

CONSUMERS' PRICE ELASTICITIES IN PRICE
COMPARISON WEBSITES - DETERMINANTS AND
IMPLICATIONS

DIPLOMARBEIT

zur Erlangung des akademischen Grades
Mag.rer.soc.oec.

im Diplomstudium
WIRTSCHAFTSWISSENSCHAFTEN

Eingereicht von
MARIO HOFER, 0255509

Eingereicht bei
PROF. DR. FRANZ HACKL

Institut für Volkswirtschaftslehre
Sozial- und Wirtschaftswissenschaftliche Fakultät
Johannes Kepler Universität, Linz

Februar 2010

Research is to see what everybody else has seen, and to think what nobody else has thought.

— Albert Szent-Gyorgyi

To Johann and Gabriela Hofer
for your patience, support and encouragement.

ABSTRACT

The intention of this thesis is a detailed analysis of the concept of the price elasticity of demand. The presented analysis is of twofold nature. In a first approach the thesis assumes that the price elasticity of demand is not exogenously given, but rather an endogenously determined variable. The goal therefore is to identify the factors which determine the price elasticity of demand. These factors consist, amongst others, of variables relating to product quality and also to the availability and accessibility of substitutes. The empirical results show that the proposed variables have a significant impact on the elasticity.

In a next step the thesis analyses the impacts of the price elasticity of demand on the actions of retailers and thus the market structure and competition. To represent competition and market structure this thesis uses the number of changes of the price leader of the respective products. The empirical results approve the formulated hypothesis that the price elasticity has an impact on the actions taken by retailers. The results are significant on statistical but not on economical grounds.

Data for all computations and estimations in this thesis were obtained from Geizhals, the largest Austrian price comparison website. Both approaches are estimated with several elasticity alternatives. Hence this thesis also proposes several models to estimate the price elasticity of demand in the context of a price comparison website.

ZUSAMMENFASSUNG

Der Zweck dieser Diplomarbeit ist eine detaillierte Analyse des Konzepts der Preiselastizität der Nachfrage. Die hier vorgestellte Analyse setzt sich aus zwei Teilen zusammen. Ausgehend von der Annahme, dass die Preiselastizität nicht exogen gegeben ist, sondern endogen bestimmt wird, werden Faktoren bestimmt, welche einen Einfluss auf die Elastizität ausüben. Zu diesen Faktoren gehören u.a. die Produktqualität, sowie die Verfügbarkeit von Substituten. Der Einfluss der vorgeschlagenen Faktoren wird durch die empirischen Resultate bestätigt.

In einem weiteren Schritt werden die Auswirkungen der Preiselastizität auf die Handlungen der Verkäufer, sowie die Art des Wettbewerbs und die Marktstruktur untersucht. Die Art des Wettbewerbs und die Marktstruktur werden durch die Anzahl der Wechsel der Preisführer des jeweiligen Produkts dargestellt. Die Resultate der empirischen Untersuchung bestätigen die aufgestellten Hypothesen, jedoch weisen diese eine zu geringe ökonomische Signifikanz auf.

Die Daten, auf denen sämtliche Berechnungen und Schätzungen in dieser Diplomarbeit basieren, stammen von Geizhals, der größten österreichischen Preisvergleichswebsite. Beide Teile der Analyse dieser Diplomarbeit bauen auf verschiedene Typen von Elastizitäten auf. Aus diesem Grund beschäftigt sich ein eigenes Kapitel mit der Schätzung von Preiselastizitäten im Rahmen einer Preisvergleichswebsite.

Sometimes our light goes out but is blown into flame by another human being. Each of us owes deepest thanks to those who have rekindled this light.

— Albert Schweitzer

ACKNOWLEDGMENTS

I want to thank Prof. Dr. Franz Hackl for providing me with the opportunity to participate in a challenging research project and for his input and support.

Many thanks to Markus Teufl for putting up with me during a myriad of discussions on technical subtleties.

I especially want to thank Isabell for her patience and for being the main source of my motivation.

CONTENTS

1	INTRODUCTION AND MOTIVATION	1
2	DETERMINING THE PRICE ELASTICITY OF DEMAND	3
2.1	Theoretical Framework	3
2.1.1	Isoelastic Demand Curve	3
2.1.2	Linear Demand Curve	4
2.2	Data	5
2.2.1	Data Structure	6
2.2.2	Variable Description	8
2.2.3	Data Overview	10
2.3	Construction Of Demand Curves	11
2.3.1	Dealing With Simultaneity	12
2.3.2	Normalization Of Clicks	13
2.3.3	Aggregation And Accumulation Of Clicks	15
2.4	Elasticity Description and Regression Setup	16
2.5	Clearinghouse Models - An Alternative Estimation Approach	20
2.6	Estimation Results	21
2.6.1	Detection And Removal Of Zero-Elasticities	22
2.6.2	General Overview Of The Full Dataset Results	26
2.6.3	Detecting Outliers	28
2.6.4	Ensuring Data Quality - The Detailed Analysis Of A Random Subsample	30
2.6.5	Improving Data Quality	39
2.7	Conclusion	42
3	FACTORS INFLUENCING THE PRICE ELASTICITY OF DEMAND	45
3.1	Previous Research	45
3.2	Dataset Description	48
3.2.1	An Excursion On Regression-Weights	51
3.3	Hypothesis	53
3.4	Estimation Results And Model Extensions	55
3.4.1	Estimation Results Of The Regressions Based On The Full Dataset	56
3.4.2	Results Of The Enhanced Model	60
3.4.3	Estimation Results For The Cleansed Dataset	63
3.5	Conclusion And Remaining Problems	66
4	THE PRICE ELASTICITY OF DEMAND AS A DETERMINANT OF MARKET STRUCTURE	71
4.1	Introduction	71
4.2	Dataset Description	72
4.3	Hypothesis	75
4.4	Estimation Results	77
4.4.1	A Short Remark Concerning The Number Of Price Leader Changes	78
4.4.2	Results Of The First Regression	79
4.4.3	Results For The Full And The Cleansed Dataset	80
4.5	Conclusion And Remaining Problems	86
5	CONCLUSION	89
A	APPENDIX - REGRESSION RESULTS AND STATISTICS	91
A.1	Summary Statistics For The Explanatory And Control Variables	91

A.2	Summary Statistics For The Estimated Elasticities	93
A.3	Summary Statistics For The Number Of Price Leader Changes	94
A.4	Estimation Results Of Chapters 3 And 4	94
B	APPENDIX - DIAGRAMS, GRAPHS AND TABLES	115
B.1	Estimated Demand Curves Of The 40-Product Sample	118
C	APPENDIX - LISTINGS	159
C.1	Creation Of Product-Specific Data For The Second Stage Regressions	159
C.2	Creation Of Subsubcategory-Specific Data	164
C.3	Outlier Detection With Grubbs' Test	167
D	MISCELLANEOUS CONCEPTS	171
D.1	Explaining The Brandrank	171
D.2	The MySQL INFORMATION_SCHEMA	172
D.3	An Alternative Criterion For Data Quality	173
	BIBLIOGRAPHY	177

LIST OF FIGURES

Figure 1	General product information of a typical Geizhals product page	6
Figure 2	List of offers of a typical Geizhals product page	6
Figure 3	Simplified conceptual schema of the Geizhals database	7
Figure 4	Number of products by category	11
Figure 5	Number of clicks by category	11
Figure 6	Tracing out the demand curve by shifting the supply curve	12
Figure 7	Graphical overview on the full dataset estimation results	27
Figure 8	Summary statistics of the elasticities and the corresponding \bar{R}^2 and t-values for the 40-product-sample	33
Figure 9	Scatterplot featuring zero-click-offers	34
Figure 10	Share of zero-click-offers	35
Figure 11	Influence of zero-click-offers at high prices	36
Figure 12	Correlation matrix of the 40-product-sample	38
Figure 13	Share of offers with zero-LCT grouped into categories	40
Figure 14	Summary statistics of the elasticities and the corresponding \bar{R}^2 and t-values for the cleansed dataset	41
Figure 15	Summary statistics of the elasticities and the corresponding \bar{R}^2 and t-values for the full dataset	42
Figure 16	Example 1 of a problematic observation from the cleansed dataset (Product: Manfrotto tripod)	67
Figure 17	Example 2 of a problematic observation from the cleansed dataset (Product: Tamron 55-200mm lens)	68
Figure 18	Example of a suitable demand curve which has been ruled out by Grubbs' test (Product HP Pavilion notebook)	69
Figure 19	Projection of the start- and end-timestamps of offers	71
Figure 20	Time line with price leaders	72
Figure 21	Histogram of the number of price leader changes	78
Figure 22	Simplified conceptual schema of the complete Geizhals database	115
Figure 23	Layout of a page presenting alternative demand curves for a product	118
Figure 24	Product: 6580, Canon BCI-3eBK Tintenpatrone schwarz (BJC-3000/6000/S400/450/600/630/6300/MultiPASS C755): Hardware \Rightarrow Verbrauchsmaterial	119
Figure 25	Product: 35137, FireWire IEEE-1394 Kabel 6pin/4pin, 1.8m: Hardware \Rightarrow KabelZubehr	120
Figure 26	Product: 77828, OkI 1103402 Toner schwarz (B4200): Hardware \Rightarrow Verbrauchsmaterial	121
Figure 27	Product: 79079, Twinhan VisionDTV DVB-S PCI Sat CI (1030A/1032A): Hardware \Rightarrow PCVideo	122

- Figure 28 Product: 86015, Sony Vaio PCGA-BP2V Li-Ionen-Akku: Hardware ⇒ Notebookzubehr 123
- Figure 29 Product: 96539, Sony MDR-DS3000: Hardware ⇒ PCAudio 124
- Figure 30 Product: 97556, Digitus DA-70129 Fast IrDa Adapter, USB 1.1: Hardware ⇒ Mainboards 125
- Figure 31 Product: 111148, Komplettsystem AMD Athlon 64 3200+, 2048 MB RAM: Hardware ⇒ Systeme 126
- Figure 32 Product: 115211, No-Name/Diverse Mousepad schwarz: Hardware ⇒ Eingabegeräte 127
- Figure 33 Product: 115433, HP Q6581A Fotopapier seidenmatt 42", 190g, 30.5m: Hardware ⇒ Verbrauchsmaterial 128
- Figure 34 Product: 120019, Konica Minolta magicolor 5430 Toner schwarz (1710582-001): Hardware ⇒ Verbrauchsmaterial 129
- Figure 35 Product: 128259, Zalman CNPS7700-Cu CPU-Kühler (Sockel 478/775/754/939/940): Hardware ⇒ Luftkhlung 130
- Figure 36 Product: 130965, FSC Pocket LOOX 710/720 Bump Case (S26391-F2611-L500): Hardware ⇒ PDAs-GPS 131
- Figure 37 Product: 132933, Apple iPod AV Kabel (M9765G/A): AudioHIFI ⇒ PortableAudio 132
- Figure 38 Product: 133848, Targus XL Metro Messenger Notebook Case Notebooktasche (TCG200): Hardware ⇒ Notebookzubehr 133
- Figure 39 Product: 134972, Kingston ValueRAM DIMM 256MB PC2-533 ECC DDR2 CL4 (KVR533D2E4/256): Hardware ⇒ Speicher 134
- Figure 40 Product: 137437, Canon DCC-80 Softcase (0021X145): VideoFotoTV ⇒ FotoVideozubehr 135
- Figure 41 Product: 141318, Cherry Linux, PS/2, DE (G83-6188): Hardware ⇒ Eingabegeräte 136
- Figure 42 Product: 142849, Philips HR1861/00 Entsafter : ⇒ 137
- Figure 43 Product: 145401, Adobe: GoLive CS2 (deutsch) (PC) (23200457): ⇒ 138
- Figure 44 Product: 160965, Creative Sound Blaster X-Fi Platinum (70SB046002000): Hardware ⇒ PCAudio 139
- Figure 45 Product: 167564, Intel Xeon DP 3.80GHz, 200MHz FSB, 2048kB Cache, 604-pin boxed passiv (BX80546KG3800FP): Hardware ⇒ CPUs 140
- Figure 46 Product: 169429, MSI K8MM3-V, K8M800 (PC3200 DDR) (MS-7181-020R): Hardware ⇒ Mainboards 141
- Figure 47 Product: 187584, Samsung SyncMaster 940N Pivot, 19", 1280x1024, analog (LS19HAATSB): Hardware ⇒ Monitore 142
- Figure 48 Product: 193387, ELO: EloOffice 7.0 Update (deutsch) (PC) (9303-70-49): Software ⇒ SicherheitBackup 143
- Figure 49 Product: 198925, Liebherr KTP 1810: Household ⇒ Kchengertegro 144
- Figure 50 Product: 200404, ASUS M2NPV-VM, GeForce 6150/MCP 430 (dual PC2-800 DDR2): Hardware ⇒ Mainboards 145

Figure 51	Product: 203899, Hama 35in1 Card Reader, USB 2.0 (55312): Hardware ⇒ SpeichermedienLesegeräte	146
Figure 52	Product: 204301, Sony HVL-F56AM Blitzgerät: VideoFotoTV ⇒ FotoVideozubehr	147
Figure 53	Product: 205234, Hähnel HL-2LHP Li-Ionen-Akku (1000 188.2): VideoFotoTV ⇒ FotoVideozubehr	148
Figure 54	Product: 210626, Gigabyte GA-945GM-S2, i945G (dual PC2-5300U DDR2): Hardware ⇒ Mainboards	149
Figure 55	Product: 210768, Acer LC.08001.001 80GB HDD: Hardware ⇒ Notebookzubehr	150
Figure 56	Product: 216426, Samsung ML-4551ND: Hardware ⇒ DruckerScanner	151
Figure 57	Product: 217675, Adobe: Photoshop Elements 5.0 (deutsch) (PC) (29230441): ⇒	152
Figure 58	Product: 219507, Canon BCI-6 Multipack Color (S800/820D/BJC 8200/9000/i9100/i9950) (4706A022): Hardware ⇒ Verbrauchsmaterial	153
Figure 59	Product: 220621, V7 Videoseven L22WD, 22", 1680x1050, analog/digital, Audio: Hardware ⇒ Monitore	154
Figure 60	Product: 224495, Apple MacBook Pro Core 2 Duo, 15.4", T7400 2x 2.16GHz, 2048MB, 160GB: Hardware ⇒ Notebooks	155
Figure 61	Product: 227850, Sony Walkman NW-S703FV 1GB violett: AudioHIFI ⇒ PortableAudio	156
Figure 62	Product: 230160, Seagate SV35 320GB (ST3320620AV): Hardware ⇒ Festplatten	157
Figure 63	Product: 230429, Trust WB-5400 4 Megapixel Webcam USB 2.0 (15007): Hardware ⇒ PCVideo	158
Figure 64	Large wedge due to inappropriate data	174
Figure 65	Acceptable data, which yields only a small wedge	175
Figure 66	Angle of intersection of two lines	175

LIST OF TABLES

Table 1	Variables of the first stage regression	9
Table 2	Aggregation and accumulation of clicks	15
Table 3	Explanation of abbreviations used in elasticity variable notations	17
Table 4	Summary statistics of two elasticities of the full dataset	28
Table 5	Summary statistics for the cleansed data set	30
Table 6	Description of the random sample (part 1)	31
Table 7	Description of the random sample (part 2)	32
Table 8	Alternative data filters	36
Table 9	Measures of product quality	48
Table 10	Measures of product substitutability	49
Table 11	Dummies indicating the brand rank of a product	50
Table 12	Dummies indicating the category for a product	50
Table 13	Miscellaneous variables	50

Table 14	Enhanced model, full dataset regression results - version 2	61
Table 15	Impact on <code>c_elast</code> if an explanatory variable changes by one standard deviation (N=14814)	63
Table 16	Impact on <code>c_elast</code> if an explanatory variable changes by one standard deviation (N=189)	66
Table 17	Measures of product quality	73
Table 18	Measures of product substitutability	73
Table 19	Dummies indicating the brand rank of a product	74
Table 20	Dummies indicating the category for a product	74
Table 21	Miscellaneous variables	75
Table 22	Exemplary summary statistics for the cleansed dataset (N=181)	80
Table 23	Summary statistics of the explaining variables (N=14814)	85
Table 24	Summary statistics for the explanatory variables of the full dataset filtered in accordance to <code>c_elast</code> and regression version 2 (N=14814)	91
Table 25	Summary statistics for the explanatory variables of the full dataset, filtered in accordance to <code>cv_lctw_c_elast</code> and regression version 2 (N=4824)	92
Table 26	Summary statistics for the explanatory variables of the cleansed dataset filtered in accordance to <code>c_elast</code> and regression version 2 (N=189)	92
Table 27	Summary statistics of the explanatory variables of the cleansed dataset, filtered in accordance to <code>cv_lctw_c_elast</code> and regression version 2 (N=56)	92
Table 28	Summary statistics of the elasticities for the full dataset, filtered in accordance to regression version 2	93
Table 29	Summary statistics of the elasticities for the cleansed dataset, filtered in accordance to regression version 2	93
Table 30	Summary statistics of the number of price leader changes (<code>num_pleader_change</code>). The variable name denotes the elasticity which has been used as the filter criteria.	94
Table 31	Summary statistics of the number of price leader changes (<code>num_pleader_change</code>). The variable name denotes the elasticity which has been used as the filter criteria.	94
Table 32	Full dataset regression results - version 1 (cf. section 3.4.1 on page 56)	95
Table 33	Full dataset regression results - version 2 (cf. section 3.4.1 on page 56)	96
Table 34	Full dataset regression results - version 3 (cf. section 3.4.1 on page 56)	97
Table 35	Enhanced model, cleansed dataset regression results - version 1 (cf. section 3.4.3 on page 63)	98
Table 36	Enhanced model, cleansed dataset regression results - version 2 (cf. section 3.4.3 on page 63)	99
Table 37	Enhanced model, cleansed dataset regression results - version 3 (cf. section 3.4.3 on page 63)	100

Table 38	Enhanced model, full dataset regression results - version 1 (cf. section 3.4.2 on page 60)	101
Table 39	Enhanced model, full dataset regression results - version 2 (cf. section 3.4.2 on page 60)	102
Table 40	Enhanced model, full dataset regression results - version 3 (cf. section 3.4.2 on page 60)	103
Table 41	Enhanced model, relaxed filter cleansed dataset regression results - version 2 (cf. section 3.4.2 on page 60)	104
Table 42	Price leader changes, Poisson first regression - version 1 (cf. section 4.4.2 on page 79)	105
Table 43	Price leader changes, Poisson first regression - version 2 (cf. section 4.4.2 on page 79)	105
Table 44	Price leader changes, Poisson first regression, cleansed dataset - version 3 (cf. section 4.4.2 on page 79)	106
Table 45	Price leader changes, Poisson first regression, Full Dataset - version 1 (cf. section 4.4.2 on page 79)	106
Table 46	Price leader changes, Poisson first regression Full Dataset - version 2 (cf. section 4.4.2 on page 79)	107
Table 47	Price leader changes, Poisson first regression full dataset - version 3 (cf. section 4.4.2 on page 79)	107
Table 48	Price leader changes, Poisson regression, full dataset - version 1 (cf. section 4.4.3 on page 80)	108
Table 49	Price leader changes, Poisson regression, full dataset - version 2 (cf. section 4.4.3 on page 80)	109
Table 50	Price leader changes, Poisson regression, full dataset - version 3 (cf. section 4.4.3 on page 80)	110
Table 51	Price leader changes, Poisson regression, cleansed dataset - version 1 (cf. section 4.4.3 on page 80)	111
Table 52	Price leader changes, Poisson regression, cleansed dataset - version 2 (cf. section 4.4.3 on page 80)	112
Table 53	Price leader changes, Poisson regression, cleansed dataset - version 3 (cf. section 4.4.3 on page 80)	113
Table 54	Subsubcategory specific second stage variables	116
Table 55	Product specific second stage variables	117

LISTINGS

Listing 1	Detection of missing offers	23
Listing 2	Creation of destination table	159
Listing 3	Database connection	159
Listing 4	Main body	160
Listing 5	Computation of product-specific data	162
Listing 6	Computation of subsubcategory-specific data	164
Listing 7	Outlier detection using Grubbs' test	167
Listing 8	Computation of clicks for a brand	171
Listing 9	Retrieving the columns of a table with the INFORMATION_SCHEMA	172

Listing 10 Retrieving a list of table names with the INFORMATION_SCHEMA 173

ACRONYMS

UML Unified Modeling Language
LCT Last Click Through
OLS Ordinary Least Squares
RMSE Root Mean Squared Error
DF Degrees of Freedom
QMLE Quasi-Maximum Likelihood Estimator
QML Quasi-Maximum Likelihood
VBA Visual Basic for Applications
RHS right-hand-side
LHS left-hand-side
JKU Johannes Kepler University
NLS Non-linear Least Squares
IV instrumental variables
2SLS two stage least squares

INTRODUCTION AND MOTIVATION

Many standard economic textbooks like [Varian \(2001\)](#) describe the price elasticity of demand as a measure of the sensitivity of demand concerning changes in price or income. As a first approach to measure sensitivity of demand, intuition would suggest to use the slope of the demand curve, because it is defined as the quotient of the change in the quantity and the change in the price. Therefore the slope of the demand curve gives information on how the demand changes, in accordance to changes in the price.

Although the slope seems to be a good measure for price sensitivity, it still features some drawbacks. The most significant drawback is that the change in demand is always denoted in a certain unit. Therefore if one measure demand in a different unit, the slope will also change. This can be easily seen if one takes a look at the following example: Assume that ink cartridges in packs, where each pack contains five cartridges. Furthermore assume that the producing company sells 100 packs at a price of €50 each and that the company could increase its sales to 120 packs if it would reduce the price per pack by €5. That yields a slope of $-1/4$. However, if one measures the quantity sold not in packs but in single cartridges, a price reduction of €5 per pack would increase the demand from 500 cartridges to 600 cartridges and hence a slope of the demand curve of $-1/20$, which is obviously different from $-1/4$.

The above described problem makes clear that one needs a dimensionless measure for price sensitivity of demand. This is exactly what the concept of an elasticity does. The price elasticity of demand tells by how many percent does demand change if the price changes by 1%. This percentage change does not depend on any unit. In the example of the ink cartridges the demand increases from 100 to 120 if measured in packs and from 500 to 600 if measured in single cartridges. Nevertheless, in both of the cases the increase accounts for 20%.

Although the concept of the price elasticity of demand is well known, there is hardly any literature which assumes this concept to be endogenous. The vast majority of empirical research just tries to estimate the price elasticity of demand and treats it as exogenously given. One of the novel strategies in this thesis is that it not only estimates elasticities, it also tries to explain them by characteristics concerning the market, the products and the retailers. This approach enables this thesis to give answers to the questions whether the price elasticity is lower for high-quality or brand products. The second novelty is the idea to examine the impacts of the price elasticity of demand on the actions of retailers concerning their price policies and hence also the type of competition prevalent on the respective market. Nevertheless one has to notice that this thesis should be seen as a preliminary study for a larger research project of the department of economics at the Johannes Kepler University (JKU) in the field of online price comparison websites. The author of this thesis is aware of the econometric problems which arise when estimating demand curves. These problems encompass topics like endogeneity, simultaneity and heteroskedasticity. Unfortunately it is beyond the scope of this thesis to incorporate detailed solutions

for all of the aforementioned problems. For this reason this thesis only tries to alleviate these problems by making simplifying assumptions.

The data for this thesis has been obtained from the Preisvergleich Internet Services AG (Geizhals), which is the largest price comparison website in Austria. Geizhals gathers data on prices and clicks several times a day, therefore the advantage of this data is its high level of detail. Because of this high level of detail one can alleviate the problem of simultaneity by choosing a short period of observation. However, the fact that the dataset only contains clicks and not actual purchases constitutes a serious drawback. This drawback provides a further interesting challenge within the context of this thesis, since it has to deal with several questions. Apart from the problem that clicks have to be converted into actual sales, one must also find a way to use the click data to generate demand curves. This thesis suggests different approaches for both, the generation of demand curves and the definition of different types for the price elasticity of demand. Finally one has to evaluate the quality of the estimated elasticities and prepare them for their application in the second stage regressions.

For this reason the work of this thesis can be decomposed into two stages. The first stage is described in chapter 2 and deals with the set of regressions which is used to construct the demand curves and estimate the respective elasticities. The regressions of chapter 2 are therefore termed *first stage regressions*. The second stage consists of chapters 3 and 4. Whereas chapter 3 tries to identify and explain the factors which influence the price elasticity of demand and 4 tries to explain the price elasticity of demand as a determinant of market structure. The regressions of these chapters are termed *second stage regressions* because they rest upon the results from the first stage regressions.

DETERMINING THE PRICE ELASTICITY OF DEMAND

This chapter is concerned with the construction of a solid basement for the two preceding chapters. In order to explain factors influencing the elasticity of demand or to explore the impacts of the elasticity of demand on market structure, one needs a solid set of elasticities for a sufficient range of products. Hence the chapter starts with an introduction of the basic theoretical framework on the computation of the price elasticity of demand. In a next step it gives a description of the Geizhals data used for the estimation procedure. The further sections and subsections explicitly deal with problems and irregularities which are specific to the data from Geizhals.

2.1 THEORETICAL FRAMEWORK

In a first step one has to define the demand structure and hence the type of the elasticities to be estimated. This thesis deals with the following two types of demand curves (respectively elasticities):

ISOELASTIC DEMAND CURVE: A convex curve where the elasticity ϵ takes on a constant value for each point on the demand curve.

LINEAR DEMAND CURVE: An affine function of the form $f(x) = ax + b$. The elasticity ϵ changes when moving along the curve.

Independent of the type of demand curve, the price elasticity of demand for a demand function $Q_i = D(P_i, X_i)$, where Q_i is the demand faced by retailer i , P_i is the price charged by retailer i and X_i is a set of additional retailer-specific variables like shipping costs or retailer ratings is given by:

$$\epsilon \equiv \frac{\partial Q/Q}{\partial P/P} = \frac{\partial Q}{\partial P} \frac{P}{Q} \quad (2.1)$$

This definition can also be found in standard economics textbooks like [Chiang \(2005\)](#).

2.1.1 Isoelastic Demand Curve

The starting point will be a simple isoelastic specification for the demand curve which uses the price of a good and additional retailer specific information as inputs. This function can be stated as

$$Q = \frac{e^X}{p^\alpha} \quad (2.2)$$

where X can be any linear combination of additional retailer specific variables, e.g. $\beta_0 + \beta_1 \text{var}_1 + \beta_2 \text{var}_2 + \dots$

Using the definition of the price elasticity of demand one can compute

$$\frac{\partial Q}{\partial P} = -\alpha e^X p^{-\alpha-1} \quad (2.3)$$

and

$$\frac{P}{Q} = \frac{P}{e^{X/P^\alpha}} = \frac{P^{\alpha+1}}{e^X} \quad (2.4)$$

Multiplying both terms yields the following elasticity ϵ

$$\epsilon = -\alpha e^X P^{-\alpha-1} \frac{P^{\alpha+1}}{e^X} = -\alpha \quad (2.5)$$

From above results one can see that every demand curve which follows the specification $Q = e^X/P^\alpha$ will have a constant elasticity ϵ of $-\alpha$.

In a further step one has to show how such a demand specification can be estimated empirically. Given that $\partial \ln x / \partial x = 1/x$ one can use the definition of ϵ and rewrite

$$\frac{1}{Q} = \frac{\partial \ln Q}{\partial Q} \quad (2.6)$$

$$P = \frac{1}{\frac{\partial \ln P}{\partial P}} \quad (2.7)$$

Using this modification one can rewrite the definition of the elasticity of demand as

$$\epsilon \equiv \frac{\partial Q}{\partial P} \frac{P}{Q} = \frac{\partial Q}{\partial P} \frac{1}{Q} P = \frac{\partial Q}{\partial P} \frac{\partial \ln Q}{\partial Q} \frac{1}{\frac{\partial \ln P}{\partial P}} = \frac{\partial \ln Q}{\partial \ln P} \quad (2.8)$$

Above relation is essential for this thesis, since it shows that the elasticity of a general isoelastic demand specification can be computed as the first derivative of the log-version of the demand equation to the log-version of the price. Taking logs of the general demand specification $Q = e^X/P^\alpha$ yields

$$\ln Q = \ln \frac{e^X}{P^\alpha} = \ln \frac{e^X}{P^\alpha} = X - \alpha \ln P \quad (2.9)$$

Looking at equation 2.9 one can observe the following facts:

- The coefficient on $\ln P$ is $-\alpha$ and hence exactly the elasticity of the demand equation. Therefore one can use the log-log version of the equation to estimate the price elasticity of demand by using the method of Ordinary Least Squares (OLS).
- Even though the natural logarithm has been applied to the complete equation 2.2, equation 2.9 shows that X is included into the equation in its level form. Therefore it is not required to build the logs of any of the additional retailer specific variables¹.

2.1.2 Linear Demand Curve

One cannot safely assume that all demand curves are strictly isoelastic. For this reason this thesis also incorporates elasticities from linear

In the log-log model α is the elasticity of Q with respect to P and hence the price elasticity of demand.

¹ This is handy for practical reasons, since some of the retailer specific variables are dummy variables, i.e. they can only take on the values 0 and 1.

demand curves. The linear demand curve used in this thesis is the same affine function as described by [Varian \(2001\)](#).

$$Q = X - \alpha P \quad (2.10)$$

where X is a linear combination of retailer specific variables which may also include the satiation level of the market, i.e. the maximal marketable quantity of the good.

Using the definition from equation 2.1 one can write the elasticity of demand for the linear demand curve as

$$\epsilon = -\frac{\alpha P}{X - \alpha P} \quad (2.11)$$

The price elasticity of a general linear demand function.

As one can observe from equation 2.11 the elasticity of a linear demand function is not constant and varies along the curve. The absolute value of the elasticity $|\epsilon|$ is approaching ∞ when moving towards the P axis intercept and 0 when moving towards the Q axis intercept.

In order to make the elasticities estimated from various products comparable they have to be put on equal footing. One has to specify an exact point on the demand curve for which the elasticity is being calculated. To achieve comparable values, the elasticities will be computed at mean values so equation 2.11 becomes

$$\epsilon = -\alpha \frac{\bar{P}}{\bar{X} - \alpha \bar{P}} \quad (2.12)$$

The advantage of a linear demand function is that the parameter α can be estimated by, regressing equation 2.10 directly. Nevertheless one has to keep in mind that the estimation of the elasticity at mean values is only an approach to get at least some sort of comparable values.

2.2 DATA

The data used in this thesis was obtained from Geizhals², which has been founded as a small price comparison platform in 1996. In the year 2000 Geizhals has been integrated in the newly funded "Preisvergleich Internet Services AG". With an estimated transaction volume of 1.5 billion Euros and roughly 2.3 million unique clients, the website has grown into one of the largest e-commerce-platforms in the German speaking world. According to the company's website the platform features over 400,000 products from over 1,700 retailers resulting in over 11,000,000 prices which get updated on an hourly basis.


The website's product list is organized in categories like *Hardware*, *Software* or *Audio/HIFI*. Each category consists of subcategories which contain further subsubcategories. Consumers who are interested in buying a product or retrieving information about a product navigate through the categories and subcategories until they reach the page for the specific product. On this page Geizhals presents general product information which also includes ratings for the specific product and a list of retailers offering the product.

Figure 1 gives an impression of how the product description of a product page on Geizhals looks like. The product description shows the title, a photo of the product and basic product metrics. Furthermore

With its 400,000 products, 1,700 retailers and 11,000,000 prices Geizhals is one of the largest e-commerce-platforms in the German speaking world.

² <http://www.geizhals.at/>.

the description also features a list of links to test reports for the specific product.



Gainward BLISS GeForce GTX 275 Golden Sample, 896MB GDDR3, VGA, DVI, HDMI, PCIe 2.0 (0582)

Übersicht & Preise | Preisentwicklung | Bewertungen | Info beim Hersteller

Chiptakt: 648MHz, Speichertakt: 1185MHz, Shadertakt: 1404MHz • Chip: GT206 (GT200b) • Speicherinterface: 448-bit • Stream-Prozessoren: 240 • Textureinheiten: 80 • Fertigung: 55nm • Maximaler Verbrauch: 216W • DirectX: 10.0 • Shader Modell: 4.0 • Bauweise: Dual-Slot • Besonderheiten: unterstützt HDCP, 3-Way-SLI, werkseitig übertaktet

Es liegen noch keine Bewertungen für dieses Produkt vor ([Produkt bewerten](#)).

Testberichte:

- computerbase.de: [Alles auf einmal](#)
- PCGH.de: [Test: ATI Radeon HD 4890 gegen nVIDIA Geforce GTX 275](#)
- HT4U.net: [Zotac GeForce GTX 275](#)
- computerbase.de: [Test: 13 aktuelle Grafikkarten - 6 x ATI gegen 7 x Nvidia](#)
- fudzilla.com: [Gainward GTX 275 896MB beats the HD 4890 1GB](#)
- tweaktown.com: [Gainward GeForce GTX 275 Graphics Card](#)

Besucher, die diesen Artikel angesehen haben, haben auch die folgenden Artikel angesehen:

[anzeigen](#)

(Die Abbildungen mit Produkt

Figure 1: General product information of a typical Geizhals product page

Geizhals lists the retailers who are offering the specific product right beneath the product description. Figure 2 shows a schematic depiction of such a product list. Each entry in the list constitutes an offer of the current product from a specific retailer. An offer consists of the price of the product (including the VAT), the retailer who offers the product, an average retailer rating, information on the availability of the product at the specific retailer, the shipping costs charged by the retailer and a textfield with the exact product specifications coming from the retailer's website.

Preis in €*	Anbieter	Händler-Bewertung	Verfügbarkeit	Artikelbezeichnung des Händlers
			Versand**	
63,88	Austriahosting Infos AGB Forumsuche	Note: 1,05 47 Bewertungen	Artikel ist bestellt! Österreich: Vorkasse € 9,90, Kreditkarte € 9,90 plus € 1,60 Zuschlag. Lieferung in weitere Länder auf Anfrage. Abholung nach Vorbestellung möglich (A-7022 Schattendorf)	Asus 90-MIB510-G0EAY00Z M3A78-EM AM2+ 780G MATX Socket AM2+, AMD 780G/SB700 Chipset, FSB: Up to 5200/MT/s HyperTransport 3.0 interface for AM2+ CPU, 4 x DDR2 1066 /800/667, Max. 8 GB, 1 x PCIe x16, 1 x PCIe x1, 2 x PCI 2.2.. (14.08.2009, 13:10)
64,98	EGW ELECTRONICS EGW-Electronics Infos AGB Forumsuche	Note: 1,23 1109 Bewertungen	Artikel ist zur Zeit nicht verfügbar Österreich: Vorkasse € 4,80 Nachnahme € 8,40 Deutschland: Vorkasse € 9,90 Abholung nach Vereinbarung möglich (A-8510 Stainz)	Asus 90-MIB510-G0EAY00Z Asus M3A78-EM AM2+ 780G MATX Socket AM2+, AMD 780G/SB700 Chipset, FSB: Up to 5200/MT/s Hy (Art# B994495) (14.08.2009, 13:24)
67,50			Zentrale St. Pölten:	ASUS 90-MIB510-G0EAY00Z

Figure 2: List of offers of a typical Geizhals product page

Geizhals does not charge the consumers for using the platform's services. The user frequency is measured in clicks. A click or referral is made if a user clicks on a retailer's offer of a specific product.

2.2.1 Data Structure

The data used in this thesis is an extract from the fully fledged dataset provided by Geizhals. Over the last 3 years Geizhals supplied the Department of Economics of the Johannes Kepler University with live dumps of their production database on a daily basis. The live dumps

received by Geizhals were preprocessed and then fed into a MySQL database³.

A simplified conceptual schema of the Geizhals database is depicted in figure 3. As one can observe the central tables in the database are the *product*-table, the *retailer*-table, the *offer*-table and the *click*-table.

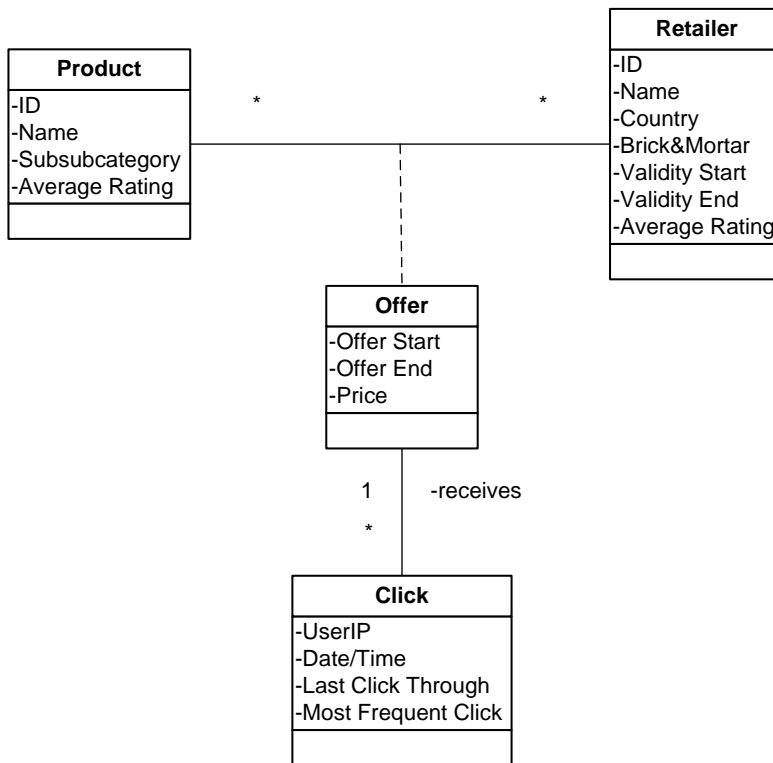


Figure 3: Simplified conceptual schema of the Geizhals database

A product is offered by at least one retailer. Furthermore a retailer offers at least one product. The link between a specific retailer and a specific product is called an offer. An offer is only available for a limited amount of time. The recorded clicks do not belong to a retailer or a product, but rather to the offer of a certain product from a certain retailer.

PRODUCT: Each record in this table represents a product in the Geizhals database and consists of a unique identifier for the product, the name of the product, the subsubcategory to which the product belongs to and average product ratings.

RETAILER: Each record in this table represents a specific retailer in the Geizhals database and consists of a unique identifier for the retailer, the country of origin of the retailer⁴, a dummy variable indicating whether the retailer has a brick and mortar presence, geographical coordinates, the validity period of the record and average retailer ratings.

OFFER: An *offer* is the link between a *product* and a *retailer* and shows that a certain retailer offers a specific product for a limited period

The most important tables in the Geizhals database contain information on products, retailers, offers and clicks.

³ An open source relational database system owned by Sun Microsystems. For more information see <http://www.mysql.com>.

⁴ Austria, Germany, the UK or the Netherlands.

of time at certain conditions. Therefore a record in this table contains the identifiers of the retailer and the product, two dates representing the beginning and the end of the validity period of the offer and the price at which the product is offered. Each row in the product list of figure 2 constitutes an offer.

CLICK: This thesis uses clicks as a measure of demand. A record in this table constitutes a specific click from a user on a specific product-offer of a retailer. An entry in the *click*-table consists of the IP of the user that made the click, the product identifier, the retailer identifier, the date and time of the click and also information whether the click was a Last Click Through (**LCT**). This concept will be explained later in this thesis. Unfortunately clicks and offers are not linked directly with a unique identifier, therefore they have to be matched by using the product- and retailer identifier in combination with the time and date of the click and the validity period of the offer.

A detailed depiction of the conceptual schema of the database in Unified Modeling Language (**UML**)⁵ notation can be found in the appendix on page 115.

2.2.2 Variable Description

The dataset used for the estimation procedure is created from the database depicted in figure 3 and contains the variables described in table 1.

The logarithmized variables are used for the isoelastic demand specification whereas the normal versions of *clicks*, *lctw* and *price* are used for the linear demand specification. Apart from the variables representing the price of an offer and the number of clicks it received the dataset also contains data on the retailer who is providing the offer. These retailer specific variables are used to control for variations in *clicks* which do not stem from variations in the *price*.

Geizhals uses a range of criteria to assess the retailers who are offering products at the Geizhals platform. These criteria include the retailer's assortment of goods, the retailer's quality of service, the retailer's time to delivery, etc. In a first estimation approach all of these criteria have been included into the first stage estimation equation. However, if one takes a closer look at the ratings, one observes that these criteria are highly correlated amongst each other. The coefficient of correlation can reach a value of 0.9 and above. Then in turn the high correlation leads to estimation coefficients which are hardly statistically significant. In order to circumvent this problem one can merge the ratings of these criteria into a single average retailer rating.

For both, the average retailer rating and the shipping costs a corresponding dummy variable exists that indicates, whether the retailer rating or respectively the shipping costs are missing for the current offer. When estimating an equation with **OLS**, observations with missing values are left out. This results in a large loss of viable information. To prevent this loss of data, a dummy variable which indicates missing values has been included into the equation and in addition the value of *avg_full* and respectively *shippingcosts* has been set to the data

⁵ Unified Modeling Language, see <http://www.uml.org>.

VARIABLE	DESCRIPTION
product_id	Identifier of the product
retailer_desc	Label of the retailer
clicks	Number of clicks the offer received
clkcum	Aggregated and accumulated version of clicks
lnclicks	Logarithmized version of clicks
lctw	Number of weekly last clicks through the offer received
lctwcum	Aggregated and accumulated version of lctw
lnlctw	Logarithmized version of last clicks through
aggr	1 if the logarithmized variables are aggregated/accumulated versions, else 0
price	Price of the offer
lnprice	Logarithmized version of price
brick	1 if the retailer has got a brick and mortar presence, else 0
foreignc	1 if the retailer is located outside of Austria, else 0
avail	1 if the product currently is in stock, else 0
avg_full	Average retailer rating
avg_full_missing	1 if the retailer has not received any ratings, else 0
shippingcosts	Shipping costs for the current offer
smissing	1 if the shipping costs for the current offer are missing, else 0

Table 1: Variables of the first stage regression

set's average of those two variables. This approach enables OLS to use observations where either one or both of the variables `avg_full` and `shippingcosts` are missing and still indicate statistical differences for missing values.

LAST CLICK THROUGH (LCT) Besides the normal clicks variable table 1 also features the so called *last clicks through*. The concept of a *last click trough* or *LCT* is similar to a normal click, but it rather represents the last click of a set of logically linked clicks. This concept has emerged from the given fact that the clicks recorded by Geizhals do not represent actual purchases. Therefore elasticity estimates based on normal clicks might be biased, though a priori one cannot tell the direction of the bias. In order to prevent any potential problems stemming from this fact, one could try to convert the number of clicks into the number of real purchases. This could either be done by using a constant conversion ratio, e.g. one could say that 40% of all clicks result in a purchase, or one uses *last clicks through*.

As mentioned before a last click through is the final click of a set of logically linked clicks. To illustrate this concept assume that a person wants to buy a digital camera. A user looking for digital cameras at Geizhals will navigate through the menu items and arrive at the subsub-category *Digital Cameras*. A typical consumer will then click through

The LCT is the last click of a set of logically linked clicks.

different cameras to obtain data on the cameras and hence build a preference order. Although it would be unrealistic to assume that every click of the consumer results in a purchase, it seems reasonable to assume that the bulk of clicks results in the purchase of one of the clicked cameras. In the preceding example the last click in the bulk of clicks on different digital cameras would be marked as the *last click through*.

From a technical point of view the detection of the LCT is done via a clustering algorithm⁶. For each consumer (identified by the IP of the webclient) and subsubcategory the algorithm tries to identify clusters of clicks. This is done by putting all clicks of e.g. consumer A (C_A) and subsubcategory B (SSC_B) on a time axis. In the next step the algorithm starts at the beginning of the time axis and applies an outlier detection procedure to the clicks. The set of clicks interrupted by an outlier constitute a potential cluster. If the number of clicks of a potential cluster is larger than a parameter n and if distance between the last click of the potential cluster and the outlier is at least one week, then the set of clicks will be marked as a cluster and the last click of the cluster is the cluster's *last click through*.

Although this approach seems to be reasonable and coherent it has one significant drawback. Even though the chance that a set of clicks of a specific person on a set of digital cameras results in a purchase is rather high, one cannot tell which digital camera has finally been bought. The LCT is always the last click of a cluster/search. However, it is not guaranteed that the consumer really bought the last product of her or his search. Hence it could be that the LCT detection allocates the purchases to wrong products. Nevertheless the LCT-approach seems to be a more adequate solution than constant conversion ratios.

2.2.3 Data Overview

As mentioned before the data used in this thesis is an extract of the full Geizhals database. The data sample for this thesis has been reduced to the period of one week. This is done for econometrical reasons which will be explained in the next section.

The period of observation goes from May 25th 2007 to June 1st 2007 and features 50,497 products offered by 905 retailers, leading to a total number of 2,117,569 offers which have generated 6,967,438 clicks. The categories used by Geizhals to classify the products are *Hardware*, *Software*, *Games*, *Films*, *Household Appliance*, *Audio/HIFI*, *Video/Foto/TV*, *Telephone&Co* and *Sports*.

The size of each category measured by the number of products within the category is depicted in figure 4. One can see that *Hardware* is by far the largest category at Geizhals. For one part this seems to be historically rooted since in the beginning the primary focus of Geizhals had been on computer-hardware. On the other hand one could argue that the retailers and consumers of computer hardware are more open for a platform like Geizhals.

Figure 5 shows that *Hardware* is not only the largest category measured by the number of products but it is even more dominant if one uses the number of clicks as a measure for the size of a category. 53% of all clicks have occurred in the category *Hardware*.

A first look at the data shows that Hardware is by far the largest category.

⁶ A discussion of different clustering algorithms is beyond the scope of this thesis and therefore omitted. For an introduction into this topic see e.g. Han and Kamber (2000).

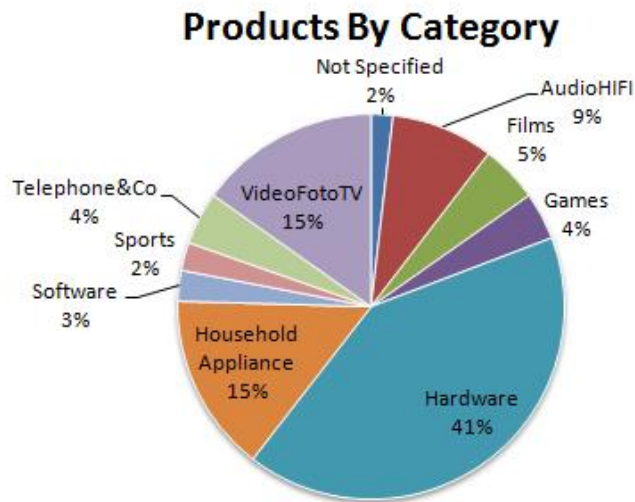


Figure 4: Number of products by category

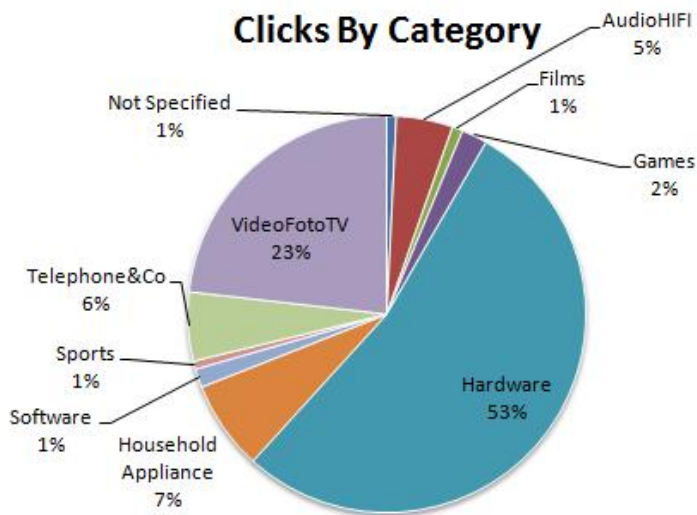


Figure 5: Number of clicks by category

2.3 CONSTRUCTION OF DEMAND CURVES

The next step towards the estimation of price elasticities is the construction of demand curves based on data from the tables of the Geizhals database. Therefore one has to create a single dataset which contains the merged information from the tables shown in figure 3. From a technical point of view this merging process seems to be simple and straightforward. Nevertheless there are some technical and econometric issues which one must not neglect. This section will cover the three most important issues, namely *simultaneity*, *normalization* and *aggregation/accumulation*.

2.3.1 Dealing With Simultaneity

This subsection deals with the econometric phenomenon of *simultaneity* which is described by Wooldridge (2003, p. 525) as a situation "when one or more of the explanatory variables is jointly determined with the dependent variable, typically through an equilibrium mechanism". A perfect example of such jointly determined variables would be the quantity (q) and the price (p) of a commodity which are determined by equating the quantity demanded with the quantity supplied.

However, if one takes a sample of prices and quantities from the market one will only observe equilibrium values. Therefore it is impossible to tell whether an equation $q = \beta p + \epsilon$, where ϵ constitutes the error term of a linear regression, is the demand or the supply equation. This problem is known as an *identification problem* and described in more detail by Koopmans (1949).

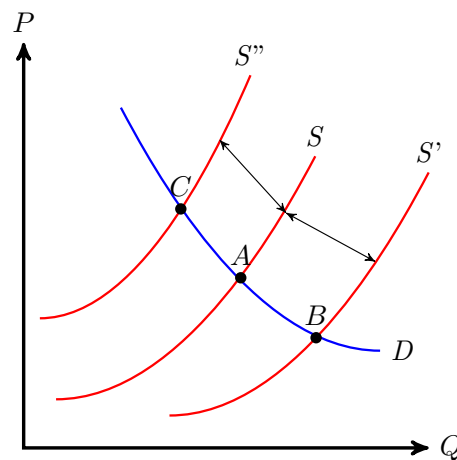


Figure 6: Tracing out the demand curve by shifting the supply curve

It is impossible to estimate a downward slope and an upward slope with a single linear regression equation involving only two variables.

The goal of this chapter is to obtain price elasticities by estimating equation 2.9 respectively 2.11 using OLS. As described in the last paragraph this leads to the problem that in general the equation cannot be identified as *the* demand curve or *the* supply curve. In order to trace out the demand curve one has to ensure that the supply function is exposed to random shifts. If this is the case then the observed equilibrium points will lie on the demand curve. This process of shifting the supply curve is depicted in figure 6. One can observe that if the demand curve is held fixed and one shifts the supply curve, then all of the resulting equilibrium points lie on the demand curve. As described by Wooldridge (2003, p. 533) this can be achieved by incorporating a shift parameter into the equation of the supply curve. This shift parameter "is the presence of an exogenous variable in the supply equation that allows us to estimate the demand equation".

Another possibility to estimate *the* demand curve would be if one could safely assume that the demand curve remains fixed for the whole dataset. In the case of Geizhals, where changes in prices can occur up to several times per day, it seems reasonable to assume that during a short period of observation (e.g. 1 week), factors influencing demand are constant and that variation in market outcome is therefore due to

variation in price setting only. Therefore the problem of identification should be circumvented or at least mitigated. For this reason the data sample used in this thesis is drawn from the period May 25th 2007 to June 1st 2007. Recall that this thesis constitutes a preliminary study and is part of a larger research project. Therefore choosing a short period of observation to circumvent simultaneity is an appropriate solution. However in subsequent studies one should deal with this problem in more detail, e.g. by using instrumental variables (IV) in the context of a two stage least squares (2SLS) estimation procedure.

2.3.1.1 Endogeneity - The Superordinate Concept Of Simultaneity

The problem of simultaneity is of great importance in the context of this chapter since it deals with the estimation of demand curves. Generally endogeneity arises if the error term u is correlated with one or more explanatory variables x_j (see Wooldridge (2003, p. 86)). A variable x_j that is correlated with u is called *endogenous explanatory variable*. If a regression model contains one or more endogenous explanatory variables all estimators will be biased and inconsistent.

Wooldridge (2003, p. 506) proposes the following test for endogeneity:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \quad (2.13)$$

In equation 2.13 y_1 denotes the explained variable, z_j the exogenous explanatory variables, u_1 the error term and y_2 a supposedly endogenous explanatory variable. Furthermore assume that there are two further exogenous explanatory variables z_3 and z_4 . To test whether y_2 really is an endogenous explanatory variable one has to perform the following two steps:

1. Regress y_2 on *all* exogenous explanatory variables and obtain the residuals from this regression (\hat{v}_2). This can be achieved by estimating equation 2.14.

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \quad (2.14)$$

2. Add \hat{v}_2 to the structural equation 2.13, carry out the regression and use a t test to test $H_0 : \delta_0 = 0$. If the coefficient on \hat{v}_2 is statistically significant then y_2 is indeed an endogenous explanatory variable⁷. The regression to be estimated in this step is given in equation 2.15.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_0 \hat{v}_2 + u_1 \quad (2.15)$$

In the case of endogeneity one should estimate equation 2.13 by using IV and the method of 2SLS. Unfortunately this is beyond the scope of this thesis. Therefore it is essential that subsequent studies use appropriate IV to test and respectively control for endogeneity.

2.3.2 Normalization Of Clicks

Although the period of observation has been set to one week one cannot expect that all offers start and end during the period of observation.

⁷ It might be advisable to use a heteroskedasticity-robust test statistic.

Given that t_{obs_b} and t_{obs_e} are the timestamps⁸ of the beginning and the end of the period of observation⁹ and t_{off_b} and t_{off_e} are the timestamps of the beginning and the end of an offer, following cases can occur¹⁰:

CASE 1: $(t_{off_b} < t_{obs_b}) \wedge (t_{obs_b} < t_{off_e} < t_{obs_e})$ In this case the offer starts before the period of observation and ends during the period of observation.

