Discussion Paper No. 101

A Description of a Household Panel Dataset with Consumption and Advertising

Christoph Nagel*

July 2010

*Christoph Nagel, Universität Mannheim, Economics Department, 68131 Mannheim, cnagel@gmx.de

Financial support from the Deutsche Forschungsgemeinschaft through Graduiertenkolleg Risk and Liquidity in Financial, Commodity, and Factor Markets is gratefully acknowledged.
A Description of a Household Panel Dataset with Consumption and Advertising

A representative German Panel for Detergents and Chocolate Purchases

Christoph Nagel *

First version: March 2010
This Version: 27 Juli 2010

Abstract

In this paper I give an overview of a German dataset collected by the marketing research company A.C. Nielsen. The paper is intended for potential users of the dataset and contains numerous details. Therefore, the focus is on describing and analyzing the data. 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 comprises 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 usefulness of the dataset.

Keywords: dataset, consumer goods, advertising

JEL Classifications: C89, M37, L69

*I thank very much Kai Kopperschmidt, Martin Schniedermeier and Marc Rossbach of A.C. Nielsen for providing the data. Universität Mannheim, Economics Department, L7, 3-5, 68131 Mannheim, Germany; contact: cnagel@gmx.de
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 paper is intended for potential users of the dataset and contains numerous details to permit profound research usage of it.

In this work 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, 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 3 describes each data file in detail, accompanied by descriptive statistics. I restrict myself to simple statistics and interesting graphics. In section 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 5 concludes.

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.

<table>
<thead>
<tr>
<th>File</th>
<th>Household appears if . . .</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cash</td>
<td>purchased anything</td>
<td>total value of purchases with time, store, zip</td>
</tr>
<tr>
<td>Wash</td>
<td>purchased detergent</td>
<td>detergent purchases with time, store, zip and product details (price, quantity, characteristics)</td>
</tr>
<tr>
<td>Demo</td>
<td>sampled</td>
<td>time constant sociodemographic Variables</td>
</tr>
<tr>
<td>Contact</td>
<td>TV telemeter equipped</td>
<td>TV advertisement, TV representation factors</td>
</tr>
</tbody>
</table>

Table 1. Overview of Consumer Data Files

See table 1 for details on the four data files that contain purchase data, category purchase data, sociodemographics and advertising contacts. Table 2

\(^1\)”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.
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. Number of Households by Required Information

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Criterion</th>
<th>No. of Households</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demo</td>
<td>sociodemographics known</td>
<td>17,978</td>
</tr>
<tr>
<td>Cash</td>
<td>any purchase</td>
<td>16,757</td>
</tr>
<tr>
<td>Cash</td>
<td>any purchase in “detergent” store</td>
<td>16,737</td>
</tr>
<tr>
<td>Cash</td>
<td>above plus demographics</td>
<td>16,737</td>
</tr>
<tr>
<td>Wash</td>
<td>any purchase of detergent</td>
<td>13,455</td>
</tr>
<tr>
<td>Wash</td>
<td>above plus demographics</td>
<td>13,455</td>
</tr>
<tr>
<td>Wash</td>
<td>TV coverage in any year</td>
<td>3,783</td>
</tr>
<tr>
<td>Wash</td>
<td>TV coverage 2004</td>
<td>2,953</td>
</tr>
<tr>
<td>Wash</td>
<td>TV coverage 2005</td>
<td>2,630</td>
</tr>
<tr>
<td>Wash</td>
<td>TV coverage 2006</td>
<td>2,571</td>
</tr>
<tr>
<td>Wash</td>
<td>TV coverage 2004 and 2005</td>
<td>2,250</td>
</tr>
<tr>
<td>Wash</td>
<td>TV coverage 2005 and 2006</td>
<td>1,993</td>
</tr>
<tr>
<td>Wash</td>
<td>TV coverage 2004 to 2006</td>
<td>1,735</td>
</tr>
</tbody>
</table>

**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 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 distribution of household residence.⁴

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

⁴This 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 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.
Figure 1. Geographic Distribution of Households according to the Zip Code of Residence
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.

2.1 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

\(^5\)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.
all the purchases they have 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

⁶In 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.
are quite satisfied with the collection process.\textsuperscript{7}

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.\textsuperscript{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.

2.3 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 paper 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

\textsuperscript{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.

\textsuperscript{8}Address: Chair of Statistics, Economics Department, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany.
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 paper. 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 price data. I deal with this issue in a similar fashion as explained later in section 5. 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. The product category of interest is laundry detergents. In principle it has the same TV advertising data as the data employed in this paper, 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 form 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.

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.

---

9This is the freely available data of the Kilts Marketing Research Center at the Graduate School of Business at Chicago.
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.\textsuperscript{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 2.1. 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.

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 30\textsuperscript{th} 2004 through June 30\textsuperscript{th} 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 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.

