	23	Date YYYY
display	1	18	Product is on display
duft*	1	44	Numerical code for scent type (coded)
durhhmean2w	1	31	Mean duration per household in days since last purchase, exclude zero durations
durhhmeanw	1	30	Mean duration per household in days since last purchase
durobs2w	1	29	Duration in days since last purchase, exclude durations of zero
durobsw	1	28	Duration in days since last purchase
dursamplew	1	34	Duration between first and last observation in this file

Notes: * indicates that the variable is numerically coded. The tables to resolve the codes are found in appendix part B. **,*** are defined at the end of the table.

Table 2.5. (continued...)

Variable	Numeric	Position	Description
edate	1	26	Date (STATA) YYYYMMDD
erg	0	3	Effective amount of detergent
extra_size	1	40	Dummy for detergent is sold in extra big size
feature	1	16	Product is featured
gimmick	1	39	Dummy for detergent is sold with gimmick
handbill	1	17	Product is hand billed
her**	1	49	Numerical code for manufacturer (coded)
herb8**	1	51	8 biggest manufacturers (coded), SONST (=Various), HANDEL (=Private Label)
hhnr	1	20	Household ID code
hhobsw	1	22	No of purchases in file for HH
id_nr	1	15	unknown ID number
inh	1	42	Packet size (l or kg)
key_acc	0	7	Name of store
keyb10	0	55	10 biggest key_accs, details see keyb7
keyb15	0	54	15 biggest key_accs, details see keyb7
keyb7	0	56	7 biggest key_accs, SONST (=Various)
kons*	1	45	Numerical code for consistency type (coded)
konzentrat	1	38	Dummy for detergent is concentrated
liquid	1	43	Dummy for product is liquid detergent
menge	1	5	Number of units purchased
mke**	1	50	Numerical code for brand (coded)
mkeb14**	1	52	14 biggest brands (coded), details see mkeb9
mkeb9**	1	53	9 biggest brands (coded), SONST (=Various), EIGENMARKE (=Private Label)
plz	1	9	German zip code, all five digits
plz1	1	10	German zip code, 1st digit
plz2	1	11	German zip code, 1st two digits
plz3	1	12	German zip code, 1st three digits
plz4	1	13	German zip code, 1st four digits
prceflag	1	19	Product is price flagged
preis	1	6	Purchase price
purchase	1	35	Dummy for wash purchase incident
quartal	1	14	Date as YYYYQQ, where QQ is quarter
sensitiv	1	37	Dummy for detergent is sensitiv
stadt	0	8	Name of city
store	1	4	Store ID code
uwg*	1	46	Numerical code for general purpose (coded)
vpa***	1	48	Numerical code for packaging type (coded)
zmke***	1	47	Numerical code for sub brand (coded)

Notes: Numeric indicates whether a variable is numeric. Position gives the column position of the variable in the file. STATA in the description means that the variable is in STATA date format. * indicates that the variable is numerically coded. The tables to resolve the codes are found in appendix part B. ** marks that variable is coded, but is not disclosed in the appendix due to the confidentiality agreement for data usage. *** indicates that variable is coded, but is not detailed in appendix.

2.3.2.1 Data Modifications

This section details the modifications done to the zip codes, store names and product characteristic variables. Some of these changes have an impact on the sample size.
Zip codes (variable: `plz`) For several observations, the zip code is missing. I fill up missing values in the same fashion as done for the general purchase data. Before the procedure I have 50,544 zip codes missing, afterwards that number is down to 23,136.
Store names (variable: `key_acc`) Just as in the general purchase data file, there are missing values for the store names. These are necessary to identify the store and more importantly retail chains. I conduct the analogue steps to fill up missing values, of 225 missing values 99 can be constructed, while 126 remain unknown.

Second, in the general purchase data the store name information is richer because there are more purchase trips in that file. Therefore, I looked for the identical store codes in the category and general purchase data, and filled up missing store names in the category data using the general purchase data. After this, there are merely 104 missing values among the store names. I drop those observations so that 94,643 observations remain.
Store type (variable: `key_acc`) Next, I checked the kind of stores in which the consumers bought detergent products by checking the store names.[14] A few of the stores are quite unusual and should be excluded from the analysis. For example, a chain that sells pet accessories appeared, but it is not a typical place to buy detergent. Another example is a delivery service for frozen food. It is not possible to verify whether these are errors, but it may be that these are part of an unusual promotion. Removing these observations reduces the sample size by 205 to 94,438.
Products (variables: `zmke`, `mke`, `her`, `duft`) In these variables several strings had a question mark or additional character, but otherwise the value was identical to an existing one. In this case I removed the character to get consistent values on the variables. This applied to 61 observations in `zmke`, 242 observations in `duft`, 87 observations in `her` and 29 observations in `mke`.[15]
Products (variable: `cnanlabl`) One observation contained an highly unusual product identified as bundle of five large single washing packets and occurred only one time in the file, so this observation was dropped.
Products (variable: `inh`) The contents variable contained quantities that did not fully correspond to the information in `cnanlabl`. These observations were dropped, in total 216, leaving us with 94,222 observations.

[14] Recall: `store` and `key_acc` have a n:1 relation, i.e. one store name of a retail chain is consistent with several store codes, as these identify the outlet.

[15] Example: `duft` had the value "UNB" and "O ANGABE" that both implies scent to be unknown. I consolidated this by setting all values to "O ANGABE", as this occurred more frequently than "UNB".

2.3.2.2 Remarks on created Variables

In this subsection I want to emphasize details of newly constructed variables.
Duration and Timing variables The same remarks as for the general purchase data apply to the variables constructed in this file.[16]
Product characteristics A number of very detailed product characteristic variables are constructed by using string functions that searched the string variable `cnanlabl`. The functions search the string for certain keywords that are unique for the product characteristic and if at least one of the keywords was found, the dummy for this newly created variable is set to one. I checked all products to be sure that a keyword used for a characteristic is uniquely identifying the characteristic. Two examples for the string variable `cnanlabl` with underlined keywords are:

1. `ARIEL SANFT + REIN 5.94 KG + SWIFFER STAUBMAGNET VOLLWASCHMITTEL`
2. `PERSIL VWM MEGAPERLS COLOR 1.418 KG NF (+15 %) VOLLWASCHMITTEL`

Product 1 is a sensitiv powder with a gimmick. Product 2 is a concentrate with color option and sold in an extra size different from the standard one.

The variable `cnanlabl` is linked to the EAN Code and therefore it is reliable.[17] This permits to extract precise information on the product and is used to construct a series of dummy variables in the manner described above: `color`, `sensitiv`, `konzentrat`, `gimmick`, `extra_size` and `bigpack`. `liquid` marks a detergent as liquid. `color` indicates whether the detergent is appropriate for colored washing. `sensitiv` declares the detergent to be suitable for allergen sensitive people. `konzentrat` flags a concentrated detergent so that less detergent is required per washing load. `gimmick` signals the product to be sold with a giveaway, e.g. a CD, a cleaning cloth or cleanser. `bigpack` marks a bundle of at least two identical products sold in one unit. `extra_size` signals the product size to be increased by 10%, 15%, 20% or 25%.

Before constructing the dummy variables, all values of `cnanlabl` had to be searched in order to find all interesting keywords that could potentially serve to identify important product characteristics. At first, for example, I did not search for keywords that identify the product to have a gimmick. It simply occurred to me while looking at all values of `cnanlabl` that many keywords indicated detergents to be sold with gimmicks. Collecting keywords in this manner, the definition of product characteristic dummy variables followed. For example, to define the variable `extra_size`, more than 10 strings indicate that the product is sold in a packaging that is bigger than the standard one.[18]

[16] Names are similar to the general purchase data, but for the ending letter `w` that indicates the timing variables to be calculated from the detergent purchase data. As before, for variables `durobsw, durhhmeanw` there exist versions `durobs2w, durhhmean2w`.

[17] See footnote 5 for details on the EAN Code.

[18] Example: String fragments that appear in `cnanlabl` and are used to construct the variable `extra_size`: "+10%","+33%","+33 %","+ 15%", "+15%", "+ 10 PRZ", "+10PRZ", "+20%", "+20 %", ...

2.3.2.3 Summary Statistics

In this section I present summary statistics for the category purchase data. Tables 2.6 and 2.7 display simple statistics for all variables per household (HH) and per purchase incident (PI), respectively. I first discuss the table with the statistics per household.

Table 2.6. Summary Statistics for Category Purchase Data

Variable		Mean	Median	SD	IQR	Min	10%	90%	Max
`hhobsw`	HH	7.02	5.00	7.54	7.00	1.00	1.00	16.00	102.00
`durhhmeanw`	HH	56.60	43.20	49.07	40.79	0.00	18.82	105.75	723.00
`durhhmean2w`	HH	64.31	50.17	52.40	44.17	1.00	22.69	118.40	723.00
`ppm`	HH	0.95	0.60	2.67	0.56	0.08	0.26	1.46	90.00
`numbrand`	HH	2.40	2.00	1.69	2.00	1.00	1.00	5.00	14.00
`brandhhi`*	HH	0.70	0.68	0.28	0.55	0.10	0.31	1.00	1.00
`brandhhiv`**	HH	0.69	0.66	0.29	0.56	0.09	0.30	1.00	1.00
`numstore`	HH	2.35	2.00	1.51	2.00	1.00	1.00	4.00	13.00
`storehhi`*	HH	0.69	0.64	0.28	0.56	0.13	0.32	1.00	1.00

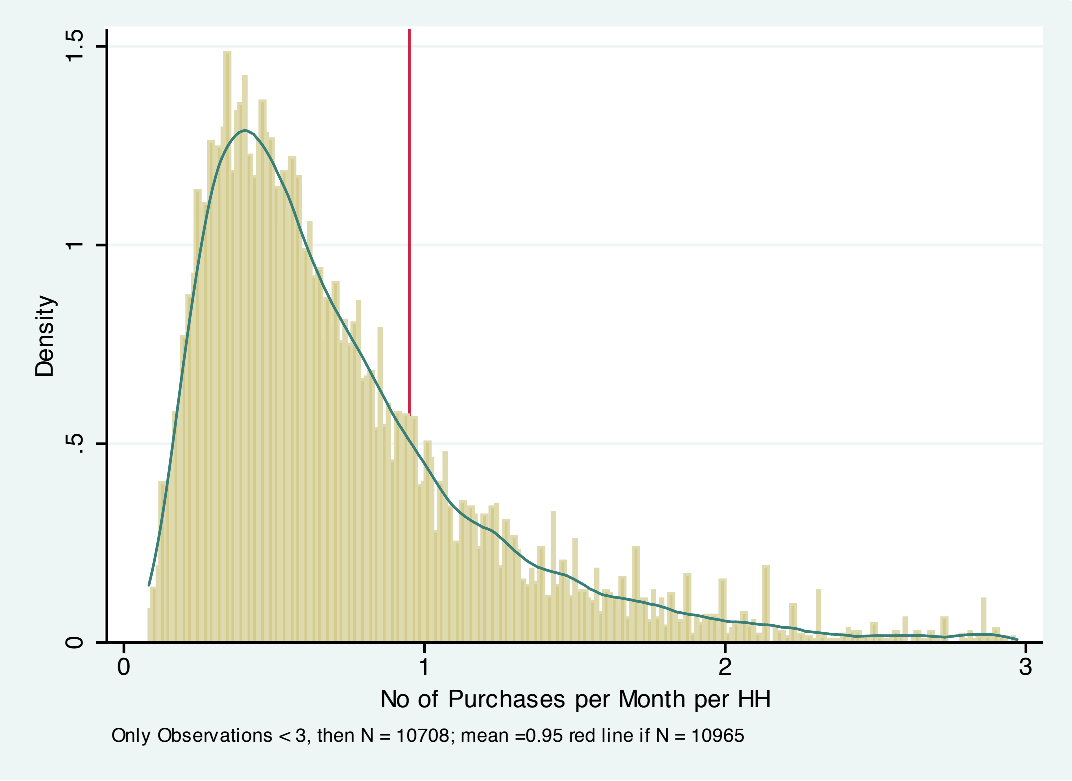
Notes: The second column indicates whether statistics are constructed by purchase incident (PI) or by household (HH). * indicates the HHI is calculated with value as weights. ** indicates the HHI is calculated with volume counts as weights.

Compared to the duration of the general purchase data, the duration measures for the category purchase data behave similarly (`durhhmeanw, durhhmean2w`). Note that the median duration from the household variable shows that there are about 40 to 50 days, roughly one and a half months, between two trips that result in a detergent purchase. Translated into purchases per month (`ppm`) this results into a median of 0.60 for all households and the density is depicted in figure 2.2. The longer duration and lower frequency naturally translate into a relatively low number of average purchases per household for the sampling period of two years (`hhobsw`). Figure 2.3 shows the density of the variable. The mass of households have less than 10 purchases in the sampling period.

Figure 2.4 displays how long households remain in the sample by measuring time between first and last recorded purchase. The following pattern emerges: There are two peaks at about half a year and one at the maximum possible time of two years and two valleys at zero and roughly one year. Presumably this is due to the entry/exit rules of the Homescan Panel.

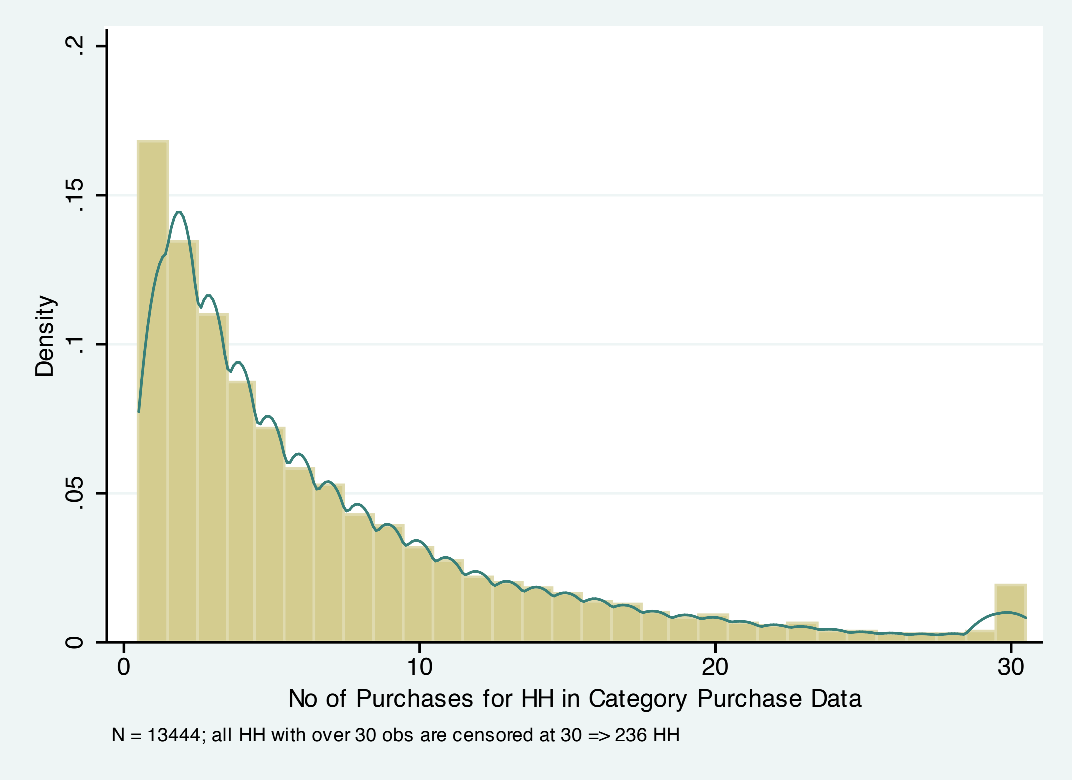
The median of the number of different brands bought per household is 2 (`numbrand`), and the concentration measures indicate that this is in fact a high concentration (`brandhhi, brandhhiv`). The number of different stores at which the household shops is low as well, having a median of 2 (`numstore`). Again, the concentration measure indicates that there is high concentration in store choice (`storehhi`). Thus, consumers seem to stick to their favorite brands and stores.

Figure 2.2. Purchases per Month per Household



Only Observations < 3, then N = 10708; mean =0.95 red line if N = 10965

Figure 2.3. Purchases in Sample per Household



N = 13444; all HH with over 30 obs are censored at 30 => 236 HH

Figure 2.4. Sample Membership Duration per Household

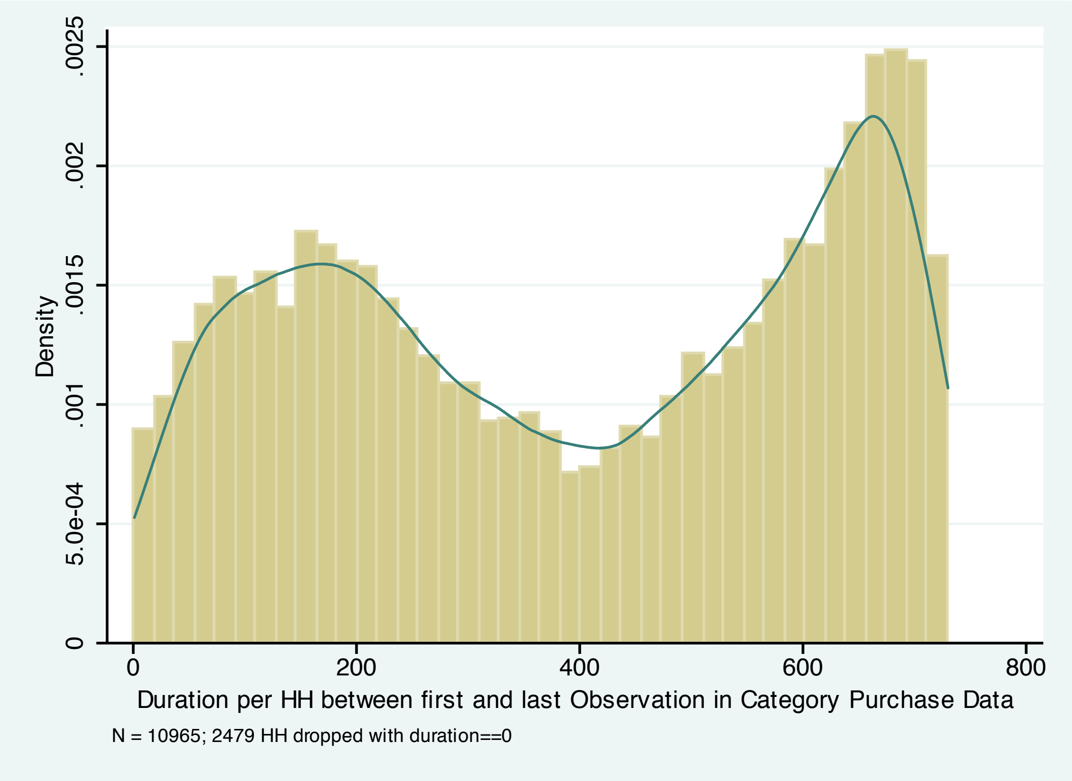


Table 2.7. Summary Statistics for Category Purchase Data

Variable		Mean	Median	SD	IQR	Min	10%	90%	Max
`durobsw`	PI	53.09	34.00	64.14	57.00	0.00	0.00	126.00	723.00
`durobs2w`	PI	60.66	41.00	65.12	57.00	1.00	9.00	135.00	723.00
`preis`	PI	3.71	2.99	2.30	1.58	0.01	2.19	6.66	32.99
`menge`	PI	1.12	1.00	0.44	0.00	1.00	1.00	1.00	16.00
`liquid`	PI	0.48	0.00	0.50	1.00	0.00	0.00	1.00	1.00
`konzentrat`	PI	0.38	0.00	0.48	1.00	0.00	0.00	1.00	1.00
`color`	PI	0.31	0.00	0.46	1.00	0.00	0.00	1.00	1.00
`sensitiv`	PI	0.02	0.00	0.13	0.00	0.00	0.00	0.00	1.00
`gimmick`	PI	0.03	0.00	0.18	0.00	0.00	0.00	0.00	1.00
`bigpack`	PI	0.02	0.00	0.13	0.00	0.00	0.00	0.00	1.00
`extra_size`	PI	0.03	0.00	0.17	0.00	0.00	0.00	0.00	1.00
`feature`	PI	0.07	0.00	0.25	0.00	0.00	0.00	0.00	1.00
`display`	PI	0.11	0.00	0.32	0.00	0.00	0.00	1.00	1.00
`prceflag`	PI	0.14	0.00	0.34	0.00	0.00	0.00	1.00	1.00
`handbill`	PI	0.11	0.00	0.32	0.00	0.00	0.00	1.00	1.00

Notes: The second column indicates whether statistics are constructed by purchase incident (PI) or by household (HH).

Now I turn to the statistics per purchase incident in table 2.7, calculated mainly for product characteristics. The duration variables calculated are lower than if calculated per household, especially the median values. Price of products (`preis`) covers broadly the expected range, with some values being extremely low (=0.01) or high (=32.99). The low values are in fact very small packages, and perhaps the given price only represents an internal price of the retailer whereas the large values are for professional packages of detergents. An internal price is merely necessary to register a product in the retailer IT system so that the price level of 1 cent is only of symbolic value and not related to the product value. Both extremes occur seldom in the data (less than fifty observations each).

Consumers mostly shop one unit, at least for 90% of all purchases as is evident from the ninetieth percentile (`menge`). This is advantageous if the researcher wants to use a simple discrete choice model that only permits choice of a single unit.

The distribution of product characteristics across all sold products segments the characteristics into two broad groups. 48%, 38% and 31% of sold products possess the three characteristics `liquid, konzentrat, color` which highlights their importance as product characteristics whereas the characteristics `sensitiv, gimmick, bigpack, extra_size` are only relevant for at most 3% of all sold units.[19] The last four retail activity variables `feature, display, prceflag, handbill` are defined as dummy variables and take values in a range from 7 to 14% of all purchases. Hence, over 80% of sales in the category happen without the product being actively promoted in the store.[20]

2.3.3 Demographic Data

The demographic data file contains sociodemographic information for each household. The raw data have 17,978 observations, each represents a household. The information is time fixed, changes are not documented. The information is updated yearly and the latest version is retained. Therefore, the present file dates from the end of the year 2005. For each household, there is a household managing person who is responsible for the household and conducts most purchases.

Consult table 2.8 for an overview of all variables in this file. If variables are numerically coded, the corresponding codes can be found in the tables of appendix part B. The tables represent a modified version of the raw data originally provided and have the same number of observations as the raw data. In the following I outline the steps undertaken.

[19] Recall that the means of these characteristic variables are defined as dummy variables and thereby represent shares.

[20] The data show (not presented here) that different retail activity measures are mostly combined, e.g. featured products are almost always price flagged.

Table 2.8. Overview of Variables in Sociogemographic Data

Variable	Numeric	Position	Description
age*	1(=yes)	8+*	Age of *th household member, *=1,...,8
bblue	1	45	Blue collar (12<bstel<16)
beginn	1	21	Entry of household into panel
bik*	1	7	BIK community size (inhabitants)
bself	1	42	Self employed (bstel<6)
bstel*	1	6	Occupational Status
bunemp	1	46	Ever since unemployed (bstel=18)
bwhite_h	1	43	White collar high (5<bstel<11)
bwhite_l	1	44	White collar low (10<bstel<13, 15<bstel<18)
eink	1	3	Household net monthly Income (€), upper bound of interval
ende	1	22	Exit of household from panel
hhfage	1	2	Age of household managing person
hhfpsex*	1	30	Gender of household managing person
hhnr	1	18	Household ID code
hhsize	1	5	Number of persons in household
kizahl	1	1	Number of children in household
ng*	1	8	Nielsen areas
plz	1	19	Zip code of household residence, all five digits
plz1	1	23	as above, 1st digit
plz2	1	24	as above, 1st two digits
plz3	1	25	as above, 1st three digits
plz4	1	26	as above, 1st four digits
prf	1	17	Scaling factor to get representative German Sample
psex**	1	30+*	Gender of *th household member, *=1,...,8
psexm	1	40	Number of men in household
psexw	1	39	Number of women in household
schulab*	1	47	Degree of houshold managing person
stadt	0(=no)	20	City name of household residence
stand*	1	4	Family status of the household managing person
tv_prf04	1	27	Scaling factor of TV panel 2004
tv_prf05	1	28	Scaling factor of TV panel 2005
tv_prf06	1	29	Scaling factor of TV panel 2006
urban	1	41	BIK community size greater 50K inhabitants

Notes: Numeric indicates whether a variable is numeric. Position gives the column position of the variable in the file. STATA in the description means that the variable is in STATA date format. * indicates that the variable is numerically coded. The tables to resolve the codes are found in appendix part B.

2.3.3.1 Data Modifications

Various variables are coded as strings and I code them with numeric values. I will not list each variable that got this treatment, as it does not change the informational content. Refer to the label tables in appendix B to get information on the characteristic values of the variables.

Family Status (variable: `stand`) The coding of the family status changed from 2005 on so that the raw variable cannot be used. The same code has different meanings depending on whether the household entered the panel before 2005 or afterwards. I correct the coding to be consistent for all households according to the codes from 2005 on. The change amounts to combining codes of 2004. Singles and people who have a partner, but each have their own household as well were distinguished in 2004 and are now labelled uniformly "Unmarried". Legally divorced couples and married couples who no longer live together were distinguished in 2004, but are now labelled altogether "Divorced". This implied 7,406 changes in the variable.

Income Variable (variable: `eink`) This is the only income variable available in the sample, representing net household monthly income. The raw data contain classified data: Each household belongs to an income class where the bounds are known for all intermediate classes. I replace the class codes by the upper bound of the class. For the last open interval class with the high income households I set the value to 8,000 € which amounts to roughly three times the average German net income per household, see footnote 21. For the first interval with low income households I set the value to 750 € which corresponds to minimal social transfers plus money transfers for the rent.

2.3.3.2 Remarks on created Variables

In this subsection I want to emphasize details of the newly constructed variables.

Gender (variables: `psexw`, `psexm`) I create two variables that directly indicate the number of female and male household members.

Urbanity (variable: `urban`) This dummy variable equals one if more than 50,000 inhabitants live in the community and indicates an urban character of the household surroundings.

Employment Variables (variables: `bstel`, `bself`, `bwhite_h`, `bwhite_l`, `bblue`, `bunemp`) To simplify the usage of the precisely coded occupational status in `bstel`, I aggregate some of the values in that variable to common groups: `bself` indicates self employed persons, `bwhite_l` and `bwhite_h` represent low and high position white collar persons, `bblue` marks blue collar workers and `bunemp` represents unemployed persons. Note that it is unknown whether this is the highest occupational status in the household, that of the working person or that of the household managing person. Presumably it will be the first of these alternatives, as is common practice.

2.3.3.3 Summary Statistics

In this section I present summary statistics for the demographic data. Table 2.9 contains descriptive statistics for all sampled households. This includes also households that have never purchased a detergent.

Table 2.9. Summary Statistics for Sociodemographic Data

Variable	Mean	Median	SD	IQR	Min	10%	90%	Max
eink	2363.75	2000.00	1481.77	1250.00	750.00	1000.00	3500.00	8000.00
urban	0.76	1.00	0.43	0.00	0.00	0.00	1.00	1.00
hhsize	2.35	2.00	1.20	2.00	1.00	1.00	4.00	10.00
kizahl	0.50	0.00	0.86	1.00	0.00	0.00	2.00	8.00
hhfage	45.73	43.00	15.12	24.00	18.00	27.00	67.00	95.00
hhfpsex	0.71	1.00	0.45	1.00	0.00	0.00	1.00	1.00
bself	0.09	0.00	0.29	0.00	0.00	0.00	0.00	1.00
bwhite_h	0.45	0.00	0.50	1.00	0.00	0.00	1.00	1.00
bwhite_l	0.24	0.00	0.42	0.00	0.00	0.00	1.00	1.00
bblue	0.21	0.00	0.41	0.00	0.00	0.00	1.00	1.00
bunemp	0.01	0.00	0.12	0.00	0.00	0.00	0.00	1.00
membership	1692.11	622.00	1766.78	2517.00	0.00	203.00	4818.00	5596.00
membership2	481.70	436.00	239.75	477.00	0.00	187.00	730.00	748.00

Notes: All statistics are constructed by household (HH).

Recall that income is classified data (eink). The average and median income are below the German average income of 2,800 €.[21] Figure 2.5 shows a bar chart of income to get an impression of the income distribution. The chart does not look smooth, since there are peaks at 1,500, 2,000, 2,500 and 3,500 €. 76% of the sampled households live in a community with more than 50,000 inhabitants (urban). The median household size is two, the average being slightly higher, and the number of children per household has a median of zero and a mean of 0.5 (hhsize, kizahl). Hence, mostly two person adult households are in the sample. The average age of a household managing person is 45 years with 71% being females (hhfage, hhfpsex). 69% of the sample are white collar employees, 21% are blue collar workers, 9% are self employed and 1% is unemployed. The detailed occupational status is visualized in figure 2.6. Next I turn to the membership duration, defined as the total time the household spent in the Homescan panel. I have the entry date and exit dates of each household to the Homescan panel (beginn, ende). I use these to calculate membership times. The average total time households were participants in the Homescan Panel is 1,692 days, with a median of 622 days (membership). If I consider only the time period for which purchase data are available the mean is 482 days and the median is 436 days (membership2). The latter fits

[21] The average monthly net income for all households in Germany was about 2,800 € in 2005. The source is a standard report of the Statistische Bundesamt, Report Name: "Nettoeinkommen und Zahl der Haushalte nach Haushaltsgruppen 1991 bis 2005". Recall that I use upper bounds of the class intervals as per class means. This biases income upwards, but top incomes are cut off by setting maximal income to 8,000 €, and naturally, the latter does not impact the median.

Figure 2.5. Net Monthly Household Income

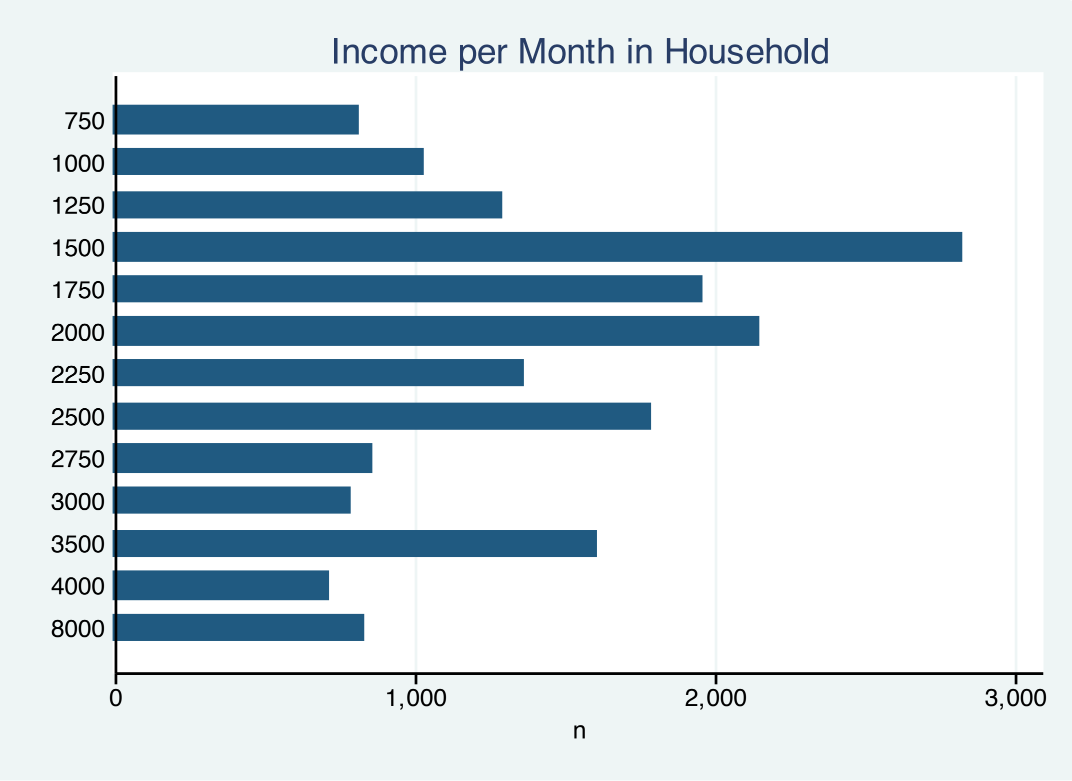
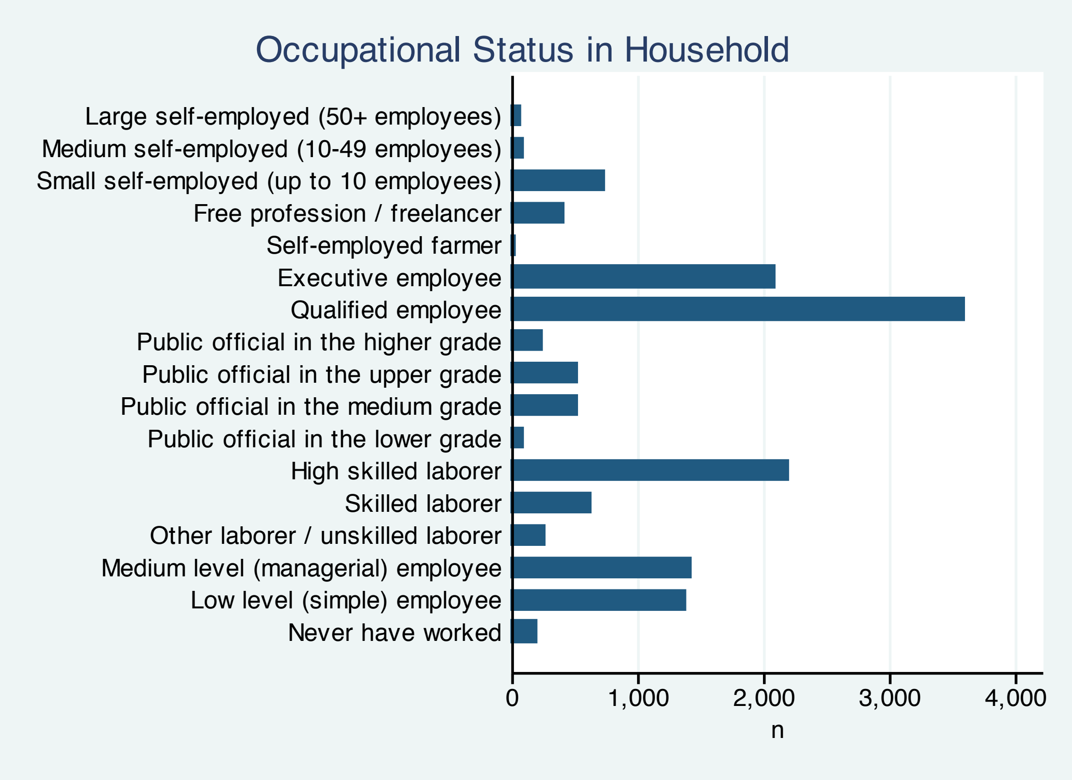
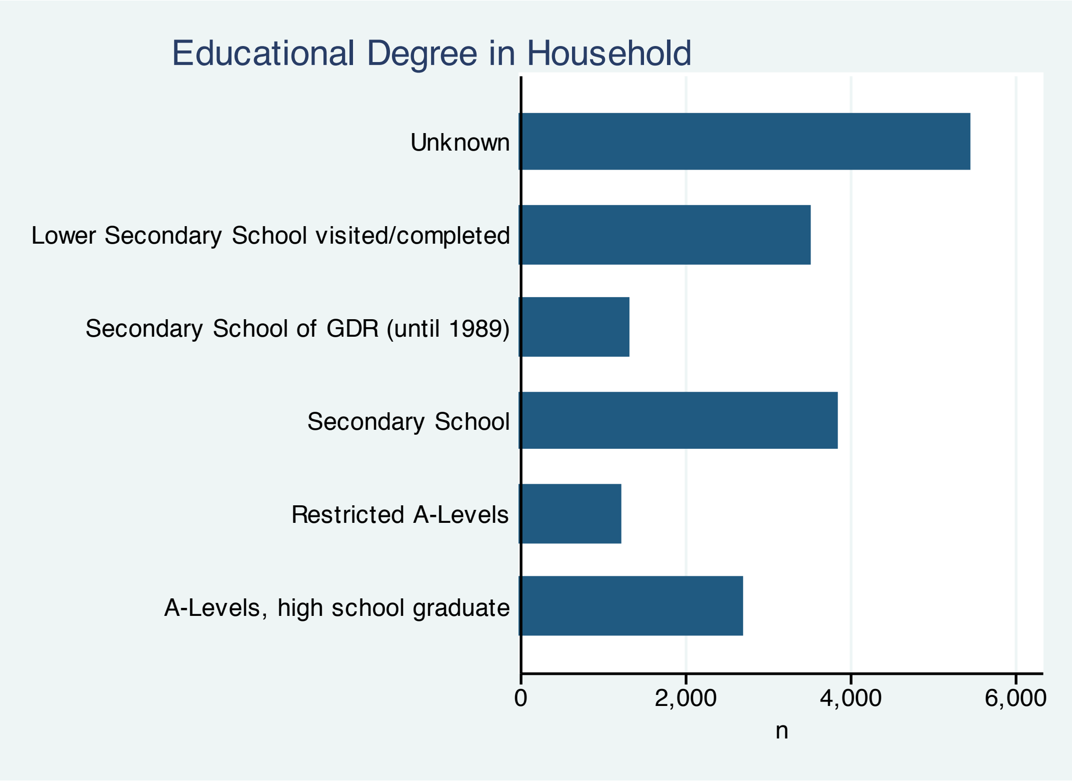


Figure 2.6. Occupational Status of Households



well the data in figure 2.4 that displays the duration between the first and last detergent purchase in the sample for each household in the category purchase data. Educational degrees for the sampled households are displayed in figure 2.7. For about a third of the sample no educational information is available. Just as in the case of the occupational status it is unknown to whom the educational degree belongs to. The possibilities are: the highest of the household, the household managing person or the working person.

Figure 2.7. Education of Households



2.3.4 Advertising Data

The raw advertising data consist of two kinds of files that are available for each year (2004 through 2006) giving six files in total. The first kind contains the contacts that households had with advertisements of specific advertising campaigns and the second provides campaign information. The consolidated data file has 1,899,852 observations. One observation represents an advertisement contact of a household with a specific spot along with the spot information. Consult table 2.10 for an overview.

Table 2.10. Overview of Variables in Advertisement Contact Data

Variable	Numeric	Position	Description
`hhnr`	1(=yes)	1	Household ID code
`kam`	1	2	Advertising Campaign ID code
`laenge`	1	4	Length of Spot
`mke`**	1	10	Advertised Brand Name
`sender2`	0(=no)	5	TV-Station
`spotimg`	1	13	Spot for Image Building
`spotliq`	1	11	Spot for Liquid Detergent
`spotpow`	1	12	Spot for Powder Detergent
`spotwash`	1	14	Spot for Detegent, not wash additive
`tvdate`	1	9	Broadcast Date (STATA) YYYY/MM/DD
`tvdateday`	1	8	Broadcast Date as DD
`tvdatemonth`	1	7	Broadcast Date as MM
`tvdateyear`	1	6	Broadcast Date as YYYY
`zeit`	1	3	Begin of Broadcast

Notes: Numeric indicates whether a variable is numeric. Position gives the column position of the variable in the file. STATA in the description means that the variable is in STATA date format. ** marks that variable is coded, but is not disclosed in the appendix due to the confidentiality agreement for data usage.

2.3.4.1 Data Modifications

The major task was to combine the 6 data files into one file. Apart from minor details it was not necessary to do any modifications to the raw data, signaling the good quality of the advertising data.

Combining The first kind of data files contain the raw advertisement contact data for each year.[22] Each observation is a household that has seen a specific spot of a campaign at a given time, identified by household id, broadcast date/time and campaign id. The second kind of data files contain the advertising campaign information for each year.[23] Each observation is a campaign with the spot information, identified by campaign id. Campaign ids are newly assigned each year. That is why campaign and contact data

[22] Total number of advertisement contacts for all households in a given year. 2004: 696,168 2005: 839,715 2006: 363,971.

[23] Total number of advertising campaigns for all products in a given year. 2004: 57 2005: 156 2006: 64.

have to be merged per year and are then appended. In the merged data, the campaign id is dropped.[24]

I merged the information from the campaign files into the contact files without dropping any information apart from campaign id so that I do not present the data file types individually.
Campaigns Two campaigns were never on air in the sample (numbers 136 and 151 in the year 2005). No contact with these campaign ids appeared so that both are dropped.
Variables dropped The following variables from the campaign data in table 2.11 are dropped: `prodnr`, `motivnr`, `prodstr` and `motivstr`. Note that I extract information from `prodstr` and `motivstr` before dropping them, see the next paragraph.

2.3.4.2 Remarks on created Variables

Spot characteristics From the campaign information in table 2.11, I extract information on the advertisement from the variables `prodstr` and `motivstr` and create new variables: a brand variable `mke`, two dummies for whether detergent advertised is powder or liquid (`spotliq` and `spotpow`), one dummy for whether the advertised detergent is a detergent but not a wash additive (`spotwash`) and one dummy for umbrella brand advertising in form of image spots (`spotimg`).[25]

Table 2.11. Campaign Information Data

Variable	Description
`kam`	Campaign ID (different numbers if content or spot changed)
`prodnr`	An internal A.C. Nielsen code
`prodstr`	A text defining the product
`motivnr`	An internal A.C. Nielsen code
`motivstr`	A text defining the spots content

2.3.4.3 Summary Statistics

This section presents the summary statistics for the advertising exposure data. Table 2.12 presents descriptive statistics for the advertisement contact data.

Average spot length is roughly 20 seconds, and 90% of all spots are less than 32 seconds long (`laenge`). Concerning the inferred informational contents of the spots, 83% of the spots advertise common detergents (`spotwash`) and the rest advertise wash additives. 53% of the spots advertise powder detergents (`spotpow`), 25% do so for liquid detergents

[24] Naturally, it is possible to use the campaign id to generate a campaign dummy to mark advertisement contacts that have been achieved with a given campaign.

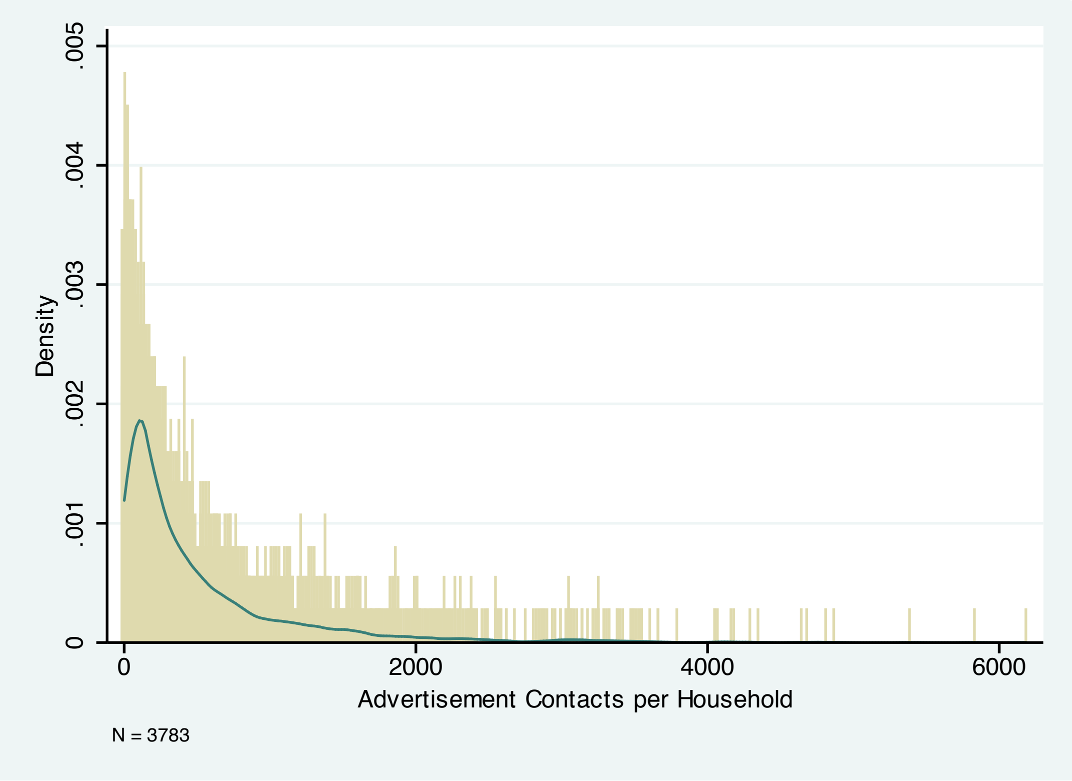
[25] The difference between `spotliq, spotpow` and `spotwash` is that the latter contains both former detergent types and additionally wash capsules which are hardly sold.

Table 2.12. Summary Statistics for Advertising Data

Variable		Mean	Median	SD	IQR	Min	10%	90%	Max
laenge	PI	21.03	20.00	9.19	15.00	5.00	10.00	32.00	101.00
spotwash	PI	0.83	1.00	0.38	0.00	0.00	0.00	1.00	1.00
spotpow	PI	0.53	1.00	0.50	1.00	0.00	0.00	1.00	1.00
spotliq	PI	0.25	0.00	0.43	1.00	0.00	0.00	1.00	1.00
spotimg	PI	0.05	0.00	0.21	0.00	0.00	0.00	0.00	1.00
hhobsa	HH	502.21	268.00	647.81	533.00	1.00	35.00	1272.00	6184.00
brandhhic*	HH	0.19	0.17	0.08	0.04	0.11	0.14	0.24	1.00
brandhhiz**	HH	0.19	0.17	0.08	0.04	0.11	0.14	0.24	1.00
senderhhic*	HH	0.35	0.31	0.16	0.19	0.14	0.19	0.56	1.00
senderhhiz**	HH	0.35	0.31	0.17	0.19	0.13	0.19	0.56	1.00

Notes: * indicates the HHI is calculated with volume counts as weights. ** indicates the HHI is calculated with volume time as weights. The second column indicates whether statistics are constructed by purchase incident (PI) or by household (HH).

Figure 2.8. Advertisement Contacts per Household

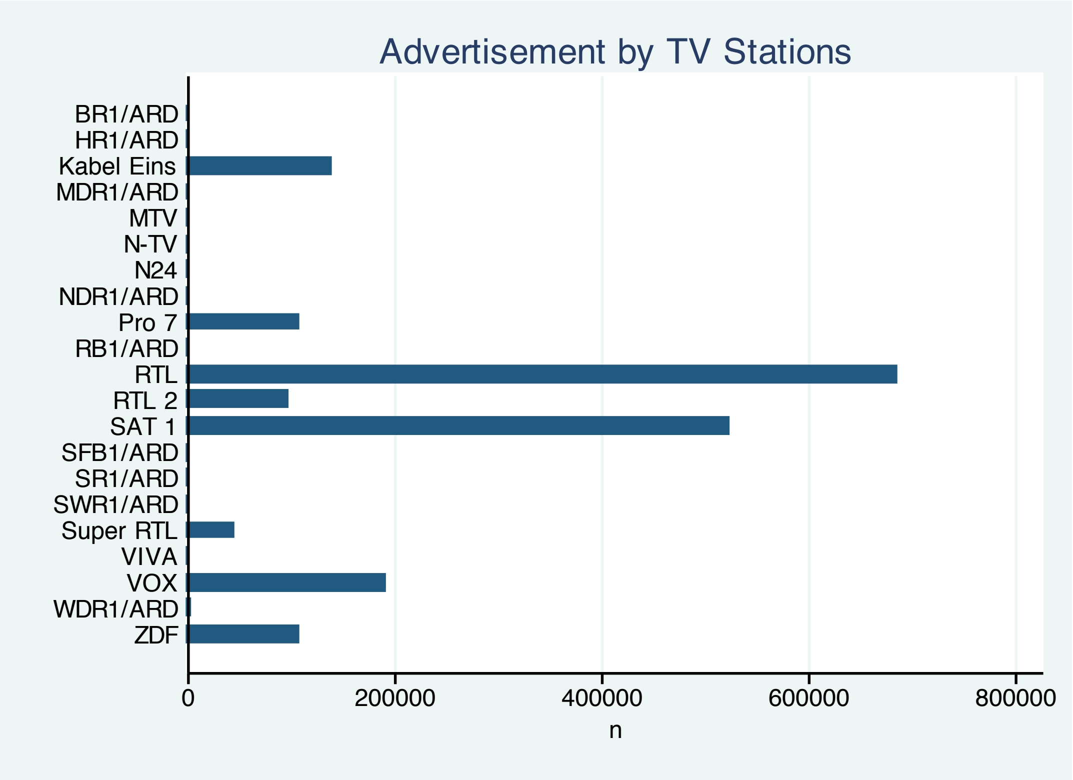


(spotliq). 5% spots are identified to be spots that advertise solely the brand in general, serving presumably for image/prestige recognition (spotimg).