CASE 2: $(t_{obs_b} < t_{off_b} < t_{obs_e}) \wedge (t_{obs_e} < t_{off_e})$ In this case the offer starts during the period of observation and ends after the period of observation.

CASE 3: $(t_{off_b} < t_{obs_b}) \wedge (t_{obs_e} < t_{off_e})$ In this case the offer starts before the period of observation and ends after the period of observation.

CASE 4: $(t_{obs_b} < t_{off_b}) \wedge (t_{off_e} < t_{obs_e})$ In this case the offer starts and also ends during the period of observation.

One should realize that it is very likely that offers and the period of observation overlap. Therefore it could be possible that the clicks would be calculated for an offer which has been valid for several weeks. In such a setup one cannot assure that the demand curve is held fixed, meaning that the estimation results will suffer from simultaneity bias. For this reason one has to prefilter the clicks and use only those, which belong to the period of observation (independent of the offer to which a click may belong to).

Furthermore it is possible that a high-price-offer is valid for six days of the week under observation whereas a low-price-offer is only valid for one day. This suggests that a the high-price-offer will receive more clicks than the low-price-offer, which would distort the construction of valid demand curves. To avoid this problem the number of clicks on a product-offer have to be *normalized*.

In order to compute the clicks for an offer which can be used to estimate the price elasticity for a product one has to execute the following steps:

STEP 1: Count all clicks that belong to the offer ($=clicks_{total}$).

STEP 2: Subtract the clicks which timestamp does not belong to the period of observation do receive the net number clicks ($=clicks_{net}$)

STEP 3: Compute the length of the offer ($=len_{total}$).

STEP 4: If $t_{off_b} < t_{obs_b}$ compute $cut_b = t_{obs_b} - t_{off_b}$.

STEP 5: If $t_{obs_e} < t_{off_e}$ compute $cut_e = t_{off_e} - t_{obs_e}$.

STEP 6: Use the variables from steps 3-5 to compute the net length of the offer $len_{net} = len_{total} - cut_b - cut_e$.

⁸ Geizhals and hence this thesis uses Unix timestamps to denote date and time variables. A Unix timestamp is the number of seconds between a particular date and January 1st 1970.

⁹ t_{obs_b} would be set to 1251900239, indicating the May 25th 2007 and t_{obs_e} would be set to 1251900294, which is June 1st 2007.

¹⁰ The cases $t_{off_e} < t_{obs_b}$ and $t_{off_b} > t_{obs_e}$ have been omitted because these cases do not belong to the period of observation at all.

STEP 7: Use the net length of the offer to normalize the net number of clicks by computing $\text{clicks}_{\text{normalized}} = (\text{clicks}_{\text{net}} / \text{len}_{\text{net}}) * (60 * 60 * 24)^{11}$

As a result of steps 1 through 7 one receives the average number of clicks on an offer during the period of observation.

2.3.3 Aggregation And Accumulation Of Clicks

This subsection is going to present an alternative view of a demand curve. So far offers with their price and their clicks received have been used directly in order to generate a demand curve. But this is not the only approach. For an alternative way to estimate the demand curves and the price elasticities, the number of clicks will be accumulated going from the highest to the lowest price of a product, i.e. if one observes 5 clicks at a price of €100, it is assumed that a price of €80 would also have received those clicks, even if no clicks were recorded for that price in the period of observation.

The reasoning behind this idea goes as follows: The clicks on a product at the geizhals.at website are similar to the purchase of a certain good in a shop. In order to derive the demand curve of such a good, one could take different shops, all of them offering that certain good. Those shops are completely identical (e.g. in terms of service, skills of employees etc.), the only difference is that they offer the product at different prices. Each shop is frequented by 100 identical consumers, whereas each consumer goes to exactly one shop, i.e. each shop has its own, unique pack of 100 identical consumers. Each consumer decides whether he/she wants to buy that good in the shop he/she is going to and how many units of that good he/she wants to purchase.

Price	Quantity Sold	Accumulated Quantity
40	10	10
35	20	30
30	30	60
25	40	100
20	50	150
15	60	210
10	70	280
5	80	360

Table 2: Aggregation and accumulation of clicks

The column 'Accumulated Quantity' in table 2 shows an accumulation of the units sold along the price. One has to accumulate the sold units, because as mentioned before, both, the shops and the consumers are identical. For this reason the 10 units which have been sold at a price of €40 would also have been sold at a price of €35, therefore those 10 sold units should also be included in the units sold at a price of €35. The same reasoning is applicable to the clicks at the Geizhals website. This thesis uses the term *normal-shop-setup* to refer to the setup described above.

¹¹ $60 * 60 * 24$ are the number of seconds per day.

Nevertheless there are reasons why the assumptions of the previous paragraphs may not hold. In the setup from the previous paragraph it is logical to assume that a customer who bought the good at a price of e.g. €40, would *ceteris paribus* also have bought the good at a price of €30. But there are differences between the setup described above and the Geizhals website. First of all, in the "normal-shop-setup" all offers exist simultaneously and secondly the "normal-shop-setup" features homogeneous shops. All of these facts are not true in the context of a price comparison website. Each user can easily click on every single offer. If a user clicked on an offer at €40, the €30 offer would have basically just been a click away. However this gets complicated by the second fact which states that in the context of a price comparison websites offers might not exist simultaneously. Thus if one observes a price-click relation from Geizhals, one does in fact look at a conglomerate of offers which have existed during specific, maybe non-congruent periods of time. Therefore it might be that a user clicked on a €40 offer because a €30 offer did not exist at that time. Heterogeneous shops are a further explanation why a user might prefer a more expensive offer to a cheaper offer, whereas *heterogeneous* refers to things like retailer quality, shipping costs etc.

Since the period of observation for the whole thesis is only one week it is very unlikely that the non-congruent offers can be stated as the general explanation. But even in the case of heterogeneous shops where one could theoretically account for things like different retailer quality or shipping costs, when accumulating and aggregating the clicks all data on retailers and shipping costs get lost, since one cannot reliably compute retailer ratings and shipping costs for aggregated and accumulated data.

All in all one has to conclude that it is ambiguous whether aggregated and accumulated clicks are suited for the estimation of price elasticities in the case of Geizhals. Nevertheless there is not enough evidence to reject this kind of demand construction. For this reason this thesis will incorporate aggregated and accumulated clicks as an alternative approach to normal clicks.

2.4 ELASTICITY DESCRIPTION AND REGRESSION SETUP

The previous sections have already indicated that one cannot estimate *the* price elasticity of demand but that there are rather several approaches and ideas for different elasticity specifications. The section [Theoretical Framework](#) introduced the difference between an isoelastic and a linear demand specification. The paragraph on page 9 explained the necessity of *last clicks through* and finally the section on page 15 brought up the idea of aggregated and accumulated clicks.

The combination of the different ideas and approaches yields a large variety of different elasticities. The purpose of this section is to introduce some further elasticities and to describe which types of elasticities will be estimated during the first stage regressions. This thesis is a result of a cyclic process with several evolutionary iterations. Therefore the following list describes the elasticities which have been estimated during at least one of the iterations. However, this does not imply that the full set of elasticities will be used for any further work. The letters and abbreviations in the name of the listed elasticities display the type of the elasticity are explained in table 3.

ABBR.	DESCRIPTION
c	Isoelastic demand with a constant elasticity
l	Elasticity from a linear demand specification
cv	The regression includes variables controlling for retailer specific data
aggr	The elasticity is based on aggregated/accumulated clicks
inv	The elasticity is estimated by regressing price on clicks rather than clicks on price
choke	The elasticity is based on a dataset without offers above the choke price
lctw	The elasticity is based on <i>last clicks through</i>

Table 3: Explanation of abbreviations used in elasticity variable notations

Each entry consists of a short description of the elasticity and of a schematic regression equation.

C_ELAST: Elasticity from an isoelastic demand (regression clicks on price).

This approach is a simple log-log regression of clicks on price. In such a regression the estimated parameter on price constitutes the elasticity. This approach only uses price as the explanatory variable; there are no further explanatory or control variables. The demand equation belonging to this regression is of the type $\text{clicks} = e^{\beta_0}/\text{price}^{\beta_1}$.

$$\ln(\text{clicks}) = \beta_0 + \beta_1 \ln(\text{price}) + \epsilon \quad (2.16)$$

C_ELAST_INV: Elasticity from an isoelastic demand (regression price on clicks).

This elasticity is estimated by regressing $\ln(\text{price})$ on $\ln(\text{clicks})$ therefore this elasticity should be the exact inverse of c_{elast} . The main idea of this approach stems from the fact that in case of imperfect data OLS does not yield the same results when estimating a demand curve or its inverse. With perfect data it should be the case that $c_{\text{elast}} = c_{\text{elast_inv}}^{-1}$. Unfortunately this is not true in the case of the Geizhals data. For this reason some of the estimation results will also be graphed in order to get an idea whether one can say that one of the two elasticities will on average be better than the other.

$$\ln(\text{price}) = \beta_0 + \beta_1 \ln(\text{clicks}) + \epsilon \quad (2.17)$$

C_ELAST_CHOKE: Demand with constant elasticity and exclusion of offers above the choke price.

This elasticity is computed in the same way as c_{elast} , i.e. by regressing $\ln(\text{clicks})$ on $\ln(\text{price})$. The only difference is the dataset which is used for the regression. Observed offers above the choke price (=lowest price which yields a demand of zero¹²)

¹² In a first attempt the choke price has also been excluded from the dataset so that the offer with the highest price exhibits a number of clicks greater than zero. The drawback of this approach is that estimated demand curves are very susceptible to poor data, because the

tend to generate a downward bias of the estimated elasticity in terms of absolute values, i.e. the estimated demand will be less elastic than the true demand. For this reason it seems reasonable to exclude offers above the choke price.

$$\ln(\text{clicks}) = \beta_0 + \beta_1 \ln(\text{price}) + \epsilon \quad (2.18)$$

AGGR_C_ELAST: Demand with constant elasticity with aggregated and accumulated clicks.

This elasticity is computed by regressing $\ln(\text{clicks})$ on $\ln(\text{price})$, but in this case the number of clicks have been aggregated across the offers and accumulated along the price axis. This ensures a strictly negative-sloped demand curve and therefore also a negative elasticity. Because of the accumulation the estimated demand of this approach should on average be more elastic than *c_elast*.

$$\ln(\text{aggr_clicks}) = \beta_0 + \beta_1 \ln(\text{price}) + \epsilon \quad (2.19)$$

AGGR_C_CHOKE: Demand with constant elasticity with aggregated and accumulated clicks and exclusion of offers above the choke price.

Similar to *c_elast_choke* this elasticity is the same as *aggr_c_elast*, except that the offers above the choke price are excluded from the regression.

$$\ln(\text{aggr_clicks}) = \beta_0 + \beta_1 \ln(\text{price}) + \epsilon \quad (2.20)$$

CV_C_ELAST: Demand with constant elasticity including retailer control variables (regressing clicks on price).

This elasticity is computed by regressing $\ln(\text{clicks})$ on $\ln(\text{price})$ and a set of variables which control for retailer properties. The constant elasticity demand equation for this regression is $\text{clicks} = e^{\beta_0 + \beta_{cv} CV}$, where *CV* denotes a vector of *control variables*.

$$\begin{aligned} \ln(\text{clicks}) = & \beta_0 + \beta_1 \ln(\text{price}) + \beta_2 \text{brick} + \\ & \beta_3 \text{foreignc} + \beta_4 \text{avail} + \beta_5 \text{avg_full} + \\ & \beta_6 \text{avg_full_missing} + \beta_7 \text{shippingcosts} + \\ & \beta_8 \text{scmissing} + \epsilon \end{aligned} \quad (2.21)$$

L_ELAST: Linear demand elasticity estimated at mean values.

This approach estimates a linear demand curve by regressing price on clicks (in level-form). Since the elasticity is not constant across a linear demand curve, one has to pick a specific point of the demand curve in order to get an exact estimate of the elasticity. Therefore the elasticity is computed at the mean price for the specific product.

$$\text{clicks} = \beta_0 + \beta_1 \text{price} + \epsilon \quad (2.22)$$

offer at the choke price more or less constitutes an anchor to enable a negative slope of the demand curve. Removing all offers above the choke price leads on average to a larger amount of positive elasticities.

AGGR_L_ELAST: Linear demand elasticity estimated at mean values on the basis of aggregated/accumulated clicks.

This approach estimates a linear demand curve based on aggregated/accumulated clicks. The price and clicks variables are used in level-form. As it is in the case of l_elast the elasticity will be estimated at mean values. Since the aggregation and accumulation flattens the demand curve the demand should be more elastic in comparison to l_elast .

$$\text{aggr_clicks} = \beta_0 + \beta_1 \text{price} + \epsilon \quad (2.23)$$

CV_L_ELAST: Linear demand elasticity estimated at mean values including retailer control variables.

This elasticity is like l_elast but the regression includes the same set of control variables as cv_c_elast .

$$\begin{aligned} \text{clicks} = & \beta_0 + \beta_1 \text{price} + \beta_2 \text{brick} + \\ & \beta_3 \text{foreignc} + \beta_4 \text{avail} + \beta_5 \text{avg_full} + \\ & \beta_6 \text{avg_full_missing} + \beta_7 \text{shippingcosts} + \\ & \beta_8 \text{scmissing} + \epsilon \end{aligned} \quad (2.24)$$

CV_C_LCTW_ELAST: Isoelastic demand equation including retailer control variables based on [LCT](#).

Similar to cv_c_elast this elasticity is computed by regressing $\ln(\text{lctw})$ on $\ln(\text{price})$ and a set of variables which control for retailer properties.

$$\begin{aligned} \ln(\text{lctw}) = & \beta_0 + \beta_1 \ln(\text{price}) + \beta_2 \text{brick} + \\ & \beta_3 \text{foreignc} + \beta_4 \text{avail} + \beta_5 \text{avg_full} + \\ & \beta_6 \text{avg_full_missing} + \beta_7 \text{shippingcosts} + \\ & \beta_8 \text{scmissing} + \epsilon \end{aligned} \quad (2.25)$$

CV_L_LCTW_ELAST: Linear demand equation including retailer control variables based on [LCT](#).

This approach is the linear demand version of $cv_c_lctw_elast$. It does include the same control variables as the other cv variables and like the other linear elasticities this elasticity is also estimated at mean values.

$$\begin{aligned} \text{lctw} = & \beta_0 + \beta_1 \text{price} + \beta_2 \text{brick} + \\ & \beta_3 \text{foreignc} + \beta_4 \text{avail} + \beta_5 \text{avg_full} + \\ & \beta_6 \text{avg_full_missing} + \beta_7 \text{shippingcosts} + \\ & \beta_8 \text{scmissing} + \epsilon \end{aligned} \quad (2.26)$$

AGGR_LCTW_C: Isoelastic demand equation based on aggregated and accumulated [LCT](#).

$$\ln(\text{aggr_lctw}) = \beta_0 + \beta_1 \ln(\text{price}) + \epsilon \quad (2.27)$$

AGGR_LCTW_L: Linear demand equation based on aggregated and accumulated LCT.

$$\text{aggr_lctw} = \beta_0 + \beta_1 \text{price} + \epsilon \quad (2.28)$$

2.5 CLEARINGHOUSE MODELS - AN ALTERNATIVE ESTIMATION APPROACH

An alternative approach for estimating demand curves for online price comparison websites was introduced by Baye et. al. (2004). They use so called *clearinghouse models*, which "view a price comparison site as an information clearinghouse where shoppers and loyals obtain price and product information to make online purchases". The main idea of such a model is that demand consists of two parts:

CLICK GENERATING PROCESS: This process generates the number of referrals from the price comparison website to the retailer's website. It is influenced by a number of characteristics like market structure, number of competitors, firm characteristics, etc. These characteristics are subsumed by Baye et. al. (2004, p. 10) under X .

CLICK CONVERSION PROCESS: A process which converts clicks into actual purchases. The probability of purchase is influenced by aforementioned set of characteristics X and in addition by a set of retailer specific information Z_i , e.g. whether the firm offers guarantee, the design and layout of the retailer's website, etc.

The expected demand D_i for a specific product sold by firm i can therefore be stated as:

$$E(D_i|X, Z_i) = \Pr(\text{sale}_i|X, Z_i)E(Q_i|X) \quad (2.29)$$

Whereas Q_i denotes the number of clicks received, which is a random variable drawn from a not specified random distribution. For this reason, one has to use the expected value of Q_i , which in turn, is computed using the *Lebesgue Integral*. The first part of equation 2.29 denotes the probability that a click will result in a purchase given the characteristics X and the retailer specific information Z_i . The last part denotes the expected number of clicks (Q_i) a product from retailer i will receive given the set of characteristics X . Multiplying the number of clicks a product receives with the probability that a click will result in an actual purchase yields the expected actual demand for the product of retailer i .

In order to retrieve the price elasticity of demand in that framework Baye et. al. (2004, p. 11) rewrite X as $X = (x_i, X_1)$ where " X_1 represents all components of X other than x_i " (recall that X contains factors like firm characteristics, the number of competitors, etc.). Furthermore they assume that $\Pr(\text{sale}_i|(x_i, X_1), Z_i) = \Pr(\text{sale}_i|(x_i', X_1), Z_i)$ for all x_i, x_i' . This assumption states that the probability of a sale is independent of the value of x . Therefore if one examines the impact of a variation in x on the demand for a specific product, the probability of sale is held constant.

Taking this assumption as granted one can continue by logarithmizing equation 2.29 and differentiating this equation with respect to x_i which yields:

$$\frac{\partial \ln E(D_i|X, Z_i)}{\partial \ln x_i} = \frac{\partial \ln \Pr(\text{sale}_i|X, Z_i)}{\partial \ln x_i} + \frac{\partial \ln E(Q_i|X)}{\partial \ln x_i} \quad (2.30)$$

Since the probability of a sale is independent of the value of x one can state $\frac{\partial \ln \Pr(\text{sale}_i|X, Z_i)}{\partial \ln x_i} = 0$ and above equation boils down to

$$\frac{\partial \ln E(D_i|X, Z_i)}{\partial \ln x_i} = \frac{\partial \ln E(Q_i|X)}{\partial \ln x_i} \quad (2.31)$$

Bearing in mind equation 2.8, one can notice that the right-hand-side of above equation denotes the elasticity of x_i to $E(\bullet)$. This equation is a central part in the work of Baye et. al. (2004, p. 11), since it shows that the price elasticity of demand can be estimated by simply using the number of clicks as a measure of demand. However, one must notice that equation 2.31 only holds in the case of constant conversion ratios. The novelty in the approach of Baye et. al. (2004, p. 11) is the insight that for the estimation of elasticities one does not strictly have to know the size of the conversion ratio, it suffices if one can safely assume that the ratio is constant. Therefore one can conclude that the approach of Baye et. al. (2004, p. 11) does not differ from the approach used in this thesis as presented in section 2.4. Furthermore one has to keep in mind that because of the normalization process the approach in this thesis offers the advantage that elasticities can be estimated using simple OLS, whereas the approach of Baye et. al. (2004, p. 11) requires the method of *pseudo-maximum likelihood*. This method is also known as quasi-maximum likelihood (QML) and yields an estimator which is according to Wooldridge (2002, p. 646) "*fully robust to distributional misspecification*". The Quasi-Maximum Likelihood Estimator (QMLE) is a Poisson estimator and therefore adequate for discrete variables as it is the case for the number of clicks received by a product¹³.

2.6 ESTIMATION RESULTS

This is the final section of this chapter and it presents the results of the first stage regressions. In order to use only valid elasticities the first part of this chapter deals with so called zero-elasticities. Zero-elasticities are elasticities which feature a value of 0. There are several reasons why an elasticities can adapt a value of 0, but all of them are of technical nature. Since zero-elasticities only arise because of technical reasons they cannot be treated as elasticities per se and will therefore be removed from the dataset and from further analysis. After this fundamental cleansing process the subsequent part of this chapter reports a general overview of the estimation results for the full dataset containing the complete range of products¹⁴. In a next step the thesis uses a small subsample of products in order to detect data problems and verify the data quality. The insights of this detailed analysis are

¹³ Note that the approach in this thesis does not rely on a Poisson estimator because due to the normalization process the various clicks variables are not discrete.

¹⁴ The complete range of products consists of all products which contain the minimum number of observations in order to estimate the elasticity equation that includes the retailer specific control variables.

then used to generate a reduced dataset of higher quality and to correct for problems of the full dataset.

2.6.1 Detection And Removal Of Zero-Elasticities

The first stage regressions yield elasticities with a value of 0. On economical grounds these elasticities are invalid and will therefore be removed from the dataset. The goal of this subsection is to give possible explanations why the first stage estimation yields elasticities with a value of 0. The analysis in this section is based on the elasticities coming from an isoelastic demand specification constructed with normal, non-aggregated or accumulated clicks (i.e. the variable c_elast)¹⁵. The analysis is split into several subsequent steps, where each step tries to detect a specific problem which may result into an elasticity with a value of 0. A further property of this approach is that the subsequent step is always finer grained than its predecessor. The following example should make clear the idea of this approach: The approach can be conducted in a step-by-step manner. Each step takes an input-dataset and filters out the records of this level and then passes the remaining records to the next level.

STEP 1: Gather all products with an elasticity of 0 (= N)

STEP 2: Using set N: gather all products which did not receive any clicks apart from those on missing offers (m)

STEP 3: Using set N – m: gather all products with only one non-zero-click offer (n)

STEP 4: Using set N – m – n: gather...

2.6.1.1 Missing Offers

The first possible reason for an elasticity of zero is a problem in the *offer* table from the Geizhals database. The *demand* dataset, which is used for the first stage regressions, is based on a list of product identifiers which have been generated by selecting the distinct product identifiers of all clicks from the *click* table which occurred during the period of observation. i.e. the list consists of products, which have received at least one click during the period of observation.

Missing offers are the main culprit for a zero-elasticity.

In a further step of the data generation process each click has to be matched to an *offer*. The source of this problem is that the *offer* table does not contain all corresponding offers (e.g. click happened on a still open webpage after the offer has changed). Therefore these offers will not find their way into the final *demand* dataset. But if such a missing offer has generated all clicks for a certain product during the period of observation, the final dataset will still contain that product, but all offers will be marked by a click-number of 0, i.e. the clicks variable will be constant. If the dependent variable is a constant, the explanatory variables will not have any explanatory power and thus yield coefficients of value 0.

In order to check for this problem, one can run a query, which computes the number of non-zero-click offers for each zero-elasticity product. If e.g. the elasticity of product 123456 is zero, one can gather all offers for 123456 which belong to the period of observation and then

¹⁵ The results of this analysis however are also applicable to the other elasticity variables.

count the records where the number of clicks is different from 0. If this counting procedure yields 0, then no offer of product 123456 in the *offer* table has received any clicks at all.

Applied to a relational database a SQL query to check for the problem of missing offers could look like the code presented in listing 1¹⁶.

```

1 SELECT *
2 FROM elasticities e
3 WHERE e.elasticity = 0
4 AND (SELECT count(*)
5      FROM demand
6      WHERE product_id = e.product_id
7      AND clicks != 0) = 0

```

Listing 1: Detection of missing offers

2.6.1.2 One Non-Zero-Click-Offer

This category constitutes a more extreme case of too little variation in clicks. A zero-elasticity which belongs to this category only contains a single offer where the number of clicks received is different from zero. The first stage estimation has therefore been carried out using just two different numbers of clicks, namely $\ln(0.01)$ and \times the non-zero-clicks record.

Like mentioned before, the detection works like a filter where each step is finer or stricter than its predecessor. For this reason the first action in this stage is to create a reduced set of elasticities, namely those which do not belong to the category of *missing offers*¹⁷. When creating the reduced set one can easily count the number of non-zero-click-offers. In order to check whether a zero-elasticity belongs to the category *One Non-Zero-Click-Offer* one has to simply test whether the number of non-zero-click-offers for a product is 1.

2.6.1.3 Degrees of Freedom Problem

This problem occurs if there are too few observations in the sample in relation to the number of parameters which should be estimated. If a regression consists of n explanatory variables then the dataset has to consist of at least $n + 1$ observations¹⁸. In order to ensure this, the *demand* relation contains a variable expressing the number of observations for the specific product. The drawback of this counter-variable is that it simply counts the number of offers and not the number of distinct offers (where distinct applies to the values of the variables used in the first stage regression). Therefore it is possible that two (or more) observations contain exactly the same values, making one (or more) of them redundant. In such a case the regression yields zero-coefficients.

DEGREES OF FREEDOM AND THE RMSE The elasticities estimated in the first stage are the main explaining variables in the second and the third chapter of this thesis. However, one has to take under consideration that the quality of the data is not equal for each product

¹⁶ This code is solely included for illustrative purposes. Although the statement in the listing is syntactically correct, it cannot be directly applied to the Geizhals database.

¹⁷ The creation of the reduced set works with temporary tables and is, because of its technical details, beyond the scope of this thesis and therefore omitted.

¹⁸ n the number of explanatory variables plus one further observation for the intercept.

and hence the elasticities will differ in their statistical expressiveness. For this reason one might be tempted to use the Root Mean Squared Error (RMSE) as a quality indicator and therefore as a weighting measure for the second stage regressions. There are two seemingly odd cases concerning the RMSE when looking at the results of the first stage regressions. The first case is when there is a positive elasticity but a RMSE of zero. The second case is coined by a non-zero RMSE in combination with a zero-elasticity.

CASE 1: ELAST > 0, RMSE = 0 This case occurs if there is a problem with the Degrees of Freedom (DF). Since the RMSE is the square root of s^2 , the *Variance of the Error Terms*, it will be corrected using the DF. This is shown in the equation below (cf. Davidson and MacKinnon (2004)):

$$s^2 \equiv \frac{1}{n-k} \sum_{t=1}^n \hat{u}_t^2 \quad (2.32)$$

Since n denotes the number of observations used for the regression and k denotes the number of the explaining variables plus 1 (+1 because of the intercept), the term $n - k$ may be zero. This is the case if " $n = 9$ ", because we use $8 + 1$ explaining variables. In such a case the calculation of the "Variance of the Error Terms" yield an division-by-zero error. Nevertheless Stata does not really raise an error, instead in just states a RMSE of 0. Such a value for the RMSE would render it useless as a weight in the second stage regression. Therefore the elasticities with a RMSE of zero should be omitted when carrying out the second stage regression. This raises the question why there are elasticities produced by regressions using 9 observations, which still display a non-zero RMSE. The rationale for this (unexpectedly not so rare) case lies in the small sample size. The regression includes two dummy variables which indicate whether some other variables are missing and have been set to the average value of that specific variable. Because of the small sample size, it is not uncommon, that none of the variables contain missing data. If there is no missing data, the dummy variables indicating missing data will be strictly a constant with the value of zero. Given that a constant has no explanatory power in a regression, the dummy variables are useless and Stata drops them when carrying out the regression. A dropped variable increases the DF by one, which in turn resolves the division-by-zero error.

CASE 2: ELAST = 0, RMSE != 0 Another oddity which can be found in the results from stage one is the case where there is a zero-elasticity with a non-zero RMSE. The source of this problem is a very small coefficient on the price variable and respectively on its log-version. The elasticity values are stored with a floating point precision of 10, i.e. the smallest number which can be expressed using this format is 0.0000000001. If the estimated elasticity is smaller than this value, it will be set to 0.0000000000. Zero-elasticities stemming from the floating point cut-off will be termed as almost-zero elasticities. Nevertheless when computing the RMSE Stata uses the full floating point precision yielding a

RMSE different from 0 (which can be in fact quite large if the regression doesn't fit very well).

2.6.1.4 Click Variation

A further reason why an elasticity might be zero is that there is just too little variation in the data. The most obvious form of this problem is an almost constant number of clicks. Given that "clicks" is the left-hand-side (LHS), it is not surprising that an almost constant number of clicks yields estimated coefficients with a value of zero or *almost-zero*.

To get a grasp of the amount of variation in clicks one can introduce a new variable called *clickvar* which is defined as follows:

$$\text{clickvar} = \frac{\text{distinctclicks} - 1}{\text{numobs}} \quad (2.33)$$

where *numobs* denotes the number of offers (observations) for a specific product and *distinctclicks* denotes the number of distinct values of the variable *clicks* for a specific product. e.g. if there are 4 offers for product A (O1, ..., O4) and given that O1 received 0 clicks, O2 2 clicks, O3 3 clicks and O4 2 clicks, then *distinctclicks* will be 3. However, this measure does not give any hints about the quantitative size of the variation in the number of clicks. In addition the emphasis of the first stage regression is on offers which have received any clicks. Therefore one is rather interested in the *distinctclicks* which are non-zero. Since all remaining zero-elasticity products or almost-zero elasticity products display at least one non-zero-click offer, we use *distinctclicks* - 1 to compute *clickvar*. If this has been done, one has to set a threshold which defines a numeric representation of the term "enough variation in clicks". Econometric literature does not give any instructions on this topic and therefore the threshold is set to 0.3.

2.6.1.5 Price Variation

The price variation problem is similar to the problem described in the previous section. But here one does not look at almost-constant-clicks, but rather at the problem of too little variation in prices. If one counts the distinct prices for each product one can observe that on average this number is larger than the number of distinct clicks. This gives the impression that there should be enough variation in prices in order to get proper estimates. Nevertheless a closer look at the prices reveals that although the prices vary across the offers, the quantitative value of the variation is rather small in most cases.

Given the fact that the dataset covers a wide range of products from different price categories, one cannot simply use the standard deviation of the price variable to check whether there is enough variation. A standard deviation of 10 for a product which costs €1000 on average cannot be compared to a standard deviation of 10 for a product whose mean price is €20. Therefore the standard deviation has to be normalized in some way. This can be achieved by using the *Coefficient of Variation* (c_v ¹⁹) which is defined as follows:

$$c_v = \frac{\sigma_p}{\mu_p} \quad (2.34)$$

¹⁹ Unfortunately the MySQL version used to host the Geizhals DB cannot compute the standard deviation. For this reason one has to use Stata to load the remaining zero-click products, which have passed all of the previous stages and inspect them manually.

where σ_p denotes the standard deviation of the price and μ_p denotes the mean price.

2.6.1.6 Further Reasons

Coming to this point there should only be a handful of records left, which can be inspected manually. This group covers the records which did not fit into the categories from the previous steps. The most prominent reason why an observation belongs to this group is that there is not enough variation in the data, but enough to be excluded from the other groups. This category would be even smaller if one chooses to loosen up the thresholds set in the previous stages. Another reason why a product can be found in this category is that there are some offers with a price of zero, which is obviously a case of erroneous data.

2.6.2 General Overview Of The Full Dataset Results

As mentioned before the first set of elasticities has been estimated with the full dataset containing roughly 40,000 products. The following elasticities have been estimated during this stage:

- `c_elast`: Elasticity from an isoelastic demand (regression clicks on price).
- `c_elast_inv`: Elasticity from an isoelastic demand (regression price on clicks).
- `c_elast_choke`: Demand with constant elasticity and exclusion of offers above the choke price.
- `l_elast`: Linear demand elasticity estimated at mean values.
- `cv_c_elast`: Isoelastic demand specification including control variables for retailer data.
- `cv_l_elast`: Linear demand specification including control variables for retailer data.
- `aggr_c_elast`: Isoelastic demand specification based on aggregated/accumulated clicks.
- `aggr_c_elast_choke`: Demand with constant elasticity with aggregated and accumulated clicks and exclusion of offers above the choke price.
- `cv_c_lctw_elast`: Isoelastic demand specification including control variables based on [LCT](#).
- `cv_l_lctw_elast`: Linear demand specification including control variables based on [LCT](#).
- `aggr_c_lctw_elast`: Isoelastic demand specification based on aggregated/accumulated [LCT](#).
- `aggr_l_lctw_elast`: Linear demand specification based on aggregated/accumulated [LCT](#).

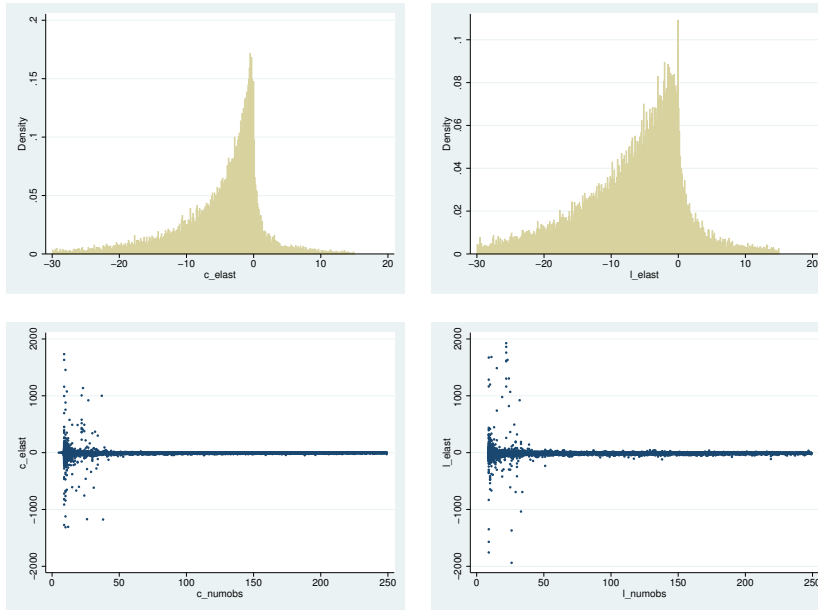


Figure 7: Graphical overview on the full dataset estimation results

A detailed description of the results of this estimation would not uncover any specific phenomena. Hence this section will only give a short overview on the resulting elasticities. An in-depth analysis is given in the section on page 30.

The upper row of the result overview in figure 7 shows the histograms for the isoelastic elasticities and the elasticities coming from a linear demand specification whereas the lower row shows the relation of the elasticities compared to the number of observations which have been used to estimate the respective elasticity. From this graphical representation one can draw the following conclusions:

SIGN OF ELASTICITIES: Both c_elast and l_elast are skewed to the right²⁰ with a negative mean and hence negative median. Simply put, the majority of the estimated elasticities is negative. Since a negative elasticity can only come from a negatively sloped demand curve the results are exactly what one would expect because as theory suggests, demand curves for normal goods will not have a positive slope.

ELASTICITY VALUES: This can be seen in the upper row of figure 7. The majority of the estimated elasticities feature a value between 0 and -10 , which seem to be reasonable values for elasticities.

QUALITY OF ELASTICITIES: The scatter plots of the respective type of elasticity versus the number of observations which have been used for estimation indicate that there is a large variation in resulting elasticities for low number of observations. This can also be checked by splitting the estimated elasticities in two groups, namely a group of elasticities which has been estimated by using 30 or less observations and a group which has been estimated by using more than 30 observations and then by computing and comparing the respective standard deviations. In the case of c_elast

²⁰ The absolute value of the mean is larger than the absolute median value.

the standard deviation of the first group is 5441.26 whereas the standard deviation of the second group is 5.35. This suggests that elasticities which have been estimated with viewer observations might be less reliable.

As an extension to the graphical overview one can also take a look at the summary statistics of the elasticities which are given in table 4.

Variable	Mean	Std. Dev.	Min.	Max.
c_elast	-2.079	4070.712	-277148.6	652898.2
l_elast	4.001	2853.878	-267785.1	338689.3

Table 4: Summary statistics of two elasticities of the full dataset

The summary statistics confirm the findings from the graphical overview and yield the following further results:

OUTLIERS: Looking at the *min* and *max* values of the respective elasticities one can see that they are enormous. In order to verify the validity of such extreme values this thesis will introduce the so called Grubbs' Test. This test is a method for the detection of outliers.

MEAN ELASTICITY: The mean of *c_elast* is clearly negative, however this is not true for *l_elast*. The mean of *l_elast* is only slightly positive. This might stem from the fact the the lowest value of *l_elast* is relatively small to its largest value ($-39,624.12$ compared to $+197,887.03$). If one subtracts the potential outliers the mean value of both elasticities becomes negative, which is exactly what theory suggests.

As a concluding remark of this section one can say that on first sight the estimated elasticities seem to be reasonable. Nevertheless the summary statistics show that there might be exceptional extreme values which might come up because of product markets with extremely low observations. In addition the estimated dataset also contains elasticities with a value of zero, which is very unusual for the type of observed goods. For this reason the next two sections are going to deal with those aforementioned problems. In a first step the thesis will explain, how data mining techniques can be used to detect outliers. The following section analyzes reasons and potential problems which lead to elasticities with a value of 0.

2.6.3 Detecting Outliers

As already mentioned in the previous section this part of the thesis deals with the detection of outliers. According to [Han and Kamber \(2000, p.381\)](#) "very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of the data, are called outliers". Based on this definition they state that outlier detection basically consists of two subproblems. First of all one has to define which data objects do not comply with the remaining data and secondly one has to find a method that efficiently detects outliers as defined in the first step.

As described by Han and Kamber (2000) contemporary literature distinguishes between the following three approaches of outlier detection²¹:

STATISTICAL-BASED OUTLIER DETECTION: Methods using this type of outlier detection assume that the data generating process for a given data set uses a specific distribution or probability model like e.g. the *normal distribution*. Based on this assumption the method uses a so called *discordancy test* to detect values which do not belong to the assumed distribution or probability model. This definition implies that the user knows the distribution of the given dataset and she or he also has information on distribution parameters like the mean and the variance. Based on the distribution and its parameters one can compute the probability that a specific value belongs to the specified distribution. A drawback of this type of methods is that most of the discordancy tests are built for single attributes whereas data will often be multidimensional.

DISTANCE-BASED OUTLIER DETECTION: The basic idea of this approach is that objects which do not have enough neighbors within a specified distance are marked as outliers. Or as expressed in a more formal way by Han and Kamber (2000, p.384): "An object o in a data set S is a distance-based (DB) outlier with parameters p and d , that is, $DB(p, d)$, if at least a fraction p of the objects in S lie at a distance greater than d from o ". The most important algorithms to mine distance-based outliers are the index-based algorithm, the nested-loop algorithm and the cell-based algorithm. The advantage of this approach is that one does not have to fit the observed distribution into a standard distribution and select adequate discordancy tests. On the other hand the quality of these approaches hinges on the selection of suitable values for p and d . The process of finding acceptable values is mainly trial and error, which can be very time consuming.

DEVIATION-BASED OUTLIER DETECTION: The main idea behind this approach is that outliers are not detected by using statistical tests or distances but rather by the extraction of the main characteristics of objects in a group and then by identifying objects that *deviate* from the discovered characteristics. Examples of this category are the *sequential exception technique* and the *OLAP data cube technique*.

Because of its simplicity and its ease of implementation this thesis will use the so called Grubbs' test which is a statistical-based outlier detection method. The main assumption of the Grubbs' test is that the data set under consideration can at least be approximated by a normal distribution. Applying the *Shapiro-Francia test* for normality yields a P-value of 0.464 indicating that the null hypothesis of a normal distribution cannot be rejected.

The demand curve of a normal good has a negative slope and thus a negative price elasticity of demand. As it will be explained in the section 2.6.4 there are some markets which feature too few observations which can lead to unusual outcomes. For this reason this thesis uses negative elasticities only. On the grounds of outlier detection this implies that a one-sided Grubbs' test has to be executed.

²¹ A detailed discussion and description of different methods for outlier detection is beyond the scope of this thesis, the interested reader may be referred to Han and Kamber (2000).

As described by Grubbs (1958) the test statistic of the one-sided Grubbs' test (G) can be computed as

$$G = \frac{\bar{X} - \min(X_i)}{S_n} \quad (2.35)$$

where \bar{X} denotes the mean of the variable X and S_n its standard deviation. Under the null hypothesis that there are no outliers in the dataset G will adopt a t-distribution. Hence one can reject the null hypothesis at a significance level of α if $G > z_\alpha$, where z_α is defined as follows:

$$z_\alpha = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{\frac{\alpha}{n}, n-2}^2}{n-2 + t_{\frac{\alpha}{n}, n-2}^2}} \quad (2.36)$$

where $t_{x,y}$ denotes the critical value of a t-distribution with a significance level of x and y degrees of freedom.

The Grubbs' test is a cyclic procedure. If the test detects an outlier, it will be excluded from the data set and the test will be executed again to search for a further outlier in the reduced data set. This process is repeated until the test does not find any further outliers. In the case of the Geizhals database the Grubbs' test has been implemented twice. In a first version Microsoft Excel has been used to implement the test via Visual Basic for Applications (VBA). The built-in statistical functions constitute the main advantage of VBA. However, it turned out that the VBA test suffered from severe performance problems. For this reason the testing routine has been optimized and implemented in a second version as a PHP²² script. The full PHP code including a short description can be found in the appendix on page 167.

The Grubbs' test detected roughly 500 outliers in the case of c_elast and 380 outliers in the case of l_elast . The new summary statistics computed after the exclusion of the outliers can be seen in table 5. One can see that the minimal elasticity is now only -37.943 respectively -41.449 instead of a value in the region of around $-300,000$.

Variable	Mean	Std. Dev.	Min.	Max.
c_elast	-6.531	6.766	-37.943	-1.00e-10
l_elast	-7.897	7.227	-41.449	-1.00e-10

Table 5: Summary statistics for the cleansed data set

The cleansed data set for which the summary statistics are reported in table 5 will be used in a first approach for the topics of chapters 3 and 4. Although the data has been cleansed from outliers the question still remains why there is a significant amount of elasticities with a value of 0. For the sake of simplicity these elasticities will be also termed *zero-elasticities* from now on.

2.6.4 Ensuring Data Quality - The Detailed Analysis Of A Random Sub-sample

The following sections show a detailed analysis of a random sample of 40 products drawn from the Geizhals database. The main purpose of

²² cf. <http://www.php.org/>.

this analysis is to get more insights into the estimated elasticities in order to possibly give any hints to the quality of the estimated elasticities. For this reason twelve alternative approaches to estimate the demand curve for a specific product and therefore also to estimate the price elasticity of demand have been chosen. The results show that the difference between the resulting elasticities from the different approaches can be drastic. A further important implication that due to econometric issues like endogeneity or heteroskedasticity the regression direction (i.e. regressing clicks on price and regressing price on clicks in order to estimate the elasticity and respectively the inverse elasticity of demand) has a huge impact on the estimated elasticities. Tables 6 and 7 describe a list of the products used in this analysis.

ID	DESCRIPTION	Subcategory	Category
6580	Canon BCI-3eBK Tintenpatrone schwarz	Consumeables	Hardware
35137	FireWire IEEE-1394 Kabel 6pin/4pin, 1.8m	Cables	Hardware
77828	OKI 1103402 Toner schwarz	Consumeables	Hardware
79079	Twinhan VisionDTV DVB-S PCI Sat CI	PC/Video	Hardware
86015	Sony Vaio PCGA-BP2V Li-Ionen-Akku	Notebook Accessories	Hardware
96539	Sony MDR-DS3000	PC/Audio	Hardware
97556	Digitus DA-70129 Fast IrDa Adapter, USB 1.1	Mainboards	Hardware
111148	Komplettsystem AMD Athlon 64 3200+, 2048 MB RAM	Systems	Hardware
115211	Diverse Mousepad schwarz	Inputdevices	Hardware
115433	HP Q6581A Fotopapier	Consumeables	Hardware
120019	Konica Minolta magicolor 5430 Toner schwarz	Consumeables	Hardware
128259	Zalman CNPS7700-Cu	PC Cooling	Hardware
130965	FSC Pocket LOOX 710/720 Bump Case	PDA/GPS	Hardware
132933	Apple iPod AV Kabel	Portable/Audio	Audio/HIFI
133848	Targus XL Metro Messenger Notebook Case	Notebook Accessories	Hardware
134972	Kingston ValueRAM DIMM 256MB PC2-4200E	Memory	Hardware

Table 6: Description of the random sample (part 1)

For this analysis each of the following elasticities have been estimated for each product:

- c_{elast}
- $c_{\text{elast_inv}}$
- $c_{\text{elast_choke}}$
- $aggr_c_{\text{elast}}$
- $aggr_c_{\text{choke}}$
- l_{elast}
- cv_c_{elast}
- cv_l_{elast}
- cv_lctw_c
- cv_lctw_l
- $aggr_lctw_c$
- $aggr_lctw_l$

ID	DESCRIPTION	Subcategory	Category
137437	Canon DCC-80 Softcase	Foto/Video Equipment	Video/Foto/TV
141318	Cherry Linux, PS/2, DE	Inputdevices	Hardware
142849	Philips HR1861 Entsafter	n/a	n/a
145401	Adobe: GoLive CS2 (deutsch) (PC)	n/a	n/a
160965	Creative Sound Blaster X-Fi Platinum, PCI	PC/Audio	Hardware
167564	Intel Xeon DP 3.80GHz,	CPUs	Hardware
169429	MSI K8MM3-V, K8M800 (PC3200 DDR)	Mainboards	Hardware
187584	Samsung SyncMaster 940N Pivot	Screens	Hardware
193387	ELO: EloOffice 7.0 Update (deutsch) (PC)	Security-Backup	Software
198925	Liebherr KTP 1810	Kitchen equipment	Household appliance
200404	ASUS M2NPV-VM, GeForce 6150/MCP 430	Mainboards	Hardware
203899	Hama 35in1 Card Reader, USB 2.0	Storage Mediums	Hardware
204301	Sony HVL-F56AM Blitzgerät	Foto/Video Equipment	Video/Foto/TV
205234	Hähnel HL-2LHP Li-Ionen-Akku	Foto/Video equipment	Video/Foto/TV
210626	Gigabyte GA-945GM-S2, i945G	Mainboards	Hardware
210768	Acer LC.08001.001 80GB HDD	Notebook Accessories	Hardware
216426	Samsung ML-4551ND	Printer-Scanner	Hardware
217675	Adobe: Photoshop Elements 5.0 (PC)	n/a	n/a
219507	Canon BCI-6 Multipack Color	Consumeables	Hardware
220621	V7 Videoseven L22WD	Screens	Hardware
224495	Apple MacBook Pro Core 2 Duo	Notebooks	Hardware
227850	Sony Walkman NW-S703FV	Portable-Audio	Audio/HIFI
230160	Seagate SV35 320GB	Harddisc	Hardware
230429	Trust WB-5400 4 Megapixel Webcam	PC-Video	Hardware

Table 7: Description of the random sample (part 2)

Since the detailed analysis focuses on the general quality of the elasticities rather than the resulting elasticities themselves, this section only presents a summary of the estimated elasticities. The summary statistics of this analysis are shown in figure 8. The figure shows for each elasticity approach the average value of the elasticity and the corresponding \bar{R}^2 and t-value. The average values are written in boldface. Additionally the figure reports for each average value the respective standard deviation and the min- and max-value. The most striking facts of these results are the difference between c_elast and c_elast_inv , the high \bar{R}^2 and t-values of the aggregated versions and that the linear demand seems to have the worst fit.

	c_elast	c_elast_inv	c_elast_choke	aggr_c_elast
Elast	-4,1287	-74,6587	-4,7762	-16,6252
Stddev	5,1226	98,6627	14,7604	11,8580
Min-Max	-22,746 0,644	-339,995 327,890	-83,330 19,705	-46,689 -1,157
Adj. R²	0,0565	0,0565	0,0408	0,6756
Stddev	0,0817	0,0817	0,1091	0,1930
Min-Max	-0,019 0,379	-0,019 0,379	-0,077 0,559	0,151 0,921
t-Val	-2,1163	-2,1163	-1,5198	-15,0183
Stddev	1,5526	1,5526	2,0442	9,1062
Min-Max	-7,995 0,389	-7,995 0,389	-7,890 2,000	-40,687 -5,082
N	35	35	25	34
	aggr_c_choke	l_elast	cv_c_elast	cv_l_elast
Elast	-18,6307	-6,9571	-4,2608	-8,1016
Stddev	24,2729	6,3954	4,7868	7,6824
Min-Max	-125,332 0,000	-24,320 4,241	-22,749 -0,047	-29,536 0,689
Adj. R²	0,6454	0,0251	0,1624	0,0868
Stddev	0,2520	0,0436	0,1814	0,1521
Min-Max	0,000 0,935	-0,029 0,157	-0,029 0,734	-0,139 0,528
t-Val	-12,9871	-1,5666	-2,0688	-1,4702
Stddev	9,5055	1,0341	1,6204	1,1052
Min-Max	-40,045 0,000	-5,031 0,513	-7,052 -0,109	-4,567 0,436
N	27	34	36	34
	cv_lctw_c	cv_lctw_l	aggr_lctw_c	aggr_lctw_l
Elast	-0,4843	-1,5290	-5,8383	-4,2156
Stddev	1,4549	4,2121	7,7269	6,0246
Min-Max	-5,744 1,394	-14,687 5,011	-32,106 0,000	-21,274 0,000
Adj. R²	0,0840	0,0075	0,2855	0,2540
Stddev	0,2195	0,2403	0,2966	0,2708
Min-Max	-0,064 1,000	-0,346 1,000	-0,013 0,916	-0,022 0,804
t-Val	-0,2276	-0,0735	-5,3725	-4,8593
Stddev	1,5490	1,4455	5,9276	5,6964
Min-Max	-3,772 5,455	-2,204 5,769	-18,776 0,000	-19,771 0,000
N	15	14	19	19

Figure 8: Summary statistics of the elasticities and the corresponding \bar{R}^2 and t-values for the 40-product-sample

After the short review on the resulting elasticities the next couple of sections deal with explanations of potential problems of the estimation approach of the elasticities respectively problems with the elasticities themselves.

2.6.4.1 Zero-Click-Offers

One of the most striking features of the products used in this sample is the huge proportion of offers, which did not generate any clicks

Zero-click-offers can have a significant impact on the slope of the estimated demand curve.

at all. So far two quality criteria for the second stage regression have been set. First of all the estimated elasticities from the first stage had to pass Grubbs' test for outliers in order to be used in the second stage. Secondly a lower limit for the number of observations for each product in the first stage has been introduced and set to 30, i.e. all products which did not feature more than 30 offers were excluded from the regressions (both first- and second stage).

Nevertheless there was no test whether these offers had received any clicks at all. Therefore it is very well possible that there are products with 30 offers but only 2 of them have received any clicks, thus making it impossible to estimate a meaningful demand curve. Even if there were a sufficient number of offers that received clicks the detailed analysis has shown that the estimated demand curves can be influenced respectively deteriorated heavily by a large proportion of zero-click-offers. A graphical example of above mentioned problem is given in figure 9. It shows a scatter plot for a Canon Softcase (product with the ID 137437).

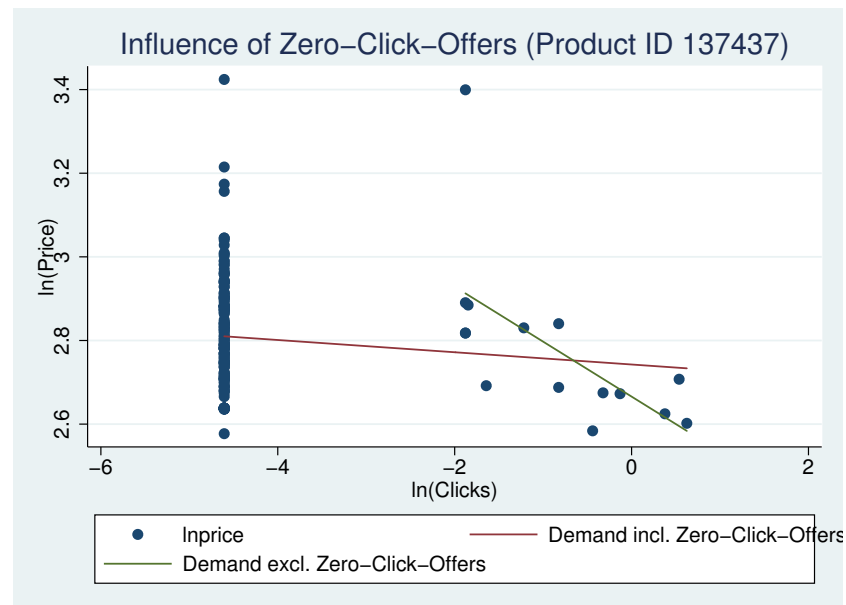


Figure 9: Scatterplot featuring zero-click-offers

The line *Demand incl. Zero-Click-Offers* shows the estimated demand curve by using all available observations for the product, i.e. also including all offers which have not received any clicks at all. As one can see, this demand curve is much flatter than the line *Demand excl. Zero-Click-Offers*, which has been generated by a regression which only uses offers with a number of clicks greater than 0. The axes in the figure are given in logs, because the difference between the two demand curves and respectively their elasticities is much easier to see in the log version. To highlight the impact of zero-click-offers one can take a look at the numerical elasticities. The demand curve with zero-click-offers has an elasticity of -68.31 . On the other hand the demand curve without zero-click-offers has an elasticity of -7.61 . Although this difference might not be as large for other products, the detailed analysis has shown that the zero-click-offers definitely do bias the estimated elasticities.

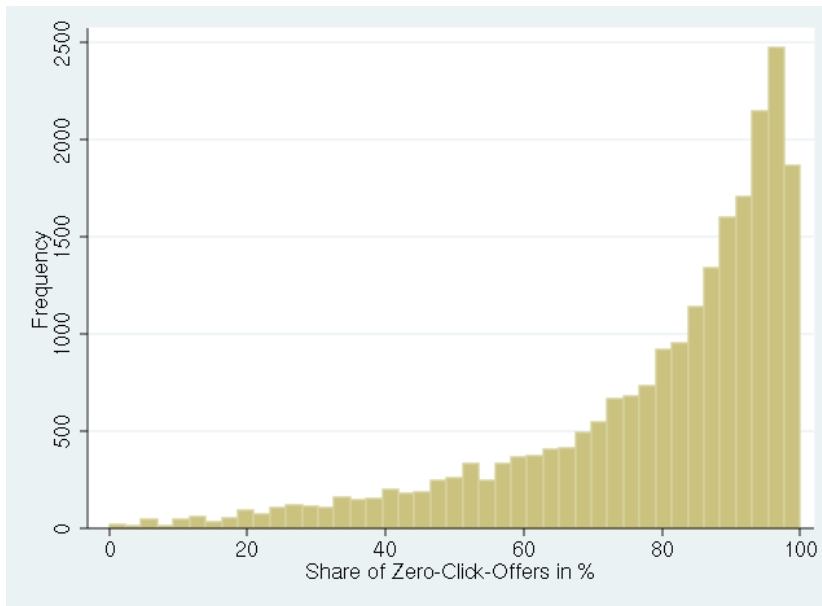


Figure 10: Share of zero-click-offers

In order to get an overview of the general situation of the data figure 10 shows a histogram where the x-axis is the percentage of offers of a product with zero clicks (=zero-click-offers).