3.1.1 Data Modifications

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

\textbf{Zip codes} (variable: \texttt{plz}) For several observations, zip codes are missing. To fill up missing values I assume that people will mostly do consumer good 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.

\textbf{Store names} (variable: \texttt{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 \texttt{store}) on other purchase trips within the same file. I replace missing \texttt{key_acc} store name values with the \texttt{key_acc} values of the identical store code if it was

\textsuperscript{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.
Table 3. Overview of Variables in General Purchase Data

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

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.

present in the file for another purchase trip.\textsuperscript{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 where 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.

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

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.\footnote{For variables durobsc, durhmean there exist versions durobs2c, durhmean2c. The c indicates the duration is calculated for the general purchase data. The zeros are “ignored” by setting the variable to missing.}

3.1.3 Summary Statistics

This section provides summary statistics for the general purchase data. Consult table 4 for descriptive statistics.\footnote{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 $> 0.7$.} 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 (durobsc, durobs2c). Therefore, the difference between the duration measures is quite
Table 4. Summary Statistics for General Purchase Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>PI Mean</th>
<th>Median</th>
<th>SD</th>
<th>IQR</th>
<th>Min</th>
<th>10%</th>
<th>90%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>18.91</td>
<td>12.67</td>
<td>19.60</td>
<td>18.77</td>
<td>0.10</td>
<td>2.99</td>
<td>42.24</td>
<td>204.51</td>
</tr>
<tr>
<td>durobsc</td>
<td>2.47</td>
<td>1.00</td>
<td>3.97</td>
<td>3.00</td>
<td>0.00</td>
<td>0.00</td>
<td>6.00</td>
<td>549.00</td>
</tr>
<tr>
<td>durobs2c</td>
<td>3.56</td>
<td>2.00</td>
<td>4.35</td>
<td>3.00</td>
<td>1.00</td>
<td>1.00</td>
<td>7.00</td>
<td>549.00</td>
</tr>
<tr>
<td>hhobsb</td>
<td>HH</td>
<td>182.77</td>
<td>133.00</td>
<td>164.26</td>
<td>188.00</td>
<td>1.00</td>
<td>29.00</td>
<td>403.00</td>
</tr>
<tr>
<td>dursamplec</td>
<td>HH</td>
<td>601.28</td>
<td>725.00</td>
<td>209.00</td>
<td>314.00</td>
<td>0.00</td>
<td>242.00</td>
<td>730.00</td>
</tr>
<tr>
<td>durhhmean</td>
<td>HH</td>
<td>2.47</td>
<td>2.05</td>
<td>1.59</td>
<td>1.55</td>
<td>0.00</td>
<td>1.09</td>
<td>4.39</td>
</tr>
<tr>
<td>durhhmean2c</td>
<td>HH</td>
<td>3.47</td>
<td>2.99</td>
<td>1.83</td>
<td>1.90</td>
<td>1.00</td>
<td>1.80</td>
<td>5.77</td>
</tr>
<tr>
<td>numstore</td>
<td>HH</td>
<td>12.06</td>
<td>12.00</td>
<td>5.46</td>
<td>8.00</td>
<td>1.00</td>
<td>5.00</td>
<td>19.00</td>
</tr>
<tr>
<td>storehhi*</td>
<td>HH</td>
<td>0.34</td>
<td>0.29</td>
<td>0.19</td>
<td>0.21</td>
<td>0.06</td>
<td>0.16</td>
<td>0.61</td>
</tr>
</tbody>
</table>

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.

...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.

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 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).