The average household has 502 contacts with spots during the sampled period, with a median of 268. This seems to reflect the difference in TV consumption: heavy consumers, e.g. consult the ninetieth percentile, have seen 1272 spots.[26] Figure 2.8 presents the density of advertising contacts. The density clearly diminishes with the increasing number of contacts.

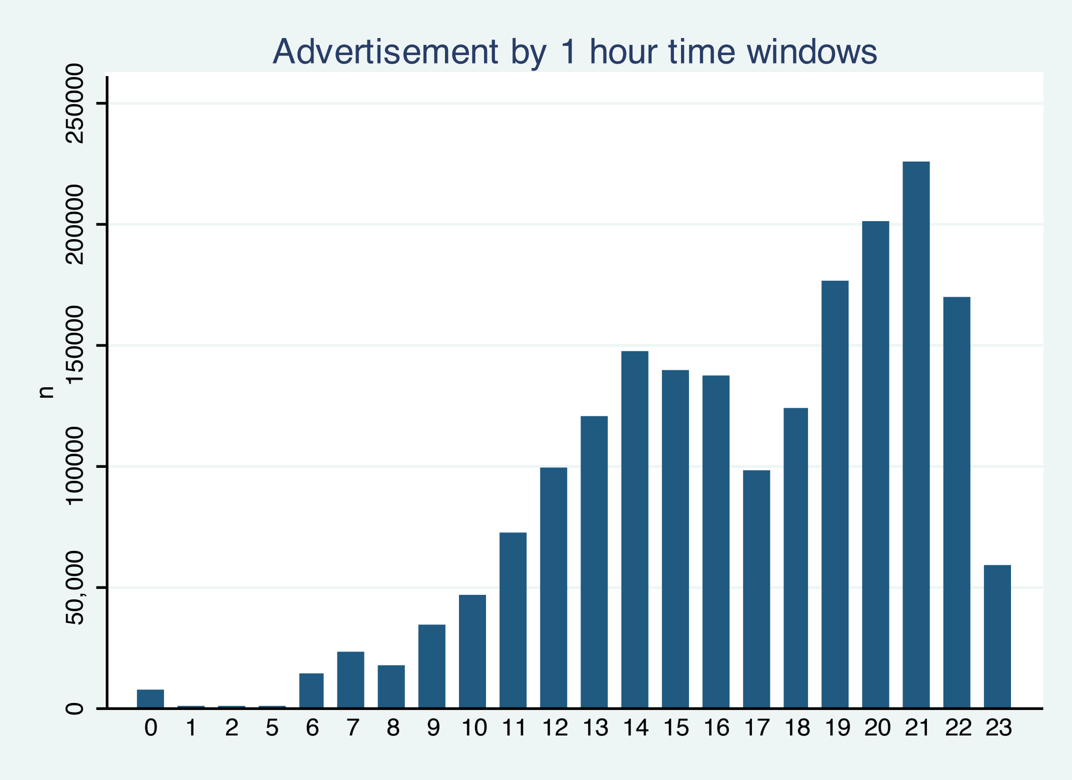
The concentration measures for brand indicate a low concentration, concluding that consumers see advertisements of almost all advertised products (`brandhhi,brandhhic`). Differently, the concentration measures for TV stations indicate higher concentration, but are still in the medium range between 0.3 to 0.7. As can be seen from figure 2.9 two TV stations broadcast most of the advertisements, consistent with the concentration measure. The other TV stations are quite equally sized concerning the total number of achieved advertising contacts. Figure 2.10 shows the intraday distribution of advertisements across one hour time windows. Visible are the expected afternoon and evening peaks, with a gap at early German evening dinner time.

Figure 2.9. Advertisement Contacts broken by TV Stations



[26] Assuming 20 seconds per spot this household will have seen 424 minutes, roughly seven hours, only of detergent advertisements.

Figure 2.10. Advertisements by Day time Windows



2.4 Combined Dataset

2.4.1 Construction

The four data files presented so far, namely general/category purchase data, sociodemographics and advertising exposure data, can be merged to yield a combined dataset that can be used for various types of analysis. I sketch the process to create a dataset for discrete choice analysis. Two exemplary applications with the data are found in chapters 3 and 4 of this dissertation.

In the final combined data an observation is a potential product choice a household had during a shopping trip, called incident thereafter, so the choice set the household faces at an incident consists of several observations. To be clear, an observation is a row in the data file. The incident is identified by household id, date of purchase, store name and store id, namely the variables `hhnr`, `edate`, `key_acc` and `store`. Distinguish purchase incidents that result in purchase of a detergent product and no-purchase incidents that result in choosing the no-purchase alternative.

A product is defined by the following product characteristics: consistency, brand, general purpose, contents, scent, color, sensitive, concentrate, sold with a giveaway, special size and bigpack.[27] This product definition differs from the definition of the EAN code, but the differences between the product definitions are almost indistinguishable in terms of product characteristics for a consumer.

Each alternative picked from the choice set at an incident is marked by the purchase dummy `purchase`. In the data file several choices can be made per incident, so before using the data for classical discrete choice analysis one has to deal with this issue. See the handbook article by Ackerberg, Benkard, Berry, and Pakes (2006) for recent approaches. If one is not willing to adjust the model for purchase of multiple units/products, the incidents have to be dropped. As a consequence, one accepts the following two assumptions: (i) choice of several goods is equal to the choice of several independent single product choices and (ii) purchase incidents with choice of several goods are not systematically different from purchase incidents with choice of a single good. Of course, the general solution without dropping observations is to enhance the choice set with the multiple unit/product choices, but this yields a situation where the interpretation of the choice set alternatives is unintuitive and problematic.[28]

The following steps are undertaken to yield the combined dataset:

1. Category purchase data: Identify and remove duplicates according to the above definitions of purchase incident and product. A duplicate is an observation that has the identical values for purchase incident and product choice as another observation, e.g. the consumer went twice to the same place and bought the same product. Of these, only one transaction is kept, but the number of units bought is increased to match the number of duplicates of an observation. The number of units of a purchased alternative is recorded in variable `menge`.

2. Category purchase data: Create a no-purchase alternative per purchase incident.

3. General purchase data: Select incidents that are to be kept (transactions with total value above 5 €). Then 601,153 of 3,058,880 incidents are dropped. A total value below 5 € indicates that the consumer was looking for a specific item and not considering the detergent category.

4. General purchase data and category purchase data: Merge both into a new file according to the definition of an incident. In the following I call it purchase data.

5. Purchase data: For each incident create observations with product alternatives that were offered to any consumer at the same time and place.

[27] The corresponding variables are listed here and detailed in section 2.3.2: `kons, mke, uwg, inh, duft, color, sensitiv, konzentrat, gimmick, extra_size, bigpack`.

[28] Example: Suppose there exist two brands A and B, and the no-purchase alternative. Choice set for exclusive choice of one brand is $\{\emptyset, A, B\}$, if the consumer can buy both brands it is $\{\emptyset, A, B, AB\}$. Now, it is difficult to explain the choice AB compared to choosing only A or B because the model tries to explain this with characteristics and it is in general not obvious how to define the characteristics of AB.

6. Purchase data: Merge with zip code and county ("Landkreis") data to add county id to the data.[29]

7. Purchase data: Merge with sociodemographic data.

8. Purchase data: Merge with advertising data, generate advertising variables.

9. Purchase data: Create durations since last purchase, state dependence variables for brand purchases.

2.4.2 Inferring Prices in Step 5

As explained earlier, household level data record prices of products that are bought from a sampled household, but do not have prices of product alternatives that were not bought by any sampled household. Therefore, to generate the alternative product choices for all incidents in the purchase data it is not possible to simply take all products bought in the sample and offer them as choices for each consumer in an incident. That is why this step needs a careful discussion and it is only possible to add alternatives for which prices can be deducted from the available data.[30]

I infer prices as Erdem and Keane (1996) and Keane (1997). Firstly, the large number of households permits to close the price gap of a product alternative faced by a consumer by filling it with prices of other consumers that shopped in the same store at the same time. Usually, all data are aggregated to weeks. With daily data at hand and the assumption of constant prices over a week at a given outlet, I can use price information from other days within a week for the same outlet to infer prices of product alternatives for many incidents.

Secondly, a subgroup of retailers have nationwide targets so that filling the gaps can be done by using the information from other outlets of that retailer.

A possible list of necessary steps to fill up missing product alternative prices looks as follows:

1. Collect all product purchases in a calendar week in a chain in a county ("Landkreis"). Fill up missing product alternative prices in all incidents within the same week, chain and county. Repeat this for all weeks, chains and counties.

[29] I use a freely downloadable file from the public domain dataset OpenGeoDB that links zip codes and county ids. Missing values on the zip codes are filled up using the DeutschePost website zip code tool to identify the city area of the zip code. Then I use publicly available administrative data to find the county in which the city is located and add the information to the purchase data. About thirty zip codes were missing in OpenGeoDB.

[30] In principle, the price data exist in the store level scanner data collected at the supermarkets, but this is not integrated into the Single Source data of A.C. Nielsen. Moreover, the store level scanner data are commonly hard to get for research.

2. If this fails, widen the time interval from a single week to two adjacent weeks, use the average price of the weeks based on the days with sales and redo step 1.

3. For the nationwide chains with national price targets, collect all product purchases in a week in a chain. Fill up as in 1.

4. If this fails, widen the time interval from a single week to the two adjacent weeks and proceed as in 3.

These steps offer a tradeoff between price information accuracy and more observations. When conducting several of the above mentioned steps, a larger dataset with possibly erroneous prices is obtained, whereas if only step 1 is conducted prices are precise at the cost of having less observations.

I restricted myself to step 1 because I want to avoid measurement error in prices. Especially if the researcher wants to use Hausman (1996) instruments this is critical for the validity of the instrumental variable approach. Hausman instruments for a given price are actually prices of similar products at the same time in other outlets and locations. Hence, it is important to clearly avoid any overlap in the two actions of inferring missing prices and construction of Hausman instruments. In the end, the consumer decision is based on the transaction price, so it should be clear what price is a potential instrument and what price is an inferred and correct transaction price.

If after step 1 no product alternative for a given no-purchase incident is found, this incident, consisting of only one observation, is dropped. This is the case for 1,836,896 observations (= no-purchase incidents). This is done without loss of generality as the informational content of these observations is low, because the no-purchase alternative is the only alternative the consumer has. The same step was not done for purchase incidents that led to a detergent purchase, as depending on the application, one may be still interested in these observations although product alternatives are missing.

Looking at the number of prices available per zip code I find that the price information is too scarce to infer prices on the zip code level. There are about fifteen thousand zip codes that map into 434 counties, i.e. “Landkreise”. The fewer number of counties than zip codes makes it much more likely to find two households in the same area.

Different from the US, in Germany during the sample period the practice of issuing price coupons in stores that reduce the retail price is not common so that no correction is needed for the imputed prices. Keane (1997) notes that existence of the US coupon-redemption system leads to exaggerated price elasticities from models that do not account for this.

2.4.3 Remarks on created Variables

Different to the single files, the combined file was especially created for the analysis within a discrete choice model in chapters 3 and 4 of this dissertation. That is why I will only sketch the generated variables and refer the interested reader for details to the work mentioned.

Duration (variables: `duration, duration2, idurhh`) The `duration` is constructed such that on each alternative the value is the time in weeks since the last store visit that led to a detergent purchase. To make the variable identifiable I set the value of the variable to zero for the no-purchase alternative. Since the first observation of each household has no prior visit, the value of these first observations is set to the missing value. `duration2` is merely the square of `duration`. In the case of multiple purchases on a certain day, duration is still weeks elapsed since previous purchase at an earlier date, therefore `duration` is never zero for a brand alternative. `idurhh` is the interaction of `duration` and household size in persons.

State Dependence (variables: `GLdumA, GLdumB, mdum*`)[31] These variables can be used to control for state dependence. It is modeled as so-called brand loyalty where this term highlights the underlying habit of the consumer. `GLdumA` and `GLdumB` are simple dummies that take the value one if the previous purchase was of the same brand as the faced alternative. `GLdumA` contains values on all incidents, `GLdumB` sets all dummies to zero for no-purchase incidents, i.e. incidents that resulted in choosing the no-purchase alternative. When looking at more than one purchase further back into the past, a series of dummies can be constructed according to the same rules as `GLdumB`. This results in a series of dummies `mdum1, mdum2, ...`. The integer numbers specify the lag.

Advertising (variables: `countc140ad`, `liqc56adr`, `countc140adpr` ...) This set of cryptic variables define the TV advertisement contact of the household with brand specific advertisement. Each variable code consists of 3 parts and I explain each component.

The variable name up to the letter `c` (i.e. `xxxxxc140ad`) defines the type of information taken from the spot: `count` indicates contact with a spot for the brand, `time` gives time length of contact in seconds with a brand spot, `liq` signals contact with a liquid detergent spot, `pow` flags contact with a powder detergent spot and `img` indicates contact with an image spot.

The number after the letter `c` (i.e. `countcxxxad`) defines up to which lag in days the advertisement contacts are cumulated. The values for the lag are multiples of 14: 14, 28, ..., 126, 140.

After that number the following keywords `ad, adr, adpr` detail variable construction (i.e. `countc140xxxx`): standard is `ad` where advertisement contacts are simply cumulated per brand for the specified lag so that the variable sums absolute contacts in a given time window. If the variable ends with `adr` it measures advertising pressure of

[31] The $*$ is a wildcard as in usual programming languages and represents an integer.

one brand relative to all competitors: it is the same as the variable ending on `ad`, but in addition it is divided by total advertisement contacts of all competitors up to the same lag length. If the variables ends with `adpr` it captures advertising pressure between two time windows, i.e. `countc140adpr` captures the relative advertising pressure in the time window of 140 to 126 days before the purchase.[32]

Contents (variables: `inh, inhp, inhl`) The `inh` variables define the contents of a detergent package: `inh` quantifies it for powder (in kilogram) or liquids (in liters), while `inhp` is a nonzero kilogram value only for powder (and zero for liquids) and `inhl` is a nonzero liter value for liquid detergents (and zero for powder).

2.4.4 Basic Analysis

As the combined data are merely a merged version of all data files, I relinquish a summary table. Instead I want to give an overview of the detergent market by brands to understand the market segmentation. Table 2.13 presents the related results. After that I want to uncover basic dependencies between the variables newly combined. For this task I use the Linear Probability Model (LPM) and the Discrete Choice Logit model

Table 2.13. Detergent Market - Market Shares, Advertising and Characteristics for all Product Purchases broken up by Brand

	Means		Shares								
Brand	Nominal Price	Size in kg or liters	Sales by Volume	TV Advertising	Liquid	Konzentrat	Color	Sensitiv	Gimmick	Extrasize	Bigpack
3	5.37	2.41	0.06	0.36	0.41	0.59	0.51	0.04	0.06	0.00	0.13
6	3.40	1.48	0.01	0.00	0.61	0.26	0.10	0.00	0.01	0.03	0.00
8	3.72	1.45	0.02	0.02	0.68	0.00	0.80	0.00	0.13	0.00	0.00
10	4.38	3.86	0.01	0.00	0.48	0.09	0.26	0.01	0.07	0.00	0.00
11	3.90	2.36	0.01	0.00	0.61	0.89	0.00	0.00	0.04	0.18	0.09
13	1.92	1.13	0.02	0.00	1.00	0.00	0.13	0.02	0.00	0.00	0.00
16	2.70	1.84	0.56	0.00	0.52	0.53	0.42	0.01	0.00	0.00	0.00
22	3.42	1.94	0.01	0.00	0.90	0.00	0.05	0.00	0.00	0.00	0.00
36	4.92	3.37	0.02	0.07	0.41	0.26	0.16	0.00	0.06	0.09	0.06
40	6.79	2.98	0.07	0.32	0.28	0.47	0.40	0.11	0.13	0.12	0.06
41	3.99	1.40	0.03	0.04	0.79	0.00	0.14	0.00	0.00	0.04	0.00
55	3.58	1.80	0.09	0.14	0.54	0.39	0.14	0.00	0.01	0.01	0.00
57	4.22	2.96	0.02	0.00	0.49	0.29	0.42	0.00	0.00	0.10	0.00
67	4.79	3.08	0.03	0.04	0.36	0.28	0.02	0.00	0.16	0.10	0.02
100	4.63	3.79	0.02	0.00	0.19	0.19	0.06	0.00	0.00	0.03	0.06
Total	3.51	2.08	0.07	0.06	0.51	0.45	0.36	0.02	0.03	0.02	0.02

[32] Note the difference: The variable countc140adr captures relative advertising pressure from 140 to zero days before the purchase incident.

(DCLM) to analyze the effect of purchase relevant variables on the purchase decision of households. The results of the estimations are displayed in table 2.14, 2.15 and 2.16. The latter two tables are in appendix part A.

Table 2.13 gives a market overview by breaking up sales, advertising and product characteristics variables by brand. The last line shows market averages. Especially, I look at 4 major brands that have 78% market share, namely brand 3, 16, 40 and 55. Together, brands 3, 40 and 55 are the biggest TV advertisers in the market. Brand 16 has 56% market share and charges prices below the market average price. Brand 16 represents the private labels of the retailers, the non-branded low priced alternatives. Naturally, there is no TV advertising for this brand. Brand 3 and 40 charge the highest prices and have high variabilities on the characteristics compared to the other brands. Hence, the product portfolios of 3 and 40 cover the full range of characteristics with the exception of `extra_size`. Different from that brand 55 covers the three most important product characteristics identified in the preceding paragraph, has an average price about equal to the average market product and represents the third largest TV advertiser.

In the LPM the purchase dummy is the dependent variable and a selection of the variables described earlier enter as explanatory variables. The purchase dummy is one for the product alternative that is chosen, this can be a detergent or a no-purchase alternative. Each alternative the consumer has is an observation. This implies that purchase trips and households are not modeled. Besides this obvious modeling limitation, it suffers from the known caveats.[33] The LPM just serves to verify intuitive expectations about the interplay of the variables.

The DCLM has also the purchase dummy as dependent variable, but different to the LPM, the variable serves to identify the alternative chosen at an incident, i.e. a shopping trip. Thus, the DCLM compares the alternatives at a shopping trip. Households are still not explicitly modeled.[34]

The tables 2.14 to 2.16 show LPM and DCLM estimates for different samples. The samples differ in terms of the (i) product alternatives that are available to the consumers and (ii) incidents used for the estimation. An incident can lead to the choice of a product or result in a no-purchase. Each of the aforementioned tables is for a specific set of product alternatives. Within each table, I use for the estimation either all incidents or only purchase incidents that result in a product purchase, as indicated at the top of each table.

[33] The linear probability model, a binary choice model, is simply an OLS regression of a binary variable on explanatory variables. Therefore, marginal effects are constant for variables that enter linearly over the whole range, which is not very intuitive in the case of a binary dependent variable. Also, an OLS regression can give predictions of the dependent variable that are not valid probabilities, e.g. values outside the unit interval. For more details consult a graduate textbook, e.g. chapter 15 of Wooldridge (2002).

[34] Note that in the DCLM the coefficients are not comparable to the LPM and are identified up to scale, i.e. the ratio of two coefficients is identified. For more details consult chapter 3 of Train (2003).

Table 2.14. Linear Probability Model and Discrete Choice Logit Model with all Product Alternatives

	LPM				DC Logit
Incidents	all	all	all	purchase	all
Variables	Coef./SE	Coef./SE	Coef./SE	Coef./SE	Coef./SE
preis	−.048***	−.062***	−.063***	.026***	−.428***
	(.00)	(.00)	(.00)	(.00)	(.01)
inh	−.068***	−.051***	−.051***	.034***	−.126***
	(.00)	(.00)	(.00)	(.00)	(.02)
feature	−.045***	−.065***	−.072***	.043***	.179***
	(.00)	(.00)	(.00)	(.01)	(.04)
display	.041***	.036***	.033***	−.025**	.155***
	(.00)	(.00)	(.00)	(.01)	(.04)
prceflag	−.025***	−.026***	−.032***	−.008	−.008
	(.00)	(.00)	(.00)	(.01)	(.04)
handbill	.031***	.024***	.028***	−.010	.021
	(.00)	(.00)	(.00)	(.01)	(.04)
liquid	−.331***	−.297***	−.295***	.204***	−1.284***
	(.00)	(.00)	(.00)	(.00)	(.02)
konzentrat	−.318***	−.323***	−.321***	.145***	−1.566***
	(.00)	(.00)	(.00)	(.01)	(.02)
color	−.120***	−.143***	−.144***	.067***	−.768***
	(.00)	(.00)	(.00)	(.01)	(.02)
sensitiv	−.031***	−.121***	−.133***	.087***	−.448***
	(.01)	(.01)	(.01)	(.02)	(.08)
gimmick	.113***	.065***	.075***	−.005	.556***
	(.00)	(.00)	(.00)	(.01)	(.06)
extrasize	.045***	−.033***	−.028***	−.001	.022
	(.01)	(.01)	(.01)	(.02)	(.07)
bigpack	.404***	.333***	.345***	−.148***	1.821***
	(.01)	(.01)	(.01)	(.02)	(.08)
countc56ad	−.001**	−.001***	−.001***	−.000	−.008*
	(.00)	(.00)	(.00)	(.00)	(.00)
GLdumB			.594***	.278***	5.640***
			(.00)	(.00)	(.04)
duration			−.003***	.002***	−.032***
			(.00)	(.00)	(.00)
constant	.793***	.804***	.797***	.044***	
	(.00)	(.00)	(.00)	(.00)	
dummies					
brand	No	Yes***	Yes***	Yes***	Yes***
No. of obs	375,448	375,448	318,170	39,726	317,890
R2	.52	.54	.59	.45	

Note: Asterisks indicate significance levels at $\alpha : * = 0.05, ** = 0.01, *** = 0.001$. Variable `countc56ad` is total number of advertisement contacts within a three month period per market.

In table 2.14 the complete combined data with all product alternatives are used. In table 2.15 the results build on the same data but with all no-purchase alternatives removed. Historically, this data setup resembles the first brand choice models that did not include a no-purchase alternative. In table 2.16 the complete combined data are freed of all no-purchase alternatives and all private label alternatives.

Before discussing the results, I want to summarize the expected dependencies among the variables. Prices impact the propensity to purchase a product negatively, whereas TV advertising, retail activity variables and product characteristics are generally beneficial to the household and affect purchases positively. State dependence should have a positive effect on purchases, as consumers like to stick to their brands as seen earlier in the descriptive statistics of the category purchase data. Duration should impact purchases positively because the more time since the last purchase has elapsed the emptier the stock of detergent at home and the more likely is a purchase of detergent.

In table 2.14, some interesting findings for the LPM occur. The price coefficient is statistically significant and negative if all incidents are used, otherwise it is positive as visible in the fourth model column. Generally, most of the variables, perform a sign switch, when considering only purchase incidents compared to all incidents. Interestingly, advertising is negatively significant. Product characteristics linked to the wash functionality (liquid, konzentrat, color, sensitiv) switch to have a positive effect on the purchase probability. The effect of state dependence is positive and significant as expected without sign switch. For example, if sticking to the previous brand choice is fancied by the consumer, the coefficient will be positive. Duration has the expected sign if only purchase incidents are considered. The intuitive argument goes as follows: If duration is higher, inventories tend to be lower so that the probability of purchase should be higher.

The underlying explanation for the sign switch is the following. If all incidents that lead to any outcome are considered, there are many incidents that lead to choice of a no-purchase alternative. Compare the sample size decrease from 318,170 to 39,726 when removing all no-purchase incidents that result in choosing the no-purchase alternative. This alternative has a low value on all variables beneficial to the household that a product usually has: product characteristics, advertising and retail activity. In addition, it has a low value on price as it costs zero. The simple binary model then compares the many no-purchases to the purchases of brands. As the number of no-purchases is very high, the model evaluates the variables that are beneficial to the household as bad, because households mostly chose not to purchase products but chose the no-purchase alternative. At the same time prices of the no-purchase alternative are low compared to products that all have positive prices and are seldom chosen. That is why the price coefficient has the right sign if all incidents are considered. It gets positive if only purchase incidents that lead to a product purchase are considered. Then consumers seem not to chose always the cheapest alternative, ceteris paribus. This may be an indication for an omitted variable bias problem, e.g. that quality is not adequately controlled by the variables in this model.[35]

The DCLM performs quite convincing, with the exception of the signs on the product characteristics and advertising, both suffering from unintuitive significant signs. Just as for the LPM, in the DCLM, the high number of chosen no-purchase alternatives leads to a counter intuitive result concerning the valuation of beneficial product characteristics. Compared to the LPM, both retail activity variables feature and display have the expected positive sign. The state dependence dummy proves to be the most important variable.

Tables 2.15 and 2.16 highlight further the importance of the state dependence variable, whereas prices turn insignificant in the DCLM model. Otherwise, no major changes occur. Generally, statistical significance of most variables and explanatory power in terms of the coefficient of determination R2 is much lower than compared to table 2.14.

Of course, these very simple models are not meant to explain the purchase patterns of all consumers, but they shed a first light on the possibilities offered by the dataset at hand.

2.5 Conclusion

This data description has demonstrated that there are rich data available for economic research. The uncommon link between detailed purchase data and advertising exposure data collected at the household level for the same households offers new research possibilities. The basic analysis shows there is underlying economic action that cannot be fully captured by simple models. Apart from demand analysis in differentiated product markets, the data may be used to study individual behavior, e.g. to assess the impact of advertising, estimate or test models of bounded rational behavior or study information processing of the consumer.

Marketing research companies seem to exert a high effort to collect data at a high quality as could be seen throughout the description. The number of obvious errors is very low compared to the number of observations available to the researcher. The described dataset is available for research in collaboration with the University of Mannheim so that it is possible to get high quality individual level data for research.

[35] The analysis of Trajtenberg (1990) is a classic example for the quality omitted variable bias story in empirical industrial organization, where CT scanners also have demand increasing in price.

Appendix

Appendix A: Tables and Figures

Figure 2.11. Geographic Distribution of Purchase Trips according to the Zip Code of Trip Destination

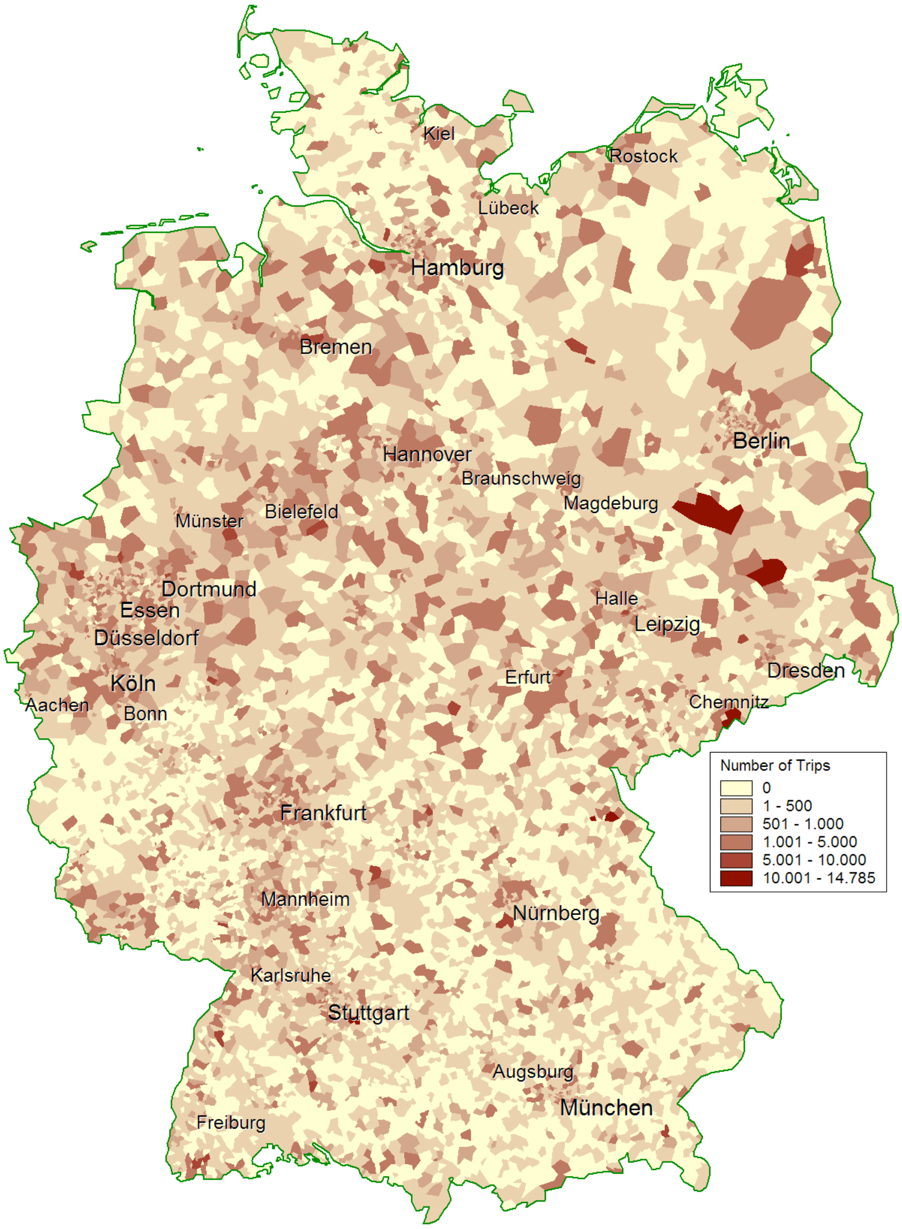


Table 2.15. Linear Probability Model and Discrete Choice Logit Model without no-purchase alternatives

	LPM				DC Logit
Incidents	all	all	all	purchase	all
Variables	Coef./SE	Coef./SE	Coef./SE	Coef./SE	Coef./SE
preis	.004***	−.004***	−.004***	−.005	−.042
	(.00)	(.00)	(.00)	(.00)	(.02)
inh	−.006***	.001	.002*	.006	.038
	(.00)	(.00)	(.00)	(.00)	(.03)
feature	.044***	.040***	.036***	−.018	.225*
	(.00)	(.00)	(.00)	(.01)	(.11)
display	.017***	.012***	.007**	−.015	−.140
	(.00)	(.00)	(.00)	(.01)	(.08)
prceflag	.018***	.008***	.006**	−.029**	.009
	(.00)	(.00)	(.00)	(.01)	(.08)
handbill	−.004	−.009***	−.005*	.001	.124
	(.00)	(.00)	(.00)	(.01)	(.09)
liquid	.002	.001	.003*	.017*	.102*
	(.00)	(.00)	(.00)	(.01)	(.05)
konzentrat	−.016***	−.012***	−.009***	−.035***	.090
	(.00)	(.00)	(.00)	(.01)	(.05)
color	−.007***	−.005***	−.005***	−.009	.024
	(.00)	(.00)	(.00)	(.01)	(.04)
sensitiv	.025***	.019***	.004	.006	.070
	(.00)	(.00)	(.00)	(.02)	(.15)
gimmick	.011**	.005	.010**	.019	.127
	(.00)	(.00)	(.00)	(.02)	(.13)
extrasize	.023***	.000	−.000	−.019	−.101
	(.00)	(.00)	(.00)	(.02)	(.15)
bigpack	.001	−.000	.001	.025	.176
	(.00)	(.00)	(.00)	(.02)	(.19)
countc56ad	.001**	−.001***	−.000*	−.000	−.002
	(.00)	(.00)	(.00)	(.00)	(.01)
GLdumB			.672***	.180***	1.696***
			(.00)	(.01)	(.07)
duration			−.000***	.000	
			(.00)	(.00)	
constant	.080***	.079***	.030***	.555***	
	(.00)	(.00)	(.00)	(.01)	
dummies					
brand	No	Yes***	Yes***	Yes***	Yes***
No. of obs	232,643	232,643	197,435	24,399	12,518
R2	.01	.01	.35	.04	

Note: Asterisks indicate significance levels at $\alpha : * = 0.05, ** = 0.01, * * * = 0.001$. Variable `countc56ad` is total number of advertisement contacts within a three month period per market.

Table 2.16. Linear Probability Model and Discrete Choice Logit Model without no-purchase alternatives and private labels

	LPM				DC Logit
Incidents	all	all	all	purchase	all
Variables	Coef./SE	Coef./SE	Coef./SE	Coef./SE	Coef./SE
preis	−.002***	−.005***	−.003***	−.002	−.038
	(.00)	(.00)	(.00)	(.00)	(.03)
inh	.001	.002*	.002	.002	.041
	(.00)	(.00)	(.00)	(.00)	(.04)
feature	.070***	.065***	.057***	.009	.163
	(.00)	(.00)	(.00)	(.01)	(.18)
display	.015***	.014***	.006*	−.022	−.063
	(.00)	(.00)	(.00)	(.01)	(.10)
prceflag	.006*	.005	.003	−.036***	.056
	(.00)	(.00)	(.00)	(.01)	(.10)
handbill	−.005	−.006	.001	.020	.085
	(.00)	(.00)	(.00)	(.01)	(.12)
liquid	−.004	−.005	−.001	−.020	−.049
	(.00)	(.00)	(.00)	(.01)	(.09)
konzentrat	−.009***	−.011***	−.008**	−.038**	.041
	(.00)	(.00)	(.00)	(.01)	(.10)
color	.003	−.004	−.003	−.016	−.027
	(.00)	(.00)	(.00)	(.01)	(.08)
sensitiv	.022***	.009	.003	.004	.053
	(.01)	(.01)	(.01)	(.02)	(.20)
gimmick	.005	.002	.010*	.016	.068
	(.00)	(.00)	(.00)	(.02)	(.14)
extrasize	.008	−.002	−.001	−.023	−.130
	(.00)	(.01)	(.00)	(.02)	(.16)
bigpack	−.005	−.007	−.002	.011	.034
	(.01)	(.01)	(.01)	(.02)	(.23)
countc56ad	−.000	−.001**	−.000	−.000	−.004
	(.00)	(.00)	(.00)	(.00)	(.01)
GLdumB			.782***	.244***	1.680***
			(.00)	(.01)	(.09)
duration			−.001***	−.000	
			(.00)	(.00)	
constant	.105***	.160***	.136***	.744***	
	(.00)	(.01)	(.01)	(.03)	
dummies					
brand	No	Yes***	Yes***	Yes***	Yes***
No. of obs	107,191	107,191	90,243	13,563	5,265
R2	.00	.01	.28	.06	

Note: Asterisks indicate significance levels at $\alpha : * = 0.05, ** = 0.01, *** = 0.001$. Variable `countc56ad` is total number of advertisement contacts within a three month period per market.

Appendix B: Value Labels of Coded Variables

In the following tables the value labels of numerically coded variables are presented with the numeric code, a description and translation if necessary. Note that the code 999 refers to the no-purchase alternative/option (=NOP), i.e. the consumer did not shop in the detergent category. This code appears when looking at the value labels of product characteristic variables.

Table 2.17. Overview of Value Label for variables `hhfpsex, psex1, ..., psex8` and `stand`: gender & famstatus

gender		famstatus	
Value	Description	Value	Description
0	male	0	No information
1	female	1	Unmarried
		2	Married
		3	Divorced
		4	Widowed

Table 2.18. Overview of Value Label for variable `bstel`: occupation

Value	Description	Combined to
1	Large self-employed (50+ employees)	self employed
2	Medium self-employed (10-49 employees)	self employed
3	Small self-employed (up to 10 employees)	self employed
4	Free profession / freelancer	self employed
5	Self-employed farmer	self employed
6	Executive employee	white collar high
7	Qualified employee	white collar high
9	Public official in the higher grade	white collar high
10	Public official in the upper grade	white collar high
11	Public official in the medium grade	white collar low
12	Public official in the lower grade	white collar low
13	High skilled laborer	blue collar
14	Skilled laborer	blue collar
15	Other laborer / unskilled laborer	blue collar
16	Medium level (managerial) employee	white collar low
17	Low level (simple) employee	white collar low
18	Never have worked	unemployed

Notes: The third column indicates the grouping that is done to simplify usage of the occupational status. The corresponding variables are: `bblue, bself, bwhite_h, bwhite_l and bunemp`.

Table 2.19. Overview of Value Label for variable `schulab`: degree

Value	Description
1	Unknown
2	Secondary (lower level) School visited or completed (HAUPTSCHULE und kein Abschluss))
3	Secondary School of GDR (until 1989) (POLYTECHN. OBERSCHULE)
4	Secondary School (REALSCHULE)
5	Restricted A-Levels (FACHHOCHSCHULREIFE)
6	A-Levels, high school graduate (ALLG HOCHSCHULREIFE)

Table 2.20. Overview of Value Label for variables `ng` and `bik`: nielsenarea & community

nielsenarea		community	
Value	Description	Value	Description
10	Hamburg, Bremen, Schleswig-Holstein, Niedersachsen	1	<2K
20	Nordrhein-Westfalen	2	2-5K
31	Hessen, Rheinland-Pfalz, Saarland	3	5-20K
32	Baden-Württemberg	4	20-50K
40	Bayern	5	50-100K
51	Westberlin	6	100-500K
52	Ostberlin	7	500K+
60	Mecklenburg-Vorpommern, Brandenburg, Sachsen-Anhalt		
70	Thüringen, Sachsen		

Table 2.21. Overview of Value Label for variables `duft`, `kons` and `uwg`: duftcode, konscode & uwgcode

duftcode	
Value	Description
1	Aktivfrisch
2	Aloe Vera
3	Alpine
4	Alpine Fresh
5	Antibakteriell
6	Apfel
7	Apfelfrische
8	Bluetenzauber
9	Citrus
10	Cotton Fields
11	Fresh Magic
12	Frische
13	Fruehlingsfrisch
14	Green Lemon
15	Kuschelduft
16	Lavendel
17	Limonenfrische
18	Mango Fresh
19	Meeresfrische
20	Neutral
21	O Angabe
22	Orange Fresh
23	Pfirsich & Limone
24	Pfirsichtraum
25	Pink Grapefruit
26	Pure Frische
27	Rosenfrische
28	Sommerfrische
29	Spring Fresh
30	Sunny Peach
31	Thai Paradies
32	Tropic
33	Wiesenfrische
34	Wilde Orchidee
35	Wildrose
36	Winterfrisch
37	Zitronenfrisch
999	NOP

konscode		
Value	Description	Description (translated)
1	Creme	Creme
2	Fluessig	Liquid
3	Gel	Gel
4	Nuesse	Nuts
5	Paste	Paste
6	Pulver	Powder
7	Schaum	Foam
8	Tabletten	Tabletts
9	Waschkissen	Cushions
999	NOP	NOP

uwgcode		
Value	Description	Description (translated)
1	Feinwaschmittel	Silk Detergent
2	Gardinenwaschmittel	Curtain Detergent
3	Kaltwaschmittel	Low Temperature Detergent
4	Vollwaschmittel	Normal Detergent
5	Wollwaschmittel	Wool Detergent
999	NOP	NOP

Notes: NOP indicates the value that the no-purchase alternative takes for the characteristic. As NOP is the choice of the outside good, it naturally has no detergent characteristics.

Chapter 3

Estimating the Magnitude of Causes for Consumer Price Endogeneity

3.1 Introduction

Understanding the reaction of consumers to price changes is at the heart of market demand analysis. Price is the classical economic variable signaling the ultimate value of any economic object. Knowledge of consumer reaction is a primitive for many more elaborate procedures to analyze markets. Therefore, if measuring the consumer reaction following price changes fails, the whole procedure may be flawed.

A prominent example is the endogeneity of prices in a differentiated products market. Product characteristics can explain the existence of many varieties in these markets, following arguments of vertical differentiation. Commonly, at least one product characteristic exists that is relevant for consumer choice but unobserved to the econometrician. Product quality is the favorite example in the literature. If product price and its characteristics have a systematic relationship, then estimating price reactions without accounting for the unobserved characteristics will bias estimates. In the empirical industrial organization literature Berry (1994) and Berry, Levinsohn, and Pakes (1995) developed a technique to deal with this scenario in a market setting. This endogeneity correction method has been introduced and applied to the marketing literature by Chintagunta, Dubé, and Goh (2005) who focus on individual consumers and naturally use individual level data. Petrin and Train (2006) use an alternative approach that works for any aggregation level of data, i.e. either market or individual level. Note that all papers treat product price as the only endogenous variable.[1] Up to this point the

[1] Theoretically all of them are based on McFadden's (1974) probabilistic discrete choice model, but they result methodologically in different estimation techniques and importantly in different economic interpretations of endogeneity for prices. In the literature several explanations are common to motivate existence and relevance of the endogeneity problem of prices. The "Berry" techniques lead to statistically

papers were more concerned with the technical part of correcting the endogeneity, but did not combine all potentially relevant causes for price endogeneity in one approach. This orientation bears the risk that one might miss a potential endogeneity cause in a specific application and misquantify price effects. This clarifies exactly the direction and need of this work, where I want to study several endogeneity causes simultaneously.

I restrict myself to the setting of individual level data, and will not discuss differences to non-individual data settings. For the case of individual level data, I focus on the following five major causes that may lead to endogenous prices: (i) "variety characteristics", i.e. time constant unobserved product characteristics for varieties within brands (ii) "retail activity" or local demand shocks, i.e. time varying unobserved product characteristics at the retail outlet (iii) TV advertising per brand at the household level (iv) individual household level inventories (v) state dependence, motivated either by habit formation or taste. Although not relevant for price endogeneity, this chapter allows for (vi) consumer heterogeneity and (vii) "disaggregated daily data", usually data are aggregated to weekly observations.[2] The papers of Chintagunta, Dubé, and Goh (2005) and Petrin and Train (2006) take care of points (i), (ii), (vi). A brand or product fixed effect in form of dummy variables captures possible endogeneity from cause (i). The endogeneity corrections in each of the two papers and the standard variables that measure observable retail activity should control cause (ii). Point (vi) is modeled by using a standard random coefficients model and point (vii) by using weekly data. Problems (iii)-(v) are not being addressed explicitly. Summing up, in previous work endogeneity problems (i)-(v) are not treated simultaneously.

In this work I want to study how additional information may solve problems (i)-(v) for individual level data while allowing for above mentioned points (vi) and (vii). Especially I want to find out what the leading causes are and whether the price endogeneity correction still is important once the other causes have been addressed. This can only be done with a very detailed dataset on consumer purchases at hand. Besides this empirical part, I discuss each cause in detail and show the relation to the endogeneity correction. In particular, I discuss which causes the endogeneity correction is not able to capture.

The novel dataset from A.C. Nielsen Germany comprises daily detergent purchases and TV viewing history of nationwide sampled German households across different retail chains. I have precise variety information for the products sold. A wide array of demographic variables is available for all households in the sample. The advertising data are available for a fraction of the national sample and are detailed down to the individual viewing contact of the household with a specific advertisement/spot. The motive of the

and economically significant different results. This usually means that the absolute value of estimates for price coefficients and elasticities are larger with the correction than without it.

[2] Causes that may be missing are national promotional activities and advertising from print, radio or internet. As I study detergents, I know that these are not relevant for detergents (and similar "simple" consumer goods), but may be important for other complex products such as cars or computers. See footnote 3 for evidence.

spot is known so that besides brand other spot characteristics are available, such as length or theme of the spot.

I control for each cause in the following way: (i) variety information is available and is included in the estimation to help in explaining the choice of different products. For example, consumers may care about the “color” characteristic, i.e. the suitability of the detergent to clean nonwhite clothing. Brand dummies are also included. (ii) retail activity is controlled for by feature and display variables. Different from previous work I have a nationwide sample, local effects cannot influence the whole sample of consumers as in previous work. Remaining retail activity and local demand shocks are captured by the endogeneity correction of Petrin and Train (2006) (iii) the effect of TV advertising is controlled for by using the individual household TV viewing history of advertisements. The detergent category under study is frequently advertised on TV, whereas print, radio and online advertising are only a minimal part of advertising expenditure.[3] (iv) duration since last detergent purchase is used to control for individual household inventories. (v) state dependence dummies are included as frequently done in the marketing literature (that does not address price endogeneity) (vi) estimation of a random coefficients mixed logit model allows unobserved heterogeneity in preferences and demographic information is used to allow observable heterogeneity. (vii) data from actual purchases are not aggregated to weekly observations as in previous approaches, but are kept on a daily basis in the data. This allows intra-week variation in product choice, something I find in my dataset. Multiple purchases per week occur frequently and are allowed because I use daily data. There are rarely multiple purchases per day.

Conducting all of these steps with the dataset at hand and comparing to the price endogeneity correction of Petrin and Train (2006), I evaluate the overall impact of the endogeneity correction and the endogeneity causes on the price coefficients and especially on price elasticities. The effects are studied for two different settings: in the first, consumer prices enter nominally, as visible for each consumer on the price label. In the second, consumer prices enter the problem as efficiency prices, i.e. as nominal price divided by product size. These are depicted on the price tag in the supermarket to comply with legal requirements.[4] It is common to display them in a smaller size than the nominal price. A priori it is not clear what consumers are looking at.

Methodologically, to compare the impact of different endogeneity causes, I use price elasticities calculated for the estimation sample. Sections 3.2.5 and 3.2.3.2 highlight why this is a natural approach in this setting.

[3] According to Nielsen Media Research advertising expenditure for TV in the detergent category in 2006 make nearly up to 100% of total advertising expenditure. See for example the publication of SevenOneMedia in 2007 on p. 30. Viewed on January 18th, 2010: `http://www.sevenonemedia.de/imperia/md/content/content/Research/Downloads/branchenreport/branchenreport_2007.pdf`

[4] Since 2000 the “Grundpreis” is mandatory according to German law and is the product price divided by quantity, usually measured in kilogram or liters. It is defined in the “Preisangabenverordnung” (PAngV) and is in effect since September 1^{st} 2000.

The chapter proceeds as follows. After a literature review, in section 3.2, the empirical model is specified, the relevant causes for price endogeneity are laid out and instruments are motivated. Section 3.3 describes the data, especially the variables used and the instrument construction. In section 3.4, the results are presented. The final section 3.5 concludes.

3.1.1 Related Literature

This work relates methodologically to two papers by Chintagunta, Dubé, and Goh (2005) and Petrin and Train (2006) who conduct price endogeneity corrections for individual level data, but both do not have national samples and advertising information as I have. The approach of Chintagunta, Dubé, and Goh (2005) follows directly the fixed effects approach of Berry (1994), whereas Petrin and Train (2006) use an alternative control function approach. Petrin and Train (2006) show that both approaches yield qualitatively identical results, yet the control function approach being easier to implement than the fixed effects approach. This work uses the control function approach. Villas-Boas and Winer (1999) use a different approach based on Rivers and Vuong (1988) to account for local demand shocks that may contaminate price effect estimates. They were the first to look at endogeneity in prices on an individual level. The papers mentioned so far ignore detailed product varieties (apart from price and size) and concentrate on brand choice, i.e. they condition the sample on the most sold subset of a product category, e.g. 4 oz. ketchup glass bottles, whereas I estimate product choice looking at the whole product category.[5] My approach is in line with Guadagni and Little (1983) who write on p. 204: “To understand such issues [. . . how price, promotion and other marketing variables affect the sales . . .] we need to model whole product categories.”

More importantly, no one has attempted to look at several endogeneity causes and the technical endogeneity correction simultaneously, something that is at the heart of this empirical work. Finally, this relates to the original work of Berry (1994) and Berry, Levinsohn, and Pakes (1995) who introduced the possibility of treating endogenous prices in market demand for differentiated products, but for the individual level demand these papers do not lay out several endogeneity causes and assess their relevance.[6]

I am of course not the first to look at the role of TV advertising for the purchase decision in a consumer goods category. But so far the focus has not been on its relation to the

[5] In addition, most papers follow common practice to concentrate only on major brands, thus ignore parts of the market. Hence, products are missing that may share some of the same variety characteristics with the products that remain in the sample. With this non-random omission, another source of bias may be present, as the choice set the modeled consumer faces is definitely incomplete and biased towards market share leaders, i.e. the consumer, but not the econometrician, sees the complete characteristics space with all products.

[6] To see the relation to the simultaneity bias in the classical homogeneous goods case, consult Berry, Levinsohn, and Pakes (1995), p. 842 bottom and p. 850, 851.

price endogeneity problem for individual level data, although advertising is suggested as endogeneity cause in most introductions of the above mentioned papers. Due to common unavailability of advertising data, this suggestion has not been investigated further. The marketing literature has been active on the empirical side to assess the impact of advertising on prices and price sensitivity. See the paper of Kaul and Wittink (1995) for a survey of the results in this literature.