As one can observe in figure 10 the vast majority of products has a share of zero-click-offers of 75% percent and more. The average product has 53 offers and a share of zero-clicks of 62%, i.e. on average the typical product consists of 20 offers which have received any clicks. To get a more complete picture of the data one has to notice that roughly 67% of all products have less than 50 offers and 80% of all offers have less than 100 offers.

Thus it would seem reasonable to introduce a maximum threshold for the share of zero-click-offers as a further data quality criterion. Although this seems a reasonable idea, one has to take into consideration that such a criterion would drastically reduce the number of products in question. Table 8 gives an overview of how the data sample size is influenced by the filter criteria.

The conclusion is that the data sample used for the estimation of the elasticities has to be modified.

ZERO-CLICK-OFFERS AT HIGH PRICES Apart from biasing the normal elasticities there is a further phenomenon which especially arises with an adequate number of zero-click-offers at high prices. Zero-click-offers at high prices tend to favor demand structures with an isoelastic demand specification over linear demand curves. The reasoning behind this is that zero-click-offers at higher prices make even a rather flat scatter plot look like a convexly shaped scatter plot. A graphical example of this phenomenon can be seen in figure 11.

2.6.4.2 Normal vs. Inverse Demand Curves

Theory suggests that the demand curve and the inverse demand curve are exactly the same. Because of imperfect data in combination with the workings of OLS this is not true in practical work. When estimating the

DATA FILTER	NO. OF PRODS.
No filter	39681
Only products with more than 30 observations	16170
Only products with a share of zero-click-offers below 40%	9789
Only products with more than 30 observations and a share of zero-click-offers below 40%	389
Only products with more than 30 observations and a share of zero-click-offers below 30%	202
Only products with more than 15 observations and a share of zero-click-offers below 40%	1231
Only products with more than 15 observations and a share of zero-click-offers below 30%	676

Table 8: Alternative data filters

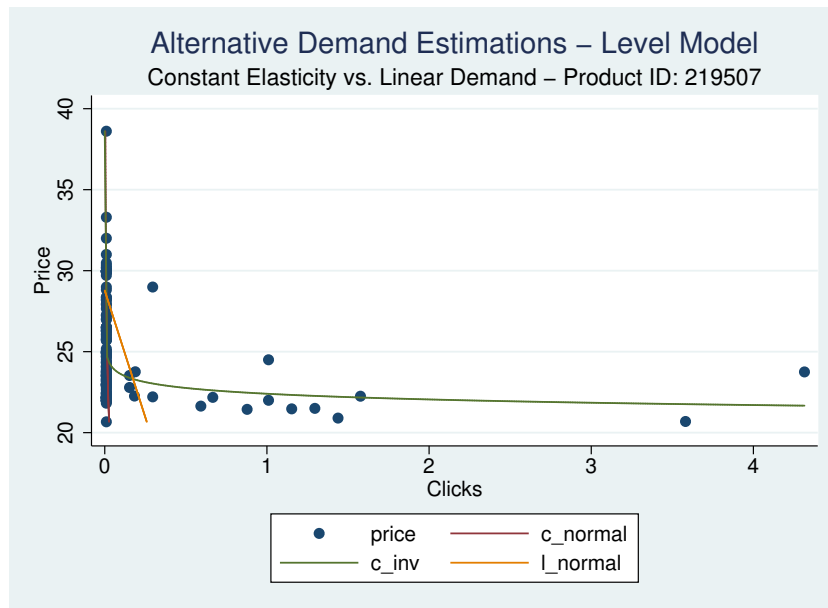


Figure 11: Influence of zero-click-offers at high prices

elasticity of a product listed at Geizhals one can either regress *clicks* on *price* (respectively *lnclicks* on *lnprice* in the case of an isoelastic demand curve) to estimate the elasticity (ϵ) or one can regress *price* on *clicks* in order to estimate the inverse elasticity ($1/\epsilon$). The summary of the estimation results on page 33 show that the difference between those two elasticities is huge. The average normal elasticity c_{elast} is -4.13 whereas the average inverted elasticity is -74.66 .

A simple difference between those two elasticity types would not constitute a problem per se because it might be that on average this difference does not have a systematic root. To check whether this difference just constitutes a kind of offset one has to take a look at the correlation coefficient of those two elasticities. Unfortunately the correlation between c_{elast} and c_{inv} is with 0.04 quite low.

Although the coefficients of those two opposing estimation approaches are totally different the statistical properties are the same. R^2 , \bar{R}^2 and t-values do not depend on the regression direction (in terms of regress-

ing clicks on price or regressing price on clicks). Thus one cannot use the statistical properties of the coefficients in order to judge the quality or correctness of those two approaches.

Therefore the only thing one can do, is to take a look at the graphical representations of the estimated demand curves and check which demand curve seems more intuitive. This is exactly what has been done during this analysis. When looking at the resulting graphs one can say that the inverse demand functions seem to fit the data better than the normal demand curves. The graphs give the impression that the normal demand is more susceptible to zero-click-offers, i.e. the bias resulting from zero-click-offers seems to be larger in the case of normal demand curves.

2.6.4.3 Quality Of Shipping Costs

For the majority of products the shipping costs range from €0 to roughly €30. Unfortunately one has to question the quality of the shipping costs variable. Especially for low price products the shipping costs can account for over 90% of the final price. In such cases the quality of the estimated elasticity is hugely dependent on the accuracy of the shipping cost data.

Unfortunately an analysis of the shipping costs shows that the accuracy might be mediocre at best.

- First of all the Geizhals database only provides shipping costs for 61% of the products offered during the period of observation.
- Secondly, when looking at the Geizhals website, one can notice that retailers quite often do not report the shipping costs as an absolute value, they rather post the cheapest value from a range of shipping costs ("starting at €5,90").
- Finally, 8 out of 390 retailers report negative shipping costs solely.

Apart from the problems listed above the analysis has shown that concerning the shipping costs there seem to be two types of retailers:

CONSTANT SHIPPING COSTS Retailers that offer rather constant shipping costs, i.e. a very small selection of shipping costs applied to a large number of products or offers. For example the retailer *1ashop.at* provided 605 offers free of any shipping costs and 20201 offers at shipping costs of €5.90.

VARIABLE SHIPPING COSTS Retailers that report a very large number of different shipping costs, but the difference between those shipping costs is negligible. Geizhals for example lists for one of the retailers 446 shipping costs in the range of €11 to €13.

Since the shipping costs are only extracted from a textfield via a scanning script it is questionable whether this data is really reliable. Especially the second type of shipping costs seems to be odd. Furthermore the regression results show that the difference in elasticities between the normal versions and the versions with control variables is only very small. This suggests that the influence of the control variables and therefore also the shipping costs is negligible.

All in all one can say that the data on shipping costs in general is very vague and one has to ask whether they should be really used in the regression process. If one includes the shipping costs one has to

be very cautious with the results and should consider whether it pays off to set minimum price for products included in the regression. This minimum price should ensure that the shipping costs do not have too much of an impact on the regression results.

2.6.4.4 Correlation Between Alternative Elasticities

In order to assess the quality and robustness of the estimated elasticity alternatives this section takes a look at the correlation matrix of the elasticities. The correlation matrix is shown in figure 12.

	c_elast	c_inv	c_choke	aggr_c	ag_chok	l_elast
c_elast	1,0000					
c_inv	0,0406	1,0000				
c_choke	0,6121	0,0205	1,0000			
aggr_c	0,4217	0,1052	0,4054	1,0000		
aggr_choke	0,2100	-0,0034	0,2154	0,7009	1,0000	
l_elast	0,5844	0,1212	0,6298	0,6867	0,5521	1,0000
cv_c	0,7942	0,0380	0,5224	0,4802	0,2190	0,5496
cv_l	0,5072	0,0303	0,5166	0,6342	0,4782	0,8321
cv_lctw_c	0,2137	0,0480	0,1047	0,0593	0,0193	0,0726
cv_lctw_l	0,2381	-0,0111	0,1760	0,4521	0,2400	0,3252
aggr_lctw_c	0,3589	0,0934	0,1178	0,5082	0,3915	0,3235
aggr_lctw_l	0,2605	0,0824	0,1121	0,6339	0,3991	0,4066
	cv_c	cv_l	cv_lct_c	cv_lct_l	ag_lct_c	ag_lct_l
cv_l	0,7652	1,0000				
cv_lctw_c	0,1480	0,0907	1,0000			
cv_lctw_l	0,2787	0,3266	0,7167	1,0000		
aggr_lctw_c	0,2010	0,1831	0,4995	0,5670	1,0000	
aggr_lctw_l	0,1725	0,2697	0,4859	0,6939	0,8377	1,0000

Figure 12: Correlation matrix of the 40-product-sample

One can draw the following conclusions from the correlation matrix :

1. Apart from two coefficients all correlation coefficients are positive in this sample. That means that almost all correlations move in the same direction. Only the correlations between *aggr_c_choke* and *c_elast_inv* and also *cv_lctw_l_elast* and *c_elast_inv* are slightly negative with values of -0.0034 and $-0,011$.
2. The correlation between the basic elasticity (*c_elast* respectively *l_elast*) and the corresponding versions which include control variables (*cv_c* respectively *cv_l*) is very high. This fact becomes even more interesting if one takes into consideration that the average value of the estimated coefficient is also almost the same in both versions (-4.13 vs. -4.26 in the case of *c_elast* vs. *c_cv*). This suggests that the control variables hardly have any statistical significance and therefore hardly have any influence on the estimated elasticities.
3. The correlation between an elasticity and its choke version is also above average. In the case of *c_elast* and *c_choke* the correlation coefficient is 0.6121 and in the aggregated version the correlation coefficient is 0.7009 . This suggests that although the estimated

elasticities are similar, there still seems to be a difference when excluding the offers above the choke price.

4. The elasticity coming from the inverse demand curve hardly correlates with any other elasticity at all. This suggests that opting for one of the two regression directions (price on clicks or clicks on price) is not only a decision between rather flat or rather steep demand curves, it is a decision between two completely different elasticity alternatives.
5. The correlation between the normal version of an elasticity and its aggregated version is lower than expected. In the case of c_elast and $aggr_c_elast$ the correlation coefficient is 0.4217. This correlation diminishes to a value of 0.21 if one excludes the offers above the choke price.
6. The correlation between the constant elasticity and the linear demand elasticity is at a moderate level with 0.5844. This suggests that both elasticities might constitute a reasonable approach.

2.6.5 Improving Data Quality

Although the estimated elasticities using the full range of products seem to have a reasonable distribution and feature reasonable values the detailed analysis has shown that there are significant disturbing factors which might result in biased estimates. As mentioned before chapters 3 and 4 will use the results from this stage to explain the factors which determine the price elasticity of demand respectively to analyze the impact of the price elasticity of demand on the level and type of competition in a market. In a first step this thesis will use the full range of elasticities for the forthcoming chapters. However, this will mainly be done in order to check the robustness of the results. The second stage will partly reveal counterintuitive results when using the full range of elasticities. To control for problems stemming from data impurities the second stage regressions will not only be carried out by using the full set of elasticities but rather the analysis will also cover a reduced dataset with an improved quality.

To determine the quality of the data this thesis uses the dimensions *number of observations* and *share of zero-click offers*.

NUMBER OF OBSERVATIONS: As it has already been discussed in section 2.6.2 the variation in the estimated elasticities correlates with the number of observations that have been used to compute the respective elasticity. Elasticities which are based on only a few number of observations tend to adopt rather extreme values and can therefore be considered as doubtful. In order to prevent any potential problems coming from too few observations one has to set a threshold on this dimension. Manual inspection has shown that a threshold of 30 seems to be a reasonable value, i.e. elasticities which have been estimated by using 30 or fewer observations will be excluded from any further computations and estimations.

SHARE OF ZERO-CLICK OFFERS: Especially during the detailed analysis in section 2.6.4 it became clear that there is quite a large amount of products which do feature a significant number of offers (and therefore also a significant number of observations)

but these offers do not generate any clicks at all. As elasticities vary with the inclusion of those zero-click offers different versions are calculated based on the maximum number of allowed offers with no clicks. However, this number has to be set in relation to the total number of offers for the specific product. For this reason the share of zero-click offers for a specific product seems to be an adequate dimension for the data quality. One has to keep in mind the imposition of thresholds reduces the set of available products, therefore the limits must not be set too tight. A maximum share of zero-click offers of 30% seems to ensure both, an adequate data quality and also an acceptable number of products in order to execute meaningful second stage regressions.

As one can observe from table 8 on page 36 setting the thresholds for above mentioned data quality dimensions reduces the data set to a total amount of 202 products. The summary statistics for this reduced data set are given in figure 14. Compared to the detailed analysis the results are very similar to the results of the reduced data set.

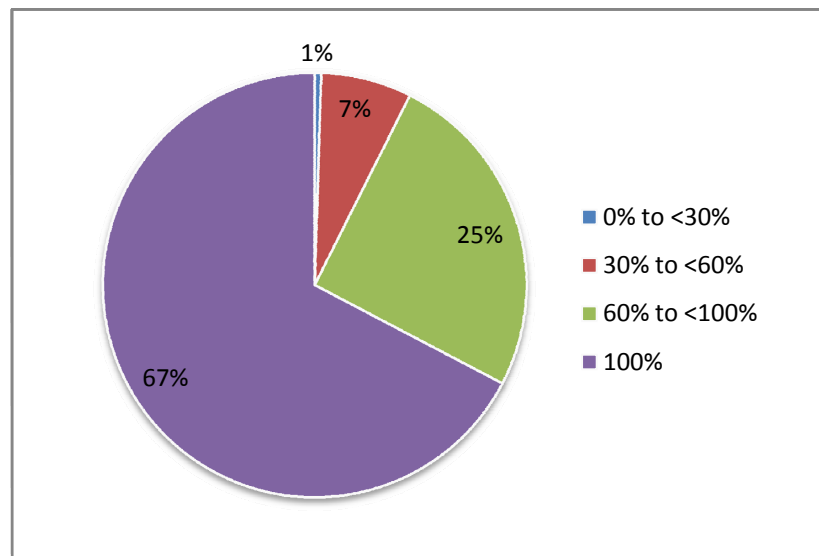


Figure 13: Share of offers with zero-LCT grouped into categories

One of the main features is again the huge difference between the normal isoelastic demand (c_elast), which shows an elasticity of -10.76 and its inverse equivalent where the elasticity is reported as -475.18 . This indicates that the cleansed and refined data is still far from perfect. The coefficient on c_elast_choke does hardly differ from the normal c_elast variable. This means that in this reduced dataset there do not seem to be many products with a large number of zero-click offers above the choke-price. Apart from the elasticities coming from linear demand structures and those based on last-clicks-through the estimated coefficients seem to be statistically significant on average.

As expected the elasticities of the aggregated and accumulated clicks display the highest \bar{R}^2 values, although this is not true for the demand curves which are based on LCT. Since LCT are a subset of all clicks, i.e. each LCT is also a normal click but not all normal clicks are also LCT it might be that the LCT versions are plagued even more by zero-click offers. This presumption can be verified by graphing the share of offers which do not possess any LCT. The results can be seen in figure 13. As

	c_elast	c_elast_inv	c_elast_choke	aggr_c_elast
Elast	-10,7596	-475,1804	-10,7088	-16,9004
Stddev	9,0840	5556,3220	9,1146	10,6835
Min-Max	-48,481 6,885	-78922,5 2977,822	-48,481 6,885	-61,573 -1,169
Adj. R²	0,1491	0,1491	0,1426	0,8341
Stddev	0,1347	0,1347	0,1316	0,1240
Min-Max	-0,033 0,707	-0,033 0,707	-0,035 0,707	0,137 0,977
t-Val	-2,9428	-2,9428	-2,8630	-17,4142
Stddev	2,0069	2,0069	1,9858	8,9636
Min-Max	-12,468 2,233	-12,468 2,233	-12,468 2,233	-64,667 -2,918
N	189	181	183	137
	aggr_c_choke	l_elast	cv_c_elast	cv_l_elast
Elast	-16,0964	-6,7965	-11,3072	-6,6347
Stddev	9,8368	6,5058	8,8858	6,5322
Min-Max	-53,467 -11,694	-34,561 3,748	-41,602 5,276	-34,914 13,925
Adj. R²	0,8455	0,0622	0,2437	0,0755
Stddev	0,1153	0,0714	0,1732	0,1476
Min-Max	0,137 0,977	-0,033 0,510	-0,204 0,735	0,000 0,903
t-Val	-17,9369	-1,8574	-2,7992	-1,6201
Stddev	9,1080	1,1852	1,8472	1,1538
Min-Max	-64,667 -2,918	-8,472 1,746	-10,222 1,344	-8,053 2,274
N	136	181	189	189
	cv_lctw_c	cv_lctw_l	aggr_lctw_c	aggr_lctw_l
Elast	-1,4447	-1,4769	-4,3723	-1,8204
Stddev	3,1994	3,2997	8,1553	3,3681
Min-Max	-17,106 5,471	-20,405 8,362	-51,258 0,000	-22,272 0,000
Adj. R²	0,0440	-0,1316	0,2513	0,1733
Stddev	0,1051	0,1514	0,3673	0,2786
Min-Max	0,000 0,585	-0,364 0,456	0,000 0,901	0,000 0,914
t-Val	-0,5575	-0,3768	-4,0660	-2,4760
Stddev	1,0903	0,7540	6,4085	4,3127
Min-Max	-5,639 1,295	-3,747 1,646	-28,604 0,000	-27,436 0,000
N	56	59	38	39

Figure 14: Summary statistics of the elasticities and the corresponding R^2 and t-values for the cleansed dataset

one can observe the vast majority of products (67%) did not receive any LCT. For 25% of the products the share of offers with no LCT is between 60% and 100%. Only 8% of the products show a share of offers with no LCT of less than 60%. This indicates that the problem of zero-click offers is especially severe in the case of LCT.

The whole picture does not change if one uses the complete range of products. Even with the complete data set 85% of the products have a share of offers without last-clicks-through of 90% and above. This indicates that the expressiveness and validity of the elasticities based on last-clicks-through is questionable. As mentioned during the detailed analysis, the occurrence of zero-click-offers is less problematic in the case of aggregated and accumulated clicks. This should be kept in mind when executing the second stage regressions.

2.6.5.1 Results of the full dataset

After the creation of a reduced and cleansed dataset, the insights of the detailed analysis have been compiled into a set of factors and rules that ensure data quality of the full dataset. Therefore the final part of this section presents the results of the full dataset. The results are presented in figure 15 with the notion already used before, i.e. for each

type of elasticity the section reports the average value of the respective elasticity and the corresponding \bar{R}^2 and t-value. These average values are denoted in boldface. Additionally the report includes the standard deviation and the min- and max-value for each average value.

	c_elast	c_elast_inv	c_elast_choke	aggr_c_elast
Elast	-4,4498	-77,2134	-7,2094	-17,5586
Stddev	4,9179	61,6832	7,3336	11,8744
Min-Max	-37,337 -0,001	-370,5 -1,528	-51,048 -0,003	-73,113 -0,155
Adj. R²	0,0734	0,0772	0,0758	0,6729
Stddev	0,1023	0,1033	0,1044	0,2047
Min-Max	-0,034 0,831	-0,033 0,831	-0,034 0,757	0,002 0,986
t-Val	-2,5910	-2,6872	-2,6012	-16,3071
Stddev	1,8381	1,8186	2,0058	9,7031
Min-Max	-14,172 -0,001	-14,172 -0,027	-14,177 -0,001	-108,647 -1,040
N	14814	14264	10615	14710
	aggr_c_choke	l_elast	cv_c_elast	cv_l_elast
Elast	-15,7207	-8,5293	-4,9894	-9,4102
Stddev	12,0027	7,3461	5,4154	8,3691
Min-Max	-69,394 -0,009	-41,449 -0,001	-42,748 -0,001	-53,475 -0,001
Adj. R²	0,7288	0,0374	0,1439	0,0791
Stddev	0,1870	0,0637	0,1638	0,1537
Min-Max	-0,004 0,986	-0,034 0,730	-0,280 1,000	-0,321 1,000
t-Val	-18,2484	-1,8716	-2,7577	-1,8463
Stddev	10,8131	1,1771	22,8003	1,4533
Min-Max	-108,647 -0,933	-11,078 -0,010	-2764,763 0,000	-75,586 0,000
N	11680	14940	14771	14795
	cv_lctw_c	cv_lctw_l	aggr_lctw_c	aggr_lctw_l
Elast	-1,3129	-4,4954	-9,6847	-7,2759
Stddev	1,7163	4,9641	7,5473	5,4810
Min-Max	-13,818 0,001	-34,103 -0,001	-44,523 -0,030	-30,961 -0,010
Adj. R²	0,0745	0,0517	0,4843	0,3836
Stddev	0,1463	0,1409	0,2404	0,2292
Min-Max	-0,269 1,000	-0,301 1,000	-0,013 0,960	-0,022 0,941
t-Val	-1,7086	-1,3712	-10,5757	-8,1166
Stddev	1,3503	1,0276	6,9424	5,2030
Min-Max	-13,450 0,000	-10,861 0,000	-65,352 -0,688	-54,146 -0,332
N	4824	4833	5329	5322

Figure 15: Summary statistics of the elasticities and the corresponding \bar{R}^2 and t-values for the full dataset

A comparison of figures 15 and 14 shows that the results of the full dataset are quite similar to those from the reduced and cleansed dataset. Therefore the properties of the results and the conclusion drawn from these properties are similar as well. The most striking feature is still the huge difference between the average value of *c_elast* and *c_elast_inv*. Again the elasticities of the aggregated and accumulated clicks display the highest \bar{R}^2 values. One can conclude by saying that these statistics show that the full dataset represents a valid basis for the estimation of the second stage regressions. The full dataset does not suffer from any flaws and provides the advantage of being a very rich dataset.

2.7 CONCLUSION

Finally one can say that this chapter has established a procedure to estimate the price elasticity of demand in the case of the online price

comparison website Geizhals. The major problem of such a website is that it only reports clicks and not actual purchases. This problem can be circumvented by setting a fixed conversion ratio or using so called last-clicks-through. However, there exist several elasticity alternatives. To ensure data quality and robust results three sets of elasticities have been estimated. The first set covers the complete range of products, the second set has been used for a detailed analysis to uncover potential data problems. Furthermore the results from the second set have been used to build a reduced third data set with a higher quality. The full data set and the reduced data set will be used for the second stage regressions in chapters 3 and 4.

Finally one has to recall that this thesis is a preliminary study that is part of a larger research project. A detailed treatment of econometric issues like heteroskedasticity, endogeneity and simultaneity is not an objective of this thesis. Therefore subsequent studies have to test for aforementioned issues and, if applicable, use IV and 2SLS to control for them.

FACTORS INFLUENCING THE PRICE ELASTICITY OF DEMAND

The work in this chapter is based on the estimated elasticities from chapter 2. In the previous chapter the elasticities have been estimated as the coefficient of a right-hand-side (RHS)-variable. In this section the elasticity will be put on the LHS, i.e. it will be used as the explained variable. As already seen in the chapter 2 the elasticity of demand does vary across products. The goal of this chapter is to trace out, whether systematic determinants of the elasticity of demand can be found or whether the elasticity of demand and hence consumer tastes are given exogenously. The structure of this chapter starts with the introduction of previous research. The second section deals with the description of the dataset. The third section formulates hypothesis based on the dataset and the final section presents the results and concludes.

3.1 PREVIOUS RESEARCH

The idea of this chapter is based on the paper by Pagoulatos and Sorensen (1986). They use data from the U.S. food and tobacco manufacturing industries in order to explain systematic differences in the elasticities of these industries. They use the work of Scitovsky and Stigler to state that *"the major factors that influence the elasticity of demand for a product include the availability and closeness of substitutes for the product, the degree to which the product is a complement to other goods, the level of information in the market regarding the product and its substitutes, and the competitive behavior of the firms in the market in which the product is sold"*.

The availability and closeness of substitutes influences the substitution effect of a price change. If the availability of close substitutes is greater, then the substitution effect of a price change will be more important and larger. The availability of complements on the other hand dampens the substitution effect and therefore also decreases the elasticity of demand. The information on the market regarding substitutes and complements can be seen as a catalyst for the availability and closeness of substitutes and complements because as Pagoulatos and Sorensen (1986, p. 242) state *"the response to a price change will be larger the more knowledge market participants have about alternatives"*.

However, it is hardly possible to obtain direct quantitative measures for above described determinants. This is even more of a problem in the case of the Geizhals database. From an empirical point of view a possible solution would be the usage of proxy variables. This is also the approach of Pagoulatos and Sorensen (1986) who use variables *"that are expected to influence the degree of substitutability and complementarity of products, the level of market information, and the competitive behavior of firms across industries"*. The list of variables contains the total number of brands, the R&D expenditures to domestic sales ratio, the advertising to domestic sales ratio, the four-firm concentration ratio, the capital requirements, the effective tariff rate and the percentage of industry output sold to the final demand sector.

The availability and closeness of substitutes are the major factors influencing the price elasticity.

TOTAL NUMBER OF BRANDS (BN): Pagoulatos and Sorensen (1986) state that the total number of brands in an industry influences the level of information on the market and also the availability of substitutes. To obtain information on price and quality of products or brands a consumer has to conduct search activities. An increasing number of brands increases the complexity of the search and hence the search costs. This in turn would imply that a greater number of brands decreases the available information on the market. The second argument of Pagoulatos and Sorensen (1986) is that an industry with more brands tends to have a lower substitutability because the brands in such an industry fill product gaps and therefore prevent customers from switching to substitutes. Both of these arguments lead to a less elastic demand.

The arguments of Pagoulatos and Sorensen (1986) seem to be pretty straightforward, however they also offer drawbacks. Concerning the available information one could argue that although a larger number of brands significantly increases the cost to obtain price and quality data on *all* (N) products, it does not increase the search costs of a specific number (n) of products. Therefore an increased number of products might lead to less informed customers in relative terms (n/N), but not in absolute terms (n). Additionally the argument of Pagoulatos and Sorensen (1986) that a larger number of brands decreases substitutability hinges on the assumption that each brand in the industry has its own niche. One could argue that this will not always be the case. If the assumption of Pagoulatos and Sorensen (1986) turns out not to be true, then an increased number of brands would imply an increased number of potential substitutes, which in turn would yield a more elastic demand.

R&D EXPENDITURES TO DOMESTIC SALES RATIO (R&D/DS): This ratio can be seen as a proxy for the turnover rate of brands. If new brands get introduced into and old brands are withdrawn from an industry consumers have to gather new information on prices and quality, i.e. a higher turnover rate of brands depreciates the information of consumers faster. This finally results in a less elastic demand.

ADVERTISING TO DOMESTIC SALES RATION (A/DS): According to Pagoulatos and Sorensen (1986) it is not possible to predict the influence of advertising on the price elasticity of demand a priori. On the one hand advertising is used to differentiate a product, on the other hand advertising is the main source of consumer information. The former effect would imply a less elastic demand, whereas the latter effect would lead to a more elastic demand.

In addition Pagoulatos and Sorensen (1986, p. 244) introduce three variables which should measure the degree of collusion within an industry, since *"a final factor that may act to prevent substitution in response to a price change is the existence of tacit or explicit collusion to restrict price competition in the market"*.

FOUR-FIRM INDUSTRY CONCENTRATION RATIO (CR₄): On the one hand the price elasticity of demand should decrease with market concentration, since tacit or explicit collusion is easier to achieve

because it enables firms to effectively monitor and enforce pricing agreements, reducing the consumer's incentives for switching products. On the other hand Pagoulatos and Sorensen (1986) state an argument from Becker's work that in a more concentrated market firms will reduce their output to raise the price, in order to maximize profits. This will shift the equilibrium point to a more elastic part of the demand curve, which would counter the above described effect. Pagoulatos and Sorensen (1986) conclude that a priori one cannot tell whether the price elasticity of demand will increase or decrease with an increasing four-firm industry concentration ratio.

However, one has to bear in mind that Becker's argument will only be true in the case of a linear demand curve. In an environment with an isoelastic demand, the elasticity will not change when moving along the demand curve.

CAPITAL REQUIREMENTS (KREQ): This variable "measures the combined effect of barriers attributable to scale economies and absolute capital requirements" (Pagoulatos and Sorensen (1986, p. 244)). Increased entry barriers should decrease the elasticity of demand because less substitution possibilities exist.

EFFECTIVE TARIFF RATE (EFFT): This variable represents the barriers for foreign entrants. Similar to KREQ an increase in this variable should decrease the elasticity of demand because less possibilities for substitution exist.

Finally Pagoulatos and Sorensen (1986, p. 244) suggest including a variable to control for differences in complementarity across industries. Intermediate goods are usually used in combination with other goods, hence such goods usually feature a higher degree of complementarity. Theory suggests that goods for which a large amount of complements are available should be subject to a less elastic demand curve. The same argument can be used to state that industries with a large proportion of intermediate goods have a less elastic demand curve. For this reason Pagoulatos and Sorensen (1986, p. 244) introduce:

SHARE OF FINAL DEMAND SECTOR (CD/S): This variable measures the percentage of industry output that is sold to the consumer sector. This variable can be interpreted as the inverse to the share of intermediate goods in an industry, thus the larger the share the more elastic the demand should be.

Therefore the variables can be compressed into the following single equation:

$$\epsilon = f(\overset{+/-}{BN}, \overset{+}{R\&D/DS}, \overset{+/-}{A/DS}, \overset{+}{CR4}, \overset{+}{KREQ}, \overset{+}{EFFT}, \overset{-}{CD/S}) \quad (3.1)$$

where a + represents a larger and – a smaller value of ϵ . Since the elasticity ϵ is a negative number a larger value of ϵ indicates a less elastic demand and a smaller value of ϵ indicates a more elastic demand curve.

The result of this section so far is a model which tries to explain differences in the price elasticity of demand between several industries. One of the goals of this chapter is to adapt the idea of Pagoulatos and Sorensen (1986, p. 244) to the case of Geizhals. Therefore this thesis will

establish an other set of variables trying to explain the price elasticity of demand. However, the focus of the analysis in this thesis is the elasticity on a product-level and not on an industry-wide-level.

3.2 DATASET DESCRIPTION

The dataset for the regressions carried out in this chapter consists of the elasticities from the first stage, of product-specific data and of category-specific data. The full list of variables can be found in the appendix in tables 54 and 55 on pages 116 and 117. However, the correlation matrix shows that quite a large amount of variables are correlated highly. If one uses the full set of explaining variables, none of them will be statistically significant because of their high correlation. Therefore the approach in this thesis is to start from a minimal regression specification and then iteratively add and/or change variables of the specification.

In order to extract a minimal set of explanatory variables the full set of variables has been split up into a set of logical cohesive groups. These groups are *product quality*, *substitutability*, *brand dummies*, *category dummies* and a set of *miscellaneous* variables. The following paragraphs will introduce each of these groups and also report the list of variables which belong to the respective group. To keep it simple and straightforward a variable description is only given for variables which will be used during the second stage regressions. Variables whose names are written in bold letters, like e.g. **prod_recommendation**, are part of the specification of the minimal regression model.

PRODUCT QUALITY The variables of this group represent different measurements of product quality. Amongst other measurements this group covers the absolute product quality and the quality in relation to the average quality of the remaining products of the respective subcategory. Furthermore this group also contains variables describing the share of users which would recommend the product under observation. However, previous work suggests that the recommendation variable distributes more to the explanation of elasticities than the other measures of product quality. In addition relative variables, i.e. variables which put a property of a product like e.g. its quality in relation to the average property value of the remaining products of the subcategory, tend to be correlated with variables from the group *substitutability*.

VARIABLE	DESCRIPTION
prod_recommendation	Share of users who recommend the current product
prod_avgmissing	1 if the current product's quality data is missing
prod_quality	See appendix page 117
prod_rel_qual	
prod_rel_recom	

Table 9: Measures of product quality

SUBSTITUTABILITY Variables in this group describe the availability/accessibility and, respectively, the quality of substitutes. Intuition suggests that the number of products offered in the same subcategory

egory seems to be the best measure for the availability of substitutes. The problem with the market size of substitutes (measured by counting the total number of clicks in the subsubcategory minus the clicks received by the respective product) is its high correlation with other substitutability variables. Other variables contain measures like the total number of products in a subsubcategory or the number of products in the same subsubcategory which also have received any clicks. Furthermore variables corresponding to information about retailers represent the accessibility of substitutes.

VARIABLE	DESCRIPTION
ssk_num_offeredprds	Number of products offered in the current subsubcategory during the week of observation
prod_subst_msize	See appendix pages 117 and 116
ssk_numtotclks	
ssk_numprods	
ssk_numretailer	
ssk_numclickedprobs	
ssk_quality	
ssk_recommendation	
ssk_qual_samplesize	
ssk_qual_miss	

Table 10: Measures of product substitutability

BRAND DUMMIES The variables in this group are dummy variables which split the products up in different groups according to their brand rank. The brand rank is a ranking of brand names in accordance to the number of clicks that products of the specific brand have received¹. Preliminary regressions have shown that both of the top brand variables are statistically significant. Since the primary focus of this variable type is to find out whether it pays off to build up a brand (i.e. reduce the elasticity of demand) this thesis will just look at top brands and use the rest as the base group. This will be achieved by generating a new variable indicating whether a product belongs to a top 20 brand. A list of the brand dummies can be found in table [11](#) on page [50](#).

CATEGORY DUMMIES The variables in this section indicate to which category a product belongs to. Roughly 70% of all products in the dataset belong to the group *Hardware*, the second largest group is *VideoFotoTV* which accounts for roughly 15% of the products. For this reason in a first step we only include dummies for the two largest groups and combine the remaining categories into the base group. The list of category dummies can be found in table [12](#) on page [50](#).

MISCELLANEOUS The variables in this group did not fit into the other groups but are also not enough to form any further cohesive groups. The retailer variable of this group is included into the minimal model specification for technical reasons. In order to control for the difference

¹ More information on the brand rank and also how it is computed can be found in the appendix on page [171](#).

VARIABLE	DESCRIPTION
brand1to10	1 if the brand rank is between 1 and 10, else 0
brand11to20	1 if the brand rank is between 11 and 20, else 0
prod_brandrank	The rank of the brand as an integer value
brand21to30	See appendix page 117
brand31to40	
brand41to50	
brand51to70	
brand71to100	
nobrandatall	

Table 11: Dummies indicating the brand rank of a product

VARIABLE	DESCRIPTION
cat4_hardware	1 if the subsubcategory belongs to the category <i>Hardware</i> , else 0
cat9_videofototv	1 if the subsubcategory belongs to the category <i>Video/Foto/TV</i>
cat1_audiohifi	See appendix page 117
cat2_films	
cat3_games	
cat5_household	
cat6_software	
cat7_sports	
cat8_telephoneco	

Table 12: Dummies indicating the category for a product

between necessities and luxury goods and to control for systematic differences in prices, one can include the variable `prod_avgprice`. Finally one should also include the number of ratings from the subsubcategory as a measure of general consumer awareness. The remaining variables can be found in table 13 on page 50.

VARIABLE	DESCRIPTION
prod_numretailer	Share of users who recommend the current product
prod_avgprice	Average price of the product during the week of observation
ssk_numratings	Number of ratings for products in the current subsubcategory
prod_numratings	See appendix page 117
prod_numclicks	

Table 13: Miscellaneous variables

To sum up, up to this point this section has established a list of variables which can be used as proxies for factors which influence the price elasticity of demand like e.g. the substitutability or the level of information on the market. Before one can start with the actual second

stage regressions, one has to deal with two further topics. The first topic is the question of regression weights, the second topic is the formulation of hypothesis on how the presented variables influence the price elasticity of demand. The choice of a suitable regression weight is a rather technical topic, which is not vital for the understanding of the results from the second stage regressions and might therefore not be interesting to all readers. In this case the reader can skip the following subsection and proceed with section 3.3 on page 53 and notice that \bar{R}^2 will be used in the form of an analytic weight.

\bar{R}^2 will be used in form of an analytic weight.

3.2.1 An Excursion On Regression-Weights

Since the results from the first stage regressions are reported including variables, which can be interpreted as quality indicators for the respective regressions, one can incorporate these indicators as a weight for the observations used in the regressions of the second stage. The four potential weights are the *number of observations*, the *root-mean-squared-error*, R^2 and \bar{R}^2 .

Upfront one cannot tell which of these weights is suited best and therefore the adequacy of the respective weighting measures has to be tested. In addition, one has to decide on how to incorporate the weights into the regression. Looking at the help file of Stata's `reg` command one obtains the following possible weighting-options (all taken from the Stata Online Help²):

FWEIGHTS: *Frequency weights indicate replicated data. The weight tells the command how many observations each observation really represents. fweights allow data to be stored more parsimoniously. The weighting variable contains positive integers. The result of the command is the same as if you duplicated each observation however many times and then ran the command unweighted. From this description, one can conclude, that fweights do not represent an adequate weighting type for our purposes.*

PWEIGHTS: *Sampling pweights indicate the inverse of the probability that this observation was sampled. Commands that allow pweights typically provide a cluster() option. These can be combined to produce estimates for unstratified cluster-sampled data. Since the observations in the second stage dataset do not differ in the chance of being included in the sample, pweights do not represent an adequate weighting type for our purposes.*

AWEIGHTS: *Analytic aweights are typically appropriate when you are dealing with data containing averages. For instance, you have average income and average characteristics on a group of people. The weighting variable contains the number of persons over which the average was calculated (or a number proportional to that amount). Those weights that are inversely proportional to the variance of an observation; i.e., the variance of the j th observation is assumed to be σ^2/w_j , where w_j are the weights. Typically, the observations represent averages and the weights are the number of elements that gave rise to the average. For most Stata commands, the recorded scale of aweights is irrelevant; Stata internally rescales them to sum to N , the number of observations in your data, when it uses them.*

² <http://www.stata.com/help.cgi?weight>.

IWEIGHTS: *This weight has no formal statistical definition and is a catch-all category. The weight somehow reflects the importance of the observation and any command that supports such weights will define exactly how such weights are treated. Furthermore the Stata User Guide (Stata 8 [U] 23.16.4) states: "iweights are treated much like aweights except that they are not normalized".*

The above descriptions suggest that *awweights* seem to be the correct choice. Nevertheless the question which variable should be used as a weight still remains. In order to answer this question the following list gives a short explanation of the potential weights:

NUMBER OF OBSERVATIONS: As shown in chapter 2 in figure 7 on page 27 the variation of the estimated elasticities depends on the number of observations which have been used to estimate the respective elasticity. Elasticities based on a larger number of observations seem to be more reliable. Hence it seems reasonable that those elasticities are assigned with a greater weight.

ROOT-MEAN-SQUARED-ERROR (RMSE): The RMSE is the standard deviation of the residuals or as expressed in a formal way by Davidson and MacKinnon (2004) by taking the root of $MSE(\hat{\beta}) \equiv E((\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)')$, where $(\hat{\beta})$ denotes a vector of the estimated coefficients and β_0 denotes a vector containing the true population values of the coefficients). The RMSE is an absolute measure of fit and therefore depends on the size of the residuals and also on the size of the coefficients. This can be easily seen when carrying out a regression in Stata which contains a single explanatory variable. Because of this drawback the RMSE is not suitable as a regression-weight, since it would favor elasticities with values close to zero.

COEFFICIENT OF DETERMINATION: A definition of the coefficient of determination, or also called goodness of fit, is given by Wooldridge (2003, p. 81) "as the proportion of the sample variation in y_i that is explained by the OLS regression line". Therefore a R^2 of 1 means that the regression line explains all of the variation in the explained variable, a R^2 of 0 on the other hand would mean that the regression line has no explanatory power at all. The coefficient of determination can be formally denoted as

$$R^2 = \frac{(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}))^2}{(\sum_{i=1}^n (y_i - \bar{y})^2)(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2)} \quad (3.2)$$

A further important fact is that R^2 usually increases with the number of explaining variables. This could imply that R^2 favors first stage regressions which did not drop any variables (e.g. the dummy variable indicating missing shipping costs if none of the shipping costs were missing for a specific product).

ADJUSTED COEFFICIENT OF DETERMINATION: The main difference between R^2 and its adjusted version \bar{R}^2 is, that the adjusted version penalizes the addition of further independent variables to

a model. A formal definition of \bar{R}^2 is given by Davidson and MacKinnon (2004, p. 117):

$$\bar{R}^2 \equiv 1 - \frac{\frac{1}{n-k} \sum_{t=1}^n \hat{u}_t^2}{\frac{1}{n-1} \sum_{t=1}^n (y_t - \bar{y})^2} \quad (3.3)$$

where u are the residuals, n is the number of observations in the dataset and k is the number of independent variables in the model. Compared to the other potential weights \bar{R}^2 seems to be the most adequate one and will therefore be used as a weight of the second stage regressions.

3.3 HYPOTHESIS

Section 3.2 has established a list of variables suitable for the second stage regressions. The goal of this section is to formulate hypothesis on the sign of the coefficient for each of the presented variables. The structure of this section is that each variable of the minimal model specification will be listed including a short repetitive description and a hypothesis for the sign will be formulated.

PROD_RECOMMENDATION: Share of ratings where the consumer recommends the purchase of the product under observation.

Products which receive a large share of recommendations might feature higher quality than products with a smaller share of recommendations. The demand for high-quality products might be less elastic, because quality may serve as a factor of product differentiation and therefore reduces the substitutability. Furthermore it might be that high-quality products are per se subject to a less elastic demand.

The expected sign of the regression coefficient is +.

SSK_NUMOFFEREDPRODS: The number of products for which there exists at least one offer during the period of observation in the same subsubcategory as the product under observation.

A subsubcategory is expected to host a group of homogeneous products with similar features and properties. Therefore it seems reasonable to expect that one can easily substitute a product in a subsubcategory by a product of the same subsubcategory. Hence this variable measures the number of potentially available substitutes. The higher the availability of substitutes, the higher the absolute value of the price elasticity of demand for a product.

The expected sign of the regression coefficient is −.

BRAND1TO20: This variable is the combination of brand1to10 and brand11to20 and therefore represents the top 20 products according to their brand rank³. The variable is a binary or dummy variable, where a value of 1 indicates that the product under observation has a brand rank of 20 or below. The variable is 0 for products with a brand rank of 21 and above.

The promotion of a brand name is costly and requires the usage of a firm's resources. Since one can observe in the real world that

³ For a more detailed discussion on the brand rank see section D.1 on page 171.

firms try to establish brand names it must be the case that the utility gained from a brand name outweighs the imposed costs. To formulate a hypothesis on how a brand name might influence the price elasticity of demand one can look at the utility gained from a brand name. One benefit of a well established brand is brand loyalty. When a consumer is faced with a purchase decision it might well be that the brand name is a more crucial factor than the price of the product. Therefore one can assume that branded products are subject to a less elastic demand.

The expected sign of the regression coefficient is +.

SSK_NUMRATINGS: The total number of ratings received by all products in the specific subsubcategory.

This variable acts as a proxy for *consumer awareness* or, as termed by Pagoulatos and Sorensen (1986, p. 244), for "information in the market regarding the product and its substitutes". Like mentioned before, information in the market can be seen as a catalyst for the availability and accessibility for substitutes. A large amount of ratings in a subsubcategory indicates that the consumers are well informed about the products of the subsubcategory respectively about the features and properties of the products. Well informed consumers will not only have a better knowledge on features and properties of a specific product, but also on the existence, availability and accessibility of substitutes.

The expected sign of the regression coefficient is –.

PROD_NUMRETAILER: The number of retailers which offer the observed product.

A larger number of retailers implies that there will be more offers for a specific product. Due to the method used for the construction of demand curves this in turn leads to flatter curves and therefore to a more elastic demand. This effect will probably be amplified in the case of demand structures based on aggregated and accumulated clicks.

The expected sign of the regression coefficient is –.

PROD_AVGPRICE: The average price of the observed product. The average price is the unweighted average of all prices of a specific product during the week of observation.

If one assumes that low-price goods can be considered as *necessities* then a rising average price should yield a more elastic demand.

The expected sign of the regression coefficient is –.

It is important to note that one has to be careful with this hypothesis. First of all one has to bear in mind, that the hypothesis is based on the assumption that low-price goods are considered as necessities. It might well be that this is not even true in the case of general goods. This drawback is even more striking in the case of Geizhals. Measured in *Clicks By Category* one can see that the categories *Hardware* and *Video/Photo/TV* account for roughly 75% of all clicks. And it is doubtful that products from this case can be considered as "necessities", even if the product under observation is a low-price product.

In addition Pagoulatos and Sorensen (1986, p. 241) argue that from a theoretical point of view it is not clear that the necessity of a good and its price elasticity of demand are related. They emphasize their argument by stating:

As Friedman (1962, p. 22) points out, if consumers are in equilibrium, thus receiving the same additional utility per dollar for each good purchased, it must be the case that each good is equally necessary or unnecessary.⁴

The aforementioned concerns are valid objections which may render this hypothesis useless.

CAT4_HARDWARE: A dummy variable which takes on the value of 1 for products which do belong to the category *Hardware*.

This variable controls for systematic differences in elasticities between different categories. Literature does not suggest a specific sign for the coefficient of this variable and therefore it is not possible to state any expectations ex ante.

CAT9_VIDEOFOTOTV: A dummy variable which takes on the value of 1 for products which do belong to the category *Video/Foto/TV*.

This variable controls for systematic differences in elasticities between different categories. Literature does not suggest a specific sign for the coefficient of this variable and therefore it is not possible to state any expectations ex ante.

3.4 ESTIMATION RESULTS AND MODEL EXTENSIONS

This section will report and interpret the estimation results of the second stage regressions. The estimation of the second stage is not a one-shot regression but rather an iterative process consisting of consecutive regressions with modifications and enhancements of the regression model between each of the iterations. A fully fledged description of the complete iteration process is not within the scope of this thesis. Therefore this section will only report the most important results and findings on the basis of the full range of elasticities. In addition the section concludes with the presentation and interpretation of the estimation results of the cleansed dataset⁵.

Independent of the dataset, all of the second stage regressions have been carried out in the following three versions:

VERSION 1: This version consists of all observations which feature a negative elasticity that has been estimated by using more than 30 observations.

VERSION 2: Observations from this version fulfill all requirements imposed by version 1, but in addition observations which have been marked as outliers by Grubbs' test will be excluded.

VERSION 3: The requirements of this version are the same as of version 2, but version 3 only contains observations from the category *Hardware*.

⁴ Friedman (1962) refers to: Friedman, M.: Price Theory. Aldine, Chicago, IL, 1962.

⁵ "Cleansed dataset" refers to the reduced dataset with an improved quality as described in section 2.6.5 on page 39.

3.4.1 Estimation Results Of The Regressions Based On The Full Dataset

As mentioned before this section starts with the presentation and interpretation of the second stage results for the full dataset. The complete list of estimates containing all three regression-versions can be found in the appendix on pages 95 to 97.

3.4.1.1 General Results And Conclusions

At this stage, this thesis will not give a detailed report on the exact coefficients, but rather a general overview which is deduced from the estimations of the different regression versions. When comparing the three versions one can notice that there are no striking differences in the estimated coefficients. Neither does the value of the coefficients change by a large amount, nor does the statistical significance of the estimated coefficients change drastically and the signs of the coefficients do change least of all.

Furthermore one should bear in mind that when moving from version 1 to version 2 one should be able to observe improved regression results, since the observations of the regressions in version 2 do not include outliers according to Grubbs' test. This can indeed be verified by comparing the R^2 values of the two regression versions. For each and every type of elasticity the R^2 of version 2 is larger than the R^2 of version one, i.e. the variation in the coefficients of the regressions of version 2 explain a larger proportion in the variation of the elasticities than the coefficients of the regressions of version 1. However, this increase in R^2 cannot be witnessed for every elasticity when moving from version 2 to version 3. Recall that version 3 only uses observations from the category *Hardware*. This in turn implies that the two category dummies *cat4_hardware* and *cat9_videofototv* will be dropped from the regressions. The main reason for the incorporation of version 3 was to observe whether the results differ significantly when only looking on products from the category *Hardware*. However, this has not been the case so far. Therefore one of the conclusions of the general results is that the focus of the second stage regressions should be put on version 2.

The signs of the coefficients are statistically significant but may come up with an unexpected sign.

When comparing the signs of the coefficients between the different types of elasticities (irrespective of the version of the regression) one should notice that they are consistent in the vast majority of the cases, i.e. the estimated coefficients are robust to changes in the type of the observed elasticity. Only in the case of *cat9_videofototv* and *prod_numretailer* one can observe some variation in the signs of the coefficients. Nevertheless especially in the case of *cat9_videofototv* this might not constitute a problem, because as shown in figure 4 on page 11 only 15% of the products belong to this category.

What seems to be rather strange and puzzling is that the coefficients on some of the variables do consistently show an unexpected sign. This is the case for *prod_recommendation*, *ssk_numofferedprods*, *brand1to20* and *prod_numretailer*.

3.4.1.2 Variable-Specific Results And Implications

As already mentioned before, the coefficients on some of the variables come up with unexpected signs. The goal of this subsection is to take a closer look at these variables and to present explanations and possible

solutions for the unexpected signs. This will be done step-by-step for each of the aforementioned variables.

PROD_RECOMMENDATION Contrary to the hypothesis, coefficients on this variable come up with a negative sign, implying that products which have received a larger amount of recommendations are subject to a more elastic demand. However, one has to notice that the coefficient on `prod_recommendation` is not statistically significant for all elasticities. Nevertheless in the cases where the coefficient is statistically significant, it features a negative sign.

A possible explanation for this problem would be erroneous data. If one compares the number of ratings for a specific product from the Geizhals database dump stored at the JKU with the number of ratings for the same product as displayed on the Geizhals website, there are quite a lot of cases where one can notice discrepancies. This results from the fact that an incomplete history of recommendations has been transferred to the JKU.

Furthermore there are products which did not receive any ratings at all. For this reason the missing quality information has been replaced by the average values of the respective quality criteria. The average has been computed by using all products that really received any ratings. In addition, a dummy variable which marks the replaced missing values has been introduced to the dataset. It might be that in this case the replacement of missing values by average values is the wrong thing to do. The share of products which did not receive any ratings is between 30% and 70%, depending on the datafilter and the type of clicks.

Missing product ratings have been replaced by average values.

Nonetheless there is one argument which favors the idea of replacing missing values by average values. It seems reasonable to state that usually consumers will rather rate "extreme" products, i.e. products which are exceedingly good or awfully bad (e.g. concerning the product quality or the product features). If this assumption was true, then one could safely argue that products which did not receive any ratings are average products, i.e. neither top notch, nor complete rubbish. In such a framework it would be perfectly fine to replace missing quality data by average quality data.

A further explanation for the unexpected sign would be that this variable suffers from a form of measurement error. This especially happens if a product has only received few ratings. For a example if three out of four consumers would recommend a specific product its `prod_recommendation` value would be 0.75. If only one further consumer rates and recommends this product the value will suddenly jump to 0.80. So the power of the marginal consumer to influence `prod_recommendation` is declining. To circumvent this problem one could control for the number of product ratings, i.e. include `prod_numratings` into the regression. An analysis of the correlation matrix shows that `prod_numratings` does not heavily correlate with any other of the explaining variables (correlation coefficients range from 0.0008 to 0.23). Therefore multicollinearity should not be a problem.

As a partial solution one should at least control for the products where the missing quality data has been replaced by average values of the quality specific variables, i.e. one should include the variable `prod_avgmissing` into the regressions. This dummy variable indicates if the quality variables of the current product were replaced by average values.

`SSK_NUMOFFEREDPRODS` The coefficients on this variable are statistically significant for the vast majority of elasticities. In the case of version 2 this is even true for all estimated elasticities. Furthermore all statistically significant coefficients are positive. A positive coefficient represents a less elastic demand. Since `ssk_numofferedprods` functions as a proxy for the availability and accessibility of substitutes to positive coefficients are counter-intuitive. From a theoretical point of view there is no reason why products which are exposed to a large number of substitutes should be subject to a less elastic demand.

The subsubcategories might not be a good way do determine the substitutive or complementary relation between products.

One reason why the estimation procedure yielded positive coefficients is that `ssk_numofferedprods` is not a good enough proxy for the availability and accessibility of substitutes. On the one hand it is not the case that all of the offered products have actually received any clicks, i.e. a product listed in the same subsubcategory as the observed product might not be perceived by the consumer as an actual substitute. On the other hand groups of products exist, where the set of possible substitutes is not only limited to the same subsubcategory as the observed product but rather the whole subcategory may contain substitutes. A PC-soundcard and a set of speakers would rather be seen as complements and not substitutes. Nonetheless both of these products are listed on the Geizhals website in the subcategory *pc-audio*. In this case one has indeed to dig deeper into subsubcategories to find substitutes for specific products. This however does not hold true if one takes a look at the subcategory *computer monitors*, which is divided into subsubcategories in accordance to screen size (e.g 21", 22", 24", etc.). In this setup the group of substitutes should rather be set to the level of the subcategory rather than the subsubcategory, because a 22" screen can very well be seen as a substitute for a 24" screen.