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 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>
<thead>
<tr>
<th>Variable</th>
<th>Numeric</th>
<th>Position</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bigpack</td>
<td>1(=yes)</td>
<td>41</td>
<td>Dummy for detergent is sold in extra big pack</td>
</tr>
<tr>
<td>calweek</td>
<td>1</td>
<td>27</td>
<td>Calendar Week (STATA) YYYYwWW</td>
</tr>
<tr>
<td>cnanlabl</td>
<td>0(=no)</td>
<td>2</td>
<td>Detailed Product Information String</td>
</tr>
<tr>
<td>color</td>
<td>1</td>
<td>36</td>
<td>Dummy for detergent is color</td>
</tr>
<tr>
<td>date</td>
<td>1</td>
<td>1</td>
<td>Date as YYYYddd, where ddd is absolute day in year YYYY</td>
</tr>
<tr>
<td>date2</td>
<td>1</td>
<td>21</td>
<td>Date as YYYYMMDD</td>
</tr>
<tr>
<td>datebegw</td>
<td>1</td>
<td>33</td>
<td>Date of first observation in this file</td>
</tr>
<tr>
<td>dateday</td>
<td>1</td>
<td>25</td>
<td>Date DD</td>
</tr>
<tr>
<td>dateendw</td>
<td>1</td>
<td>32</td>
<td>Date of last observation in this file</td>
</tr>
<tr>
<td>datemonth</td>
<td>1</td>
<td>24</td>
<td>Date MM</td>
</tr>
<tr>
<td>dateyear</td>
<td>1</td>
<td>23</td>
<td>Date YYYY</td>
</tr>
<tr>
<td>display</td>
<td>1</td>
<td>18</td>
<td>Product is on display</td>
</tr>
<tr>
<td>duft</td>
<td>1</td>
<td>44</td>
<td>Numerical code for scent type (coded)</td>
</tr>
<tr>
<td>durhhmean2w</td>
<td>1</td>
<td>31</td>
<td>Mean duration per household in days since last purchase, exclude zero durations</td>
</tr>
<tr>
<td>durhhmeanw</td>
<td>1</td>
<td>30</td>
<td>Mean duration per household in days since last purchase</td>
</tr>
<tr>
<td>durobs2w</td>
<td>1</td>
<td>29</td>
<td>Duration in days since last purchase, exclude durations of zero</td>
</tr>
<tr>
<td>durobsw</td>
<td>1</td>
<td>28</td>
<td>Duration in days since last purchase</td>
</tr>
<tr>
<td>dursamplew</td>
<td>1</td>
<td>34</td>
<td>Duration between first and last observation in this file</td>
</tr>
<tr>
<td>edate</td>
<td>1</td>
<td>26</td>
<td>Date (STATA) YYYYMMDD</td>
</tr>
<tr>
<td>erg</td>
<td>0</td>
<td>3</td>
<td>Effective amount of detergent</td>
</tr>
<tr>
<td>extra_size</td>
<td>1</td>
<td>40</td>
<td>Dummy for detergent is sold in extra big size</td>
</tr>
<tr>
<td>feature</td>
<td>1</td>
<td>16</td>
<td>Product is featured</td>
</tr>
<tr>
<td>gimmick</td>
<td>1</td>
<td>39</td>
<td>Dummy for detergent is sold with gimmick</td>
</tr>
<tr>
<td>handbill</td>
<td>1</td>
<td>17</td>
<td>Product is hand billed</td>
</tr>
<tr>
<td>her**</td>
<td>1</td>
<td>49</td>
<td>Numerical code for manufacturer (coded), SONST (=Various), HANDEL (=Private Label)</td>
</tr>
<tr>
<td>herb8**</td>
<td>1</td>
<td>51</td>
<td>8 biggest manufacturers (coded), SONST (=Various), HANDEL (=Private Label)</td>
</tr>
<tr>
<td>hhnr</td>
<td>1</td>
<td>20</td>
<td>Household ID code</td>
</tr>
<tr>
<td>hhobsw</td>
<td>1</td>
<td>22</td>
<td>No of purchases in file for HH</td>
</tr>
<tr>
<td>id_nr</td>
<td>1</td>
<td>15</td>
<td>unknown ID number</td>
</tr>
<tr>
<td>inh</td>
<td>1</td>
<td>42</td>
<td>Packet size (1 or kg)</td>
</tr>
<tr>
<td>key_acc</td>
<td>0</td>
<td>7</td>
<td>Name of store</td>
</tr>
<tr>
<td>keyb10</td>
<td>0</td>
<td>55</td>
<td>10 biggest key_acc, details see keyb7</td>
</tr>
</tbody>
</table>