The data used are novel in the sense that this combination of information has not been used before. Especially the level of disaggregation combining product and TV advertising information is a novelty in this literature. A similar dataset has been used by Ackerberg (2001, 2003). His work is not concerned with price endogeneity but with the prestigious and informational aspects of advertising. Erdem and Keane (1996) estimate models similar to my work, but have to use a dataset where the advertising information is known to be of a lower quality than in this work.[7] Their focus is not on assessing the impact of price endogeneity corrections. Shum (2004) studies the interplay of advertising and habits, but uses only national advertising data. Keane (1997) does a very detailed analysis on the interplay of consumer taste heterogeneity and state dependence using data without advertising. Dubé, Hitsch, and Rossi (2009) take a closer look at state dependence and try to find out whether there are other explanations for consumer inertia, the tendency to stick to a favorite product. Horsky, Misra, and Nelson (2006) show the benefit of using additional stated-preference information for modeling consumer heterogeneity. In this chapter the inclusion of advertising leads to an effect that is in line with their work. This inclusion reduces the need for the model to capture the consumer heterogeneity randomly. Advertising levels vary per consumer, per product and over time, thus introducing more observed heterogeneity that needs not to be accommodated by the parametric distribution of the random coefficients. This helps the model to fit the data without leading to overdispersion in the random coefficients, i.e. inflated standard deviations of the coefficients. This effect can be achieved as well by having additional stated-preference information on the consumer level.

This work zooms in on detergents, the product category that has also been studied in a recent paper by Hendel and Nevo (2006). As detergents are storable goods, the authors can quantify the influence of individual inventory holdings on individual demand. They find large corrections of their procedure for price elasticities, i.e. state that usual models overestimate these elasticities, whereas usual endogeneity corrections of the above mentioned papers make the opposite statement. The net effect is unknown yet. Different from their dynamic optimization multi-stage approach, I use a reduced form discrete choice specification and try to control directly for inventory holdings that may be an endogeneity cause.

[7] The data coming from the ERIM database from the Kilts Center for Marketing at Chicago Booth School of Business are a dataset offered from A.C. Nielsen in the US, where the telemeter technique (to measure advertising exposure of a household) was firstly introduced and thereby missing reliability manifests itsself in inconsistencies, missing data and reporting errors.

There are recent papers that are related by topic or method, but do not have their focus on price endogeneity. Ching, Erdem, and Keane (2009) have a discrete choice model that explicitly models the observation that consumers do not consider all prices of all products when doing their consumption decision. Leaving the traditional discrete choice world, the papers of Berry and Pakes (2007) and Bajari and Benkard (2005) use individual consumer models that have no random error term, opening a new branch of literature termed “pure” characteristics models.

Finally, the conclusions drawn in this work are for a specific product category, but may extend from the detergent category to other categories. Several papers in marketing and economics have studied this category, because historically detergents were one of four product categories of the ERIM database.[8] A mere conjecture is that this will most likely be non-food fast moving consumer goods. Some of these categories are heavily advertised as detergents, their structure is simple and they are products of daily life, where the consumption decision is not a complex procedure. Nonetheless it is important to understand the underlying mechanism as these products make up a notable fraction of household consumption.

3.2 Model

To start I want to highlight the level of detail in my data using a simple notation, then I present the model. The price a consumer i faces (precisely speaking an unidentified representative of a household i that goes for the shopping trip) at the purchase time t is:

$$p_{v_b lst} = p_{jot} \tag{3.1}$$

where $b = 1, \dots B$ denotes brand, v_b denotes variety v of brand b and $v_b \in V_b$, which is the set of all varieties of brand b. Introduce an index j of all available products down to the variety level where $j = 1, \dots J$ and $j \in (V_1, V_2, \dots, V_B)$. l denotes geographic location, which is coded as postal/zip or county code (i.e. a German “Landkreis”), and s denotes the store/chain. Introduce an index o of all available store/location combinations from $(l \times s)$. This boils down the left hand side of (3.1) to the simpler notation on the right hand side, bearing in mind the complexity of the data involved therein.

All models estimated are random coefficient logit models. Thus, the degree of generality is bounded by this model class. Details of the precise specification will be presented alongside the model.

[8] The database is publicly available at the website of the Kilts Center for Marketing at University of Chicago Booth School of Business. For details see footnote 7.

3.2.1 The Empirical Model of Choice

In this section I am agnostic about possible endogeneity of any variable and think of the model as completely specified. The model is a mixed logit model and thereby permits a random coefficients interpretation. Note that the model permits also an equivalent error correction formulation, see chapter 6 of Train (2003). It is a model of product choice, where a product is defined down to a specific variety, not a model of brand choice as Guadagni and Little (1983), who introduced a logit model without random coefficients used extensively in marketing applications. It is a simplification of the mixed logit model used in this chapter if the i subscript on the parameters is removed and if all varieties of a brand are aggregated to the brand level or if the product category is conditioned on product characteristics such that only a brand choice decision for a subset of all available products is left. For example, if I condition on a subset of the product category, then doing this amounts to index j simply representing brand b. A common running example is the focus on unseasoned tomato ketchup glass bottles of a certain size, which is a subset of the ketchup category.

Each consumer i at time t derives utility from a choice among one product $j = 1, \ldots, J$ or a no-purchase option $J+1$ in location o, thus conditional on store visit, according to the following latent utility form based on McFadden (1974):

$$
\begin{aligned}
U_{ijt} &= \beta^1_{ij} + X_{ijot}\beta_i + p_{jot}\alpha_i + a_{ijt}\gamma_i + \epsilon_{ijt} && (3.2)\\
X_{ijot}\beta_i &= X^A_{jot}\beta^2_i + X^S_{ijt}\beta^3_i + X^V_{jot}\beta^4_i + X^I_{it}\beta^5_i && (3.3)\\
U_{i(J+1)t} &= \epsilon_{i(J+1)t} && (3.4)
\end{aligned}
$$

where the row vector X^A contains marketing/retail variables, row vector X^S contains state dependence/habit formation variables, row vector X^V contains product characteristics, the "variety information", scalar X^I contains duration since last detergent purchase, scalar p is price, row vector a is TV advertising and ϵ is an i.i.d. extreme value error. For ease of notation, let $X_{ijot} = (X^A_{jot}, X^S_{ijt}, X^V_{jot}, X^I_{it})$.[9] The scalar β^1_{ij}, the column vectors $\beta^l_i, l = 2, 3, 4, 5$, their stacked version $\beta_i = (\beta^{2'}_i, \beta^{3'}_i, \beta^{4'}_i, \beta^{5'}_i)'$ and especially scalar α_i and column vector γ_i are the parameters of interest. Commonly, the β^1_{ij} are termed as brand dummies. I have one price coefficient, as I do not assume promotional price reductions to be different than long term price changes.[10]

State dependence X^S is constructed using observations from previous time periods, yielding a dynamic model. Dynamic is to be understood in the sense of a reduced form

[9] Note that the o in the subscript of X_{ijot} is not strictly necessary, because there is only one o possible for each pair of i and t if we remember the fact that one consumer i can only be in one location o at a time t.

[10] See the paper by Briesch, Chintagunta, and Matzkin (2002) that uses a non-separable nonparametric function for price and promotional price changes to analyze the consumer reaction. A possibility to capture this effect partially is to add a variable indicating extreme price changes, for example a dummy if prices cross a certain threshold.

approach, as there is no unique behavioral explanation to rationalize this construction. Keane (1997) details how to motivate dynamics by habit formation, inertia or tastes. If the dynamics can be entirely captured by observables (also lagged dependent variables) and the standard i.i.d. assumption (across i, j, t) on the extreme value error ϵ holds, the model can be estimated just as a standard mixed logit. This is a feature of the mixed logit model compared to the discrete choice probit, see Train (2003), p. 150 middle. I model state dependence as general variant of the exponentially smoothed weighted average of past purchases suggested by Guadagni and Little (1983), which are also used in Keane (1997). For the exponentially smoothed case the vector X^S is a scalar and defined by:

$$\begin{aligned} d_{ijt} &= \mathbf{1}\{\text{consumer } i \text{ purchases product } j \text{ at time } t\} && (3.5)\\ d^S_{ijt} &= \mathbf{1}\{\exists b, j^\star : (j \in V_b : d_{ijt} = 1 \wedge j^\star \in V_b : d_{ij^\star t-1} = 1)\} && (3.6)\\ X^S_{ijt} &= \phi X^S_{ijt-1} + (1-\phi) d^S_{ijt} && (3.7)\end{aligned}$$

Equation (3.5) simply defines a product purchase dummy. Equation (3.6) constructs a dummy that is one if the product $j^\star$ bought in the last period has the same brand as the product j. Thus, the consumer gets a potential benefit from sticking to the same brand. Equation (3.7) states that this value exponentially decays for the following periods. Setting $\phi = 0$ gives first-order dependence.[11] This setting corresponds to Chintagunta, Dubé, and Goh (2005). Petrin and Train (2006) do not allow for state dependence in their application. As there is no a priori reason to be restrictive I use a variant that gives a K-dimensional row vector X^S and its k-th component is defined by:

$$(X^S_{ijt})_{k=(1,\ldots,K)} = d^S_{ijt-(k-1)} \qquad (3.8)$$

The dummy d^S is defined as in (3.6). Thus, the vector X^S is a vector of brand purchase dummies of the last K purchase occasions, and not of product purchase dummies. Moreover, the pattern of lagged dummies is not restricted by any assumption, e.g. the exponential form of (3.7). Thereby this formulation allows flexible K-th order dependence.[12]

Advertising variables a are constructed similarly to the state dependence variable:

$$\begin{aligned} d^a_{ibt} &= h \times \mathbf{1}\{\text{consumer } i \text{ sees } h \text{ advertisements of } b \text{ at time } t\} && (3.9)\\ d^a_{ijt} &= h \times \mathbf{1}\{\exists b : (j \in V_b \wedge d^a_{ibt} \geq 0\} && (3.10)\end{aligned}$$

[11] Guadagni and Little (1983) suggest two options for setting ϕ from (3.7): calibration or estimation. From calibration Guadagni and Little (1983) get $\phi = 0.875$. Keane (1997) includes ϕ as parameter in his joint estimation procedure. He initializes $X^S_{ijt} = 0 \, \forall i, j$ and $t = 0$. I also estimate ϕ.

[12] In addition to product brands, Guadagni and Little (1983) add state dependence for product sizes. In their ketchup setting there were few sizes, whereas the detergent category is more rich in size variation, making this unfeasible for this work.

where h is the number of advertisement contacts. First, I define two variables that indicate how often the household had contact to an advertisement of a specific brand (i.e. (3.9)) and assume these contacts to have an effect on all varieties of a brand (i.e. (3.10)). Building on these variables, I construct measures for absolute advertising pressure in a given time span before a detergent purchase:

$$a^A_{ijt}(Q) = \sum_{q=0}^{Q} d^a_{ijt-q} \tag{3.11}$$

$$a^A_{ibt}(Q) = \sum_{q=0}^{Q} d^a_{ibt-q} \tag{3.12}$$

This is done by cumulation of the variables from (3.9) and (3.10) for the last Q days preceding a purchase occasion, where the first sum is per product and the second sum is per brand. This is constant for all varieties within the brand so that $a^A_{ijt}(Q) = a^A_{ibt}(Q)$ if $j \in V_b \quad \forall Q, i, t$. As far as households are concerned the above quantities do not take into account the relative frequency of a specific advertisement to all its competitor advertisements. That is why I construct relative advertising pressure, by adjusting the absolute level per product by the total amount of advertisements seen in a given time span:

$$a^R_{ijt}(Q) = \frac{a^A_{ijt}(Q)}{\sum\limits_{b=1}^{B} a^A_{ibt}(Q)} \tag{3.13}$$

Obviously the intuition is that not only the frequency of advertising contact matters, but the frequency relative to all advertising contacts experienced in a given time window.

The random coefficients of the model parameters in the utility function (3.2) are defined by:

$$\begin{pmatrix} \beta^1_{i1} \\ \vdots \\ \beta^1_{iJ} \\ \beta^2_i \\ \vdots \\ \beta^5_i \\ \alpha_i \\ \gamma_i \end{pmatrix} = \begin{pmatrix} \beta^1_1 \\ \vdots \\ \beta^1_J \\ \beta^2 \\ \vdots \\ \beta^5 \\ \alpha \\ \gamma \end{pmatrix} + \begin{pmatrix} \kappa'_1 \\ \vdots \\ \kappa'_m \end{pmatrix} \left(X^D_i\right) + \Sigma^{\frac{1}{2}} \eta_i \tag{3.14}$$

where all parameters are stacked into one column vector of dimension m. The first right hand side summand contains constants which are the mean value of the parameter, i.e. the mean effect of the variable. In the second right hand side summand observed heterogeneity of the parameters is modeled. X_i^D is a column vector of demographic variables of dimension d relevant for the model. $\kappa_k, k = 1, \ldots, m$ is a column vector of dimension d denoting which demographic variables are relevant for each parameter. Thereby the demographic variables that enter each random coefficient can vary. The choice of demographics depends on the application. As this part is observable its significance be tested by standard test procedures. In section 3.3, I explain the choice for this chapter. The first two summands represent the deterministic part of the random coefficients for observed heterogeneity in the form of component linear indices per random coefficient. I denote the second right hand side summand of (3.14) as κX_i^D to simplify notation. The third right hand side term contains a random normal vector $\eta_i \sim N(0, I_m)$ and $\Sigma^{\frac{1}{2}}$ is a Cholesky decomposition of a simple covariance matrix $\Sigma = diag(\sigma_1^2, \ldots, \sigma_m^2)$. The diagonal elements of the Cholesky decomposition contain the standard deviation and can be interpreted as unobserved heterogeneity in the parameters. This summarizes the form of the random coefficients as sum of a constant, a linear index on demographics and a random normal variable.

I vary the specifications. The simplest variant has $\kappa = 0$ and only the price parameter α_i has a random normal distribution. Although there are more convincing parametric forms such as the lognormal that do not allow consumers to experience the good as Giffen good I stick to the normal.[13] The most complex variant will exhibit demographic variables for the coefficients of price α_i and inventory β_i^5 and a random normal η_i with a diagonal covariance matrix Σ.[14]

Now I illustrate the choice probabilities. All parameters to be estimated are denoted by $\theta = (\beta_1^1, \ldots, \beta_J^1, \beta^2, \beta^3, \beta^4, \beta^5, \alpha, \gamma, \kappa, \Sigma)$. 1_j is a J-dimensional row vector with value one at column j and all other entries zero. The choice probabilities for consumer i picking product j at time t in the mixed logit model take the usual logit form conditional on realizations of η_i that determine the stochastic part of the random coefficients:

$$\begin{aligned} P_{ijt|\eta_i}(\theta) &= \frac{exp(\varpi_{ijt} + \nu_{ijt}^D + \nu_{ijt}^S)}{1 + \sum\limits_{j=1}^{J} exp(\varpi_{ijt} + \nu_{ijt}^D + \nu_{ijt}^S)} && (3.15) \\ \varpi_{ijt} &= \beta_j^1 + X_{jot}^A\beta^2 + X_{ijt}^S\beta^3 + X_{jot}^V\beta^4 + X_{it}^I\beta^5 + p_{jot}\alpha + a_{ijt}\gamma && (3.16) \\ \nu_{ijt}^D &= [1_j, X_{jot}^A, X_{ijt}^S, X_{jot}^V, X_{it}^I, p_{jot}, a_{ijt}]\kappa X_i^D && (3.17) \\ \nu_{ijt}^S &= [1_j, X_{jot}^A, X_{ijt}^S, X_{jot}^V, X_{it}^I, p_{jot}, a_{ijt}]\Sigma^{\frac{1}{2}}\eta_i && (3.18) \end{aligned}$$

[13] As the normal distribution has full support on the real line, there are always realizations in the positive domain for the price coefficient. However, Chintagunta, Dubé, and Goh (2005) or Petrin and Train (2006) used this parametric form as well.

[14] The author estimated models with non-diagonal covariance matrix. The result was a sharp increase in estimation time, but no other results arose.

The standard logit expression is given by (3.15), see chapter 6 of Train's (2003) book. Equation (3.16) collects all expressions that contain mean parameters, i.e. without subscript i, that would appear in a standard logit without consumer heterogeneity. The deterministic part and the stochastic part of consumer heterogeneity are summarized in equations (3.17) and (3.18), respectively.

Different to models estimated frequently, in this model important variables exhibit more than usual variation across the panel dimensions. In the typical literature, the data come from one or from narrow markets so that marketing variables do not vary anymore across i, as all individuals are in the same location o, TV advertising is unobserved and "variety information" is omitted which implies the following restrictions for typical models relative to the one formulated here:

$$\begin{aligned} X^A_{jot} &= X^A_{jt}, \forall o && (3.19)\\ p_{jot} &= p_{jt}, \forall o && (3.20)\\ a_{ijt} &= 0, \forall i && (3.21)\\ X^V_{jot} &= 0, \forall i && (3.22)\\ X^I_{it} &= 0, \forall i && (3.23) \end{aligned}$$

The first two simplifications (3.19) and (3.20) do not matter for models without endogeneity correction, but having them makes the construction of Hausman instruments infeasible and thereby complicates use of an endogeneity correction, see section 3.2.4.

Now, as the individual choice probabilities are given, I can specify the probability that a consumer did a specific sequence of product choices conditional on his individual draw of η_i:

$$L^S_{i|\eta_i}(\theta) = \prod_{t=1}^{T_i} \prod_{j=1}^{J+1} P^{d_{ijt}}_{ijt|\eta_i}(\theta) \qquad (3.24)$$

where d_{ijt} is the dummy from (3.5). As the sample durations per household differ, T_i is usually very different across households. Moreover, I can specify the unconditional likelihood function for a purchase sequence S of a household:

$$L^S_i(\theta) = \int L^S_{i|\eta_i}(\theta) f(\eta_i) d\eta_i \qquad (3.25)$$

where $f(\eta_i)$ is the standard normal density in my case, but choice of this mixing distribution is in principle not restricted.

Then the log likelihood function to be maximized given a sample of I consumers is:

$$LL(\theta) = \sum_{i=1}^{I} log[L^S_i(\theta)] \qquad (3.26)$$

Estimation of this expression is conducted by Simulated Maximum Likelihood, where the integration in (3.25) is approximated by simulation methods. Consult chapter 10 of Train (2003) for details. Conditional on parameters and a draw of η_i, the inner (logit) part of the integral in (3.25) for a household is calculated. This is repeated many times and the results are averaged to approximate the integral.

The presented model is the standard panel random coefficients model, where η_i, and thereby all parameters it affects according to (3.14) (recall α, γ and all βs) are not allowed to vary across each purchase observation of a consumer so that only one draw of η_i is used for a consumer's choice sequence defined in (3.24).[15]

3.2.2 Endogeneity Correction Model

In this section I introduce the endogeneity problem into the empirical model. The proposed correction follows closely the control function approach of Petrin and Train (2006). Assume there exists an unobserved variable ξ_{ijt} that is the sole cause of the price endogeneity. $g()$ is some monotone and unknown function. In the next section, I will introduce the major motivation for the existence of ξ_{ijt}, namely unobserved retail activity. Recall (3.2), in which now an unobservable is visible as ξ_{ijt}:

$$\begin{aligned} U_{ijt} &= \beta^1_{ij} + X_{ijot}\beta_i + p_{jot}\alpha_i + a_{ijt}\gamma_i + g(\xi_{ijt}) + \epsilon_{ijt} && (3.27)\\ X_{ijot}\beta_i &= X^A_{jot}\beta^2_i + X^S_{ijt}\beta^3_i + X^V_{jot}\beta^4 + X^I_{it}\beta^5_i && (3.28)\\ U_{i(J+1)t} &= \epsilon_{i(J+1)t} && (3.29) \end{aligned}$$

Petrin and Train (in the following PT) set $a_{ijt} = X^S_{ijt} = X^V_{jot} = X^I_{it} = 0$ in their application. If ξ_{ijt} is unobserved, correlated with prices and the econometrician does not control for its presence, the standard procedure leads to inconsistent results. PT suggest the following procedure: A control function equation is defined to recover the control that is used as proxy for $g(\xi_{ijt})$. Using the control function residuals as these controls in the above utility specification then alleviates the endogeneity problem.

My data have another structure than PT so that I need to adjust their approach that is formulated for local data to my setting for nationwide data. There are two ways to define the additively separable control function of PT. I will illustrate both, but only the first is feasible and thereby implemented. The second is presented for illustrative purposes. For the control function specification an assumption on the relationship between endogenous prices and instruments/exogenous variables is necessary.[16] The first alter-

[15] In the estimation, I alternatively estimated the cross sectional version, where η_i is allowed to vary across each purchase occasion for a consumer and compared both versions, yielding no new results but demanding only much more computational resources. See appendix part A for the pooled case.

[16] This assumption is closely linked to the recent literature on hedonic models, see Bajari and Benkard (2005). The analogue of the control function is termed hedonic pricing function in that literature.

native to specify the control function equation uses variation in time, product variety and across consumers, leading to B regressions:

$$\begin{aligned} p_{jot} &= E[p_j|z_{jot}] + g_b(\xi_{jot}),\ b = 1, \ldots, B, b \text{ s.t. } j \in V_b & (3.30) \\ z_{jot} &= [1, p^{\star}_{jot}, w_{jot}] & (3.31) \end{aligned}$$

where g_b is a some unknown, monotone function for each brand b. z collects instrumental and exogenous variables $p^{\star}$ and w respectively. Note the notation of ξ used here: For two consumers i and k, $\xi_{ijt} = \xi_{kjt} = \xi_{jot}$ if both are in location o, thus only the location of the consumer but not his immanent properties enter into (3.30). This is an important property of ξ in this approach. Therefore, only endogeneity causes that stem from the location can be controlled through this approach.

In practice equation (3.30) is estimated as a pooled regression per brand b over i, j and t. See the section 3.2.4 for the precise instrumental and exogenous variables $p^{\star}$ and w used in this chapter. Different from PT, I use instruments that vary across locations o. The fitted values of the residual $\widehat{g_b(\xi_{jot})}$ are used as proxy in the estimation of (3.27) to alleviate the endogeneity of prices. As evident from the subscripts of the used control, the endogeneity correction cannot capture endogeneity causes that stem from the specific individual itsself as the information used for its estimation comes from the level of the product, location and time. This property is important for the next section, where I introduce the endogeneity causes.

The second alternative is to use solely the variation across time and brand variety, leading to $I \times B$ equations:

$$\begin{aligned} p_{jot} &= E[p_{ij}|z_{jot}] + g_b(\xi_{jot}),\ b = 1, \ldots, B,\ i = 1, \ldots, I, b \text{ s.t. } j \in V_b & (3.32) \\ z_{jot} &= [1, p^{\star}_{jot}, w_{jot}] & (3.33) \end{aligned}$$

Note that the main difference is the i subscript of p in the conditional expectation. The control function is specified per consumer and brand. This alternative leads to very "individual" controls because every consumer gets his own fitted parametric model. Several consumers may be at the same store at the same time, but have overall a different purchase history. This in turn implies that the sample on which equation (3.32) is estimated differs per consumer and delivers a per individual per product control. Although this approach is theoretically possible, in practice there are not enough observations, as each consumer is required to have a purchase history per product, something that is widely unavailable. The instruments of PT are wholesale prices $p^{\star}_{jot} = p^{\star}_{jt}$ that do not vary per location. If the instruments of PT are used, I would have the same instruments in all control function regressions. So there exists only variation across consumers because of different purchase histories.

3.2.3 Causes of Price Endogeneity

In this section I detail the empirical structure of the consumer choice problem. Given this structure, I discuss consumer and industry specific causes that may lead to endogeneity of prices in the individual consumer choice problem. Each discussion of a cause ends with an empirical hypothesis that is pursued later in the application.

The literature on price endogeneity in industrial organization has emphasized that usually the positive correlation between demand shocks/unobserved product attributes (typically "quality" is the running example) and prices leads to an attenuation bias of the price coefficient. In that case, without endogeneity correction the effect of price is understated. This line of argument is mainly motivated by considering a market setting, where the researcher solely has access to market level data, whereas I focus on the individual level due to access to the necessary data.

I do not expect that the endogeneity correction can capture endogeneity causes which stem from the individual itself. Why? Because information in the procedure comes from the product, time and location level, but not from the individual as has been emphasized in section 3.2.2.

3.2.3.1 Industry Structure

I assume there is a dichotomy between manufacturers and retailers. Thus, the price from (3.1) can be decomposed to:

$$p_{jot} = p^W_{jot} + p^R_{jot} \tag{3.34}$$

Manufacturers, i.e. producers of detergent, are assumed to set wholesale prices p^W and plan/run TV advertising campaigns. Retailers set markup p^R and have discretionary power to do in-store advertising, typically captured by the feature and display variables in scanner data.[17] Retail chains are assumed to operate independently of each other. Note that the exact functional form of the decomposition is not important, as I merely use it for expositional purposes.[18] The price components p^W and p^R are unobserved. Note that in Germany a price agreement between retailer and manufacturer is directly negotiated and unobservable to the econometrician. In the US, *Promodata Leemis Services* collects this price information. Chintagunta, Dubé, and Goh (2005) and Petrin and Train (2006) use wholesale price information as instrument for retail price p.

[17] This assumption may not hold for several reasons. Firstly, in the case of product introductions all marketing instruments are likely to be used in a concerted action, i.e. marketing mix. Secondly, in larger retail chains, manufacturers tightly interfere in the retailers' domain, in the sense that manufacturers compete for shelf space at the retailer.

[18] Of course, this does not allow to use information on the relative magnitudes of the wholesale and retail price component to assess the theoretical size of biases that originate in the different causes discussed throughout this section.

3.2.3.2 Cause 1 - Product Characteristics and Variety Information

Detergent is a differentiated product that is very homogeneous in its functioning. Any detergent product removes basic stains from clothes and "refreshes" clothing. Each product has a brand name and some specific characteristics, making up varieties within a brand. Products have different scent compositions (i.e. summer, april,...), come in different consistency (i.e. liquid, powder, tablet, gel), have different optimal purpose of use other than basic cleaning (i.e. for color, wool, silk or black clothing), have one-time special packing sizes (i.e. now +xx%, double pack) or come with a gimmick (i.e. a CD or another cleaning item). Generally, the price of a product does not only depend on own characteristics, but also on the characteristics of the nearest neighbor. As Nevo (2001) argues, the markup a firm can set depends on the position of the nearest neighbor which in turn depends on characteristics.

However, the standard assumption in the literature is: own product characteristics, but also characteristics of other products are exogenous or predetermined and therefore contemporaneously uncorrelated to prices. The relevance of this assumption is different for market versus individual level approaches. For the market level case, the assumption may be necessary to get estimates at all. In the iterated algorithm of Berry, Levinsohn, and Pakes's (1995) application this assumption delivers the moment condition for the GMM objective function because the characteristics of other products are used as instruments. In my individual level case it is necessary to identify and obtain parameter estimates (the α, βs and γ), but price sensitivities/elasticities can be calculated without the exogeneity assumption on characteristics.[19] If all product characteristics are included into the estimation, there should be no problem in the individual level case.

In common literature though, a crucial point arises that leads to price endogeneity: The focus is only on part of a product category, choices per consumer are aggregated to a week and all product variety information is lumped into one time constant brand coefficient.[20] Define the price of a variety of a brand as follows:

$$p_{jot} = p_{bot} + \nu_{v_b ot}, \qquad v \in V_b \tag{3.35}$$

where p_{bot} is the average price for brand b and ν is discount or a markup for variety v (short for v_b) within the brand b. Suppose now a consumer buys on two occasions the same brand, but of a different variety, e.g. "sensitive" and "color" and prices may differ. If all varieties within brands cost the same, $\nu \equiv 0 \, \forall v \in V_b$. If not, then the model should attribute any perceived benefit of the characteristic of the chosen variety over

[19] See Nevo (2001) for further explanation.

[20] Size and consistency of the detergent are identified as varieties, but usually the sample is conditioned on them and thereby the focus lies on a narrow product niche of the category. In this setup, the choice is one of brand and not of product, i.e. a variety of a brand. Moreover, the characteristics space is not complete, as some products are missing.

other available choices to the specific characteristic of the chosen variety. If this variety information is omitted, but pricing is correlated with variety information, then the price variable will pick up part of the variation that is due to the omitted characteristic.

To solve the described problem in this work, I add the available variety information, as almost each brand has a “liquid”, “concentrated”, “color”, et cetera version. For the argument, it does not matter with which component of equation (3.34) the markup of variety v, $\nu_{v_b ot}$, is correlated: it may correlated with p^W because production costs differ or with p^R because the retailer reshifts his range of detergents sold and thereby puts it on sale. The current literature deals implicitly with this argument by conditioning on available characteristics to get a homogeneous product such that only brand choice matters. This leads to a problematic setup: the model is only valid for a specific product type within a narrow product category. It is then legitimate to pose the question what the model purpose is after all.

Interestingly, under certain assumptions, data aggregation and the price endogeneity correction can solve the problem of omitted variety information. Why? The argument goes as follows. Consumers always only choose one variety of a brand per week. The price endogeneity correction removes all causes that may be correlated with prices on an individual level, also variety information. The instrumental variable approach breaks the correlation of the unobserved characteristic and price, if I use instruments detailed in a later section 3.2.4.[21] This has so far not been recognized in the literature.

What direction of bias is expected? Suppose there are two products with identical characteristics, but one characteristic is unobserved (to the econometrician) and prices of both products are up to unsystematical deviations equal. In that case the model can only accommodate choice differences through price differentials. If the researcher could add the unobserved characteristic and it is choice relevant (and need not be related to price), then the information on the characteristic and not the price differential would rationalize the choice in the model. Thereby, the explanatory power of the characteristic is higher, and this implies that the explanatory power of the price variable diminishes as it is by construction overstated in the example. Following this argument, I expect a decrease in price effects, once I add previously unused variety information.[22]

[21] Illustrating example: From equation (3.35), the price of product 1, namely, of brand 1, variety 1 in location 5 in time 4 is $p_{154} = p_{154} + \nu_{154}$ and of product 2, namely brand 1, variety 2 in location 6 and time 4 is $p_{264} = p_{164} + \nu_{264}$. Naturally, ν_{154} is correlated to the variety information of product 1. Consumers face both when doing their choice in store. If ν_{264} is uncorrelated to ν_{154} and thereby also uncorrelated to the omitted variety information, then another variety of the same brand in another location can serve as instrument. Relevance of the instrument is established by the correlation of the brand part of the prices. However, as the price scheme depends heavily on the product development of the manufacturer, it is unlikely that this simple procedure will work because there is correlation among markups ν and variety information across varieties within a brand.

[22] One might be tempted to follow the traditional Berry (1994) unobserved quality argument to determine the bias direction. It is not applicable here, because variety information is made of several variables that have no obvious a priori known correlation with price.

Summarizing, it seems obvious that the inclusion of brand dummies to capture only brand specific effects is not enough to capture variety information that is choice relevant.

Empirical Hypothesis 1 *Controlling for variety characteristics after controlling for brand decreases price effects.*

3.2.3.3 Cause 2 - Retail Activity

Unobserved retail or marketing activity in general is a further source of price endogeneity. An example for general activity may be TV advertising as in problem (iii). This is discussed in the next subsection. Commonly used retail activity variables may not capture the whole activity present at the moment of purchase.[23] For this chapter it is important to distinguish the location of a store and the type of a store.

In previous papers, the sample of consumers is restricted to a narrow geographical area. Consequently, a local unobserved "demand shock" or local unobserved retail activity is shared by virtually all consumers in the sample and can be the cause for price endogeneity.[24] There is another possibility for retail activity to influence consumers in this work: unobserved retail activity in a specific chain can have an effect across several locations on many consumers.

In this work I have a nationwide sample with different chains and use the dispersion in "unobserved demand shocks" across locations and chains for the construction of instrumental variables in section 3.2.4.

Retail activity may be manufacturer driven, retailer driven or a joint outcome of both. For big chains with large subsidiaries both players, manufacturers and retailers, are more likely to cooperate than in chains of small average store size. Additionally, there may be effects across locations within the same chain.

What are the sources of unobserved retail activity?

An obvious example are promotions. This kind of retail activity is typically manufacturer driven and takes place in larger subsidiaries. Promotions come with a certain pricing strategy, inducing a correlation between retail activity and prices, especially the wholesale component p^W. Promotions are not recorded in the data, but usually are quite rare in Germany and especially for the detergent category.

An important example is the competition between manufacturers for shelf space, especially at the large retailers, where this bargaining process is unobserved. Bargaining for shelf space can be accompanied by a pricing strategy so that this may induce correlation between retail activity and prices, especially the wholesale component p^W.

[23] See the data section 3.3 for more information on included retail variables.

[24] An example for a demand shock can be a special offer day with price rebates during sales time.

Thus, mostly for large subsidiaries I expect some potentially unobserved retail activity, whereas prices in small subsidiaries seem free of this problem.

Measurement error in the reporting of commonly used retail activity variables may lead to underreporting of retail activity. If this is the case, there may be situations with unobserved retail activity and a temporary price change, resulting in a correlation of unobserved retail activity with prices, especially in the retail component p^R. Summarizing, there are several retail activities that are either unobserved or imperfectly measured and may cause price endogeneity.

The direction of the bias is inferred by the same argument as for the quality example in the market data setting of Berry (1994).[25]

Empirical Hypothesis 2 *Controlling for unobserved retail activity increases price effects.*

3.2.3.4 Cause 3 - TV Advertising

Manufacturers may set (wholesale) prices and advertising jointly in a strategic fashion, while the econometrician does not account for TV advertisements on the consumer level. Hence, this causes price endogeneity due to unaccounted correlation of prices and advertising levels. The available endogeneity correction is not able to correct this, as the cause is linked to the individual consumption of advertisements. Recall the discussion in section 3.2.2.

In this chapter TV advertising is taken to be exogenous. Advertising for simple goods like detergent is unlikely to make a consumer go to a store, but should influence his behavior in a store. Think of two decisions the consumer has to take: the first is whether to buy in the category and the second is what product to choose. For the first decision advertising influences the consumer in the store just as all other covariates do, e.g. prices and promotions, whereas for the second decision it may play a more prominent role. I observe all on air TV advertisements for detergent products. Presumably, the aggregate level of advertising influences the consumer in first decision, while only own brand advertising is relevant for the second decision.[26]

[25] Assume prices and unobserved retail activity are positively correlated. Note that this correlation should hold on average for all products together. Example 1: Higher priced (branded) products are more often combined with retail activity than lower priced (private label) products. Example 2: This occurs if a product is either on display without price reduction (a high price) or not on display with price reduction (a low price). For both examples, a price increase is not as negative in its effect, as consumer benefitting retail activity increases, too. If retail activity is unobserved, prices will not be measured by the model as negative as they truly are, because retail activity counteracts. If retail activity is added to the model, the effect of prices can be identified as being worse on consumption, because the benefitting effect of retail activity is controlled for.

[26] Referring to Bagwell (2005) or Lauga (2008) advertising enters the consumption decision in many ways. A crude concept distinguishes informational versus prestigious advertising, the point of interest

If TV advertising is set by the manufacturer, it will be correlated with prices, at the least through the wholesale price component p^W. It may be uncorrelated to other demand side factors that still contaminate prices, as the remaining marketing variables operate on retailer level and are not directly influenced by the manufacturer.

Under these assumptions, the mere use of advertising data in the estimation alleviates part of the price endogeneity problem.

Direction of Advertising Bias TV advertising is assumed to be set by the manufacturer, and is likely to be set simultaneously with prices according to an unobserved decision calculus. Consider the following two thought experiments to understand the interplay of advertising and prices.[27]

Firstly, suppose manufacturers set prices and TV advertising such that I would observe a positive correlation in the number of advertising contacts and price level variables over time. Then omitting advertising this introduces attenuation bias. This is the same direction of bias that results from unobserved retail activity. The conventional price endogeneity correction à la Berry should fix this if all consumers were exposed to the same level, but this is obviously not the case as TV viewing patterns differ across consumers. Consequently, inclusion of advertising increases price effects. This case of positive correlation may arise as firms try to increase their prestige or brand recognition to create local market power under monopolistic competition so that firms can charge higher prices.

Secondly, suppose firms set prices and TV advertising such that I would observe a negative correlation in the number of advertising contacts and price level. This would cause an amplification bias. If observed advertising data are added to the model, price effects should decrease. In this case of negative correlation, firms may inform the consumer by advertising about pricing events, such as sales, certain rebates or jubilees.

Under the assumption that the first case fulfills the intuitive expectation of the role of TV advertising, this line of reasoning implies a positive correlation of TV advertising and prices.

Finally, note that for our discussion TV advertising is treated as exogenous, although it is presumably set by firms along with prices. With advertising added as additional endogenous variable, finding instruments for both endogenous variables is challenging and beyond the scope of this chapter.

Empirical Hypothesis 3 *Controlling for TV advertising increases price effects.*

in the work of Ackerberg (2001, 2003). I will not elaborate on the mechanism as this is not the aim of this chapter, but follow a reduced form approach.

[27] The argument builds on the intuition from a linear regression model that carries through if only one endogenous variable is present.

3.2.3.5 Cause 4 - Individual Inventories

Not accounting for individual level household inventories can lead to serious errors in estimating price effects. In a structural model Hendel and Nevo (2006) illustrate that for a simple storable consumer good: detergents. Different from that model, the discrete choice model in this chapter does not feature an explicit dynamic model of inventory holding, but controls for inventories by using a reduced form approach.

In most of the literature, however, individual inventory holdings (at the consumer level) are mostly ignored and seldom proxied. If they are proxied, a variable is added that captures the duration from the last purchase, the so-called interpurchase time. Whereas in Guadagni and Little (1983) it enters in the latent utility for each brand, approaches with no-purchase option let only the utility of the no-purchase option depend on the duration. Different from this, duration is implemented in this work by setting the duration variable equal to zero for the no-purchase alternative and retaining its value for all product alternatives at the purchase occasion.[28]

Why should individual inventories lead to price endogeneity? Inventory behavior induces a correlation of price and inventory on the individual level. I make two assumptions on simple consumer behavior for the reasoning: (i) Consumers look for a low price. (ii) There is not always a product on sale. Let subscript l indicate a low value and subscript h indicate a high value of a variable.

Then the reasoning goes as follows: When inventory is low ($=inv_l$), the consumer is likely to shop. If a product is on sale ($=p_l$ is paid), he will choose this. Sometimes there is no sale, then he chooses a product at normal price ($=p_h$ is paid). He will not choose the outside option (the no-purchase, where 0 is paid), as he needs a product due to his low inventory.

When inventory is high ($=inv_h$), the consumer is less likely to shop, even if a product is on sale, but more likely to choose the outside option (=0 is paid). He is most unlikely to shop if the price is high.

If I could observe price and inventory pairs (p, inv) in the consumer data, what would I expect to see? Under the two assumptions (i) and (ii) on consumer behavior from above, there will be a unique ranking of the frequency, with which the (price, inventory) pairs will appear.
1. The low inventory case: For the simple consumer behavior, there would be more (p_l, inv_l) pairs than (p_h, inv_l) pairs and no (0, inv_l) pairs, because consumers look for low price (due to (i)), and there exist (p_h, inv_l) pairs, as there is not always a sale (due to (ii)). Thus, for low inventory levels, I observe mostly low prices.
2. The high inventory case: There would be more (0, inv_h) than (p_l, inv_h) than (p_h, inv_h)

[28] It is a requirement for identification to normalize a variable that is constant across all alternatives of a purchase occasion to some fixed value for one alternative from the choice set.

pairs, where the latter is never realized if inventory is high and prices are high. Overall there will be few purchases if the individual has inv_h and most purchase occasions will result in a no-purchase.

Summing up, low inventories mostly come along with low prices in the data, and high inventories come along with the no-purchase option that has a price of zero. Then I have a negative correlation of price level and inventory level for the observed price and inventory pairs.

What is bias direction? As argued, the correlation of price and inventory is negative. The effect of price and inventory on latent utility are both negative for a purchase. Consequently, not accounting for inventory attributes the effect to the low price in comparison to other available alternatives and overstates the effect of the low price. Thus, I expect an amplification bias. Controlling for inventory should reduce the price effects. This conjecture is absolutely in line with the empirical results of Hendel and Nevo's (2006) structural model.

Duration can serve as proxy for the unobserved inventory level. The idea is that the longer ago is a purchase, the lower the inventory level will be. Then the purchase may be done on the basis of an actual demand (price plays its role in determining product choice) and not due to a stockpiling motive on the side of the consumer. Following a more formal argument I expect a positive sign on the duration coefficient. Let $P(j)$ be the probability of choosing good j, inv is the level of individual inventories and dur be the duration since the last purchase. It follows:

$$\begin{aligned} \frac{\partial P(j)}{\partial inv} &< 0, \forall j = 1, \ldots, J && (3.36)\\ \frac{\partial inv}{\partial dur} &< 0, \forall j = 1, \ldots, J && (3.37)\\ \Rightarrow \frac{\partial P(j)}{\partial dur} &= \frac{\partial P(j)}{\partial inv}\frac{\partial inv}{\partial dur} > 0, \forall j = 1, \ldots, J && (3.38) \end{aligned}$$

Empirical Hypothesis 4 *Controlling for individual inventories decreases price effects.*

3.2.3.6 Cause 5 - State Dependence

State dependence is modeled as a positive effect on latent consumer utility of choosing again a product of the same brand that was chosen at the previous purchase.[29] I do not elaborate on the mechanism that makes sticking to the brand have a positive effect on latent consumer utility. Why does this generate price endogeneity? First, recall the two assumptions on consumer behavior from the previous subsection on inventories. The

[29] Previously being chosen can be interpreted as an additional beneficial characteristic of a product. If the correlation is positive with prices, then I will have essentially the same type of bias as in the market level data case with unobserved product quality.

explanation is best illustrated by two examples.
1. Consumer sticks to a brand: Due to state dependence, a product has a higher probability of being purchased at whatever price if a product from the same brand has been bought before. The state dependence variable has a value of one.[30] Therefore, a consumer will tend to stick to a product from the same brand and is not likely to pick a similar product of a different brand, even if the latter may be offered at a lower price. The data contain data pairs for state dependence and price, where the state dependence variable is positive (=1, b/c it is a dummy) and any price (that means at high and low prices).
2. Consumer switches the brand: If a product is similar and very cheap compared to the previously chosen one the consumer may switch. When switching the data pairs look as follows: the state dependence variable is zero and price is very low.

Consequently, summing up both examples, price and state dependence are positively correlated.

What is the bias direction? The effect of price on latent utility is negative and that of state dependence is positive. When the consumer switches from one product to another brand a very small price comes along with a low value for state dependence. As a result, not accounting for state dependence underestimates the effect of the low price because price had to outweigh a zero state dependence value that is lower than if the consumer had stuck to the previous brand. Thus, I expect an attenuation bias. Controlling for state dependence should increase the price effects.

Empirical Hypothesis 5 *Controlling for state dependence increases price effects.*

3.2.3.7 Point 6 - Consumer Heterogeneity

As mentioned before, consumer heterogeneity is not an endogeneity cause, but is an important ingredient of any consumer choice model. The heterogeneity of consumers is captured in two different ways. The linear index structure of the discrete choice model can have random coefficients that capture the unobservable portion of heterogeneity in a parametric way. The parameters are assumed to follow a standard normal distribution.

The observed heterogeneity part uses demographic variables such as household size or income and these are interacted with price or duration variables. There is no survey data available as in Horsky, Misra, and Nelson (2006) to improve the observed heterogeneity part of the models by adding stated preference information.

I expect that adding observed heterogeneity to the model will decrease the standard deviations of the random coefficients. I do not have any a priori expectations about the impact on price effects.

[30] The state dependence variable consists either of a single dummy (one lag) or of a vector of dummies (several lags) as defined in section 3.2.

3.2.3.8 Point 7 - Data Aggregation

This work uses disaggregated data, a point worth to be mentioned as this deviates from common practice, but is unrelated to the central endogeneity questions of the chapter. It is common practice to aggregate purchase data to get weekly observations. This prohibits firstly intra-week variation in product choice and secondly does not allow more than a single unit purchase per week for the simple discrete choice models to work. In the data I observe both issues. Of course, this is in principle feasible in a discrete choice model with weekly data, but (a) it may lead to a huge choice set if pairs and triples of brands need to be considered and (b) the interpretation of the choice set elements from (a) is not clear. I have no a priori expectation about changes of price effects that stem from using daily instead of weekly aggregated purchase data.

3.2.4 Choice and Discussion of Instrumental Variables

To find instruments with two potentially endogenous variables, namely prices and advertising, is a difficult task. I have argued in section 3.2.3.4 that there are even in the most simplistic case several strategic firm patterns for the detergent industry. Due to this I only treat price as endogenous variable.

In the literature unobserved brand-time specific effects, i.e. local unobserved retail activity, cause price endogeneity. This retail activity is the favored explanation Chintagunta, Dubé, and Goh (2005) have for introducing the unobserved brand characteristic ξ. Thus, an instrument should at least break this correlation of unobserved retail activity and prices.

Before I motivate the usage of the instruments in this chapter, recall that the data are a nationwide sample and that the restrictions from equations (3.19) and (3.20) do not hold here. I have variation in local retail activity and product prices that previous papers did not have. This will prove helpful for the construction of instruments that is outlined in the following paragraphs.

I use instruments of the Hausman (1996) type, but modify the approach further to get better instruments.[31] The modification amounts to using a split sample idea detailed in the construction description of the instruments. Nevo (2001) uses the Hausman instruments for his market level data application, too.[32] The instruments are constructed as follows and I shall start with the general idea.

For a given product price, suppose that the choice for the instrument $p^\star$ is price for a similar product at the same time in another outlet. As all underlying manufacturer

[31] See Petrin and Train (2006) on p. 22. They propose another instrumental variable approach in the manner of Pakes (1994).

[32] A discussion of these instruments is on p. 309 top and p. 320 bottom of his paper. For potential weaknesses of Hausman instruments see p. 321.

specific characteristics are the same due to the similarity of both products, this delivers instrument relevance, e.g. production costs enter their retail prices. The difference between price and its instrument stems from the retail activity in the outlet and any agreement between manufacturer and chain that influences price setting. I summarize these two latter factors under the term local brand-time effects. If these factors are unrelated across outlets instrument exogeneity prevails. Hence, if Hausman instrument $p^\star$ are prices of similar goods in another outlet, this instrument breaks the correlation of ξ and the endogenous product price. The underlying assumption is that local brand-time effects are independent across outlets. This is a problematic assumption for large retail chains, as they may have centralistic coordinated retail activity, but this is not problematic for small stores. My instrument construction uses this structural difference between large and small outlets. This is the point where the split sample idea is used.

In the empirical application I construct Hausman instruments in the following way. I restrict myself to the 10 biggest chains for the analysis and use the mean prices of the remaining 160 chains in the same week, for a "similar" product, in other geographic regions (identified by postal/zip code or county "Landkreis") as instrument.[33] Two sources of endogeneity are eliminated using this approach. Local demand effects are eliminated by using prices in other geographical regions. The correlation between similar kinds of retail outlets in different regions is broken up by using prices of small retailers as instruments for prices in big retailers. Thus, not only the retailers have different names, but also structurally small retailers are surely different from large retailers. Chintagunta, Dubé, and Goh (2005) cannot do this, as they focus on a narrow geographical area.

Petrin and Train (2006) favor wholesale prices, but informal interviews of the author with market participants indicate that there are only direct unobserved negotiations between retailer and manufacturer in Germany so that no "wholesale" price exists for a consumer good like detergents as in the US.[34] None of the major marketing companies collect this price information for the German market. One drawback of wholesale prices is the missing local variation that is present if Hausman instruments are used. In addition, wholesale prices might be set strategically in accordance with TV advertising, another source of price endogeneity. Therefore, it is not obvious why wholesale prices are first choice instruments.

In the following I discuss two other hypothetical approaches for finding instruments. A potential set of instruments are aggregate advertisements by a manufacturer, i.e. all national broadcast TV advertisements, that are correlated with price but not with local retail activity. Then the advertisements per consumer (i.e. the filtered aggregate

[33] "Similar" means that brand, approximate packaging size and consistency match. I resolve using the mean over all other geographic regions than that region with the endogenous price. Instead one could only use the "near" regions to increase variability of the instruments.

[34] Such market participants are people working in marketing agencies, marketing research companies and market research at retailers.

advertisements that depend on what program the consumer watches) could be used to see the impact of advertising on consumers, while aggregate advertisements control for unobserved retail activity. The data for this procedure are available. Problematic is the missing possibility to create local variability of the instruments.

An alternative way to find instruments is the "unobservable instruments" approach of Matzkin (2004).[35] There is random variation between total advertising and advertising received by individual consumer. Not seeing a certain spot is random, as TV choice decision is not based on the choice of advertising but on the TV program. However, this variation may be correlated to the advertising level so that it is not an instrument for the advertisements by a manufacturer. At least it is unrelated to the unobserved retail activity that - according to Chintagunta, Dubé, and Goh (2005) - is the primary source of endogeneity in the case of individual panel data. To qualify as instrument it must be correlated with prices and the conjecture is that it is less relevant than the Hausman instruments presented here.

To find instruments when both prices and advertising are treated as endogenous is still an outstanding challenge.

Finally, there is a specific reason why I do not follow the methodical approach of Chintagunta, Dubé, and Goh (2005) for the endogeneity correction. Their approach requires estimation of one parameter per product, time and location triple (that is for each j,o,t triple) and a subsequent decomposition of these parameters with linear IV methods according to the original Berry (1994) procedure.