In order to test whether `ssk_numofferedprods` does represent the availability and accessibility of substitutes poorly, one should use one of the other substitute-proxies in order to test for their quality. This approach has actually been followed during the course of this diploma thesis, however it turned out that the other proxy variables do not yield better results. Therefore one can stick with `ssk_numofferedprods` as a proxy for substitutes.

`BRAND1TO20` One would have expected a positive sign on the coefficients of this variable, because a brand product should be subject to a less elastic demand. The unexpected sign could be the result of a bad choice of the base group for this variable. Products of the top 20 brands account for roughly 25% of the observations. Since the brand rank has been generated by counting the number of clicks per products it is reasonable to assume that the share of the top 20 brand products is larger in the filtered datasets.

A further potential source of the problem might be the way of how the brand rank respectively the list of brands has been computed. Although names like "No-name" or null-values which obviously do not constitute a brand name have been filtered out, it might very well be the case that the data set still contains erroneous brand names like "Antennensplitter" or "ausblenden"⁶.

Concerning the computation of the brand rank, it could be the case that the total number of clicks received by the products of a brand is not a good enough proxy for the strength of a brand. It seems

⁶ For a more detailed discussion on the brand rank see [D.1](#) on page 171.

reasonable to assume that this total number of clicks emerges from the interaction of the products offered at Geizhals' website and the target group of Geizhals. This can be seen when looking at the size of the various categories listed at the Geizhals website. The vast majority of products listed at Geizhals can be allocated to the category *Hardware*. This indicates that this category will also attract the most consumers or that computer-hardware is the main focus of attention of the primary target group of Geizhals. This would also explain why brands like e.g. *SanDisk*⁷ receive a better brand rank than well renowned German brands like *Bosch* or *Miele*. The same argument is true in the case of *MSI* vs. *Apple*, where *MSI* received the better brand rank. This seems rather counter-intuitive, especially in terms of recognition value. Another curiosity is the brand rank of *Nike*, which can be found in the region above 400.

Finally one cannot be clear about the question whether it is reasonable to squeeze all brands into a single ranking, because the strength of the brand also depends on the context of the category under observation. Furthermore a brand should also be viewed in relation to the strength of its substitutes. If one takes a look at e.g. the brands *Samsonite* or *Kärcher* one should notice that although their brand rank is rather low, they are very renowned and strong brands in their respective subcategory.

Unfortunately, at the moment there are no efficient and effective solutions to address the majority of the problems explained in above paragraphs. However, one can easily alleviate the problem of the too large base group by replacing `brand1to20` by the actual brand rank. When generating the brand rank one must bear in mind that brands which have received the same number of clicks also have to receive the same brand rank.

`PROD_NUMRETAILER` The oddity concerning this variable is the fact that the sign of the coefficients of the elasticities coming from aggregated/accumulated clicks are consistently negative, which is exactly what theory and intuition would suggest. Though in the case of non-aggregated/accumulated clicks the opposite is true, except from `l_elast` and `c_elast_inv`.

The problem with this variable is that there are no obvious reasons for the unexpected sign on the coefficients of the non-aggregated and non-accumulated variables. One attempt of explanation is rooted on technical grounds. The only difference between non-aggregated/accumulated clicks and aggregated/accumulated clicks is the process of aggregation/accumulation. Amongst other dimensions the offers are aggregated/accumulated along their price. Because of this aggregation/accumulation the dimension of the retailer is removed completely. However, this explanation does not give an answer to the question why the coefficients of `prod_numretailer` for `l_elast` and `c_elast_inv` have a negative sign. Since the correlation matrix of the elasticities which is given on page 38 shows that there is hardly any correlation between `c_elast` and `c_elast_inv`, it might be the case that `c_elast_inv` would really be a better representation of the estimated elasticities. Apart from this there are no further approaches to explain or solve the problem of the unexpected sign on the coefficients for this variable.

⁷ A manufacturer which produces flash memory cards.

3.4.1.3 *Enhancing The Minimum Model Specification*

This section shortly sums up the present insights of the second stage regressions and tries to deduce further implications. The minimal model will be modified in accordance with the problems and explanations given for the unexpected signs. One of the modifications is that the variable `prod_avgmissing` will be included in the regression. This should improve the regression outcomes because replaced missing values will be marked explicitly. Furthermore the variable `brand1to20` will be replaced by the actual brand rank (`prod_brandrank`), whereas a rank of 1 represents the strongest brand. This should bring more flexibility concerning the brand into the regression equations. Taking a look at the correlation matrix of the new set of explaining variables shows that multicollinearity is not an issue and therefore the coefficients should not suffer from too low statistical significance.

3.4.2 *Results Of The Enhanced Model*

An extract of the results of the enhanced model estimated using the full dataset can be seen in table 14 on page 61. The results are only presented for a chosen subset of elasticities and also only in form of version 2, i.e. the dataset only contains elasticities which have been estimated by using more than 30 observations and only elasticities which passed Grubbs' test. The full results for all versions can be found in appendix A on pages 101 to 103.

3.4.2.1 *General Results And Implications*

When comparing the R^2 of the enhanced model with the minimal specification one can notice that it has increased across the board. However, in the majority of the cases this increase is very small and the enhanced model contains one additional variable which therefore might explain the increase in R^2 . In addition one can notice that in general the values of the coefficients have only changed by small margins when going from the minimal model to the enhanced model. The results in general are very robust and did not change a lot.

3.4.2.2 *Variable-Specific Results And Implications*

After the previous subsection summed up the most important general results, this section deals with variable-specific results.

PROD_RECOMMENDATION Concerning the introduction of the variable `prod_avgmissing` one can say that it is statistically significant in the vast majority of the cases on a 1%-significance level. Furthermore it does not render `prod_recommendation` as statistically insignificant. However, the coefficients of `prod_recommendation` have decreased in absolute value across the board. Unfortunately, the sign remained negative on all statistically significant coefficients. The coefficients on this variable range from -0.43 to -1.68 , which means that if the share of recommendations increases by 10 percentage points, the elasticity ϵ will fall by 0.043 to 0.168. The empirical results suggest that products with a higher share of recommendations are subject to a more elastic demand.

LHS-variable:	(1)	(2)	(3)	(4)	(5)
	c_elast	aggr- c_elast	l_elast	cv_lctw- c_elast	aggr_lctw- c_elast
prod_recommendation	-1.5407*** (0.2513)	-0.5716 (0.4560)	-1.6789*** (0.3784)	-0.4342** (0.1795)	-1.3748*** (0.4441)
prod_avgmissing	1.6493*** (0.1104)	-1.5398*** (0.2010)	-0.4870*** (0.1618)	0.7079*** (0.0863)	1.1452*** (0.2172)
ssk_numofferedprods	0.0027*** (0.0002)	0.0043*** (0.0003)	0.0013*** (0.0003)	0.0006*** (0.0001)	0.0029*** (0.0004)
prod_brand rank	0.0025*** (0.0002)	0.0064*** (0.0004)	0.0030*** (0.0003)	0.0000 (0.0001)	0.0033*** (0.0003)
ssk_numratings	-0.0068*** (0.0002)	-0.0093*** (0.0004)	-0.0038*** (0.0003)	-0.0019*** (0.0002)	-0.0038*** (0.0005)
prod_numretailer	0.0201*** (0.0016)	-0.0362*** (0.0030)	-0.0365*** (0.0024)	0.0031** (0.0012)	-0.0414*** (0.0029)
prod_avgprice	-0.0013*** (0.0001)	-0.0025*** (0.0001)	-0.0021*** (0.0001)	-0.0001*** (0.0000)	-0.0013*** (0.0002)
cat4_hardware	1.4144*** (0.1485)	2.2632*** (0.2864)	-0.0080 (0.2153)	1.1066*** (0.1082)	2.4189*** (0.3010)
cat9_videofototv	-0.6380*** (0.1759)	2.8494*** (0.3690)	-0.3599 (0.2543)	-0.1078 (0.1295)	0.0082 (0.3827)
Constant	-9.7508*** (0.2607)	-17.240*** (0.4978)	-9.547*** (0.3902)	-2.9084*** (0.1844)	-10.648*** (0.4903)
Observations	12216	14710	11512	3349	5316
R ²	0.2148	0.1108	0.1143	0.1639	0.1478

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 14: Enhanced model, full dataset regression results - version 2

SSK_NUMOFFEREDPRODS Even in the enhanced model the coefficient on this variable remained positive across the board. Therefore it seems that one has to conclude that `ssk_numofferedprods` is not a suitable proxy for the availability and accessibility of substitutes. Nonetheless the results suggest that an additional product in a subsubcategory increases the elasticity by 0.0006 to 0.0043.

PROD_BRANDRANK When looking at the coefficients of the variable `prod_brandrank` one can immediately see that the coefficients have decreased in absolute value. This is due to the measurement of this variable. The `brand1to20` indicated whether a product belongs to a brand with a brand rank of 20 or below, therefore the coefficient on this variable showed the difference in elasticities of top20 brand products and products which do not belong to a top20 brand. `prod_brandrank` on the other hand directly shows the rank of the brand of a product. Therefore the coefficient on `prod_brandrank` shows how the elasticity changes when moving from a product with rank x to a product with rank $x + 1$.

Even in the enhanced model which directly uses the brand rank, the conclusion of the second stage regression is that brand products are subject to a more elastic demand. However, this statement depends on the assumption that the brand rank is indeed a good measure of the strength of a brand. Unfortunately, in the context of this thesis no alternative measures for brand strength are available.

Although the reported coefficients of this variable feature an unexpected sign, it might be explained on economical grounds. Economic literature suggests that one of the incentives for firms to engage in advertisement and the establishment of a brand name is that it reduces the substitutability of a good. In turn standard economic textbooks like [Varian \(2001\)](#) state that a low substitutability should imply a less elastic demand. However, when working with data from Geizhals one does not look at competition between producers but rather at competition between retailers. Let us assume that a brand product X does not have any substitutes at all. Furthermore assume that there are n retailers offering X. This setup can also be viewed as a situation where n retailers offer completely homogeneous products of type X. Since the products of type X are assumed to be completely homogeneous they cannot be differentiated by their product characteristics. The only two remaining dimensions for differentiation are the price of the product and retailer specific characteristics. A premium price could e.g. be justified if a retailer warrants a 24 hour delivery. Bear in mind that a comparison between the coefficients of the variables `c_elast` and `cv_c_elast` has shown that controlling for retailer specific data has no significant impact on the estimated elasticities. This suggests that the price of a product is the most important factor in the determination of the demand, which in turn could imply that consumers could react more price sensitive. So if a renowned brand name implies that it is hard to find a substitute for a product of that brand, the retailers competing for their market share within the context of this product might be subject to a more elastic demand.

In the case of Geizhals one focuses at the competition on the retailer-side of the market and not the producer-side.

`PROD_NUMRETAILER` Unfortunately, the signs remain the same in the enhanced model. Nonetheless the coefficient on `prod_numretailer` is negative for the majority of the elasticities. `c_elast`, `c_elast_choke`, `cv_c_elast` and `cv_lctw_c_elast` are the only elasticities which come up with a positive coefficient on `prod_numretailer`. This indeed indicates that the hypothesis that a larger number of retailers induces further competition among retailers and therefore increases the absolute value of elasticity of demand is actually true.

The results show that the coefficient on `prod_numretailer` ranges from -0.008 to -0.1771 , which means that *ceteris paribus* a further retailer who offers a specific product decreases the price elasticity of demand of that specific product by 0.008 to 0.1771 .

3.4.2.3 *Putting The Results Into Context By Using The Standard Deviation*

The previous sections reported the estimation results of the full dataset. This section tries to put these results into a context by computing the impact for each explanatory variable if the respective variable changes by one standard deviation. The mean and the standard deviation of the respective variables depend on the elasticity under observation because the sample size of the dataset differs between the respective elasticities. Therefore the standard deviation of the explanatory variables is

computed by using the example of `c_elast`. The results are shown in table 15, which reports the mean and the standard deviation for each explanatory variable in the first two columns. The third column lists the estimated coefficient for each variable. The final column shows how `c_elast` changes, if the explanatory variable changes by one standard deviation, hence this column features $\beta_i * \text{stddev}_i$.

Variable	Mean	Std. Dev.	Coeff.	$ \Delta c_{\text{elast}} $
prod_recommendation	0.699	0.201	-1.5407	0.31
prod_avgmissing	0.62	0.485	1.649	0.7998
ssk_numofferedprods	381.123	306.95	0.0027	0.8288
prod_brandrank	140.545	274.936	0.0025	0.6873
ssk_numratings	146.046	235.408	-0.0068	1.6008
prod_numretailer	61.86	33.937	0.0201	0.6821
prod_avgprice	359.476	810.888	-0.0013	1.0542

Table 15: Impact on `c_elast` if an explanatory variable changes by one standard deviation (N=14814)

The numbers in table 15 show that the resulting impacts seem to be rather small. However, if one accounts for the fact that the mean of `c_elast` is -4.450 and its median is -2.73 , one cannot deny that economical significance is on hand. Expressed as a percentage-change, the impact of `prod_recommendation` (0.31) implies a change of `c_elast` with regard to its mean by 6.97% and with regard to its median by 11.36%.

3.4.3 Estimation Results For The Cleansed Dataset

The previous sections reported the results for the regressions which are based on the full range of products. This section deals with the reduced and cleansed dataset. Recall that the criteria for the quality improving reduction of the dataset are the number of observations which have been used to estimate the respective elasticity and the share of zero-click-offers of the respective product. The thresholds for these criteria have been set to a minimum of 31 observations and a maximum share of zero-click-offers of 30%, which reduces the dataset to 202 observations⁸.

The estimation results for the cleansed datasets are given in tables 35 to 37 on pages 98 to 100. Though the main point of interest is table 36 on page 99, since the quality of the estimation results of regression-version 2 is better than the quality of the results of regression-version 1. Furthermore the cleansed dataset only contains a rather small number of observations, which has a negative impact on the statistical significance of the estimated coefficients. Therefore it is not surprising that version 3 hardly delivers any statistical significant results.

⁸ The actual number of observations used in the regressions are even less, since there are still elasticities with a value of 0 or more. Furthermore there are observations with a negative \bar{R}^2 . The problem linked to a negative \bar{R}^2 is that stata requires an analytic weight to be a positive number. Observations with negative analytic weights will therefore be dropped from the estimation procedure.

3.4.3.1 General Results And Interpretations

One of the most striking features of the result table of the cleansed dataset is that compared to the previous result tables the proportion of statistically insignificant coefficients is much higher. This is especially true for the category dummies and the variables `ssk_numofferedprods`, `ssk_numratings` and `prod_avgprice`. However, this does not have to be a reason to worry, since e.g. Davidson and MacKinnon (2004, p. 101) explain that the variance of the estimated coefficients and therefore also their statistical significance partly depends on the sample size used for the regression. They state the variance of the coefficients as given in equation 3.4, where n denotes the sample size and σ_0^2 denotes the true variance of the error terms.

$$\text{Var}(\hat{\beta}) = \left(\frac{1}{n}\sigma_0^2\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \quad (3.4)$$

An increase in sample size leads to a proportional increase in the variance of the estimated coefficients.

In the context of this equation Davidson and MacKinnon (2004, p. 101) state that "the second factor on the right-hand side [...] does not vary much with the sample size n , at least not if n is reasonably large". Since the true variance of the error terms is constant, the first factor of the right-hand side and therefore the complete right-hand side is proportional to $1/n$. The number of observations of the regressions based on the full dataset ranges from roughly 8200 to roughly 12600⁹. In the case of the cleansed dataset these numbers decrease to 137 and 178. This reduction in the sample size leads to a proportional increase in the variance of the estimated coefficients in accordance to equation 3.4. Apart from the aforementioned variables the majority of the estimated coefficients is statistically significant on a 1%-level.

A decrease of the sample size may improve R^2 .

In addition, the increased R^2 -values cannot go unnoticed. Although this increase is not unexpected, the R^2 -values of the cleansed dataset results seem to be exceptionally high. On one hand the cleansed dataset has been generated by applying data filters to the complete dataset with the aim to improve the data quality at the expense of the sample size. Under the assumption that the cleansed dataset really features an improved data quality, an increase in R^2 seems to be perfectly reasonable. On the other hand it might be that the increase of R^2 is partly due to the reduction in the sample size. The phenomenon of a decrease in R^2 stemming from a smaller data sample is explained by Cornell and Berger (1986, p. 65). They argue that the replication of different values of y for one or more values of x can in fact reduce the coefficient of determination. To see this one has to take a look at the definition of the coefficient of determination which is given in equation 3.5.

$$R^2 \equiv 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.5)$$

The nominator in the right-most term is the regression sum of squares, whereas the denominator is the total sum of squares. A replication of y -values for one or more x -values has no effect on the predicted

⁹ These numbers apply to the regressions based on normal clicks. The numbers for the LCT are lower.

values, i.e. the nominator of the right-most part of 3.5 remains the same. The denominator, however, does increase, because although the replicated values do not change \hat{y} they do have an arbitrary distance to \bar{y} . Finally this effect results in a reduction of R^2 based on an increase in the sample size.

3.4.3.2 Variable-Specific Results And Interpretations

The variable-specific results show some interesting changes compared to the full dataset. First and foremost one can notice that all statistically significant coefficients of the variable `prod_recommendation` have a positive sign. So with the cleansed sample the sign of the estimated coefficient complies with the expected sign from the hypothesis. These results suggest that products that receive a larger proportion of recommendations are indeed subject to a less elastic demand. Recall that the proportion of recommendations in this case is a proxy variable for the quality of a product and therefore for its substitutability. One should still keep in mind that a large share of users, who recommend the purchase of a product, does not only imply a premium quality per se, but it could also be a reference to other properties which could somehow be subsumed under the term "quality" like e.g. a good value for money. A further peculiarity concerning `prod_recommendation` is the size of the coefficients, which are quite high, especially `c_elast` (11.64) and `c_elast_choke` (13.05).

The variable `ssk_numofferedprods` is statistically insignificant across the board, except for `c_elast_inv` and `cv_l_elast`, where the coefficient on the latter features a negative sign. However, both variables are only significant on a 5% level. Nonetheless this could be interpreted as a small hint that `ssk_numofferedprods` does indeed represent the availability of substitutes.

Even though the coefficients of the cleansed dataset suffer from a reduced statistical significance this has hardly any impact on the significance of the coefficients of `prod_brandrank`. Unfortunately, the sign is still positive and thus the opposite to what the hypothesis suggests. A further remarkable result is that all of the statistical significant coefficients on `prod_numretailer` are negative, suggesting that a further retailer does indeed induce a more elastic demand.

To conclude this subsection one can say that the reduction of the dataset has indeed improved the quality as some of the coefficients now show the sign which theory would suggest.

3.4.3.3 Putting The Results Into Context By Using The Standard Deviation

Using the example of `c_elast` this section shows how the elasticity changes if one changes the explanatory variables by their standard deviation. The results are shown in table 16, which reports the mean and the standard deviation for each explanatory variable in the first two columns. The third column lists the estimated coefficient for each variable. The final column shows how `c_elast` changes, if the explanatory variable changes by one standard deviation, hence this column features $\beta_i * stddev_i$.

One can notice that on average the coefficients in table 16 are larger than the coefficients in table 15. However one has to consider the fact that the mean and median of `c_elast` have also increased (measured in absolute values) to -10.715 respectively -9.421 . Furthermore table 16

Variable	Mean	Std. Dev.	Coeff.	$ \Delta c_{\text{elast}} $
prod_recommendation	0.745	0.153	11.6379	1.7806
prod_avgmissing	0.275	0.448	-2.0752	0.9297
ssk_numofferedprods	420.101	281.263	0.0017	0.4781
prod_brandrank	207.915	442.981	0.0073	3.2338
ssk_numratings	359.317	314.253	-0.0010	0.3142
prod_numretailer	34.656	17.856	-0.1337	2.3873
prod_avgprice	656.313	672.956	-0.0001	0.0673

Table 16: Impact on c_{elast} if an explanatory variable changes by one standard deviation (N=189)

does not only show that the coefficient of `prod_recommendation` now features the expected sign, but it is also highly significant on economical grounds. A change in `prod_recommendation` by one standard deviation implies a change of c_{elast} in respect of the mean by 16.56% and in respect of the median by 18.9%

3.5 CONCLUSION AND REMAINING PROBLEMS

This chapter has established a dataset with variables that emulate factors which are supposed to influence the price elasticity of demand. Such factors are e.g. the accessibility and availability of substitutes or the consumers' awareness. The estimation results of this stage show that there indeed exists statistical evidence that the price elasticity of demand is not solely determined exogenously in form of consumers' tastes. However, the question remains whether the factors identified during this section are triggered through the channel of retailer-competition or whether they really represent means of how producers can systematically influence the price elasticity of demand.

The variables `prod_recommendation` and `prod_brandrank` constitute an example of the aforementioned argument. The former variable features negative coefficients across the board in the case of the full dataset. Only in the case of the reduced, cleansed dataset the signs of the estimated coefficients comply with the expected signs from the established hypothesis. This either accentuates the need for [IV](#) and [2SLS](#) or it shows that this variable influences the retailer side of the competition to at least some extent. The reason for the importance of the retailer-competition has been argued along with the explanation why `prod_brandrank` might yield an unexpected sign on page [62](#). The idea of this argument is that product differentiation reduces the homogeneity and therefore the substitutability of products which might turn the retailer side of the competition into a "winner-takes-it-all" game. Since it has been argued that the quality of a product can also be interpreted as a way of product differentiation and therefore also as a factor which reduces the substitutability, it could also be the case that `prod_recommendation` will in fact yield a series of negative coefficients.

Another unexpected result is related to `ssk_numofferedprods`, which should be a measure for the availability of substitutes. Economic theory clearly states that a larger number of substitutes leads to a more elastic demand. However, the estimation results show that an increase in the

listed number of products in a specific subsubcategory decreases the absolute value of the price elasticity of demand for products in the respective subsubcategory. These results are a rather strong evidence for either a poor quality of the variable or a poor quality of the complete dataset in general (even in the case of the cleansed dataset), i.e. that either `skk_numofferedprods` is in fact not a good representation of the availability of substitutes or that the dataset in general suffers from impurities.

Concerning the quality of the cleansed dataset one has to ask whether the chosen criteria and also the chosen thresholds for the respective criteria were a good decision and whether the quality of the dataset has really improved because of this filter. Especially if one only focuses on the number of observations used to estimate the respective elasticity and the share of zero-click-offers, it might turn out that one still gets poor results. If either the criteria or the thresholds were a bad choice the following two things could happen:

- The filter fails to exclude a poor or even corrupt observation.
- The filter excludes a good observation which in fact should be part of the cleansed dataset.

The former problem is recognizable in two ways. First of all neither the number of observations used to estimate an elasticity nor the share of zero-click-offers ensure that the estimated elasticity is, as suggested by standard economic theory, negative. Though in the first place one cannot systematically rule out positive demand curves, hence such cases require further inspection. An example of such a product is given in figure 16, which depicts the demand curve for a *Manfrotto* tripod. To make the graph more readable its axes are denoted in logarithmic form. The graph shows a rather unsystematic scatterplot of the relation between price and demand, which finally results in positively sloped demand price curves. However, manual inspection of figure 16 shows that it seems rather implausible that the true demand for this product increases with its price.

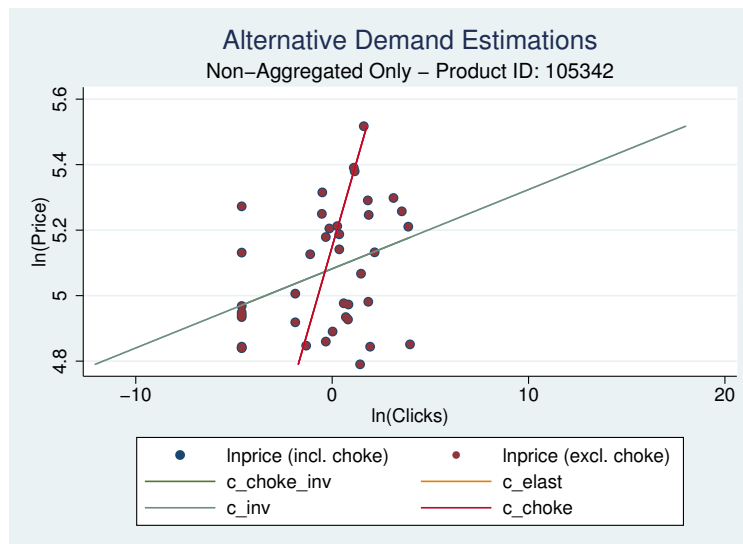


Figure 16: Example 1 of a problematic observation from the cleansed dataset (Product: Manfrotto tripod)

Secondly, the wedge between c_elast and c_elast_inv is also an indicator of poor price/click data for a product. As it has already been mentioned before the coefficient on x coming from a regression of y on x should be the exact inverse of the coefficient on y coming from the regression x on y . This will only not hold true in the case of poor data. Therefore it seems reasonable to assume that a large wedge between those two elasticities is an indicator of poor data quality. This assumption can be verified by looking at figure 17, which shows the scatterplot for a *Tamron 55-200mm* lens that actually passed the cleansing process. One should notice from figure 17 that even a small number of unlucky zero-click-offers can have a large impact on the final outcome. The scatterplot in figure 17 features a rather rectangular shape,

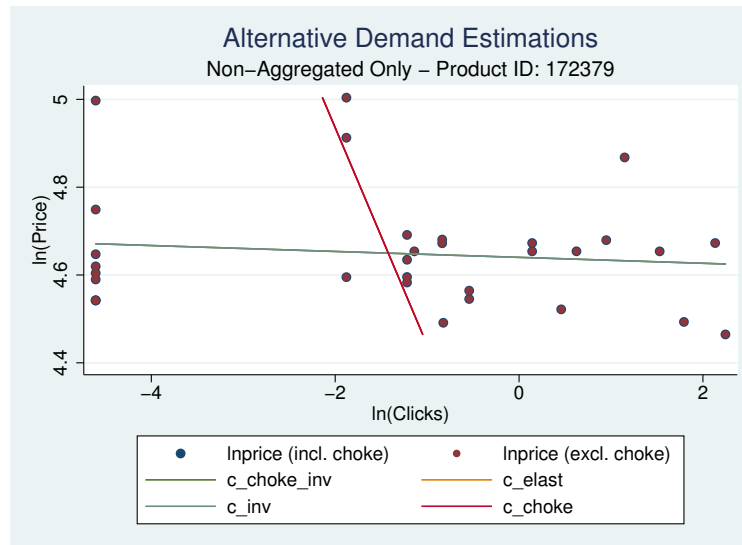


Figure 17: Example 2 of a problematic observation from the cleansed dataset (Product: Tamron 55-200mm lens)

preventing OLS from estimating a meaningful demand curve. Since the outcome of this estimation is hugely dependent on the direction of the regression, it is not surprising if the elasticities yield poor results in the course of the second stage regressions.

As mentioned before the second source for a poor reduced dataset is that the cleansing process excludes elasticities which in fact seem to be useful. It would not be target-oriented to draw scatterplots for the complete range of roughly 40,000 products, therefore these plots have only been drawn for the 202 products of the cleansed dataset. Hence there is no graphical presentation which would enable the user to manually inspect the data in order to find suitable observations which have been excluded because of their share of zero-click-offers or their number of observations in the first stage. Nonetheless the cleansed dataset also contains observations which did not make it into the final regression dataset, namely those that have been ruled out by Grubbs' test. An example of such a product, a HP Pavilion, is given in figure 18.

The wedge between the demand curve coming from a regression of price on clicks and the demand curve coming from the regression of clicks on price is rather small. Manual inspection also shows that the scatterplot for this product does actually have the looks and structure

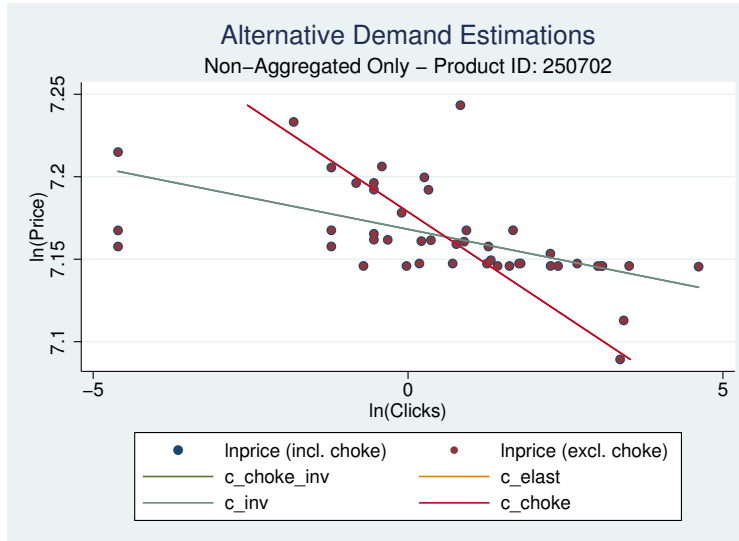


Figure 18: Example of a suitable demand curve which has been ruled out by Grubbs' test (Product HP Pavilion notebook)

of a negatively shaped linear curve¹⁰. However, the scatterplot is rather flat leading to a quite elastic demand with a c_elast of -39.44 and a c_elast_inv of -131.12 . Although the estimated demand curves for this product seem to be perfectly reliable from a graphical point of view, the absolute values of their elasticities are extraordinarily high compared to the elasticities from other products. The noticeable difference between the absolute values of the elasticities for this product and the absolute values of the elasticities for the other products is the reason why this product does not pass Grubbs' test and will therefore be excluded from any second stage regression.

Despite the problems described in the previous paragraphs, the estimation results still provide enough evidence to state that the majority of the formulated hypothesis turn out to be true. Especially the coefficient on the variable $ssk_numratings$, which represents the consumer awareness, is negative across the board, irrespective of the type of elasticity or the dataset used for the second stage regression. The same argument is true for $prod_numretailer$, which does come up with some unexpected signs in the case of the complete dataset. However, in the case of the reduced, cleansed dataset the coefficients are consistently negative, implying that a larger number of retailers does lead to a more elastic demand.

To conclude this chapter one can say that the examinations and analysis are definitely on the right track. Nonetheless further work needs to be done to achieve more significant and meaningful results. First of all one has to improve the quality of the elasticities estimated in the first stage. The previous paragraphs have shown that there is still a lot of room for improvement concerning the quality criteria of the estimated elasticities. One potential approach for an alternative quality criterion is presented in the appendix on page 173. Furthermore the results suggest that the explanatory variables are not perfect proxies for the underlying factors which are supposed to influence the price

¹⁰ One has to notice that the axes in figure 18 are denoted in logs, which implies that the final demand curve has an isoelastic and not a linear structure.

elasticity of demand. The most obvious case for a blurry proxy variable is `prod_brandrank`. However, these remaining points go beyond the scope of this thesis and their treatment has to be delayed to any further work on this topic.

THE PRICE ELASTICITY OF DEMAND AS A DETERMINANT OF MARKET STRUCTURE

Similar to chapter 3 the work in this chapter is also based on the elasticities which have been estimated in the course of chapter 2. However, this chapter focuses on the role of the price elasticity of demand as a determinant of market structure, where *market structure* refers to the type of competition, i.e. Bertrand or Cournot competition. In Bertrand competition *price* is the strategic variable from the firm's point of view, in Cournot competition *quantity* is the strategic variable. Or simply spoken, in Bertrand competition firms set the price of the product and in Cournot competition they set the quantity which will be produced.

4.1 INTRODUCTION

The idea for this chapter stems from the work of Caves (1964) found in Johnson and Helmberger (1967, p. 1218). Caves argues that in an elastic market it is more likely that a firm engages in price cutting in order to increase its market share, because total industry sales will also rise, even if the remaining firms also choose to reduce their prices. This is not true in the case of an inelastic demand curve. If multiple firms engage in price cutting in an inelastic setup, they all might end up selling almost the same output but just at a lower price. The goal of this chapter is to analyze the impact of the price elasticity of demand on the type and intensity of competition.

An elastic demand provides incentives for undercutting.

To measure the type of competition this thesis uses the number of changes of the price leader of a certain product. The number of changes of the price leader is an indicator for the toughness of the competition on a market. More frequent changes of the price leader could either mean that retailers try to undercut their rivals, or it means that price leaders can only reduce their price for a limited time period. After this period their cost structure forces them to increase the price again, leaving another retailer as the price leader. This number is computed for the complete range of offers from the Geizhals DB and not only for the week of observation. To compute this number one has to split the complete timespan into several intervals. These intervals are determined by using the start- and end-timestamps of all offers from the Geizhals DB.

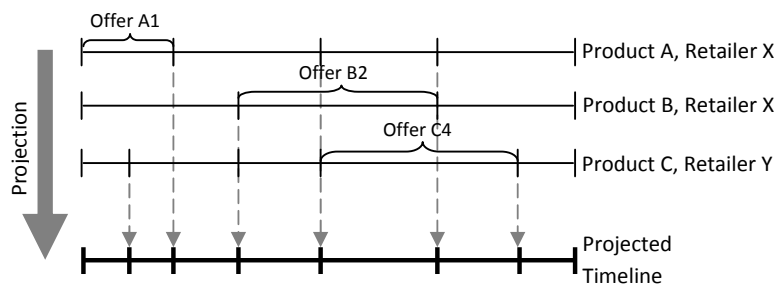


Figure 19: Projection of the start- and end-timestamps of offers

The process of determining the final intervals is illustrated in figure 19. Basically the script which determines the final intervals takes each and every offer from the *offer*-table of the Geizhals DB and projects them onto a new time line. This process also removes duplicate start- or end-timestamps. At the end of this script the new time line contains every possible time-interval. The thick line in figure 19 depicts this new time line and also shows the offers.

In a next step an additional script iterates through all time-intervals and determines the price leader for each and every product for the actual time-interval. The result of this step is a list for each product, which contains the price leaders in a chronological order. In a final step one has to iterate through those chronologically ordered lists and count the changes of the price leader of the respective product.

There are four changes of the price leader for product A.

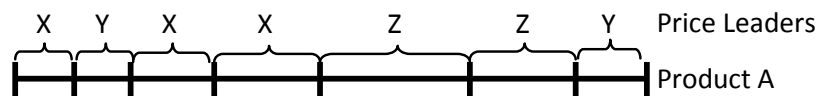


Figure 20: Time line with price leaders

Figure 20 shows the time line for product A and the price leaders X, Y and Z. Recall that the variable to be determined is the number of changes of the price leader and not the number of distinct price leaders. So when looking at the first three intervals one should notice that there are two changes of the price leader, namely the change from retailer X to retailer Y, between the first and the second interval, and also the change from Y back to X between the second and the third interval. In addition, there is no change of the price leader between the third and the fourth interval. Therefore the number of changes of the price leader for product A is 4.

4.2 DATASET DESCRIPTION

The dataset is very similar to the dataset of the previous chapter, i.e. it also consists of category- and product-specific data. Furthermore the variables have been condensed to exactly the same groups, namely *product quality*, *substitutability*, *brand dummies*, *category dummies* and *miscellaneous*. The following paragraphs will again give a short explanation for each group and they will also list the variables of the specific groups.

This chapter features a further similarity to the previous chapter - it also starts with the specification of a minimal model checking the relationship between the elasticity of a product and the toughness of competition. Therefore the first estimation will simply be carried out by regressing the number of price leader changes on the respective elasticity.

PRODUCT QUALITY The variables of this group represent different measurements of product quality. Amongst other measurements this group covers the absolute product quality and the quality in relation to the average quality of the remaining products of the respective subsub-category. Furthermore this group also contains variables describing the share of users which would recommend the product under observation. However, previous work suggests that the recommendation variable dis-

tributes more to the explanation of elasticities than the other measures of product quality. In addition, relative variables, i.e. variables which put a property of a product like e.g. its quality in relation to the average property value of the remaining products of the subsubcategory, tend to be correlated with variables from the group *substitutability*.

VARIABLE	DESCRIPTION
prod_recommendation	Share of users who recommend the current product
prod_avgmissing	1 if the current product's quality data is missing
prod_quality	See appendix page 117
prod_rel_qual	
prod_rel_recom	

Table 17: Measures of product quality

SUBSTITUTABILITY Compared to the previous chapter where the elasticities have been used as the LHS-variable, substitutes are not that much of a big point when looking at price leader changes. In this case it is more interesting to look at the factors, which attract new retailers or provoke incumbents to undercut their rivals. The main factor would be the attractiveness of the market (i.e. the size), which would rather be measured in clicks than in number of substitutes.

Substitutes are less important in the context of price leader changes.

VARIABLE	DESCRIPTION
prod_subst_msize	Size of the market for substitutes measured as the total number of clicks received by products of the current subsubcategory but with exclusion of clicks of the observed product. The variable is given in 1000 clicks.
ssk_num_offeredprds	Number of products offered in the current subsubcategory during the week of observation
ssk_numclickedprds	Number of products in the current subsubcategory which have received any clicks during the period of observation
ssk_numtotclks	See appendix pages 117 and 116
ssk_numprods	
ssk_numretailer	
ssk_quality	
ssk_recommendation	
ssk_qual_samplesize	
ssk_qual_miss	

Table 18: Measures of product substitutability

BRAND DUMMIES The variables in this group are dummy variables which split the products in different groups according to their brand rank. The brand rank is a ranking of brand names in accordance to

the number of clicks that products of the specific brand have received¹. However, these variables seem to be less important in the context of this chapter. Nevertheless there might be systematic differences between important brands and rather unknown ones. The results from the previous chapter have shown that the variable `prod_brandrank` offers more flexibility than the dummy variables, without suffering from any major drawbacks. For this reason, the regressions in this chapter will also use `prod_brandrank`.

VARIABLE	DESCRIPTION
prod_brandrank	The rank of the brand divided by 100
brand1to10	See appendix page 117
brand11to20	
brand21to30	
brand31to40	
brand41to50	
brand51to70	
brand71to100	
nobrandatall	

Table 19: Dummies indicating the brand rank of a product

CATEGORY DUMMIES Similar to the previous chapter these variables control for systematic differences between the categories. Additionally the regressions of this stage will only include the variables `cat4_hardware` and `cat9_videofototv` because *Hardware* and *VideoFotoTV* are by far the largest categories.

VARIABLE	DESCRIPTION
cat4_hardware	1 if the subsubcategory belongs to the category <i>Hardware</i> , else 0
cat9_videofototv	1 if the subsubcategory belongs to the category <i>Video/Foto/TV</i>
cat1_audiohifi	See appendix page 117
cat2_films	
cat3_games	
cat5_household	
cat6_software	
cat7_sports	
cat8_telephoneco	

Table 20: Dummies indicating the category for a product

MISCELLANEOUS The number of clicks is a representation of the market size for the current product and therefore included in the regression. The number of retailers offer the current product is an indicator for the toughness of competition. The price of the product is an indicator for the size of the profit margin.

¹ More information on the brand rank and also on how it is computed can be found in the appendix on page 171.

VARIABLE	DESCRIPTION
prod_numretailer	Number of retailers who offer the current product
prod_avgprice	Average price of the product during the week of observation given in €100
prod_numclicks	Number of clicks received by the observed product divided by 100
prod_numratings } ssk_numratings }	See appendix page 117

Table 21: Miscellaneous variables

4.3 HYPOTHESIS

The goal of this section is to formulate hypothesis on the sign of the coefficient for each of the presented variables. The structure of this section is that each variable of the model specification will be listed including a short repetitive description and a hypothesis for the sign will be formulated.

ELASTICITY: The product's estimated elasticity of the respective type of click.

Products and therefore also markets which are subject to an elastic demand provide incentives for undercutting, because in such a setup undercutting leads to a large increase in the revenue. Therefore it is tempting for retailers to lower prices and also to undercut their rivals. Recall that elasticities are reported as negative numbers, i.e. the lesser the value of the elasticity the more elastic the respective demand.

The expected sign of the regression coefficient is $-$.

PROD_RECOMMENDATION: Share of ratings where the consumer recommends the purchase of the product under observation.

Concerning this variable there are two effects which go in the opposite direction. First of all the share of consumers which recommend a product is an indicator for product quality. It might be that the price competition is not as tough for high-quality products as it is in the case of low-quality products. This would suggest that the importance of other marketing tools increases, rendering the position of the price-leader as less attractive. To sum up, this differentiation effect suggests that a larger share of consumers who recommend the purchase of the respective product leads to a smaller number of changes of the price-leader.

On the other hand it might be the case that mark-ups are larger for high quality products. Higher mark-ups imply a larger profit for each unit sold. Thus high mark-ups provide an incentive to retailers to reduce the price of the respective product. The essence of this effect is that a larger share of consumers who recommend the product leads to a larger number of changes of the price-leader. However, one cannot a priori tell which of the two effects has the greater impact.

The expected sign of the regression coefficient is therefore undetermined.

There are two opposing effects concerning product quality - a differentiation effect and a mark-up effect.

PROD_SUBST_MSIZE: Size of the market for substitutes measured as the total number of clicks received by products of the current subcategory but with exclusion of clicks of the observed product.

A large number of available substitutes indicates that the market is large and that mark-ups might be high. On account of this, undercutting should increase profits and would therefore be a desirable action. Furthermore one has to bear in mind that if retailer A becomes the price leader for product B he or she will not only lure away consumers that would have bought product B from other retailers, but also from consumers which would have bought substitutes of B. Both of these effects suggest that the number of changes of the price leader should be positively correlated to the market size of the substitutes of the respective product.

On the other hand one has to consider that in this thesis one observes the actions of retailers and not producers. The substitutive relationship between products of the same subcategory could result in cannibalization effects. If a retailer lowers the price for product A, it might very well be that the demand for product B (which is also offered by the same retailer that lowered the price for A) decreases. Therefore a retailer who lowers the price of the product does at least to some extent cut off his nose to spite his face. This cannibalization effect suggests a smaller mark-up effect as assumed in the previous paragraph.

As the cannibalization effect is only dampening the mark-up effect the expected sign of the regression coefficient is +.

PROD_BRANDRANK: The rank of the brand as an integer value.

The establishment of a brand name is costly and requires the producer of a product to invest resources to build it up. Since one can observe that companies are obviously willing to bear these costs one has to conclude that the benefits of a brand name outweigh the costs. One of the benefits of a brand name is the possible reduction of the number of products which consumers perceive as substitutes, simply spoken: the brand becomes a product or even a complete group of products.

Firms that built up a brand name might try to recover their expenses by charging a premium price with a higher mark-up. It seems reasonable to assume that retailers will, at least to some extent, also benefit from this higher mark-up. This renders the market of brand products more attractive, which would result in more frequent changes of the price leader.

In this context one should bear in mind that producers will try to counteract this undercutting in order to avoid a ruinous competition. This could be achieved e.g. by supplying only those retailers which accept the pricing policy of the producer. This would dampen the competition and hence result in less frequent changes of the price leader.

As one cannot tell which effect will be larger it is impossible to make a clear statement about the sign of the resulting coefficient.

PROD_NUMCLICKS: Number of clicks received by the observed product.

The number of clicks on a product constitutes especially in the case of *LCT* the demand for the product and therefore also the market volume. A greater market volume leads *ceteris paribus* to larger profits, which renders undercutting as a very attractive action. This suggests that if a product receives a large number of clicks, one should also observe a rather high number of changes of the price leader.

The expected sign of the regression coefficient is +.

PROD_NUMRETAILER: Number of retailers that are offering the observed product.

If more retailers offering a specific product, the competition for the market of this product will be tougher. A tougher price competition leads in turn to more frequent changes of the price leader. Furthermore literature suggests that an increasing number of retailers reduces the chance for explicit or tacit collusion, because it becomes harder to monitor agreements and it also becomes more difficult to detect traitors. This effect also fosters the idea that a larger number of retailers leads to more frequent changes of the price leader.

The expected sign of the regression coefficient is +.

PROD_AVGPRICE: The average price of the observed product. The average price is the unweighted average of all prices of a specific product during the week of observation.

If it is true that more expensive products offer a higher mark-up, then one could conclude that a higher price leads to more frequent changes of the price leader.

The expected sign of the regression coefficient is +.

CAT4_HARDWARE: A dummy variable which takes on the value of 1 for products which do belong to the category *Hardware*.

This variable controls for systematic differences in elasticities between different categories. Literature does not suggest a specific sign for the coefficient of this variable and therefore it is not possible to state any expectations *ex ante*.

CAT9_VIDEOTV: A dummy variable which takes on the value of 1 for products which belong to the category *Video/Foto/TV*.

This variable controls for systematic differences in elasticities between different categories. Literature does not suggest a specific sign for the coefficient of this variable and therefore it is not possible to state any expectations *ex ante*.

4.4 ESTIMATION RESULTS

This section presents the results of the regressions of this stage. All regressions from this chapter have been carried out in the following three versions:

VERSION 1: This version consists of all observations which feature a negative elasticity that has been estimated by using more than 30 observations.

VERSION 2: Observations from this version fulfill all requirements imposed by version 1, but in addition observations which have been marked as outliers by Grubbs' test will be excluded.

VERSION 3: The requirements of this version are the same as version 2, but version 3 only contains observations from the category *Hardware*.

4.4.1 A Short Remark Concerning The Number Of Price Leader Changes

Count variables take on nonnegative integer values only.

As already mentioned before, this chapter tries to explain the number of changes of the price leader for a product, in order to analyze the influence of the price elasticity of demand on the type of competition prevalent in the market of the respective product. However, there is one significant difference between the explained variable of this chapter and the explained variable of the previous chapter, namely `num_pleader_change` is classified as a *count variable*, i.e. it only takes on nonnegative integer values. Furthermore in the vast majority of the applications using count data, the count variable will take on the value of 0 for at least some of the observations. In the case of large numbers one might still be able to use OLS to estimate the respective regressions, but the majority of the values of `num_pleader_change` is rather small.

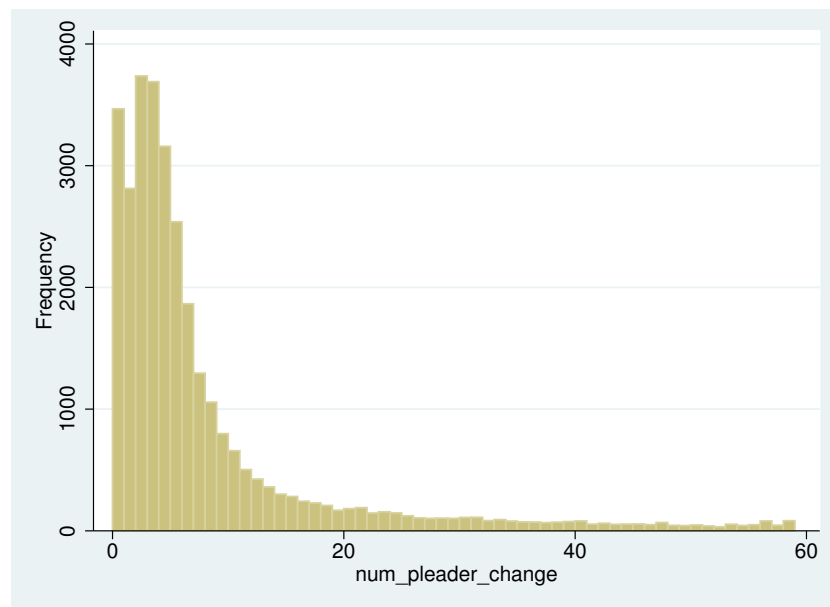


Figure 21: Histogram of the number of price leader changes

If one uses OLS to estimate a model $y = \beta x$, it is very likely that there are some x where $x\hat{\beta} < 0$, i.e. some of the predicted values \hat{y} will be negative. One possible solution to prevent negative predicted values would be to apply a logarithmic transformation to the explained variables. However, this solution is not adequate in the context of count data, because, as mentioned before, count variables usually take on a value of 0 for at least some observations. In some cases the majority of the observations will feature a count variable with a value of 0, e.g. if y measures the number of children in a family who are high school graduates. Wooldridge (2002, p. 645) mentions that one could also use

transformations which allow $y = 0$ like e.g. $\log(1 + y)$. Yet in the same breath he states that “ $\log(1 + y)$ itself is nonnegative, and it is not obvious how to recover $E(y|x)$ from a linear model for $E[\log(1 + y)|x]$ ”.

Instead of applying any transformations Wooldridge (2002) suggests that one should model $E(y|x)$ directly and ensure positive predicted values by choosing an appropriate functional form like e.g. $E(y|x) = \exp(x\beta)$. However, since a logarithmic transformation is not possible in this case, one cannot use OLS to estimate this model specification. Therefore one could either use Non-linear Least Squares (NLS) or a Poisson regression model. According to Wooldridge (2002) NLS would be the inferior solution, because NLS-estimators are relatively inefficient in the context of count data. Hence the regressions in this chapter will be estimated by using a Poisson regression model which will be estimated by QMLE.

As mentioned above the functional form of the models used in this chapter can be denoted as $E(y|x) = \exp(x\beta)$, where x is a $1 \times K$ -vector containing the $K - 1$ explanatory variables of the model² and β is a $K \times 1$ -vector containing the estimation coefficients. To interpret the parameter β_j one has to differentiate the model with respect to x_j . This is shown in equation 4.1.

$$\frac{\partial E(y|x)}{\partial x_j} = \exp(x\beta)\beta_j \quad (4.1)$$

Using the definition of the model specification one can rearrange equation 4.1 and apply the chain-rule of differentiation to obtain the following representation of β_j :

$$\beta_j = \frac{\partial E(y|x)}{\partial x_j} \frac{1}{E(y|x)} = \frac{\partial \log[E(y|x)]}{\partial x_j} \quad (4.2)$$

Equation 4.2 states that the Poisson specification used in this chapter can be interpreted as a log-level model. According to Wooldridge (2002, p. 648) $100\beta_j$ can therefore be interpreted as the semi elasticity of $E(y|x)$ with respect to x_j . Therefore he states that “for small changes Δx_j , the percentage change in $E(y|x)$ is roughly $(100\beta_j)\Delta x_j$ ”³.

A detailed discussion on Poisson regression models and QMLE-estimators (respectively their properties) is beyond the scope of this thesis. The interested reader may be referred to Wooldridge (2002) - Chapter 19.

4.4.2 Results Of The First Regression

As already mentioned before the first regression is a simple estimation where the LHS consists of the `num_pleader_changes` (the number of changes of the price leader) and the RHS consists of the elasticity of the respective click type. This regression has been carried out in all three regression versions. Both, the full dataset and also the reduced, cleansed dataset have been used.

The full results of these regressions are given in appendix A on pages 105 to 107. The results are quite robust to variations in the

Count data require a Poisson regression model which is estimated by the method of Quasi-Maximum Likelihood (QML).

² $K - 1$ because the first element of x is set to unity to account for the intercept.

³ If a more accurate estimate is required one can compute $\% \Delta E(y|x) = 100(\exp(\beta_j \Delta x_j) - 1)$, as given in Wooldridge (2003, p. 574).

regression version and to variations in the dataset. The intercept of the regression line is positive and statistically significant on a 1% level across the board. The vast majority of the coefficients on elasticity is statistically significant on a 1% level as well. However, especially in the case of the cleansed dataset the coefficients show a positive sign for the elasticities coming from the first stage regressions which incorporate control variables for retailer specific information. One of the concluding results of chapter 3 was that although the quality of the cleansed dataset is better than the quality of the full dataset, it is still far from perfect. This suggests that the positive signs are a result of a special selection of data, because this problem is more incisive in the context of the cleansed dataset.

Despite some positive coefficients the results are statistically and economically significant.

So all in all the few positive coefficients might not be a problem. Despite those positive coefficients the results look promising. Although some of the coefficients seem to be low. For example, the coefficient on `l_elast` in table 43 is -0.0049 . Looking at the summary statistics of `l_elast` and `num_pleader_change` in table 22, we are shown that the average value of the elasticity is -6.185 , the standard deviation is -4.298 , its minimum value is -20.61 and its maximum value is -0.429 . The difference between the maximum and the minimum value is therefore 20.181 , i.e. in the extreme case of a change from the minimal to the maximal elasticity we would observe a change of the variable `num_pleader_change` of 9.89% . If one applies this change to the mean value of `num_pleader_change` one receives an absolute change in this variable of 5.56 . Even if one uses the median of `num_pleader_change` as a reference point, the absolute value is still 1.78 . Recall that this variable is an integer value. So from an economical point of view a meaningful minimal change of `num_pleader_change` should be at least 1 or greater.

Variable	Mean	Median	Std. Dev.	Min.	Max.
<code>num_pleader_change</code>	56.238	18	183.341	0	2318
<code>l_elast</code>	-6.185	-5.511	4.298	-20.161	-0.429

Table 22: Exemplary summary statistics for the cleansed dataset (N=181)

Despite the good economical significance one must not forget that these first regressions contain the elasticity as the only explaining variable. Therefore it might be that the estimated coefficient on the elasticity is influenced by other variables which have not been included into the regression so far. Furthermore the results might suffer from an omitted variable bias. The first regressions suggest that the elasticity has indeed an influence on the frequency of the changes of the price leader, however one needs a more detailed model to give further information. This is done in the next sections.

4.4.3 Results For The Full And The Cleansed Dataset

This section presents the results of the complete model for both, the full and the cleansed dataset. The results for both datasets are combined into one section because those datasets do in fact yield very similar results. However, the results of the cleansed dataset are less clear-cut.

The results are very robust to changes in the regression version. For this reason this section focuses on regression version 2 because it seems to be the most comprehensive version. Nevertheless the appendix

provides a full report of the regression results for each and every regression version. The tables containing the exact results can be found on pages 108 to 113.

Before going into the detailed analysis of the respective variables one should notice the following general results: First of all one can state that especially in the case of the full dataset the coefficients are statistically significant on a 1% level. In addition, the results are robust to changes in the type of the used elasticity. This supports the notion that the estimated elasticity types are numerical representations of the same underlying phenomenon of the price sensitivity of consumers. A further striking result is that the vast majority of the coefficients possesses the expected sign. Therefore one can a priori say that the results do support the formulated hypothesis. Some of the results seem to feature a rather low economical significance. Therefore section 4.4.3.1 explicitly deals with this topic.