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 5. (continued...)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Numeric</th>
<th>Position</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyb15</td>
<td>0</td>
<td>54</td>
<td>15 biggest key_accs, details see keyb7</td>
</tr>
<tr>
<td>keyb7</td>
<td>0</td>
<td>56</td>
<td>7 biggest key_accs, SONST (=Various)</td>
</tr>
<tr>
<td>kons'</td>
<td>1</td>
<td>45</td>
<td>Numerical code for consistency type (coded)</td>
</tr>
<tr>
<td>konzentrat</td>
<td>1</td>
<td>38</td>
<td>Dummy for detergent is concentrated</td>
</tr>
<tr>
<td>liquid</td>
<td>1</td>
<td>43</td>
<td>Dummy for product is liquid detergent</td>
</tr>
<tr>
<td>menge</td>
<td>1</td>
<td>5</td>
<td>Number of units purchased</td>
</tr>
<tr>
<td>mke**</td>
<td>1</td>
<td>50</td>
<td>Numerical code for brand (coded)</td>
</tr>
<tr>
<td>mkeb14**</td>
<td>1</td>
<td>52</td>
<td>14 biggest brands (coded), details see mkeb9</td>
</tr>
<tr>
<td>mkeb9**</td>
<td>1</td>
<td>53</td>
<td>9 biggest brands (coded), SONST (=Various), EIGENMARKE (=Private Label)</td>
</tr>
<tr>
<td>plz</td>
<td>1</td>
<td>9</td>
<td>German zip code, all five digits</td>
</tr>
<tr>
<td>plz1</td>
<td>1</td>
<td>10</td>
<td>German zip code, 1st digit</td>
</tr>
<tr>
<td>plz2</td>
<td>1</td>
<td>11</td>
<td>German zip code, 1st two digits</td>
</tr>
<tr>
<td>plz3</td>
<td>1</td>
<td>12</td>
<td>German zip code, 1st three digits</td>
</tr>
<tr>
<td>plz4</td>
<td>1</td>
<td>13</td>
<td>German zip code, 1st four digits</td>
</tr>
<tr>
<td>prceflag</td>
<td>1</td>
<td>19</td>
<td>Product is price flagged</td>
</tr>
<tr>
<td>preis</td>
<td>1</td>
<td>6</td>
<td>Purchase price</td>
</tr>
<tr>
<td>purchase</td>
<td>1</td>
<td>35</td>
<td>Dummy for wash purchase incident</td>
</tr>
<tr>
<td>quartal</td>
<td>1</td>
<td>14</td>
<td>Date as YYYYQQ, where QQ is quarter</td>
</tr>
<tr>
<td>sensitiv</td>
<td>1</td>
<td>37</td>
<td>Dummy for detergent is sensitiv</td>
</tr>
<tr>
<td>stadt</td>
<td>0</td>
<td>8</td>
<td>Name of city</td>
</tr>
<tr>
<td>store</td>
<td>1</td>
<td>4</td>
<td>Store ID code</td>
</tr>
<tr>
<td>uwg'</td>
<td>1</td>
<td>46</td>
<td>Numerical code for general purpose (coded)</td>
</tr>
<tr>
<td>vpa***</td>
<td>1</td>
<td>48</td>
<td>Numerical code for packaging type (coded)</td>
</tr>
<tr>
<td>zmke***</td>
<td>1</td>
<td>47</td>
<td>Numerical code for sub brand (coded)</td>
</tr>
</tbody>
</table>

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.

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. 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. 

**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 where dropped, in total 216, leaving us with 94,222 observations.

### 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.

---

14Recall: 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.

15Example: 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”.

16Names 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
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. 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.

---

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 %", ....
3.2.3 Summary Statistics

In this section I present summary statistics for the category purchase data. Tables 6 and 7 display simple statistics for all variables per household (HH) and per purchase incident (PI), respectively. I first discuss the table with the

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

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.

statistics per household. 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. 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 3 shows the density of the variable. The mass of households have less than 10 purchases in the sampling period.

Figure 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...
Figure 2. Purchases per Month per Household

Figure 3. Purchases in Sample per Household
in store choice (storehhi). Thus, consumers seem to stick to their favorite brands and stores.

Now I turn to the statistics per purchase incident in table 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.
Table 7. Summary Statistics for Category Purchase Data

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

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

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. 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.

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 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.

### 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 €. For the first interval with low income households I set the value to 750 €.

### 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
Table 8. Overview of Variables in Sociogemographic Data

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

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.
groups: \textit{bself} indicates self employed persons, \textit{bwhite\_l} and \textit{bwhite\_h} represent low and high position white collar persons, \textit{bblue} marks blue collar workers and \textit{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.

### 3.3.3 Summary Statistics

In this section I present summary statistics for the demographic data. Table 9 contains descriptive statistics for all sampled households. This includes also households that have never purchased a detergent.

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

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

Recall that income is classified data (\textit{eink}). The average and median income are below the German average income of 2,800 €.\textsuperscript{21} Figure 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 (\textit{urban}). The median household size is two, the average being slightly higher, and the number of children per household has

\textsuperscript{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 5. Net Monthly Household Income

Figure 6. Occupational Status of Households
a median of zero and a mean of 0.5 \( (\text{hhsize, kizahl}) \). 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 \( (\text{hhfage, hhfpsex}) \). 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 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 \( (\text{beginn, ende}) \). 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 \( (\text{membership}) \). If I consider only the time period for which purchase data are available the mean is 482 days and the median is 436 days \( (\text{membership2}) \). The latter fits well the data in figure 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 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 7.** Education of Households
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 10 for an overview.

Table 10. Overview of Variables in Advertisement Contact Data

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

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.

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.\textsuperscript{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

\textsuperscript{22}Total number of advertisement contacts for all households in a given year. 2004: 696,168 2005: 839,715 2006: 363,971.
the advertising campaign information for each year.\textsuperscript{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 have to be merged per year and are then appended. In the merged data, the campaign id is dropped.\textsuperscript{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 11 are dropped: `prodnr`, `motivnr`, `prodstr` and `motivstr`. Note that I extract information from `prodstr` and `motivstr` before dropping them, see the next paragraph.