As their data have many observations in only one location that method apparently worked although the high number of parameters does not make the optimization trivial. In this work, the original Berry procedure is not feasible with a widely dispersed nationwide sample. There are not enough observations per product, time and location triple, i.e. the ratio of observations per parameter to be estimated is unfavorable. In addition, the total number of parameters is very high.

3.2.5 Method and Interpretation

In this section I want to explain the methodology used to compare the discussed endogeneity causes and argue that it is a suitable procedure to compare the many different estimated specifications.

It is not informative to compare coefficients in logit and mixed logit models because the level of the parameters is not identified. Coefficients are identified up to scale, but typically other parameters than the parameter of interest change as well when switching from one specification to another. This prohibits a good interpretation of results based

[35] The idea originates from Richard Blundell.

on parameter estimates of logit models across specifications. Therefore, comparison of price elasticities is suitable to understand price effects. Price elasticities are expected to point into the same direction of interpretation as the coefficients suggest. So far, every paper in this literature uses both means to report results.

Direction and magnitude have a clear interpretation for elasticities and are comparable across specifications. Especially, the economic interpretation is lucid if the results are presented as elasticities. There are clear expectations from classical microeconomic theory about price elasticities in the homogeneous goods case under perfect competition: (i) the value should be in the negative domain and (ii) it should be elastic: absolute size is larger than 1.

Typically, an elasticity is defined as the effect of a relative marginal price change of alternative $j \in \{1, \ldots, J\}$ on the probability of purchase for any product $k \in \{1, \ldots, J\}$. As the number of products is very high I resolve to calculate the elasticities for brands: What is the average elasticity of a brand, i.e. aggregated from all products of that brand, following a price change for all products of another brand. This enables a parsimonious analysis of endogeneity causes.

3.3 Data

3.3.1 Data Overview

The data are an extensive household level panel supplied by A.C. Nielsen, Germany. The "Single Source" panel provides household, daily purchase and real-time advertising exposure over a period of 2 years from June 30^{th} 2004 through June 30^{th} 2006.[36] The name Single Source highlights the fact that daily purchase and high frequency TV advertising exposure are each recorded for the same household.[37] The A.C. Nielsen competitor GfK (Gesellschaft für Konsumforschung AG, Nürnberg, Germany) does not supply these data based on the same households, but tries to combine the information from two panels using matching procedures.

The dataset is collected nationwide throughout Germany and consists of two components: a household panel where purchases are followed and a subsample of the former where additionally all TV advertising contacts are recorded automatically. As the data consist of several collected data files from A.C. Nielsen, sample sizes of the merged files

[36] Precisely speaking the purchase information "Homescan" is collected by A.C. Nielsen, and the media information for the same households is collected by Nielsen Media Research, both companies belonging to the Nielsen group. A.C. Nielsen supplies the combined data. Single Source and Homescan are registered trademarks of A.C. Nielsen.

[37] See the reviews about the data collection procedure for Single Source authored from various sampled consumers at the Website of Ciao GmbH in 2007. Viewed January 14th 2007: `http://www.ciao.de/ACNielsen_Werbeforschungsunternehmen__942530`.

vary with the desired information. The detailed construction of the employed dataset based on the underlying dataset from A.C. Nielsen and its full contents are described in detail in chapter 2 of this dissertation. Note that a few tables are repeated in this chapter for the reader's convenience.

Table 3.1. Overview of Consumer Data Files

File	Household appears if	Description
Cash	sampled	total value of purchases with time, store, zip
Wash	purchased detergent	detergent purchases with time, store, zip and product details (price, quantity, characteristics)
Demo	sampled	time constant socio-demographic Variables
TV	TV telemeter equipped	TV advertisement, TV representation factors

See table 3.1 for details on the four files containing the purchase trip information, detergent purchases, sociodemographics and advertising exposure.

Table 3.2 reports the number of households for which relevant information is available. Sociodemographics are available for all households. 80% of the sampled and purchasing households buy detergents. 23% of the purchasing households have participated in recording advertising exposure.

In appendix B table 3.11 lists the variables used in this work with a short description. An more exhaustive description of the important variables is given in section 3.3.2.

Table 3.2. Number of Households by Required Information

Dataset	Criterion	No. of Households
Demo	sociodemographics known	17,978
Cash	any purchase	16,757
Cash	any purchase in "detergent" store	16,737
Cash	above plus demographics	16,737
Wash	any purchase of detergent	13,455
Wash	above plus demographics	13,455
Wash	TV coverage in any year	3,783
Wash	TV coverage 2004	2,953
Wash	TV coverage 2005	2,630
Wash	TV coverage 2006	2,571
Wash	TV coverage 2004 and 2005	2,250
Wash	TV coverage 2005 and 2006	1,993
Wash	TV coverage 2004 to 2006	1,735

Notes: A “detergent” store is defined as store where it is possible to buy a detergent.

The product category under consideration is detergent. Data for the chocolate category is available. For detergents, Erdem and Keane (1996) argue that (a) detergents are “frequently and regularly purchased products”; (b) brands are frequently introduced; (c) “firms heavily advertise in this category”; (d) detergents are “low in variety seeking”.

In addition to this, detergents have an objective functionality that is to clean clothes, and are thereby required by any type of consumer. Detergents are storable products so that individual inventories have to be considered. Chocolate represents a potentially addictive product and does not provide a basic function as detergents do. Therefore, advertising may have distinct effects on both categories and consumers may decide differently following advertising exposure. In this chapter I restrict my attention to the detergent category.

I observe daily visits to supermarkets and the amounts spent at each visit for two product categories: chocolate and detergent. The following data are recorded daily per purchase: aggregate amount spent per visit, exact brand-size combinations with quantity, price and precise characteristics, name of outlet and location. As usual, retail activity variables (i.e. feature, display, priceflag, handbill) are contained in the data as well.

Table 3.3. Summary Statistics of Household level Raw Data

Variables	Units	Mean	Median	Std.	Min	10%	90%	Max
Demographics								
Income	HH	2566	2000	1930	750	1250	3500	10000
Urbanity	HH	5.6	6	1.55	1	3	7	7
No. persons	HH	2.43	2	1.20	1	1	5	10
No. kids	HH	0.53	0	0.88	0	0	2	8
MA Age	HH	47	45	14.8	18	28	68	93
MM Female	HH	0.76	1	0.43	0	0	1	1
Self employed	HH	0.09	0	0.28	0	0	0	1
White-collar high	HH	0.46	0	0.50	0	0	1	1
White-collar low	HH	0.23	0	0.42	0	0	1	1
Blue-collar	HH	0.21	0	0.41	0	0	1	1
Unemployed	HH	0.01	0	0.1	0	0	0	1
Days in sample 1*	HH	320	278	255	0	0	674	730
Days in sample 2	HH	392	396	227	1	84	683	730
Detergents								
Price	PI	3.71	2.99	2.30	0.01	2.19	6.66	32.99
Quantity	PI	1.11	1	0.44	1	1	1	16
Duration	PI	53	34	64.2	0	0	126	723
Duration 1**	HH	81.7	60.7	74.7	0	22.7	162	723
Duration 2**	HH	91.6	69.4	77.4	1	29.4	177	723
No. Brands bought	HH	2.4	2	1.7	1	1	5	14
Brand HHI***	HH	0.69	0.68	0.29	0.10	0.31	1	1
Brand HHI****	HH	0.69	0.66	0.29	0.09	0.29	1	1
Store visits								
No. Stores visited	HH	2.34	2	1.49	1	1	4	13
Store HHI***	HH	0.69	0.64	0.28	0.07	0.32	1	1

Notes: Statistics are constructed per household (HH) or per purchase incident (PI). * Households that purchase only one time are included: 2247 observations. ** Duration 1 includes more than one purchase a day as duration zero values. Duration 2 combines all multiple purchases per day to one purchase. *** indicates the HHI is calculated with value as weights. **** indicates the HHI is calculated with volume as weights.

Table 3.3 contains descriptive statistics calculated per household and per purchase incident for the category purchase data. The sample does not look representative for Germany. For example, income is too low compared to official figures just as is average number of kids. As in most datasets of this size I encounter coding anomalies that are partly very obvious. For example, some detergents are sold for one € cent, perhaps an internal price of the retailer, or a household has a duration of over 700 days between two purchase occasions. I exclude observations with these obvious errors to obtain the samples used later for the estimation. For a detailed discussion of these data issues consult chapter 2 of this dissertation. Interestingly the Herfindahl-Hirschmann Index concentration measures, brand HHI and store HHI, indicate that consumers stick to their favorite stores and brands.[38]

The TV advertising variable contained in the data needs some explanation. It is possible to see if a household TV is tuned to a channel with a product specific advertisement. The broadcast time, the TV channel and the topic/motive of the spot are recorded automatically. I assume that this spot has an impact on the household, in particular on the member who is responsible for purchases, but it is unknown which person of the household was tuned in to the program.[39] This assumption is of course nothing but a mere approximation, but potentially the data contain information on the household that

Table 3.4. Market Shares and Means per Brand for the full Estimation Sample

	Shares					Means		
Brand	Sales	Value	Advertising	Feature	Display	Price	Size (kg)	Size (l)
3	0.06	0.10	0.36	0.15	0.14	5.37	2.75	1.93
6	0.01	0.01	0.00	0.04	0.02	3.40	1.57	1.41
8	0.02	0.02	0.02	0.07	0.05	3.71	1.14	1.59
10	0.01	0.02	0.00	0.00	0.03	4.38	6.03	1.46
11	0.01	0.01	0.00	0.03	0.03	3.90	3.64	1.56
13	0.02	0.01	0.00	0.00	0.00	1.92	–	1.13
16	0.56	0.43	0.00	0.19	0.11	2.70	2.16	1.54
22	0.01	0.01	0.00	0.02	0.02	3.42	1.35	2.00
36	0.02	0.03	0.07	0.03	0.05	4.92	4.32	2.01
40	0.07	0.14	0.32	0.21	0.17	6.79	3.25	2.27
41	0.03	0.03	0.04	0.04	0.03	3.99	0.83	1.56
55	0.09	0.09	0.14	0.09	0.13	3.58	1.84	1.77
57	0.02	0.03	0.00	0.05	0.08	4.22	4.19	1.69
67	0.03	0.04	0.04	0.07	0.10	4.79	3.79	1.83
100	0.02	0.02	0.00	0.01	0.05	4.63	4.42	1.13
Total	0.07	0.07	0.06	0.15	0.10	3.51	2.57	1.62

[38] HHI measures are calculated in normalized form so that values range from $\frac{1}{b}$ to 1, where a higher number indicates higher concentration and b is the number of alternatives.

[39] In fact A.C. Nielsen records this information, but it is not part of the available raw data. There is a recent literature that emphasizes the importance of looking into the households and not simply take the whole household as one decision making unit, see e.g. Cherchye, Rock, and Vermeulen (2007).

might be used to mitigate this assumption, i.e. number of household members, gender and age.

Consult table 3.4 to get an overview of the detergent market used in the full estimation sample without the anomalies visible in table 3.3. The table shows shares and means of interesting variables broken down by brands. The last line shows market averages. The first column is the brand code, the following five columns show market shares and the last three means. The market is segmented into two broad groups: common branded products and private labels. The latter are indicated by brand number 16 in the table. Private labels are products that have a brand that is specific to the chain where it is sold and usually represent the "discounter" alternative that is low priced. Note that the two most heavy advertisers by share in the market charge the highest average price. Moreover, the biggest brand by sales, i.e. the private labels of the retailers, do not make any advertising and charge a price below average. In table 3.5 the variety characteristics that differentiate products are split up by brand. The last line shows market shares. There is high variation across brands in the characteristics "Liquid", "Concentrate" and "Color". Many brands do not offer any product variety with one of the characteristics from the last 4 columns. This offers the possibility for some brands to differentiate from the rest of the market due to a characteristic and possibly charge higher prices. For example, brands that have a nonzero entry in the column "Gimmick" mostly charge prices above average.

Table 3.5. Characteristics for all Product Purchases broken up by Brand

	Means		Shares								
Brand	Nominal Price	Size in kg or liters	Sales by Volume	TV Advertising	Liquid	Konzentrat	Color	Sensitiv	Gimmick	Extrasize	Bigpack
3	5.37	2.41	0.06	0.36	0.41	0.59	0.51	0.04	0.06	0.00	0.13
6	3.40	1.48	0.01	0.00	0.61	0.26	0.10	0.00	0.01	0.03	0.00
8	3.72	1.45	0.02	0.02	0.68	0.00	0.80	0.00	0.13	0.00	0.00
10	4.38	3.86	0.01	0.00	0.48	0.09	0.26	0.01	0.07	0.00	0.00
11	3.90	2.36	0.01	0.00	0.61	0.89	0.00	0.00	0.04	0.18	0.09
13	1.92	1.13	0.02	0.00	1.00	0.00	0.13	0.02	0.00	0.00	0.00
16	2.70	1.84	0.56	0.00	0.52	0.53	0.42	0.01	0.00	0.00	0.00
22	3.42	1.94	0.01	0.00	0.90	0.00	0.05	0.00	0.00	0.00	0.00
36	4.92	3.37	0.02	0.07	0.41	0.26	0.16	0.00	0.06	0.09	0.06
40	6.79	2.98	0.07	0.32	0.28	0.47	0.40	0.11	0.13	0.12	0.06
41	3.99	1.40	0.03	0.04	0.79	0.00	0.14	0.00	0.00	0.04	0.00
55	3.58	1.80	0.09	0.14	0.54	0.39	0.14	0.00	0.01	0.01	0.00
57	4.22	2.96	0.02	0.00	0.49	0.29	0.42	0.00	0.00	0.10	0.00
67	4.79	3.08	0.03	0.04	0.36	0.28	0.02	0.00	0.16	0.10	0.02
100	4.63	3.79	0.02	0.00	0.19	0.19	0.06	0.00	0.00	0.03	0.06
Total	3.51	2.08	0.07	0.06	0.51	0.45	0.36	0.02	0.03	0.02	0.02

3.3.1.1 Relation to other Datasets

Similar datasets were already used in the literature, where each one has his comparative advantages for his application.[40] Most important for the current chapter are the differences in the quality and quantity of pricing and advertisement information. The data of Hendel and Nevo (2006) track households that purchase in one store and a complete store level dataset is available to deliver all prices during purchase decisions of the consumer in a simple fashion, but it lacks the TV advertisement information of the dataset in this work. Concerning prices, my data are household level data of the same kind as Keane (1997), where he has only a few regional markets in the US in his dataset. He has to impute product alternatives the consumers face on each shopping trip. I deal with this issue in a similar fashion to him as explained in the next section. Erdem and Keane (1996) have a dataset that tracks households daily in two stores from 1986 to 1988. In principle it has the same TV advertising data, but only for 51 weeks and for only 1800 households and due to experimental recording technology there is a lot of missing data in the advertising history. My data are very precise and complete in this respect, as the technology used for measuring advertising exposure has matured ever since. Their data also lack the precise clip information I have to identify image and product specific clips. The product category of interest is also laundry detergents.

3.3.2 Variables of Interest in the Analysis

This section deals with the specialities of the central variables of the chapter, especially prices. In the data I face an incomplete price list for the available alternatives per consumer at each store visit. Only prices of purchased products are reported and thereby price information needs to be inferred from within the dataset to fill up the missing prices. Partly, variable details have already been given in chapter 2 but have to be repeated to emphasize their role for the application and illustrate the link to the empirical model in section 3.2.

Inferring Prices I follow Erdem and Keane (1996) and Keane (1997) in inferring prices. The large number of households permits to construct the price of an alternative that a consumer faces by filling in the price gaps with prices of other consumers who shopped in the same store at the same time. Usually, all data are aggregated to weeks. With daily data at hand and the assumption of constant prices over a week at a given outlet, I can use price information from other days within a week for the same outlet to get prices for many product alternatives at different visits.

I use the following procedure: Collect all product purchases in a calendar week in a chain in a county ("Landkreis"). Fill up missing product alternative prices in all store visits within the same week, chain and county. Repeat this for all weeks, chains and counties.

[40] In chapter 2 of this dissertation several datasets are listed and compared.

This procedure minimizes the measurement error in prices that would flaw the analysis. Of course this comes at the cost of having fewer elements in the choice set at each store visit. It is not feasible to use zip codes instead of counties in the procedure. The fewer number of counties than zip codes makes it much more likely to find two households in the same area.

Different from the US, in Germany the practice of issuing price coupons in stores that reduce the retail price is not common during the sample period so that no correction is needed for the imputed prices.

There are other possible ways. A more detailed discussion with suggestions can be found in chapter 2 of this dissertation.

TV Advertising Variables Those variables capture the individual household exposure to TV advertising. The data are precise enough to record the following spot characteristics for all broadcast times per household: length of the spots in seconds, advertised brand and dummy variables indicating a powder ad spot, a liquid ad spot and image spot (for whole brand, no specific product). For all mentioned characteristics I construct variables that cumulate the number of brand-specific advertising contacts a household had over an interval prior to a specific date. The intervals for the cumulation increase by 14 day steps. I use the following interval values: 14, 28, 42, 56, 70, 84, 98, 112, 126 and 140 days. The interval size 126 days means that the constructed variable for a date t contains the cumulated contact history for a spot characteristic per brand for 126 days prior to the date t. The spot characteristics are those mentioned at the start of the paragraph. The number of different intervals allow to discriminate short-term versus medium-term versus long-term effects of advertising. As mentioned earlier advertising is completely measured with precision at the household level, there is no imputation necessary.

In the following paragraph I explain the variables used in the analysis: `countc140ad, liqc56adr, countc140adpr` ... This set of cryptic variables define the TV contact of the household. Each variable code consists of 3 parts and I explain each component.

The word up to the letter `c` (i.e. `xxxxxc140ad`) defines the type of spot: `count` indicates contact with a spot for the brand, `time` time length of contact in sec with a brand spot, `liq` contact with a liquid detergent spot, `pow` contact with a powder detergent spot, `img` with an image spot.

The number after the letter `c` (i.e. `countcxxxad`) defines up to which lag in days the contacts are cumulated. The values for the lag are multiples of 14: 14, 28, ..., 126, 140. After that number the following keywords `ad, adr, adpr` detail variable construction (i.e. `countc140xxxx`): standard is `ad` where advertisement contacts are simply cumulated for the specified lag so that the variable sums absolute contacts in a given time window as in equation (3.11). If the variable ends with `adr`, it is the same variable ending on `ad`, but in addition it is divided by total advertisement contacts of all competitors up to the same lag according to (3.13). The latter construction can be thought of as measuring advertising pressure over a time period of a certain brand relative to all competitors. If the variables ends with `adpr` it captures advertising pressure between two time windows, i.e. `countc140adpr` captures the relative advertising pressure in the time window of 140 to 126 days before the purchase.[41] It is possible that advertising has different effects given its timing before the purchase occasion. This should allow to distinguish if advertising short or long time before the purchase plays an important role.

Duration Variables (`duration, duration2, idurhh`) These variables can be used to control for inventories, due to the intimate link between inventories and duration. The variable `duration` is constructed such that on each alternative the value is the time in weeks since the last store visit that led to a detergent purchase. To make the variable identifiable I set the value of the variable to zero for the no-purchase alternative. The first observation of each household has no prior visit, so the variable is set to missing. `duration2` is merely the square of `duration`. In the case of multiple purchases on a certain day, duration is weeks elapsed since a previous purchase at an earlier date, therefore `duration` is never zero for a brand alternative. `idurhh` is the interaction of `duration` and household size in persons. This interaction captures the effect of `duration` for differently sized households on consumption of detergent. For example, big family households will have an empty inventory at a lower value of `duration` compared to a single person household.

State Dependence Variables (`GLdumA, GLdumB, mdum*`[42]) These variables can be used to control for state dependence. It is modeled as so-called brand loyalty where this term highlights the underlying habit of the consumer. `GLdumA` and `GLdumB` are defined according to equation (3.7) when setting $\phi = 0$ so that the exponential specification is simplified to contain only one lag. Thus, the variables are simple dummies that take the value one, if the previous purchase was of the same brand as the faced alternative. In case of a store visit that leads to a detergent purchase this is fine. If a visit results in a no-purchase, it makes sense to set the dummies to zero, as the detergent category is not under consideration and brand loyalty is not relevant. Thereby I define two sets of dummys: `GLdumA` contains values on all store occasions, `GLdumB` sets all dummys to zero for no-purchase occasions. When looking at more than one purchase further back into the past, a series of dummies can be constructed according to the same rules as `GLdumB`. This resulting series of dummies `mdum1, mdum2, ...` compared to the exponential specification in (3.7) is additively separable and does not have parameters that are a priori restricted. The integer numbers specify the lag.

[41] Note the difference: The variable countc140adr captures relative advertising pressure from 140 to zero days before the purchase occasion.

[42] The $*$ is a wildcard as in usual programming languages and represents an integer.

3.3.3 Construction of Instrumental Variables

This section details the construction of instruments that are used for the analysis. Recall that these instruments solve the endogeneity caused by unobserved retail activity, i.e. cause 2. Only instruments for the 10 largest chains of the dataset are generated, where the remaining 160 chains serve as origin for the instruments. Two assumptions are invoked: (1) Retail activity that might hit consumers is independent between the largest 10 and the rest of the 160 chains.[43] (2) There are no demand shocks that operate independently of the chains and influence consumers across locations. Assumption (1) qualifies prices from small retailers as valid instruments for prices at big retailers if unobserved retail activity that is linked to the retailer causes an endogeneity problem. By assumption (2) prices from all other regions other than a given region are valid instruments for the given region if local demand shocks that are unrelated to the retailer cause the endogeneity.

I give three reasons that invoking assumptions (1) and (2) is a good approach to construct an instrumental variable.

1. Assumption (1) is credible as the 160 stores may follow other strategies than the 10 big chains when allocating the products in their limited shelf space. It is common practice that manufacturers buy themselves in or competitors out in the shelf space of the big national chains, i.e. the 10 big chains in the dataset. Manufacturers might do that in a concerted action with price setting. Under assumption (1), the instruments break this link.

2. Outlet level effects, such as stock clearance or any discretionary measures of the single outlet may lead to price cuts and special exposition that are not recorded in the retail activity variables in the marketing datasets. Prices in other smaller outlets are independent of such actions. This thought invigorates both assumptions (1) and (2). In smaller outlets in other regions the mentioned activities are unlikely to be recognized at all.

3. Retail activity is often combined with price changes. Then any reporting error in the retail activity variable may attribute the effect to prices if prices were modified at the same time. As detergent is a very active product category with entries and exits of specific brand varieties this is credible. Under assumption (1), independence of prices in small outlets from retail activity in big chains solves this problem.

The instruments for product price are constructed in the following way. I calculate the mean of product price for "similar" products in all other geographic regions (than the one with the endogenous price) over small outlets in the same calendar week. A product is "similar" if (i) it has the same brand (ii) size rounded to integers in kilograms or liters

[43] Most of the largest chains operate nationally, whereas the majority of the smaller chains are restricted to a geographic area.

is the same (iii) it has the same consistency (liquid, powder, ...) (iv) it has the same general purpose washing function (general detergent, for sensitive fabrics, ...).

These variables are used as regressors in an OLS regression of endogenous prices on instruments and product characteristics of the good, i.e. size, dummys for sensitive skin, color, concentrated detergent according to equation (3.30). As explained previously, for each brand a separate OLS regression is estimated. The fitted residuals of these OLS regressions for all store visits constitute estimates $\widehat{g_b(\xi_{jot})}$ of a function of the unobserved effect ξ_{jot} from the endogeneity correction model of section 3.2. Adding the corresponding values to the model alleviates the endogeneity in prices caused by unobserved retail activity.

3.3.4 Definition of Estimation Samples

In total, three different samples are used throughout the chapter: a full sample, an advertising sample and a comparison sample. To get results that are free from outliers or data artifacts, households/observations for the estimation of the models from section 3.2 are selected according to the following criteria: (a) in the choice situation, the consumer's choice set had at least 4 elements. (b) all shopping trips that resulted in no detergent purchase and had a total shopping volume below 5 € are dropped. (c) all shopping trips that resulted into no detergent purchase are dropped and that were on the same day as another shopping trip with a detergent purchase (d) all shopping trips are dropped that resulted into detergent purchases and that were on the same day as other shopping trips with a detergent purchase (e) all households that had no detergent purchases were dropped. (f) the shopping trip was to one of the 10 biggest retail chains.

All purchase occasions that fulfill criteria (a)-(f) are part of the full estimation sample. Condition (f) is needed, as it is part of the instrumental variable construction. Condition (d) is necessary since it is impossible to model these trips convincingly in the simple discrete choice models used here.[44]

As advertising is not available for all of the households, a subsample of the full sample termed advertising sample, fulfills in addition the following criterion: (g) the household was registered as TV household and had to be subject to TV advertising during the sampling period.

Apart from the daily data samples, a comparison sample on weekly data for liquid detergents is constructed to match the setup of Petrin and Train (2006). As several identical branded products can be bought per week, I have taken the median of the price if several purchases occur, which introduces measurement error. When constructing instruments, I aggregate to weekly price data by taking median values.

[44] I discuss related issues in more detail in section 4 of chapter 2 when describing the construction of the dataset.

3.4 Results

First, I present a summary of the results and details of the elasticity results. The rest of the results section digs deeper in detail into the parameter estimates of mixed logit models. Table 3.6 illustrates the main results at a glance. The hypotheses formulated earlier get differing support: three out of five are affirmed, whereas the remaining two are unconfirmed. The impact of a cause is defined as large when upon introduction of a cause into the model the price elasticities change by more than 20%.

Table 3.6. Empirical Hypothesis and Summary of Results

Hypothesis			Empirical Results	
No.	Cause	Expected Effect on Elasticities	Direction⋆	Magnitude⋆⋆
1	varieties	decrease	affirmed	large
2	unob. retail activity	increase	affirmed	large
3	advertising	increase	indecisive	small
4	inventories	decrease	indecisive	small
5	habits	increase	affirmed	large

Note: ⋆ The direction is affirmed if in all specifications the change of the price effect corresponds to the hypothesis. ⋆⋆ The change is large if the price elasticity changes by more than 20% upon introduction of a control for the endogeneity cause.

Variety information, unobserved retail activity and habits have a large economic impact on price elasticities. Variety information reduces price effects, whereas unobserved retail activity and habits increase price effects. Not accounting for one of them will seriously misquantify price elasticities. The result for those three causes is in line with the arguments brought up in section 3.2.3. In contrast to that, TV advertising and inventories only have a negligible impact on price effects.[45] Considering that advertising is one of the leading examples frequently brought up in papers on price endogeneity, this is a surprising result.

3.4.1 Price Elasticity Results

The tables 3.7 to 3.10 give a summary of the results for mean own price elasticities across brands for the full and advertising sample.[46] The first two columns show for each specification the implied price elasticities without and with endogeneity correction. The right

[45] Recall that in the detergent category almost all manufacturer advertising expenditures are spent on TV advertising, see footnote 3 in the introduction. Therefore, synonymity of TV advertising and advertising is justifiable.

[46] Results for cross price elasticities across brands are also available, but do not add substantial insights into the questions raised in this chapter over own price elasticities. Outside option price elasticities are also not additionally informative.

section of each table shows which additional causes are being controlled for. All models contain at least the variables from the literature benchmark model.[47] Importantly, bear in mind that comparison of the causes is done relative to the standard literature case that does not encompass controls for the five causes. From an econometric standpoint the standard case is a misspecified model so that comparisons of different misspecified models with additional controls is not a sound methodology. However, comparison to the standard case is interesting nonetheless and indicates the direction of change.

Next I discuss the results for the full sample. Table 3.7 contains the results.[48] The upper panel A shows the results for nominal prices and the lower panel B for efficiency prices. It is interesting to see that dependent on the causes accounted for I have a very wide bandwidth of elasticity estimates. Note that due to the computation time for each model it was so far not possible to estimate bootstrap standard errors, but due to the large sample size an interpretation should be allowed. Elasticities for the efficiency price case are always greater than for the corresponding nominal price case specification. Several results are interesting in this table. Inventory effects are small, as elasticities hardly change from row 1 to row 2. Variety information decrease price elasticities, compare row 1 to row 3. Habits/state dependence increase price effects strongly, compare row 1 to row 4, and even more pronounced when efficiency prices are used.

Table 3.7. Mean own price Elasticities - Full sample

Elasticities		Specifications			Model
normal	corrected	Inventory	Varieties	Habits	Code
		A: Nominal Prices			
-1.46	-2.25				s445b,a
-1.46	-2.19	X			s448b,a
-1.04	-1.62		X		s449b,a
-2.56	-4.26			X	s44GLb5b,a
-1.93	-3.04	X	X	X	s44GLb10b,a
		B: Efficiency Prices			
-2.25	-2.54				es445b,a
-2.42	-2.58	X			es448b,a
-1.85	-2.07		X		es449b,a
-5.78	-6.23			X	es44GLb5b,a
-4.75	-5.57	X	X	X	es44GLb10b,a

Note: Model code is explained in section 3.4.2.

[47] To get the precise specification with included variables check table 3.18 in the appendix and look at the last two columns.

[48] Sample size and coefficient estimates can be identified according to the model code and are found in tables 3.18, 3.19, 3.20, 3.22 for the nominal price case and in tables 3.27, 3.28, 3.29, 3.31 for the efficiency price case. The model code is explained in section 3.4.2.

Comparing specifications, when controlling for 3 out of 5 causes without endogeneity correction versus no control, the overall effect is an increase of own price elasticities of about 32% for the nominal price case and 111% for the efficiency price case. The analogue comparison for adding 3 causes with the endogeneity correction yields numbers 35% and 119%, respectively. It is comforting to get the same relative changes under both price regimes when accounting for the 3 causes, independent of the endogeneity correction.[49] Comparing the no cause corrected elasticity with the 4 causes corrected elasticity in this table I get a 108% increase for the nominal and a 147% for the efficiency price case.[50] Using only the endogeneity correction I get as numbers 54% and 12%. Thus, accounting for the additional endogeneity causes in addition to the endogeneity correction proves to be quantitatively important.

Table 3.8. Alternative mean own price Elasticities - Full sample

Elasticities		Specifications			Model
normal	corrected	Inventory	Varieties	Habits	Code
		A: Nominal Prices			
-2.01	-2.94		X	X	s44GLb11b,a
-2.26	-3.83	X		X	s44GLb12b,a
-1.13	-1.65	X	X		s44GLb13b,a
-1.93	-3.04	X	X	X	s44GLb10b,a
		B: Efficiency Prices			
-4.85	-5.42		X	X	es44GLb11b,a
-5.69	-6.21	X		X	es44GLb12b,a
-1.92	-2.19	X	X		es44GLb13b,a
-4.75	-5.57	X	X	X	es44GLb10b,a

Note: Model code is explained in section 3.4.2. Note that I do not provide the parameter estimates for the models presented in this table in the appendix.

To come back to the issue raised before of comparing the impact of the causes relative to the literature specification, I did the comparison relative to the correctly specified model with all endogeneity causes controlled for and always removed one cause. Table 3.8 reports the results. For inventories, the findings from before change. Under both price regimes without endogeneity correction, inventories reduce price elasticities as expected. But with endogeneity correction the effect is reversed to lead to an increase in elasticities. Variety information, state dependence and the endogeneity correction have the same effects shown earlier. Therefore, I will present the rest of the results in the format of table 3.7, where I compare all endogeneity causes relative to the literature benchmark case.

[49] Percentage number are the relative change from row 1 to row 5 (nominal prices) respectively row 6 to row 10 (efficiency prices). Without endogeneity correction, use column 1, with it use column 2.

[50] The number for the nominal case stems from the relative increase from -1.46 to -3.04, for the efficiency case the numbers are -2.25 and -5.57.

Table 3.9. Mean own price Elasticities - Advertising Sample

Elasticities		Specifications					
normal	corrected	Inventory	Varieties	Habits	Ads	RC Habits	RC Ads
-1.44	-1.96						
-1.26	-2.07	X					
-1.00	-1.38		X				
-2.19	-3.80			X			
-1.70	-2.50	X	X	X			
-1.76	-2.71	X	X	X	X		
-1.69	-2.58	X	X			X	
-1.56	-2.67	X	X			X	X

Note: RC indicates that the variable is specified to have a random coefficient.

Table 3.9 presents results for the nominal price case and the advertising sample, where advertising can be controlled for. The numbers differ from the full sample results, since the advertising sample is a subset of the full sample and has about 30% of its sample size. As for the full sample in table 3.8, inventories have an indecisive effect on the price elasticities. Price elasticities increase as expected with the endogeneity correction. Variety information and habits/state dependence operate as before.

The first five rows are for comparison to the full sample table 3.7, the interesting part are the last three rows. Row 6 contains all five endogeneity causes. The overall increase in the price elasticity from no cause to 5 causes controlled is 88%, as opposed to 36% if only the endogeneity correction is conducted. Consequently, the main quantitative result from table 3.7 still holds. Comparing rows 5 and 6 advertising has an increasing effect on the price elasticity, but the relative change is below 10% and much smaller than for varieties or habits. In rows 7 and 8 the same comparison can be made for advertising and habits with random coefficients. Note that only with the endogeneity correction present adding advertising with or without random coefficients increases elasticities as expected. Adding habits in rows 7 or 8 as random coefficients does not have any major impact as compared to a fixed coefficient with endogeneity correction, evident from comparison of row 5 versus 7 and 6 versus 8. In spite of this result without endogeneity correction a reduction upon introduction of a random coefficient for habits for the same rows is visible.

Table 3.10 shows the results for the advertising sample and efficiency prices. Variety information and habits work as before. Inventories have an increasing effect which contradicts our empirical hypothesis 4. In the first two rows the endogeneity correction decreases price elasticities which contradicts the empirical hypothesis 2. The overall increase from no cause to all 5 controlled is 94% which is in line with the former tables. Advertising has a decreasing effect and contradicts hypothesis 3, compare rows 5 and 6. In the nominal price case in table 3.9 the effect of advertising was in line with hypoth-

Table 3.10. Mean own price Elasticities - Advertising Sample - Efficiency Prices

Elasticities		Specifications					
normal	corrected	Inventory	Varieties	Habits	Ads	RC Habits	RC Ads
-2.47	-2.33						
-2.93	-2.65	X					
-1.69	-1.87		X				
-5.73	-6.46			X			
-4.50	-5.08	X	X	X			
-4.48	-4.80	X	X	X	X		
-4.43	-4.83	X	X			X	
-4.56	-4.65	X	X			X	X

Note: RC indicates that the variable is specified to have a random coefficient.

esis 3. Specifying advertising or habits as random coefficients does not add anything noteworthy.

Detailed results by brand can be taken from tables 3.12 to 3.17 in the appendix B. The brands are listed in descending order according to their frequency in the dataset. Every second column indicates the increase of the elasticity that comes from doing the endogeneity correction. The last row shows the mean over all brands, the numbers presented in the summary tables of the current section. The codes indicate the model estimated and the specification with variables can be taken from the tables with the parameter estimates. Generally, going from left to right more causes are controlled for and I see the same patterns as in the summary tables, now for each brand. Interestingly, for very simple models, elasticities take on unconvincing values, but this changes the more complex the model gets. Indeed, for the full sample in table 3.12 price elasticities take on convincing values. In the advertising sample, however, for some products even in the more complex models unintuitive positive or zero values of price elasticities persist. This may have partly caused the differing results between the full sample and the advertising sample discovered earlier.

3.4.2 Parameter Estimates and Estimation Details

Now I discuss the parameter estimates of the models from section 3.2. I restrict myself to standard panel models where the consumer has a constant taste over time so that each consumer gets only one draw of the random coefficient for his entire choice sequence in the sample.[51] Due to the high number of tables, they are placed at the end of the chapter in appendix part B. Statistical significance is indicated by the asterisks on the

[51] I estimated pooled mixed logit models with full taste flexibility across time so that each consumer got for each choice a new draw of the random coefficient. In this pooled case the estimation does not use the information that individuals do several choices, but treats each choice as being from a potentially

coefficient estimates. The asterisks mark significance to the following significance levels p: * for p<0.05, ** for p<0.01, and *** for p<0.001.

The sample sizes indicate the estimation sample being used: for the full sample $n = 269,942$ and for the advertising sample $n = 77,112$.

The table headline specifies the theme of the estimated specifications and whether prices are measured as efficiency prices. Each model is estimated twice with and without endogeneity correction (indicated by "IV" in the headline) so that each two columns can be compared pairwise. All specifications allow for multiple random coefficients, therefore the headline "Multiple RC". In the literature there is no agreement whether to enter nominal price or efficiency price defined as price divided by contents in liters or kg of the detergent package, so I estimate both versions.

Variables To understand the tables better, I comment on the variables found therein. The control function enters as variable `xi` into the specifications, followed by a number coding the anonymous brand name. Due to the high number of variables it is only indicated in the regression whether an endogeneity correction is conducted by `endogeneity correction`. Variables `duration,duration2, idurhh` control for inventory effects. Whether the interaction of price with household size and of income group dummies with price is included in the estimation is signaled by `observed demography`. The income group dummies differentiates between 13 income groups. TV advertising variables contain the keyword `ad`. The keywords `bdum` and `brand dummies` indicate when brand dummies are present in the regression. `feature` and `display` measure retail activity. Variables `inhp, inhl` refer to the amount of detergent (in kg or liters) in the package. For a more detailed explanation on variables see section 3.3.

Instrumental Relevance To estimate the instruments I use the approach explained in section 3.3.3, where also the exogeneity assumptions are introduced in detail. The control function OLS regressions per brand all have very high F test values, indicating good instrumental quality for a linear model.

Model Codes I briefly explain the model code used to identify a model. The code has the form `aestVWXXXYZ`. I comment on each letter: `a` marks the advertising sample, without `a` it is the full sample. `e` indicates prices enter as efficiency prices, without `e` prices enter nominally. `st` or only `s` have no meaning. `V` takes integer values 1,. . . ,5 and defines the sample of retail chains used, default is 4.[52] `W` defines the minimal number of alternatives that are available at each purchase occasion, default is 4. `XXX` takes the values `GLb` and `GLc` and marks the type of the state dependence. `GLb` is the one period lag dummy, `GLc` the 10 period lag dummy vector. If `XXX` is none of the values, no state dependence variable is contained in the model. `Y` specifies the remaining variable

different individual. Apart from affecting estimation time negatively and absolute coefficient levels, all later results were unaffected. That is why I relinquish to present the results.

[52] For a definition of retail chains used see section 3.4.3.1.

specification. `Z` specifies whether an endogeneity correction is conducted, where `a` means yes and `b` means no.

In the following I summarize results for the specifications with nominal prices and thereafter I will look at specifications with efficiency prices.

3.4.3 Specifications with nominal Prices

Table 3.18 contains the basic random coefficient specification estimated on the full sample. Each specification is conducted once without and with endogeneity correction to reveal the effect on the price parameter. Thus, odd columns contain the specifications without and even columns with endogeneity correction. In the upcoming paragraphs I discuss each cause.

Retail activity Adding the observed retail activity variables (`feature, display`) in column 3 and 4 in table 3.18 has no major impact on the price coefficient. As expected from the literature, the price coefficient increases clearly when doing the endogeneity correction. The last two columns reveal that the most prominent specification in the literature (with brand dummies) shows the highest impact with the correction. Hence, the observed retail activity does not seem to capture anything relevant for price reactions. In general, once brand dummies are introduced, the effect of the endogeneity correction is intensified.

Inventories It is apparent from table 3.19 that individual household inventory levels proxied by the variable `duration, duration2` are not significant, but `ihurhh`, the interaction of duration and household size, is significant. The elasticities from before reveal that this effect is not large. This result holds for both the full sample in columns 1 and 2 and weakly for the advertising sample in columns 3 and 4.

Variety information In table 3.19 variety information (variables: `color`, `sensitiv`, `konzentrat`, `gimmick`, `extrasize`, `bigpack`) is included that cannot be captured by brand dummies, as variety information has intra brand variation as explained in section 3.2. In columns 5 to 8 most variety information variables are highly significant for both samples. It is evident that the endogeneity correction is still at work, but the price coefficient is strongly reduced. Note that the price coefficient in column 6 with variety information and endogeneity correction is lower than in column 7 of table 3.18 without endogeneity correction (the standard specification in the literature).

Habits/state dependence The role of habits is undoubted in the literature on consumer choice. That is why it is surprising why this has not received more attention when analyzing price endogeneity. Table 3.20 shows the effect when including consecutively more flexible patterns for state dependence. It is obvious from the table that the price coefficient is increasing in the flexibility of the specification. Two versions are estimated: with a one lag dummy `GLdumB` and with ten dummies `mdum*`. Variables `GLdumB` and up to the eighth lag mdum* are highly significant and big in magnitude. The results hold for the full sample in the first four columns and for the advertising sample in the last four columns.

Advertising Unobserved advertising, such as the precise TV advertising information of the employed dataset, is always mentioned when considering causes for price endogeneity. Table 3.21 presents the results after adding advertising. The first four columns show the results for relative advertising pressure in the last 56 days prior to purchase. The variable is not significant. When adding partial advertising pressure per week in the last four columns I get significance with counterintuitive signs. The results are not stable.

In tables 3.22, 3.23 the specifications are displayed, when endogeneity causes are controlled for simultaneously. These tables are hard to interpret and basically all effects mentioned before come up again. Noteworthy is that duration is highly significant in table 3.22 for the first two columns where the full sample is used. Overall, using random coefficients for duration, habits or advertising delivers no new insights. The models in these tables are used to estimate the price elasticities which are easier to interpret.

3.4.3.1 Robustness Checks

Stability across samples As the estimation with endogeneity correction is not conducted on all 170 retail chains, there may be worries that the data are very different across different samples. Recall, I split the sample into small retailers that deliver the Hausman instruments for the large retailers, see section 3.2.4. Table 3.24 shows the results for the literature benchmark model. The estimated specification is ordered increasingly by sample code, the third digit in the model code. Sample definitions are as follows: 1: all retailers. 2: 15 biggest retailers. 3: sample 2 ex biggest retailer. 4: 10 biggest retailers. 5: sample 4 ex biggest retailer. For the first three samples no instruments are constructed so that there are only versions without endogeneity correction in the first three columns. For the last two samples I present estimates with and without endogeneity correction. The results illustrate the high stability of the price coefficient. In samples 3 and 5 the biggest German discounter that sells only products that are never advertised on TV is excluded. There is hardly any difference between columns 4 and 6 or 5 and 7, respectively, although the biggest discounter is excluded in columns 6 and 7.

Sample Selection The tables presented so far are based on choice decisions where the consumer had at least 4 alternatives to choose from in the store. One may worry that unattractive alternatives that were not chosen frequently are missing and the sample is biased towards the chosen alternatives. To analyze this problem I re-estimated one specification with a different minimal amount of alternatives. This specification is the literature benchmark case plus the state dependence dummy that had by far the highest impact on price elasticities. Table 3.25 presents the results. In the first two columns, at least 4 alternatives had to be available, in the next two columns at least 5 and in the last two columns at least 6 alternatives had to be available in a choice situation per consumer so that the situation was considered in the estimation. I observe a high stability in price effects despite strongly decreasing sample sizes.

3.4.3.2 Comparison to PT Setup

To compare the used dataset to the literature, I mimic the setup of Petrin and Train (2006). Remember that different to the standard literature, I operate on daily data. To match the PT setup, I aggregate the data to weekly observations, keep only liquid detergents and construct Hausman instruments. Table 3.26 shows the results analogue to table 3.18. The results are not convincing. Prices are hardly significant and the endogeneity correction does hardly matter. Maybe by aggregation the Hausman instruments lose their bite due to the amount of measurement error that enters.

3.4.4 Specifications with efficiency Prices

As it is disputable how consumers evaluate prices when buying detergents, I re-estimated all models from the previous section when prices are measured as efficiency prices. In the following I briefly discuss the results for efficiency prices relative to the nominal price case.

Concerning observed retail activity variables there are no new findings when looking at efficiency prices. Interestingly, the pattern of increasing the price coefficient is still present when I conduct the endogeneity correction. Despite the statistical significance, the magnitude of the effect is almost negligible, as can be seen by comparing the columns of table 3.27 pairwise.

Individual inventories are now highly significant and it is not the interaction with household size, but the duration variables themselves as can be seen from table 3.28. Variety characteristics are again highly significant and I get the same result for the inclusion of variety information as in the nominal price case: The price coefficient drops sharply upon inclusion of the variety information.

Table 3.29 affirms the earlier results for state dependence dummies of the nominal price case: they are highly significant and have an increasing effect on the price coefficient. The habit coefficient magnitude is comparably high (2-3 times the price coefficient), but as opposed to the nominal price case the standard deviation of the price coefficient is halved compared to the benchmark model from table 3.27.

In table 3.30 the results for advertising are more promising than in the nominal price case. Advertising is highly significant without brand dummies present for both advertising variables with the right sign. With brand dummies present, the effect is almost wiped out. It seems that brand dummies pick up the effect of advertising in the data.

In tables 3.31, 3.32 all endogeneity causes are combined and additional variables get random coefficients. Basically, all causes behave as before and no new insights can be taken from these tables, as they are hard to interpret, but they are the basis of the price elasticity estimations.

3.4.5 Relation to Literature

If I stick to nominal prices the direction of the price change caused by the endogeneity correction is broadly in line with the literature (i.e. Chintagunta, Dubé, and Goh (2005), Petrin and Train (2006)), but the size of the price parameter is absolutely smaller. In addition, also in relative terms the results differ: average elasticities change by 54% in this work versus a change of 20-35% in the paper of Petrin and Train when a model with endogeneity correction is compared to the benchmark case.[53] When adding more endogeneity causes, I get increases of up to 108%, which is far more than estimated before.

Concerning advertising I can only partly affirm the results of Erdem and Keane (1996) who do not find any significant impact of advertising in a reduced form model of a similar kind as in this chapter. I find that advertising matters in a model without brand dummies and efficiency prices, but loses significance once brand dummies are present. That this happens despite the excellent quality of the advertising information in the employed dataset reveals that this reduced form model is either not well suited to capture advertising effects or that they may not exist. Perhaps the following argument sheds light on the result: In the long run, advertising can be seen as influencing brand perception of the consumer and thereby modify average brand dummies in the population. This process is clearly outside the current model and by adding brand dummies the model simply absorbs this long run effect.

Nevo (2001, p. 326 footnote 24) finds for market level data that once the usual endogeneity corrections are done, effects of advertising for price sensitivities are negligible. I cannot confirm this. In this work, introduction of brand dummies renders parameter estimates of advertising variables insignificant. I want to stress that this could be due to the nature of aggregation in the model, as the Nevo paper does not have individual level consumer data. Hence, it may be the case that the weakened influence of advertising on price elasticities caused by the endogeneity correction is only present on higher aggregation levels, i.e. for market level data. Moreover, Nevo has national advertising data in mind. In chapter 4 of this dissertation, I do the first step to see whether local advertising matters at all on the market level and this could be extended to study implied price elasticities.

[53] Check the earlier discussion of table 3.7 and table 3 of Petrin and Train (2006).

However, with the endogeneity correction under the nominal price regime, advertising increases price elasticities as predicted, contradicting the results of Erdem and Keane (1996) and Nevo (2001). Interestingly, without endogeneity correction, the impact of advertising on price elasticities is unclear.

Relating to the overdispersion results of Chintagunta, Dubé, and Goh (2005) I find that the price endogeneity correction does not always reduce the standard deviations of the random coefficients on price. This finding is in line with the arguments laid out in Horsky, Misra, and Nelson (2006) who only find a reduction once stated-preference data are available. In my work, introduction of the state dependence variables reduces the standard deviation of the price coefficient, but only for the efficiency price case.

3.5 Conclusion

In this chapter I have estimated several discrete choice specifications to assess the effect of five endogeneity causes on biasing price effects, measured as price elasticities: Variety information, unobserved retail activity, advertising, individual inventories and habits/state dependence. Partly, those causes have so far been ignored in previous work on price endogeneity. I compare, in particular, the price endogeneity correction that is to control for unobserved retail activity to the other four causes. The employed novel dataset is rich in detail and allows to evaluate the merits of each cause relative to the literature benchmark case. New to the literature, I can estimate a model with all five causes being considered simultaneously. I am able to find other causes that have a larger impact than the endogeneity correction suggested by Petrin and Train (2006). Related to the correction, the suggested construction of Hausman instruments for a nationwide sample combined with the split sample idea is a successful approach that can clearly compensate the lack of wholesale prices for the national sample.

Moreover, I show that it depends on the kind of prices the consumers look at to assess the usefulness of the endogeneity correction. The result is twofold in this respect:

1. For nominal prices as price variable, the price endogeneity correction works. The relative change induced for a standard specification is even higher than in previous results in the literature. TV advertising and individual inventories are neutral with respect to the correction and generally of minor importance, being hardly statistically significant. Advertising slightly increases price elasticities, especially if the endogeneity correction is included. Variety information heavily dampens the price effects. State dependence amplifies price effects. The net effect is an overall high increase of price effects when comparing the situation of no cause controlled to all five causes controlled. Additionally, the price effect is clearly higher when all five causes are controlled for than if only the endogeneity correction is applied.