To present the results of the estimation in a structured way this section will give a short analysis for each variable in a separate paragraph. At the beginning of each paragraph there is a short note on the expected sign of the respective variable.

ELASTICITY One would expect a negative sign on this variable. Similar to the first regressions the coefficient on the respective elasticities shows a negative sign for almost every type of elasticity in the case of the full dataset. Therefore these results support the hypothesis that a more elastic demand provides an incentive to retailers to undercut their rivals in order to be the price leader. These results also indicate that an elastic demand also fosters a Bertrand type of competition.

The cleansed dataset on the other hand shows slightly different results. The elasticities which have been estimated by incorporating retailer specific data consistently show a statistically significant and positive sign. However, one has to consider that the estimations using the elasticities coming from LCT only consist of 56 respective 59 observations. Since the problem of a positive sign only arises for regressions which have been carried out with a small number of observations it might very well be that a shortage in the degrees of freedom is the culprit of the problem.

In the case of the full dataset the coefficients on this variable range from -0.0004 for `c_elast_inv` to -0.0135 for `c_elast`. The results for the cleansed dataset show coefficients between -0.003 and -0.0237 . The strikingly small coefficient on `c_elast_inv` can be explained by the large absolute average value of this variable⁴. An increase of `c_elast_inv` by one standard deviation lowers the number of price leader changes by 2.48%. Applied to the mean of `num_pleader_change` this yields an absolute change of 0.716. This value is indeed rather small.

This potential lack of economical significance of this variable might not be a problem per se. Although the mean of `num_pleader_change` is roughly 29, the variable is not distributed normally. The histogram on page 78 showed that the values of the variable form a heavily positively skewed distribution. For over 70% of the products the number of changes of the price leader is less than 15. So one can state that this variable is rather small in size. Therefore it might be the case that

⁴ A comparison of the average values of `c_elast` and `c_elast_inv` shows that the average value of the former is -4.45 , whereas the average value of `c_elast_inv` is -77.21 . The same is true for the standard deviation, which is 4.918 in the primer case and 61.683 in the latter case.

changes in one explaining variable are simply not enough to result in a change of `num_pleader_change` of 1 or even more.

PROD_RECOMMENDATION There is no expected sign for this variable. The results for this variable feature significant differences when comparing the regressions based on the full dataset with those from the cleansed dataset.

If one compares the results from table 52 on page 112 with those from table 49 on page 109 one will notice two things. First of all, the statistical significant coefficients for the full dataset feature a negative sign across the board with `c_elast_choke` as the only exception. Almost the opposite is true in the case of the cleansed dataset. Here seven out of twelve coefficients show a positive sign. The sign is negative only in the case of `cv_lctw_l_elast` and for all four elasticities based on aggregated/accumulated clicks. Secondly, the absolute value of the coefficients from the full dataset is on average significantly smaller than the absolute value of the coefficients from the cleansed dataset.

Although the picture drawn by the results for the cleansed dataset seems to be rather blurry, one can state that the results for the full dataset approve the existence of a differentiation effect. The differentiation effect states that highly recommended products possess a high quality which decreases the importance of the price in relation to other marketing tools. The results of the cleansed dataset partially support the idea of a mark-up effect. This effect states that highly recommended products possess a higher mark-up, rendering the position of the price leader as very attractive. However, one has to notice that this hypothesis is based on the assumption that recommended products feature high-markups. To verify this assumption one would have to obtain data on costs in order to compute mark-ups. The results for the cleansed dataset might also be a hint that the process of aggregation and accumulation is the culprit of the inconsistent results concerning the sign of the coefficients.

Concerning the size of the absolute value of the coefficients, it turns out that the vast majority of the coefficients is greater than one in the case of the cleansed dataset. The absolute value of the coefficients for the full dataset ranges from 0.0156 to a maximum of 0.1042. These results imply that a change of `prod_recommendation` by a value of 1 changes the number of price leader changes by about 1.56% to 10.42%. However one has to recall that `prod_recommendation` is the share of users who recommend the respective product. The definition of this variable implies that its values are bound to the range from 0 to 1, whereas 1 means that 100% of the users recommend the purchase of the current product. A coefficient of -0.1042 therefore predicates that an increase of recommendations by 10 percentage-points will decrease the number of changes of the price leader by only 1.042%.

PROD_AVGMISSING There is no expected sign for this variable. This variable controls for missing quality data on a product. It is questionable whether one can meaningfully compare the results for this variable across the full and the cleansed dataset. Recall that this variable takes on 1 if the data on product quality is missing for the current product and else 0. Since this variable is logically connected to `prod_recommendation` it is interesting to notice that the coefficients of this variable behave in the opposite way compared to the coefficients on

prod_recommendation. In the case of prod_recommendation the vast majority of the coefficients on the results for the full dataset were negative, whereas the cleansed dataset featured positive signs. For prod_avgmissing it is exactly the other way round, the cleansed dataset predominantly features negative signs, whereas the coefficients of the full dataset are positive to a large extent.

A closer look at this variable reveals that it might be the answer to the differing results for prod_recommendation. As mentioned before, this variable controls for missing quality data. To be more precise it does not only indicate whether the data on product quality is missing for a specific product, but it rather indicates that the missing data on product quality has been replaced by a dataset-wide average value for product quality. Potential pros and cons of the replacement of missing values by average values have already been discussed on page 57. Nevertheless the summary statistics of prod_avgmissing show that there is a significant difference between the full dataset and the cleansed dataset. The mean of prod_avgmissing is roughly 0.62 for the full dataset and 0.28 in the case of the cleansed dataset, i.e. data on product quality is missing for 62% respective 28% of the observations. This suggests that it is very likely that there is a systematic difference between the results of the full dataset and the cleansed dataset independently of the question whether the replacement of missing values by average values is the correct thing to do, simply because the data on product quality for the full dataset contains a much larger proportion of artificial values.

There is a significant difference between the availability of data representing product quality for the full and the cleansed dataset.

PROD_SUBST_MSIZESIZE The hypothesis suggests a positive sign for this variable. The coefficients which are positive almost across the board for both datasets confirm this hypothesis. The values range from 0.0124 to 0.0148 for the full dataset and from 0.0207 to 0.0328 for the cleansed dataset. Recall that prod_subst_msize is measured in 1000-clicks. So using the coefficient of 0.0328 one can state that if the clicks received by substitutes of a product increase by 1000 the number of changes of the price leader for this product will increase by 3.28%, which, applied to the mean of num_pleader_change of 29, corresponds to an absolute increase of roughly 1 change of the price leader.

To sum up one can say that the results clearly support the hypothesis that an increase in the size of the market for substitutes of a specific product intensifies the competition for that specific product.

PROD_BRANDRANK The hypothesis suggests an unambiguous sign for this variable. The coefficients on this variable are statistically significant on a 1% level and negative across the board for the full dataset. A statistically significant negative coefficient is also predominant in the case of the cleansed dataset. This suggests that the higher mark-ups of brand products attract retailers and intensify the competition.

The coefficients range from -0.0187 to -0.0378 for the full dataset and from -0.0119 to -0.2125 for the cleansed dataset. Recall that the brand rank variable is measured in 100 ranks in this chapter, i.e. the reported coefficients apply to a change in prod_brandrank by 100. Considering the fact that the strongest brands according to the brand rank are located in the area of 30 and below the impact of -1.87% to -3.78% seems to be rather low. This is an indicator that the producers counteract the effect of the higher mark-ups to at least some extent.

PROD_NUMCLICKS The coefficients for this variable feature the expected positive sign across the board. Hence the results verify the hypothesis that the number of clicks on a product and thus the size of the market of the respective product intensifies competition.

In case of the full dataset they range from 0.0066 to 0.0418 and in the case of the cleansed dataset they range from 0.0229 to 0.0684. Since *prod_numclicks* is denoted in 100-clicks, the reported coefficients apply to an increase or decrease of the number of clicks by 100. Applied to the mean of *num_pleader_change* of roughly 29 the coefficient of 0.0418, which belongs to *l_elast*, indicates that an increase of the number of clicks of a product by 100 increases the number of price leader changes by 1.21 (or 4.18%). In the context of *l_elast* the variable *prod_numclicks* features a mean of 3.431 and a standard deviation of 4.58.

PROD_NUMRETAILER One would expect a positive sign on this variable. The results for this variable from the full and the cleansed dataset are contradictory. The resulting coefficients from the full dataset are positive and statistically significant on a 1% level across the board. Therefore these results would support the hypothesis, that an increasing number of retailers intensifies the competition between retailers which leads to more frequent changes of the price leader. The coefficients range from 0.0002 to 0.0014. The standard deviation of *prod_numretailer* is roughly 17 to 20. So even if one changes *prod_numretailer* by an average standard deviation of 18.5 this would imply a change of the number of price leader changes by only 0.37% to 2.59%. On the grounds of economical significance one has to say that these coefficients are rather small.

The results for the cleansed dataset show a different picture. Although the coefficients are predominantly statistically significant, they feature a negative sign in almost every case. Since the second stage regressions do not control for retailer specific data it might be that there are systematic differences between the average retailer of the full dataset and the average retailer of the cleansed dataset. Nevertheless the results of the cleansed dataset do not support the formulated hypothesis.

PROD_AVGPRICE The coefficients feature the expected positive sign in the majority of the cases for both, the full and the cleansed dataset. The coefficients of the full dataset range from 0.001 to 0.0074 and the ones for the cleansed dataset range from 0.0425 to 0.0577. The reported coefficients apply to a change of *prod_avgprice* by €100. A change of *prod_avgprice* by €100 will therefore change the number of price leader changes by 0.1% to 0.74% in case of the full dataset and by 4.25% to 5.77% in the case of the cleansed dataset.

This supports the hypothesis that high-price products have a higher mark-up which provides an incentive to retailers to be the price leader.

CAT4_HARDWARE There are no expectations for the sign of this variable. This variable only controls for systematic differences between categories. The reported coefficients are all negative and statistically significant for both datasets. This suggests that compared to the base group products from the category *Hardware* feature less frequent changes of the price leader.

`CAT9_VIDEOSFOTOTV` There is no expected sign for this variable. Just as `cat4_hardware` this variable controls for systematic differences between categories. Apart from the coefficient on `aggr_c_elast_choke` coefficients are all negative. This suggests that compared to the base group products from the category *Video/Foto/TV* feature less frequent changes of the price leader.

4.4.3.1 Putting The Results Into Context With The Help Of An Example

The previous paragraphs have shown that each coefficient for each variable is statistically significant, however there are some elasticities where there seem to be problems with the economical significance. Although the impact of a single variable on `num_pleader_change` might be only very small, one has to bear in mind, that the values of `num_pleader_change` itself are also rather small. Therefore it might require more than only a change in a single variable to produce a change of `num_pleader_change` by 1. Furthermore since the resulting coefficients constitute semi elasticities it can be difficult to interpret them without the context of the explained variable.

Variable	Mean	Std. Dev.	Coeff.	$\Delta E(y x)$
<code>c_elast</code>	-4.45	4.918	-0.0135	6.64%
<code>prod_recommendation</code>	0.699	0.201	-0.0156	0.31%
<code>prod_avgmissing</code>	0.62	0.485	0.0284	1.38%
<code>prod_subst_msize</code>	4.84	7.612	0.0132	10.05%
<code>prod_brandrank</code>	1.405	2.749	-0.0192	5.28%
<code>prod_numclicks</code>	0.291	0.956	0.0305	2.92%
<code>prod_numretailer</code>	61.86	33.937	0.0012	4.07%
<code>prod_avgprice</code>	3.595	8.109	0.0056	4.54%

Table 23: Summary statistics of the explaining variables (N=14814)

This subsection tries to give further insights into the results by conducting the following simple experiment: To check the economical impact of the respective variables, one changes the value of each explaining variable by its standard deviation. This puts the coefficient into the context of the statistical properties of a variable, which helps to illustrate the size of the resulting change. Furthermore the impacts of the respective variables are summed up to get a grasp of the economic significance of the full set of variables. In order to prevent opposing effects from canceling out each other the direction of the standard deviation will be chosen in accordance to the sign of the respective coefficient. Since the dataset used for the estimation of the second stage regressions is also always restricted by the values of the respective elasticity, the experiment in this section will be conducted by using `c_elast`. Therefore table 23 shows the summary statistics for the explaining variables, but only for those observations which fulfill the requirements of regression version 2. The final column of table 23 shows the expected percentage change of the number of price leader changes computed as $\% \Delta E(y|x) = 100(\exp(\beta_j \Delta x_j) - 1)$.

The summary statistics for the explaining variables are reported with the same scaling as it has been used for the regressions. To get the total change in `num_pleader_change` one can sum up the impacts of the

respective variables either additively or multiplicatively. The primer is shown in equation 4.3, the latter in equation 4.4.

$$\Delta E(y|x) = \sum_{i=1}^n (\exp(\sigma_i |\hat{\beta}_i|)) - n \quad (4.3)$$

Where in both equations σ_i denotes the standard deviation and β_i the coefficient of the respective variable with index i . The evaluation of equation 4.3 yields a total change `num_pleader_change` of roughly 36.32%, or applied to the mean of `num_pleader_change` of 29 an absolute change of 10.53.

$$\Delta E(y|x) = \prod_{i=1}^n (\exp(\sigma_i |\hat{\beta}_i|) - 1) \quad (4.4)$$

The evaluation of equation 4.4 yields a total change `num_pleader_change` of roughly 42.17%, or applied to the mean of `num_pleader_change` of 29 an absolute change of 12.23.

4.5 CONCLUSION AND REMAINING PROBLEMS

This chapter has established a model to show the impact of the price elasticity of demand on the market structure. Market structure has been quantified by the number of changes of the price leader for a product. In addition to the elasticity the model proposed in this thesis uses further explaining variables like the market size for substitutes, the market size of the product and also the quality of the product.

The model has been estimated with two datasets, the full dataset, which contains the full range of products from the Geizhals database and also a reduced, cleansed dataset which is supposed to feature better data quality. The estimation of the model has shown that the coefficients support the formulated hypothesis as suggested by economic theory. Nevertheless some of the coefficients of the cleansed dataset did not behave as expected. This however does not constitute a problem because for one thing, it has been shown that the cleansed dataset only provides a small number of observations and therefore only a small number of degrees of freedom. For another thing the summary statistics presented in this chapter show that there is a systematic difference between the full and the cleansed dataset. Both of these arguments would explain the unexpected results of the cleansed dataset.

Apart from improving the quality of the estimated elasticities (as it has already been proposed in the conclusion of chapter 3) one could try a different measure for market structure. The number of changes of the price leader is a very good measure for the intensity of the competition, but only if the changes of the price leader are a result from undercutting. The problem of `num_pleader_change` as used in this chapter can be explained by the following example: Assume that retailer A undercuts the price leader B only temporarily, e.g in the form of a temporary discount for the time period of one week. After this week A raises his price which leaves B as the price leader. In the end the price has not changed, but the number of price leader changes has gone up by two. Even though in such a setup the number of changes of the

price leader might represent the intensity of competition in some way, it will be a poor indicator whether the market is subject to a Bertrand- or a Cournot-competition.

A solution to the problem mentioned above would be the introduction of an alternative measure for market structure and intensity of competition. A simple way would be to use a form of `num_pleader_change` which only counts price leader changes which are a result from undercutting. This could e.g. be implemented by excluding price leader changes which can be attributed to an increase of the price by the current price leader. So if retailer A would undercut B temporarily and then increase the price again, this will only count as one price leader change.

The variable `num_pleader_change` as used in this thesis has been generated by counting the number of price leader changes for the complete range of offers stored in the dump of the Geizhals production database, ranging from Oct 28, 2006 to July 17, 2007. This timespan contains a huge number of offers which leaves the generation of the variable as a very cumbersome and time-intensive number-crunching process. The processing-time for the first version of `num_pleader_change` was roughly one month. So one should have a clear plan before computing alternatives to `num_pleader_change`.

CONCLUSION

This thesis has shown an approach for the estimation of the price elasticity of demand and hence of the respective demand curves in the context of a price comparison website, namely www.geizhals.at. The main challenge in such a context is the fact that price comparison websites rather store clicks on products than actual purchases. Therefore one has to find a way for the conversion of clicks to purchases. The concept of LCT constitute one possible solution to this challenge. This diploma thesis has discussed the advantages and drawbacks of the estimation of demand curves based on last-clicks-through. In addition, it has pointed out the importance of a well-balanced relationship between sample size and the quality of the data sample. Concerning the quality of the data sample this thesis has presented several alternative criteria and approaches to quantify the quality of the estimated price elasticities. Nevertheless it became apparent that the Geizhals dataset is very rich measured in the number of products and hence markets, but the number of observations within each market is rather small, hampering the estimation procedure.

In a next step the thesis has tried to explain the price elasticity of demand by market- and product characteristics. It has turned out that firms do indeed have an influence on the price elasticity of demand e.g. by setting up a brand name. However, especially a brand name is established by a producer, whereas at Geizhals one observes data from retailers. Therefore it is not clear whether the identified determinants should be attributed to the retailer- or the producer side of the market.

Finally this thesis dealt with the question whether the price elasticity of demand is a determinant of market structure. This question has been answered by explaining the number of changes of the price leader of a product by the price elasticity and further control-variables. It was possible to establish results which show a significant economical impact. These results suggest that the price elasticity of demand does at least partly determine the type of competition prevalent on the market.

As a concluding remark one can say that this thesis has shown that good-quality-elasticities are vital in order to use them in further regression stages. Although this thesis has introduced several criteria to assess the quality of the estimated elasticities, it was not possible to extensively check for common econometric phenomena like heteroskedasticity and endogeneity. Furthermore there some markets suffer from a low number of observations. Subsequent studies should therefore try to overcome aforementioned problems. Especially controlling for simultaneity facilitates the usage of a longer period under review, which would reduce the problems stemming from a low number of observations.



APPENDIX - REGRESSION RESULTS AND STATISTICS

A.1 SUMMARY STATISTICS FOR THE EXPLANATORY AND CONTROL VARIABLES

This section presents the summary statistics for the explanatory- and control variables used in chapters 3 and 4. The exact dataset depends on the respective type of elasticity. However, the differences between the datasets of the respective elasticities are negligibly. Therefore this section will only differentiate between normal clicks and LCT. The summary statistics of the normal clicks will be shown using the example of `c_elast`, the LCT are represented by `cv_lctw_c_elast`. The same argument holds true for the regression version, hence the summary statistics will only be reported for regression version 2. Recall that this version consists of all observations which feature a negative elasticity that have been estimated by using more than 30 observations and which have passed Grubbs' test. The regressions of chapters 3 and 4 have been carried out on the basis of two datasets, a full dataset and a reduced, cleansed dataset. Since there is a drastic difference in the sample size of those two datasets, the summary statistics will be reported for each of them. The explanatory variables are similar in both chapters, therefore their descriptive statistics will be merged into a single table. The respective elasticities constitute the LHS-variable in chapter 3 and they are part of the RHS-variables in chapter 4. In order to keep the summary statistics for the elasticities manageable they are reported in a separate section in tables 28 and 29.

Variable	Mean	Std. Dev.	Med.	Min	Max
prod_recommendation	0.699	0.201	0.693	0	1
prod_avgmissing	0.62	0.485	1	0	1
ssk_numofferedprods	381.123	306.95	323	1	1689
prod_brandrank	140.545	274.936	49	1	1506
ssk_numratings	146.046	235.408	41	0	1220
prod_numretailer	61.86	33.937	55	3	231
prod_avgprice	359.476	810.888	103.627	1.142	18797.434
prod_substitute_marketsize	4839.848	7611.912	2297	0	41977
prod_numclicks	29.098	95.578	7	1	3681
cat4_hardware	0.701		1	0	1
cat9_videofototv	0.154		0	0	1

Table 24: Summary statistics for the explanatory variables of the full dataset filtered in accordance to `c_elast` and regression version 2 (N=14814)

Variable	Mean	Std. Dev.	Med.	Min	Max
prod_recommendation	0.709	0.221	0.693	0	1
prod_avgmissing	0.429	0.495	0	0	1
ssk_numofferedprods	369.714	297.147	323	3	1689
prod_brandrank	154.558	315.386	50	1	1506
ssk_numratings	140.561	220.848	50	0	1220
prod_numretailer	68.134	37.422	61	5	231
prod_avgprice	263.631	713.428	82.437	1.779	18797.434
prod_substitute_marketsize	4722.789	7295.352	2600	2	41977
prod_numclicks	34.268	85.158	13	1	3028
cat4_hardware	0.687		1	0	1
cat9_videofototv	0.168		0	0	1

Table 25: Summary statistics for the explanatory variables of the full dataset, filtered in accordance to cv_lctw_c_elast and regression version 2 (N=4824)

Variable	Mean	Std. Dev.	Med.	Min	Max
prod_recommendation	0.745	0.153	0.714	0	1
prod_avgmissing	0.275	0.448	0	0	1
ssk_numofferedprods	420.101	281.263	407	2	1005
prod_brandrank	207.915	442.981	47	1	1506
ssk_numratings	359.317	314.253	346	0	1220
prod_numretailer	34.656	17.856	30	11	112
prod_avgprice	656.313	672.956	401.416	3.363	3849.009
prod_substitute_marketsize	12710.45	10786.602	9899	32	41929
prod_numclicks	329.767	450.567	158	12	3059
cat4_hardware	0.455		0	0	1
cat9_videofototv	0.402		0	0	1

Table 26: Summary statistics for the explanatory variables of the cleansed dataset filtered in accordance to c_elast and regression version 2 (N=189)

Variable	Mean	Std. Dev.	Med.	Min	Max
prod_recommendation	0.757	0.126	0.76	0.4	1
prod_avgmissing	0.125	0.334	0	0	1
ssk_numofferedprods	475.089	283.263	448	29	1005
prod_brandrank	232.714	457.094	56	5	1506
ssk_numratings	331.232	361.421	219	0	1220
prod_numretailer	32.071	15.109	27.5	16	103
prod_avgprice	384.281	392.398	246.839	23.421	2406.538
prod_substitute_marketsize	12939.161	12517.34	6500	472	41929
prod_numclicks	222.107	215.208	137.5	12	1019
cat4_hardware	0.375		0	0	1
cat9_videofototv	0.411		0	0	1

Table 27: Summary statistics of the explanatory variables of the cleansed dataset, filtered in accordance to cv_lctw_c_elast and regression version 2 (N=56)

A.2 SUMMARY STATISTICS FOR THE ESTIMATED ELASTICITIES

The following two tables report the summary statistics for the respective elasticities. Since the dataset is always restricted by a filter that checks whether the respective elasticity is negative and has been estimated by using more than 30 observations, the summary statistic for each elasticity features a different size of N. The statistics show that the absolute value of the mean is always larger than the absolute value of the median. This implies that the distribution of the estimated elasticities is negatively skewed or skewed to the right.

Variable	Mean	Std. Dev.	Med.	Min	Max	N
c_elast	-4.45	4.918	-2.73	-37.337	-0.001	14814
c_elast_inv	-77.213	61.683	-60.144	-370.452	-1.528	14264
c_elast_choke	-7.209	7.334	-4.934	-51.048	-0.003	10615
aggr_c_elast	-17.559	11.874	-15.159	-73.113	-0.155	14710
aggr_c_elast_choke	-15.721	12.003	-12.929	-69.394	-0.009	11680
l_elast	-8.529	7.346	-6.526	-41.449	0	14940
cv_c_elast	-4.989	5.415	-3.143	-42.748	0	14771
cv_l_elast	-9.41	8.369	-7.064	-53.475	0	14795
cv_lctw_c_elast	-1.313	1.716	-0.672	-13.818	0	4824
cv_lctw_l_elast	-4.495	4.964	-2.775	-34.103	0	4833
aggr_lctw_c_elast	-9.685	7.547	-7.836	-44.523	-0.03	5329
aggr_lctw_l_elast	-7.276	5.481	-6.075	-30.961	-0.01	5322

Table 28: Summary statistics of the elasticities for the full dataset, filtered in accordance to regression version 2

Variable	Mean	Std. Dev.	Med.	Min	Max	N
c_elast	-10.715	7.488	-9.420	-35.174	-0.081	189
c_elast_inv	-85.568	59.705	-67.117	-288.282	-7.901	181
c_elast_choke	-10.745	7.567	-9.420	-35.174	-0.081	183
aggr_c_elast	-17.172	10.37	-15.369	-48.279	-1.169	137
aggr_c_elast_choke	-16.983	10.052	-15.617	-48.279	-1.169	136
l_elast	-6.185	4.298	-5.511	-20.161	-0.429	181
cv_c_elast	-11.981	8.236	-10.449	-39.789	-0.281	189
cv_l_elast	-6.965	5.559	-5.721	-24.193	-0.201	189
cv_lctw_c_elast	-4.565	2.775	-4.738	-12.064	-0.159	56
cv_lctw_l_elast	-4.357	2.901	-4.326	-11.585	-0.276	59
aggr_lctw_c_elast	-11.099	6.137	-10.49	-28.442	-1.82	38
aggr_lctw_l_elast	-4.877	3.272	-4.173	-12.583	-0.486	39

Table 29: Summary statistics of the elasticities for the cleansed dataset, filtered in accordance to regression version 2

A.3 SUMMARY STATISTICS FOR THE NUMBER OF PRICE LEADER CHANGES

The following tables report the summary statistics for the number of price leader changes, which constitutes the LHS-variable in chapter 4. The dataset varies with the type of the respective elasticity. This section will therefore report the summary statistics of num_pleader_change with respect to every type of elasticity. One can notice that neither the mean, nor the median change significantly. However, this is not true in case of the cleansed dataset. But this can possibly be attributed to the small sample size of the cleansed dataset, especially in the case of LCT.

Filter	Mean	Std. Dev.	Med.	Min	Max	N
c_elast	28.935	65.828	7	0	2318	14814
c_elast_inv	28.887	65.823	7	0	2318	14264
c_elast_choke	29.581	66.431	8	0	2318	10615
aggr_c_elast	28.802	61.551	7	0	1053	14710
aggr_c_choke	29.314	60.889	8	0	1053	11680
l_elast	28.922	65.428	7	0	2318	14940
cv_c_elast	29.264	66.483	7	0	2318	14771
cv_l_elast	29.293	66.398	7	0	2318	14795
cv_lctw_c	27.638	64.451	6	0	899	4824
cv_lctw_l	28.053	64.943	6	0	899	4833
aggr_lctw_c	27.582	63.178	7	0	899	5329
aggr_lctw_l	27.598	63.213	7	0	899	5322

Table 30: Summary statistics of the number of price leader changes (num_pleader_change). The variable name denotes the elasticity which has been used as the filter criteria.

Filter	Mean	Std. Dev.	Med.	Min	Max	N
c_elast	60.011	184.267	19	0	2318	189
c_elast_inv	60.547	187.431	19	0	2318	181
c_elast_choke	55.656	181.538	19	0	2318	183
aggr_c_elast	34.153	53.189	18	0	353	137
aggr_c_choke	31.684	45.744	17.5	0	255	136
l_elast	56.238	183.341	18	0	2318	181
cv_c_elast	59.667	183.741	19	0	2318	189
cv_l_elast	54.407	178.983	18	0	2318	189
cv_lctw_c	54.893	94.05	12	1	387	56
cv_lctw_l	56.644	93.595	13	1	387	59
aggr_lctw_c	19.474	31.616	7.5	1	170	38
aggr_lctw_l	24.026	42.206	8	1	197	39

Table 31: Summary statistics of the number of price leader changes (num_pleader_change). The variable name denotes the elasticity which has been used as the filter criteria.

A.4 ESTIMATION RESULTS OF CHAPTERS 3 AND 4

This section provides the detailed estimation results for all regressions estimated during the analysis in chapters 3 and 4.

Table 32: Full dataset regression results - version 1 (cf. section 3.4.1 on page 56)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c_elast-choke	L_elast	cv_c_elast	cv_L_elast	cv_Lctw-c_elast	cv_Lctw-L_elast	aggr_Lctw-c_elast	aggr_Lctw-L_elast
prod_recommendation	-2.2007*** (0.2801)	-1.9186 (1.5316)	-0.4379 (0.5096)	-0.5599 (0.4921)	-0.9694* (0.5715)	-2.1760*** (0.5191)	-2.0279** (1.0288)	-0.0674 (5.2198)	-0.5667** (0.2218)	-2.1635*** (0.7986)	-1.7031*** (0.4578)	-0.3389 (0.3153)
ssk_numofferedprods	0.0034*** (0.0002)	0.0113*** (0.0011)	0.0035*** (0.0004)	0.0035*** (0.0004)	0.0016*** (0.0005)	0.0014*** (0.0004)	0.0031*** (0.0008)	0.0050 (0.0036)	0.0010*** (0.0002)	0.0023*** (0.0006)	0.0034*** (0.0004)	0.0021*** (0.0002)
brand1to20	-1.6516*** (0.1269)	-9.2057*** (0.6938)	-2.0162*** (0.2374)	-2.9062*** (0.2170)	-1.9175*** (0.2587)	-0.6943*** (0.2245)	-2.3701*** (0.4784)	-6.8796*** (2.3080)	0.3539*** (0.1144)	0.9542** (0.4131)	-0.6918*** (0.2249)	-0.4966*** (0.1542)
ssk_numratings	-0.0078*** (0.0002)	-0.0263*** (0.0012)	-0.0061*** (0.0004)	-0.0083*** (0.0004)	-0.0057*** (0.0005)	-0.0037*** (0.0004)	-0.0079*** (0.0009)	-0.0108** (0.0045)	-0.0024*** (0.0002)	-0.0042*** (0.0008)	-0.0045*** (0.0005)	-0.0024*** (0.0003)
prod_numretailer	0.0232*** (0.0017)	-0.1594*** (0.0095)	0.0183*** (0.0032)	-0.0298*** (0.0032)	-0.0586*** (0.0038)	-0.0369*** (0.0033)	0.0369*** (0.0067)	0.0781** (0.0354)	0.0042*** (0.0015)	0.0040 (0.0059)	-0.0470*** (0.0030)	-0.0307*** (0.0021)
prod_avgprice	-0.0013*** (0.0001)	-0.0072*** (0.0004)	-0.0022*** (0.0002)	-0.0025*** (0.0001)	-0.0029*** (0.0002)	-0.0029*** (0.0001)	-0.0013*** (0.0003)	-0.0025** (0.0012)	-0.0002*** (0.0001)	-0.0012*** (0.0002)	-0.0013*** (0.0002)	-0.0015*** (0.0001)
cat4_hardware	0.6821*** (0.1655)	-2.8804*** (0.9048)	-1.0707*** (0.3333)	0.7799** (0.3037)	0.1708 (0.3729)	-0.1431 (0.2931)	0.4182 (0.6046)	-0.7744 (2.8962)	1.1826*** (0.1355)	0.7928 (0.4942)	1.8649*** (0.3088)	1.1155*** (0.2087)
cat9_videofootv	-1.0103*** (0.1927)	5.1881*** (1.0539)	-0.2536 (0.3848)	1.9530*** (0.3897)	2.3021*** (0.4773)	0.6116* (0.3415)	-2.2811*** (0.7295)	-6.8137* (3.5071)	-0.1882 (0.1543)	-0.6628 (0.5658)	-0.9228** (0.3849)	-1.3927*** (0.2596)
Constant	-7.3686*** (0.2673)	-31.5944*** (1.4617)	-11.7518*** (0.5240)	-15.4260*** (0.4904)	-10.6806*** (0.5925)	-9.9389*** (0.4874)	-6.9264*** (0.9810)	-12.5027** (4.8698)	-2.8927*** (0.2144)	-5.8903*** (0.7864)	-8.3643*** (0.4698)	-5.6581*** (0.3208)
Observations	12238	12238	8206	14735	11751	11617	12657	10619	3356	3003	5322	5302
R ²	0.1780	0.1343	0.0752	0.0911	0.0761	0.0761	0.0185	0.0036	0.1108	0.0345	0.1229	0.1382

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 33: Full dataset regression results - version 2 (cf. section 3.4.1 on page 56)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c_elast-choke	L_elast	cv_c_elast	cv_L_elast	cv_Lctw-c_elast	cv_Lctw-L_elast	aggr_Lctw-c_elast	aggr_Lctw-L_elast
prod_recommendation	-2.0078*** (0.2532)	-2.8257* (1.5005)	-1.3230*** (0.4399)	-0.4965 (0.4583)	-0.7322 (0.5038)	-1.7491*** (0.3783)	-2.0416*** (0.2625)	-1.4881*** (0.4472)	-0.5418*** (0.1805)	-0.7725 (0.4933)	-1.6757*** (0.4487)	-0.2340 (0.3048)
ssk_numofferedprods	0.0031*** (0.0002)	0.0113*** (0.0011)	0.0033*** (0.0004)	0.0036*** (0.0003)	0.0021*** (0.0004)	0.0011*** (0.0003)	0.0025*** (0.0002)	0.0009*** (0.0003)	0.0009*** (0.0001)	0.0013*** (0.0004)	0.0032*** (0.0004)	0.0021*** (0.0002)
brand1to20	-1.1641*** (0.1151)	-8.9569*** (0.6795)	-1.5138*** (0.2045)	-2.8253*** (0.2023)	-1.9483*** (0.2285)	-0.4263*** (0.1647)	-1.1576*** (0.1222)	-0.4907** (0.1986)	0.2841*** (0.0933)	0.9219*** (0.2545)	-0.8058*** (0.2206)	-0.5352*** (0.1491)
ssk_numratings	-0.0074*** (0.0002)	-0.0261*** (0.0012)	-0.0062*** (0.0004)	-0.0080*** (0.0004)	-0.0060*** (0.0005)	-0.0035*** (0.0003)	-0.0071*** (0.0002)	-0.0046*** (0.0004)	-0.0024*** (0.0002)	-0.0033*** (0.0005)	-0.0043*** (0.0005)	-0.0024*** (0.0003)
prod_numretailer	0.0165*** (0.0016)	-0.1630*** (0.0093)	0.0149*** (0.0027)	-0.0334*** (0.0030)	-0.0561*** (0.0033)	-0.0385*** (0.0024)	0.0044** (0.0017)	-0.0549*** (0.0030)	0.0009 (0.0012)	-0.0124*** (0.0037)	-0.0471*** (0.0029)	-0.0303*** (0.0021)
prod_avgprice	-0.0012*** (0.0001)	-0.0071*** (0.0004)	-0.0020*** (0.0001)	-0.0024*** (0.0001)	-0.0026*** (0.0001)	-0.0022*** (0.0001)	-0.0011*** (0.0001)	-0.0016*** (0.0001)	-0.0001*** (0.0000)	-0.0009*** (0.0001)	-0.0013*** (0.0002)	-0.0014*** (0.0001)
cat4_hardware	1.0404*** (0.1496)	-2.6518*** (0.8860)	-0.5337* (0.2865)	0.9614*** (0.2829)	0.1723 (0.3290)	-0.4544** (0.2145)	1.1807*** (0.1543)	0.5993** (0.2489)	1.1512*** (0.1108)	1.5912*** (0.3048)	1.8142*** (0.3030)	1.1084*** (0.2018)
cat9_videofootv	-1.1097*** (0.1741)	5.1928*** (1.0320)	-0.2952 (0.3306)	1.8604*** (0.3691)	1.8326*** (0.4210)	-0.9253*** (0.2491)	-0.7105*** (0.1861)	0.4165 (0.3007)	-0.2054 (0.1265)	-0.2529 (0.3484)	-0.9387** (0.3777)	-1.3519*** (0.2513)
Constant	-7.2988*** (0.2418)	-30.9136*** (1.4318)	-11.0555*** (0.4508)	-15.3548*** (0.4568)	-10.7148*** (0.5223)	-8.6120*** (0.3563)	-5.7102*** (0.2503)	-6.7489*** (0.4181)	-2.5522*** (0.1747)	-5.7112*** (0.4848)	-8.2217*** (0.4609)	-5.7173*** (0.3102)
Observations	12216	12218	8183	14710	11678	11512	12640	10574	3349	2984	5316	5288
R ²	0.1946	0.1386	0.0864	0.1007	0.0894	0.1050	0.1518	0.0820	0.1493	0.0640	0.1254	0.1404

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 34: Full dataset regression results - version 3 (cf. section 3.4.1 on page 56)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c_elast-choke	L_elast	cv_c_elast	cv_L_elast	cv_Lctw-c_elast	cv_Lctw-L_elast	aggr_Lctw-c_elast	aggr_Lctw-L_elast
prod_recommendation	-2.0872*** (0.2898)	-1.9572 (1.8625)	-1.2242** (0.5489)	-0.8543 (0.5427)	-0.7732 (0.6056)	-2.3972*** (0.4646)	-1.9019*** (0.3024)	-1.8845*** (0.5358)	-0.6784*** (0.1759)	-1.0566* (0.5393)	-2.1515*** (0.5324)	-0.5238 (0.3560)
ssk_numofferedprods	0.0041*** (0.0003)	0.0181*** (0.0018)	0.0055*** (0.0006)	0.0076*** (0.0005)	0.0044*** (0.0006)	0.0036*** (0.0004)	0.0025*** (0.0003)	0.0018*** (0.0005)	-0.0000 (0.0002)	-0.0004 (0.0006)	0.0040*** (0.0006)	0.0025*** (0.0004)
brand1to20	-1.1469*** (0.1362)	-12.5554*** (0.8694)	-2.0045*** (0.2592)	-3.3098*** (0.2438)	-2.0503*** (0.2799)	-1.3357*** (0.2068)	-1.1485*** (0.1477)	-1.3302*** (0.2523)	0.0716 (0.1068)	0.3145 (0.3286)	-0.5822** (0.2691)	-0.5918*** (0.1799)
ssk_numratings	-0.0155*** (0.0003)	-0.0720*** (0.0020)	-0.0149*** (0.0006)	-0.0184*** (0.0006)	-0.0127*** (0.0007)	-0.0114*** (0.0005)	-0.0123*** (0.0004)	-0.0092*** (0.0006)	-0.0021*** (0.0003)	-0.0040*** (0.0010)	-0.0103*** (0.0008)	-0.0057*** (0.0006)
prod_numretailer	0.0182*** (0.0017)	-0.1617*** (0.0106)	0.0238*** (0.0032)	-0.0280*** (0.0033)	-0.0513*** (0.0039)	-0.0344*** (0.0027)	0.0073*** (0.0018)	-0.0506*** (0.0034)	0.0030*** (0.0011)	-0.0099*** (0.0036)	-0.0438*** (0.0033)	-0.0294*** (0.0022)
prod_avgprice	-0.0015*** (0.0001)	-0.0082*** (0.0005)	-0.0025*** (0.0002)	-0.0028*** (0.0001)	-0.0033*** (0.0002)	-0.0029*** (0.0001)	-0.0014*** (0.0001)	-0.0022*** (0.0001)	-0.0003*** (0.0001)	-0.0019*** (0.0003)	-0.0023*** (0.0002)	-0.0024*** (0.0002)
Constant	-5.1930*** (0.2586)	-27.2065*** (1.6571)	-11.0055*** (0.5145)	-14.0750*** (0.4869)	-10.3314*** (0.5584)	-7.8772*** (0.4102)	-3.8348*** (0.2640)	-5.2332*** (0.4610)	-1.0750*** (0.1565)	-3.1224*** (0.4779)	-5.6924*** (0.5085)	-3.9741*** (0.3394)
Observations	8533	8534	5704	10584	8511	7989	8806	7401	2238	1983	3781	3760
R ²	0.2745	0.2376	0.1600	0.1598	0.1163	0.1801	0.1777	0.1167	0.0422	0.0467	0.1347	0.1535

Estimated using OLS. Standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1

Table 35: Enhanced model, cleansed dataset regression results - version 1 (cf. section 3.4.3 on page 63)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c_elast_choke	L_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_L_elast	aggr_lctw_c_elast	aggr_lctw_L_elast
prod_recommendation	7.1460 (4.4012)	19.7329 (19.7033)	8.6232* (4.5837)	-1.1475 (5.7703)	0.9507 (5.4291)	4.5767 (3.4858)	8.1856** (3.7779)	5.9262 (3.6319)	11.5594* (5.7361)	0.0115 (7.5105)	-1.5586 (14.7544)	-4.1623 (6.7225)
prod_avgmissing	-3.4079* (1.7362)	-14.3193* (7.7724)	-3.8835** (1.8329)	-4.4798** (2.2099)	-3.6140* (2.0569)	-3.8725*** (1.1812)	-0.9995 (1.4805)	-3.6465*** (1.2778)	1.8473 (3.7983)	1.2034 (6.5666)	-9.2660 (6.1631)	-5.5136* (2.8834)
ssk_numofferedprods	0.0001 (0.0032)	0.0220 (0.0143)	0.0014 (0.0032)	0.0012 (0.0040)	0.0015 (0.0037)	0.0016 (0.0022)	0.0017 (0.0026)	-0.0022 (0.0023)	0.0049 (0.0033)	0.0038 (0.0050)	0.0061 (0.0065)	0.0006 (0.0029)
prod_brandrank	0.0091*** (0.0018)	0.0294*** (0.0080)	0.0079*** (0.0018)	0.0111*** (0.0024)	0.0099*** (0.0022)	0.0058*** (0.0012)	0.0070*** (0.0015)	0.0056*** (0.0013)	0.0017 (0.0018)	0.0002 (0.0036)	0.0092* (0.0050)	0.0053** (0.0022)
ssk_numratings	0.0012 (0.0027)	0.0012 (0.0121)	-0.0005 (0.0027)	-0.0012 (0.0037)	-0.0023 (0.0035)	-0.0010 (0.0019)	-0.0024 (0.0022)	-0.0026 (0.0022)	-0.0018 (0.0023)	0.0020 (0.0033)	-0.0047 (0.0049)	-0.0027 (0.0022)
prod_numretailer	-0.1773*** (0.0365)	-0.7781*** (0.1633)	-0.1750*** (0.0375)	-0.1737*** (0.0475)	-0.1510*** (0.0443)	-0.1126*** (0.0288)	-0.1358*** (0.0322)	-0.1208*** (0.0296)	-0.0099 (0.0540)	0.0218 (0.0891)	-0.0928 (0.1098)	-0.0338 (0.0533)
prod_avgprice	-0.0015 (0.0011)	-0.0126** (0.0051)	-0.0012 (0.0011)	-0.0011 (0.0014)	-0.0011 (0.0013)	0.0001 (0.0008)	-0.0007 (0.0010)	0.0005 (0.0007)	0.0023 (0.0024)	0.0023 (0.0033)	0.0003 (0.0037)	-0.0002 (0.0019)
cat4_hardware	-1.1093 (2.0861)	-0.2940 (9.3389)	-0.3064 (2.1519)	3.3985 (3.2134)	4.0774 (3.1089)	-0.3496 (1.5764)	-0.7347 (1.8031)	0.4030 (1.6222)	-0.0678 (2.0136)	0.3647 (2.7989)	16.1474*** (5.1023)	7.4034*** (2.2295)
cat9_videofotv	0.8014 (2.3268)	10.6231 (10.4168)	1.4325 (2.3646)	6.6211* (3.4839)	7.1787** (3.3431)	2.2818 (1.6597)	1.8424 (1.9417)	2.7755* (1.5912)	-0.9488 (1.7772)	-4.1145 (2.5190)	14.0117** (6.3091)	6.1008** (2.8529)
Constant	-14.0201*** (4.3386)	-56.4557*** (19.4227)	-15.5566*** (4.4690)	-15.1865** (5.9233)	-17.2438*** (5.7080)	-10.2741*** (3.4666)	-15.7972*** (3.6709)	-7.8514** (3.6517)	-17.5000*** (5.5739)	-8.8629 (7.6122)	-24.5185* (14.2366)	-7.6922 (6.9545)
Observations	173	173	168	139	137	160	179	133	52	34	40	40
R ²	0.3146	0.3078	0.2977	0.2989	0.2926	0.3024	0.2702	0.3445	0.1994	0.2021	0.5032	0.5004

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 36: Enhanced model, cleansed dataset regression results - version 2 (cf. section 3.4.3 on page 63)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c-elast_choke	L_elast	cv_c_elast	cv_l_elast	cv_lctw-c_elast	cv_lctw-L_elast	aggr_lctw-c_elast	aggr_lctw-L_elast
prod_recommendation	11.6379*** (3.5262)	20.0199 (19.2692)	13.0545*** (3.6682)	2.9569 (5.0233)	1.5641 (5.1256)	4.5325* (2.2984)	8.6059** (3.7019)	6.9583** (3.0475)	5.8880 (3.9033)	0.1347 (4.2970)	-2.1859 (10.7832)	1.3969 (4.5405)
prod_avgmissing	-2.0752 (1.4086)	-13.4537* (7.6069)	-2.3883 (1.4929)	-4.5479** (1.9076)	-3.6881* (1.9412)	-2.0138** (0.8072)	-0.3723 (1.4659)	-2.8641*** (1.0789)	1.6367 (2.3466)	2.1654 (3.7638)	-5.6749 (4.0647)	-3.6966* (1.9319)
ssk_numofferedprods	0.0017 (0.0026)	0.0233* (0.0140)	0.0030 (0.0025)	0.0003 (0.0034)	0.0011 (0.0035)	0.0011 (0.0015)	0.0029 (0.0026)	-0.0036* (0.0019)	0.0011 (0.0022)	-0.0040 (0.0031)	0.0003 (0.0044)	-0.0008 (0.0019)
prod_brandrank	0.0073*** (0.0014)	0.0285*** (0.0078)	0.0061*** (0.0014)	0.0108*** (0.0020)	0.0095*** (0.0021)	0.0039*** (0.0008)	0.0065*** (0.0015)	0.0054*** (0.0011)	0.0011 (0.0011)	0.0016 (0.0021)	0.0078** (0.0033)	0.0045*** (0.0015)
ssk_numratings	-0.0010 (0.0021)	0.0017 (0.0118)	-0.0027 (0.0021)	-0.0020 (0.0032)	-0.0020 (0.0033)	-0.0020 (0.0013)	-0.0025 (0.0022)	-0.0010 (0.0019)	-0.0015 (0.0016)	0.0011 (0.0020)	0.0001 (0.0034)	-0.0023 (0.0015)
prod_numretailer	-0.1337*** (0.0294)	-0.7907*** (0.1597)	-0.1324*** (0.0303)	-0.1841*** (0.0412)	-0.1673*** (0.0420)	-0.0422** (0.0202)	-0.1178*** (0.0322)	-0.1273*** (0.0248)	-0.0090 (0.0377)	-0.0888 (0.0527)	-0.1398* (0.0729)	-0.0852** (0.0362)
prod_avgprice	-0.0001 (0.0009)	-0.0119** (0.0050)	0.0002 (0.0009)	-0.0012 (0.0012)	-0.0009 (0.0012)	0.0003 (0.0005)	-0.0002 (0.0010)	0.0000 (0.0006)	0.0002 (0.0015)	0.0004 (0.0019)	0.0015 (0.0024)	0.0010 (0.0013)
cat4_hardware	-0.2109 (1.6388)	-0.4317 (9.1332)	0.5997 (1.6923)	0.6196 (2.8514)	1.1311 (3.0216)	1.4318 (1.0222)	-0.8243 (1.7658)	1.0222 (1.3591)	0.8790 (1.2769)	0.1222 (1.6517)	2.5061 (3.9423)	1.6369 (1.7395)
cat9_videofotv	0.7365 (1.8779)	9.5236 (10.1942)	1.0140 (1.9060)	4.8913 (3.0928)	4.1232 (3.2427)	1.7077 (1.0814)	1.1248 (1.9175)	2.8142** (1.3283)	1.2181 (1.2658)	0.4338 (1.6456)	0.8817 (4.5883)	0.9692 (2.0593)
Constant	-18.8950*** (3.4757)	-56.7626*** (18.9949)	-20.2792*** (3.5769)	-14.2317*** (5.1679)	-14.1055** (5.4414)	-11.5035*** (2.2957)	-17.1401*** (3.6245)	-8.0367*** (3.0603)	-10.5969*** (3.8070)	-1.1136 (4.5498)	-8.3753 (10.1902)	-4.5301 (4.6344)
Observations	169	172	164	137	136	150	178	131	49	31	38	39
R ²	0.3420	0.3178	0.3325	0.3680	0.3131	0.2611	0.2538	0.4049	0.1346	0.2361	0.4438	0.5986

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 37: Enhanced model, cleansed dataset regression results - version 3 (cf. section 3.4.3 on page 63)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c_elast-choke	L_elast	cv_c_elast	cv_L_elast	cv_Lctw-c_elast	cv_Lctw-L_elast	aggr_Lctw-c_elast	aggr_Lctw-L_elast
prod_recommendation	4.6435 (4.9233)	-3.7837 (27.9462)	5.5263 (5.0635)	-2.2823 (6.8905)	-1.1931 (6.5521)	6.6112** (3.2354)	8.3913 (5.5485)	6.7331 (6.9253)	4.1121 (6.5676)	-6.3266 (14.0437)	5.2695 (17.0013)	-0.7059 (7.1939)
prod_avgmissing	0.1883 (1.8333)	-13.5512 (9.8643)	0.1458 (1.9373)	-1.1549 (2.4735)	0.1627 (2.3326)	-2.3211** (1.0953)	0.3810 (1.8877)	-3.1888* (1.6653)	4.3524 (2.8823)	-11.3177 (8.5194)	-3.9499 (4.2420)	-3.8913* (1.9254)
ssk_numofferedprods	-0.0005 (0.0050)	0.0074 (0.0281)	0.0007 (0.0050)	0.0028 (0.0061)	0.0058 (0.0058)	-0.0008 (0.0027)	0.0109** (0.0050)	-0.0051 (0.0055)	0.0084 (0.0052)	-0.0242 (0.0149)	0.0013 (0.0121)	-0.0097* (0.0054)
prod_brandrank	0.0082*** (0.0026)	0.0368** (0.0145)	0.0070** (0.0027)	0.0092*** (0.0033)	0.0069** (0.0031)	0.0057*** (0.0015)	0.0048* (0.0027)	0.0070** (0.0027)	-0.0060 (0.0045)	0.0227 (0.0126)	0.0026 (0.0069)	0.0074** (0.0030)
ssk_numratings	-0.0001 (0.0050)	-0.0175 (0.0287)	-0.0028 (0.0054)	-0.0037 (0.0065)	-0.0069 (0.0062)	-0.0026 (0.0027)	-0.0081 (0.0052)	0.0002 (0.0058)	-0.0031 (0.0040)	0.0098 (0.0120)	-0.0048 (0.0112)	0.0032 (0.0048)
prod_numretailer	-0.1954*** (0.0421)	-0.8786*** (0.2272)	-0.2023*** (0.0437)	-0.2292*** (0.0579)	-0.2125*** (0.0547)	-0.0660** (0.0256)	-0.1176** (0.0495)	-0.1574*** (0.0432)	0.3652** (0.1464)	-0.4495 (0.2615)	-0.0290 (0.1280)	-0.0853 (0.0573)
prod_avgprice	-0.0048** (0.0019)	-0.0310*** (0.0100)	-0.0043** (0.0021)	-0.0097*** (0.0025)	-0.0099*** (0.0024)	-0.0007 (0.0010)	-0.0077*** (0.0020)	-0.0016 (0.0015)	0.0030 (0.0047)	-0.0116 (0.0161)	0.0101 (0.0097)	0.0013 (0.0037)
Constant	-10.4409** (4.6003)	-17.3264 (25.0693)	-10.1379** (4.6035)	-6.0979 (6.3199)	-6.9021 (5.9979)	-9.6239*** (2.9489)	-16.3956*** (5.1104)	-4.6063 (6.2826)	-20.8397** (9.4207)	21.2344 (19.8778)	-14.5743 (16.1542)	0.4027 (6.9918)
Observations	78	80	75	72	71	68	78	58	19	12	23	23
R ²	0.5721	0.5340	0.5788	0.6076	0.6156	0.6188	0.5061	0.5930	0.6024	0.6395	0.4623	0.5823