### 3.4.2 Remarks on created Variables

**Spot characteristics** From the campaign information in table 11, I extract information on the advertisement from the variables `prodstr` and `motivstr` and created 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`).\textsuperscript{25}

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>kam</td>
<td>Campaign ID (different numbers if content or spot changed)</td>
</tr>
<tr>
<td>prodnr</td>
<td>An internal A.C. Nielsen code</td>
</tr>
<tr>
<td>prodstr</td>
<td>A text defining the product</td>
</tr>
<tr>
<td>motivnr</td>
<td>An internal A.C. Nielsen code</td>
</tr>
<tr>
<td>motivstr</td>
<td>A text defining the spots content</td>
</tr>
</tbody>
</table>

\textsuperscript{23}Total number of advertising campaigns for all products in a given year. 2004: 57 2005: 156 2006: 64.

\textsuperscript{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.

\textsuperscript{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.
3.4.3 Summary Statistics

This section presents the summary statistics for the advertising exposure data. Table 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 (\textit{laenge}). Concerning the inferred informational contents of the spots, 83% of the spots advertise common detergents (\textit{spotwash}) and the rest advertise wash additives. 53% of the spots advertise powder detergents (\textit{spotpow}), 25% do so for liquid detergents (\textit{spotliq}). 5% spots are identified to be spots that advertise solely the brand in general, serving presumably for image/prestige recognition (\textit{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.\footnote{Assuming 20 seconds per spot this household will have seen 424 minutes, roughly seven hours, only of detergent advertisements.}

The concentration measures for brand indicate a low concentration, concluding that consumers see advertisements of almost all advertised products (\textit{brandhhi}, \textit{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 9 two TV stations broadcast most of the advertisements, consistent with the concentration measure. The other

---

**Table 12. Summary Statistics for Advertising Data**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>IQR</th>
<th>Min</th>
<th>10%</th>
<th>90%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{laenge}</td>
<td>PI</td>
<td>21.03</td>
<td>20.00</td>
<td>9.19</td>
<td>15.00</td>
<td>5.00</td>
<td>10.00</td>
<td>32.00</td>
</tr>
<tr>
<td>\textit{spotwash}</td>
<td>PI</td>
<td>0.83</td>
<td>1.00</td>
<td>0.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>\textit{spotpow}</td>
<td>PI</td>
<td>0.53</td>
<td>1.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>\textit{spotliq}</td>
<td>PI</td>
<td>0.25</td>
<td>0.00</td>
<td>0.43</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>\textit{spotimg}</td>
<td>PI</td>
<td>0.05</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>\textit{hhobsa}</td>
<td>HH</td>
<td>502.21</td>
<td>268.00</td>
<td>647.81</td>
<td>533.00</td>
<td>1.00</td>
<td>35.00</td>
<td>1272.00</td>
</tr>
<tr>
<td>\textit{brandhhic}</td>
<td>HH</td>
<td>0.19</td>
<td>0.17</td>
<td>0.08</td>
<td>0.04</td>
<td>0.11</td>
<td>0.14</td>
<td>0.24</td>
</tr>
<tr>
<td>\textit{brandhhiz}</td>
<td>HH</td>
<td>0.19</td>
<td>0.17</td>
<td>0.08</td>
<td>0.04</td>
<td>0.11</td>
<td>0.14</td>
<td>0.24</td>
</tr>
<tr>
<td>\textit{senderhhic}</td>
<td>HH</td>
<td>0.35</td>
<td>0.31</td>
<td>0.16</td>
<td>0.19</td>
<td>0.14</td>
<td>0.19</td>
<td>0.56</td>
</tr>
<tr>
<td>\textit{senderhhiz}</td>
<td>HH</td>
<td>0.35</td>
<td>0.31</td>
<td>0.17</td>
<td>0.19</td>
<td>0.13</td>
<td>0.19</td>
<td>0.56</td>
</tr>
</tbody>
</table>

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).
TV stations are quite equally sized concerning the total number of achieved advertising contacts. Figure 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 9. Advertisement Contacts broken by TV Stations

Figure 10. Advertisements by Day time Windows
4 Combined Dataset

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 Nagel (2010).

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. 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.

27The corresponding variables are listed here and detailed in section 3.2: kons, mke, uwg, inh, duft, color, sensitiv, konzentrat, gimmick, extra_size, bigpack.

28Example: Suppose there exist two brands A and B, and the no-purchase alternative. Choice set for exclusive choice of one brand is \{∅, A, B\}, if the consumer can buy both
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.

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.

²⁹I 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.
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.\textsuperscript{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.

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.

\textsuperscript{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.
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.

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 Nagel (2010). 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 vari-
able 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*)
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 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

---

31 The * is a wildcard as in usual programming languages and represents an integer.
the time window of 140 to 126 days before the purchase. 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).

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 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 (DCLM) to analyze the effect

Note the difference: The variable countc140adr captures relative advertising pressure from 140 to zero days before the purchase incident.
of purchase relevant variables on the purchase decision of households. The results of the estimations are displayed in table 14, 15 and 16. The latter two tables are in appendix part A.