2. For efficiency prices as price variable the correction has a lower effect than in the nominal price case. TV advertising and individual inventories are partly statistically significant, but their overall impact on price effects is negligible. Variety information and habits work as in the nominal case. The five causes versus the sole endogeneity correction show a much bigger impact on price elasticities, just as in the nominal price case.

Importantly, in both cases and across all specifications, controlling for variety information dampens price effects and adding state dependence raises price effects intensely. Inventories and advertising play a minor role. Hence, I conclude that two other important causes have to be considered when studying price endogeneity alongside unobserved retail activity.

I infer that the usefulness of the endogeneity correction depends on the amount of information available. If brand varieties exist but the information is omitted, the endogeneity correction alone may be overestimating the effect of prices, as the inclusion of variety information would counteract it. Adding the state dependence information is crucial given its economic magnitude. Thus, if I consider my results for the combined five causes, using solely the endogeneity correction in a simple specification would underestimate price effects.

I find moreover that in both cases, TV advertising is neutral to the correction and of minor importance. Additionally, its effect is not consistent across the two price regimes. This clarifies a repeatedly stated presumption in the literature. There, the endogeneity correction is introduced to counteract - as one leading example - the lack of advertising information. Advertising effects require a further investigation in a more suitable model.

Inventory effects that have proven important in recent structural models, are not identified as important in the model type used in this chapter.

Regarding the scope of this work, it is realistic that the results can be carried over from the detergent category studied to other storable non-food consumer products.

Concluding, the endogeneity correction merely resolves one of several endogeneity causes. Relying on price effects that are calculated without considering the causes identified here - variety information and habit formation - will seriously underestimate price effects.

Appendix

Appendix A: Model Details - Estimation of Pooled Case

The following lines derive the analogue log likelihood formulas of the panel case for the pooled case. Intuitively, the difference between both cases is the following: In the panel case the consumer is not allowed to have random taste switches across his choice sequence, whereas in the pooled case, the consumer suffers from a taste shock each time he conducts a purchase. The shocks are only linked through the common and constant distribution of η_i from which they are drawn. Both cases represent polar cases and intermediate cases can be constructed, but this is beyond the scope of this chapter, for more consult Train's (2003) mixed logit chapter 6.

I start again from equation (3.15). The conditional likelihood is given by

$$L_{it|\eta_i}(\theta) = \prod_{j=1}^{J+1} P_{ijt|\eta_i}^{d_{ijt}}(\theta) \tag{3.39}$$

with d_{ijt} as before. The unconditional likelihood for a purchase at time t by consumer i is:

$$L_{it}(\theta) = \int L_{it|\eta_i}(\theta) f(\eta_i) d\eta_i \tag{3.40}$$

with $f(\eta_i)$ as before. Then the function to maximize given a sample of I consumers for T periods is:

$$LL(\theta) = \sum_{t=1}^{T} \sum_{i=1}^{I} d(i,t) log[L_{it}(\theta)] \tag{3.41}$$

where $d(i,t)$ is an indicator function $d(i,t) = \mathbf{1}\{i \text{ did a purchase at time } t\}$.

Appendix B: Tables

Overview of Variables

Table 3.11. Overview of Variables

Variable	Description
Variety Characteristics	
bigpack	Dummy for detergent is sold as bundle of multiple units of the same size
color	Dummy for detergent is suitable for color fabrics
extra_size	Dummy for detergent is sold in a package increased by up to 33% relative to the regular size

Table 3.11. (continued...)

Variable	Description
`gimmick`	Dummy for detergent is sold with a gimmick (CD, Cleanser, . . .)
`konzentrat`	Dummy for detergent is concentrated
`liquid`	Dummy for product is liquid detergent
`sensitiv`	Dummy for detergent is sensitiv (recommended for allergic persons)
Inventory	
`duration`	Duration in days since last purchase
`duration2`	Squared duration
`idurhh`	Interaction of duration and household size in persons, capture effect for differently sized households
Retail activity	
`display`	Product is on display: the brand was promoted via a display, e.g. lobby, aisle (front, end, back) and specialty/shipper
`feature`	Product is featured: the brand of the product was featured in the newspaper circulars for the store
`handbill`	Product is hand billed
`prceflag`	Product is price flagged
State dependence	
`GLdumA`	Dummy for previous purchase had the same brand as current alternative, including no-purchase incidents
`GLdumB`	Dummy for previous purchase had the same brand as current alternative, skipping no-purchase incidents (more intuitive)
`mdum*`	Is a series of `GLdumB` variables constructed up to lag $*$, thus looking at purchased brand that are $*$ purchases ago
Advertising	
`countc140ad`	Number of advertising contacts consumer had in last 140 days prior to purchase incident with advertisement of the same brand as the faced alternative (variable ends with `ad`)
`liqc56adr`	Relative pressure of advertising contacts for liquid products consumer had in last 56 days prior to purchase incident with advertisement of the same brand as the faced alternative relative to all contacts he had (variable ends with `adr`)
`countc140adpr`	Partial relative advertising pressure consumer had in time window 140 to 126 days prior to the purchase incident, windows are always 14 days wide (variable ends with `adpr`)

Table 3.11. (continued...)

Variable	Description
Others	
`price`	Purchase price, either in nominal or efficiency units (=nominal price divided by contents, i.e. variable `inh`)
`observed demography`	Yes/No signals whether the interaction of price with household size and of income group dummies and price is included in the estimation. Additionally, income group dummies differentiate between 13 income groups
`brand dummies`	This is a set of brand dummies that is inserted in the specification indicated by Yes/No
`endogeneity correction`	To save space, `xi3, xi5,` . . . , the control function residuals are not displayed but it is indicated by Yes/No whether the endogeneity correction is performed
`inh`	Packet size (l or kg)
`inhl`	Same as `inh`, but contains only for liquid detergents (l)
`inhp`	Same as `inh`, but contains only for non-liquid detergents (kg)

Elasticity Estimates

Table 3.12. Own price Elasticities per brand - Full Sample

Model	s445		s448		s449		s44GLb5		s44GLb10	
Brand	norm	corr	norm	corr	norm	corr	norm	corr	norm	corr
16	-1.36	-2.01	-1.36	-1.97	-0.97	-1.47	-2.28	-3.70	-1.73	-2.69
55	-1.68	-2.49	-1.71	-2.46	-1.22	-1.80	-2.95	-4.56	-2.15	-3.29
3	-1.63	-2.74	-1.64	-2.74	-1.18	-1.93	-3.30	-6.15	-2.51	-4.22
40	-1.52	-3.05	-1.59	-2.94	-1.03	-2.04	-3.71	-7.26	-2.61	-4.63
41	-1.77	-2.73	-1.66	-2.34	-1.30	-1.91	-2.70	-4.37	-2.20	-3.07
67	-1.64	-3.06	-1.65	-2.93	-1.12	-2.17	-3.41	-6.33	-2.56	-4.45
13	-1.27	-2.05	-1.39	-1.61	-1.07	-1.37	-2.24	-3.08	-1.49	-2.22
8	-1.77	-2.68	-1.79	-2.74	-1.27	-1.97	-2.91	-5.11	-2.17	-3.05
57	-1.57	-2.62	-1.56	-2.56	-1.02	-1.85	-2.99	-5.18	-2.17	-3.71
100	-1.73	-2.89	-1.74	-2.71	-1.21	-1.87	-3.21	-4.53	-2.41	-3.92
36	-1.65	-2.77	-1.67	-2.68	-1.11	-2.02	-3.25	-5.86	-2.43	-4.26
11	-1.60	-2.57	-1.63	-2.58	-1.16	-1.90	-2.86	-5.03	-2.16	-3.48
22	-1.56	-2.18	-1.47	-2.13	-1.22	-1.55	-2.38	-4.56	-1.97	-3.06
10	-1.64	-2.51	-1.65	-2.51	-1.14	-1.81	-2.91	-4.69	-2.12	-3.19
6	-1.76	-2.68	-1.71	-2.40	-1.17	-1.98	-2.43	-3.78	-1.93	-3.07
Mean	-1.46	-2.25	-1.46	-2.19	-1.04	-1.62	-2.56	-4.26	-1.93	-3.04

Table 3.13. Own price Elasticities per brand - Advertising Sample

Model	as445		as448		as449		as44GLb5		as44GLb10	
Brand	norm	corr	norm	corr	norm	corr	norm	corr	norm	corr
16	-1.43	-1.87	-1.24	-1.90	-0.99	-1.33	-2.08	-3.50	-1.62	-2.32
55	-1.71	-2.26	-1.63	-2.49	-1.22	-1.61	-2.58	-4.28	-2.01	-2.85
3	-1.73	-2.69	-1.38	-3.04	-1.22	-1.87	-3.00	-6.16	-2.39	-3.72
40	-1.77	-2.70	-1.48	-3.07	-1.06	-1.84	-3.16	-6.57	-2.47	-3.92
41	-0.18	-0.15	-0.16	-0.07	-0.10	-0.13	-0.00	0.01	-0.03	-0.00
67	0.60	0.61	0.66	0.43	0.49	0.49	0.07	0.01	0.11	0.01
13	-1.51	-1.80	-1.40	-1.96	-1.09	-1.35	-2.04	-3.02	-1.45	-2.00
8	-1.77	-2.30	-1.42	-2.78	-1.25	-1.65	-2.75	-4.76	-2.24	-3.31
57	-1.73	-2.51	-1.44	-2.68	-1.05	-1.63	-2.42	-4.39	-1.95	-2.97
100	-1.59	-2.75	-1.68	-2.80	-1.32	-1.84	-2.72	-4.30	-1.79	-3.41
36	-1.81	-2.76	-1.50	-3.04	-1.10	-1.80	-3.16	-5.44	-2.32	-3.35
11	-1.72	-2.49	-1.49	-2.73	-1.18	-1.66	-2.75	-4.98	-2.16	-3.18
22	-1.57	-2.25	-1.64	-2.35	-1.25	-1.67	-2.67	-3.48	-1.84	-2.77
10	-1.51	-2.36	-1.46	-2.56	-1.15	-1.61	-2.51	-4.09	-1.76	-2.80
6	-1.78	-2.45	-1.48	-2.29	-1.19	-1.72	-2.48	-4.27	-2.01	-2.86
Mean	-1.44	-1.96	-1.26	-2.07	-1.00	-1.38	-2.19	-3.80	-1.70	-2.50

Table 3.14. Own price Elasticities per brand - Advertising Sample

Model	as44GLb10		as44GLb22		as44GLb42		as44GLb52	
Brand	norm	corr	norm	corr	norm	corr	norm	corr
16	-1.62	-2.32	-1.68	-2.53	-1.61	-2.40	-1.50	-2.49
55	-2.01	-2.85	-2.06	-3.05	-2.00	-2.93	-1.82	-3.00
3	-2.39	-3.72	-2.47	-4.15	-2.27	-3.79	-2.03	-3.94
40	-2.47	-3.92	-2.47	-4.29	-2.40	-3.99	-1.95	-4.18
41	-0.03	-0.00	-0.03	-0.03	-0.03	-0.02	-0.06	-0.03
67	0.11	0.01	0.11	0.11	0.11	0.04	0.22	0.13
13	-1.45	-2.00	-1.55	-2.16	-1.60	-2.14	-1.56	-2.19
8	-2.24	-3.31	-2.24	-3.12	-2.06	-3.27	-1.99	-3.38
57	-1.95	-2.97	-1.97	-3.50	-1.81	-3.11	-1.77	-3.17
100	-1.79	-3.41	-1.93	-3.48	-2.15	-3.51	-2.01	-3.63
36	-2.32	-3.35	-2.48	-3.77	-2.29	-3.73	-2.09	-3.80
11	-2.16	-3.18	-2.24	-3.42	-2.09	-3.26	-1.95	-3.33
22	-1.84	-2.77	-1.59	-2.33	-2.12	-3.05	-1.97	-3.15
10	-1.76	-2.80	-1.93	-2.94	-1.96	-2.91	-1.81	-2.93
6	-2.01	-2.86	-1.95	-2.95	-1.97	-2.79	-1.85	-2.96
Mean	-1.70	-2.50	-1.76	-2.71	-1.69	-2.58	-1.56	-2.67

Table 3.15. Own price Elasticities per brand - Full Sample - Efficiency Prices

Model	es445		es448		es449		es44GLb5		es44GLb10	
Brand	norm	corr	norm	corr	norm	corr	norm	corr	norm	corr
16	-2.03	-2.28	-2.16	-2.29	-1.71	-1.90	-4.92	-5.28	-4.07	-4.73
55	-2.51	-2.73	-2.70	-2.86	-2.11	-2.33	-6.01	-6.45	-4.95	-5.75
3	-2.72	-3.22	-3.01	-3.30	-2.16	-2.51	-8.34	-9.17	-6.81	-8.34
40	-3.26	-3.89	-3.71	-3.80	-2.12	-2.63	-12.07	-13.15	-9.71	-11.79
41	-2.53	-2.85	-2.72	-2.85	-2.12	-2.35	-5.69	-6.76	-4.83	-5.35
67	-2.71	-3.23	-3.02	-3.28	-2.06	-2.42	-8.38	-9.06	-6.73	-8.07
13	-1.91	-1.95	-2.00	-2.22	-1.61	-1.65	-3.45	-3.40	-3.08	-3.27
8	-2.65	-2.91	-2.84	-3.02	-2.17	-2.44	-6.78	-7.33	-5.57	-6.57
57	-2.53	-2.89	-2.76	-3.03	-1.99	-2.30	-6.81	-7.08	-5.39	-6.53
100	-2.84	-3.22	-3.16	-3.30	-2.18	-2.50	-7.38	-7.98	-6.01	-7.07
36	-2.83	-3.18	-3.11	-3.28	-2.02	-2.39	-8.21	-8.87	-6.61	-8.01
11	-2.50	-2.84	-2.70	-2.91	-2.06	-2.31	-6.55	-7.02	-5.37	-6.30
22	-2.59	-2.78	-2.87	-2.97	-2.20	-2.36	-6.11	-6.65	-5.06	-5.89
10	-2.59	-2.84	-2.74	-2.93	-2.15	-2.39	-6.03	-6.45	-4.95	-5.78
6	-2.45	-2.75	-2.63	-2.79	-2.09	-2.32	-5.33	-6.36	-4.63	-5.45
Mean	-2.25	-2.54	-2.42	-2.58	-1.85	-2.07	-5.78	-6.23	-4.75	-5.57

Table 3.16. Own price Elasticities per brand - Advertising Sample - Efficiency Prices

Model	aes445		aes448		aes449		aes44GLb5		aes44GLb10	
Brand	norm	corr	norm	corr	norm	corr	norm	corr	norm	corr
16	-2.33	-2.17	-2.68	-2.46	-1.62	-1.79	-5.08	-5.69	-4.04	-4.55
55	-2.83	-2.78	-3.44	-3.16	-2.14	-2.31	-6.17	-6.88	-5.01	-5.50
3	-3.43	-3.27	-4.50	-3.90	-2.13	-2.44	-9.26	-10.86	-7.07	-8.20
40	-4.41	-3.94	-5.68	-4.67	-2.48	-2.75	-12.76	-14.91	-9.74	-11.31
41	-0.25	-0.26	-0.28	-0.40	-0.31	-0.29	-0.02	-0.01	-0.05	-0.03
67	1.51	1.43	2.28	2.32	1.36	1.39	0.98	1.06	1.64	1.37
13	-2.21	-2.17	-2.32	-2.24	-1.81	-1.91	-3.56	-3.83	-3.10	-3.25
8	-2.83	-2.78	-3.59	-3.38	-2.03	-2.22	-7.13	-8.17	-5.52	-6.32
57	-3.09	-2.99	-3.78	-3.44	-2.10	-2.36	-7.19	-8.07	-5.31	-5.59
100	-3.35	-3.08	-4.07	-4.04	-2.17	-2.39	-7.77	-8.13	-5.91	-6.48
36	-3.23	-3.05	-4.14	-3.69	-2.01	-2.27	-8.76	-10.03	-6.81	-7.88
11	-2.99	-2.91	-3.44	-3.18	-2.02	-2.29	-7.17	-8.21	-5.60	-6.38
22	-3.09	-2.93	-3.62	-3.48	-2.22	-2.41	-6.31	-7.06	-5.05	-5.45
10	-2.73	-2.65	-3.18	-2.97	-1.98	-2.19	-6.02	-6.47	-4.70	-5.31
6	-2.61	-2.51	-3.05	-2.85	-1.89	-2.09	-6.23	-7.11	-4.68	-5.59
Mean	-2.47	-2.33	-2.93	-2.65	-1.69	-1.87	-5.73	-6.46	-4.50	-5.08

Table 3.17. Own price Elasticities per brand - Advertising Sample - Efficiency Prices

Model	aes44GLb10		aes44GLb22		aes44GLb42		aes44GLb52	
Brand	norm	corr	norm	corr	norm	corr	norm	corr
16	-4.04	-4.55	-4.03	-4.31	-3.97	-4.33	-4.07	-4.17
55	-5.01	-5.50	-4.91	-5.13	-4.86	-5.29	-5.03	-5.09
3	-7.07	-8.20	-7.09	-7.61	-6.93	-7.61	-7.20	-7.28
40	-9.74	-11.31	-9.74	-10.45	-9.65	-10.26	-10.10	-9.80
41	-0.05	-0.03	-0.04	-0.03	-0.03	-0.04	-0.04	-0.03
67	1.64	1.37	1.40	1.18	1.19	1.15	1.30	1.10
13	-3.10	-3.25	-2.97	-3.28	-3.11	-3.39	-3.14	-3.28
8	-5.52	-6.32	-5.52	-5.91	-5.44	-5.93	-5.65	-5.71
57	-5.31	-5.59	-5.07	-6.06	-5.01	-5.94	-5.15	-5.76
100	-5.91	-6.48	-5.81	-6.44	-6.09	-6.48	-6.15	-6.24
36	-6.81	-7.88	-6.99	-7.21	-6.83	-7.43	-7.03	-7.08
11	-5.60	-6.38	-5.60	-5.95	-5.47	-5.97	-5.72	-5.79
22	-5.05	-5.45	-5.03	-5.33	-5.04	-5.49	-5.17	-5.30
10	-4.70	-5.31	-4.75	-5.02	-4.73	-5.15	-4.85	-5.04
6	-4.68	-5.59	-4.70	-5.23	-4.76	-5.07	-4.82	-4.99
Mean	-4.50	-5.08	-4.48	-4.80	-4.43	-4.83	-4.56	-4.65

Parameter Estimates - Nominal Prices

Table 3.18. Multiple RC IV Mixed Logit - Basic Models

Model	s441b	s441a	s443b	s443a	s444b	s444a	s445b	s445a
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Mean								
price	–1.10***	–1.16***	–1.14***	–1.19***	–1.43***	–1.72***	–1.44***	–1.69***
inhp	–.61***	–.50***	–.61***	–.47***	–.20***	.22***	–.28***	.17***
inhl	.38***	.52***	.34***	.49***	.64***	1.05***	.52***	.97***
iprhh	.07***	.07***	.07***	.07***	.08***	.09***	.08***	.07***
liquid	–3.01***	–3.05***	–2.95***	–2.98***	–2.74***	–2.74***	–2.69***	–2.79***
feature			.02	.09			.05	.12
display			.28**	.33**			.25**	.28***
Standard Deviation								
price	.56***	.55***	.57***	.56***	.52***	.51***	.54***	.50***
liquid	–1.72***	–1.72***	–1.70***	–1.70***	–1.70***	1.77***	1.72***	1.80***
feature			1.23***	1.24***			.84***	.75***
display			–1.01***	–.99***			–.72***	–.66***
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	No	No	No	No	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	269,942	269,942	269,942	269,942	269,942	269,942
Log-Likelihood	–21,498	–21,396	–21,415	–21,304	–20,945	–20,755	–20,897	–20,696

Note: Asterisks indicate significance levels. Random coefficients are independent.

Table 3.19. Multiple RC IV Mixed Logit - Inventory and Varieties

Model	s448b	s448a	as448b	as448a	s449b	s449a	as449b	as449a
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Model								
price	–1.44***	–1.57***	–1.42***	–1.53***	–1.04***	–1.26***	–1.03***	–1.18***
inhp	–.26***	.18***	–.34***	.11	–.37***	–.07	–.45***	–.20*
inhl	.51***	.97***	.73***	1.15***	.35***	.69***	.43***	.71***
iprhh	.10***	.08***	.08***	.07***	.08***	.07***	.05***	.04***
liquid	–2.62***	–2.76***	–3.07***	–3.06***	–2.77***	–2.96***	–3.12***	–3.22***
feature	.06	.13	–.12	.20	.03	–.04	–.30	–.18
display	.22**	.38***	.15	.00	.30***	.28***	.09	–.03
idurhh	–.01***	–.00**	–.01*	–.01				
duration	.00	.00	–.00	.01				
duration2	.00	–.00	–.00	–.00				
color					–.44***	–.34***	–.47***	–.43***
sensitiv					–.60***	–.48***	–.43	–.45
konzentrat					–.94***	–.86***	–.80***	–.79***
gimmick					1.24***	1.15***	1.32***	1.10***
extrasize					.21	.19	–.21	–.11
bigpack					2.19***	1.90***	1.32**	1.68***
Standard Deviation								
price	.53***	–.50***	.51***	.46***	.44***	.42***	.43***	.42***
liquid	1.71***	1.76***	1.65***	1.66***	–1.79***	1.93***	1.96***	1.97***
feature	.86***	.62**	1.47***	–.88**	.39	.85***	1.75***	1.51***
display	–.74***	.14	–.32	.24	.34*	–.29*	.37	.79**
duration	–.02***	.02**	.04***	.05***				
duration2	.00	–.00	.00	–.00				
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	77,112	77,112	269,942	269,942	77,112	77,112
Log-Likelihood	–20,869	–20,696	–5,940	–5,907	–20,472	–20,338	–5,845	–5,818

Note: Asterisks indicate significance levels. Random coefficients are independent.

Table 3.20. Multiple RC IV Mixed Logit - Habit Formation

Model	s44GLb5b Coef.	s44GLb5a Coef.	s44GLc5b Coef.	s44GLc5a Coef.	as44GLb5b Coef.	as44GLb5a Coef.	as44GLc5b Coef.	as44GLc5a Coef.
Mean								
price	–1.79***	–2.27***	–1.89***	–2.41***	–1.49***	–1.93***	–1.56***	–2.04***
inhp	–.19***	.58***	–.15**	.68***	–.33***	.38***	–.35***	.38**
inhl	.97***	1.61***	.97***	1.78***	1.04***	1.80***	1.17***	1.90***
iprhh	.08***	.07***	.07***	.07***	–.01	.01	–.00	–.01
liquid	–3.57***	–3.44***	–3.63***	–3.66***	–3.87***	–4.00***	–4.35***	–4.46***
feature	–.03	.08	–.01	.06	.10	.14	.00	.27
display	.36***	.37***	.47***	.46***	.04	.07	.23	–.06
GLdumB	7.27***	7.30***			7.17***	7.16***		
mdum1L			5.53***	5.59***			4.96***	5.16***
mdum2L			4.06***	3.95***			3.99***	4.15***
mdum3L			1.89***	2.04***			2.10***	2.22***
mdum4L			1.90***	1.80***			1.23***	1.39***
mdum5L			1.47***	1.62***			2.21***	2.51***
mdum6L			1.27***	1.28***			.99*	.94*
mdum7L			1.37***	1.32***			1.58***	1.47**
mdum8L			1.26***	1.24***			1.87***	1.65***
mdum9L			–.21	–.12			1.30**	1.65**
mdum10L			.95**	.74*			.43	.31
Standard Deviation								
price	.51***	.44***	.51***	.48***	.50***	.41***	.53***	.46***
liquid	1.52***	1.49***	1.51***	1.48***	1.47***	1.43***	1.53***	1.60***
feature	1.10***	.97***	1.04***	.98***	–.91***	.83*	1.14***	–.67*
display	.64**	.65**	–.29	–.14	–1.10***	–.91**	–.34	–1.05**
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	269,942	269,942	77,112	77,112	77,112	77,112
Log-Likelihood	–14,276	–13,936	–12,591	–12,181	–4,109	–4,023	–3,448	–3,342

Note: Asterisks indicate significance levels. Random coefficients are independent.

Table 3.21. Multiple RC IV Mixed Logit - Advertising

Model	as4420b Coef.	as4420a Coef.	as4421b Coef.	as4421a Coef.	as4423b Coef.	as4423a Coef.	as4424b Coef.	as4424a Coef.
Mean								
price	–1.10***	–1.33***	–1.35***	–1.63***	–1.28***	–1.31***	–1.40***	–1.62***
inhp	–.58***	–.43***	–.35***	.04	–.53***	–.36***	–.34***	.02
inhl	.57***	.74***	.77***	.99***	.64***	.80***	.85***	1.12***
iprhh	.07***	.05*	.05***	.06***	.02	.04**	.06**	.05**
liquid	–3.31***	–3.40***	–3.26***	–3.07***	–3.34***	–3.47***	–3.41***	–3.23***
feature	.34	–.34	–.06	–.21	–.43	.50***	–.03	–.33
display	–.18	.27*	.14	–.01	.18	–.02	.06	–.00
countc56adr	–.36	.57	.01	–.17				
countc56adpr					.99**	1.00**	.79*	.57
countc84adpr					–2.05**	–1.50*	–1.02	–2.24**
countc98adpr					.45	.94*	.24	.18
Standard Deviation								
price	.56***	.55***	.53***	.49***	.57***	.57***	.56***	.51***
liquid	–1.79***	1.94***	1.86***	1.85***	1.89***	1.94***	1.95***	1.82***
feature	–.80*	2.05***	1.29***	1.84***	2.01***	.47	–1.09***	1.74***
display	1.47***	.58*	–.22	.73*	.76***	1.19***	.75**	.79***
countc56adr	4.95***	4.83***	1.98**	–1.85**				
countc56adpr					–.63	–.61	.15	–1.76*
countc84adpr					5.01***	–4.53***	3.51***	4.74***
countc98adpr					–.23	.48	.38	.24
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	No	No	Yes	Yes	No	No	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	77,112	77,112	77,112	77,112	77,112	77,112	77,112	77,112
Log-Likelihood	–6,076	–6,050	–5,955	–5,914	–6,058	–6,030	–6,033	–5,974

Note: Asterisks indicate significance levels. Random coefficients are independent. Insignificant advertising variables dropped from table.

Table 3.22. Multiple RC IV Mixed Logit - Combined Models

Model	s44GLb10b Coef.	s44GLb10a Coef.	as44GLb10b Coef.	as44GLb10a Coef.	as44GLb22b Coef.	as44GLb22a Coef.
Mean						
price	–1.21***	–1.54***	–1.04***	–1.34***	–1.08***	–1.41***
GLdumB	7.17***	7.19***	7.08***	7.07***	7.09***	7.13***
inhp	–.29***	.15**	–.35***	.03	–.38***	.04
inhl	.62***	1.07***	.79***	1.23***	.76***	1.26***
iprhh	.06***	.06***	.05*	.01	.06***	.06**
idurhh	–.00	–.00	–.01	–.01	–.01*	–.01*
color	–.64***	–.50***	–.74***	–.59***	–.73***	–.60***
sensitiv	–.94***	–.88***	–.79**	–.66*	–.90**	–.78**
konzentrat	–1.28***	–1.14***	–1.23***	–1.07***	–1.21***	–1.09***
gimmick	1.58***	1.74***	1.41***	1.22***	1.71***	1.55***
extrasize	.19	.12	–.21	–.13	–.14	–.20
bigpack	2.92***	2.63***	2.63***	2.74***	2.49***	2.87***
liquid	–3.32***	–3.36***	–3.74***	–3.83***	–3.73***	–3.85***
duration	–.05***	–.04***	–.03*	–.03	.00	.00
duration2	.00***	.00***	.00***	.00**		
countc56adr					–.10	.04
Standard Deviation						
price	.35***	.32***	.35***	.37***	.36***	.34***
liquid	1.56***	1.55***	1.69***	1.62***	1.59***	1.58***
duration	.00	–.00	–.00	.01	–.00	–.01
duration2	.00	–.00	.00	–.00		
countc56adr					–3.15**	–2.88*
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	77,112	77,112	77,112	77,112
Log-Likelihood	–13,467	–13,288	–3,910	–3,863	–3,918	–3,864

Note: Asterisks indicate significance levels. Random coefficients are independent. Insignificant feature and display variables dropped from table.

Table 3.23. Multiple RC IV Mixed Logit - Combined Models continued

Model	as44GLb42b	as44GLb42a	as44GLb52b	as44GLb52a
	Coef.	Coef.	Coef.	Coef.
Mean				
price	–1.16***	–1.43***	–1.21***	–1.48***
inhp	–.34***	.03	–.32***	.07
inhl	.73***	1.27***	.82***	1.25***
iprhh	.02	.04*	.04*	.04*
idurhh	–.01*	–.01*	–.01*	–.01*
color	–.73***	–.60***	–.73***	–.61***
sensitiv	–.80**	–.63*	–.85**	–.71*
konzentrat	–1.19***	–1.06***	–1.20***	–1.06***
gimmick	1.53***	1.33***	1.32***	1.25***
extrasize	–.14	–.26	–.06	–.21
bigpack	2.58***	2.87***	2.02***	3.06***
liquid	–3.57***	–3.88***	–3.55***	–3.74***
GLdumB	8.04***	8.14***	8.69***	8.14***
duration	.00	.00	–.00	.00
countc56adr			–.04	.37
Standard Deviation				
price	.42***	.40***	.46***	.39***
liquid	1.64***	1.69***	1.61***	1.64***
GLdumB	2.95***	2.99***	4.16***	2.84***
duration	.01	.00		
countc56adr			–2.70**	1.32
Observed Demography	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes
No. of Observations	77,112	77,112	77,112	77,112
Log-Likelihood	–3,881	–3,836	–3,879	–3,829

Note: Asterisks indicate significance levels. Random coefficients are independent. Insignificant feature and display variables dropped from table.

Table 3.24. Multiple RC IV Mixed Logit - Stability across Samples

Model	st142b Coef.	st242b Coef.	st342b Coef.	st442b Coef.	st442a Coef.	st542b Coef.	st542a Coef.
Mean							
price	–1.52***	–1.46***	–1.35***	–1.44***	–1.69***	–1.47***	–1.68***
inhp	–.17***	–.18***	–.07	–.28***	.17***	–.16***	.19***
inhl	.63***	.63***	.60***	.52***	.97***	.53***	.85***
iprhh	.09***	.09***	.06***	.08***	.07***	.10***	.08***
liquid	–2.71***	–2.75***	–2.55***	–2.69***	–2.79***	–2.61***	–2.61***
feature	.01	.02	.15	.05	.12	.38**	.43***
display	.37***	.36***	.23**	.25**	.28***	.24**	.31***
Standard Deviation							
price	.52***	.53***	.59***	.54***	.50***	.59***	.53***
liquid	1.62***	1.72***	1.60***	1.72***	1.80***	–1.64***	1.55***
feature	–.80***	.84***	–1.16***	.84***	.75***	.78*	–.50*
display	–.06	.14	.36*	–.72***	–.66***	–.57***	–.31*
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction	No	No	No	No	Yes	No	Yes
No. of Observations	299,976	278,411	172,201	269,942	269,942	163,732	163,732
Log-Likelihood	–23,836	–21,946	–14,284	–20,897	–20,696	–13,327	–13,161

Note: Asterisks indicate significance levels. Random coefficients are independent.

Table 3.25. Multiple RC IV Mixed Logit - Varying Number of Choice Alternatives

Model	s44GLb5b	s44GLb5a	s45GLb5b	s45GLb5a	s46GLb5b	s46GLb5a
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Mean						
price	–1.79***	–2.27***	–1.73***	–2.30***	–1.79***	–2.09***
GLdumB	7.27***	7.30***	6.74***	6.84***	6.41***	6.47***
inhp	–.19***	.58***	–.40***	.38***	–.36**	.20
inhl	.97***	1.61***	.50***	1.29***	.54**	1.35***
iprhh	.08***	.07***	.09***	.09***	.10***	.08***
liquid	–3.57***	–3.44***	–3.00***	–3.15***	–3.01***	–3.59***
feature	–.03	.08	.13	.15	–.00	.14
display	.36***	.37***	.29	.38*	.75***	.72***
Standard Deviation						
price	.51***	.44***	.57***	.51***	.61***	.55***
liquid	1.52***	1.49***	1.13***	1.18***	1.13***	1.19***
feature	1.10***	.97***	–.70*	–.83*	–.91*	–1.01*
display	.64**	.65**	1.05***	.90***	–.12	.70
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	131,790	131,790	64,875	64,875
Log-Likelihood	–14,276	–13,936	–6,255	–6,088	–2,706	–2,635

Note: Asterisks indicate significance levels. Random coefficients are independent.

Table 3.26. Petrin and Train Setup - Liquid Detergents Weekly

Model	s441b Coef.	s441a Coef.	s443b Coef.	s443a Coef.	s444b Coef.	s444a Coef.	s445b Coef.	s445a Coef.
Mean								
price	–.29**	–.29**	–.29**	–.29**	–.33*	–.19	–.29*	–.24
inhl	–.23**	–.28**	–.26**	–.24**	–.51***	–.58***	–.52***	–.57***
iprhh	.03*	.04*	.03*	.03*	.05*	.04*	.05*	.04*
liquid	–3.89***	–3.75***	–3.97***	–3.97***	–3.64***	–3.77***	–3.71***	–3.71***
feature			–.22	–.22			–.01	–.16
display			.09	.06			–.03	.16
Standard Deviation								
price	.18***	.23***	.19***	.19***	.27***	.23***	.24***	.24***
liquid	1.91***	–1.80***	1.95***	1.91***	1.93***	1.87***	1.91***	1.90***
feature			1.63***	1.60***			1.52***	1.62***
display			.38	.35			.78	–.25
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	No	No	No	No	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	50,248	50,248	50,248	50,248	50,248	50,248	50,248	50,248
Log-Likelihood	–4,229	–4,226	–4,202	–4,193	–4,136	–4,125	–4,120	–4,111

Note: Asterisks indicate significance levels. Random coefficients are independent.

Parameter Estimates - Efficiency Prices

Table 3.27. Multiple RC IV Mixed Logit - Efficiency Prices - Basic Models

Model	es441b Coef.	es441a Coef.	es443b Coef.	es443a Coef.	es444b Coef.	es444a Coef.	es445b Coef.	es445a Coef.
Mean								
price	–2.07***	–2.09***	–2.08***	–2.04***	–2.22***	–2.24***	–2.22***	–2.23***
inhp	–.52***	–.50***	–.56***	–.55***	–.55***	–.53***	–.57***	–.55***
inhl	–.37***	–.34***	–.45***	–.45***	–.77***	–.77***	–.81***	–.77***
iprhh	.09***	.09***	.09***	.09***	.10***	.10***	.09***	.08***
liquid	–1.27***	–1.27***	–1.22***	–1.22***	–.63***	–.58***	–.66***	–.67***
feature			–.25	.10			.09	.18
display			.31**	.17			.05	.11
Standard Deviation								
price	.93***	.93***	.92***	.89***	.76***	.73***	.72***	.68***
liquid	–1.58***	–1.58***	–1.56***	–1.63***	–1.65***	–1.66***	1.82***	1.81***
feature			1.82***	1.42***			.73***	.72***
display			–1.27***	1.69***			–.82***	.68***
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	No	No	No	No	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	269,942	269,942	269,942	269,942	269,942	269,942
Log-Likelihood	–20,992	–20,914	–20,868	–20,766	–20,020	–19,933	–20,027	–19,925

Note: Asterisks indicate significance levels. Random coefficients are independent.

Table 3.28. Multiple RC IV Mixed Logit - Efficiency Prices - Inventory and Varieties

Model	es448b	es448a	aes448b	aes448a	es449b	es449a	aes449b	aes449a
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Mean								
price	–2.32***	–2.39***	–2.67***	–2.89***	–1.91***	–1.98***	–2.23***	–2.29***
inhp	–.60***	–.57***	–.66***	–.65***	–.59***	–.57***	–.64***	–.62***
inhl	–.82***	–.86***	–.55***	–.65***	–.70***	–.71***	–.52***	–.53***
iprhh	.09***	.09***	.14***	.14***	.08***	.08***	.10***	.11***
liquid	–.72***	–.61***	–1.25***	–1.11***	–1.04***	–.95***	–1.34***	–1.22***
feature	.14	.15	–.04	.16	.03	.10	–.09	–.02
display	.05	.20**	–.11	–.01	.06	.09	–.07	–.05
idurhh	–.00	.00	–.01	–.00				
duration	.03***	.04***	.06***	.06***				
duration2	–.00***	–.00***	–.00***	–.00***				
color					–.14***	–.10**	–.23**	–.19**
sensitiv					–.18	–.13	–.19	–.17
konzentrat					–.44***	–.36***	–.33***	–.26***
gimmick					.24	.23	.39	.26
extrasize					–.04	.05	–.17	–.12
bigpack					.91***	.78***	.55	.60
Standard Deviation								
price	.72***	.73***	.58***	.69***	.67***	.66***	.71***	.69***
liquid	1.84***	1.84***	1.99***	1.83***	1.84***	1.85***	1.90***	1.87***
feature	.56*	.75***	1.32***	1.13***	.84***	.83***	–1.30***	–1.20***
display	–.83***	–.35	–.55	–.46	–.67***	–.65***	.42	.41
duration	–.03***	.01	.03***	.06***				
duration2	–.00**	–.00***	–.00**	–.00				
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	77,112	77,112	269,942	269,942	77,112	77,112
Log-Likelihood	–20,004	–19,909	–5,721	–5,682	–19,992	–19,917	–5,680	–5,657

Note: Asterisks indicate significance levels. Random coefficients are independent.

Table 3.29. Multiple RC IV Mixed Logit - Efficiency Prices - Habit Formation

Model	es44GLb5b	es44GLb5a	es44GLc5b	es44GLc5a	aes44GLb5b	aes44GLb5a	aes44GLc5b	aes44GLc5a
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Mean								
price	–2.87***	–2.87***	–2.89***	–2.93***	–2.69***	–2.74***	–2.74***	–2.85***
inhp	–.78***	–.76***	–.82***	–.80***	–.84***	–.84***	–.94***	–.90***
inhl	–1.03***	–1.04***	–1.15***	–1.14***	–.70***	–.79***	–.83***	–.86***
iprhh	.06***	.05**	.02	.02	.03	.03	–.01	–.03
liquid	–.64***	–.54***	–.57***	–.50***	–1.35***	–1.24***	–1.41***	–1.31***
feature	.09	.21	.09	.25*	.31	.36	.23	.34
display	.03	.01	.05	.10	–.49	–.20	–.35	–.12
GLdumB	7.05***	7.10***			6.96***	7.03***		
mdum1L			5.52***	5.54***			5.04***	4.98***
mdum2L			3.92***	3.88***			3.75***	3.79***
mdum3L			1.82***	1.87***			2.21***	2.16***
mdum4L			1.93***	1.80***			1.34***	1.46***
mdum5L			1.76***	1.82***			2.03***	1.95***
mdum6L			1.32***	1.34***			1.65***	1.30**
mdum7L			1.38***	1.37***			1.68***	1.97***
mdum8L			1.21***	1.11***			1.37**	1.42**
mdum9L			.02	.11			1.00*	1.11*
mdum10L			.67*	.67*			.22	.35
Standard Deviation								
price	.46***	.42***	.31***	.34***	.39***	.27***	.25***	.31***
liquid	1.44***	1.42***	1.44***	1.40***	1.53***	1.61***	1.46***	1.54***
feature	.97***	.94***	.83***	.75***	.82*	.86**	.61	–.80**
display	.72***	.93***	–.41	–.48*	1.46***	1.04***	1.21***	–.05
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	269,942	269,942	77,112	77,112	77,112	77,112
Log-Likelihood	–12,859	–12,663	–11,011	–10,787	–3,729	–3,664	–3,028	–2,965

Note: Asterisks indicate significance levels. Random coefficients are independent.

Table 3.30. Multiple RC IV Mixed Logit - Efficiency Prices - Advertising

Model	aes4420b Coef.	aes4420a Coef.	aes4421b Coef.	aes4421a Coef.	aes4423b Coef.	aes4423a Coef.	aes4424b Coef.	aes4424a Coef.
Mean								
price	–2.18***	–2.19***	–2.45***	–2.53***	–2.22***	–2.22***	–2.41***	–2.46***
inhp	–.61***	–.60***	–.63***	–.60***	–.62***	–.60***	–.58***	–.56***
inhl	–.28**	–.28**	–.56***	–.67***	–.38***	–.41***	–.52***	–.50***
iprhh	.11***	.12***	.08**	.10***	.09***	.07*	.07**	.09***
liquid	–1.63***	–1.59***	–1.19***	–.89***	–1.51***	–1.40***	–1.16***	–1.20***
feature	.01	.04	–.20	.35*	–.30	–.13	.10	–.15
display	–.07	–.04	–.25	–.28	–.40	.01	–.50*	–.11
countc56adr	2.06***	2.12***	.05	–.21				
countc56adpr					1.17***	1.43***	.70*	.72*
countc84adpr					–.64	–.46	–.53	–2.39**
countc98adpr					.62	.63	.22	.17
Standard Deviation								
price	.88***	.87***	.69***	.76***	.83***	.82***	.70***	.71***
liquid	1.72***	1.71***	1.82***	1.73***	1.72***	1.70***	1.84***	1.98***
feature	1.54***	1.50***	1.50***	–.29	2.10***	1.85***	1.00**	–1.48***
display	–1.51***	–1.48***	–1.03***	1.08*	1.88***	1.27***	1.41***	–.71*
countc56adr	4.11***	4.09***	1.41*	1.43				
countc56adpr					.03	.94*	.65	–.00
countc84adpr					–2.95***	2.79***	1.53*	–3.90***
countc98adpr					–.59	–.20	.51	.03
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	No	No	Yes	Yes	No	No	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes	No	Yes
No. of Observations	77,112	77,112	77,112	77,112	77,112	77,112	77,112	77,112
Log-Likelihood	–5,938	–5,910	–5,741	–5,667	–5,908	–5,880	–5,786	–5,741

Note: Asterisks indicate significance levels. Random coefficients are independent. Insignificant advertising variables dropped from table.

Table 3.31. Multiple RC IV Mixed Logit - Efficiency Prices - Combined Models

Model	es44GLb10b	es44GLb10a	aes44GLb10b	aes44GLb10a	aes44GLb22b	aes44GLb22a
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Mean						
price	–2.31***	–2.47***	–2.52***	–2.59***	–2.46***	–2.57***
GLdumB	7.03***	7.08***	6.92***	7.02***	7.02***	7.05***
inhp	–.80***	–.78***	–.88***	–.87***	–.88***	–.83***
inhl	–.89***	–.91***	–.68***	–.69***	–.65***	–.68***
iprhh	.03	.04*	.07*	.04	.05	.03
idurhh	.00	.00	–.01	–.01	–.01	–.01
color	–.21***	–.16***	–.31***	–.24**	–.30***	–.25**
sensitiv	–.36*	–.28	–.38	–.34	–.40	–.36
konzentrat	–.59***	–.45***	–.49***	–.36***	–.48***	–.37***
gimmick	.39**	.32*	.27	.18	.35	.15
extrasize	–.30	–.07	–.33	–.45	–.45	–.39
bigpack	1.24***	.97***	1.37***	1.36***	1.38***	1.30***
liquid	–1.08***	–.92***	–1.52***	–1.48***	–1.60***	–1.47***
duration	–.02**	–.01*	.01	.02	.01	.01
duration2	.00*	.00	–.00	–.00		
countc56adr					.35	.25
Standard Deviation						
price	.41***	–.36***	.41***	.38***	.41***	.44***
liquid	1.46***	1.48***	1.44***	1.56***	1.48***	1.55***
duration	–.00	–.00	.01	.01	.01	.02*
duration2	.00	–.00	–.00	–.00		
countc56adr					–1.27	–1.05
Observed Demography	Yes	Yes	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes	No	Yes
No. of Observations	269,942	269,942	77,112	77,112	77,112	77,112
Log-Likelihood	–12,741	–12,577	–3,702	–3,647	–3,690	–3,652

Note: Asterisks indicate significance levels. Random coefficients are independent. Insignificant feature and display variables dropped from table.

Table 3.32. Multiple RC IV Mixed Logit - Efficiency Prices - Combined Models continued

Model	aes44GLb42b	aes44GLb42a	aes44GLb52b	aes44GLb52a
	Coef.	Coef.	Coef.	Coef.
Mean				
price	–2.44***	–2.59***	–2.47***	–2.55***
inhp	–.85***	–.83***	–.85***	–.82***
inhl	–.63***	–.69***	–.65***	–.67***
iprhh	.03	.03	.03	.02
idurhh	–.01	–.00	–.01	–.00
color	–.34***	–.24**	–.33***	–.27***
sensitiv	–.40	–.32	–.41	–.37
konzentrat	–.50***	–.35***	–.50***	–.39***
gimmick	.21	.03	.23	.05
extrasize	–.44	–.32	–.35	–.31
bigpack	1.36***	1.26***	1.34***	1.37***
liquid	–1.63***	–1.39***	–1.54***	–1.44***
GLdumB	7.69***	7.50***	7.84***	7.67***
duration	.01	.00	.01	.00
countc56adr			.42	.40
Standard Deviation				
price	.44***	.45***	.42***	.46***
liquid	1.62***	1.54***	1.57***	1.55***
GLdumB	2.04***	1.81***	2.22***	2.12***
duration	.00	–.00		
countc56adr			–.42	1.19
Observed Demography	Yes	Yes	Yes	Yes
Brand Dummies	Yes	Yes	Yes	Yes
Endogeneity Correction	No	Yes	No	Yes
No. of Observations	77,112	77,112	77,112	77,112
Log-Likelihood	–3,670	–3,637	–3,669	–3,633

Note: Asterisks indicate significance levels. Random coefficients are independent. Insignificant feature and display variables dropped from table.

Chapter 4

Local Advertising in a Discrete Choice Demand Model

4.1 Introduction

The prevalence of advertising in consumer goods markets is an indication that firms use it as an effective instrument to interact and communicate with their customers. So far, the focus has been on evaluating the impact of national advertising and lately as in Goeree (2008), on differentiating advertising channels, i.e. TV, print, radio or online. I want to study the question whether local TV advertising is valuable in a standard market demand model. It is of interest to know whether advertising can only explain market share differences nationally or whether it is possible to explain local market differences. Advertising expenditure of firms in the consumer detergent and cleaner industry amount to 300 million € in 2006 with almost 100% share of expenditure on TV advertising.[1]

The chapter makes contributions in several dimensions. Previously, TV advertising has not been studied at this level of detail and not for local markets in this model class. Construction of the necessary market level data from an individual level panel is a novelty. Methodologically, the chapter discusses thoroughly not only the construction, but also the model setup that urgently requires a careful discussion in this model class and is commonly skipped. In terms of estimation method I consider the recent advances made and apply them to real-life data.

Define the exposure of consumers to TV advertising for a regional market within a country as local TV advertising. As TV advertising is broadcast via national TV channels, it is not obvious how to obtain local TV advertising data. I suggest to use an individual level panel like the Single Source panel of A.C. Nielsen. The term individual is used as

[1] Consult the publication of SevenOneMedia in 2007 on p. 30 produced by Nielsen Media Research. Viewed on January 18th, 2010: `http://www.sevenonemedia.de/imperia/md/content/content/Research/Downloads/branchenreport/branchenreport_2007.pdf`

synonym to indicate that information comes from households or consumers. Moreover, this contrasts with information obtained only for a whole market. The speciality of the panel is that it links product purchases to advertising exposure on an individual (i.e. household) level. I will exploit this link by aggregating individuals in a geographic market with their purchase behavior and advertising exposure. This delivers a sample of geographically dispersed markets where purchases and advertising exposure vary across markets. In principle it is possible to combine this information differently. Suppose you had a sample of geographically dispersed consumers for which TV exposure is given and a sample of markets for which sales are known. Commonly it is difficult to obtain the latter data, e.g. marketing research companies are restrictive about this kind of data due to contracts with their customers who are the producers of the monitored goods.

Given the newly aggregated data, I estimate two discrete choice models: a standard logit model and a random coefficients logit model, both of which are workhorses used especially in competition analysis and surveyed in Ackerberg, Benkard, Berry, and Pakes (2006). The demand setup follows Nevo (2001). This model allows to control for endogenous prices.[2] The endogeneity stems from an unobserved variable that is correlated with prices while influencing consumer choice. The standard example is product quality or advertising. Note that the latter is available in my data so as by-product I conduct an endogeneity correction by adding previously missing information to the model. However, advertising is not the sole cause of price endogeneity as became evident form chapter 3 of this work so that treating prices as endogenous even with advertising in the model is warranted. Different to the mentioned chapter, I will not expand my discussion to these issues in detail. So-called Hausman instruments used in the approach are as in Hausman (1996) and applied in Nevo (2001), Petrin and Train (2006) or chapter 3 of this work.