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 38: Enhanced model, full dataset regression results - version 1 (cf. section 3.4.2 on page 60)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c-elast_choke	L_elast	cv_c_elast	cv_l_elast	cv_lctw-c_elast	cv_lctw-L_elast	aggr_lctw-c_elast	aggr_lctw-L_elast
prod_recommendation	-1.7411*** (0.2794)	-1.5115 (1.5425)	-0.4257 (0.5096)	-0.6483 (0.4897)	-1.0530* (0.5688)	-2.0786*** (0.5199)	-1.7877* (1.0319)	-0.5314 (5.2340)	-0.4772** (0.2217)	-2.1668*** (0.8006)	-1.4083*** (0.4532)	-0.2671 (0.3129)
prod_avgmissing	1.5466*** (0.1226)	-0.5641 (0.6767)	-0.5351** (0.2318)	-1.7037*** (0.2158)	-1.4312*** (0.2538)	-0.5741*** (0.2212)	1.0866** (0.4595)	-4.6082** (2.2626)	0.6224*** (0.1064)	0.3965 (0.3844)	1.0771*** (0.2216)	-0.0988 (0.1511)
ssk_numofferedprods	0.0030*** (0.0002)	0.0123*** (0.0012)	0.0039*** (0.0004)	0.0043*** (0.0004)	0.0022*** (0.0005)	0.0017*** (0.0004)	0.0031*** (0.0008)	0.0066* (0.0036)	0.0008*** (0.0002)	0.0021*** (0.0006)	0.0031*** (0.0004)	0.0022*** (0.0002)
prod_brandrank	0.0028*** (0.0002)	0.0128*** (0.0011)	0.0049*** (0.0004)	0.0066*** (0.0004)	0.0060*** (0.0005)	0.0041*** (0.0004)	0.0022*** (0.0007)	0.0046 (0.0035)	0.0001 (0.0001)	0.0005 (0.0006)	0.0034*** (0.0003)	0.0025*** (0.0002)
ssk_numratings	-0.0073*** (0.0002)	-0.0277*** (0.0013)	-0.0067*** (0.0004)	-0.0097*** (0.0005)	-0.0068*** (0.0005)	-0.0041*** (0.0004)	-0.0076*** (0.0009)	-0.0141*** (0.0046)	-0.0020*** (0.0002)	-0.0040*** (0.0008)	-0.0040*** (0.0005)	-0.0026*** (0.0003)
prod_numretailer	0.0255*** (0.0017)	-0.1737*** (0.0095)	0.0175*** (0.0032)	-0.0330*** (0.0032)	-0.0588*** (0.0038)	-0.0344*** (0.0033)	0.0348*** (0.0067)	0.0520 (0.0354)	0.0066*** (0.0015)	0.0084 (0.0058)	-0.0411*** (0.0030)	-0.0283*** (0.0021)
prod_avgprice	-0.0014*** (0.0001)	-0.0075*** (0.0004)	-0.0023*** (0.0002)	-0.0026*** (0.0001)	-0.0029*** (0.0002)	-0.0028*** (0.0001)	-0.0014*** (0.0003)	-0.0027** (0.0012)	-0.0002*** (0.0001)	-0.0012*** (0.0002)	-0.0013*** (0.0002)	-0.0015*** (0.0001)
cat4_hardware	1.1677*** (0.1652)	0.0259 (0.9119)	-0.1857 (0.3346)	2.1319*** (0.3075)	1.2899*** (0.3768)	0.4662 (0.2942)	1.0220* (0.6150)	0.6736 (2.9468)	1.1207*** (0.1330)	0.6093 (0.4859)	2.4520*** (0.3070)	1.5384*** (0.2074)
cat9_videofootv	-0.5295*** (0.1957)	6.7607*** (1.0805)	0.3880 (0.3913)	2.9707*** (0.3961)	3.2873*** (0.4840)	1.3464*** (0.3483)	-1.9590*** (0.7518)	-6.7767* (3.6150)	-0.0843 (0.1586)	-0.4808 (0.5814)	0.0374 (0.3902)	-0.8091*** (0.2637)
Constant	-9.9854*** (0.2898)	-38.2160*** (1.5998)	-13.6240*** (0.5565)	-17.2820*** (0.5345)	-12.3208*** (0.6416)	-11.2710*** (0.5353)	-9.1543*** (1.0785)	-12.2850** (5.3987)	-3.1885*** (0.2275)	-5.9819*** (0.8312)	-10.7307*** (0.5001)	-6.7700*** (0.3427)
Observations	12238	12238	8206	14735	11751	11617	12657	10619	3356	3003	5322	5302
R ²	0.1904	0.1310	0.0823	0.1011	0.0864	0.0849	0.0178	0.0033	0.1175	0.0334	0.1445	0.1547

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 39: Enhanced model, full dataset regression results - version 2 (cf. section 3.4.2 on page 60)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c_elast_choke	L_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_l_elast	aggr_lctw_c_elast	aggr_lctw_l_elast
prod_recommendation	-1.5407*** (0.2513)	-2.4555 (1.5110)	-1.3002** (0.4393)	-0.5716 (0.4560)	-0.7900 (0.5015)	-1.6789*** (0.3784)	-1.7532*** (0.2614)	-1.4143*** (0.4479)	-0.4342** (0.1795)	-0.7646 (0.4947)	-1.3748*** (0.4441)	-0.1596 (0.3023)
prod_avgmissing	1.6495*** (0.1104)	-0.6746 (0.6626)	-0.3933** (0.1995)	-1.5398*** (0.2010)	-1.0845*** (0.2241)	-0.4870*** (0.1618)	1.5506*** (0.1165)	0.1098 (0.1941)	0.7079*** (0.0863)	0.4204* (0.2376)	1.1452*** (0.2172)	-0.0638 (0.1460)
ssk_numofferedprods	0.0027*** (0.0002)	0.0123*** (0.0011)	0.0037*** (0.0004)	0.0043** (0.0003)	0.0027*** (0.0004)	0.0013*** (0.0003)	0.0022*** (0.0002)	0.0010*** (0.0003)	0.0006*** (0.0001)	0.0011*** (0.0004)	0.0029*** (0.0004)	0.0022*** (0.0002)
prod_brandrank	0.0025*** (0.0002)	0.0127*** (0.0011)	0.0045*** (0.0004)	0.0064*** (0.0004)	0.0057*** (0.0004)	0.0030*** (0.0003)	0.0017*** (0.0002)	0.0017*** (0.0003)	0.0000 (0.0001)	0.0006* (0.0003)	0.0033*** (0.0003)	0.0024*** (0.0002)
ssk_numratings	-0.0068*** (0.0002)	-0.0275*** (0.0012)	-0.0066*** (0.0004)	-0.0093*** (0.0004)	-0.0070*** (0.0005)	-0.0038*** (0.0003)	-0.0065*** (0.0002)	-0.0046*** (0.0004)	-0.0019*** (0.0002)	-0.0031*** (0.0005)	-0.0038*** (0.0005)	-0.0026*** (0.0003)
prod_numretailer	0.0201*** (0.0016)	-0.1771*** (0.0093)	0.0153*** (0.0027)	-0.0362*** (0.0030)	-0.0561*** (0.0034)	-0.0365*** (0.0024)	0.0066*** (0.0017)	-0.0535*** (0.0030)	0.0031** (0.0012)	-0.0086** (0.0036)	-0.0414*** (0.0029)	-0.0280*** (0.0020)
prod_avgprice	-0.0013*** (0.0001)	-0.0075*** (0.0004)	-0.0020*** (0.0001)	-0.0025*** (0.0001)	-0.0026*** (0.0001)	-0.0021*** (0.0001)	-0.0012*** (0.0001)	-0.0016*** (0.0001)	-0.0001*** (0.0000)	-0.0008*** (0.0001)	-0.0013*** (0.0002)	-0.0014*** (0.0001)
cat4_hardware	1.4144** (0.1485)	0.2080 (0.8927)	0.2226 (0.2868)	2.2632*** (0.2864)	1.2202*** (0.3325)	-0.0080 (0.2153)	1.5435*** (0.1558)	0.9349*** (0.2530)	1.1066*** (0.1082)	1.4299*** (0.3001)	2.4189*** (0.3010)	1.5360*** (0.2005)
cat9_videofootv	-0.6380*** (0.1759)	6.7575*** (1.0578)	0.3512 (0.3355)	2.8494** (0.3690)	2.7458*** (0.4270)	-0.3599 (0.2543)	-0.3534* (0.1905)	0.8152*** (0.3097)	-0.1078 (0.1295)	-0.0478 (0.3584)	0.0082 (0.3827)	-0.7714*** (0.2551)
Constant	-9.7508*** (0.2607)	-37.3340*** (1.5670)	-12.7391*** (0.4781)	-17.2399*** (0.4978)	-12.4733*** (0.5655)	-9.5468*** (0.3902)	-7.7889*** (0.2731)	-7.6982*** (0.4630)	-2.9084*** (0.1844)	-5.8608*** (0.5128)	-10.6479*** (0.4903)	-6.8558*** (0.3312)
Observations	12216	12218	8183	14710	11678	11512	12640	10574	3349	2984	5316	5288
R ²	0.2148	0.1357	0.0970	0.1108	0.0993	0.1143	0.1633	0.0844	0.1639	0.0619	0.1478	0.1577

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 40: Enhanced model, full dataset regression results - version 3 (cf. section 3.4.2 on page 60)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c_elast-choke	L_elast	cv_c_elast	cv_L_elast	cv_Lctw_c_elast	cv_Lctw-L_elast	aggr_Lctw_c_elast	aggr_Lctw-L_elast
prod_recommendation	-1.6870*** (0.2903)	-1.7298 (1.8843)	-1.2804** (0.5488)	-0.8275 (0.5413)	-0.7350 (0.6024)	-2.4499*** (0.4649)	-1.6156*** (0.3025)	-1.7196*** (0.5372)	-0.5464*** (0.1754)	-0.9677* (0.5411)	-1.8144*** (0.5279)	-0.4590 (0.3547)
prod_avgmissing	1.0023*** (0.1274)	-3.4545*** (0.8239)	-1.1584*** (0.2479)	-1.8024*** (0.2391)	-1.2948*** (0.2697)	-1.1723*** (0.1976)	0.9247*** (0.1354)	-0.0428 (0.2347)	0.5265*** (0.0851)	0.4953* (0.2664)	1.0209*** (0.2584)	-0.2195 (0.1722)
ssk_numofferedprods	0.0036*** (0.0003)	0.0182*** (0.0019)	0.0053*** (0.0006)	0.0076*** (0.0005)	0.0044*** (0.0006)	0.0039*** (0.0004)	0.0021*** (0.0003)	0.0018*** (0.0006)	-0.0002 (0.0002)	-0.0006 (0.0006)	0.0034*** (0.0006)	0.0024*** (0.0004)
prod_brandrank	0.0021*** (0.0002)	0.0170*** (0.0015)	0.0055*** (0.0005)	0.0074*** (0.0005)	0.0071*** (0.0006)	0.0041*** (0.0004)	0.0016*** (0.0003)	0.0020*** (0.0004)	0.0002 (0.0001)	0.0002 (0.0004)	0.0034*** (0.0004)	0.0022*** (0.0003)
ssk_numratings	-0.0145*** (0.0003)	-0.0732*** (0.0021)	-0.0150*** (0.0006)	-0.0192*** (0.0007)	-0.0130*** (0.0007)	-0.0117*** (0.0005)	-0.0116*** (0.0004)	-0.0091*** (0.0007)	-0.0014*** (0.0003)	-0.0033*** (0.0010)	-0.0084*** (0.0008)	-0.0055*** (0.0006)
prod_numretailer	0.0196*** (0.0016)	-0.1858*** (0.0106)	0.0225*** (0.0032)	-0.0332*** (0.0034)	-0.0516*** (0.0039)	-0.0350*** (0.0027)	0.0077*** (0.0018)	-0.0516*** (0.0034)	0.0041*** (0.0011)	-0.0082** (0.0035)	-0.0381*** (0.0032)	-0.0284*** (0.0022)
prod_avgprice	-0.0016*** (0.0001)	-0.0087*** (0.0005)	-0.0025*** (0.0002)	-0.0029*** (0.0001)	-0.0034*** (0.0002)	-0.0028*** (0.0001)	-0.0016*** (0.0001)	-0.0023*** (0.0001)	-0.0003*** (0.0001)	-0.0019*** (0.0003)	-0.0023*** (0.0002)	-0.0024*** (0.0002)
Constant	-6.6741*** (0.2828)	-29.9617*** (1.8312)	-11.5802*** (0.5525)	-14.5041*** (0.5356)	-11.0735*** (0.6097)	-8.1891*** (0.4494)	-5.0735*** (0.2875)	-5.9400*** (0.5057)	-1.4806*** (0.1677)	-3.4777*** (0.5137)	-7.4785*** (0.5443)	-4.4820*** (0.3663)
Observations	8533	8534	5704	10584	8511	7989	8806	7401	2238	1983	3781	3760
R ²	0.2818	0.2309	0.1701	0.1658	0.1276	0.1906	0.1810	0.1160	0.0608	0.0483	0.1548	0.1643

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 41: Enhanced model, relaxed filter cleansed dataset regression results - version 2 (cf. section 3.4.2 on page 60)

LHS-Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast-inv	c_elast-choke	aggr_c_elast	aggr_c_elast-choke	L_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_L_elast	aggr_lctw_c_elast	aggr_lctw_L_elast
prod_recommendation	-0.3268 (1.7889)	-3.1216 (9.3242)	-0.5850 (2.2978)	-2.1310 (2.9169)	-0.4800 (2.7098)	-4.2037** (1.8017)	-0.1910 (2.0548)	-3.4751* (2.0378)	0.7375 (1.2218)	5.3162*** (1.7615)	4.1811 (3.0082)	1.5114 (1.6668)
prod_avgmissing	-2.7451*** (0.7772)	-16.0180*** (3.9305)	-5.8351*** (0.9735)	-4.1859*** (1.1827)	-3.5960*** (1.1130)	-3.6842*** (0.7043)	-2.1403** (0.8759)	-3.4130*** (0.8466)	0.6370 (0.8482)	1.7861 (1.1904)	-0.4613 (1.4610)	-0.6380 (0.7920)
ssk_numofferedprods	0.0041*** (0.0014)	0.0266*** (0.0068)	0.0062*** (0.0017)	0.0054*** (0.0020)	0.0044** (0.0018)	0.0001 (0.0012)	0.0024 (0.0015)	-0.0006 (0.0014)	-0.0011 (0.0010)	-0.0006 (0.0016)	0.0020 (0.0018)	0.0005 (0.0009)
prod_brandrank	0.0055*** (0.0008)	0.0235*** (0.0041)	0.0070*** (0.0010)	0.0076*** (0.0012)	0.0070*** (0.0011)	0.0051*** (0.0008)	0.0058*** (0.0009)	0.0045*** (0.0010)	0.0010* (0.0006)	0.0007 (0.0009)	0.0023** (0.0011)	0.0020*** (0.0006)
ssk_numratings	-0.0032*** (0.0012)	-0.0123** (0.0060)	-0.0048*** (0.0015)	-0.0055*** (0.0018)	-0.0052*** (0.0017)	-0.0011 (0.0012)	-0.0028** (0.0013)	-0.0025* (0.0015)	-0.0017* (0.0009)	0.0003 (0.0017)	-0.0043** (0.0017)	-0.0023** (0.0009)
prod_numretailer	-0.0801*** (0.0170)	-0.4795*** (0.0881)	-0.0989*** (0.0211)	-0.0755*** (0.0261)	-0.0798*** (0.0244)	-0.0700*** (0.0189)	-0.0714*** (0.0201)	-0.0768*** (0.0233)	0.0181 (0.0173)	0.0226 (0.0302)	-0.0492 (0.0318)	-0.0044 (0.0175)
prod_avgprice	-0.0018*** (0.0006)	-0.0096*** (0.0031)	-0.0023*** (0.0007)	-0.0026*** (0.0009)	-0.0025*** (0.0008)	-0.0015*** (0.0005)	-0.0019*** (0.0007)	-0.0009 (0.0007)	0.0007 (0.0008)	-0.0015 (0.0015)	0.0012 (0.0015)	-0.0000 (0.0008)
cat4_hardware	0.8646 (0.8807)	-9.8467** (4.5796)	-0.5819 (1.1254)	-0.5773 (1.4983)	-0.3279 (1.4286)	-0.8560 (0.8605)	1.3432 (1.0583)	0.9200 (1.0562)	1.2667** (0.6385)	0.0449 (0.9852)	1.9811 (1.3513)	1.1655* (0.7032)
cat9_videofotv	0.7133 (0.9986)	0.9435 (5.1786)	1.6916 (1.2505)	3.7034** (1.7164)	3.5758** (1.6248)	2.8820*** (0.9494)	2.6997** (1.1564)	3.8402*** (1.1372)	0.4275 (0.7070)	-0.8531 (1.1102)	-0.2468 (1.7392)	0.0503 (0.8607)
Constant	-10.8410*** (1.8017)	-40.0831*** (9.3614)	-10.4218*** (2.2637)	-12.6145*** (2.9463)	-12.5330*** (2.7666)	-3.7156** (1.7975)	-11.5803*** (2.0791)	-3.4672 (2.1420)	-5.8436*** (1.2396)	-9.1712*** (1.7062)	-13.7114*** (2.8852)	-7.0134*** (1.5731)
Observations	463	471	448	406	391	456	494	375	157	129	147	149
R ²	0.2289	0.2249	0.2535	0.2020	0.1997	0.2119	0.1401	0.1482	0.1122	0.1358	0.2098	0.2180

Estimated using OLS. Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 42: Price leader changes, Poisson first regression - version 1 (cf. section 4.4.2 on page 79)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast-choke	aggr_c_elast	aggr_c_elast	aggr_c_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_l_elast	aggr_lctw_c_elast	aggr_lctw_l_elast
elasticity	-0.0035*** (0.0010)	-0.0000*** (0.0000)	-0.0001 (0.0011)	-0.0122*** (0.0012)	-0.0147*** (0.0014)	0.0050*** (0.0016)	0.0066*** (0.0011)	0.0113*** (0.0017)	0.0049 (0.0046)	0.0693*** (0.0052)	-0.0381*** (0.0022)	-0.0632*** (0.0060)
Constant	4.0433*** (0.0153)	4.0643*** (0.0095)	4.0081*** (0.0161)	3.2955*** (0.0273)	3.1855*** (0.0298)	4.0512*** (0.0148)	4.1676*** (0.0163)	4.0807*** (0.0151)	4.0576*** (0.0294)	4.3019*** (0.0274)	2.5748*** (0.0528)	2.7829*** (0.0521)
Observations	193	193	187	139	137	191	190	191	59	62	40	40

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 43: Price leader changes, Poisson first regression - version 2 (cf. section 4.4.2 on page 79)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast-choke	aggr_c_elast	aggr_c_elast	aggr_c_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_l_elast	aggr_lctw_c_elast	aggr_lctw_l_elast
elasticity	-0.0103*** (0.0012)	-0.0032*** (0.0001)	-0.0053*** (0.0013)	-0.0179*** (0.0013)	-0.0180*** (0.0014)	-0.0049** (0.0023)	0.0072*** (0.0012)	0.0193*** (0.0019)	0.0523*** (0.0068)	0.0394*** (0.0062)	-0.0193*** (0.0058)	-0.1471*** (0.0090)
Constant	3.9806*** (0.0165)	3.8115*** (0.0171)	3.9609*** (0.0173)	3.2059*** (0.0292)	3.1322*** (0.0308)	3.9992*** (0.0175)	4.1734*** (0.0165)	4.1252*** (0.0157)	4.2340*** (0.0337)	4.2023*** (0.0304)	2.7484*** (0.0783)	2.3362*** (0.0681)
Observations	189	181	183	137	136	181	189	189	56	59	38	39

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 44: Price leader changes, Poisson first regression, cleansed dataset - version 3 (cf. section 4.4.2 on page 79)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast-choke	aggr_c_elast	aggr_c_elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_l_elast	aggr_lctw_c_elast	aggr_lctw_l_elast
elasticity	-0.0180*** (0.0016)	-0.0129*** (0.0002)	-0.0165*** (0.0016)	-0.0281*** (0.0016)	-0.0274*** (0.0017)	-0.0155*** (0.0032)	0.0143*** (0.0017)	0.0221*** (0.0025)	0.0991*** (0.0145)	0.1430*** (0.0141)	-0.0177 (0.0109)	-0.0060 (0.0233)
Constant	3.9435*** (0.0227)	2.9266*** (0.0293)	3.8992*** (0.0238)	2.8724*** (0.0404)	2.9056*** (0.0413)	4.0339*** (0.0244)	4.2945*** (0.0226)	4.1876*** (0.0212)	3.9609*** (0.0593)	3.9526*** (0.0485)	2.4562*** (0.1242)	2.6091*** (0.1084)
Observations	86	81	82	72	71	78	83	83	21	25	23	23

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 45: Price leader changes, Poisson first regression, Full Dataset - version 1 (cf. section 4.4.2 on page 79)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast-choke	aggr_c_elast	aggr_c_elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_l_elast	aggr_lctw_c_elast	aggr_lctw_l_elast
elasticity	-0.0231*** (0.0002)	-0.0000*** (0.0000)	-0.0116*** (0.0002)	-0.0035*** (0.0001)	-0.0049*** (0.0001)	-0.0071*** (0.0002)	-0.0015*** (0.0001)	0.0001** (0.0000)	-0.0341*** (0.0013)	0.0024*** (0.0005)	-0.0069*** (0.0003)	-0.0006 (0.0005)
Constant	3.2540*** (0.0020)	3.3649*** (0.0015)	3.2973*** (0.0023)	3.2998*** (0.0024)	3.2995*** (0.0026)	3.3052*** (0.0022)	3.3682*** (0.0016)	3.3783*** (0.0016)	3.2724*** (0.0033)	3.3430*** (0.0035)	3.2485*** (0.0042)	3.3127*** (0.0043)
Observations	14836	14836	10638	14735	11753	15046	14789	14845	4831	4854	5335	5336

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 46: Price leader changes, Poisson first regression Full Dataset - version 2 (cf. section 4.4.2 on page 79)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast-choke	aggr_c_elast	aggr_c_elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_l_elast	aggr_lctw_c_elast	aggr_lctw_l_elast
elasticity	-0.0245*** (0.0003)	-0.0008*** (0.0000)	-0.0157*** (0.0002)	-0.0031*** (0.0001)	-0.0048*** (0.0001)	-0.0058*** (0.0002)	-0.0199*** (0.0002)	-0.0053*** (0.0002)	-0.0361*** (0.0015)	0.0000 (0.0005)	-0.0071*** (0.0003)	-0.0015*** (0.0005)
Constant	3.2479*** (0.0021)	3.3037*** (0.0025)	3.2667*** (0.0025)	3.3058*** (0.0028)	3.3002*** (0.0028)	3.3141*** (0.0023)	3.2707*** (0.0021)	3.3269*** (0.0023)	3.2697*** (0.0034)	3.3342*** (0.0037)	3.2466*** (0.0043)	3.3071*** (0.0043)
Observations	14814	14264	10615	14710	11680	14940	14771	14795	4824	4833	5329	5322

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 47: Price leader changes, Poisson first regression full dataset - version 3 (cf. section 4.4.2 on page 79)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast-choke	aggr_c_elast	aggr_c_elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw_c_elast	cv_lctw_l_elast	aggr_lctw_c_elast	aggr_lctw_l_elast
elasticity	-0.0251*** (0.0004)	-0.0020*** (0.0000)	-0.0173*** (0.0003)	-0.0075*** (0.0001)	-0.0081*** (0.0002)	-0.0096*** (0.0002)	-0.0225*** (0.0003)	-0.0113*** (0.0002)	0.0237*** (0.0029)	-0.0027*** (0.0008)	-0.0123*** (0.0004)	-0.0075*** (0.0006)
Constant	3.1950*** (0.0025)	3.1293*** (0.0031)	3.1953*** (0.0030)	3.1747*** (0.0033)	3.1996*** (0.0034)	3.2156*** (0.0028)	3.2051*** (0.0025)	3.2071*** (0.0028)	3.1786*** (0.0045)	3.1662*** (0.0047)	3.0725*** (0.0053)	3.1381*** (0.0053)
Observations	10382	9984	7453	10584	8511	10500	10362	10413	3312	3326	3791	3787

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 48: Price leader changes, Poisson regression, full dataset - version 1 (cf. section 4.4.3 on page 80)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast- choke	aggr_c_elast	aggr_c- elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw- c_elast	cv_lctw- l_elast	aggr_lctw- c_elast	aggr_lctw- l_elast
elasticity	-0.0137*** (0.0003)	-0.0000*** (0.0000)	-0.0066*** (0.0002)	-0.0015*** (0.0001)	-0.0032*** (0.0001)	-0.0026*** (0.0002)	-0.0007*** (0.0001)	0.0004*** (0.0001)	0.0012 (0.0017)	0.0122*** (0.0005)	-0.0037*** (0.0004)	0.0084*** (0.0005)
prod_recommendation	-0.0315*** (0.0076)	-0.0134* (0.0076)	0.0491*** (0.0086)	-0.0524*** (0.0076)	-0.0287*** (0.0083)	-0.0531*** (0.0074)	0.0093 (0.0076)	-0.0469*** (0.0075)	-0.1205*** (0.0124)	-0.0961*** (0.0123)	-0.0566*** (0.0121)	-0.0523*** (0.0120)
prod_avgmissing	0.0272*** (0.0033)	0.0039 (0.0033)	0.0350*** (0.0037)	0.0411*** (0.0033)	0.0723*** (0.0036)	0.0075** (0.0033)	0.0259*** (0.0033)	0.0211*** (0.0033)	0.0772*** (0.0059)	0.0669*** (0.0058)	0.1438*** (0.0056)	0.1322*** (0.0055)
prod_subst_msize	0.0130*** (0.0002)	0.0148*** (0.0002)	0.0148*** (0.0002)	0.0126*** (0.0002)	0.0136*** (0.0002)	0.0145*** (0.0002)	0.0145*** (0.0002)	0.0151*** (0.0002)	0.0118*** (0.0003)	0.0142*** (0.0003)	0.0125*** (0.0003)	0.0129*** (0.0003)
prod_brandrank	-0.0193*** (0.0006)	-0.0216*** (0.0006)	-0.0384*** (0.0008)	-0.0311*** (0.0007)	-0.0320*** (0.0008)	-0.0187*** (0.0006)	-0.0223*** (0.0006)	-0.0239*** (0.0006)	-0.0297*** (0.0010)	-0.0315*** (0.0010)	-0.0336*** (0.0010)	-0.0365*** (0.0010)
prod_numclicks	0.0300*** (0.0012)	0.0422*** (0.0011)	0.0387*** (0.0011)	0.0070*** (0.0015)	0.0064*** (0.0015)	0.0417*** (0.0011)	0.0423** (0.0011)	0.0419*** (0.0011)	-0.0011 (0.0031)	0.0011 (0.0028)	-0.0173*** (0.0032)	-0.0160*** (0.0031)
prod_numretailer	0.0012*** (0.0000)	0.0010*** (0.0000)	0.0002*** (0.0001)	0.0014*** (0.0000)	0.0007*** (0.0001)	0.0009*** (0.0000)	0.0008*** (0.0000)	0.0008*** (0.0000)	0.0008*** (0.0001)	0.0007*** (0.0001)	0.0002*** (0.0001)	0.0006*** (0.0001)
prod_avgprice	0.0056*** (0.0002)	0.0064*** (0.0002)	0.0076*** (0.0002)	0.0033*** (0.0001)	0.0043*** (0.0002)	0.0045*** (0.0001)	0.0059*** (0.0002)	0.0051*** (0.0001)	0.0004 (0.0004)	0.0011*** (0.0004)	-0.0017*** (0.0004)	-0.0000 (0.0004)
cat4_hardware	-0.2907*** (0.0044)	-0.3042*** (0.0043)	-0.2180*** (0.0054)	-0.2283*** (0.0045)	-0.1090*** (0.0054)	-0.2909*** (0.0043)	-0.3051*** (0.0043)	-0.2982*** (0.0043)	-0.5157*** (0.0076)	-0.5108*** (0.0076)	-0.3816*** (0.0076)	-0.4035*** (0.0077)
cat9_videofotov	-0.2451*** (0.0057)	-0.2388*** (0.0057)	-0.1238*** (0.0068)	-0.1285*** (0.0059)	0.0165** (0.0068)	-0.2261*** (0.0057)	-0.2474*** (0.0057)	-0.2588*** (0.0057)	-0.1422*** (0.0093)	-0.1443*** (0.0093)	-0.0341*** (0.0093)	-0.0299*** (0.0093)
Constant	3.3918*** (0.0078)	3.4644*** (0.0076)	3.3422*** (0.0092)	3.3956*** (0.0081)	3.2741*** (0.0090)	3.4741*** (0.0076)	3.4596*** (0.0077)	3.5130*** (0.0076)	3.6584*** (0.0128)	3.7017*** (0.0126)	3.4961*** (0.0129)	3.5815*** (0.0129)
Observations	14836	14836	10638	14735	11753	15046	14789	14845	4831	4854	5335	5336

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 49: Price leader changes, Poisson regression, full dataset - version 2 (cf. section 4.4.3 on page 80)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast- choke	aggr_c_elast	aggr_c- elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw- c_elast	cv_lctw- l_elast	aggr_lctw- c_elast	aggr_lctw- l_elast
elasticity	-0.0135*** (0.0003)	-0.0004*** (0.0000)	-0.0085*** (0.0002)	-0.0002 (0.0001)	-0.0024*** (0.0001)	0.0004** (0.0002)	-0.0103*** (0.0003)	-0.0006*** (0.0002)	0.0020 (0.0018)	0.0122*** (0.0006)	-0.0039*** (0.0004)	0.0079*** (0.0005)
prod_recommendation	-0.0156** (0.0076)	0.0091 (0.0078)	0.0607*** (0.0087)	-0.0548*** (0.0077)	-0.0342*** (0.0083)	-0.0590*** (0.0075)	-0.0021 (0.0076)	-0.0474*** (0.0075)	-0.1042*** (0.0124)	-0.0971*** (0.0123)	-0.0599*** (0.0121)	-0.0526*** (0.0121)
prod_avgmissing	0.0284*** (0.0033)	0.0219*** (0.0034)	0.0421*** (0.0038)	0.0403*** (0.0033)	0.0728*** (0.0036)	0.0107*** (0.0033)	0.0409*** (0.0033)	0.0170*** (0.0033)	0.0723*** (0.0059)	0.0682*** (0.0058)	0.1460*** (0.0056)	0.1340*** (0.0056)
prod_subst_msize	0.0132*** (0.0002)	0.0143*** (0.0002)	0.0147*** (0.0002)	0.0128*** (0.0002)	0.0136*** (0.0002)	0.0147*** (0.0002)	0.0132*** (0.0002)	0.0148*** (0.0002)	0.0124*** (0.0003)	0.0145*** (0.0003)	0.0126*** (0.0003)	0.0131*** (0.0003)
prod_brandrank	-0.0192*** (0.0006)	-0.0187*** (0.0006)	-0.0378*** (0.0008)	-0.0316*** (0.0007)	-0.0319*** (0.0008)	-0.0190*** (0.0006)	-0.0211*** (0.0006)	-0.0237*** (0.0006)	-0.0292*** (0.0010)	-0.0316*** (0.0010)	-0.0335*** (0.0010)	-0.0364*** (0.0010)
prod_numclicks	0.0305*** (0.0012)	0.0405*** (0.0011)	0.0374*** (0.0011)	0.0072*** (0.0015)	0.0066*** (0.0015)	0.0418*** (0.0011)	0.0339*** (0.0011)	0.0417*** (0.0011)	-0.0248*** (0.0039)	0.0004 (0.0029)	-0.0179*** (0.0032)	-0.0170*** (0.0032)
prod_numretailer	0.0012*** (0.0000)	0.0012*** (0.0000)	0.0003*** (0.0001)	0.0014*** (0.0000)	0.0007*** (0.0001)	0.0011*** (0.0000)	0.0010*** (0.0000)	0.0008*** (0.0000)	0.0009*** (0.0001)	0.0007*** (0.0001)	0.0002** (0.0001)	0.0006*** (0.0001)
prod_avgprice	0.0056*** (0.0002)	0.0049*** (0.0002)	0.0074*** (0.0002)	0.0033*** (0.0001)	0.0043*** (0.0002)	0.0044*** (0.0001)	0.0052*** (0.0002)	0.0051*** (0.0001)	0.0004 (0.0004)	0.0010*** (0.0004)	-0.0017*** (0.0004)	0.0005 (0.0004)
cat4_hardware	-0.2926*** (0.0044)	-0.3210*** (0.0044)	-0.2171*** (0.0054)	-0.2328*** (0.0045)	-0.1143*** (0.0054)	-0.3065*** (0.0043)	-0.2956*** (0.0043)	-0.3014*** (0.0043)	-0.5152*** (0.0077)	-0.5111*** (0.0076)	-0.3826*** (0.0076)	-0.4066*** (0.0077)
cat9_videofotov	-0.2465*** (0.0057)	-0.2367*** (0.0058)	-0.1255*** (0.0068)	-0.1325*** (0.0059)	0.0164** (0.0068)	-0.2295*** (0.0057)	-0.2492*** (0.0057)	-0.2586*** (0.0057)	-0.1355*** (0.0093)	-0.1439*** (0.0093)	-0.0340*** (0.0093)	-0.0326*** (0.0094)
Constant	3.3794*** (0.0079)	3.4092*** (0.0081)	3.3134*** (0.0093)	3.4204*** (0.0081)	3.2873*** (0.0091)	3.4977*** (0.0076)	3.4030*** (0.0078)	3.5118*** (0.0077)	3.6458*** (0.0129)	3.7007*** (0.0126)	3.4984*** (0.0129)	3.5778*** (0.0129)
Observations	14814	14264	10615	14710	11680	14940	14771	14795	4824	4833	5329	5322

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 50: Price leader changes, Poisson regression, full dataset - version 3 (cf. section 4.4.3 on page 80)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast- choke	aggr_c_elast	aggr_c- elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw- c_elast	cv_lctw- l_elast	aggr_lctw- c_elast	aggr_lctw- l_elast
elasticity	-0.0125*** (0.0004)	-0.0015*** (0.0000)	-0.0094*** (0.0003)	-0.0043*** (0.0002)	-0.0050*** (0.0002)	-0.0012*** (0.0003)	-0.0130*** (0.0004)	-0.0054*** (0.0002)	0.0229*** (0.0032)	0.0006 (0.0008)	-0.0114*** (0.0004)	-0.0028*** (0.0006)
prod_recommendation	-0.0270*** (0.0097)	-0.0474*** (0.0098)	0.0321*** (0.0109)	-0.1183*** (0.0094)	-0.0705*** (0.0101)	-0.1181*** (0.0094)	-0.0278*** (0.0096)	-0.0947*** (0.0093)	-0.1106*** (0.0165)	-0.0782*** (0.0163)	-0.0899*** (0.0154)	-0.0667*** (0.0153)
prod_avgmissing	0.1119*** (0.0041)	0.1052*** (0.0042)	0.1131*** (0.0046)	0.0991*** (0.0041)	0.1206*** (0.0044)	0.1001*** (0.0041)	0.1223*** (0.0041)	0.0940*** (0.0040)	0.1192*** (0.0077)	0.1338*** (0.0076)	0.1833*** (0.0070)	0.1601*** (0.0070)
prod_subst_msize	0.0147*** (0.0003)	0.0147*** (0.0003)	0.0165*** (0.0003)	0.0117*** (0.0003)	0.0132*** (0.0003)	0.0156*** (0.0003)	0.0125*** (0.0003)	0.0141*** (0.0003)	0.0183*** (0.0007)	0.0172*** (0.0007)	0.0151*** (0.0006)	0.0157*** (0.0006)
prod_brandrank	-0.0414*** (0.0011)	-0.0352*** (0.0011)	-0.0547*** (0.0014)	-0.0305*** (0.0010)	-0.0218*** (0.0011)	-0.0382*** (0.0010)	-0.0431*** (0.0011)	-0.0380*** (0.0011)	-0.0295*** (0.0016)	-0.0285*** (0.0016)	-0.0226*** (0.0014)	-0.0253*** (0.0014)
prod_numclicks	0.0458*** (0.0011)	0.0529*** (0.0011)	0.0502*** (0.0011)	0.0194*** (0.0016)	0.0187*** (0.0017)	0.0546*** (0.0011)	0.0463*** (0.0011)	0.0544*** (0.0011)	-0.0041 (0.0047)	-0.0004 (0.0035)	-0.0055 (0.0036)	0.0014 (0.0034)
prod_numretailer	0.0018*** (0.0001)	0.0019*** (0.0001)	0.0011*** (0.0001)	0.0016*** (0.0001)	0.0010*** (0.0001)	0.0016*** (0.0001)	0.0015*** (0.0001)	0.0012*** (0.0001)	0.0019*** (0.0001)	0.0018*** (0.0001)	0.0010*** (0.0001)	0.0012*** (0.0001)
prod_avgprice	0.0111*** (0.0002)	0.0098*** (0.0002)	0.0136*** (0.0002)	0.0063*** (0.0001)	0.0108*** (0.0002)	0.0106*** (0.0002)	0.0102*** (0.0002)	0.0107*** (0.0002)	0.0101*** (0.0003)	0.0097*** (0.0004)	0.0068*** (0.0004)	0.0076*** (0.0004)
Constant	2.9923*** (0.0091)	2.9329*** (0.0095)	2.9855*** (0.0107)	3.0894*** (0.0091)	3.0648*** (0.0099)	3.1147*** (0.0088)	3.0277*** (0.0091)	3.1125*** (0.0088)	3.0032*** (0.0162)	2.9845*** (0.0158)	2.9426*** (0.0151)	3.0020*** (0.0151)
Observations	10382	9984	7453	10584	8511	10500	10362	10413	3312	3326	3791	3787

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 51: Price leader changes, Poisson regression, cleansed dataset - version 1 (cf. section 4.4.3 on page 80)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast- choke	aggr_c_elast	aggr_c- elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw- c_elast	cv_lctw- l_elast	aggr_lctw- c_elast	aggr_lctw- l_elast
elasticity	-0.0158*** (0.0013)	-0.0000*** (0.0000)	-0.0094*** (0.0014)	0.0021 (0.0014)	0.0017 (0.0015)	-0.0041** (0.0019)	0.0083*** (0.0014)	0.0090*** (0.0020)	0.0275*** (0.0059)	0.1368*** (0.0075)	0.0234*** (0.0038)	0.1041*** (0.0137)
prod_recommendation	1.3760*** (0.0691)	1.1696*** (0.0694)	1.4218*** (0.0715)	-1.7463*** (0.0823)	-1.5337*** (0.0882)	1.1183*** (0.0697)	1.3255*** (0.0710)	0.9551*** (0.0694)	-0.4828*** (0.1436)	-0.3664*** (0.1397)	-2.8656*** (0.3541)	-2.2550*** (0.3793)
prod_avgmissing	-0.2140*** (0.0265)	-0.1676*** (0.0266)	-0.6665*** (0.0322)	0.1066*** (0.0371)	0.1282*** (0.0380)	-0.5174*** (0.0301)	-0.0578** (0.0262)	-0.6431*** (0.0324)	0.1157* (0.0679)	-0.0327 (0.0713)	0.2635 (0.2044)	0.5389** (0.2109)
prod_subst_msize	0.0238*** (0.0010)	0.0263*** (0.0010)	0.0308*** (0.0010)	0.0195*** (0.0014)	0.0204*** (0.0014)	0.0274*** (0.0010)	0.0315*** (0.0010)	0.0313*** (0.0010)	-0.0257*** (0.0020)	-0.0159*** (0.0021)	-0.0031 (0.0043)	0.0020 (0.0043)
prod_brandrank	-0.0185*** (0.0025)	-0.0245*** (0.0024)	-0.0133*** (0.0028)	-0.0807*** (0.0049)	-0.0741*** (0.0053)	-0.0069*** (0.0026)	-0.0275*** (0.0025)	-0.0184*** (0.0028)	0.0441*** (0.0038)	0.0488*** (0.0038)	-0.2006*** (0.0207)	-0.2200*** (0.0204)
prod_numclicks	0.0651*** (0.0019)	0.0673*** (0.0019)	0.0631*** (0.0019)	-0.0060 (0.0037)	-0.0028 (0.0036)	0.0622*** (0.0019)	0.0624*** (0.0019)	0.0653*** (0.0019)	0.0671*** (0.0054)	0.0719*** (0.0050)	0.0542*** (0.0060)	0.0643*** (0.0065)
prod_numretailer	-0.0316*** (0.0009)	-0.0284*** (0.0008)	-0.0271*** (0.0009)	0.0066*** (0.0009)	0.0102*** (0.0009)	-0.0249*** (0.0008)	-0.0229*** (0.0008)	-0.0258*** (0.0008)	-0.0236*** (0.0016)	-0.0031* (0.0017)	-0.0007 (0.0033)	0.0028 (0.0035)
prod_avgprice	0.0469*** (0.0016)	0.0471*** (0.0016)	0.0584*** (0.0016)	0.0012 (0.0025)	0.0030 (0.0026)	0.0510*** (0.0016)	0.0431*** (0.0016)	0.0518*** (0.0016)	-0.0659*** (0.0076)	-0.0811*** (0.0074)	0.0457*** (0.0077)	0.0412*** (0.0077)
cat4_hardware	-1.1478*** (0.0270)	-1.1924*** (0.0275)	-0.9875*** (0.0304)	-0.9846*** (0.0426)	-0.6438*** (0.0483)	-0.8258*** (0.0297)	-1.1063*** (0.0276)	-0.9698*** (0.0309)	-1.3657*** (0.0547)	-1.8128*** (0.0561)	-1.4378*** (0.1384)	-1.7416*** (0.1520)
cat9_videofotov	-1.7084*** (0.0321)	-1.7550*** (0.0321)	-1.6445*** (0.0347)	-1.3396*** (0.0507)	-1.0408*** (0.0557)	-1.5058*** (0.0347)	-1.7039*** (0.0327)	-1.5947*** (0.0354)	-0.4453*** (0.0477)	-0.7852*** (0.0527)	-1.4046*** (0.1484)	-1.6510*** (0.1585)
Constant	4.2001*** (0.0712)	4.4126*** (0.0704)	3.8147*** (0.0756)	5.3231*** (0.0829)	4.6248*** (0.0918)	4.0329*** (0.0731)	4.1495*** (0.0720)	4.3235*** (0.0739)	6.0751*** (0.1470)	6.0043*** (0.1419)	6.5432*** (0.2878)	6.3659*** (0.3005)
Observations	193	193	187	139	137	191	190	191	59	62	40	40

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 52: Price leader changes, Poisson regression, cleansed dataset - version 2 (cf. section 4.4.3 on page 80)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast- choke	aggr_c_elast	aggr_c- elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw- c_elast	cv_lctw- l_elast	aggr_lctw- c_elast	aggr_lctw- l_elast
elasticity	-0.0237*** (0.0015)	-0.0030*** (0.0002)	-0.0148*** (0.0017)	-0.0020 (0.0016)	-0.0018 (0.0016)	-0.0006 (0.0026)	0.0091*** (0.0014)	0.0200*** (0.0022)	0.0644*** (0.0075)	0.1174*** (0.0087)	0.0153* (0.0083)	0.0863*** (0.0184)
prod_recommendation	1.3905*** (0.0700)	1.0664*** (0.0708)	1.4122*** (0.0723)	-1.6812*** (0.0835)	-1.4727*** (0.0888)	1.1474*** (0.0711)	1.3157*** (0.0710)	0.9880*** (0.0704)	0.7394*** (0.1553)	-0.3072** (0.1397)	-2.1102*** (0.4191)	-2.1283*** (0.3864)
prod_avgmissing	-0.2283*** (0.0269)	-0.1657*** (0.0271)	-0.6951*** (0.0329)	0.0798** (0.0374)	0.1093*** (0.0381)	-0.6637*** (0.0327)	-0.0677** (0.0265)	-0.7265*** (0.0339)	-0.0121 (0.0734)	-0.0209 (0.0709)	0.2448 (0.2085)	0.4792** (0.2146)
prod_subst_msize	0.0221*** (0.0010)	0.0233*** (0.0010)	0.0292*** (0.0011)	0.0207*** (0.0014)	0.0211*** (0.0014)	0.0281*** (0.0010)	0.0312*** (0.0010)	0.0328*** (0.0010)	-0.0253*** (0.0020)	-0.0170*** (0.0021)	-0.0017 (0.0043)	0.0017 (0.0043)
prod_brandrank	-0.0160*** (0.0026)	-0.0161*** (0.0025)	-0.0119*** (0.0028)	-0.0769*** (0.0049)	-0.0715*** (0.0052)	-0.0030 (0.0026)	-0.0276*** (0.0025)	-0.0174*** (0.0028)	0.0581*** (0.0038)	0.0492*** (0.0037)	-0.1897*** (0.0207)	-0.2125*** (0.0210)
prod_numclicks	0.0633*** (0.0019)	0.0659*** (0.0020)	0.0621*** (0.0019)	-0.0074** (0.0038)	-0.0041 (0.0036)	0.0618*** (0.0020)	0.0630*** (0.0019)	0.0652*** (0.0020)	0.0229** (0.0114)	0.0684*** (0.0050)	0.0520*** (0.0061)	0.0619*** (0.0066)
prod_numretailer	-0.0324*** (0.0009)	-0.0318*** (0.0008)	-0.0283*** (0.0009)	0.0055*** (0.0009)	0.0094*** (0.0009)	-0.0251*** (0.0009)	-0.0233*** (0.0008)	-0.0244*** (0.0008)	-0.0448*** (0.0022)	-0.0041** (0.0017)	-0.0050 (0.0036)	0.0007 (0.0038)
prod_avgprice	0.0465*** (0.0016)	0.0425*** (0.0017)	0.0577*** (0.0016)	0.0020 (0.0025)	0.0032 (0.0026)	0.0555*** (0.0016)	0.0428*** (0.0016)	0.0554*** (0.0016)	-0.0719*** (0.0080)	-0.0814*** (0.0074)	0.0513*** (0.0080)	0.0438*** (0.0080)
cat4_hardware	-1.1054*** (0.0271)	-1.0129*** (0.0281)	-0.9567*** (0.0306)	-1.0543*** (0.0431)	-0.7064*** (0.0487)	-0.8103*** (0.0311)	-1.1115*** (0.0277)	-0.9890*** (0.0317)	-1.2069*** (0.0581)	-1.7627*** (0.0568)	-1.5919*** (0.1430)	-1.7443*** (0.1504)
cat9_videofotov	-1.6706*** (0.0325)	-1.6638*** (0.0334)	-1.6129*** (0.0351)	-1.4028*** (0.0511)	-1.0903*** (0.0557)	-1.5345*** (0.0359)	-1.6999*** (0.0327)	-1.6543*** (0.0359)	-0.2125*** (0.0522)	-0.7337*** (0.0534)	-1.5488*** (0.1536)	-1.6586*** (0.1576)
Constant	4.1388*** (0.0738)	4.3580*** (0.0724)	3.8191*** (0.0777)	5.2947*** (0.0835)	4.6052*** (0.0915)	4.0101*** (0.0776)	4.1769*** (0.0729)	4.3025*** (0.0763)	5.8195*** (0.1618)	5.9082*** (0.1430)	6.1433*** (0.3103)	6.2688*** (0.3086)
Observations	189	181	183	137	136	181	189	189	56	59	38	39

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 53: Price leader changes, Poisson regression, cleansed dataset - version 3 (cf. section 4.4.3 on page 80)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	c_elast	c_elast_inv	c_elast- choke	aggr_c_elast	aggr_c- elast_choke	l_elast	cv_c_elast	cv_l_elast	cv_lctw- c_elast	cv_lctw- l_elast	aggr_lctw- c_elast	aggr_lctw- l_elast
elasticity	-0.0234*** (0.0026)	-0.0119*** (0.0003)	-0.0187*** (0.0032)	-0.0240*** (0.0026)	-0.0135*** (0.0026)	0.0755*** (0.0053)	0.0516*** (0.0022)	0.0757*** (0.0039)	0.4588*** (0.0475)	0.8175*** (0.0569)	0.0694*** (0.0159)	0.3478*** (0.0478)
prod_recommendation	2.3717*** (0.1336)	2.0545*** (0.1236)	2.1523*** (0.1399)	-2.0573*** (0.1146)	-1.9589*** (0.1134)	1.6886*** (0.1410)	2.1723*** (0.1354)	1.4684*** (0.1492)	1.9706*** (0.7442)	-9.0710*** (1.0839)	-3.0833*** (1.0882)	-3.7039*** (1.3162)
prod_avgmissing	-0.9293*** (0.0480)	-0.6062*** (0.0452)	-1.3571*** (0.0531)	-0.1375** (0.0555)	-0.1379** (0.0568)	-1.4274*** (0.0562)	-1.1090*** (0.0520)	-1.5256*** (0.0551)	3.6556*** (0.2202)	6.0039*** (0.3111)	0.1214 (0.2806)	1.0409*** (0.3358)
prod_subst_msize	-0.0101*** (0.0018)	-0.0313*** (0.0021)	-0.0042* (0.0022)	0.0050** (0.0025)	0.0035 (0.0025)	-0.0035* (0.0020)	-0.0110*** (0.0019)	0.0094*** (0.0021)	-0.1680*** (0.0070)	-0.0497*** (0.0072)	-0.0251*** (0.0074)	-0.0046 (0.0077)
prod_brandrank	-0.0367*** (0.0058)	-0.0091 (0.0056)	-0.0730*** (0.0083)	-0.0561*** (0.0078)	-0.0845*** (0.0084)	-0.0470*** (0.0060)	-0.0674*** (0.0054)	-0.0897*** (0.0085)	-0.5308*** (0.0365)	-0.6762*** (0.0373)	0.0118 (0.0387)	-0.1032** (0.0420)
prod_numclicks	0.1057*** (0.0025)	0.1158*** (0.0027)	0.1119*** (0.0026)	0.0429*** (0.0040)	0.0394*** (0.0040)	0.1069*** (0.0029)	0.1019*** (0.0027)	0.1036*** (0.0029)	0.0171 (0.0292)	0.3891*** (0.0209)	0.0711*** (0.0093)	0.1117*** (0.0115)
prod_numretailer	-0.0542*** (0.0017)	-0.0573*** (0.0015)	-0.0559*** (0.0022)	-0.0154*** (0.0017)	-0.0119*** (0.0017)	-0.0453*** (0.0019)	-0.0418*** (0.0017)	-0.0356*** (0.0016)	-0.2055*** (0.0100)	-0.1117*** (0.0072)	-0.0411*** (0.0108)	-0.0029 (0.0112)
prod_avgprice	0.1849*** (0.0027)	0.1421*** (0.0028)	0.2044*** (0.0029)	0.0304*** (0.0050)	0.0378*** (0.0050)	0.2209*** (0.0030)	0.1938*** (0.0028)	0.2236*** (0.0029)	0.1276*** (0.0427)	0.3225*** (0.0397)	-0.5381*** (0.0950)	-0.3889*** (0.0833)
Constant	2.4271*** (0.1377)	2.4028*** (0.1182)	2.4490*** (0.1508)	4.7051*** (0.1121)	4.7656*** (0.1122)	3.2296*** (0.1533)	3.2433*** (0.1380)	2.8647*** (0.1538)	10.2329*** (0.6789)	13.9937*** (0.8767)	7.7124*** (1.0582)	7.0561*** (1.1247)
Observations	86	81	82	72	71	78	83	83	21	25	23	23

LHS-variable: num_pleader_change. Estimated using Poisson regression.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

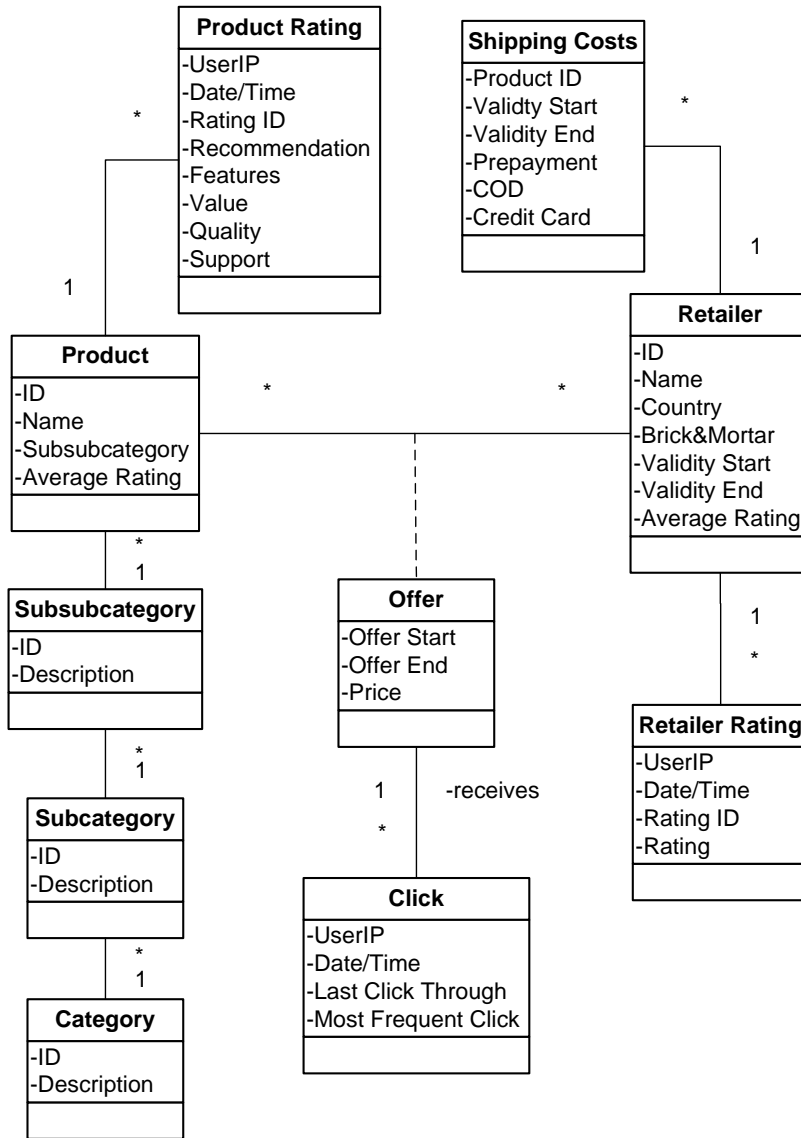


Figure 22: Simplified conceptual schema of the complete Geizhals database

VARIABLE	DESCRIPTION
ssk_numtotclks	Total number of clicks received by all products from the current subsubcategory
ssk_numprods	Number of products listed in the current subsubcategory
ssk_numretailer	Number of retailers offering products in the current subsubcategory
ssk_numoffered_prds	Number of products offered in the current subsubcategory during the week of observation
ssk_numclickedprods	Number of products in the current subsubcategory which have received any clicks during the period of observation
ssk_quality	Average quality of all products in the current subsubcategory
ssk_recommendation	Average share of recommendations for products in the current subsubcategory
ssk_qual_samplesize	Number of records used to compute ssk_quality
ssk_qual_miss	Number of records in ssk_qual_samplesize with missing quality data
ssk_numratings	Number of ratings for products in the current subsubcategory
cat1_audiohifi	1 if the subsubcategory belongs to the category Audio/HIFI, else 0
cat2_films	1 if the subsubcategory belongs to the category Films, else 0
cat3_games	1 if the subsubcategory belongs to the category Games, else 0
cat4_hardware	1 if the subsubcategory belongs to the category Hardware, else 0
cat5_household	1 if the subsubcategory belongs to the category Household-appliance, else 0
cat6_software	1 if the subsubcategory belongs to the category Software, else 0
cat7_sports	1 if the subsubcategory belongs to the category Sports, else 0
cat8_telephoneco	1 if the subsubcategory belongs to the category Telephone, else 0
cat9_videofototv	1 if the subsubcategory belongs to the category Video/Foto/TV
dummysmissing	1 if the information on categories is missing, else 0

Table 54: Subsubcategory specific second stage variables

VARIABLE	DESCRIPTION
elasticity	Elasticity which has been estimated during the first stage regression
prod_quality	Average quality rating of the product ¹
prod_recommendation	Share of users who recommend the current product
prod_avgmissing	1 if the current product's quality data is missing
prod_rel_qual	Ratio of the average quality of the current product to the average quality of the sub-subcategory of the product under exclusion of the current product
prod_rel_recom	Ratio of the share of users who recommend the current product to the average share of recommendations of the sub-subcategory under exclusion of the current product
prod_subst_msize	Size of the market for substitutes measured as the total number of clicks received by products of the current subsubcategory but with exclusion of clicks of the observed product
prod_numclicks	Number of clicks received by the observed product
prod_numretailer	Number of retailers that are offering the observed product
prod_avgprice	Average price of the product during the week of observation.
prod_numratings	Number of product ratings which the product received during the period of observation and during 90 days before that period.
prod_brandrank	The rank of the brand as an integer value
brand1to10	1 if the brandrank is between 1 and 10, else 0
brand11to20	1 if the brandrank is between 11 and 20, else 0
brand21to30	1 if the brandrank is between 21 and 30, else 0
brand31to40	1 if the brandrank is between 31 and 40, else 0
brand41to50	1 if the brandrank is between 41 and 50, else 0
brand51to70	1 if the brandrank is between 51 and 70, else 0
brand71to100	1 if the brandrank is between 71 and 100, else 0
brand101toEnd	1 if the brandrank is above 101, else 0
nobrandatall	1 if no brand information is available for the current product

Table 55: Product specific second stage variables

B.1 ESTIMATED DEMAND CURVES OF THE 40-PRODUCT SAMPLE

This section shows the estimated demand curves of the 40 product sample of section 2.6.4 on page 30.