Table 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. 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.

The tables 14 to 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

---

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

<table>
<thead>
<tr>
<th>Variables</th>
<th>LPM</th>
<th>DC Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef./SE</td>
<td>Coef./SE</td>
</tr>
<tr>
<td>Incidents</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>preis</td>
<td>-.048***</td>
<td>-.062***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>inh</td>
<td>-.068***</td>
<td>-.051***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>feature</td>
<td>-.045***</td>
<td>-.065***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>display</td>
<td>.041***</td>
<td>.036***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>prooflag</td>
<td>-.025***</td>
<td>-.026***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>handbill</td>
<td>.031***</td>
<td>.024***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>liquid</td>
<td>-.331***</td>
<td>-.297***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>konzentrat</td>
<td>-.318***</td>
<td>-.323***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>color</td>
<td>-.120***</td>
<td>-.143***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>sensitiv</td>
<td>-.031***</td>
<td>-.121***</td>
</tr>
<tr>
<td></td>
<td>(.01)</td>
<td>(.01)</td>
</tr>
<tr>
<td>gimmick</td>
<td>.113***</td>
<td>.065***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>extrasize</td>
<td>.045***</td>
<td>-.033***</td>
</tr>
<tr>
<td></td>
<td>(.01)</td>
<td>(.01)</td>
</tr>
<tr>
<td>bigpack</td>
<td>.404***</td>
<td>.333***</td>
</tr>
<tr>
<td></td>
<td>(.01)</td>
<td>(.01)</td>
</tr>
<tr>
<td>countc56ad</td>
<td>-.001**</td>
<td>-.001***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>GLdumB</td>
<td>.594***</td>
<td>.278***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>duration</td>
<td>-.003***</td>
<td>.002***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>constant</td>
<td>.793***</td>
<td>.801***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>dummies</td>
<td>brand</td>
<td>No</td>
</tr>
<tr>
<td>No. of obs</td>
<td>375,448</td>
<td>375,448</td>
</tr>
<tr>
<td>R2</td>
<td>.52</td>
<td>.54</td>
</tr>
</tbody>
</table>

Note: Asterisks indicate significance levels at the significance levels $\alpha = 0.05, **= 0.01, ***=0.001$. Variable countc56ad is total number of contacts within a 3 month period per market.
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.

In table 14 the complete combined data with all product alternatives are used. In table 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 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 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.\footnote{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.}

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 15 and 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 $R^2$ is much lower than compared to table 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.

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.
References


A Tables and Figures

**Figure 11.** Geographic Distribution of Purchase Trips according to the Zip Code of Trip Destination
Table 15. Linear Probability Model and Discrete Choice Logit Model without no-purchase alternatives

<table>
<thead>
<tr>
<th>Incidents</th>
<th>LPM</th>
<th>DC Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>Variables</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
</tr>
<tr>
<td>preis</td>
<td>.004***</td>
<td>-.004***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>inh</td>
<td>-.006***</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>feature</td>
<td>.044***</td>
<td>.040***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>display</td>
<td>.017***</td>
<td>.012***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>preeflag</td>
<td>.018***</td>
<td>.008***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>handbill</td>
<td>-.004</td>
<td>-.009***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>liquid</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>konzentrat</td>
<td>-.016***</td>
<td>-.012***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>color</td>
<td>-.007***</td>
<td>-.005***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>sensitiv</td>
<td>.025***</td>
<td>.019***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>gimmick</td>
<td>.011***</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>extrasize</td>
<td>.023***</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>bigpack</td>
<td>.001</td>
<td>-.000</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>countc56ad</td>
<td>.001***</td>
<td>-.001***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>GLdumB</td>
<td>.672***</td>
<td>.180***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.01)</td>
</tr>
<tr>
<td>duration</td>
<td>-.000***</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>constant</td>
<td>.080***</td>
<td>.079***</td>
</tr>
<tr>
<td></td>
<td>(.00)</td>
<td>(.00)</td>
</tr>
<tr>
<td>dummies</td>
<td>brand No</td>
<td>Yes***</td>
</tr>
<tr>
<td>No. of obs</td>
<td>232,643</td>
<td>232,643</td>
</tr>
<tr>
<td>R2</td>
<td>.01</td>
<td>.01</td>
</tr>
</tbody>
</table>

Note: Asterisks indicate significance levels at the significance levels $\alpha^* = 0.05$, $** = 0.01$, $*** = 0.001$. Variable countc56ad is total number of contacts within a 3-month period per market.
### Table 16. Linear Probability Model and Discrete Choice Logit Model without no-purchase alternatives and private labels