Methodologically the random coefficients model is not as effortless to implement. I follow Knittel and Metaxoglou (2008) and Dubé, Fox, and Su (2009) to circumvent the hurdles in the traditional estimation procedure of Berry, Levinsohn, and Pakes (1995). Summarizing, this amounts to setting the right stopping rules for the numerical methods used, to generate randomized starting values and the choice of the right solver when applying the original BLP estimation procedure.[3]

Before estimation it is necessary to construct the market level dataset. There are many possibilities of aggregation to construct markets and define products from the individual level data. This fact delivers many potential setups that yield data for the estimation of the model. As the random coefficients logit model is computationally expensive it is not

[2] See Berry, Levinsohn, and Pakes (1995) on p. 850f for further explanations of endogeneity in this model.

[3] I use the standard matlab built-in medium and large scale (gradient) solvers (i.e. `fminunc`), the simplex method (i.e. `fminsearch`), the gradient solver for constrained problems (i.e. `fmincon`) and the commercial optimization solver KNITRO of Ziena Inc.. The large scale solver `fminunc` and KNITRO perform similar if `fminunc` options are adjusted. Details are provided during the discussion of the results of the random coefficients model in section 4.4.

possible to compare all setups using the full model. That is why I first estimate all generated aggregation setups with the simple logit model. Since the setups are not nested, I cannot simply test specifications against each other. However, I use summary statistics and economic reasoning to preselect some setups to be estimated as random coefficients model. I want to highlight that this step is usually not detailed in previous work where commonly one economically convincing ad hoc setup is used. Usually sensitivity of the estimation results following a modification of a chosen setup is ignored.

Advertising on the market level has been introduced recently by Goeree (2008) to model personal computer demand. The advertising data used there are not local as in this work, but contain details for different advertising channels beyond TV, e.g. newspaper, magazine, TV and radio. In her model, advertising has the primary role of informing the consumer about product existence in the choice set, which is an extension to the original Berry, Levinsohn, and Pakes (1995) model. She argues that this is necessary due to the high frequency of innovations in the personal computer market. Barroso (2009) has looked at advertising in the Spanish car market, but only with aggregate advertising data as well, where advertising also has an impact on the current choice set. Other studies have not looked at market models and advertising, but studied individual level models.[4] I do not depart from the original Berry, Levinsohn, and Pakes (1995) model, but merely introduce advertising as an utility enhancing effect. In the spirit of Hendel and Nevo (2006) this can be interpreted as adding a present value of future consumption that is anticipated by the consumer at the moment of purchase. The advertising variable measures this effect at the moment of purchase. Concerning the choice set effect of advertising for detergents, I do not think this is necessary because "real" innovation is less than in the personal computer market and the decisions are financially small compared to buying PCs or cars. In these latter markets, I believe the consumers to anticipate ex ante that not knowing the choice set might damage them seriously. This is surely not the case for simple goods of daily life as detergents.

The chapter proceeds as follows. In section 4.2, I introduce the random coefficients model and, as special subcase, the simple logit model. In section 4.3, I give an introduction to the data and specify the aggregation procedure. In section 4.4, I give results and a discussion of issues found during the implementation. The final section 4.5 concludes.

4.2 Model

I start with an introduction to the discrete choice model of demand and detail the specification used in this chapter. See the handbook article of Ackerberg, Benkard, Berry, and Pakes (2006) or the practical guide of Nevo (2000) for more on this model

[4] Typical examples include Erdem and Keane (1996), Ackerberg (2003), Shum (2004), Anand and Shachar (2010).

class. The researcher observes $t = 1, \ldots, T$ markets, each inhabited with $i = 1, \ldots, I_t$ consumers. In each market, $j = 1, \ldots, J$ products are offered to the consumers. The researcher knows the following minimal information for each market: aggregate sales, average prices and product characteristics per product j. The market definition differs according to the application. A market can be a national market in a specific time period or if a panel of local markets over time is available, a market can be a location-time combination. In this work it is the latter. See Berry, Levinsohn, and Pakes (1995) for the former case and Nevo (2001) for the latter case. Market demand is the aggregated result of the choices made by consumers who in turn take their decision by maximizing the utility of consumption. The consumer i in market t benefits from his choice of one unit product j according to the indirect utility function:[5]

$$U(x_j, \xi_{jt}, p_{jt}, D_i, \nu_i; \theta) \tag{4.1}$$

Observed and unobserved characteristics to the researcher are given by vector x_j and scalar ξ_{jt}, respectively. Vectors are assumed to be column vectors if not stated else. x_j comprises only market invariant components, whereas ξ_{jt} allows also variation across markets. Consumers know all characteristics x_j and ξ_{jt}. The standard assumption is that x_j is exogenous, ξ_{jt} is independent across t, and ξ_{jt} is mean independent of x_j: $E[\xi_{jt}|x_j] = E[\xi_{jt}]$.

p_{jt} is average price of product j in market t.[6] D_i and ν_i are observed and unobserved consumer characteristics where the latter require a parametric distributional assumption and θ collects demand parameters. The researcher needs to specify functional forms for the utility function U and the cumulative distribution function $F_\nu(\nu)$. See the appendix A for a description of the general case. In this chapter U takes the following form:

$$u_{ijt} = -p_{jt}\alpha_i + x_j'\beta_i + a_{jt}'\gamma_i + \xi_{jt} + \epsilon_{ijt} \tag{4.2}$$

This is the standard random coefficients model, where the parameters α_i, β_i and γ_i are random coefficients. x_j, ξ_{jt} and p_{jt} are defined as before. Note that for x_j and the interpretation of ξ_{jt} it is important to distinguish between brand and product dummies that are commonly part of x_j, see sections 4.2.1 and 4.2.2 for details. New to this aggregation level is the vector a_{jt} that collects local advertising information. In this chapter it has two components: retail advertising and TV advertising. Note that different to previous models both vary over markets t. ϵ_{ijt} is an i.i.d. stochastic error term,

[5] The assumption of buying only one unit is a limitation, but necessary for this model. See p. 32 of Ackerberg, Benkard, Berry, and Pakes (2006) for a discussion of papers that generalize to multiple units of demand.

[6] Implicitly, I assume that all consumers face the same price, and it is the average price. If not, this introduces measurement error that is accommodated automatically in this model according to Berry (1994), as long as it is limited to the consumer invariant utility part δ_{jt} of the decomposition in (4.6) presented later in this section.

assumed to follow an type I extreme value distribution with cumulative distribution function $F_\epsilon(\epsilon)$. This form of indirect utility in (4.2) is free of wealth-effects and that is why income of the individual does not enter the utility function. Adding an income term $y_i\alpha_i$ is possible but superfluous as it drops out eventually given this functional form. Berry, Levinsohn, and Pakes (1995) or Petrin (2002) use utility functions that allow for wealth effects, which is warranted given that they study the car market, not detergents as I do.

The model must allow the possibility for the consumer not to buy any of the J products. This outside good ($j = 0$) yields the following utility:

$$u_{i0t} = \epsilon_{i0t} \tag{4.3}$$

Prices and observed characteristics do not appear, as none of the J products is bought. ξ_{0t} is normalized to zero.

The random coefficients take the following form:

$$\begin{pmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} + \Pi D_i + \Sigma \nu_i \tag{4.4}$$

The random coefficients can be decomposed into three components as visible from the right hand side of the equation: a mean value common to all consumers, an observed heterogeneity part and an unobserved random heterogeneity part.[7] Let the stacked vector of α_i, β_i and γ_i be $K \times 1$.

The observed heterogeneity part consists of a $d \times 1$ vector of demographics D_i and $K \times d$ matrix Π that selects which of the demographics from D_i affect the parameters α_i, β_i and γ_i. Typical examples are income, age, household size or education. The distribution of D_i, $F_D(D)$ is either a nonparametric or a parametric cumulative distribution function and has to be estimated from consumer data. For the estimation, all individuals i are drawn from the pre-estimated distribution.

The unobserved heterogeneity part is made up of a random $K \times 1$ vector ν_i. Its distribution is frequently assumed to be a multivariate standard normal with mean zero and a $K \times K$ covariance matrix Σ, commonly assumed to be diagonal so that the diagonal elements are interpreted as variances and off-diagonal covariances are zero. $F_\nu(\nu)$ is the corresponding cumulative distribution function. $F_\nu(\nu)$, $F_D(D)$ and $F_\epsilon(\epsilon)$ are assumed to be independent.

[7] Observed does not imply the researcher knows a particular consumer, it merely means the researcher uses knowledge about the distribution of consumer characteristics. The researcher does not have such information for the unobserved heterogeneity part.

Combining equations (4.2) and (4.4), I can rewrite the model in the following form:

$$\begin{aligned} u_{ijt} &= \delta_{jt}(p_{jt}, x_j, a_{jt}, \xi_{jt}; \theta_1) + \mu_{ijt}(p_{jt}, x_j, a_{jt}, \nu_i, D_i; \theta_2) + \epsilon_{ijt} && (4.5) \\ \delta_{jt} &= -p_{jt}\alpha + x_j'\beta + a_{jt}'\gamma + \xi_{jt} && (4.6) \\ \mu_{ijt} &= [-p_{jt}, x_j', a_{jt}'](\Pi D_i + \Sigma \nu_i) && (4.7) \end{aligned}$$

This decomposition separates individual utilities into a part common to all consumers, i.e. δ_{jt} and an individual part, i.e. μ_{ijt}, that depends on consumer characteristics. θ_1 summarizes the parameters in δ_{jt} and θ_2 the parameters in μ_{ijt} and $\theta = (\theta_1, \theta_2)$. The common part δ_{jt} can be interpreted as mean effect for all consumers. The mean zero individual part μ_{ijt} allows for deviations from the mean effect and summarizes the effects of the random coefficients. The deviations are built from interactions of observed product characteristics and both, observed and unobserved consumer characteristics. This completes the individual level. Now I move from the consumer perspective to the market perspective. If we could calculate δ_{jt}, then market shares per product and market are given by the following integral conditional on δ_{jt}, θ_2 and market data:

$$S_{jt}(\delta_{jt}, \theta_2, p_{jt}, x_j, a_{jt}) = \int_\nu \int_D \frac{exp(\delta_{jt} + \mu_{ijt}(.))}{1 + \sum_{l=1}^{J} exp(\delta_{lt} + \mu_{ilt}(.))} dF_D(D) dF_\nu(\nu) \qquad (4.8)$$

Equation (4.8) is derived by calculating the probability that the consumer chooses product j and no other product. To calculate the market shares I have to integrate out the variables ϵ, D and ν. The integrand of the double integral takes the well-known logit form, because there is an analytic solution to the integral given the i.i.d. extreme value assumption on ϵ. This part corresponds to the classical derivation of logit choice probabilities, see Train (2003) on p. 40. The arguments of μ_{ijt} are the market data, random draws from the two distributions of D, ν and parameters θ_2 as in (4.5). The market share in (4.8) can be calculated by simulation.

I briefly outline the GMM estimation algorithm of the model based on Berry (1994).[8] It is an iterative procedure that uses a GMM criterion as outer loop and requires at each iteration step g of the GMM objective function evaluation an computationally expensive simulation as inner loop. The GMM estimation delivers the estimates $\theta = (\theta_1, \theta_2)$. The GMM objective function is constructed from the moment conditions $E[\xi_{jt}|\boldsymbol{z_{jt}}] = 0 \,\forall j, t$ and the M-dimensional vector $\boldsymbol{z_{jt}}$ contains M instruments for each market and product. Moreover, for the moment conditions the value of ξ_{jt} must be known to construct the GMM objective function. ξ_{jt} is calculated in the inner loop and its estimate is used in the GMM criterion function.

[8] Estimation is explained in detail in Nevo (2000) and its separate appendix. Beware of the use of the terms product and brand: they are sometimes used interchangeably.

The inner loop is an iterative procedure that matches actual and predicted market shares from the simulated equation (4.8), $\boldsymbol{s_t}$ and $\boldsymbol{S_t(\cdot)}$, to calculate the mean utility level δ_{jt}:[9]

$$\boldsymbol{s_t} = \boldsymbol{S_t}(\boldsymbol{\delta_t}, \theta_2, \boldsymbol{p_t}, \boldsymbol{x}, \boldsymbol{a_t})\,\forall t \tag{4.9}$$

Bold letters denote vectors, where the component values are the per product values of the variable, consequently the vectors collect the values for all products per market t. For this system of equations a contraction mapping exists that can be used to recover a (consistent) unique value of $\boldsymbol{\delta_t}$:

$$\boldsymbol{\delta_t^{(k+1)}} = \boldsymbol{\delta_t^{(k)}} + ln(\boldsymbol{s_t}) - ln(\boldsymbol{S_t}(\boldsymbol{\delta_t}, \theta_2^{(g)}, \boldsymbol{p_t}, \boldsymbol{x}, \boldsymbol{a_t})\,\forall t, k \tag{4.10}$$

At each step k of the inner loop a new value of $\boldsymbol{\delta_t^{(k+1)}}\forall t$ is calculated until a stopping criterion terminates the iterative procedure. The whole inner loop is conditional on the current value of $\theta_2^{(g)}$ in step g of the outer loop. Then an estimate of ξ_{jt} is calculated from equation (4.6) conditional on the current $\theta^{(g)}$ parameters.

Back in the outer loop, from these estimated $\hat{\xi}_{jt}^{(g)}$ and the instruments $\boldsymbol{z_{jt}}$ I can construct the GMM objective function at each step, now for the next step $g+1$. The outer loop also needs a stopping criterion to terminate the estimation. At the first iteration of the outer loop, the starting values for δ_{jt} are given by the logit model in section 4.2.4. So far, this is the standard routine to estimate this model. Berry (1994) presents the necessary assumptions. Dubé, Fox, and Su (2009) developed a representation of the model as mathematical program that can be estimated without iterations with the MPEC algorithm. The latter step is beyond the scope of this chapter. In the following two sections I want to detail the implementation if either brand or product dummies are used in x_j.

4.2.1 Product dummies

If product dummies d_j are contained in x_j, all time invariant product characteristics that differentiate products are modeled by these variables. Let $\tilde{x}_j$ denote the time invariant product characteristics without product dummies. This has consequences for the interpretation of the unobserved characteristic ξ_{jt} and for the estimation of the product characteristic parameters. First, I turn to the unobserved characteristic. Assume that ξ_{jt} can be decomposed according to the following equation:

$$\xi_{jt} = \xi_j + \xi_t + \Delta\xi_{jt} \tag{4.11}$$

[9] Simulation is conducted by drawing many individuals from the distributions $F_D(D)$ and $F_\nu(\nu)$, calculating the integrand and finally approximating the integral by averaging the results.

The unobserved "ξ-components" do not vary across consumers, implying the assumption that consumers face the same unobservables. Although this is restrictive, note that the presented model permits heterogeneity for the observed product characteristics. This renders the lack of heterogeneity on the latent variable ξ more acceptable. The three parts on the right hand side have the following interpretation. ξ_j refers to the unobserved product characteristic in the original sense of Berry (1994), as it is attached to product j and is time/market invariant. ξ_t represents a common market component that affects all products and all consumers in market t, i.e. a local demand shock. Assume that $\xi_t = 0$. This assumption simplifies the model, since capturing the local shocks with fixed effects (dummies) is possible, but would increase the number of parameters by the number of markets. $\Delta\xi_{jt}$ takes the role of a residual to capture deviations from the other two "ξ-components" and can be idiosyncratic for each product-market combination. Examples of the latter are local retail activity or local advertising for a specific product. Local retail activity can be thought of as any kind of measure inside a retail store to affect consumption, for example putting the product in a different place, placing in-store advertisements or running promotional activities. Local advertising can be the amount of TV advertising the consumers in a specific market absorbed. In total, product dummies d_j absorb all time invariant product characteristics so that $d_j = \tilde{x}_j'\beta + \xi_j$.[10] A direct conclusion is that the time invariant product characteristics are not identified in presence of product dummies.

In the estimation, product dummies are included in x_j and thereby are a part of δ_{jt}, but they cannot be used in μ_{ijt} to model the deviations from the mean effect as this would deliver an infeasible routine.[11] Instead of product dummies only product characteristics $\tilde{x}_j$ are used in μ_{ijt}. Consequently, the mean effects of the product characteristics are not identified, because the included product dummies in δ_{jt} capture all time invariant characteristics as mentioned earlier. But due to the interaction of consumer characteristics and product characteristics in μ_{ijt}, the parameters on these interactions of consumer and product characteristics are identified. Apart from this, the estimation of the model is unchanged by these issues and is conducted as explained in the previous section.

An additional step is necessary to recover the mean effects on product characteristics by using the following model:

$$d_j = \tilde{x}_j'\beta + \xi_j, \qquad j = 1, \ldots, J \tag{4.12}$$

According to the model, product dummies are composed of product characteristics and the unobserved characteristic ξ_j plays the role of an residual. By assuming exogeneity of $\tilde{x}_j$ this regression recovers the mean taste parameters on product characteristics. If

[10] Note that β now only contains parameters for the product characteristics, not the product dummies, because in this equation $\tilde{x}_j$ contains no product dummies.

[11] Introducing product dummies as component of μ_{ijt} would correspond to estimating the model with an error variable that has an unrestricted covariance matrix. See Nevo (2000) on p. 528 for details.

mean independence $E[\xi_j|\tilde{x}_j] = 0$ holds, then a GLS regression with weighting matrix based on the covariance matrix of the estimated d_j delivers the mean taste parameters β. See Nevo (2000) on p. 537 for details.

4.2.2 Brand dummies

If it is not possible to follow the product dummies approach, one can resolve to adding brand dummies b_j to the model. In contrast to the product dummy approach, brand dummies and fixed product characteristics enter the δ_{jt} term. The μ_{ijt} expression looks as in the product dummy case. Now, brand dummies neither capture unobserved ξ_j nor observed characteristics $\tilde{x}_j'\beta$ fully as product dummies would do. Thus, there may be unobserved product specific deviations from the brand effect ζ_j that is constant for all products of the same brand. Assume the decomposition of ξ to be as follows:

$$\xi_{jt} = \zeta_j + \Delta\xi_j + \xi_t + \Delta\xi_{jt} \tag{4.13}$$

Brand dummies only capture the brand effect ζ_j. Compared to the product dummy case, the researcher must account in addition for the unobserved deviations from the brand effect per product, the $\Delta\xi_j$. It is important to understand the composition of ξ_{jt} to be able to search and argue for the right instruments. A central drawback of the brand dummy approach is that exogeneity of $\tilde{x}_j$ is necessary for the whole model, not only for the recovery of the mean taste parameters as in the product dummy case. This is an important distinction of the two approaches, as it is up to the application to define and choose product characteristics. Just as $\tilde{x}_j$, brand dummies have to be exogenous as well. From the perspective of the previous argument this assumption is truly not innocuous, if the existence of the unaccounted effect $\Delta\xi_j$ is taken into account. The latter might be linked to product characteristics or to relevant decision variables of the consumer, such as price, and cause endogeneity. For example, $\Delta\xi_j$ could be product specific quality. Therefore, estimation with product dummies is recommended practice.

4.2.3 Instruments

I use Hausman's (1996) standard instruments that exploit the panel structure of the dataset. The crucial assumption is that market specific deviations $\Delta\xi_{jt}$ are independent across t for all j. As the endogeneity problem in this model is caused by correlation/dependence of prices p_{jt} and $\Delta\xi_{jt}$ within a market, this assumption qualifies all prices for similar products in other markets than t as instrument for price of a specific product in market t. This delivers exogeneity of the instruments. If similar products generate similar production costs, instrumental relevance prevails by using price of similar products as instrument. For example, production costs are one exemplary price component that enter final retail price. Both are present in the endogenous price and the price that is used as instrumental variable. Then the endogenous price is correlated with the instrument through the production cost component of price.

The assumption can be violated by having economic activities that induce correlation of $\Delta\xi_{jt}$ across t, e.g. advertising or promotional activities. TV advertising, a perfect candidate to induce this kind of correlation that makes over 99% of advertising expenses, is controlled for in this work.[12] Promotional activities apart from those that I control for are not common for the product category under study. A remaining difficulty may be little variation in prices. I come back to this issue in the data section.

I need no instruments for brand or product dummies in the estimation. Product dummies capture all time invariant unobserved product characteristics. In the product dummy case the necessary assumption is that ξ_j and $\Delta\xi_{jt}$ are uncorrelated, see (4.11). In the brand dummy case the additional assumption of exogeneity of $\tilde{x}_j$ and uncorrelatedness of $\Delta\xi_{jt}$ with both ζ_j and $\Delta\xi_j$ is necessary, see (4.13).

The construction of instruments is implemented in two different procedures. In procedure 1 a set of instruments for a product price in all locations in a quarter is price of the same product in other locations in the same quarter. Those other locations are once randomly picked, because the invoked assumptions on the instruments permit this. Obviously, then there is no instrument variation per product across locations within a quarter but only across quarters. In procedure 2 a set of instruments for a product price in a specific location in a quarter is price of the same product in other locations in the same quarter. As before, other locations are randomly picked, but since this procedure is done for each location-time combination, i.e. a market in the terminology of this chapter, there is variation in the instruments across both the location and the time dimension. Therefore, this is the preferred procedure.

[12] According to Nielsen Media Research advertising expenditure for TV in the detergent category make nearly up to 100% of total advertising expenditure. See footnote 1 in the introduction for a reference.

4.2.4 The Simple Logit Model

There is a restricted market model that is as easy to estimate as an ordinary least squares regression. In the logit market model set $\mu_{ijt} = 0$ so that no random coefficients are present.[13] Then market shares for products and the outside good are given as:

$$S_{jt} = \frac{exp(\delta_{jt})}{1 + \sum_{l=1}^{J} exp(\delta_{lt})} \tag{4.14}$$

$$S_{0t} = \frac{1}{1 + \sum_{l=1}^{J} exp(\delta_{lt})} \tag{4.15}$$

Calculating $ln(S_{jt}) - ln(S_{0t})$ delivers:

$$ln(S_{jt}) - ln(S_{0t}) = -p_{jt}\alpha + x_j'\beta + a_{jt}'\gamma + \xi_{jt} \tag{4.16}$$

Equation (4.16) corresponds to equation (4.6) and the left hand side is a direct estimate for δ_{jt}, the mean utility level for a product in a specific market. x_j can include either brand dummies b_j and product characteristics $\tilde{x}_j$ or product dummies d_j. With an additive error term this equation can be estimated as pooled regression model with data $(S_{jt}, S_{0t}, p_{jt}, x_j, a_{jt})$ for $j = 1, \ldots, J$ and $t = 1, \ldots, T$, where e_{jt} represents the error term that contains ξ_{jt}:

$$ln(S_{jt}) - ln(S_{0t}) = -p_{jt}\alpha + x_j'\beta + a_{jt}'\gamma + e_{jt} \tag{4.17}$$

This model also suffers from omitted variable bias due to the correlation of "ξ-components" and price. A linear instrumental variables approach with the same instruments as in the random coefficients case is adequate to cope with this problem.

4.3 Data

4.3.1 Data Overview

The "Single Source" data I employ are an extensive household level panel supplied by A.C. Nielsen, Germany. It provides daily purchase and real-time advertising exposure by household over a period of 2 years from June 30th 2004 through June 30th 2006. The name Single Source stems from the fact that daily purchase and high frequency TV advertising exposure are recorded for each sampled household. The dataset is collected nationwide throughout Germany and consists of two components: a household panel where purchases are followed and a subsample of the former where additionally all TV

[13] This model exhibits the restricted substitution patterns, for which it has been widely criticized, see the introduction of Ackerberg, Benkard, Berry, and Pakes (2006).

advertising contacts are recorded automatically. For a more thorough description of the data consult chapter 2 of this dissertation. Note that a few tables are repeated in this chapter for the reader's convenience.

Table 4.1. Overview - Market Shares / Means broken up by Brand

	Shares						Means		
Brand	Sales	Value	Ads	Feature	Display	Retail	Price	Size (kg)	Size (l)
3	0.06	0.10	0.33	0.15	0.14	0.15	5.37	2.75	1.93
6	0.01	0.01	0.00	0.04	0.02	0.03	3.40	1.57	1.41
8	0.02	0.02	0.02	0.07	0.05	0.06	3.72	1.14	1.59
10	0.01	0.02	0.00	0.00	0.03	0.02	4.38	6.03	1.46
11	0.01	0.01	0.00	0.03	0.03	0.03	3.90	3.64	1.56
13	0.02	0.01	0.00	0.00	0.00	0.00	1.92	0.00	1.13
16	0.56	0.43	0.00	0.19	0.11	0.14	2.70	2.16	1.54
22	0.01	0.01	0.00	0.02	0.02	0.02	3.42	1.35	2.00
36	0.02	0.03	0.05	0.03	0.05	0.04	4.92	4.32	2.01
40	0.07	0.14	0.40	0.21	0.17	0.19	6.79	3.25	2.27
41	0.03	0.03	0.04	0.04	0.03	0.04	3.99	0.83	1.56
55	0.09	0.09	0.13	0.09	0.13	0.11	3.58	1.84	1.77
57	0.02	0.03	0.00	0.05	0.08	0.06	4.22	4.19	1.69
67	0.03	0.04	0.04	0.07	0.10	0.09	4.79	3.79	1.83
100	0.02	0.02	0.00	0.01	0.05	0.03	4.63	4.42	1.13
Total	0.07	0.07	0.06	0.15	0.10	0.12	3.51	2.57	1.62

Consult table 4.1 to get an overview of the detergent product market that is used to generate the market dataset. The table shows shares and means of interesting variables broken down by brands. The last line shows the mean for all brands. The first column is the brand code. The following six columns show market shares and the last three means. The following variables are contained in the table: Sales is the number of units sold, Value is the value of all units sold, Ads is TV advertising pressure relative to competitors within a 3 month time window, Feature and Display are the common retail activity variables and Retail is the sum thereof. Price is nominal transaction price and finally the Size is measured in kilogram for non-liquid and in liter for liquid detergents.

The market is segmented into two broad groups: common branded products and private labels. The latter are indicated by brand number 16 in the table. Private labels are products that have a brand that is specific to the chain where it is sold and usually represent the "discounter" alternative that is low priced. Note that the two most heavy advertisers by share in the market, brand 3 and 40, charge the highest average price. Therefore, the Value share is higher than the Sales share. Moreover, the biggest brand by sales, i.e. the private labels of the retailers, do not make any advertising and charge a price below average. The brands with the most active retail activity are the private label and the earlier mentioned two heavy TV advertisers.

In table 4.2 the variety characteristics that differentiate products are split up by brand. There is high variation across brands in the characteristics “Liquid”, “Concentrate” and “Color”. That is why these three characteristics should be kept in the empirical application to differentiate products. Many brands do not offer any product variety with one of the characteristics from the last 4 columns and the variation is low. These characteristics are not further used in the analysis.

Table 4.2. Characteristics of all Purchases broken up by Brand

	Means		Shares							
Brand	Nominal Price	Size in kg or liters	Sales by Volume	Liquid	Konzentrat	Color	Sensitiv	Gimmick	Extrasize	Bigpack
3	5.37	2.41	0.06	0.41	0.59	0.51	0.04	0.06	0.00	0.13
6	3.40	1.48	0.01	0.61	0.26	0.10	0.00	0.01	0.03	0.00
8	3.72	1.45	0.02	0.68	0.00	0.80	0.00	0.13	0.00	0.00
10	4.38	3.86	0.01	0.48	0.09	0.26	0.01	0.07	0.00	0.00
11	3.90	2.36	0.01	0.61	0.89	0.00	0.00	0.04	0.18	0.09
13	1.92	1.13	0.02	1.00	0.00	0.13	0.02	0.00	0.00	0.00
16	2.70	1.84	0.56	0.52	0.53	0.42	0.01	0.00	0.00	0.00
22	3.42	1.94	0.01	0.90	0.00	0.05	0.00	0.00	0.00	0.00
36	4.92	3.37	0.02	0.41	0.26	0.16	0.00	0.06	0.09	0.06
40	6.79	2.98	0.07	0.28	0.47	0.40	0.11	0.13	0.12	0.06
41	3.99	1.40	0.03	0.79	0.00	0.14	0.00	0.00	0.04	0.00
55	3.58	1.80	0.09	0.54	0.39	0.14	0.00	0.01	0.01	0.00
57	4.22	2.96	0.02	0.49	0.29	0.42	0.00	0.00	0.10	0.00
67	4.79	3.08	0.03	0.36	0.28	0.02	0.00	0.16	0.10	0.02
100	4.63	3.79	0.02	0.19	0.19	0.06	0.00	0.00	0.03	0.06
Total	3.51	2.08	0.07	0.51	0.45	0.36	0.02	0.03	0.02	0.02

4.3.2 Advertising

The variable a_{jt} mentioned in the theoretical model contains the retail activity in the retailers’ stores and the TV advertising exposure. The latter is to be seen as a filtered series of national TV advertising. In the detergent category all advertising is coordinated nationally. Local deviations from national advertising simply arise as households watch or do not watch an advertisement. As markets are built up from households in a geographic region, the amount of advertising seen in a market and by construction associated sales are known. This allows to see whether on markets with already different demographics advertising still adds explanatory power to the consumption choice. As retail activity is also available on the household level, this is also aggregated and used to control for marketing activity in the retail outlets.

Variation in Prices and Marketing Table 4.3 displays the standard deviations of price, retail activity, TV advertisement pressure relative to competitors for a 3 month period within the market for all transactions. Retail activity is calculated as sum of the feature and display variable. I use only the retail activity variable in the estimation. There is variation across markets for each brand, so there is no reason to believe the market for detergents is controlled on a national level. Hence, it makes sense to look at local geographic markets as I do in this work.

Table 4.3. Variation of Prices and Marketing

Brand	Price	Retail	Feature	Display	Advertisement
3	2.77	0.71	0.42	0.45	0.15
6	1.04	0.79	0.47	0.45	0.00
8	1.10	0.73	0.45	0.43	0.08
10	1.92	0.42	0.03	0.42	0.00
11	2.16	0.70	0.41	0.45	0.00
13	0.42	0.11	0.00	0.11	0.00
16	0.73	0.29	0.17	0.15	0.00
22	0.38	0.61	0.34	0.37	0.00
36	2.56	0.63	0.37	0.46	0.08
40	3.33	0.70	0.45	0.46	0.16
41	1.59	0.57	0.32	0.36	0.08
55	1.92	0.53	0.30	0.38	0.07
57	2.44	0.67	0.38	0.49	0.00
67	2.30	0.67	0.41	0.48	0.06
100	2.07	0.54	0.17	0.47	0.04
Total	2.04	0.51	0.29	0.33	0.14

Note: Values denote standard deviations of variables calculated for the non-aggregated data, i.e. across individual purchases.

Construction of different Advertising Variables In table 4.14 in the appendix I give an overview of the different TV advertising variables used later in the estimation to verify the impact of local advertising. The most important differences among them are in the following dimensions: (i) the time window that is used to construct the advertising variable, e.g. advertising contacts 56, 98 or 140 days prior to a given moment in time are used to construct the variable (ii) contacts are summed up or the relative advertising pressure of a brand relative to all other brands is calculated or a dummy is constructed that is one if the relative advertising pressure of a brand crosses a certain threshold (iii) the value in (ii) is logarithmized.

4.3.3 Construction of a Market Level Dataset

The market level dataset that is used in the analysis is aggregated from the individual level data introduced in the previous section. There are three fundamental objects that have to be defined during this course: the market, the market size and the product. I turn to each now separately.

Markets Recall a market in this chapter is a time-location combination. The data are detailed down to the zip code level, which would at the maximum allow almost 15,000 geographic locations per time interval. Moreover, the purchases are recorded daily so that given the data recording period of two years, more than 700 points in time are available to observe all the geographic locations. To go in both, time and geographic, dimensions to the maximum level of detail when defining the market is infeasible, because the tracked 16,000 households are not enough to have sufficient observations per market. Many observations per market are necessary to construct precise market shares from the individual purchases. This is a challenge for this work. For example, if I aggregate such that a market is a week-county combination (i.e. county is a German "Landkreis"), there are $104 \times 434 = 45,136$ markets.

I construct four versions for the market definition: (1) a market is a quarter-county combination (i.e. county is a German "Landkreis"), there are $8 \times 434 = 3,472$ markets (2) a market is a quarter-zipcode combination, where I aggregate up to the first two digits of the German five digit zip code ($8 \times 95 = 760$ markets) (3) a market is a quarter-zipcode combination, where neighboring zipcodes from (2) are combined to yield at least 10,000 purchases in the newly combined region over the whole sampling period ($8 \times 35 = 280$ markets) (4) a market is a quarter-county observation, where counties are combined to the first two digits of the five digit county code ($8 \times 16 = 128$ markets).

In definition (1), I expect market shares to be imprecisely calculated due to the small amount of purchases per market relative to the number of available products.

From version (1) to (4) when aggregation increases, there are two counteracting forces. Firstly, measurement errors in market shares and missing market shares due to seldom purchased products are reduced. This mitigates data sparseness. Albeit, secondly, the number of observations, i.e. markets, for the discrete choice market model falls at the same time. This is caused by over-aggregation, i.e. creating very large markets. Thus, both data sparseness of (1) and over-aggregation of (4) are serious challenges.

Market Size The discrete choice demand model requires the existence of an outside good, see Berry (1994). This and all product purchases constitute the whole market. Commonly, product sales data are available but not sales of an outside good. Therefore, by deriving a potential market size and subtracting all product sales, I can derive the sales and share of the outside good. Precisely, I do the following: I count the number of individuals per geographic location. I assume there are 6 purchases per household per quarter.[14] Multiplying both gives the potential sales per market. Then variation in market size stems mainly from the number of households in the geographic location.

[14] According to Berry (1994) the precise value of this constant, e.g. setting it to 6 purchases, is of minor importance.

Alternatively, I add up all product purchases and no-purchases in the individual level data to get the total amount of sales in the market. The no-purchases are store visits, where the households did not shop in the detergent category. This can be interpreted as proxy for the outside good. Both measures are very similar and the first is used.

Products The definition of a product is decisive as it has to mimic the actual characteristics space the consumers face when doing their choices. I will vary the product definition to see the impact of a specific choice. The characteristics brand, “Liquid”, “Concentrate” and “Color” should be present in all models estimated. From table 4.2 the characteristics “Liquid”, “Concentrate” and “Color” show the highest variation per brand, so these can potentially rationalize different choices. Product size is another characteristic that is definitely important, but considering all sizes as a distinct product characteristic leads to an intractably high number of products in this product category. That is why I implemented two alternatives for the size variable: (i) aggregation of similar sized products into a size class, where I round the size variable to the nearest positive integer and (ii) price is divided by the size variable to get efficiency prices. With “real” market data with sales from distinct markets, approach (i) is more favorable, as we keep products of different sizes as distinct in the choice set, a one liter detergent is in fact a different choice than the 3 liter packaging, with all other characteristics equal. Moreover, market shares are precise. In my setting of constructed data, I have to keep the number of distinct elements in the choice set as small as possible to calculate sufficiently precise market shares from the geographically dispersed sample of consumers. Note that one of the strength of the discrete choice model is the capability of treating many products, while being parsimonious at the same time. A prerequisite is truly that then market shares are correct and without error.[15]

In both cases (i) and (ii), some products will be sold seldom. Consequently, it is necessary to set a sales threshold for products to have enough sales so that products appear in several markets. Additionally, I have to set a sales threshold for the minimum number of sales transactions in a market such that the market is considered for the estimation. Otherwise, without setting both thresholds, market shares will be calculated with high measurement error. Note that Goeree (2008) drops products with small market shares without giving detailed information. It is not crucial to have all products in all markets, as my implementation of the random coefficients model can accommodate varying numbers of products per market. This is different from publicly available source codes that require all products in all markets.

In table 4.4 I present seven different product definitions implemented later in the estimation. The first column indicates the product setup number, columns 2 to 6 the characteristics considered that define a product and column 7 whether retail activity

[15] Note that a growing choice set has no impact on the number of parameters being estimated, as the consumer choice takes place in characteristics space. This is only true if the model is estimated without product dummies.

Table 4.4. Overview of Product Definition Setups

Product	Product Characteristics					Retail	# of
Setup No.	brand	inh3	liquid	konzentrat	color		Products
1	1	1	1	1	1	1	162
2	1	.	1	1	1	1	76
3	1	.	1	1	1	.	76
4	1	.	1	1	.	.	46
5	1	.	1	.	.	1	29
6	1	1	1	1	.	1	108
7	1	1	1	.	.	1	78

Note: brand indicates brand information, inh3 is the size class variable, liquid, konzentrat and color are dummies that mark liquid, concentrated and color detergents. Retail indicates whether retail activity information is used if the product definition is employed in a market setup.

information is used in the estimation.[16] Column 8 gives the number of products if all products were used in the estimation, i.e. without setting a threshold mentioned before. A "1" in the column indicates that products that differ in the characteristic marked with "1" are kept as different products. If there is a "." the characteristic is not used to differentiate products. For example, color has a ".". In that case an all else equal color product and a non-color product are treated as one product so that sales and retail activity are added up. TV advertising is not added up, as it is measured on the brand level and brand is always included as characteristic, see table 4.4.

In the first row in setup 1 products differ along all characteristics marked with a "1" and retail activity information is used in the estimation. In rows 2 to 5, for setups 2 to 5 products are aggregated so that they do not differ by size, but price is normalized by the size variable according to case (ii) explained previously. In rows 6 to 7, setups 6 and 7 keep the size information, but lump products that differ in characteristic color and concentrate/color together, respectively.

Market Data Setups Given the definition of markets, market size and products there are obviously several combinations to combine the three components. I constructed 120 combinations that may differ in several dimensions. The long table 4.15 in the appendix shows a list of all setups. The first column is the number of the setup. All setups are sorted by columns 2 to 5. In column 2 the market definition of the setup is given. Column 3 indicates the product definition. The thresholds to exclude "small" markets and products are contained in columns 4 and 5. Column 6 marks whether instruments are constructed according to procedure 1 or procedure 2. Column 7 shows the number of individuals for which demographic information is available per market. Columns 8 to 10 show how prices enter into the model, as nominal, as price divided by the size class variable or as price divided by the true size. Column 11 states whether the mean or median price is calculated when determining the average price for a market during

[16] Retail activity is not always contained to see whether there are joint effects with advertising or not.

aggregation. Column 12 shows whether retail activity is employed in the estimation and column 13 whether all available retail activity variables are used. Recall that all retail activity variables are feature, display, priceflag and handbill and the common two are feature and display which is indicated by a dot in column 13. Column 14 indicates whether the mean thereof is calculated during aggregation or whether variable values are summed up.

4.4 Results

This section presents the results from the numerously constructed market setups. As is common practice I start with the estimates of the market model for the simple logit case explained in section 4.2.4. Thereafter I turn to the estimates of the random coefficients discrete choice model. As it is not feasible to estimate all setups in the random coefficient version, I do an extensive analysis on all 120 setups for the simple model and then turn to a few selected setups for the random coefficients model. Importantly, this will not be an arbitrary choice. Estimation of simple logit model takes as long as a linear regression, whereas the estimation of the random coefficients model can take up to 2 days even when using latest hardware.

4.4.1 Simple Logit Model

I estimated each setup with six specifications. Consult table 4.5 for the model setup 116 with product/brand dummies and the plain OLS estimates.[17] The first three columns show OLS results, the latter three make use of the Hausman instruments.[18]

The results are positive regarding the central question of relevance of local marketing information. Though not displayed, first stage regressions of endogenous price on exogenous variables and instruments exhibit high and significant F-Test statistics for joint significance of all varying variables and joint significance of the instruments. Thus, instrumental relevance is established. Retail variables are always highly significant, TV advertising is significant in specifications with dummies, importantly in the instrumental variables (IV) product dummy specification, the typical IV logit model in column 7. The price coefficient is only significant with no dummies present when estimation is done with OLS, but always significant under the IV procedure. Product characteristics matter and have the expected positive signs when brand dummies are present. In column 6 the size class variable inh3 loses its significance with brand dummies and IV. Now the question arises whether this type of result is a mere artifact of the aggregation procedure

[17] OLS on equation (4.17) without instrumental variables delivers the plain OLS estimates.

[18] The setup details are in appendix table 4.15 and given in short here. # of markets/products: 280/57, price: nominal, geographic aggregation: 3 (2 digit zip, >10,000 sales).

Table 4.5. Logit Market Models for specific Setup

	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Coef./SE	Coef./SE	Coef./SE	Coef./SE	Coef./SE	Coef./SE
preis	–.166***	.010	–.016*	–.280***	–.036***	–.146***
	(.01)	(.01)	(.01)	(.01)	(.01)	(.02)
retail	.098***	.116***	.086***	.094***	.113***	.081***
	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)
ad5	–.001***	.001***	.001***	.000	.001***	.001***
	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)
inh3	.218***	–.070***		.397***	.006	
	(.01)	(.02)		(.02)	(.02)	
liquid	.085***	.275***		–.074**	.220***	
	(.02)	(.02)		(.02)	(.02)	
konzentrat	.352***	.220***		.326***	.208***	
	(.02)	(.02)		(.02)	(.03)	
color	.258***	.162***		.264***	.156***	
	(.02)	(.02)		(.02)	(.02)	
constant	–6.296***	–5.573***	–6.555***	–6.135***	–5.559***	–6.039***
	(.04)	(.04)	(.05)	(.03)	(.04)	(.10)
dummies						
brand	No	Yes	No	No	Yes	No
product	No	No	Yes	No	No	Yes
No. of obs	10,683	10,683	10,683	10,683	10,683	10,683
R^2	.23	.46	.64	.21	.46	.63

Note: Asterisks indicate significance levels at α: $* = 0.05, ** = 0.01, *** = 0.001$. Brand and product dummies are always highly significant. 15 Hausman instruments used. Variable ad5 is total number of contacts within a 3 month period per market. Variable retail is the sum of the variables feature, display priceflag and handbill. For the other variables consult the data section.

outlined in the data section, just chance or a constant finding for most setups. In the following I will present summaries of all 120 models estimated under varying conditions to confirm the results from setup 116.[19]

As it is most important to verify that the significance of the key variables price, retail and advertising is robust across all setups, table 4.6 presents first evidence for this presumption using the six model specifications from table 4.5. This table counts the statistical significance of the coefficient of price, retail and advertising for all 120 setups at different significance levels. Since the setups are not nested, this is the most basic approach to check the general direction of the parameters. Of course, in case of misspecified models, the numbers are less reliable, however I resolve to this approach to compare the models. The table is organized as follows. For each variable, there is a panel of five rows. The first three rows show occurrence of statistical significance. In the last two rows the frequency of positive and negative signs of the coefficient is counted if it was significant at the 5% level.

[19] Nevertheless, the interested reader can consult table 4.16 for the t-values of the variables price, retail and advertising under all 120 setups.

Table 4.6. Frequency of Variable Significance for all 120 Setups

	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
α	Price					
0.050	120	44	46	120	63	48
0.010	120	31	23	120	56	33
0.001	119	19	17	120	45	25
0.05 & Coef. >0	0	18	19	0	12	0
0.05 & Coef. <0	120	26	27	120	51	48
α	Retail					
0.050	111	120	118	115	117	104
0.010	111	118	113	113	114	99
0.001	111	112	109	112	108	98
0.05 & Coef. >0	96	120	118	100	117	104
0.05 & Coef. <0	15	0	0	15	0	0
α	Advertising Totals, 98 day window (ad5)					
0.050	67	93	96	93	95	91
0.010	51	89	88	76	91	86
0.001	38	81	85	61	82	83
0.05 & Coef. >0	24	84	90	73	84	86
0.05 & Coef. <0	43	9	6	20	11	5

Note: Numbers indicate how many out of 120 setups show statistical significance for a given significance level. The last two rows count the sign of the coefficient if it was significant at $\alpha = 0.05$. The models estimated correspond to table 4.5. Ad5 is number of total advertising contacts using a 98 day time window.

The previous results of model 116 are supported. The price coefficient is always significantly negative in the models without dummies, (1) and (4), whereas retail and advertising variables have both signs. When moving to the brand dummy specifications (2) and (5) both retail and advertising signs change to be more convincing: Retail is never negative and significant, advertising has more significant positive signs and less negative ones. In contrast to that, price coefficients are now sometimes positive and significant. For the product dummy specifications (3) and (6), retail loses few of its positive signs in some models, advertising gains positive signs and loses negative ones. Prices have again only negative and significant signs in specification (6), as in the case without dummies, but only for 48 out of 120 models. There is no major change for prices from specification (2) to (3). Focusing on specification (6), prices either have significantly negative or insignificant coefficients, retail has only significant and positive and advertising has mostly significant and positive signs, but also a few negative ones.

A problem might be that due to sparseness of the data, market shares are imprecisely calculated because there are only few purchases for a product. I call these “small” market shares. In table 4.7 I drop all product market shares that have less than 3 purchases

Table 4.7. Frequency of Variable Significance without “small” Market Shares for all 120 Setups

	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
α	Price					
0.050	120	53	40	120	67	36
0.010	119	36	30	120	62	17
0.001	118	29	22	120	53	14
0.05 & Coef. >0	0	22	25	0	9	1
0.05 & Coef. <0	120	31	15	120	58	35
α	Retail					
0.050	92	120	120	95	119	115
0.010	90	117	118	94	116	105
0.001	85	114	114	87	110	93
0.05 & Coef. >0	76	120	120	80	119	115
0.05 & Coef. <0	16	0	0	15	0	0
α	Advertising Totals, 98 day window (ad5)					
0.050	66	80	86	60	84	88
0.010	55	62	76	47	69	76
0.001	40	49	60	29	56	61
0.05 & Coef. >0	10	64	72	36	68	74
0.05 & Coef. <0	56	16	14	24	16	14

Note: Numbers indicate how many out of 120 setups show statistical significance for a given significance level. The last two rows count the sign of the coefficient if it was significant at $\alpha = 0.05$. The models estimated correspond to table 4.5. Ad5 is number of total advert contacts using a 98 day time window.

and recalculate the remaining. If these small market shares would drive the results, we would expect a change relative to table 4.6.[20]

As we can see from comparison of the tables, significance of the variables price and advertising is reduced for some models, but has surprisingly increased for the retail variable. All in all, the results support the main findings from table 4.5.

Since the key variable of interest is advertising, I will vary the construction of the advertising variable. See table 4.14 in the appendix for an overview of the different advertising variables constructed.[21]

Table 4.8 presents first evidence for differently constructed advertising variables for model specification (6) of table 4.5, the IV logit model with product dummies. This specification is the base model for the remaining analysis as it has product dummies

[20] The number of observations dropped depends naturally on the setup. Setups with low thresholds and small markets lose more observations than setups with highly aggregated markets and higher thresholds.

[21] For explanations consult the data section 4.3.2.

and controls for price endogeneity with the Hausman instruments. Hence, it is the most elaborate of the simple logit models.

Table 4.8. Frequency of Advertising Significance for all 120 Setups

	Advertising 98 days time window				
	Pressure				Total
	as is	>0.2	>0.4	>0.6	as is
	ad1	ad2	ad3	ad4	ad5
α	All Shares				
0.050	78	63	15	56	91
0.010	63	43	10	35	86
0.001	45	24	9	21	83
0.05 & Coef. >0	76	50	1	6	86
0.05 & Coef. <0	2	13	5	37	5
α	Small Shares dropped				
0.050	16	18	43	56	88
0.010	8	5	23	46	76
0.001	2	1	10	37	61
0.05 & Coef. >0	14	11	12	12	74
0.05 & Coef. <0	2	7	22	22	14

Note: Numbers indicate how many out of 120 setups show statistical significance for a given significance level. The last two rows count the sign of the coefficient if it was significant at $\alpha = 0.05$. The model estimated is model (6) of table 4.5.

As before, this table counts the statistical significance of the advertising coefficient for all 120 setups at different significance levels. All 5 advertising variables have a 98 day time window. As can be seen in the upper panel, the first (advertising pressure) and last (total contacts) variable have the highest number of significant parameters and also mostly the expected positive sign. The middle variables, the three dummy versions deduced from advertising pressure, show less significance and comparatively many unexpected signs.

As before to account for sparseness in the data and to check the robustness of the results, in the lower panel of table 4.8 all product market shares that are calculated from less than 3 purchases are dropped and shares are recalculated for the remaining. As evident from the table, significance of the advertising pressure variable ad1 is clearly diminished, so are ad2 through ad4, but the effect of advertising variable ad5, the total counts of advertising, stay remarkably high given the changes in the other variables. In the appendix the tables 4.17 and 4.18 present the same comparison for the advertising variables using 56 and 140 days time windows, respectively. These tables also support the use of the total counts for the advertising variable, the versions based on advertising pressure always lose significance once the “small” market shares are dropped.

Therefore, the remainder of the analysis focuses on advertising variables that represent total counts of advertising contacts, although the other measures also have sound economic rationales.

To account for simple nonlinearities the estimations were redone using selected logarithmized advertising variables. See table 4.9 for the evidence. The first three columns are total contacts for all three time windows used and the last three columns are the logarithmized values thereof.