The layout of the forthcoming pages is given in figure 23 and can

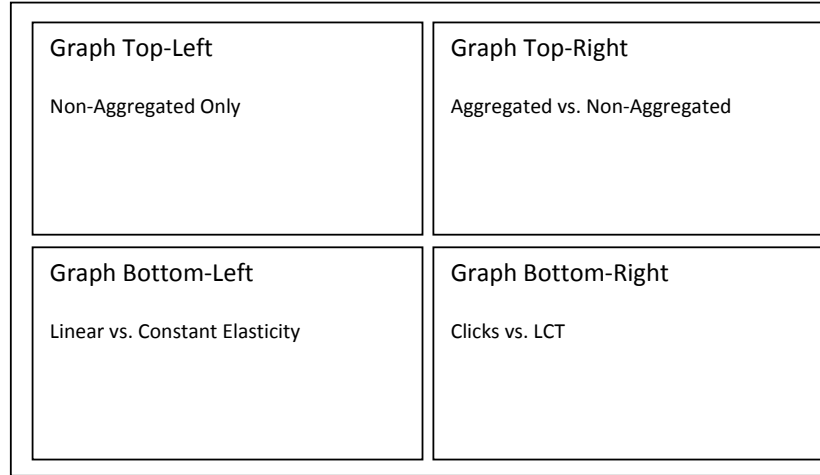


Figure 23: Layout of a page presenting alternative demand curves for a product be described as follows:

GRAPH TOP-LEFT: Main purpose of this graph is to highlight the difference between the normal demand (coming from the regression of clicks on price) and the inverse demand (coming from the regression of price on clicks).

GRAPH TOP-RIGHT: The purpose of this graph is to give an impression on the difference between a linear demand specification and an isoelastic demand specification. Therefore the axes are not denoted in logs.

GRAPH BOTTOM-LEFT: This graph compares an isoelastic demand curve based on normal clicks with an isoelastic demand curve base on aggregated/accumulated clicks.

GRAPH BOTTOM-RIGHT: Purpose of this graph is to compare an isoelastic demand curve based on normal clicks with an isoelastic curve based on [LCT](#).

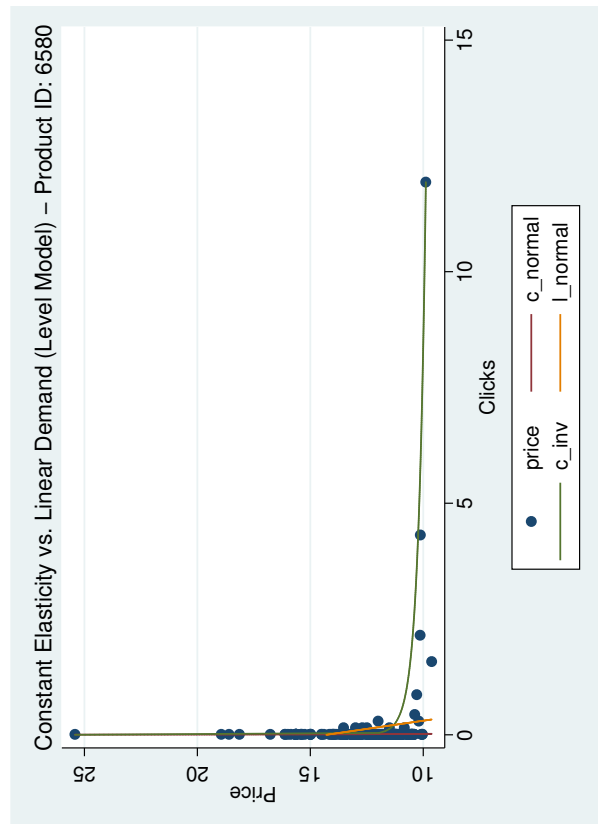
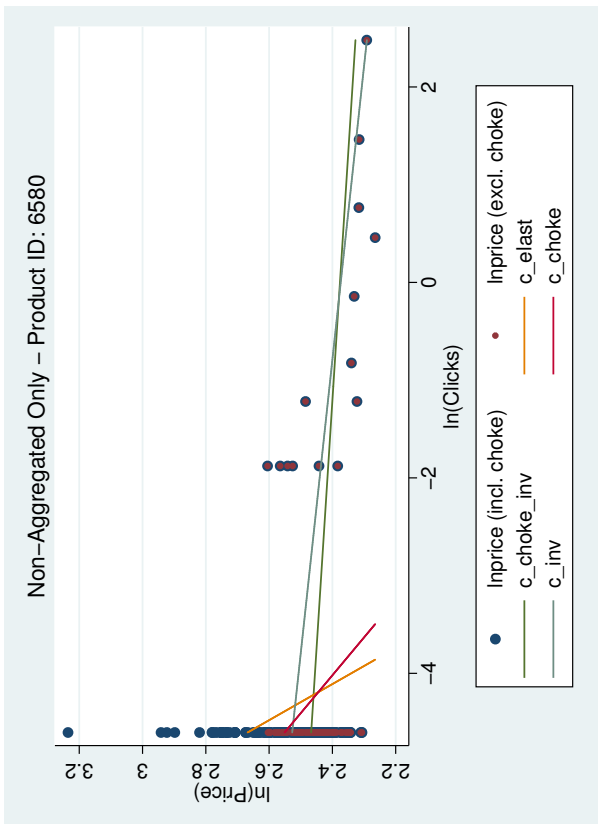
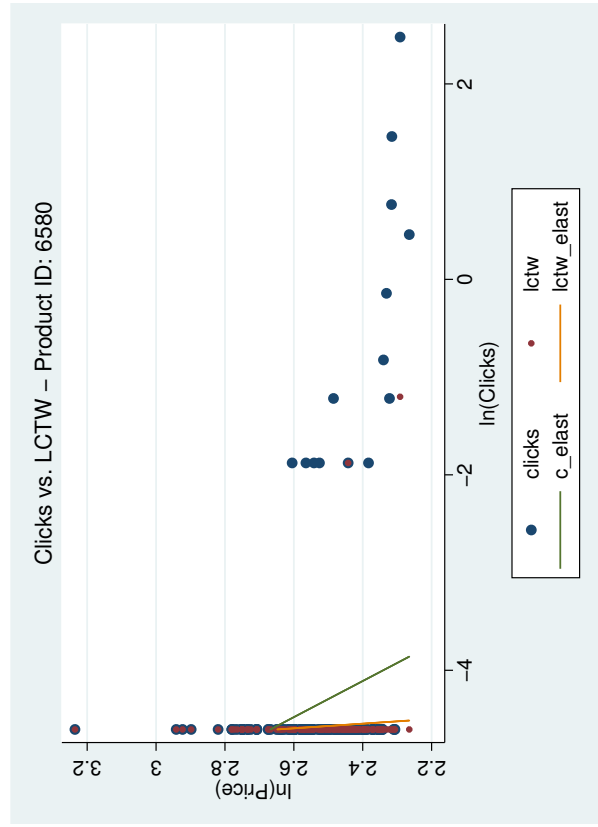
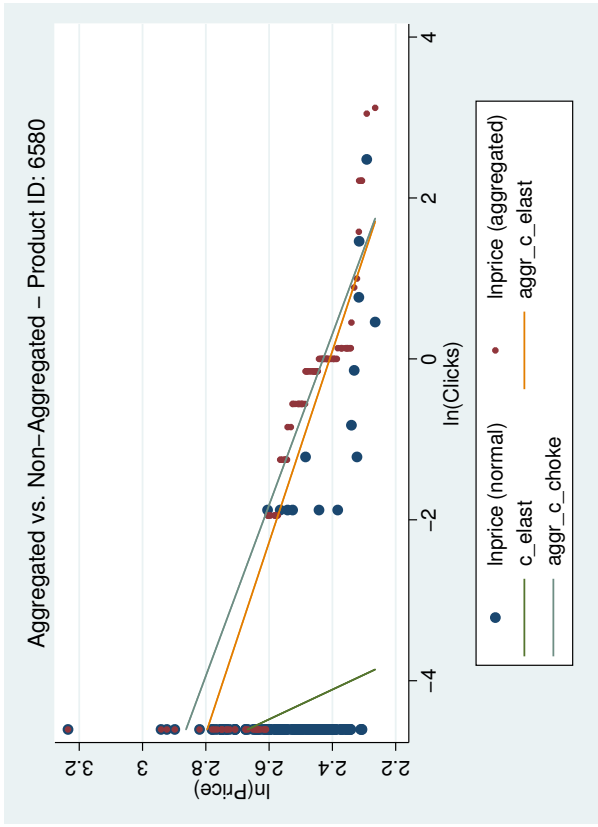


Figure 24: Product: 6580, Canon BCI-3eBK Tintenpatrone schwarz (BJC-3000/6000/S400/450/600/630/6300/MultiPASS C755): Hardware ⇒ Verbrauchsmaterial

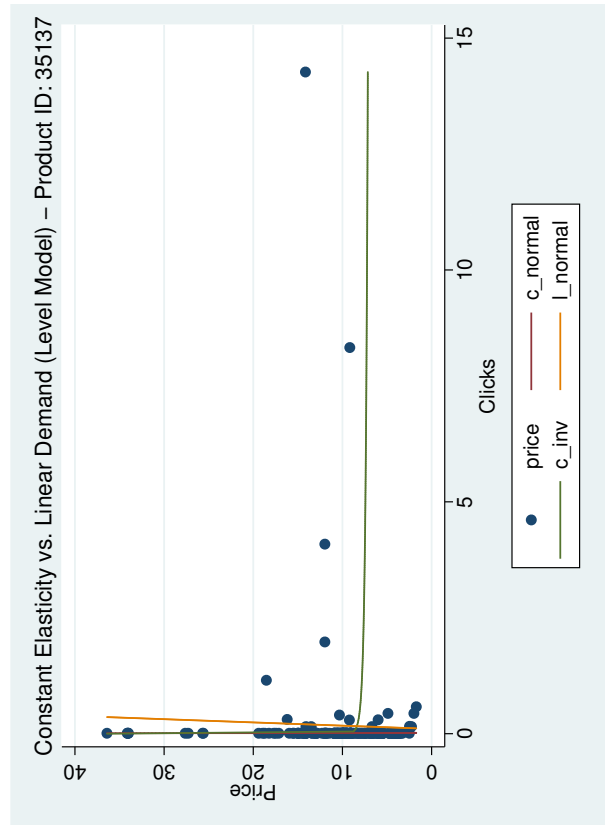
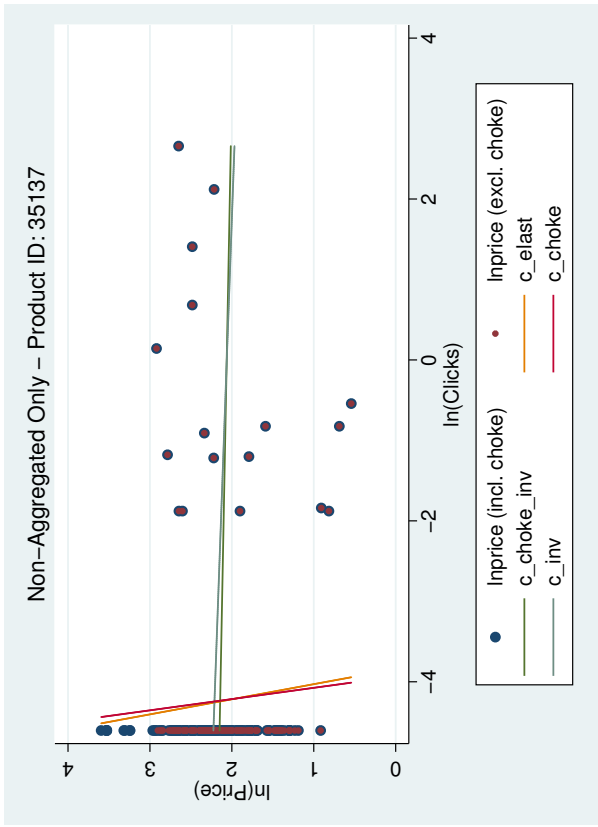
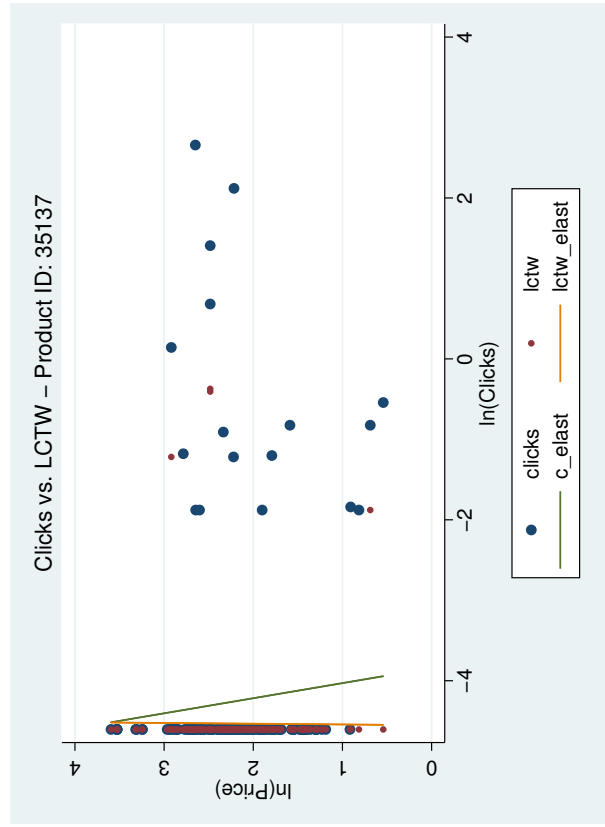
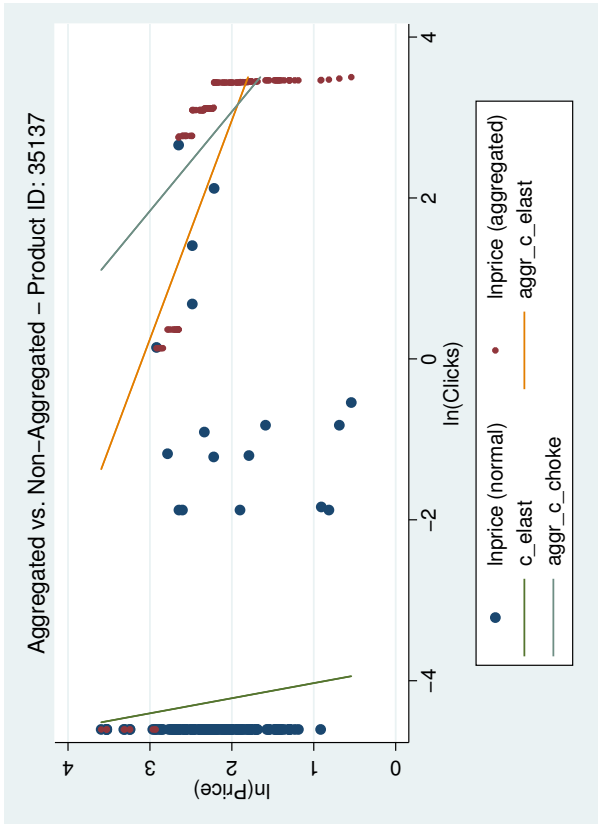


Figure 25: Product: 35137, FireWire IEEE-1394 Kabel 6pin/4pin, 1.8m: Hardware \Rightarrow KabelZubehr

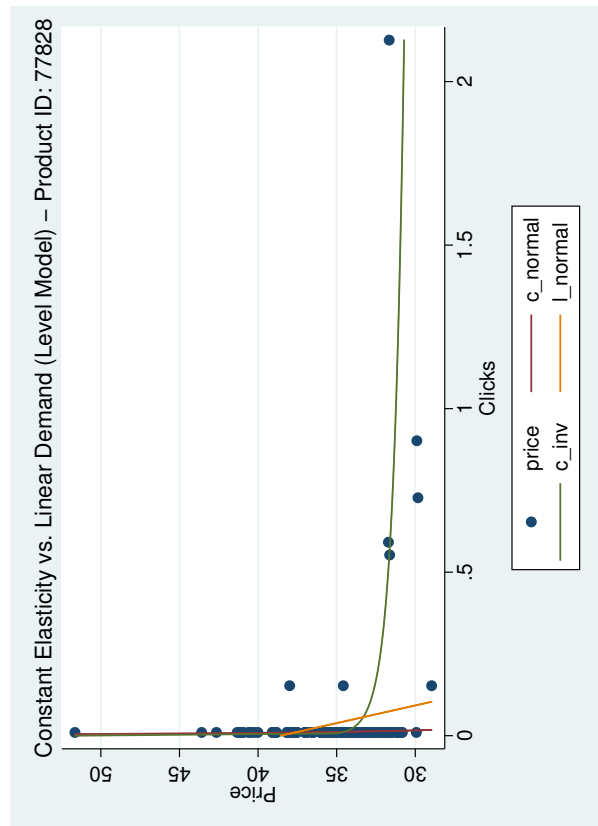
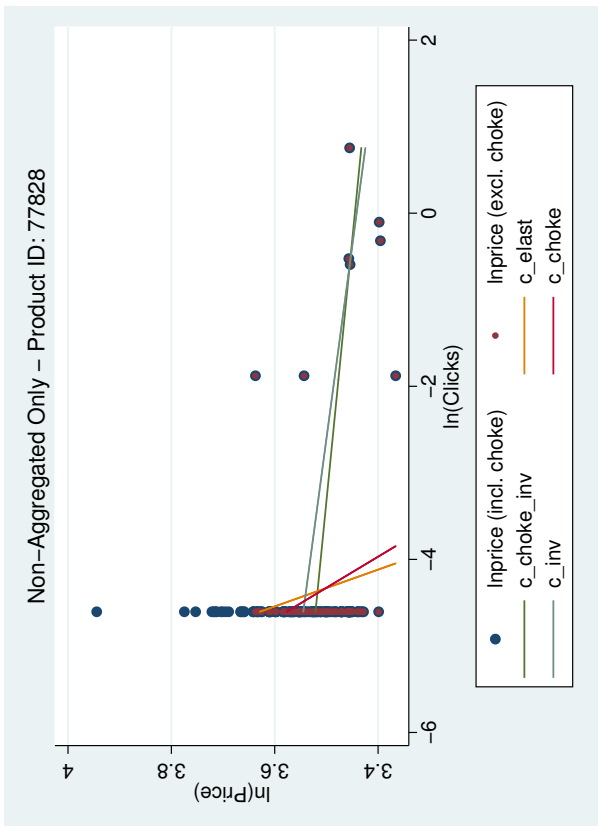
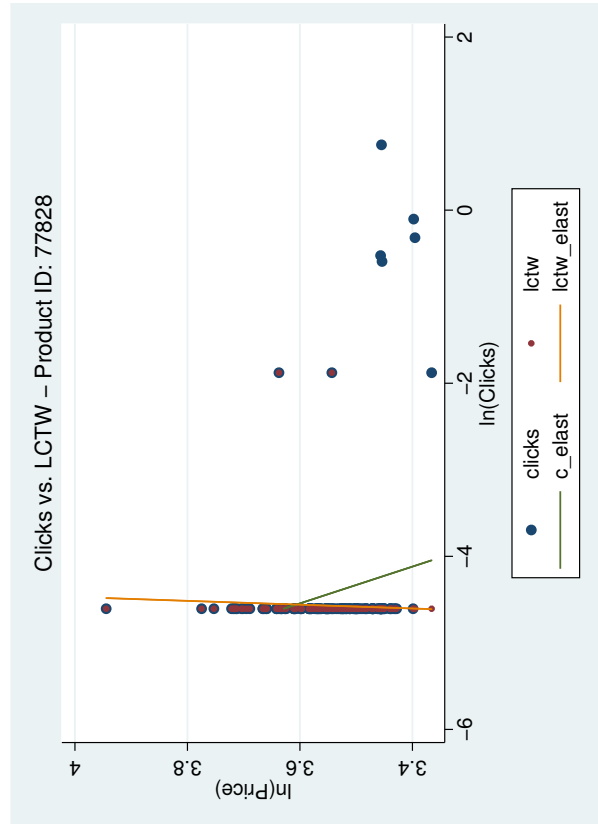
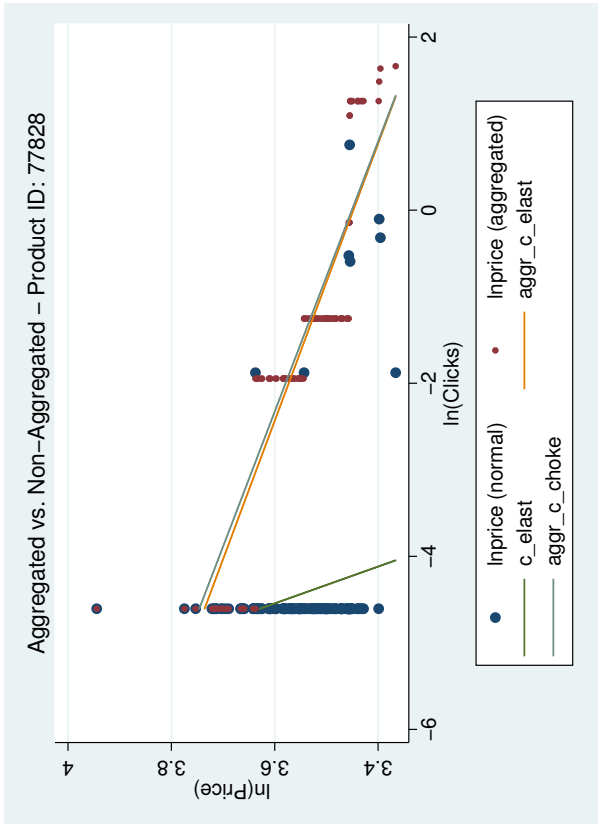


Figure 26: Product: 77828, OkI 1103402 Toner schwarz (B4200): Hardware \Rightarrow Verbrauchsmaterial

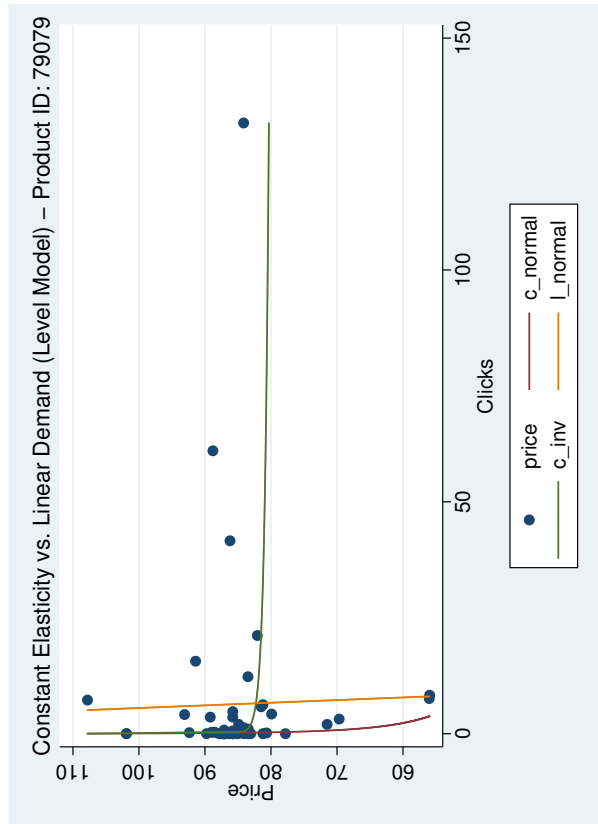
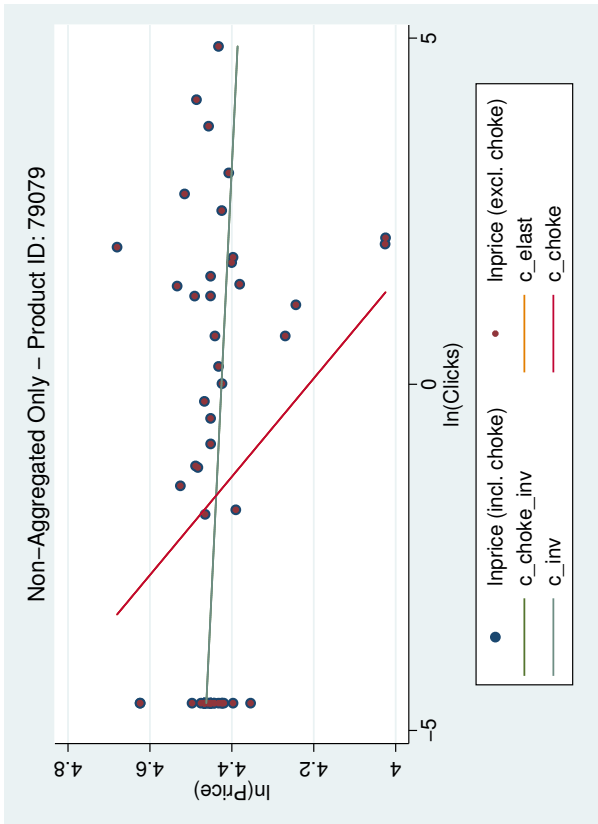
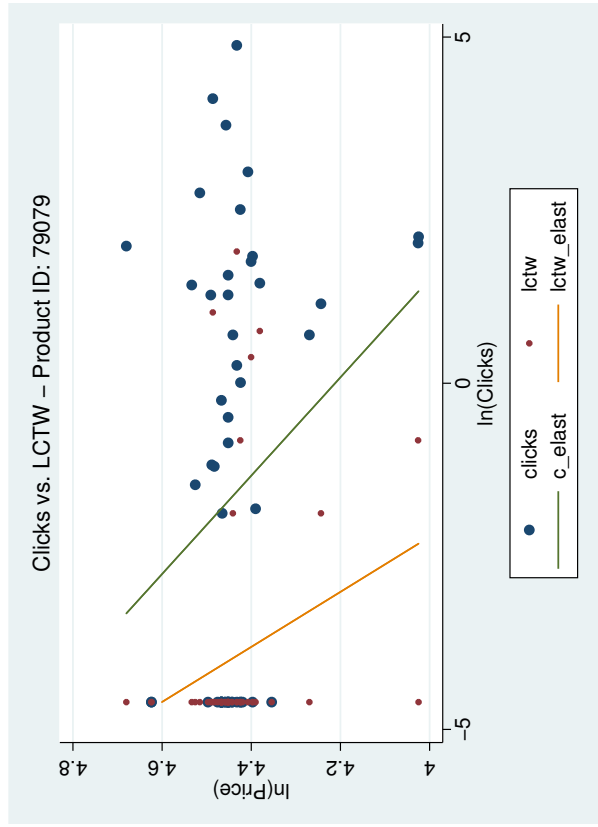
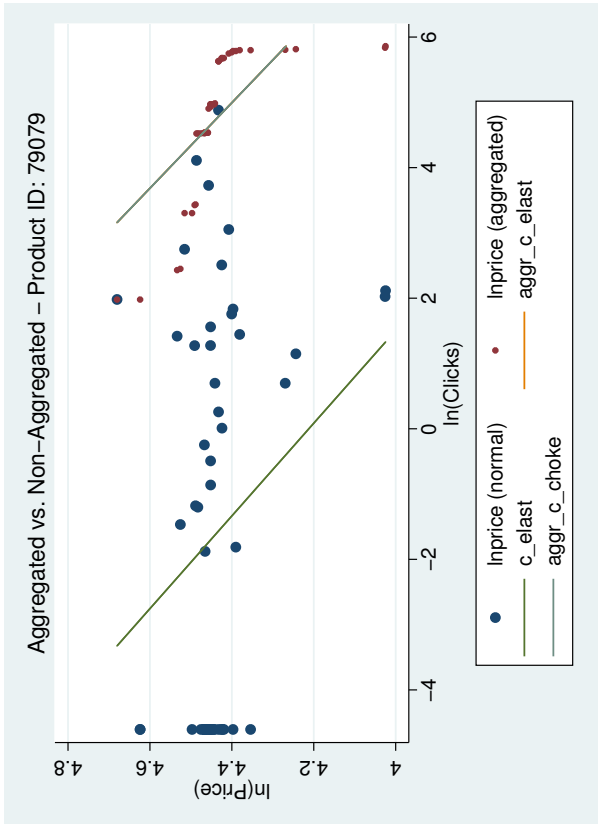


Figure 27: Product: 79079, Twinhan VisionDTV DVB-S PCI Sat CI (1030A/1032A): Hardware \Rightarrow PCVideo

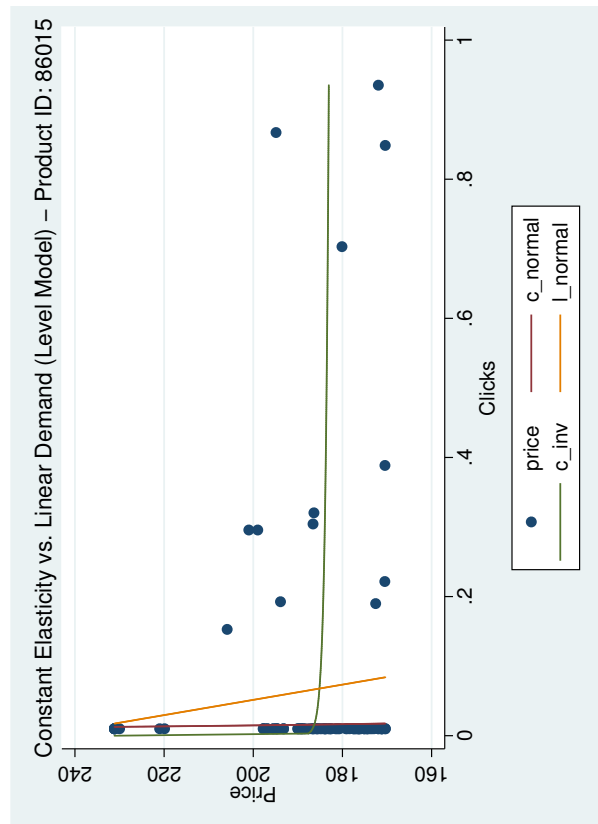
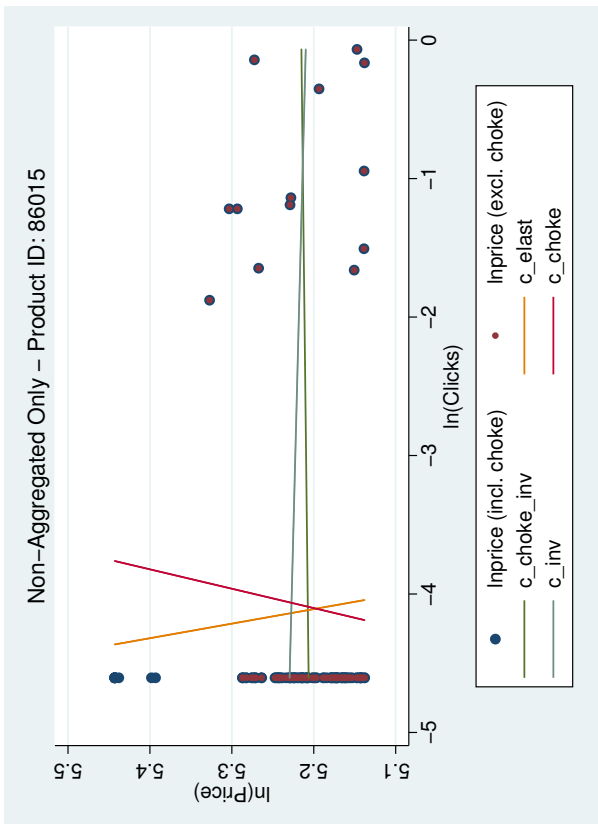
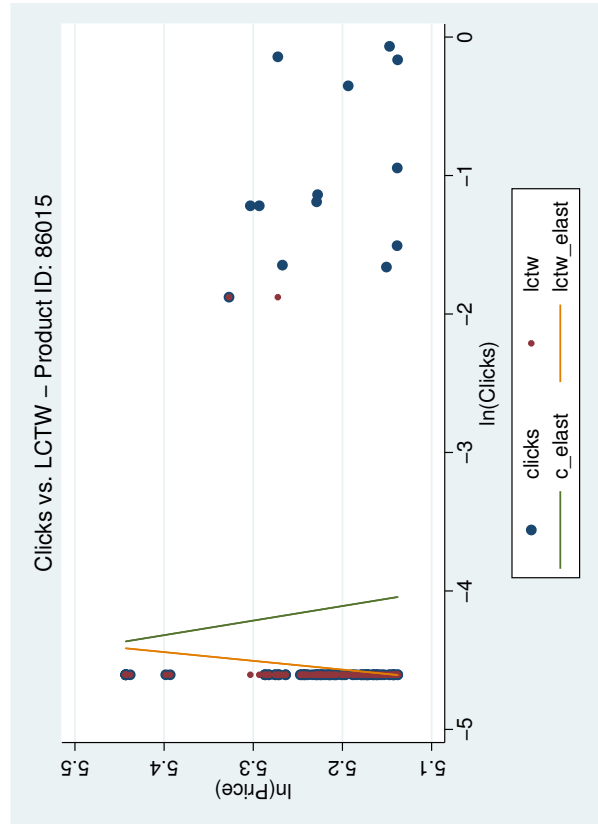
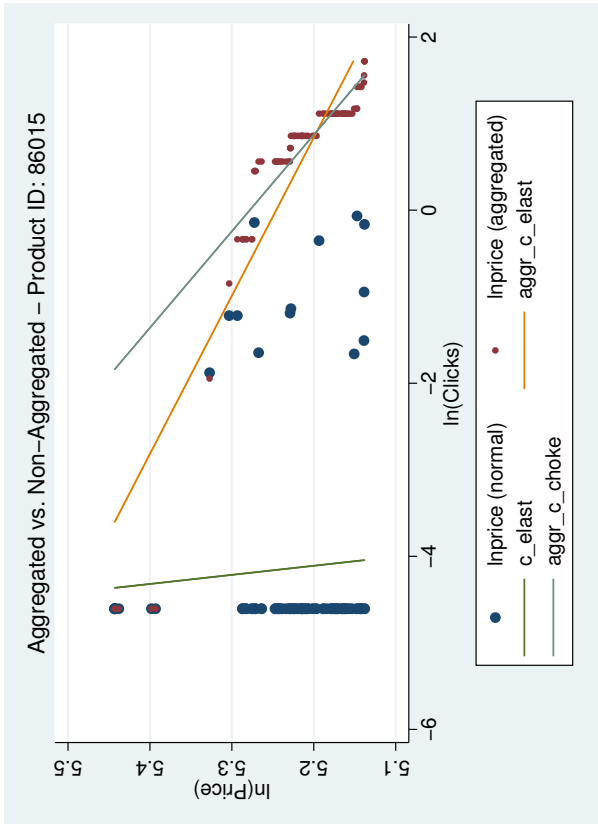


Figure 28: Product: 86015, Sony Vaio PCGA-BP2V Li-Ionen-Akku: Hardware ⇒ Notebookzubehr

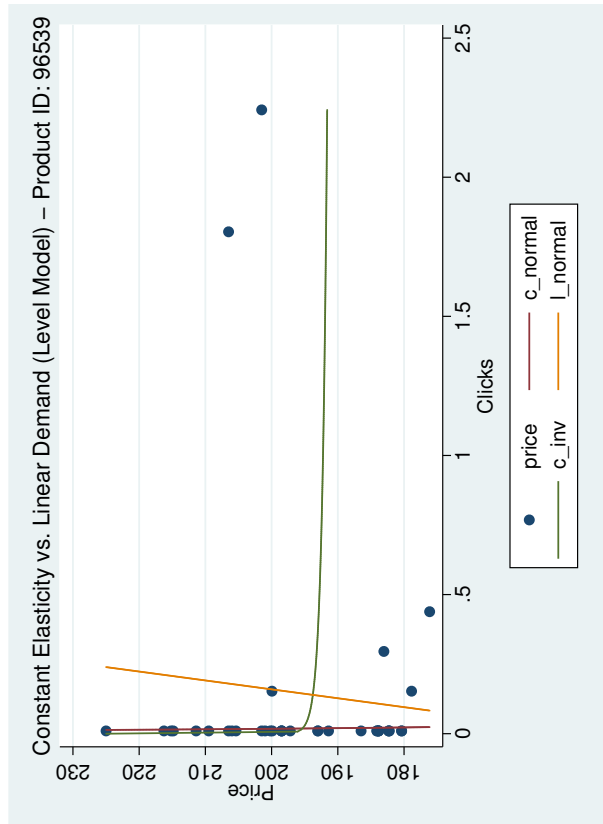
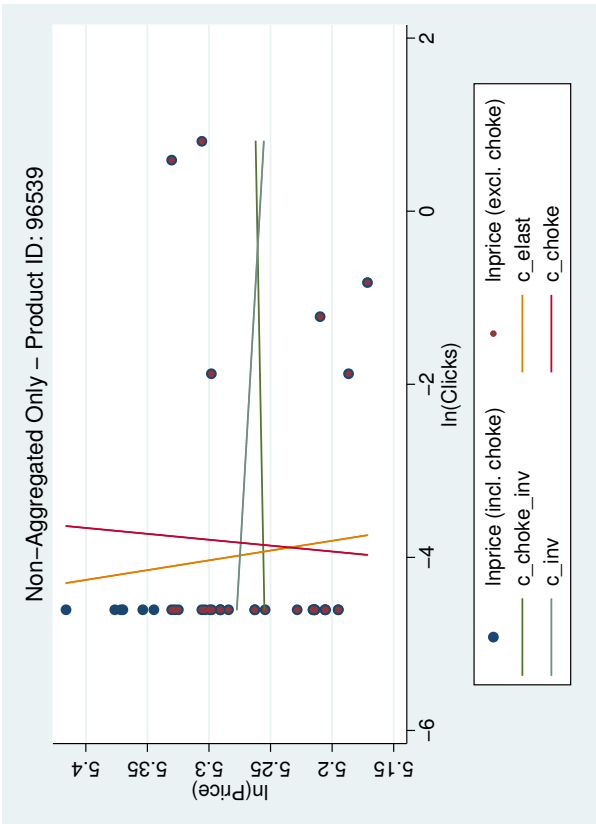
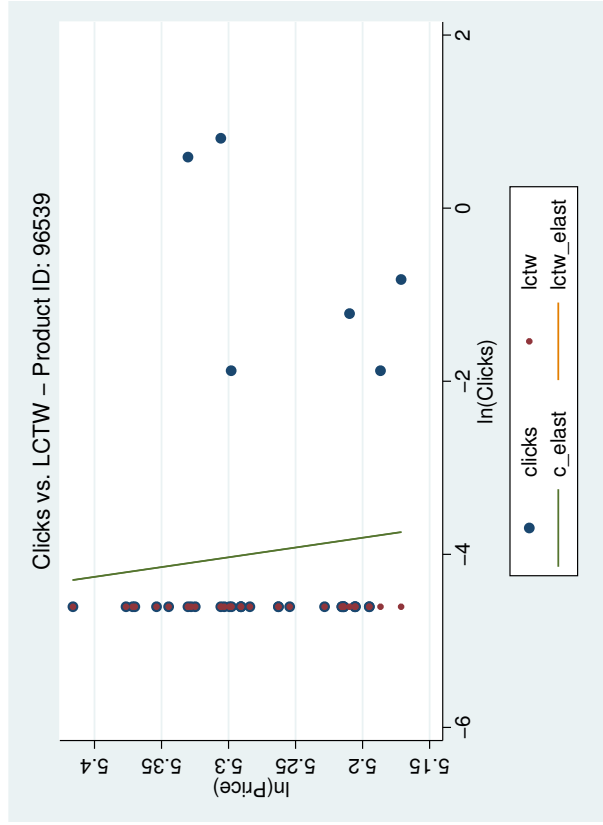
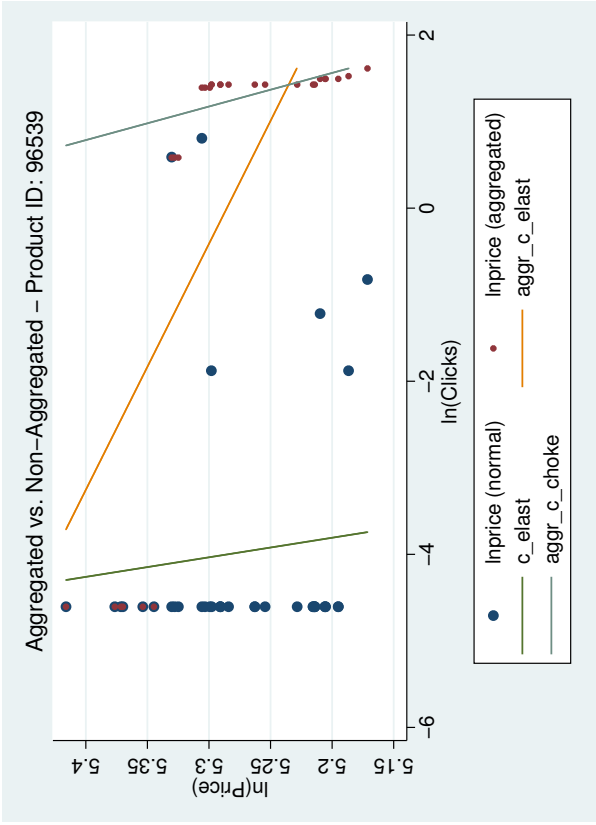


Figure 29: Product: 96539, Sony MDR-DS3000: Hardware ⇒ PCAudio

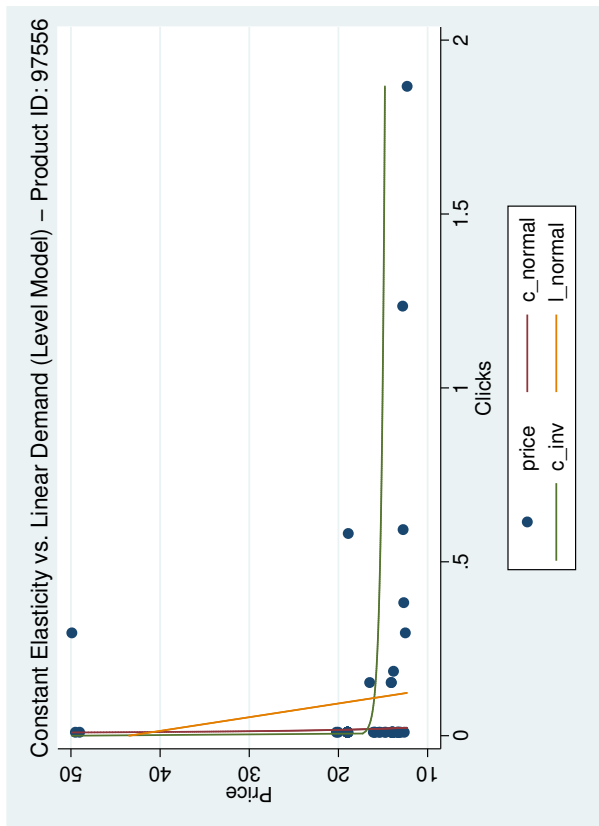
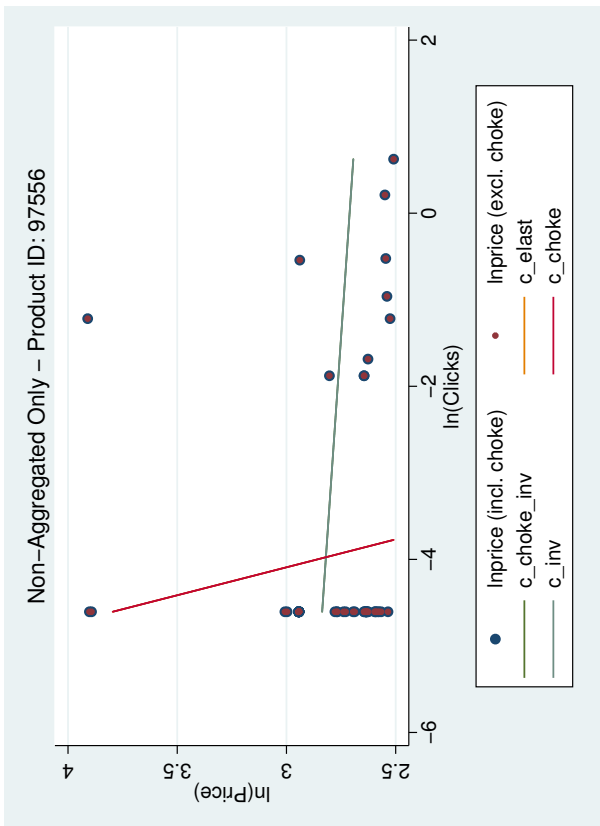
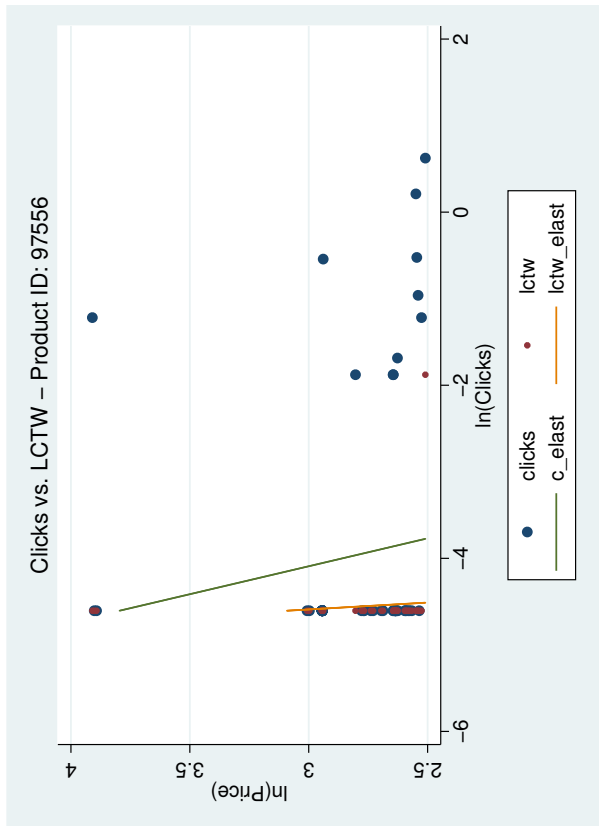
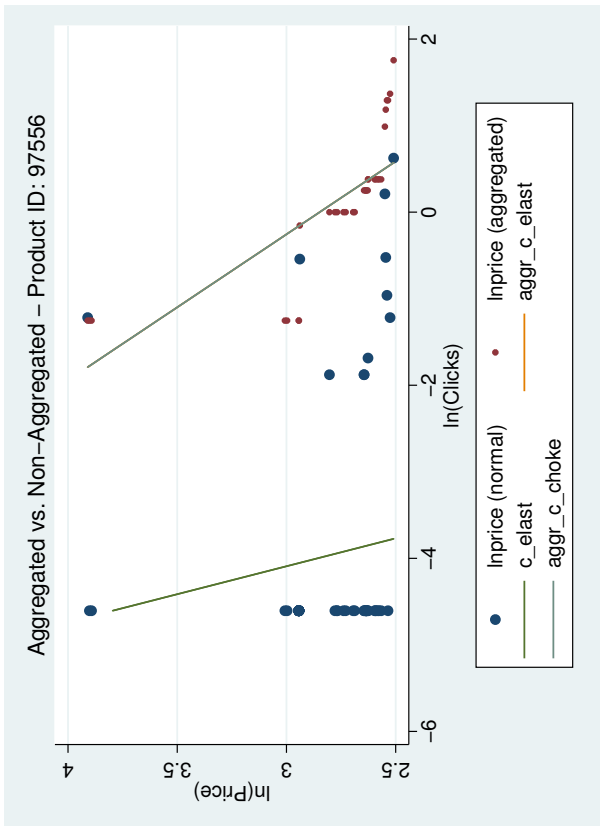


Figure 30: Product: 97556, Digitus DA-70129 Fast IrDa Adapter, USB 1.1: Hardware \Rightarrow Mainboards

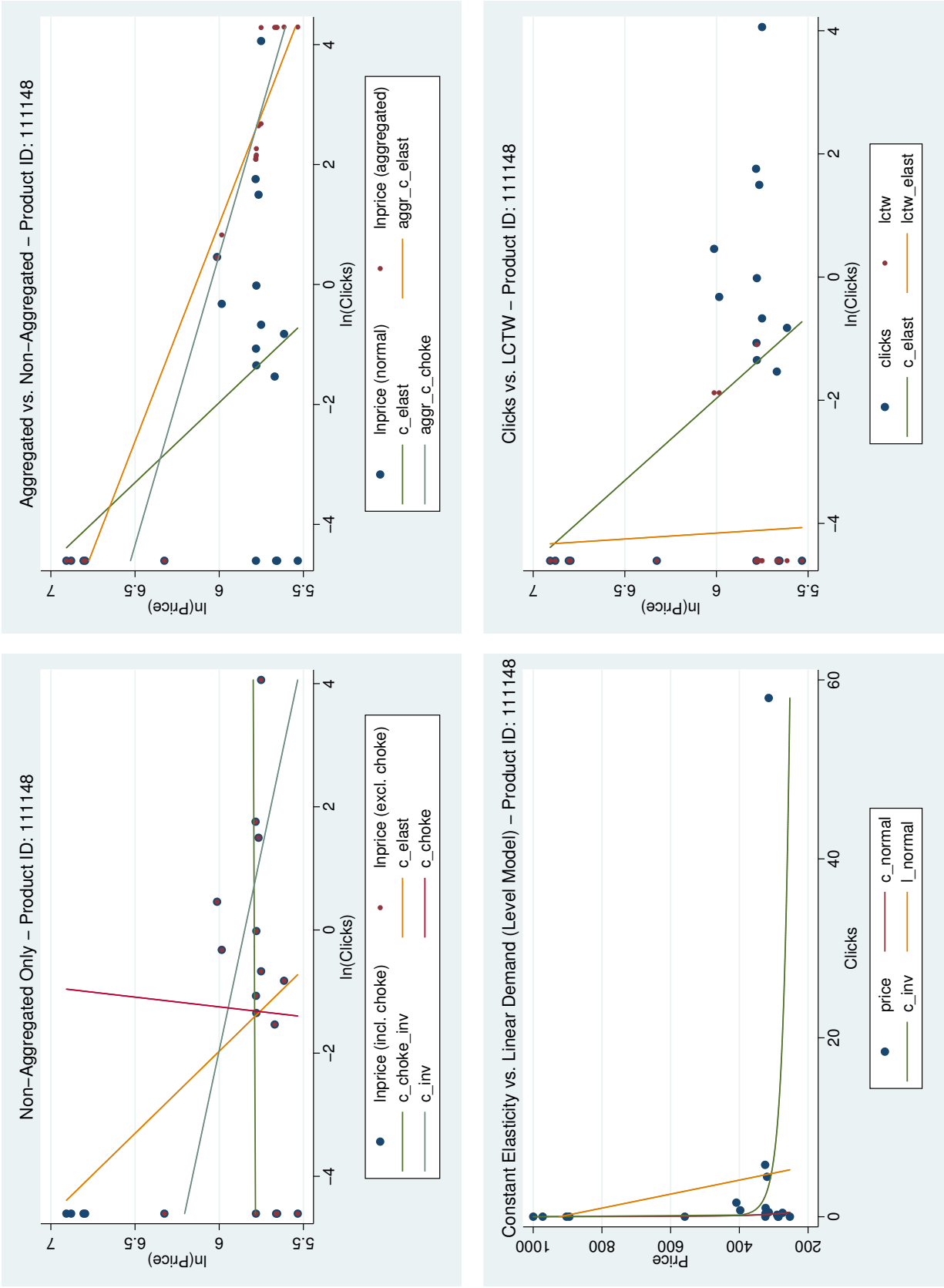


Figure 31: Product: 111148, Komplettsystem AMD Athlon 64 3200+, 2048 MB RAM: Hardware => Systeme