<table>
<thead>
<tr>
<th>Variables</th>
<th>LPM</th>
<th>DC Logit</th>
<th>LPM</th>
<th>DC Logit</th>
<th>LPM</th>
<th>DC Logit</th>
<th>LPM</th>
<th>DC Logit</th>
<th>LPM</th>
<th>DC Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef./SE</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
<td>Coef./SE</td>
</tr>
<tr>
<td>preis</td>
<td>-.002***</td>
<td>-.005***</td>
<td>-.003***</td>
<td>-.002</td>
<td>-.038</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.03)</td>
</tr>
<tr>
<td>inh</td>
<td>.001</td>
<td>.002*</td>
<td>.002</td>
<td>.002</td>
<td>.041</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.04)</td>
</tr>
<tr>
<td>feature</td>
<td>.070***</td>
<td>.065***</td>
<td>.057***</td>
<td>.009</td>
<td>.163</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.18)</td>
</tr>
<tr>
<td>display</td>
<td>.015***</td>
<td>.014***</td>
<td>.006*</td>
<td>-.022</td>
<td>-.063</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.10)</td>
</tr>
<tr>
<td>priceflag</td>
<td>.006*</td>
<td>.005</td>
<td>.003</td>
<td>-.036***</td>
<td>.056</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.10)</td>
</tr>
<tr>
<td>handbill</td>
<td>-.005</td>
<td>-.006</td>
<td>.001</td>
<td>.020</td>
<td>.085</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.12)</td>
</tr>
<tr>
<td>liquid</td>
<td>-.004</td>
<td>-.005</td>
<td>-.001</td>
<td>-.020</td>
<td>-.049</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.09)</td>
</tr>
<tr>
<td>konzentrat</td>
<td>-.009***</td>
<td>-.011***</td>
<td>-.008***</td>
<td>-.038**</td>
<td>.041</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.10)</td>
</tr>
<tr>
<td>color</td>
<td>.003</td>
<td>-.004</td>
<td>-.003</td>
<td>-.016</td>
<td>-.027</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.08)</td>
</tr>
<tr>
<td>sensitiv</td>
<td>.022***</td>
<td>.009</td>
<td>.003</td>
<td>.004</td>
<td>.053</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.20)</td>
</tr>
<tr>
<td>gimmick</td>
<td>.005</td>
<td>.002</td>
<td>.010*</td>
<td>.016</td>
<td>.068</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.14)</td>
</tr>
<tr>
<td>extrasize</td>
<td>.008</td>
<td>-.002</td>
<td>-.001</td>
<td>-.023</td>
<td>-.130</td>
<td>(.00)</td>
<td>(.01)</td>
<td>(.00)</td>
<td>(.02)</td>
<td>(1.6)</td>
</tr>
<tr>
<td>bigpack</td>
<td>-.005</td>
<td>-.007</td>
<td>-.002</td>
<td>.011</td>
<td>.034</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.02)</td>
<td>(2.3)</td>
</tr>
<tr>
<td>countc56ad</td>
<td>-.000</td>
<td>-.001**</td>
<td>-.000</td>
<td>-.000</td>
<td>-.004</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>GLdumB</td>
<td>.782***</td>
<td>.244***</td>
<td>1.680***</td>
<td>(.00)</td>
<td>(.01)</td>
<td>(.00)</td>
<td>(.01)</td>
<td>(.00)</td>
<td>(.01)</td>
<td>(.09)</td>
</tr>
<tr>
<td>duration</td>
<td>-.001***</td>
<td>-.000</td>
<td>(.00)</td>
<td>(.00)</td>
<td>(.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>.105***</td>
<td>.160***</td>
<td>.136***</td>
<td>.744***</td>
<td>(.00)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>brand</td>
<td>No</td>
<td>Yes***</td>
<td>Yes***</td>
<td>Yes***</td>
<td>Yes***</td>
<td>Yes***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of obs</td>
<td>107,191</td>
<td>107,191</td>
<td>90,243</td>
<td>13,563</td>
<td>5,265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>.00</td>
<td>.01</td>
<td>.28</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Asterisks indicate significance levels at the significance levels α* = 0.05, *** = 0.01, **** = 0.001. Variable countc56ad is total number of contacts within a 3-month period per market.

47
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 17. Overview of Value Label for variables hhfpsex, psex1, ..., psex8 and stand: gender & famstatus

<table>
<thead>
<tr>
<th>gender</th>
<th>Value</th>
<th>Description</th>
<th>famstatus</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>male</td>
<td>0</td>
<td>No information</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>female</td>
<td>1</td>
<td>Unmarried</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
<td>Married</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>3</td>
<td>Divorced</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>4</td>
<td>Widowed</td>
<td></td>
</tr>
</tbody>
</table>