Table 4.9. Frequency of Advertising Significance for all 120 Setups

	Total Counts			Logged Total Counts		
	56	98	140	56	98	140
	ad15	ad5	ad10	ad20	ad18	ad19
α	All Shares					
0.050	92	91	94	98	98	99
0.010	86	86	86	93	94	95
0.001	83	83	82	92	92	89
0.05 & Coef. >0	86	86	86	97	96	95
0.05 & Coef. <0	6	5	8	1	2	4
α	Small Shares dropped					
0.050	88	88	88	84	83	77
0.010	73	76	73	77	74	71
0.001	59	61	61	66	65	60
0.05 & Coef. >0	74	74	74	74	72	64
0.05 & Coef. <0	14	14	14	10	11	13

Note: Numbers indicate how many out of 120 setups show statistical significance for a given significance level. The last two rows count the sign of the coefficient if it was significant at $\alpha = 0.05$. The model estimated is model (6) of table 4.5.

In the top panel both, the advertising variables generated as total counts and the log thereof show a very good performance, with the logarithmized variants being superior: there are less setups with negative significant signs and more with positive signs. If we drop the "small" shares as before, the logarithmized variants lose significant positive signs and gain negative ones so that all get very similar. In the following, as advertising variables the logged total counts variable with 56 day time window (i.e. ad20) and the total counts variable with 98 days time window (i.e. ad5) are used. Both show the smallest number of negative signs of their variants (non-log vs. log) across all models.

Now, to understand more deeply some of the results seen so far and uncover their dependence on the data setup used, a glance at the detailed table 4.16 of the appendix is necessary. There is much room for further elaboration, but I will give the main findings in a nutshell that can be verified by intensive study of the sorted table. I find that generally the setups with the market setups (1) and (4) show low significances or unexpected signs. Note that (1) is the least and (4) the most aggregated market. My presumption is that under (1) data are sparse, i.e. markets are small and the number of observations per product is small or zero so that market shares are imprecisely calculated or missing. Then the market is incomplete that randomness rules and under (4) markets are so large that many interesting effects are averaged out.

Finally, to make the selection for the random coefficients model I again go back to the detailed table 4.16 of the appendix and check the setups that fulfill roughly the following criteria: t-statistics for price, retail and advertising are high relative to all others whether "small" market shares are dropped or not. That amounts to checking the last six columns. The number of products should be as high as possible to represent the whole market so that all narrow product definition setups drop out. Product definition setups 1, 2 and 6 remain. The market should have market shares calculated as precisely as possible, then market setup (1) and (4) drop out. Market setup (3) is superior to (2) because more sales are in each market. Some setups have similar values but only differed in minor setup details. In this case, the setup where price enters as median and retail enters as totals is picked. The following setups fulfill these criteria and will be estimated as random coefficients model: 116, 49, 47, 111. A detailed description is part of the next section.

4.4.2 Random Coefficients Model

The random coefficients model is computationally much more demanding than the simple logit model and due to its nonlinear nature has some known caveats. Following Knittel and Metaxoglou's (2008) recommendation I try several optimization routines and randomized starting values.[22] Additionally, following Dubé, Fox, and Su (2009) I use thight exit criterions on the loops and a state of the art solver.[23]

The benchmark setups for this results section are motivated by the simple logit results as stated there. The setups are 116, 49, 47 and 111. See table 4.10 for a description of the model setups and of market characteristics. For example, in setup 116 the raw number of products is 162, and by setting the product threshold to 200 sold units only 57

[22] Starting values of the outer loop to estimate θ are initialized around a researcher guess with a mean zero random normal distribution with standard deviation of two to reflect the fact that the researcher's a priori knowledge is small, whereas Knittel and Metaxoglou (2008) set it to one. I also randomize the starting values of the inner loop. The starting values of the inner loop to recover the δ_{jt} are constructed as follows: The values are drawn from a normal distribution, where the mean is the δ_{jt} estimate from the simple logit model and the standard deviation being the standard error from the regression equation.

[23] The KNITRO solver from Ziena Inc. outperforms unadjusted Matlab solvers in my tests with the Nevo (2000) cereal data. I used my specific code implementation. However, if I adjust the Matlab large scale `fminunc` solver appropriately by adjusting mainly the termination rules both perform remarkably similar. The threshold for the δ inner-loop is `1e-14` and `1e-8` for the GMM outer-loop. The standard value of Matlab for both is `1e-6`.

Table 4.10. Overview of Market Data Setups

Setup No.		116	49	47	111
		Setup Details			
Market Setup	Location	3 plz2+	3 plz2+	3 plz2+	3 plz2+
	Product	1	1	2	6
Minimum Sales	Market	50	100	100	100
	Product	200	700	700	300
Price	Nominal	1	1	.	1
	p/inh	.	.	1	.
Retail	full	1	.	.	1
		Characteristics			
# Markets	unrestricted	280	280	280	280
	with threshold	280	259	259	259
# Products	unrestricted	162	162	76	108
	with threshold	57	20	26	43
# Brands	unrestricted	15	15	15	15
	with threshold	15	9	10	15

Notes: The Product Setup No. is given in table 4.4. IV code for all setups is 2. Prices are aggregated as median. Retail variable is present in all. Retail full: retail = feature+display+priceflag+handbill, otherwise retail =feature+display.

remain.[24] This reflects the sparseness of the data mentioned earlier in the data section. However, note that still all major 14 brands are present, so only product varieties are lost besides small minor brands that are collected under a joint fifteenth brand name. This holds for both setups 116 and 111, whereas in setups 49 and 47 brands are lost due to the product threshold being high. There is no obvious way how to circumvent the loss of product varieties or brands in these setups.

All setups use the same location aggregation level, i.e. 3 plz2+, which means that geographic locations are aggregated up to the first two digits of the zip code such that at least 10,000 sales are in the location for the whole sampling period.

Intuitively, the setups are characterized as follows. 116 is the low threshold specification that modifies the data as little as possible. Compared to this, setup 49 loses markets and products by imposing higher thresholds to get more reliable data. Reliable is meant in the sense that market shares of the remaining products are calculated with a sufficient number of sales so that the outcome is less erroneous. In addition to the high thresholds of setup 49, setup 47 uses price divided by size as price variable. Compared to 116, setup 111 uses higher thresholds and drops the product characteristic “color” so that products are not distinguished in this dimension. I drop color because it is the least important of the product characteristics in the model. See the data section 4.3.1 for an explanation.

[24] Recall this threshold selects products such that only products that have in the sampling period more than 200 sold units across all markets are used for the estimation.

I estimate each setup as random coefficients logit model with and without observed demographics. The most basic benefit of the random coefficients model is to capture unobserved taste heterogeneity, represented by the sigma parameter.[25] Recall from section 4.2 that the sigma parameter is the standard deviation of the normally distributed random coefficient. Additionally, the random coefficient model captures observed heterogeneity by adding interactions of product characteristic variables and consumer demographics as outlined in section 4.2. I first discuss the results without demographics and then summarize the results for the specifications with demographics. It turns out that estimation of this model is not trivial, even when considering recent recommendations from the relevant literature cited at the beginning of this section.

4.4.2.1 Results without Demographics

Table 4.11. Random Coefficients Model with Setup 116 and 49 without Demographics and logarithmized Advertising Totals (ad20)

	116		49	
Variables	mean	sigma	mean	sigma
constant	-6.265***	0.788	-5.484***	0.345
	(0.264)	(1.789)	(0.339)	(1.802)
price	-0.400	0.220*	-0.636*	0.282*
	(0.216)	(0.099)	(0.286)	(0.114)
retail	0.094	0.013	-0.045	0.219
	(0.053)	(0.325)	(0.366)	(0.324)
ad20	0.051*	1.537	0.073	4.317*
	(0.021)	(3.505)	(0.055)	(2.187)
inh3	0.293***	0.381	1.732***	0.228
	(0.011)	(0.404)	(0.066)	(0.409)
liquid	0.642***	3.771	-7.863***	2.374
	(0.021)	(2.936)	(0.488)	(3.227)
konzentrat	0.729***	1.072	-6.759***	2.187
	(0.025)	(3.206)	(0.381)	(3.871)
color	0.656***	7.818	44.194***	5.754
	(0.019)	(5.712)	(3.287)	(6.045)
GMM Objective	6.6106		1.6948	
Runs, exit if	30		30	
delta X < TolX	24		7	
delta f < TolFun	5		23	
Iter > MaxIter	1		0	
Runtime (min)	4,151		3,008	

Note: Asterisks indicate significance levels at α: $* = 0.05, ** = 0.01, *** = 0.001$. 20 Hausman instruments used.

[25] Note that the sign of the sigma parameter does not matter, as only its absolute value is used in the model, but the parameter is unbounded during the optimization.

Tables 4.11 and 4.12 present the four setups of the random coefficients model without demographics. The results are very different from the simple logit model of section 4.4.1. The means of the variables price, retail and advertising are not anymore significant for all setups. Across all setups, retail is never significant. In all four setups, the means of the product characteristics, i.e. inh, liquid, konzentrat and color, are significant. Interestingly, the unobserved heterogeneity captured by the sigma is only significant for the price and the advertising variable, indicating different responses of consumers to these variables. In setup 116, where thresholds are low, mean advertising and the sigma of price are significant. Therefore, the average market response to advertising is positive. Note that the mean of price is almost significant at the 5% level. The signs of the characteristics are plausible. In setup 49, mean price and sigma are significant. Due to the normal distribution, 1.2% of all price coefficients are in the positive domain.[26] Obviously, this is very low and an intuitive result. The sigma of advertising is significant and in

Table 4.12. Random Coefficients Model with Setup 47 and 111 without Demographics and logarithmized Advertising Totals (ad20)

	47		111	
Variables	mean	sigma	mean	sigma
constant	-5.899***	0.864	-5.292***	0.040
	(0.620)	(2.545)	(0.066)	(1.340)
price	-1.925	0.896	-0.142*	0.009
	(1.130)	(0.611)	(0.071)	(0.156)
retail	0.086	0.081	0.081	0.051
	(0.100)	(0.181)	(0.054)	(0.142)
ad20	0.044	5.205*	0.022	3.747
	(0.049)	(2.617)	(0.039)	(3.044)
inh3			0.032*	0.701
			(0.014)	(0.592)
liquid	0.569*	2.963	-0.059*	4.439
	(0.280)	(2.655)	(0.027)	(4.359)
konzentrat	0.273**	0.114	0.203***	7.189
	(0.093)	(2.249)	(0.031)	(4.515)
color	-7.827***	6.420		
	(1.659)	(5.079)		
GMM Objective	3.5782		5.0865	
Runs, exit if	40		40	
delta X < TolX	10		27	
delta f < TolFun	28		11	
Iter > MaxIter	2		2	
Runtime (min)	3,250		3,353	

Note: Asterisks indicate significance levels at α: $* = 0.05, ** = 0.01, *** = 0.001$. 20 Hausman instruments used.

[26] This is easily calculated as the area in the positive domain under the normal density. Formula: $\Phi(\mu/\sigma)$, where Φ is the standard normal cumulative distribution function, μ is the mean and σ the standard deviation.

magnitude very high, indicating a big dispersion in responses of markets to advertising. Since the mean is insignificant, it can be imagined to be near zero so that the population is divided into two large groups that react positively and negatively to advertising. This finding is not fully in line with setup 116. The signs on the characteristics are significant, but partly have switched signs. The reason is that many products are now dropped from the setup compared to setup 116 due to the higher thresholds and there are less products in the characteristics space. In this case the remaining products are not compared with the dropped products. Naturally, this changes the parameters. In setup 47, mean and sigma of price, now it is the efficiency price, are insignificant, whereas advertising still has a significant sigma, again high in magnitude. The color characteristic has an unintuitive and significant sign. In setup 111, mean price is significant again but no component of advertising.

Summarizing the results without demographics for the advertising variable, the sign of advertising is always right as expected. In particular, the mean of advertising is positive and significant for the most plausible setup 116. There is evidence for heterogeneity indicated by the significant sigmas in setup 49 and 47. The results corroborate that advertising matters in the market model.

So far, the discussion is based on the estimation run for each setup that delivered the lowest GMM objective function value. The lower panel of tables 4.11 and 4.12 show the exit criterions that terminated each run. Although the same algorithm settings are used for all four setups, mostly different exit criterions terminate the estimation. Namely change in the parameter values is below a certain threshold (deltaX<TolX) is the most frequent exit criterion for setups 116 and 111, whereas the little change in the objective function (delta f < TolFun) is most relevant for setups 49 and 47.[27] For no setup the algorithm terminated because the analytical gradient was close to zero. Therefore, it is not obvious that all terminated estimations that end at different GMM objective function values deliver results that are similar to the run with the minimal GMM objective function value presented in tables 4.11 and 4.12.

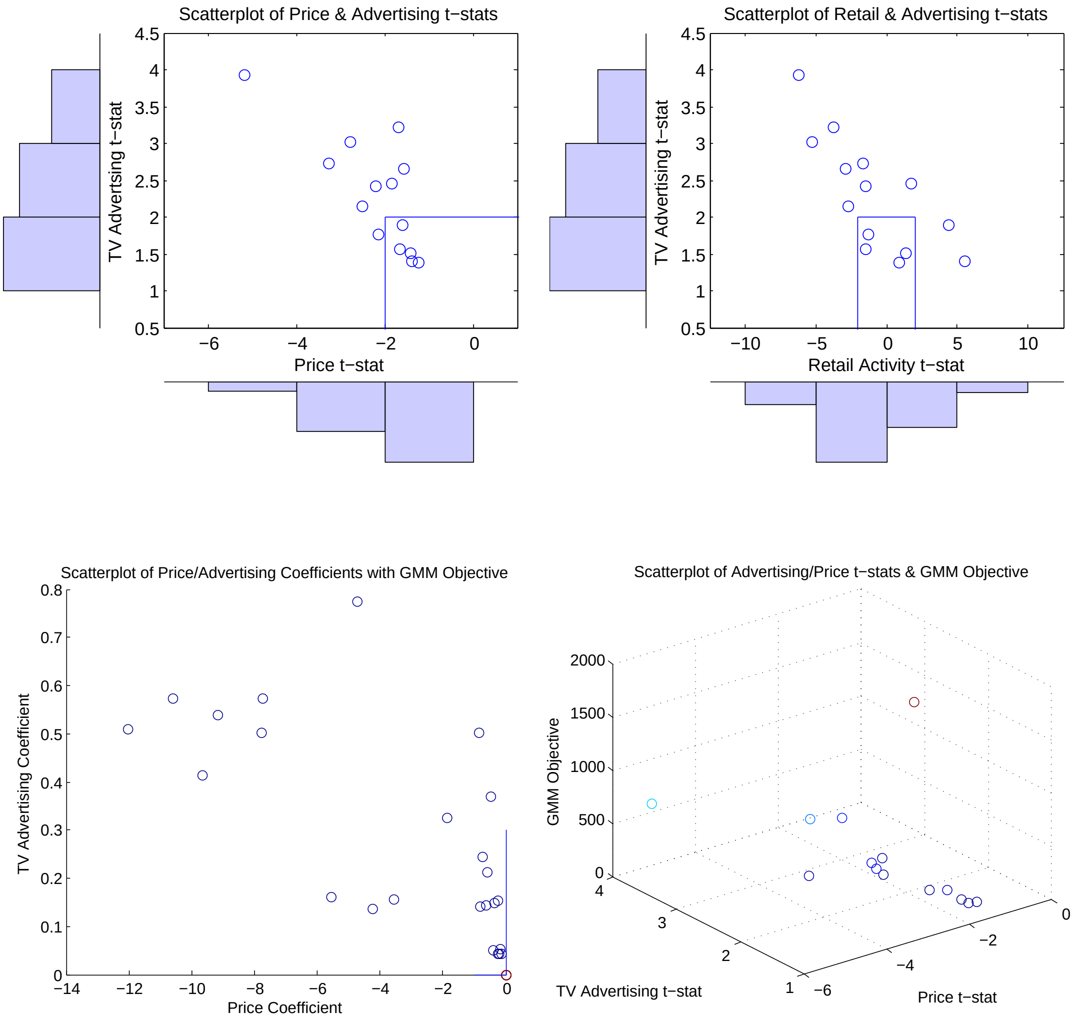
To see whether the results for the means of the variables change for the different estimation runs, I present in the following visualizations of the estimation runs per setup. The figures give an overview of coefficient values, t-statistics and GMM objective function values at the termination of the each optimization. Consult the figures 4.1 and 4.3 to 4.5 for the corresponding graphs. The latter three are found in the appendix. Each figure consists of 4 panels.

The upper left panel displays the t-statistics of the coefficients of the price and TV advertising variable with attached histograms to see the empirical distribution. Note that there is no point for estimation runs that result into a degenerate covariance matrix

[27] Some starting values did not lead to convergence and the estimation terminated because the maximum number of iterations was reached (Iter > MaxIter).

of the parameters. Then t-statistics cannot be calculated and it seems that there are only few points in the scatterplot. A bounding box shows the t-value threshold of 1.96 for statistical significance at the 5% level. Analogously, the upper right panel presents the same information for the retail activity and TV advertising variables. These panels illustrate whether the relation of the coefficients changes across runs. The lower left panel shows coefficient values of the price and TV advertising variable. The darker the circle dot, the lower the GMM objective function value. The lower right panel combines the information of the t-statistics from the upper left panel with the GMM objective function value in a three dimensional scatterplot.

Figure 4.1. Overview of 30 Runs for Setup 116 without Demographics with logarithmized Advertising Variable



I expect that t-statistics get more similar the lower the GMM objective function value is. Intuitively, this corresponds to a graphical representation of convergence that points with similarly low GMM objective function values are close to each other and form clusters.

Now I discuss the figures. In figure 4.1 the panels confirm the previous results from the minimal GMM objective function value estimation. Means of price and TV advertising have always negative and positive signs, respectively. Many values lie outside the bounding box in the top panels. Retail activity shows to be undetermined as values lie in the positive and in the negative domain. This fits to the tabulated run with the minimal objective function value where it is not significant. The two top panels suggest presence of a negative correlation between the t-statistics of the displayed variable. Probably, this is a hint for some underlying interaction between these two variables. The lower left panel shows that although most estimations terminate at similar GMM objective function values, the values of the price and TV advertising coefficient are quite dispersed. The lower right panel demonstrates that price t-statistics are all close to significance at the 5% level for low GMM objective function values and far off for higher values. However, the TV advertising t-statistics are quite dispersed even for low GMM objective function values.

The figures for the remaining three setups are situated in the appendix. In figure 4.3 almost the same patterns show up as for setup 116. The only major difference is that the concentration of values in the two lower panels is even more pronounced than in setup 116. In figure 4.4 the upper panels reveal that less runs end with a significant and positive TV advertising coefficient. The mean of retail activity is now only negative if it is significant. The dark cloud of points in the lower panels show very nicely the concentration of parameter values and t-statistics for low GMM objective function values as in figure 4.3. This concentration is not visible anymore in figure 4.5, but the basic patterns as in the other three figures still remain. Looking at all three figures in the appendix together, it seems that the correlation among t-statistics of the two top panels is not visible anymore.

I calculated all four setups with fewer runs using the advertising variable ad5 (same as ad20 but not logarithmized) and found no differences, therefore I do not report details.

To conclude the discussion of the results without demographics, there is a significant amount of evidence that shows TV advertising to be an important determinant in the model and price effects to be as economic intuition suggests. As expected, the results depend on the number of existing products in the setup. If many products are present product characteristics have positive effects.

4.4.2.2 Results with Demographics

Now I turn to the results of the random coefficients model with demographics. Potential influences of demographics on the parameter estimates of choice relevant variables can be captured by calculating the interactions of both and adding them to the observed heterogeneity part of the model. For example, the price reaction of consumers may be smaller for high incomes so that a positive sign on the new interaction coefficient is expected. Household size is interacted with the contents variable and the concentrate dummy, as bigger households with more members have higher demand for detergent. Bigger households are expected to buy bigger contents packages and are less likely to buy concentrates that only come in very small packages. For the first household effect I expect a positive sign, for the latter a negative sign.

Table 4.13. Random Coefficients Model with Setup 116 with Demography and logarithmized Advertising Totals (ad20)

			interaction with demographics	
Variables	mean	sigma	income	household size
constant	-2.545***	0.499	0.587	
	(0.213)	(5.685)	(12.533)	
price	-0.422	0.256	-0.156	
	(0.435)	(0.190)	(0.668)	
retail	-0.003	0.046	0.587	
	(0.312)	(0.419)	(0.827)	
ad20	0.083	2.853	12.672	
	(0.136)	(10.307)	(28.450)	
inh3	0.290***	0.080		1.633
	(0.013)	(0.648)		(3.299)
liquid	0.620***	1.746		
	(0.027)	(3.455)		
konzentrat	0.603***	1.098		-8.837
	(0.031)	(7.305)		(11.127)
color	0.641***	10.930		
	(0.025)	(20.845)		
GMM Objective		1.5638		
Runs, exit if		9		
delta X < TolX		5		
delta f < TolFun		1		
Iter > MaxIter		3		
Runtime (min)		1,666		

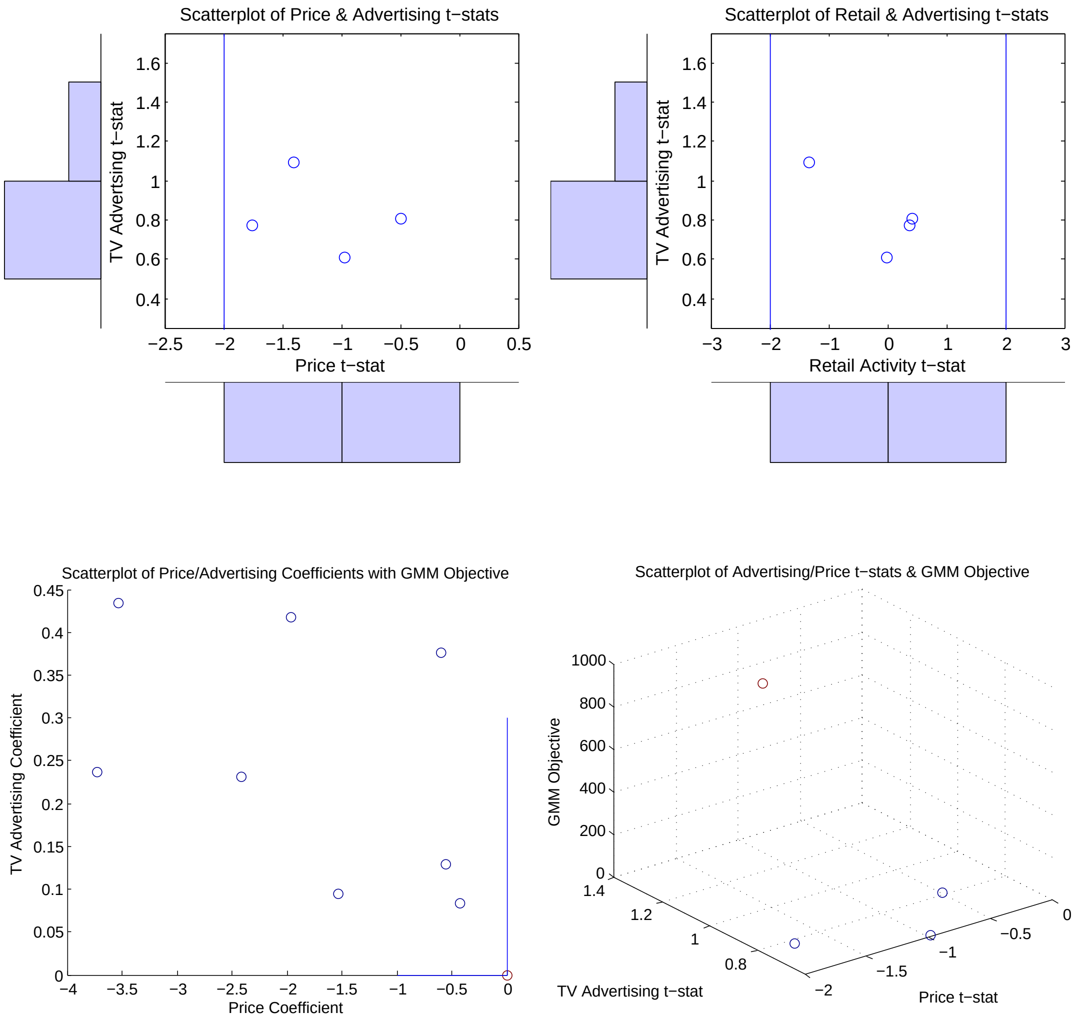
Note: Asterisks indicate significance levels at α: $* = 0.05, ** = 0.01, *** = 0.001$. 20 Hausman instruments used.

Tables 4.13 and 4.19 to 4.21 present results of the four chosen setups from table 4.10 with demographic variables. The latter three tables are placed in the appendix. As in the previous section, the results are displayed for the run with the minimal GMM objective function value. Table 4.13 presents the result for setup 116. Different from the model without demographics, no mean or sigma of the variables price, retail or TV advertising is significant. The interaction of demographics with selected variables are as well all insignificant. This holds true for all four setups. Only the product characteristics are significant as in the case without demographics, where the signs of the means make intuitive sense only for setup 116. It seems that the introduction of demographics is not working well, maybe due to high correlation among the variables, or markets may not respond strongly or differently to the price, retail and advertising variables. It seems that choice of products is strongly driven by product characteristics.

Concerning the exit criterions of the estimations the same patterns as without demographics are visible, with the difference that total computation time goes sharply up and memory requirements lead to difficulties in the estimation, e.g. that is why there are only 9 runs for setup 116 with demographics.

To see the results for all estimation runs for each setup, I use the same type of visualization as already explained in the previous case without demographics. Figures 4.2 and 4.6 to 4.8 present the corresponding results, where the latter three figures are placed in the appendix.

Figure 4.2. Overview of 9 Runs for Setup 116 with Demographics with logarithmized Advertising Variable


Scatterplot of Price & Advertising t−stats

Scatterplot of Retail & Advertising t−stats

Scatterplot of Price/Advertising Coefficients with GMM Objective

Scatterplot of Advertising/Price t−stats & GMM Objective

In figure 4.2 the top panels reveal that for no run a significant parameter is found. The lower panels show the dispersion of coefficients despite similar GMM objective function values. As evident from the number of points in the top panel scatterplots, for only 4 out of 9 runs a covariance matrix of the parameters was computable. Turning to the remaining three figures 4.6 to 4.8, the results differ from setup 116. There the top panels show the means of price, advertising and retail to be sometimes significant. Checking the lower panel of figure 4.6 it is visible that significance occurs at low and high levels of the GMM objective function value, illustrated by the red, yellow and the dark blue points, respectively. In the lower panel of figure 4.7 significance of mean advertising is very often close to the 5% level, mean prices cross the threshold for significance. Of all 4 figures, only in this one the concentration of similar t-statistics for low GMM objective function values is as in the case without demographics. In figure 4.8 the lower panel shows that advertising is never significant at low GMM objective function values but price is so only once.

All in all, I can conclude that the interpretation of the figures for the model with demographics counteracts the discouraging results of the run with the minimal GMM objective function value. Although the results are not stable, comparing the result to the model without demographics, the results are in line with the findings, but are not very supportive. The results for the demographics do not change when local advertising is removed or when the specification in terms of demographics is modified. Thus, there seems to be a problem with the demographics. Maybe the level of detail of the variables is not high enough or there is simply no segmentation of the local detergent markets with demographics possible. The latter could mean that markets are similar in terms of demographics or that demographics simply do not explain market differences. Introducing them in this case seems to “disturb” the rest of the model.

Partly, the results may also be driven by the dataset conditions: Due to sparse data (i.e. not enough consumers per market) the construction of precise market shares during the aggregation is not possible so that the model is not able to identify all relations in the data. Therefore, the results seem sensitive to the specification and the chosen setups, but a couple of findings occurred repeatedly: Significant positive means and standard deviations for the TV advertising variable.

4.4.3 Discussion of Details for other Model Variants

Apart from the estimation results presented above I experimented with the model in many dimensions, but it was not feasible to make an exhaustive analysis of each variant due to computational restrictions. Experimenting is necessary, because the researcher has many degrees of freedom when building the data setup for the demand model and his choice impacts the findings. In that sense, the flexibility of the model is a curse at the same time and a source of intransparency. I want to emphasize that these issues are typically not discussed in the results oriented literature. In the following I summarize some of the variants.

Demographics Apart from the reported results in the previous section I estimate random coefficients models with the following demographic variables: income, household size, a dummy for presence of children under 16 years and age of the household managing person. I try several specifications with these four variables where I use all or a subset thereof. In doing this, I choose different interactions of the demographic variables with the other variables, e.g. price, advertising, retail activity and product characteristics. I get no significant results. I also estimate a version without local marketing activity and remove thereby all variables but price that vary across markets, but this does not render interactions with demographics significant.

Brand Dummies I replace product by brand dummies, but this shows no changes concerning the coefficients on price, retail or TV advertising. As mentioned before, with brand dummies more assumptions are required for the procedure to be valid. Since the product dummies work, I follow the recommendation of Nevo (2000) and use the product dummy approach.[28]

Price and Retail Variable During the aggregation, the researcher must choose how to aggregate prices. I switch between taking the median or arithmetic mean of a product price in a market. Recall that I have this choice as market prices are aggregated from the transaction price of individual choices. In the simple logit model, I find no meaningful difference, e.g. see setups 49/63 in table 4.16.

I change the scaling of the retail variable from totals per product and market to means per product and market. When doing this, the change renders the formerly highly significant variable to completely insignificant, e.g. setups 93/95 and 89/91 in table 4.16 illustrate this for the simple logit model.

If prices enter as efficiency prices defined as price divided by size, I estimate a version with the true continuous size variable and one version with an integer version of the size variable where the size measure is rounded to the nearest positive integer. This in turn is done with both prices as mean and median so that four versions with efficiency prices are estimated. As can be seen from table 4.16 of the simple logit model in the appendix, comparing specifications that only differ in the construction of this variable reveal that significance of prices is much higher when using the continuous size variable, e.g. setups 41/47, 55/61, 44/48 and 58/62.

Instruments I vary the number of instruments. Since the models have varying number of parameters I use globally 20 instruments to ensure identification for each model. When using less instruments and moving from over to just identification, I observe that in general statistical significance diminishes and the GMM objective function achieves lower values at termination. But in terms of presented results, no interesting changes

[28] See the mentioned publication on p. 527, bottom.

take place. I also vary the construction of instruments as presented before in the model section. The primitive procedure 1 results in different t-statistics for the price coefficient than when using the candidate procedure 2. The change is not always in the same direction. See setups 1/2, 3/4, 5/6, 7/8, 11/12, 13/14 or 15/16 in table 4.16 for mixed evidence.

Missing Product Sales In the presented models the estimation is done with a varying number of products per market. Commonly, the random coefficients model is estimated for simplicity with the same number of products per markets by using a subsample of the market. Given the data in this work, the latter requires some imputation in terms of missing price and sales data. I experiment with this, but abandon this approach, as the results are highly dependent on the imputation procedure. Given a subsample of the products, it is possible to generate imputed market shares, prices and media exposure for non-purchased products. Imputed market shares are a constant factor of the smallest market share in a given market, a typical constant being 10e-3. Intuitively, the model got as information that the product was not sold a lot, which approximately is right as it was not purchased in the sample. Prices and retail are set to the mean of the current quarter across all markets. Following this procedure, significant of price, retail and advertising are even higher, but the effect varies non-monotonically with the chosen constant.

4.5 Conclusion

This chapter investigates the question whether local marketing activity that stays typically uncontrolled in commonly estimated market models is statistically significant once it is introduced into a standard model. The models used are the simple logit model, suffering from the well-known IIA problem, and the random coefficients logit model that alleviates this restriction and was augmented additionally by consumer demographics of each market. The employed dataset for the analysis is generated using a German national panel of consumers for which detailed detergent purchase and media exposure information is available. From this panel I construct geographically distinct markets that have in terms of media exposure superior information compared to commonly used datasets, but in terms of sales suffer of information sparseness due to the little number of consumers per market. I build 120 data setups to do the aggregation to markets with sufficient care so that results are not driven by lucky or random construction of the data and to have the sparseness problems controllable. I find that under the simple logit model all setups that do not suffer from sparseness or over-aggregation report highly significant effects for local retail activity and local TV advertising on the market level. Of several constructed advertising measures the versions counting total advertisement contacts are better than advertising pressure measures. The results remain very stable when imprecisely and small market shares are dropped.

The random coefficients logit model is estimated on four benchmark setups out of the 120 available ones that have performed very well in the simple logit model. If the model is estimated without demographics, the results for advertising point into the same direction as in the simple logit model and additionally suggest that there is heterogeneity across markets in the response to advertising exposure. Retail activity does not prove its importance as in the simple logit model and is mostly insignificant or the evidence is contradictory. Compared to the simple logit results the evidence is not as sharp. In the case with demographics the results compared to the model without demographics are weaker, while still pointing into the same direction. Demographics are not significant for the responses of the consumer and seem to disturb the "rest" of the model. Presumably, the level of detail in the data is not sufficiently high to allow such rich modeling.

To sum up, local marketing activity matters and must be accounted for when estimating market models where products are being advertised.

Appendix

Appendix A: Berry's (1994) Model

Berry starts from the general utility specification in (4.1). Omit the demographic information and the market subscripts t for the exposition. Vectors containing values for all products j are bold. Consumer i chooses product j instead of any k given x, ξ, p if:

$$U(x_j, \xi_j, p_j, \nu_i; \theta) > U(x_k, \xi_k, p_k, \nu_i; \theta) \quad \forall j \neq k \tag{4.18}$$

ν_i is the only remaining random variable. Assume that utility U takes the following form and can be decomposed to a deterministic and random part: $u_{ij} = \delta_j + \nu_{ij}$. Collect the vectors $\boldsymbol{\nu_i}$ that lead to choice of product j in a set:

$$A_j(\boldsymbol{\delta}) = \{\boldsymbol{\nu_i} | \delta_j + \nu_{ij} > \delta_k + \nu_{ik}, \quad \forall j \neq k\} \tag{4.19}$$

The market share of good j is given by $P(\boldsymbol{\nu_i} \in A_j(\boldsymbol{\delta}))$, the probability of $\boldsymbol{\nu_i}$ being in the specified set. If we assume a parametric form for ν (that does not differ across i), we can get the cdf $F(\nu, \boldsymbol{x}, \sigma_\nu)$ and density $f(\nu, \boldsymbol{x}, \sigma_\nu)$. The functions may depend on the characteristics $\boldsymbol{x}$ and the parameters of the distribution σ_ν because all deviations from the market mean $\boldsymbol{\delta}$ are captured by ν, e.g. random coefficients. The market share has the following form where it can be written as a function of $\boldsymbol{\delta}$:

$$S_j(\boldsymbol{\delta}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\xi}, \theta)) = \int\limits_{A_j(\boldsymbol{\delta})} f(\nu, \boldsymbol{x}, \sigma_\nu) d\nu, \qquad \forall j \tag{4.20}$$

Suppose given data on characteristics and prices $(\boldsymbol{x}, \boldsymbol{p})$, unobservables $\boldsymbol{\xi}$, parameters θ and known distribution of ν (forget σ_ν for the moment), market share is a function of $\boldsymbol{\delta}$. At the true values of the market shares $\boldsymbol{s}$ and $\boldsymbol{\delta}$ it must hold:

$$s_j = S_j(\boldsymbol{\delta}) \longleftrightarrow \boldsymbol{\delta} = S_j^{-1}(s_j) \quad \forall j \tag{4.21}$$

Berry shows that this inversion works if $S_j(\boldsymbol{\delta})$ is continuously differentiable, and the following restrictions on derivatives hold: $\frac{\partial S_j}{\partial \delta_j} > 0$ and $\frac{\partial S_j}{\partial \delta_k} < 0, \forall j \neq k$. Thus, the procedure needs monotonicity of the market share in the mean effect δ.

This leads to the following procedure for estimation:

1. Calculate $\forall j \quad \widehat{\boldsymbol{\delta}} = S_j^{-1}(s_j)$ from observed data of market shares s_j.

2. Regress with instrumental variables the mean utility level on prices and controls: $\widehat{\delta_j} = x_j\beta - \alpha p_j + \xi_j$. The specification depends on the parametric form of $U()$.

This works under two sets of assumptions: (a) independence across markets r, s: ξ_{js} independent of ξ_{jr}, but ξ_{jr} not independent ξ_{kr} (b) independence across firms in one market: ξ_j independent ξ_s.

Since ν is commonly unknown, Berry suggests to specify a parametric family for ν with parameters σ_ν, e.g. the normal distribution. Then $f(\nu, \boldsymbol{x}, \sigma_\nu)$ as above and $S_j = S_j(\boldsymbol{\delta}, \sigma_\nu)$, $\boldsymbol{\delta} = \boldsymbol{\delta}(s, \sigma_\nu)$, so both depend on choice of the parameter σ_ν that is estimated along with the rest of the parameters in the model. In section 4.2 I sketch the estimation steps. More details and variants with the random coefficients model as example are found in Berry's (1994) paper.

Appendix B: Tables and Figures

Model Setup Information

Table 4.14. Overview of Advertising Variable Definitions

Advertising Variable No.	Time Window in days	Total Counts	Relative Pressure	Log	Dummy if Pressure > threshold
ad1	98	.	1	.	.
ad2	98	.	.	.	0.2
ad3	98	.	.	.	0.4
ad4	98	.	.	.	0.6
ad5	98	1	.	.	.
ad6	140	.	1	.	.
ad7	140	.	.	.	0.2
ad8	140	.	.	.	0.4
ad9	140	.	.	.	0.6
ad10	140	1	.	.	.
ad11	56	.	1	.	.
ad12	56	.	.	.	0.2
ad13	56	.	.	.	0.4
ad14	56	.	.	.	0.6
ad15	56	1	.	.	.
ad16	140	.	1	1	.
ad17	56	.	1	1	.
ad18	98	1	.	1	.
ad19	140	1	.	1	.
ad20	56	1	.	1	.

Note: If a variable x is logarithmized, it is transformed monotonically by log(x+1) to get well defined values even if advertising total or pressure is zero.

Table 4.15. Overview of Market Data Setups

Setup	Market Setup		Minimum Sales		IV	Individuals	Price				Retail		
No.	Location	Product	Market	Product	Setup		nominal	p/inh3	p/inh	mean	in	full	mean
5	1 kreis	1	30	300	1	20	1	.	.	.	1	1	.
6	1 kreis	1	30	300	2	20	1	.	.	.	1	1	.
105	1 kreis	1	30	400	2	20	1	.	.	.	1	1	.
1	1 kreis	1	30	500	1	20	1	.	.	.	1	1	.
2	1 kreis	1	30	500	2	20	1	.	.	.	1	1	.
106	1 kreis	1	30	600	2	20	1	.	.	.	1	1	.
7	1 kreis	1	30	700	1	20	1	.	.	.	1	1	.
8	1 kreis	1	30	700	2	20	1	.	.	.	1	1	.
117	1 kreis	1	50	200	2	100	1	.	.	.	1	1	.
3	1 kreis	1	60	500	1	20	1	.	.	.	1	1	.
4	1 kreis	1	60	500	2	20	1	.	.	.	1	1	.
9	1 kreis	1	60	700	1	20	1	.	.	.	1	1	.
10	1 kreis	1	60	700	2	20	1	.	.	.	1	1	.
17	1 kreis	2	30	500	2	30	.	1	.	.	1	1	.
21	1 kreis	2	30	500	2	30	.	1	.	1	1	1	.
22	1 kreis	2	30	500	2	30	.	.	1	.	1	1	.
23	1 kreis	2	30	500	2	30	.	.	1	1	1	1	.
24	1 kreis	2	30	500	2	30	.	1	.	.	1	1	1
79	1 kreis	2	30	700	2	50	.	.	1	1	1	.	1
18	1 kreis	2	30	1000	2	30	.	1	.	.	1	1	.
80	1 kreis	2	30	1000	2	50	.	.	1	1	1	.	1
81	1 kreis	2	30	2000	2	50	.	.	1	1	1	.	1
82	1 kreis	2	40	700	2	50	.	.	1	1	1	.	1
83	1 kreis	2	40	1000	2	50	.	.	1	1	1	.	1
84	1 kreis	2	40	2000	2	50	.	.	1	1	1	.	1
85	1 kreis	2	50	700	2	50	.	.	1	1	1	.	1
86	1 kreis	2	50	1000	2	50	.	.	1	1	1	.	1
87	1 kreis	2	50	2000	2	50	.	.	1	1	1	.	1

Notes: The Product Setup No. of column 3 is given in table 4.4.

Table 4.15. (continued...)

Setup	Market Setup		Minimum Sales		IV	Individuals	Price				Retail		
No.	Location	Product	Market	Product	Setup		nominal	p/inh3	p/inh	mean	in	full	mean
20	1 kreis	2	100	500	2	30	.	1	.	.	1	1	.
19	1 kreis	2	100	1000	2	30	.	1	.	.	1	1	.
11	1 kreis	5	30	500	1	20	.	.	1	.	1	1	.
12	1 kreis	5	30	500	2	20	.	.	1	.	1	1	.
97	1 kreis	5	40	700	2	50	.	.	1	1	1	.	.
99	1 kreis	5	40	700	2	50	.	.	1	1	1	.	1
98	1 kreis	5	40	1000	2	50	.	.	1	1	1	.	.
100	1 kreis	5	40	1000	2	50	.	.	1	1	1	.	1
101	1 kreis	5	50	700	2	50	.	.	1	1	1	.	.
103	1 kreis	5	50	700	2	50	.	.	1	1	1	.	1
102	1 kreis	5	50	1000	2	50	.	.	1	1	1	.	.
104	1 kreis	5	50	1000	2	50	.	.	1	1	1	.	1
13	1 kreis	6	30	500	1	20	1	.	.	.	1	1	.
14	1 kreis	6	30	500	2	20	1	.	.	.	1	1	.
120	1 kreis	6	50	200	2	100	1	.	.	.	1	1	.
15	1 kreis	7	30	500	1	20	1	.	.	.	1	1	.
16	1 kreis	7	30	500	2	20	1	.	.	.	1	1	.
115	2 plz2	1	50	200	2	50	1	.	.	.	1	1	.
29	2 plz2	1	70	700	2	50	1	.	.	.	1	.	.
30	2 plz2	1	70	1000	2	50	1	.	.	.	1	.	.
107	2 plz2	1	100	400	2	50	1	.	.	.	1	1	.
108	2 plz2	1	100	600	2	50	1	.	.	.	1	1	.
35	2 plz2	2	50	700	2	30	.	1	.	.	1	1	.
51	2 plz2	2	50	700	2	50	.	1	.	.	1	.	.
65	2 plz2	2	50	700	2	50	.	1	.	1	1	.	.
73	2 plz2	2	50	700	2	50	.	.	1	1	1	.	.
76	2 plz2	2	50	700	2	50	.	.	1	1	1	.	1
36	2 plz2	2	50	1000	2	30	.	1	.	.	1	1	.
52	2 plz2	2	50	1000	2	50	.	1	.	.	1	.	.

Notes: The Product Setup No. of column 3 is given in table 4.4.

Table 4.15. (continued...)

Setup	Market Setup		Minimum Sales		IV	Individuals	Price				Retail		
No.	Location	Product	Market	Product	Setup		nominal	p/inh3	p/inh	mean	in	full	mean
66	2 plz2	2	50	1000	2	50	.	1	.	1	1	.	.
74	2 plz2	2	50	1000	2	50	.	.	1	1	1	.	.
77	2 plz2	2	50	1000	2	50	.	.	1	1	1	.	1
75	2 plz2	2	50	2000	2	50	.	.	1	1	1	.	.
78	2 plz2	2	50	2000	2	50	.	.	1	1	1	.	1
53	2 plz2	2	70	700	2	50	.	1	.	.	1	.	.
67	2 plz2	2	70	700	2	50	.	1	.	1	1	.	.
54	2 plz2	2	70	1000	2	50	.	1	.	.	1	.	.
68	2 plz2	2	70	1000	2	50	.	1	.	1	1	.	.
71	2 plz2	2	70	2000	2	50	.	1	.	.	1	.	.
72	2 plz2	2	70	2000	2	50	.	1	.	1	1	.	.
38	2 plz2	3	50	700	2	30	.	1	.	.	.	1	.
40	2 plz2	4	50	700	2	30	.	1	.	.	.	1	.
93	2 plz2	5	50	700	2	50	.	.	1	1	1	.	.
95	2 plz2	5	50	700	2	50	.	.	1	1	1	.	1
94	2 plz2	5	50	1000	2	50	.	.	1	1	1	.	.
96	2 plz2	5	50	1000	2	50	.	.	1	1	1	.	1
118	2 plz2	6	50	200	2	50	1	.	.	.	1	1	.
116	3 plz2+	1	50	200	2	100	1	.	.	.	1	1	.
109	3 plz2+	1	100	300	2	100	1	.	.	.	1	1	.
110	3 plz2+	1	100	500	2	100	1	.	.	.	1	1	.
31	3 plz2+	1	100	700	2	200	1	.	.	.	1	.	.
49	3 plz2+	1	100	700	2	200	1	.	.	.	1	.	.
63	3 plz2+	1	100	700	2	200	1	.	.	1	1	.	.
32	3 plz2+	1	100	1000	2	200	1	.	.	1	1	.	1
41	3 plz2+	2	100	700	2	200	.	1	.	.	1	.	.
47	3 plz2+	2	100	700	2	200	.	.	1	.	1	.	.
55	3 plz2+	2	100	700	2	200	.	1	.	1	1	.	.
61	3 plz2+	2	100	700	2	200	.	.	1	1	1	.	.

Notes: The Product Setup No. of column 3 is given in table 4.4.

Table 4.15. (continued...)

Setup	Market Setup		Minimum Sales		IV	Individuals	Price				Retail		
No.	Location	Product	Market	Product	Setup		nominal	p/inh3	p/inh	mean	in	full	mean
44	3 plz2+	2	100	1000	2	200	.	1	.	.	1	.	.
48	3 plz2+	2	100	1000	2	200	.	.	1	.	1	.	.
58	3 plz2+	2	100	1000	2	200	.	1	.	1	1	.	.
62	3 plz2+	2	100	1000	2	200	.	.	1	1	1	.	.
64	3 plz2+	2	100	1000	2	200	.	.	1	1	1	.	.
69	3 plz2+	2	100	2000	2	200	.	1	.	.	1	.	.
70	3 plz2+	2	100	2000	2	200	.	1	.	1	1	.	.
42	3 plz2+	3	100	700	2	200	.	1	.	.	.	.	.
56	3 plz2+	3	100	700	2	200	.	1	.	1	.	.	.
45	3 plz2+	3	100	1000	2	200	.	1	.	.	.	.	.
59	3 plz2+	3	100	1000	2	200	.	1	.	1	.	.	.
25	3 plz2+	4	50	200	2	200	.	.	1	1	.	.	.
43	3 plz2+	4	100	700	2	200	.	1	.	.	.	.	.
57	3 plz2+	4	100	700	2	200	.	1	.	1	.	.	.
46	3 plz2+	4	100	1000	2	200	.	1	.	.	.	.	.
60	3 plz2+	4	100	1000	2	200	.	1	.	1	.	.	.
26	3 plz2+	5	50	200	2	200	.	.	1	1	1	.	.
89	3 plz2+	5	100	700	2	200	.	.	1	1	1	.	.
91	3 plz2+	5	100	700	2	200	.	.	1	1	1	.	1
90	3 plz2+	5	100	1000	2	200	.	.	1	1	1	.	.
92	3 plz2+	5	100	1000	2	200	.	.	1	1	1	.	1
119	3 plz2+	6	50	200	2	100	1	.	.	.	1	1	.
111	3 plz2+	6	100	300	2	100	1	.	.	.	1	1	.
112	3 plz2+	6	100	500	2	100	1	.	.	.	1	1	.
50	3 plz2+	6	100	1000	2	200	1	.	.	.	1	.	.
113	3 plz2+	7	100	300	2	100	1	.	.	.	1	1	.
114	3 plz2+	7	100	500	2	100	1	.	.	.	1	1	.
27	4 kreis2	1	100	400	2	200	1	.	.	.	1	1	.
33	4 kreis2	2	100	700	2	200	.	1	.	.	1	1	.

Notes: The Product Setup No. of column 3 is given in table 4.4.

Table 4.15. (continued...)

Setup	Market Setup		Minimum Sales		IV	Individuals	Price				Retail		
No.	Location	Product	Market	Product	Setup		nominal	p/inh3	p/inh	mean	in	full	mean
34	4 kreis2	2	100	1000	2	200	.	1	.	.	1	1	.
37	4 kreis2	3	100	1000	2	200	.	1	.	.	.	1	.
39	4 kreis2	4	100	1000	2	200	.	1	.	.	.	1	.
28	4 kreis2	6	100	400	2	200	1	.	.	.	1	1	.
88	4 kreis2	7	100	400	2	200	1	.	.	.	1	1	.

Notes: The Product Setup No. of column 3 is given in table 4.4.

Simple Logit Model Results

Table 4.16. Overview of t-statistics for Price, Retail and Advertising for all Market Setups given two different Minimal Sales Requirements per Product

Setup	Market Setup		Minimum Sales		Min 1 product sold in market			Min 3 products sold in market		
No.	Location	Product	Market	Product	price	retail	advertising (ad5)	price	retail	advertising (ad5)
5	1 kreis	1	30	300	−3.51	16.96	−2.33	−2.89	2.79	−2.60
6	1 kreis	1	30	300	−2.83	16.63	−2.54	−0.94	3.31	−2.41
105	1 kreis	1	30	400	−2.23	15.40	−0.85	−0.58	3.79	−1.14
1	1 kreis	1	30	500	−2.33	15.41	0.17	−0.06	3.44	−0.22
2	1 kreis	1	30	500	−2.42	14.85	0.29	−1.57	3.12	−0.41
106	1 kreis	1	30	600	−0.73	14.57	0.22	−0.93	3.55	−0.23
7	1 kreis	1	30	700	−0.33	14.89	1.66	−0.11	5.01	−0.03
8	1 kreis	1	30	700	−1.30	14.71	1.71	0.30	4.90	−0.05
117	1 kreis	1	50	200	−2.75	10.43	−4.02	−2.93	9.50	−6.04
3	1 kreis	1	60	500	−1.24	13.25	−0.75	−2.39	4.68	−1.42
4	1 kreis	1	60	500	−0.77	13.26	−0.66	−1.41	5.08	−1.37
9	1 kreis	1	60	700	−1.08	11.52	1.28	−1.96	5.51	−0.83
10	1 kreis	1	60	700	−0.16	11.43	1.57	0.55	6.33	−0.69
17	1 kreis	2	30	500	−0.14	15.73	−1.83	−0.13	3.75	−2.81
21	1 kreis	2	30	500	−0.09	16.16	−1.83	−0.03	3.90	−2.8
22	1 kreis	2	30	500	−2.08	0.86	0.76	0.71	3.26	−2.14
23	1 kreis	2	30	500	−2.08	0.95	0.81	0.86	3.62	−2.18
24	1 kreis	2	30	500	−0.56	5.35	0.97	−0.29	3.93	−2.30
79	1 kreis	2	30	700	−1.95	2.23	1.83	−1.10	2.44	−1.48
18	1 kreis	2	30	1000	1.38	12.00	1.68	0.35	3.81	0.82
80	1 kreis	2	30	1000	−2.25	0.02	3.92	−1.19	2.05	1.68
81	1 kreis	2	30	2000	1.52	2.14	5.33	1.24	3.27	3.10
82	1 kreis	2	40	700	−1.70	1.83	3.10	−0.85	2.16	−0.03
83	1 kreis	2	40	1000	−3.66	−1.56	3.42	−2.26	0.65	2.15
84	1 kreis	2	40	2000	−1.07	1.88	4.98	0.35	2.82	2.84

Notes: The Product Setup No. of column 3 is given in table 4.4. The advertising variable is explained in table 4.14.

Table 4.16. (continued...)

Setup	Market Setup		Minimum Sales		Min 1 product sold in market			Min 3 products in market		
No.	Location	Product	Market	Product	price	retail	advertising (ad5)	price	retail	advertising (ad5)
85	1 kreis	2	50	700	−0.47	1.61	2.86	0.02	2.22	−0.53
86	1 kreis	2	50	1000	−2.40	−0.64	2.12	−1.03	1.57	0.5
87	1 kreis	2	50	2000	−0.87	1.93	4.79	0.12	3.30	2.85
20	1 kreis	2	100	500	0.61	12.04	−1.91	−1.43	5.33	−2.67
19	1 kreis	2	100	1000	0.59	9.48	−1.08	1.89	5.51	−0.51
11	1 kreis	5	30	500	−2.41	11.07	1.74	−0.16	2.62	−0.24
12	1 kreis	5	30	500	−0.96	10.11	1.59	0.23	2.42	−0.04
97	1 kreis	5	40	700	−1.30	11.20	3.89	0.52	3.28	1.89
99	1 kreis	5	40	700	−1.29	1.57	6.57	0.59	2.36	2.55
98	1 kreis	5	40	1000	0.10	11.58	4.07	−1.54	3.26	1.92
100	1 kreis	5	40	1000	0.12	2.74	6.57	−1.39	1.80	2.57
101	1 kreis	5	50	700	−0.72	9.19	3.18	1.32	3.90	1.69
103	1 kreis	5	50	700	−0.93	0.92	5.68	1.27	2.31	2.54
102	1 kreis	5	50	1000	0.74	9.95	3.35	0.77	4.24	1.6
104	1 kreis	5	50	1000	0.95	2.31	5.8	0.84	2.09	2.48
13	1 kreis	6	30	500	−3.30	15.34	1.00	−1.43	3.09	0.74
14	1 kreis	6	30	500	−2.03	15.68	1.01	−1.74	2.96	0.76
120	1 kreis	6	50	200	−2.44	11.96	−4.2	−2.51	10.98	−6.07
15	1 kreis	7	30	500	−3.78	14.70	−0.57	−1.23	3.55	−0.39
16	1 kreis	7	30	500	−3.78	14.70	−0.57	−1.23	3.55	−0.39
115	2 plz2	1	50	200	−4.03	40.43	2.35	−2.36	22.21	−0.54
29	2 plz2	1	70	700	−1.47	23.13	7.04	0.20	16.35	3.89
30	2 plz2	1	70	1000	1.63	16.21	4.92	2.33	15.29	4.31
107	2 plz2	1	100	400	−1.20	27.99	7.01	−0.81	19.71	3.16
108	2 plz2	1	100	600	0.75	25.78	6.23	1.64	19.68	4.71
35	2 plz2	2	50	700	−0.77	30.06	5.75	−1.98	18.79	1.52
51	2 plz2	2	50	700	−1.31	27.47	7.33	−2.20	15.91	2.31
65	2 plz2	2	50	700	−1.40	27.95	7.36	−2.39	16.12	2.43
73	2 plz2	2	50	700	−2.81	20.49	7.3	−1.13	13.48	2.32

Notes: The Product Setup No. of column 3 is given in table 4.4. The advertising variable is explained in table 4.14.

Table 4.16. (continued...)

Setup	Market Setup		Minimum Sales		Min 1 product sold in market			Min 3 products in market		
No.	Location	Product	Market	Product	price	retail	advertising (ad5)	price	retail	advertising (ad5)
76	2 plz2	2	50	700	−3.47	0.77	8.34	−1.53	2.76	2.90
36	2 plz2	2	50	1000	−1.02	23.92	7.47	−0.23	18.22	3.02
52	2 plz2	2	50	1000	−1.32	22.49	8.85	−0.40	16.06	3.93
66	2 plz2	2	50	1000	−1.38	22.78	8.93	−0.74	16.31	3.95
74	2 plz2	2	50	1000	−3.66	16.55	8.65	−1.29	12.60	3.95
77	2 plz2	2	50	1000	−4.82	−0.93	8.53	−2.24	1.48	4.38
75	2 plz2	2	50	2000	−0.94	16.95	5.55	0.38	14.93	2.93
78	2 plz2	2	50	2000	−1.70	2.35	7.32	−0.22	4.42	4.27
53	2 plz2	2	70	700	−1.33	25.74	9.39	−1.18	16.36	4.76
67	2 plz2	2	70	700	−1.22	26.11	9.42	−0.96	16.65	4.79
54	2 plz2	2	70	1000	0.38	20.79	9.22	−0.03	15.99	5.34
68	2 plz2	2	70	1000	0.46	21.18	9.15	0.12	16.26	5.31
71	2 plz2	2	70	2000	1.35	14.47	5.28	0.75	13.65	2.93
72	2 plz2	2	70	2000	1.55	15.57	5.05	0.88	14.17	2.73
38	2 plz2	3	50	700	−1.14		8.51	−2.30		2.62
40	2 plz2	4	50	700	−2.73		10.86	−2.26		5.69
93	2 plz2	5	50	700	−0.68	26.21	9.61	0.07	14.01	6.41
95	2 plz2	5	50	700	−2.10	1.87	10.51	−0.67	2.36	7.08
94	2 plz2	5	50	1000	0.62	25.23	9.27	0.95	16.72	6.21
96	2 plz2	5	50	1000	0.21	3.78	10.15	0.63	4.64	7.02
118	2 plz2	6	50	200	−5.09	40.41	3.78	−1.38	22.47	1.06
116	3 plz2+	1	50	200	−6.68	29.52	7.59	−4.04	24.46	4.23
109	3 plz2+	1	100	300	−4.26	28.86	8.77	−1.89	25.32	5.62
110	3 plz2+	1	100	500	−3.62	21.46	9.00	−2.20	20.99	5.21
31	3 plz2+	1	100	700	−3.69	19.68	10.97	−2.25	18.09	6.60
49	3 plz2+	1	100	700	−3.56	19.89	11.12	−2.35	18.57	6.79
63	3 plz2+	1	100	700	−3.76	19.91	11.31	−2.31	18.20	6.89
32	3 plz2+	1	100	1000	1.04	2.15	8.65	0.93	4.96	5.85
41	3 plz2+	2	100	700	−1.24	23.24	10.95	−1.44	20.69	8.52

Notes: The Product Setup No. of column 3 is given in table 4.4. The advertising variable is explained in table 4.14.

Table 4.16. (continued...)

Setup	Market Setup		Minimum Sales		Min 1 product sold in market			Min 3 products in market		
No.	Location	Product	Market	Product	price	retail	advertising (ad5)	price	retail	advertising (ad5)
47	3 plz2+	2	100	700	−3.34	15.10	10.05	−4.29	11.82	7.12
55	3 plz2+	2	100	700	−1.32	23.39	11.04	−1.40	20.71	8.61
61	3 plz2+	2	100	700	−3.26	17.46	10.62	−4.55	13.49	7.95
44	3 plz2+	2	100	1000	0.37	18.98	9.53	0.72	18.43	8.47
48	3 plz2+	2	100	1000	−2.48	11.36	8.92	−2.05	12.72	7.50
58	3 plz2+	2	100	1000	0.03	18.89	9.54	0.57	18.24	8.28
62	3 plz2+	2	100	1000	−2.54	13.04	9.52	−1.71	14.29	8.29
64	3 plz2+	2	100	1000	−4.62	25.60	11.12	−4.53	16.29	7.57
69	3 plz2+	2	100	2000	0.64	12.83	8.16	0.86	13.28	7.48
70	3 plz2+	2	100	2000	0.55	13.26	8.44	0.06	13.52	7.29
42	3 plz2+	3	100	700	−1.05		12.29	−1.34		9.21
56	3 plz2+	3	100	700	−1.14		12.42	−1.31		9.32
45	3 plz2+	3	100	1000	0.24		10.17	0.63		9.00
59	3 plz2+	3	100	1000	−0.07		10.23	0.51		8.88
25	3 plz2+	4	50	200	−3.66		13.34	−4.80		10.49
43	3 plz2+	4	100	700	−1.44		11.65	−2.47		10.20
57	3 plz2+	4	100	700	−1.28		11.83	−2.29		10.67
46	3 plz2+	4	100	1000	−0.64		11.55	−1.17		10.29
60	3 plz2+	4	100	1000	−0.95		11.75	−1.37		10.55
26	3 plz2+	5	50	200	−1.69	27.10	9.49	−0.11	23.49	10.23
89	3 plz2+	5	100	700	−2.93	17.86	10.07	−3.79	12.18	9.53
91	3 plz2+	5	100	700	−3.25	−0.80	11.41	−3.84	−0.90	10.21
90	3 plz2+	5	100	1000	−1.38	19.90	10.09	−1.52	17.33	10.20
92	3 plz2+	5	100	1000	−1.79	1.10	11.18	−1.94	2.35	10.88
119	3 plz2+	6	50	200	−3.88	32.81	6.44	−2.37	27.76	4.78
111	3 plz2+	6	100	300	−4.97	32.41	7.86	−3.83	28.59	6.16
112	3 plz2+	6	100	500	−4.66	28.02	7.53	−4.39	25.44	5.90
50	3 plz2+	6	100	1000	−3.51	22.03	9.40	−3.90	19.48	6.87
113	3 plz2+	7	100	300	−2.36	32.53	7.83	−2.82	26.94	6.28

Notes: The Product Setup No. of column 3 is given in table 4.4. The advertising variable is explained in table 4.14.

Table 4.16. (continued...)

Setup	Market Setup		Minimum Sales		Min 1 product sold in market			Min 3 products in market		
No.	Location	Product	Market	Product	price	retail	advertising (ad5)	price	retail	advertising (ad5)
114	3 plz2+	7	100	500	−2.54	28.05	8.35	−4.03	24.25	6.07
27	4 kreis2	1	100	400	−3.33	9.54	−0.33	−3.53	8.64	−3.31
33	4 kreis2	2	100	700	0.13	10.26	0.04	−0.15	9.85	−3.74
34	4 kreis2	2	100	1000	0.39	8.54	1.91	−0.73	8.53	−1.15
37	4 kreis2	3	100	1000	0.40		5.31	−0.82		1.81
39	4 kreis2	4	100	1000	−0.74		6.58	0.13		3.46
28	4 kreis2	6	100	400	−4.63	11.15	−1.98	−4.95	10.19	−3.99
88	4 kreis2	7	100	400	−3.19	10.52	−0.77	−3.73	10.08	−2.89

Notes: The Product Setup No. of column 3 is given in table 4.4. The advertising variable is explained in table 4.14.

Table 4.17. Frequency of Advertising Significance for all 120 Setups

	Advertising 56 days time window				
	Pressure				Total
	as is ad11	>0.2 ad12	>0.4 ad13	>0.6 ad14	as is ad15
α	All Shares				
0.050	71	48	22	34	92
0.010	57	27	5	21	86
0.001	43	19	3	13	83
0.05 & Coef. >0	69	41	6	4	86
0.05 & Coef. <0	2	7	14	17	6
α	Small Shares dropped				
0.050	12	13	32	71	88
0.010	7	2	23	54	73
0.001	2	2	12	37	59
0.05 & Coef. >0	11	4	1	17	74
0.05 & Coef. <0	1	9	29	39	14

Note: Numbers indicate how many out of 120 setups show statistical significance for a given significance level. The last two rows count the sign of the coefficient if it was significant at $\alpha = 0.05$. The model estimated is model (6) of table 4.5.

Table 4.18. Frequency of Advertising Significance for all 120 Setups

	Advertising 140 days time window				
	Pressure				Total
	as is ad6	>0.2 ad7	>0.4 ad8	>0.6 ad9	as is ad10
α	All Shares				
0.050	77	71	11	48	94
0.010	62	59	10	45	86
0.001	49	37	9	43	82
0.05 & Coef. >0	73	54	0	5	86
0.05 & Coef. <0	4	17	2	2	8
α	Small Shares dropped				
0.050	19	14	21	74	88
0.010	8	8	11	67	73
0.001	3	5	10	57	61
0.05 & Coef. >0	13	8	7	15	74
0.05 & Coef. <0	6	6	5	14	14

Note: Numbers indicate how many out of 120 setups show statistical significance for a given significance level. The last two rows count the sign of the coefficient if it was significant at $\alpha = 0.05$. The model estimated is model (6) of table 4.5.

Random Coefficient Model Results without Demographics

Figure 4.3. Overview of 30 Runs for Setup 49 without Demographics with logarithmized Advertising Variable

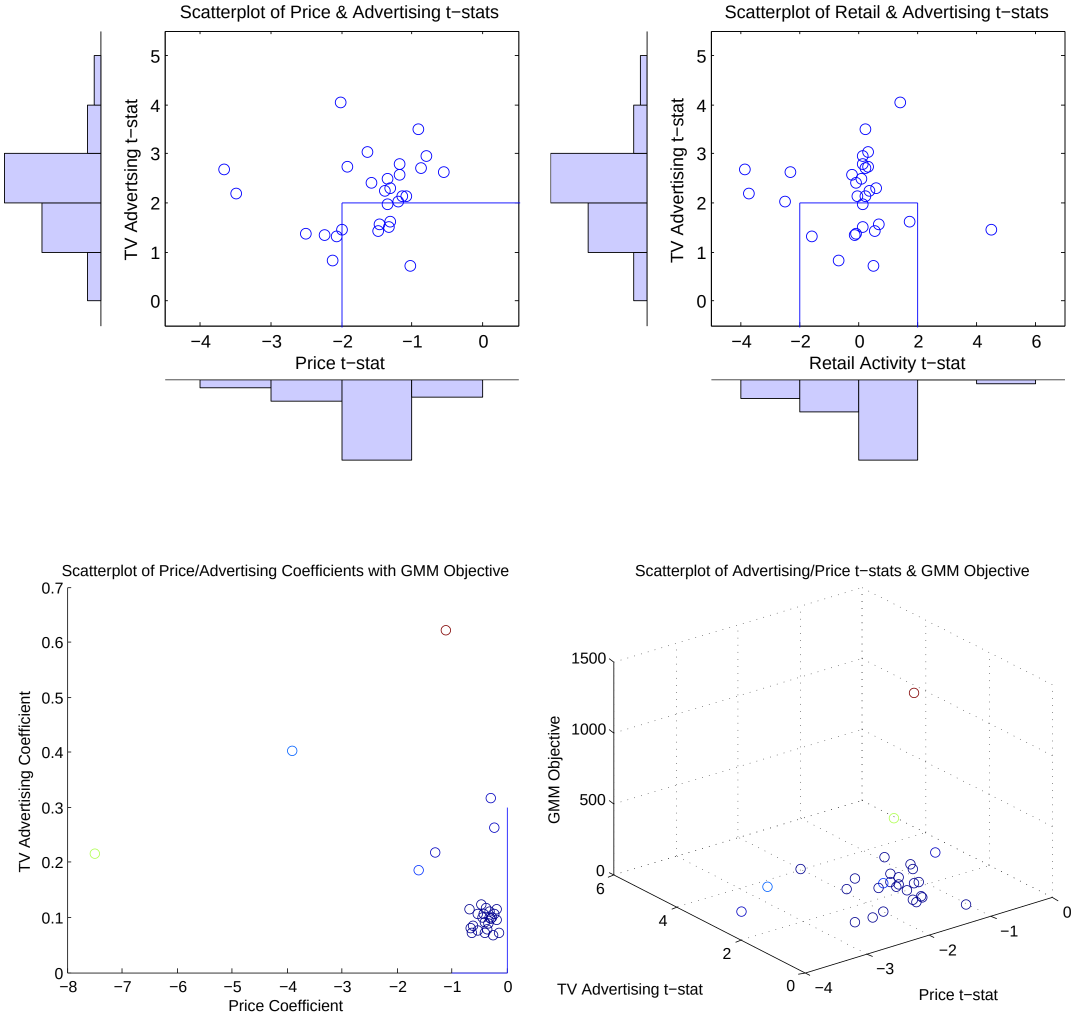


Figure 4.4. Overview of 40 Runs for Setup 47 without Demographics with logarithmized Advertising Variable

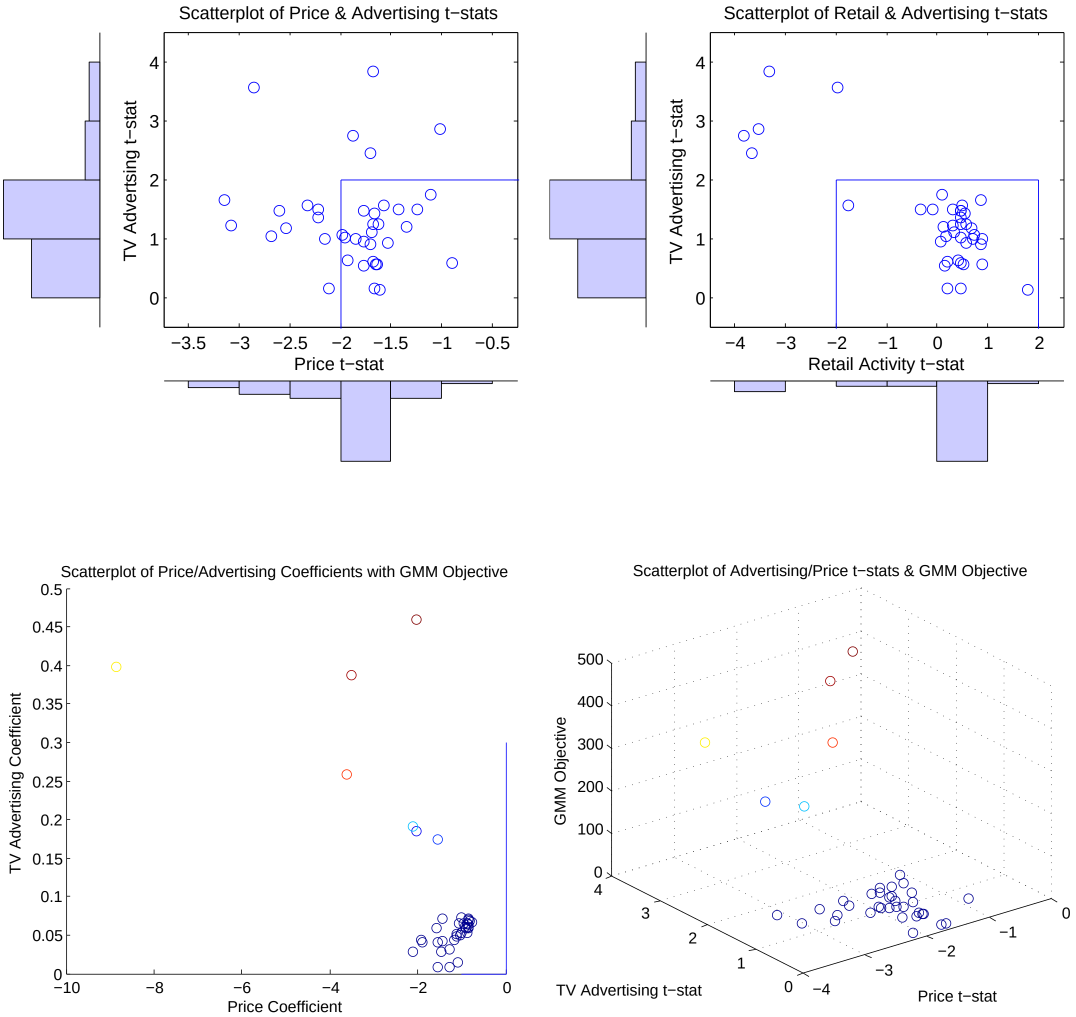
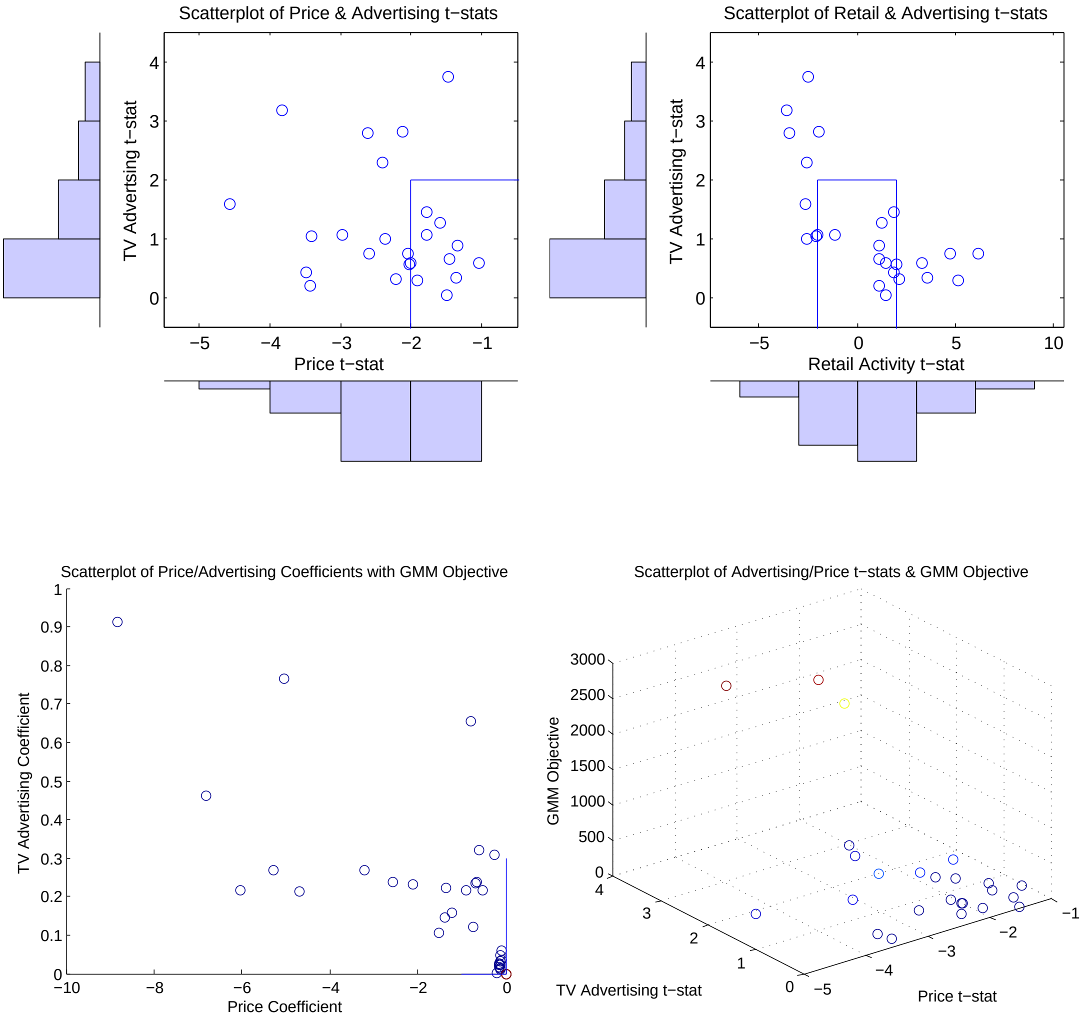


Figure 4.5. Overview of 40 Runs for Setup 111 without Demographics with logarithmized Advertising Variable



Random Coefficient Model Results with Demographics

Table 4.19. Random Coefficients Model with Setup 49 with Demographics and logarithmized Advertising Totals (ad20)

			interaction with demographics	
Variables	mean	sigma	income	household size
constant	-6.143***	0.024	-4.739	
	(1.044)	(2.665)	(7.459)	
price	-0.475	0.033	0.580	
	(0.410)	(0.108)	(0.477)	
retail	0.118	0.013	0.073	
	(0.079)	(0.627)	(0.738)	
ad20	0.066	3.047	4.047	
	(0.133)	(4.041)	(8.667)	
inh3	0.532***	0.292		0.590
	(0.108)	(0.345)		(0.640)
liquid	-2.594***	4.143		
	(1.208)	(7.305)		
konzentrat	-8.982***	0.022		3.847
	(0.908)	(7.408)		(7.375)
color	-1.314***	1.090		
	(0.152)	(5.280)		
GMM Objective			0.71753	
Runs, exit if			15	
delta X < TolX			6	
delta f < TolFun			8	
Iter > MaxIter			1	
Runtime (min)			2,075	

Note: Asterisks indicate significance levels at α: $* = 0.05, ** = 0.01, * * * = 0.001$. 20 Hausman instruments used.

Table 4.20. Random Coefficients Model with Setup 47 with Demographics and logarithmized Advertising Totals (ad20)

			interaction with demographics	
Variables	mean	sigma	income	household size
constant	-4.463***	0.413	-0.492	
	(0.489)	(2.907)	(5.778)	
price	-2.298	0.956	-1.891	
	(2.020)	(0.972)	(1.358)	
retail	0.018	0.160	0.455	
	(0.220)	(0.387)	(0.477)	
ad20	0.164	1.684	19.466	
	(0.186)	(9.978)	(18.697)	
inh3	–	–		
	–	–		
liquid	-7.688***	3.571		
	(0.618)	(3.716)		
konzentrat	-15.983***	4.734		0.293
	(1.855)	(3.966)		(5.454)
color	20.706***	5.968		
	(1.944)	(5.866)		
GMM Objective			2.4081	
Runs, exit if			30	
delta X < TolX			8	
delta f < TolFun			21	
Iter > MaxIter			1	
Runtime (min)			6,366	

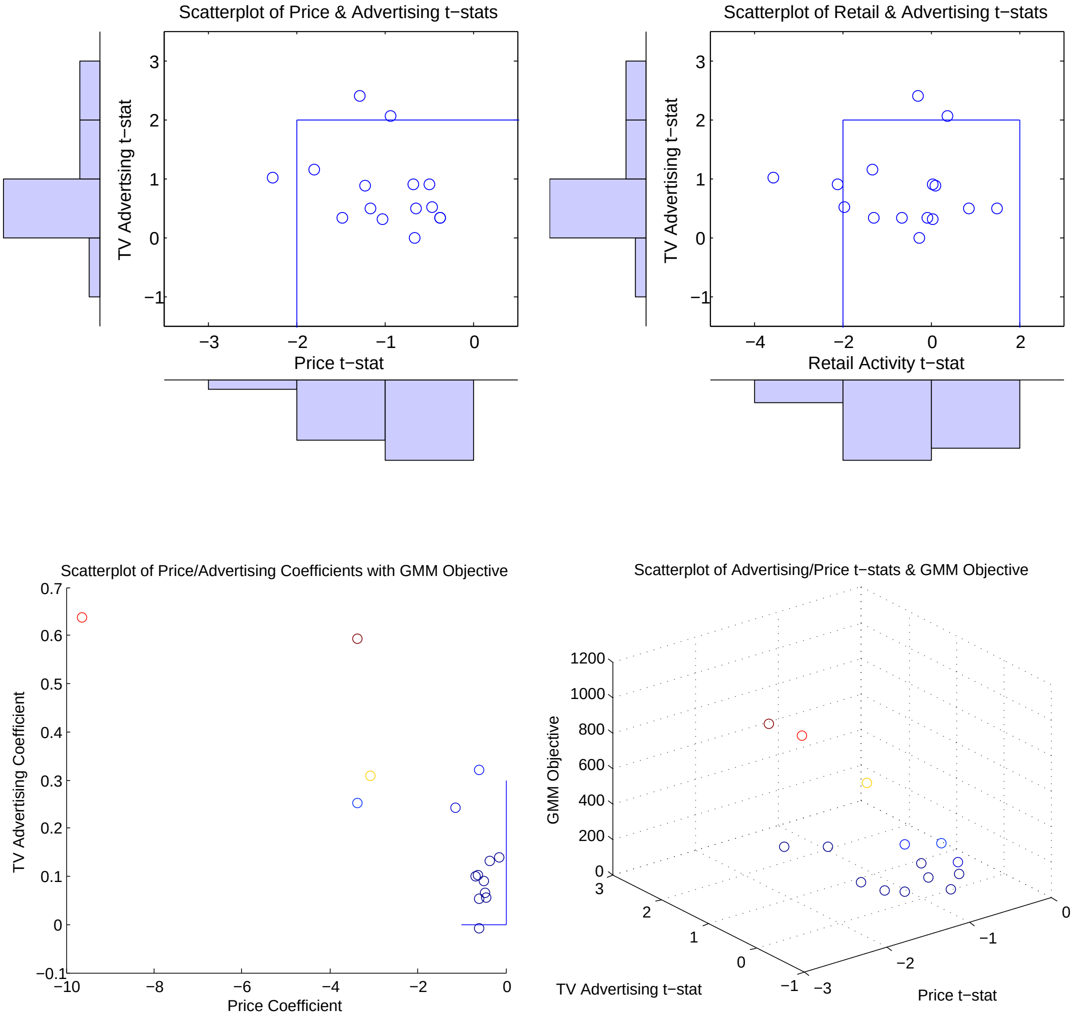
Note: Asterisks indicate significance levels at α: $* = 0.05, ** = 0.01, *** =$ 0.001. 20 Hausman instruments used.

Table 4.21. Random Coefficients Model with Setup 111 with Demographics and logarithmized Advertising Totals (ad20)

			interaction with demographics	
Variables	mean	sigma	income	household size
constant	-2.641***	0.751	0.461	
	(0.291)	(3.849)	(10.720)	
price	-0.186	0.076	-0.034	
	(0.170)	(0.353)	(0.887)	
retail	0.079	0.030	0.005	
	(0.200)	(0.638)	(0.472)	
ad20	0.021	4.873	-5.131	
	(0.200)	(5.458)	(28.007)	
inh3	0.091***	0.997		-0.152
	(0.014)	(0.965)		(0.643)
liquid	-0.150***	1.837		
	(0.025)	(4.206)		
konzentrat	0.574***	2.501		-6.318
	(0.032)	(7.265)		(11.295)
color	–	–		
	–	–		
GMM Objective			4.8156	
Runs, exit if			30	
delta X < TolX			25	
delta f < TolFun			5	
Iter > MaxIter			0	
Runtime (min)			5,945	

Note: Asterisks indicate significance levels at α: $* = 0.05, ** = 0.01, *** = 0.001$. 20 Hausman instruments used.

Figure 4.6. Overview of 15 Runs for Setup 49 with Demographics with logarithmized Advertising Variable



Scatterplot of Price & Advertising t−stats

Scatterplot of Retail & Advertising t−stats

Scatterplot of Price/Advertising Coefficients with GMM Objective

Scatterplot of Advertising/Price t−stats & GMM Objective

Figure 4.7. Overview of 30 Runs for Setup 47 with Demographics with logarithmized Advertising Variable

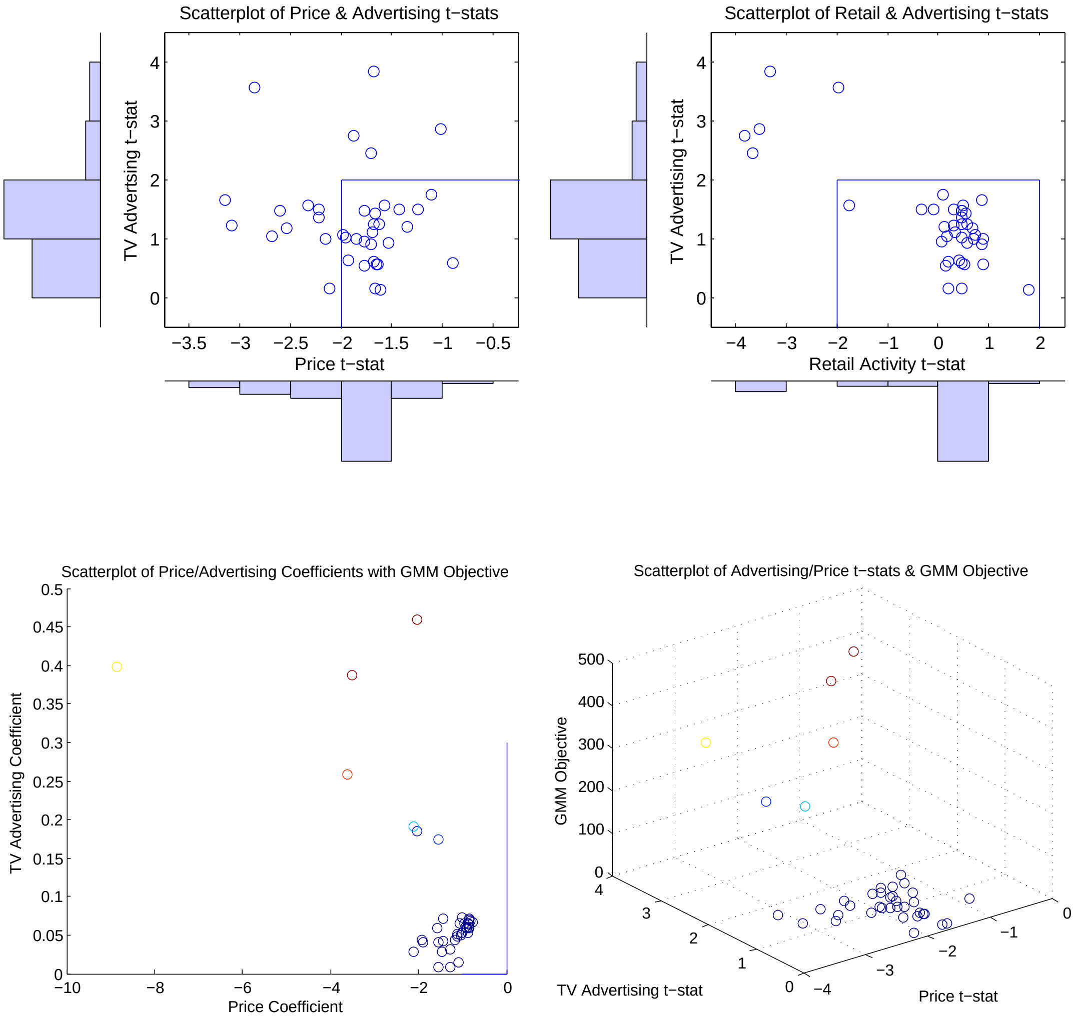
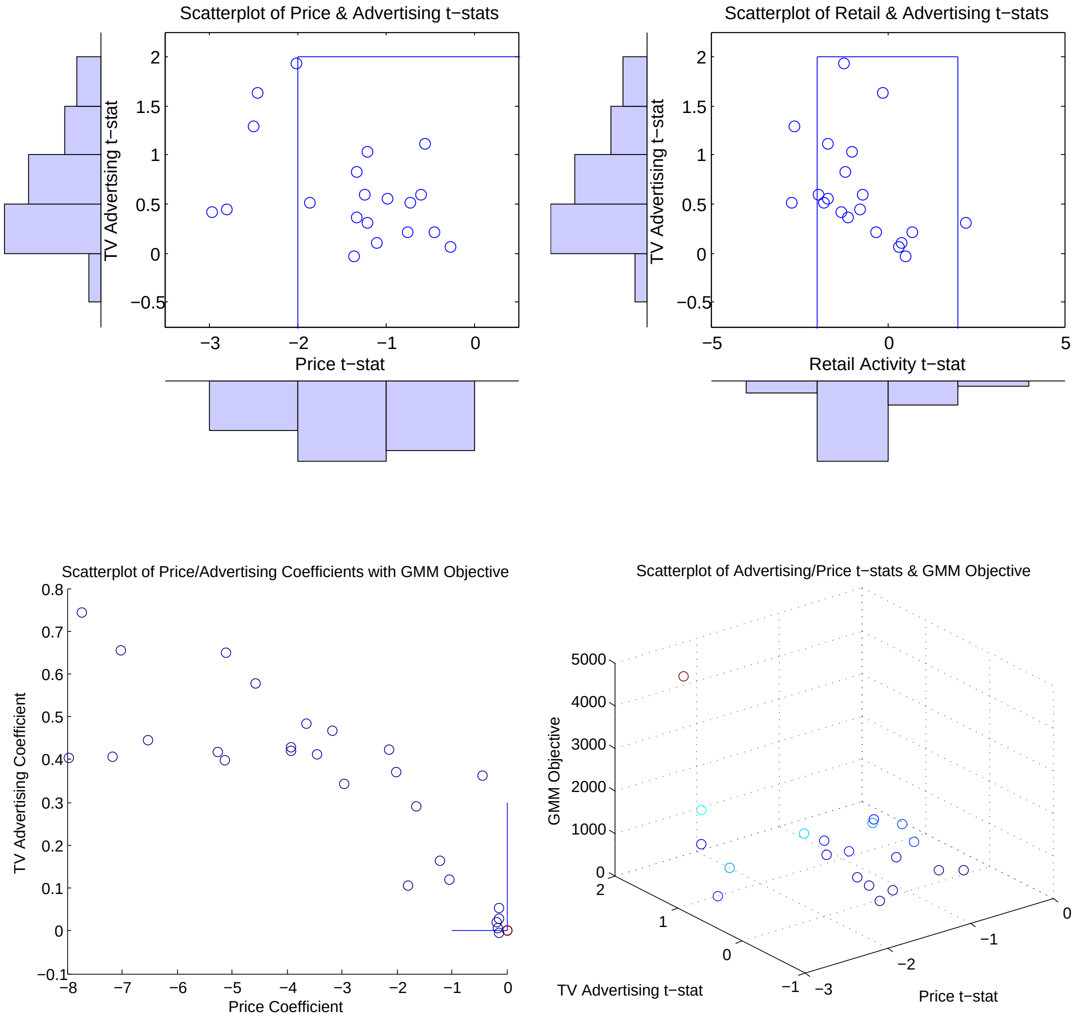


Figure 4.8. Overview of 30 Runs for Setup 111 with Demographics with logarithmized Advertising Variable



Bibliography

Ackerberg, D., L. Benkard, S. Berry, and A. Pakes (2006): "Econometric Tools for Analyzing Market Outcomes," *Handbook of Econometrics*, 6, forthcoming.

Ackerberg, D. A. (2001): "Empirically Distinguishing Informative and Prestige Effects of Advertising," *The RAND Journal of Economics*, 32(2), 316–333.

——— (2003): "Advertising, Learning, and Consumer Choice in Experience Good Markets: An Empirical Examination," *International Economic Review*, 44(3), 1007–1040.

Allenby, G. M., and P. E. Rossi (1998): "Marketing models of consumer heterogeneity," *Journal of Econometrics*, 89(1-2), 57–78.

Anand, B., and R. Shachar (2010): "Advertising, the Matchmaker," *RAND Journal of Economics*, forthcoming.

Anderson, S., A. de Palma, and J.-F. Thisse (1992): *Discrete Choice Theory of Product Differentiation.* MIT Press.

Bagwell, K. (2005): "The Economic Analysis of Advertising," forthcoming in Mark Armstrong and Rob Porter (eds.), Handbook of Industrial Organization, Vol. 3, North-Holland: Amsterdam.

Bajari, P., and L. Benkard (2005): "Demand Estimation With Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach," *Journal of Political Economy*, 113, 1239–1276.

Barroso, A. (2009): "Advertising and Consumer Awareness of a New Product," *Working Paper.*

Berry, S. (1994): "Estimating discrete-choice models of product differentiation," *Rand Journal of Economics*, 25, 242–262.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–889.

Berry, S., and A. Pakes (2007): "The Pure Characteristics Demand Model," *International Economic Review*, 48(4), 841–889.

Bijwaard, G. E., P. H. Franses, and R. Paap (2003): "Modeling Purchases as Repeated Events," *Econometric Institute Report EI 2003-45*.

Blundell, R. (1988): "Consumer Behavior: Theory and Empirival Evidence - A Survey," *Economic Journal*, 98, 16–65.

Briesch, R. A., P. K. Chintagunta, and R. L. Matzkin (2002): "Semiparametric Estimation of Brand Choice Behavior," *Journal of the American Statistical Association*, 97(460), 973–982.

Cherchye, L., B. D. Rock, and F. Vermeulen (2007): "The Collective Model of Household Consumption: A Nonparametric Characterization," *Econometrica*, 75(2), 553–574.

Ching, A., T. Erdem, and M. Keane (2009): "The Price Consideration Model of Brand Choice," *Journal of Applied Econometrics*, 24(3), 393–420.

Chintagunta, P., J. P. Dubé, and K. Y. Goh (2005): "Beyond the Endogeneity Bias: The Effect of Unmeasured Brand Characteristics on Household-Level Brand Choice Models," *Management Science*, 51(5), 832–849.

Dubé, J. P., J. T. Fox, and C.-L. Su (2009): "Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients Demand Estimation," *NBER Working Paper*, 14991.

Dubé, J. P., G. J. Hitsch, and P. E. Rossi (2009): "State Dependence and Alternative Explanations for Consumer Inertia," *NBER Working Paper*, 14912.

Einav, L., E. Leibtag, and A. Nevo (2008): "Not-So-Classical Measurement Errors: A Validation Study of Homescan," *NBER Working Paper*, 14436.

Erdem, T., and M. Keane (1996): "Decision Making under Uncertainty: Capturing Dynamic Brand Choice Porcesses in Turbulent Consumer Goods Markets," *Marketing Science*, 15(1), 1–20.

Goeree, M. S. (2008): "Limited Information and Advertising in the U.S. Personal Computer Industry," *Econometrica*, 76(5), 1017–1074.

Gowrisankaran, G., and M. Rysman (2007): "Dynamics of Consumer Demand for New Durable Goods," Working paper, Boston University.

Griffith, R., E. Leibtag, A. Leicester, and A. Nevo (2009): "Consumer Shopping Behavior: How Much Do Consumers Save?," *Journal of Economic Perspectives*, 23(2), 99–120.

GUADAGNI, P., AND J. LITTLE (1983): "A logit model of brand choice calibrated on scanner data," *Marketing Science*, 2, 203–238.

HAUSMAN, J. (1996): "Valuation of New Goods Under Perfect and Imperfect Competition," in *The Economics of New Goods, Studies in Income and Wealth Vol. 58*, ed. by T. Bresnahan, and R. Gordon. Chicago: National Bureau of Economic Research.

HECKMAN, J., R. MATZKIN, AND L. NESHEIM (2005): "Nonparametric Estimation of Nonadditive Hedonic Models," Working paper.

HENDEL, I., AND A. NEVO (2006): "Measuring the Implications of Sales and Consumer Inventory Behavior," *Econometrica*, 74(6), 1637–1673.

HODERLEIN, S. (2007): "How many Consumers are Rational?," Working paper.

HORSKY, D., S. MISRA, AND P. NELSON (2006): "Observed and Unobserved Preference Heterogeneity in Brand-Choice Models," *Marketing Science*, 25(4), 322–335.

KAUL, A., AND D. R. WITTINK (1995): "Empirical Generalizations about the Impact of Advertising on Price Sensitivity and Price," *Marketing Science*, 14(3), 151–160.

KEANE, M. (1997): "Modeling heterogeneity and state dependence in consumer choice behvior," *Journal of Business and Economic Statistics*, 15(3), 310–327.

KNITTEL, C. R., AND K. METAXOGLOU (2008): "Estimation of Random Coefficient Demand Models: Challenges, Difficulties and Warnings," *NBER Working Paper*, 14080.

LANCASTER, K. (1966): "A New Approach to Consumer Theory," *Journal of Political Economy*, 74, 132–157.

LAUGA, D. O. (2008): "Persuasive Advertising with Sophisticated but Impressionable Consumers," Working paper.

MATZKIN, R. L. (2004): "Unobservable Instruments," *Working Paper*.

MCFADDEN, D. (1974): "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, chap. 4, pp. 105–142. New York: Academic Press.

MCFADDEN, D., AND K. TRAIN (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15(5), 447–470.

MELNIKOV, O. (2001): "Demand for differentiated products: The case of the U.S. computer printer market," *Unpublished manuscript. Cornell University*.

NEVO, A. (2000): "A practioner's guide to estimation of random-coefficients logit models of demand," *Journal of Economics and Management Strategy*, 9(4), 513–548.

——— (2001): "Measuring market power in the ready-to-eat cereal industry," *Econometrica*, 69(2), 307–342.

PAKES, A. (1994): "Dynamic structural models, problems and prospects: mixed continuous discrete controls and market interaction," in *Advances in Econometrics, Sixth World Congress*, ed. by C. Sims, pp. 171–259. New York. Cambridge.

PETRIN, A. (2002): "Quantifying the Benefits of New Products: The Case of the Minivan," *The Journal of Political Economy*, 110(4), 705729.

PETRIN, A., AND K. TRAIN (2006): "Control Function Corrections for Unobserved Factors in Differentiated Product Models," mimeo.

RIVERS, D., AND Q. H. VUONG (1988): "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, 39(3), 347 – 366.

SHAKED, A., AND J. SUTTON (1983): "Natural Oligopolies," *Econometrica*, 51(5), 1469–1483.

SHUM, M. (2004): "Does Advertising Overcome Band Loyalty? Evidence From the Breakfast Cereals Market," *Journal of Economics and Management Strategy*, 13(2), 241–272.

TRAIN, K. (2003): *Discrete Choice Methods with Simulation.* Cambridge University Press.

TRAJTENBERG, M. (1990): *Economic Analysis of Product Innovation: The Case of CT Scanners.* Harvard University Press.

VILLAS-BOAS, J. M., AND R. S. WINER (1999): "Endogeneity in Brand Choice Models," *Management Science*, 45(10), 1324–1338.

WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data.* MIT Press.

Lebenslauf

Persönliche Daten

Christoph Nagel

geboren 1977

Ausbildung

10/2004 - heute Promotion in Volkswirtschaftslehre am Center for Doctoral Studies in Economics (CDSE), Universität Mannheim.

03/2006 - 08/2006 Forschungsaufenthalt am University College London (UCL), Institute for Fiscal Studies (IFS), Center for Microdata Methods and Practice (cemmap), UK.

10/1998 - 08/2003 Diplom-Volkswirt, Rheinische Friedrich-Wilhelms Universität Bonn.

10/2001 - 07/2002 Gaststudium im Economics Ph.D Programm, University of California, Berkeley, USA.

06/1997 Abitur, Gymnasium an der Hermann-Böse-Straße, Bremen.

Mannheim, 28. Juli 2010

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die Dissertation selbstständig angefertigt und mich anderer als der in ihr angegebenen Hilfsmittel nicht bedient habe, insbesondere, dass aus anderen Schriften Entlehnungen, soweit sie in der Dissertation nicht ausdrücklich als solche gekennzeichnet und mit Quellenangaben versehen sind, nicht stattgefunden haben.

Mannheim, 28. Juli 2010