Table 18. Overview of Value Label for variable schulab: degree

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unknown</td>
</tr>
<tr>
<td>2</td>
<td>Secondary (lower level) School visited or completed (HAUPTSCHULE und kein Abschluss))</td>
</tr>
<tr>
<td>3</td>
<td>Secondary School of GDR (until 1989) (POLYTECHN. OBERSCHULE)</td>
</tr>
<tr>
<td>4</td>
<td>Secondary School (REALSCHULE)</td>
</tr>
<tr>
<td>5</td>
<td>Restricted A-Levels (FACHHOCHSCHULREIFE)</td>
</tr>
<tr>
<td>6</td>
<td>A-Levels, high school graduate (ALLG HOCHSCHULREIFE)</td>
</tr>
</tbody>
</table>
### Table 19. Overview of Value Label for variable bstel: occupation

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

*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 20. Overview of Value Label for variables ng and bik: nielsenarea & community

<table>
<thead>
<tr>
<th>nielsenarea</th>
<th>Value</th>
<th>Description</th>
<th>community</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Hamburg, Bremen, Schleswig-Holstein, Niedersachsen</td>
<td>1</td>
<td>&lt;2K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Nordrhein-Westfalen</td>
<td>2</td>
<td>2-5K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Hessen, Rheinland-Pfalz, Saarland</td>
<td>3</td>
<td>5-20K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Baden-Württemberg</td>
<td>4</td>
<td>20-50K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Bayern</td>
<td>5</td>
<td>50-100K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>Westberlin</td>
<td>6</td>
<td>100-500K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Ostberlin</td>
<td>7</td>
<td>500K+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Mecklenburg-Vorpommern, Brandenburg, Sachsen-Anhalt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>Thüringen, Sachsen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 21. Overview of Value Label for variables duft, kons and uwg: duftcode, konscode & uwgcode

<table>
<thead>
<tr>
<th>duftcode</th>
<th>Description</th>
<th>Value</th>
<th>Description</th>
<th>Description (translated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aktivfrisch</td>
<td>1</td>
<td>Creme</td>
<td>Creme</td>
</tr>
<tr>
<td>2</td>
<td>Aloe Vera</td>
<td>2</td>
<td>Fluessig</td>
<td>Liquid</td>
</tr>
<tr>
<td>3</td>
<td>Alpine</td>
<td>3</td>
<td>Gel</td>
<td>Gel</td>
</tr>
<tr>
<td>4</td>
<td>Alpine Fresh</td>
<td>4</td>
<td>Nuesse</td>
<td>Nuts</td>
</tr>
<tr>
<td>5</td>
<td>Antibakteriell</td>
<td>5</td>
<td>Paste</td>
<td>Paste</td>
</tr>
<tr>
<td>6</td>
<td>Apfel</td>
<td>6</td>
<td>Pulver</td>
<td>Powder</td>
</tr>
<tr>
<td>7</td>
<td>Apfelfrische</td>
<td>7</td>
<td>Schaum</td>
<td>Foam</td>
</tr>
<tr>
<td>8</td>
<td>Bluetenzauber</td>
<td>8</td>
<td>Tabletten</td>
<td>Tabletts</td>
</tr>
<tr>
<td>9</td>
<td>Citrus</td>
<td>9</td>
<td>Waschkissen</td>
<td>Cushions</td>
</tr>
<tr>
<td>10</td>
<td>Cotton Fields</td>
<td>999</td>
<td>NOP</td>
<td>NOP</td>
</tr>
<tr>
<td>11</td>
<td>Fresh Magic</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Frische</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Fruehlingsfrisch</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Green Lemon</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Kuschelduft</td>
<td>14</td>
<td>Feinwaschmittel</td>
<td>Silk Detergent</td>
</tr>
<tr>
<td>16</td>
<td>Lavendel</td>
<td>15</td>
<td>Gardinenwaschmittel</td>
<td>Curtain Detergent</td>
</tr>
<tr>
<td>17</td>
<td>Limonenfrische</td>
<td>16</td>
<td>Kaltwaschmittel</td>
<td>Low Temperature Detergent</td>
</tr>
<tr>
<td>18</td>
<td>Mango Fresh</td>
<td>17</td>
<td>Vollwaschmittel</td>
<td>Normal Detergent</td>
</tr>
<tr>
<td>19</td>
<td>Meeresfrische</td>
<td>18</td>
<td>Wollwaschmittel</td>
<td>Wool Detergent</td>
</tr>
<tr>
<td>20</td>
<td>Neutral</td>
<td>19</td>
<td>NOP</td>
<td>NOP</td>
</tr>
<tr>
<td>21</td>
<td>O Angabe</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Orange Fresh</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pfirsch &amp; Limone</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Pfirsichtraum</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Pink Grapefruit</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Pure Frische</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Rosenfrische</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Sommerfrische</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Spring Fresh</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Sunny Peach</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Thai Paradies</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Tropic</td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Wiesenfrische</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Wilde Orchidee</td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Wildrose</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>Winterfrisch</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Zitronenfrisch</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>999</td>
<td>NOP</td>
<td>37</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.