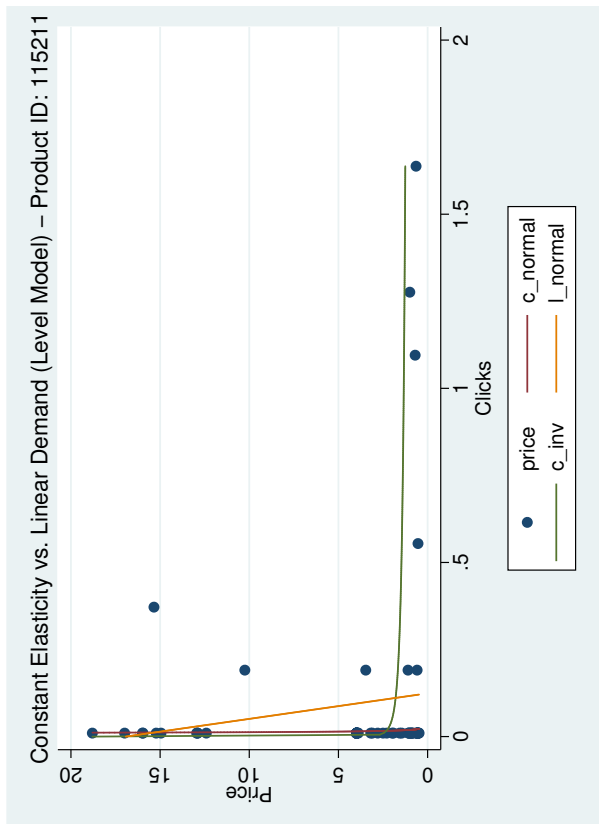
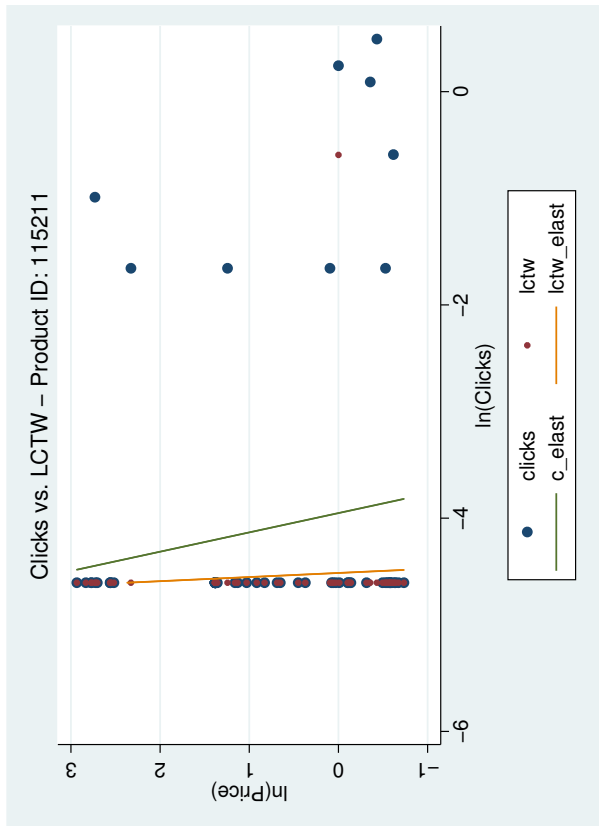
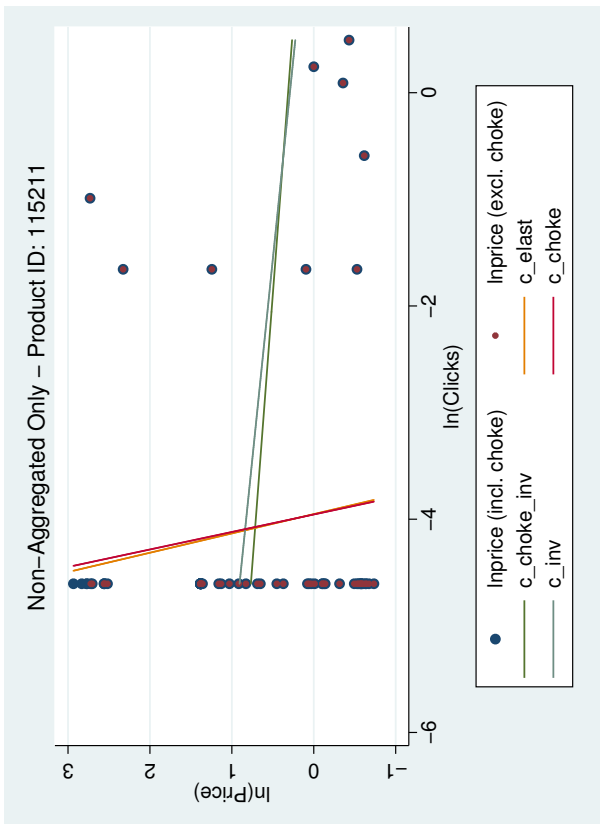
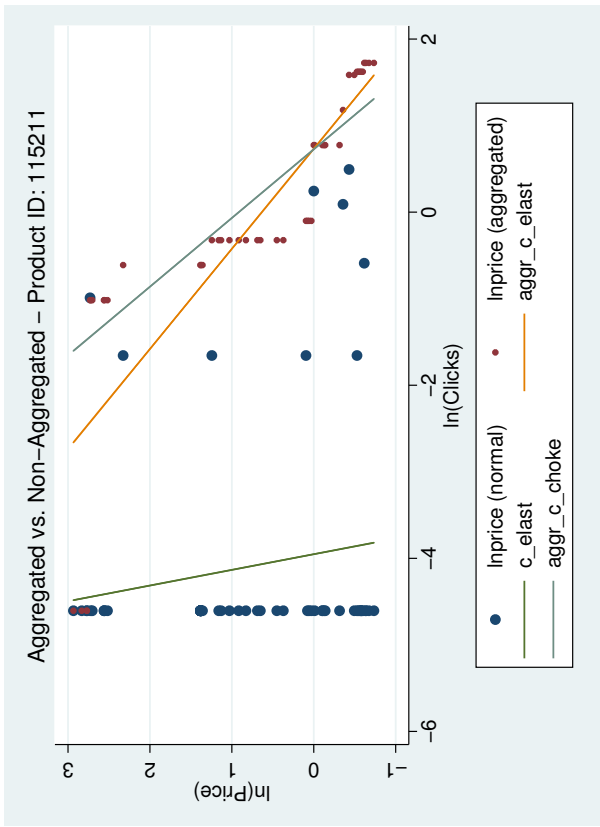


Figure 32: Product: 115211, No-Name/Diverse Mousepad schwarz: Hardware \Rightarrow Eingabegeräte

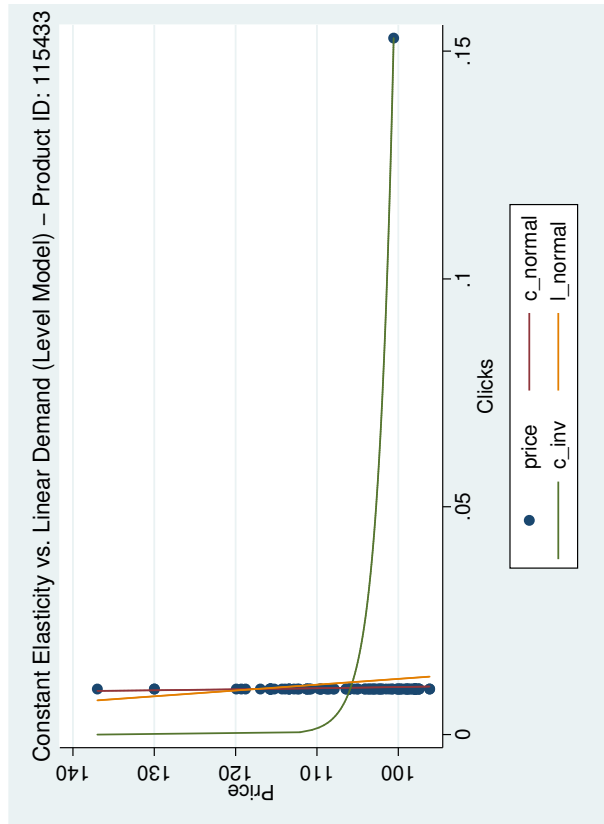
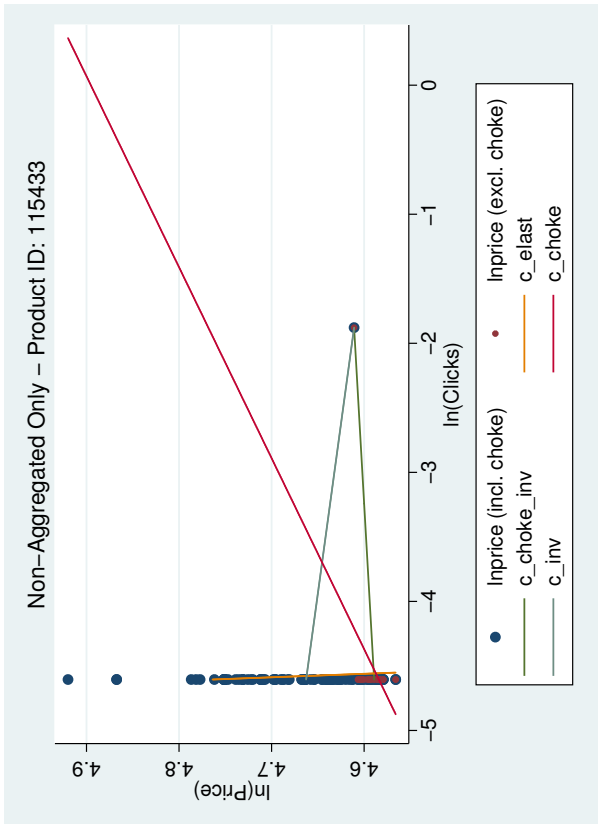
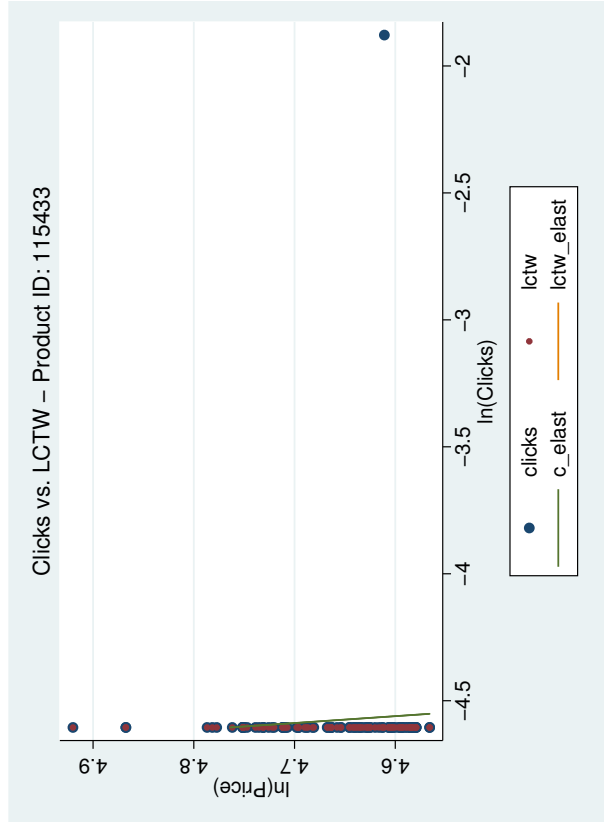
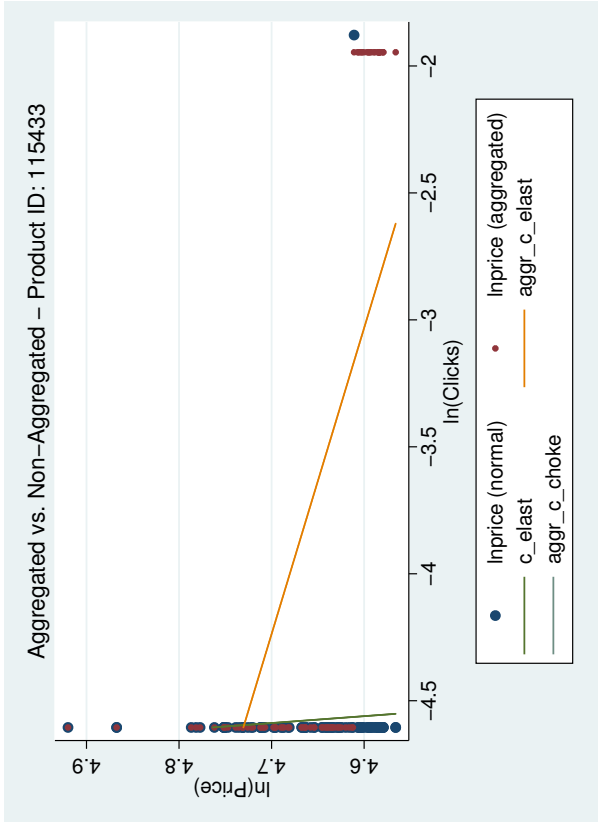


Figure 33: Product: 115433, HP Q6581A Fotopapier seidenmatt 42", 190g, 30.5m: Hardware ⇒ Verbrauchsmaterial

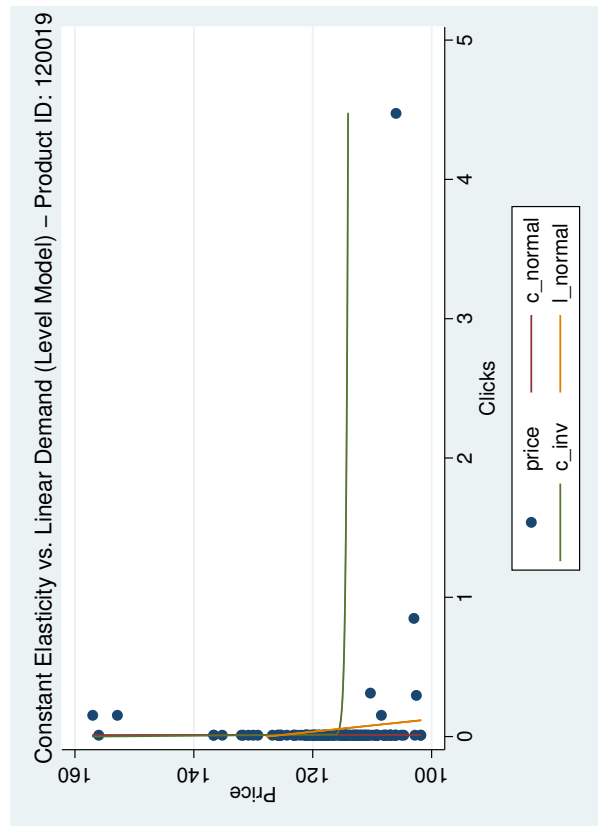
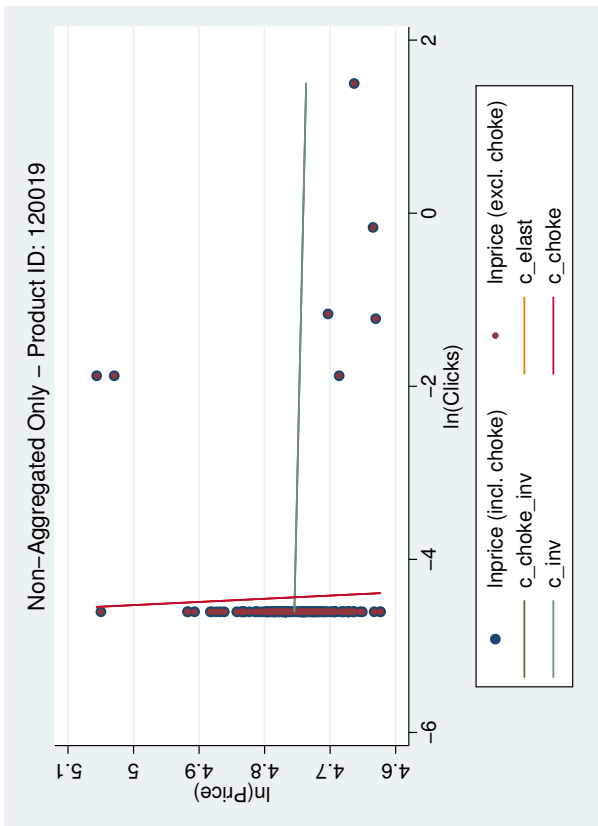
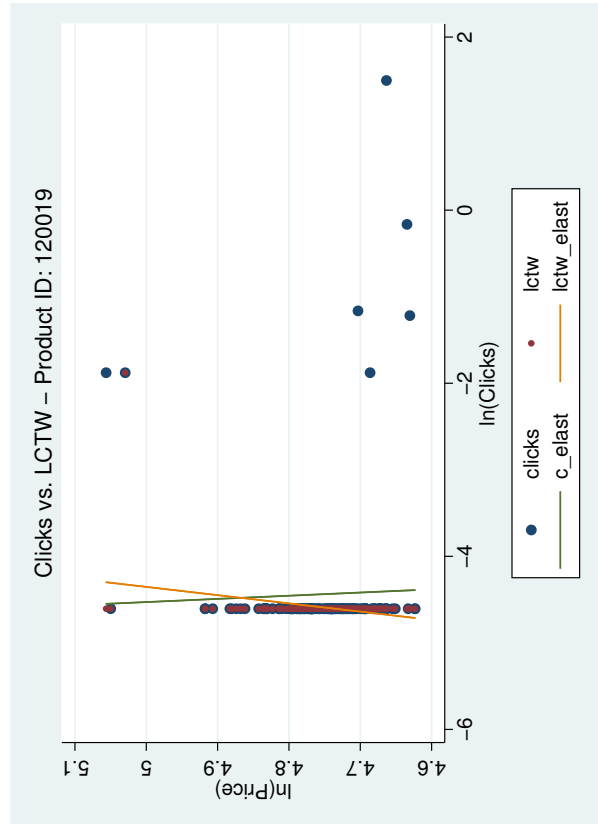
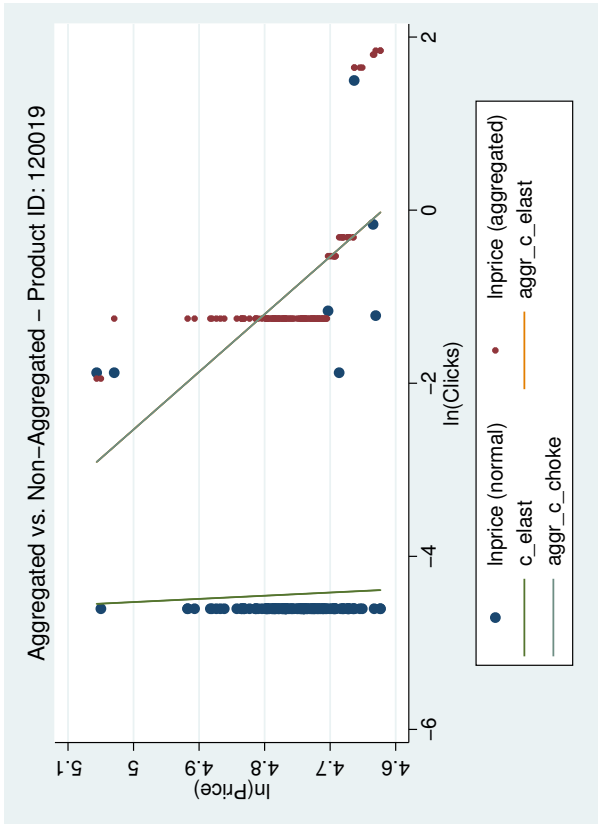


Figure 34: Product: 120019, Konica Minolta magicolor 5430 Toner schwarz (1710582-001): Hardware \Rightarrow Verbrauchsmaterial

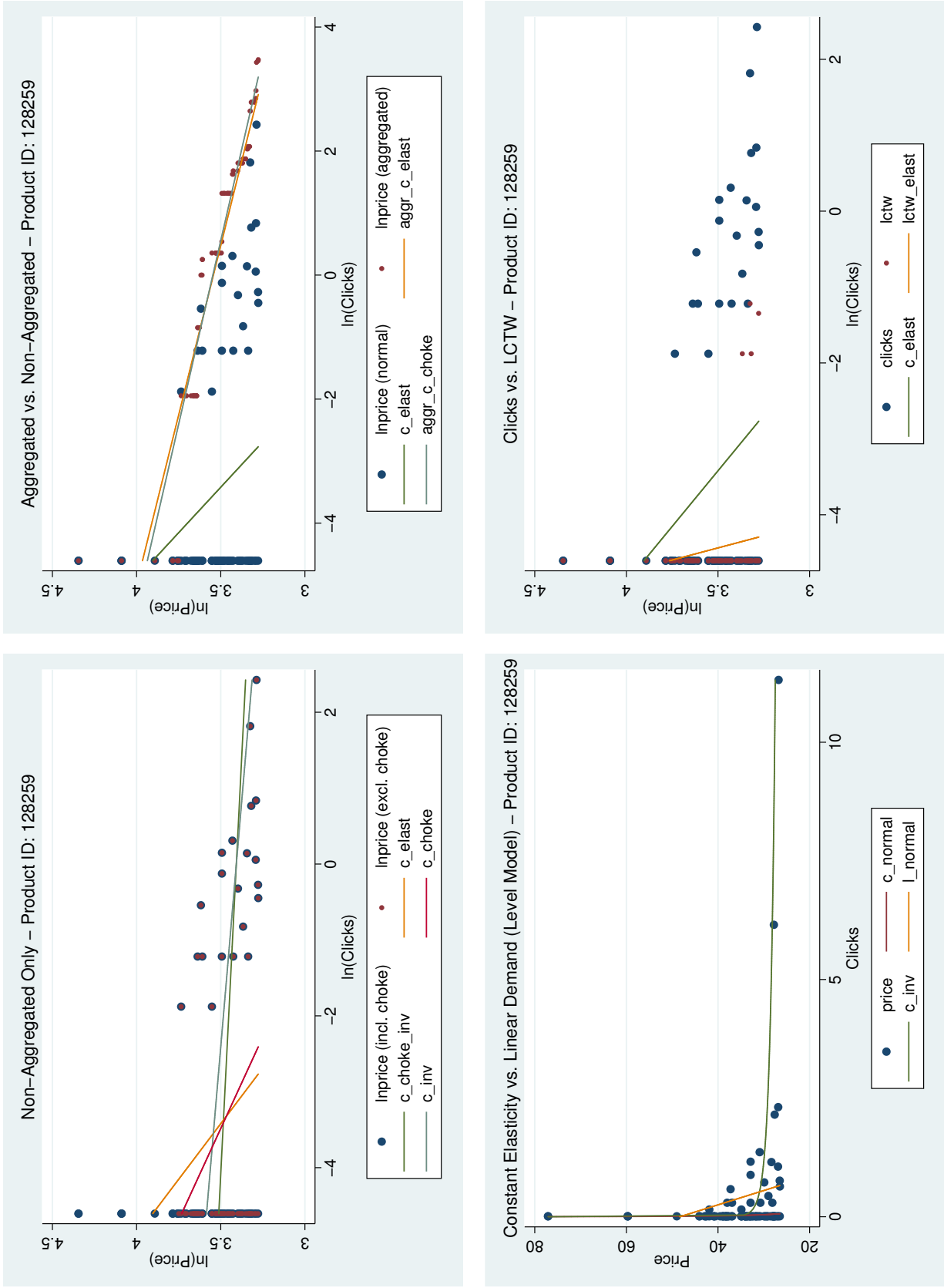


Figure 35: Product: 128259, Zalman CNPS7700-Cu CPU-Kühler (Sockel 478/775/754/939/940): Hardware ⇒ Luftkühlung

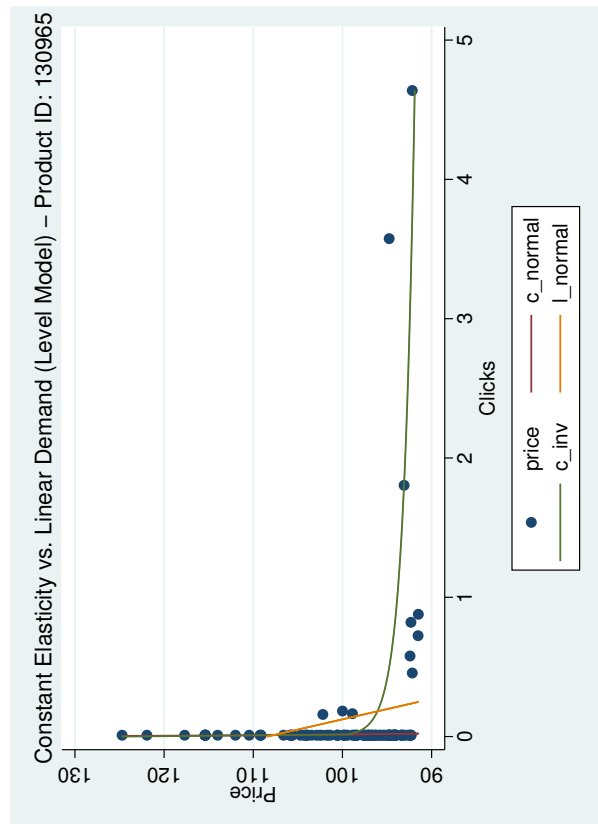
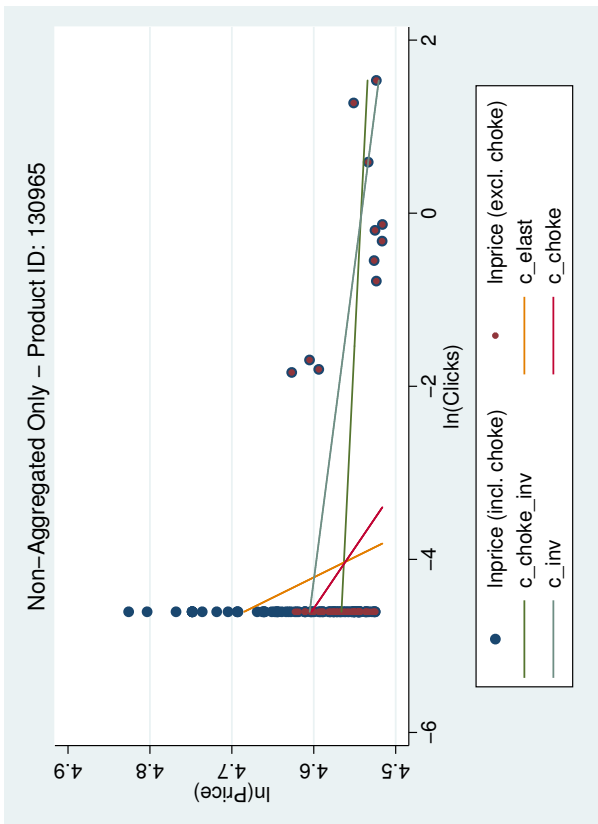
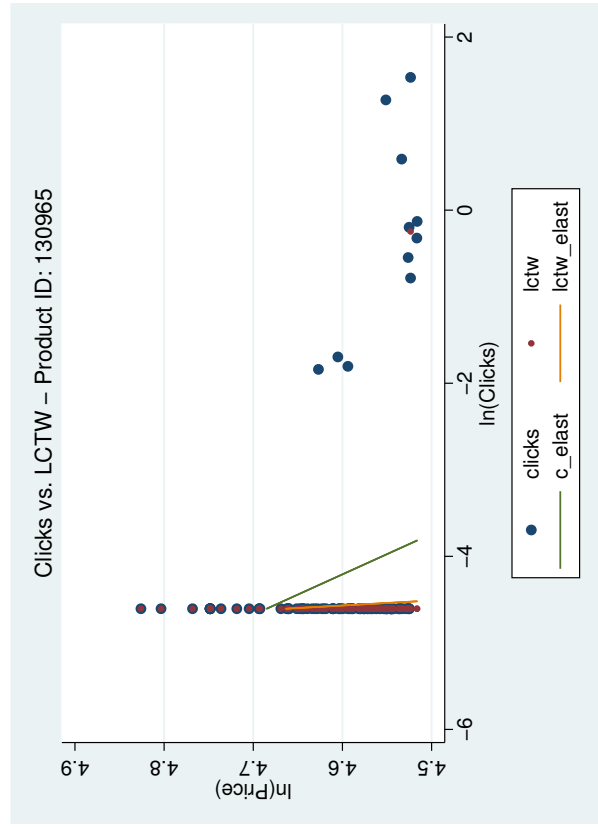
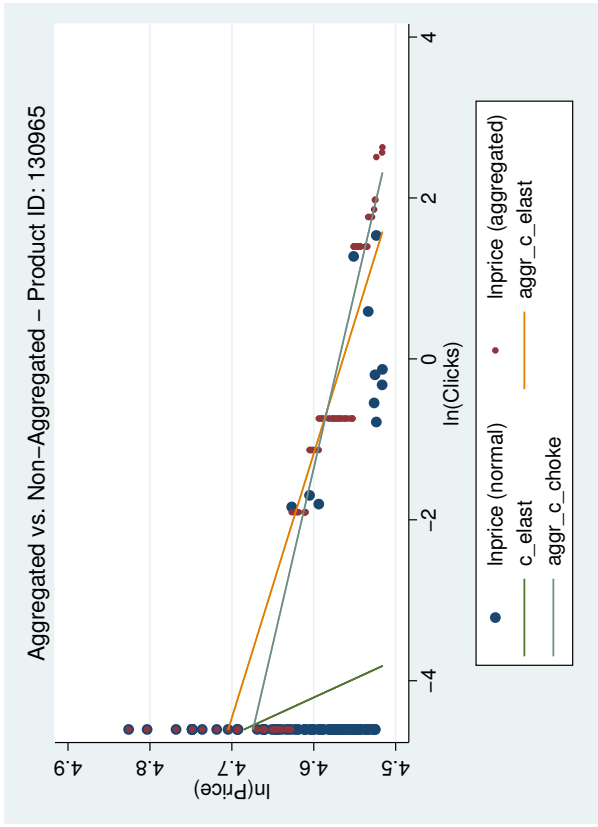


Figure 36: Product: 130965, FSC Pocket LOOX 710/720 Bump Case (S26391-F2611-L500): Hardware ⇒ PDAsGPS

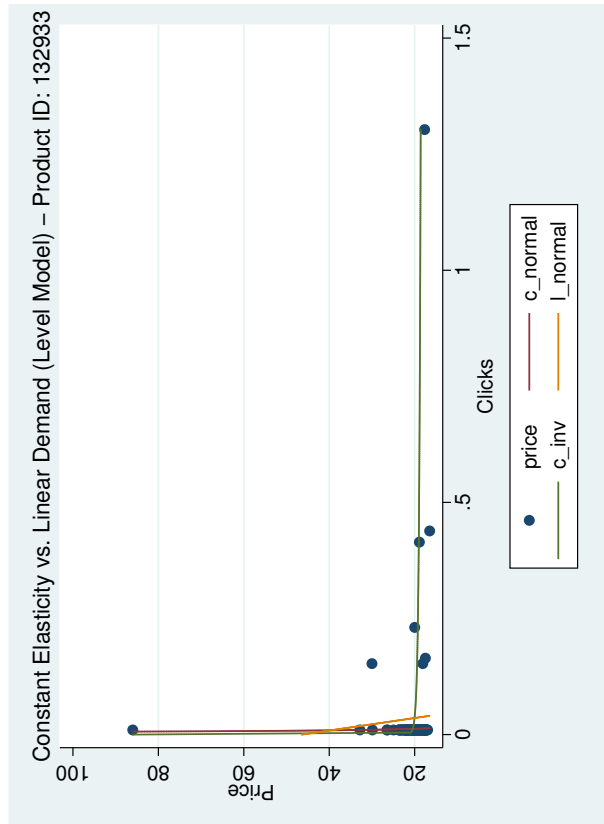
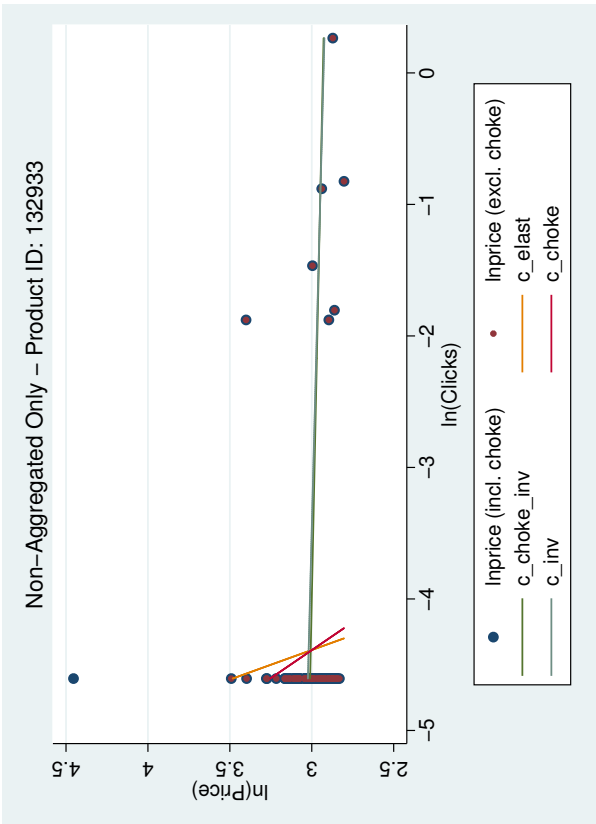
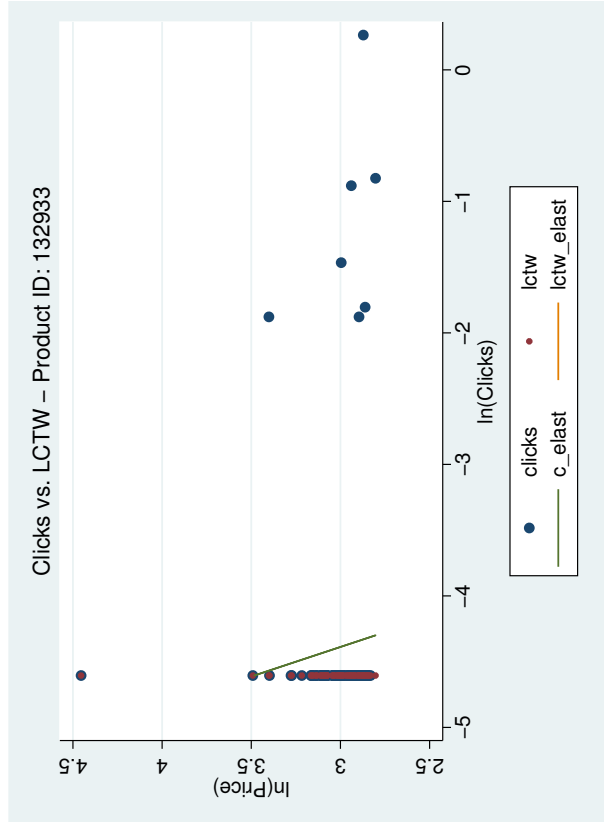
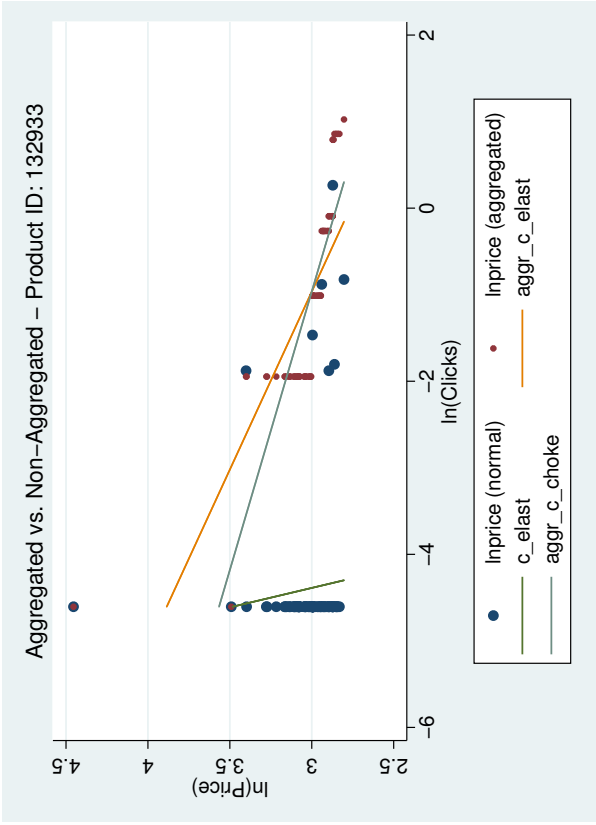


Figure 37: Product: 132933, Apple iPod AV Kabel (M9765G/A): AudioHiFi ⇒ PortableAudio

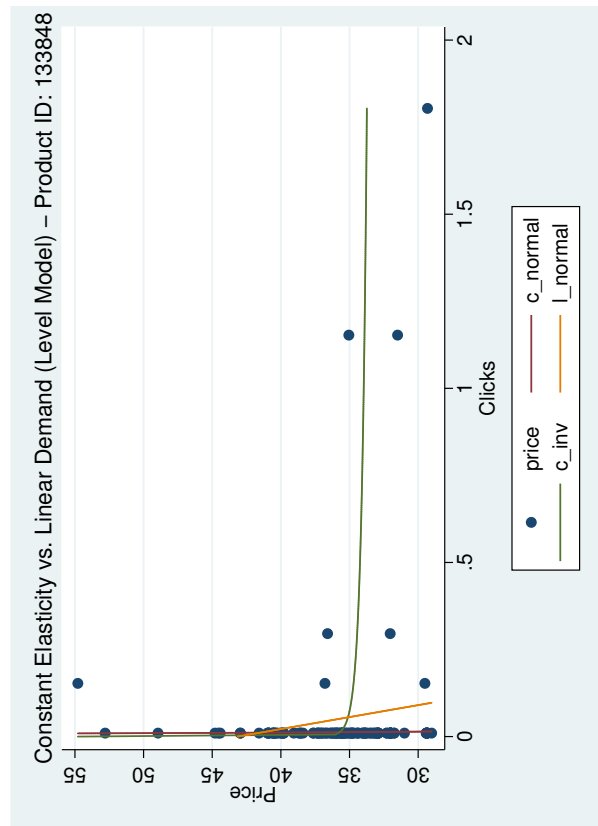
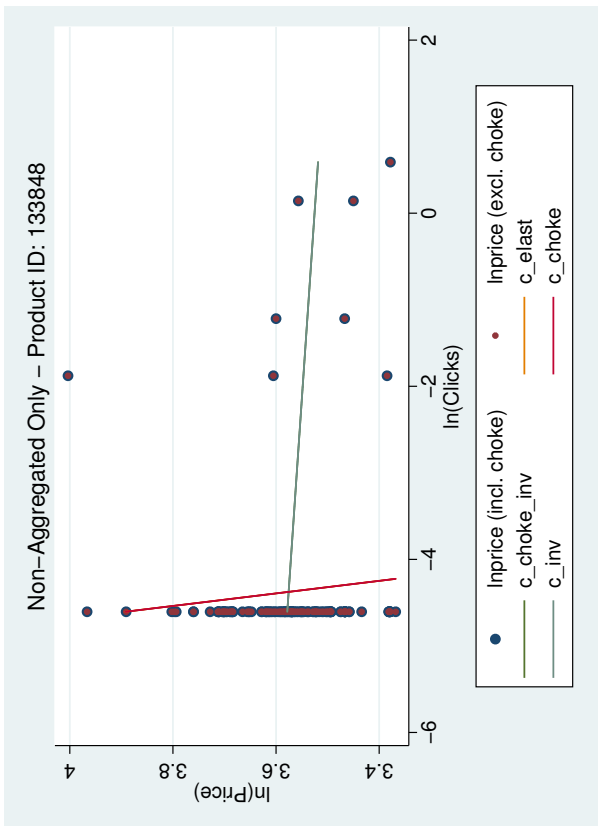
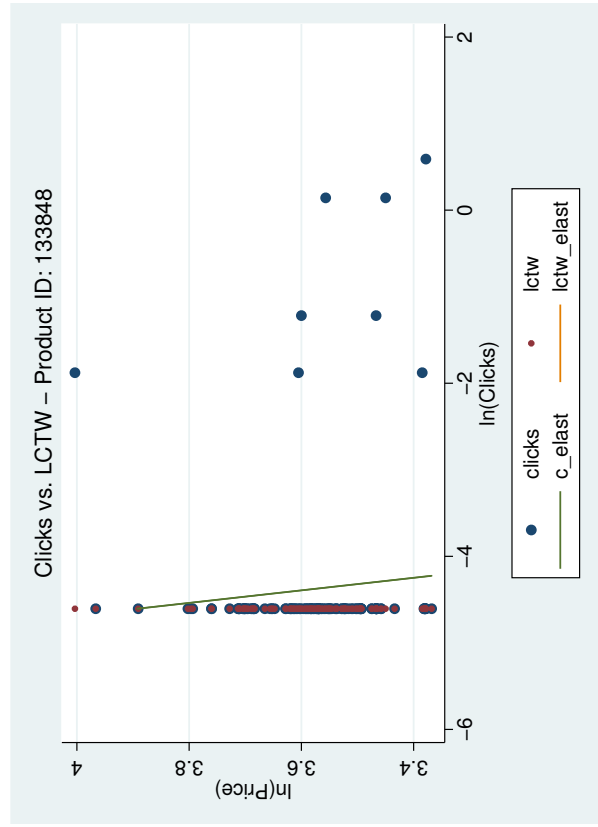
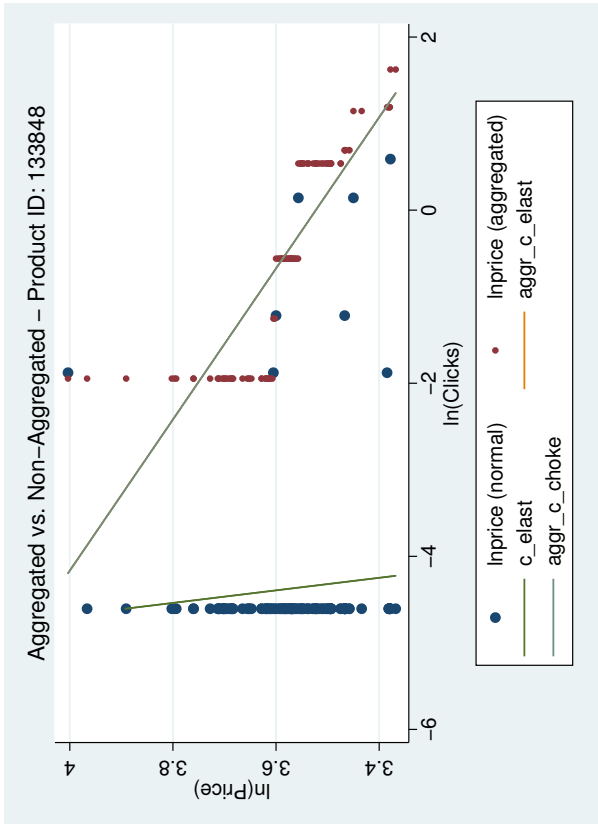


Figure 38: Product: 133848, Targus XL Metro Messenger Notebook Case Notebooktasche (TCG200): Hardware \Rightarrow Notebookzubehor

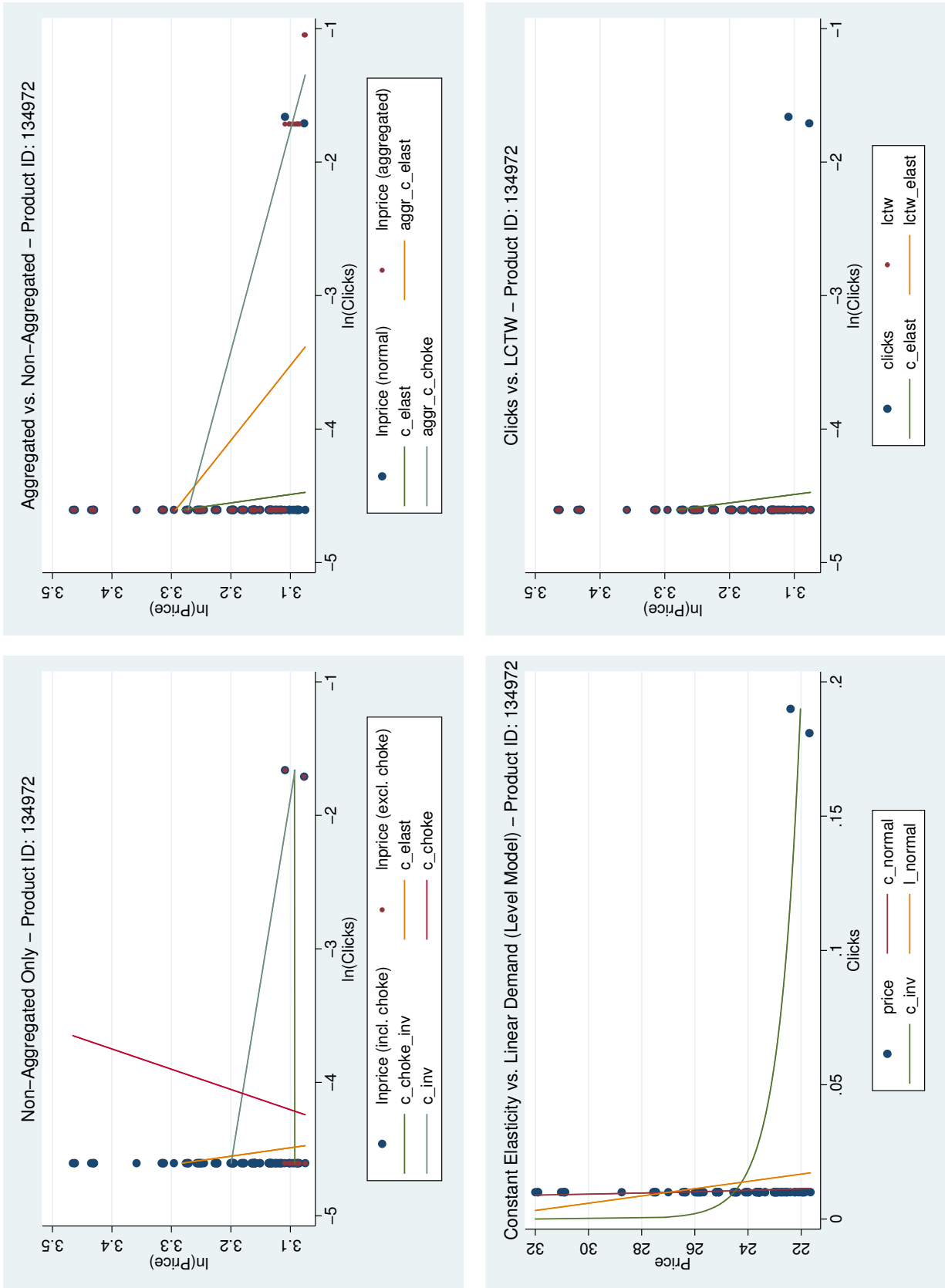


Figure 39: Product: 134972, Kingston ValueRAM DIMM 256MB PC2-533 ECC DDR2 CL4 (KVR533D2E4/256): Hardware \Rightarrow Speicher

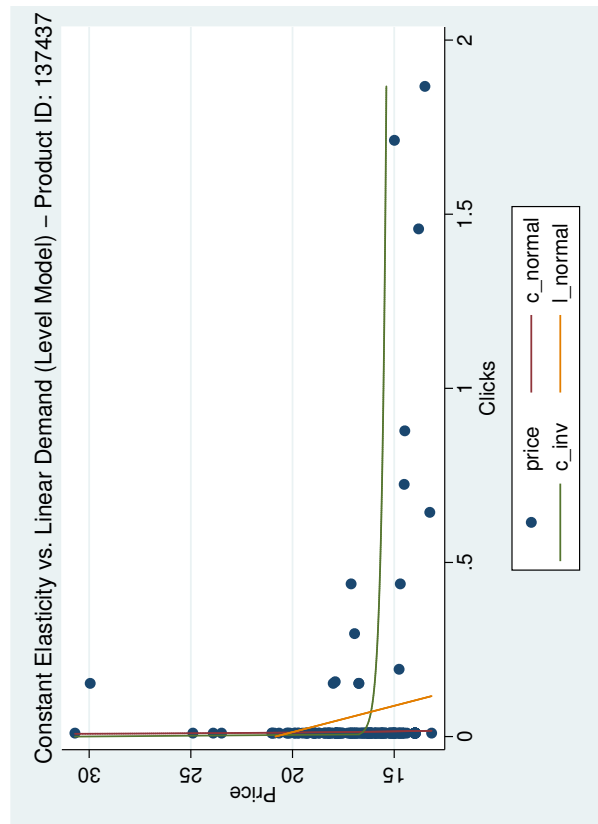
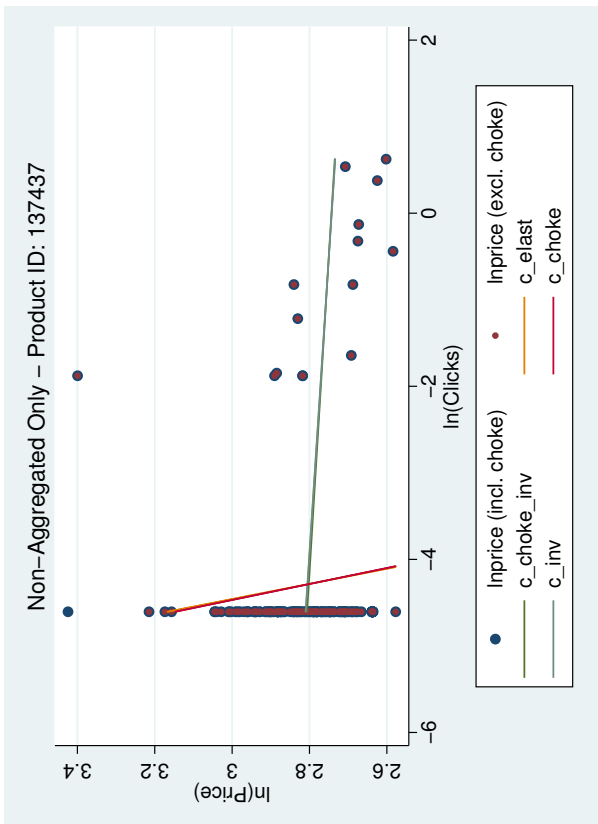
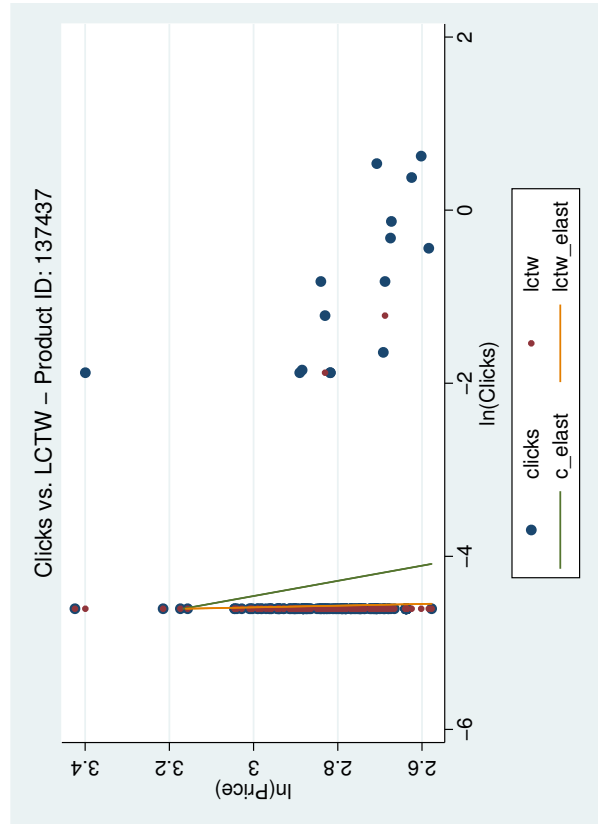
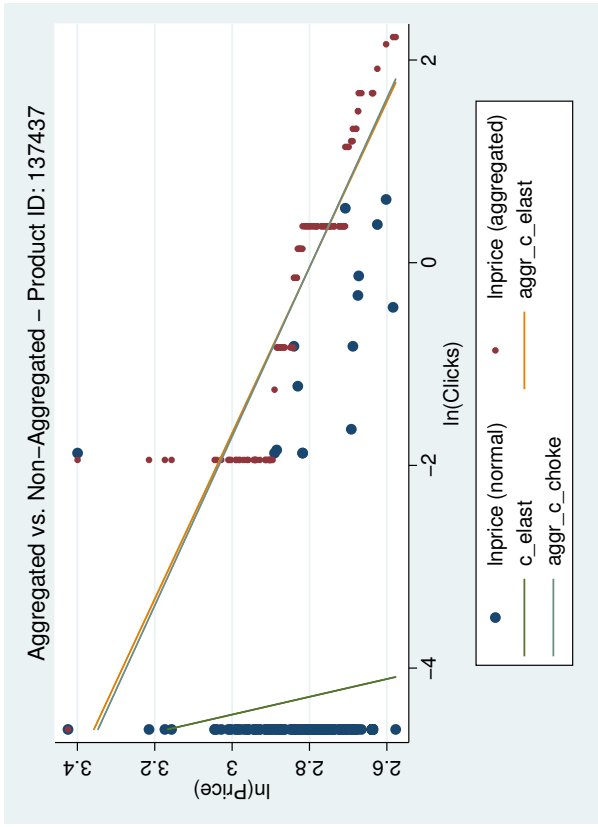


Figure 40: Product: 137437, Canon DCC-80 Software (0021X145): VideoFotoTV ⇒ FotoVideozubehr

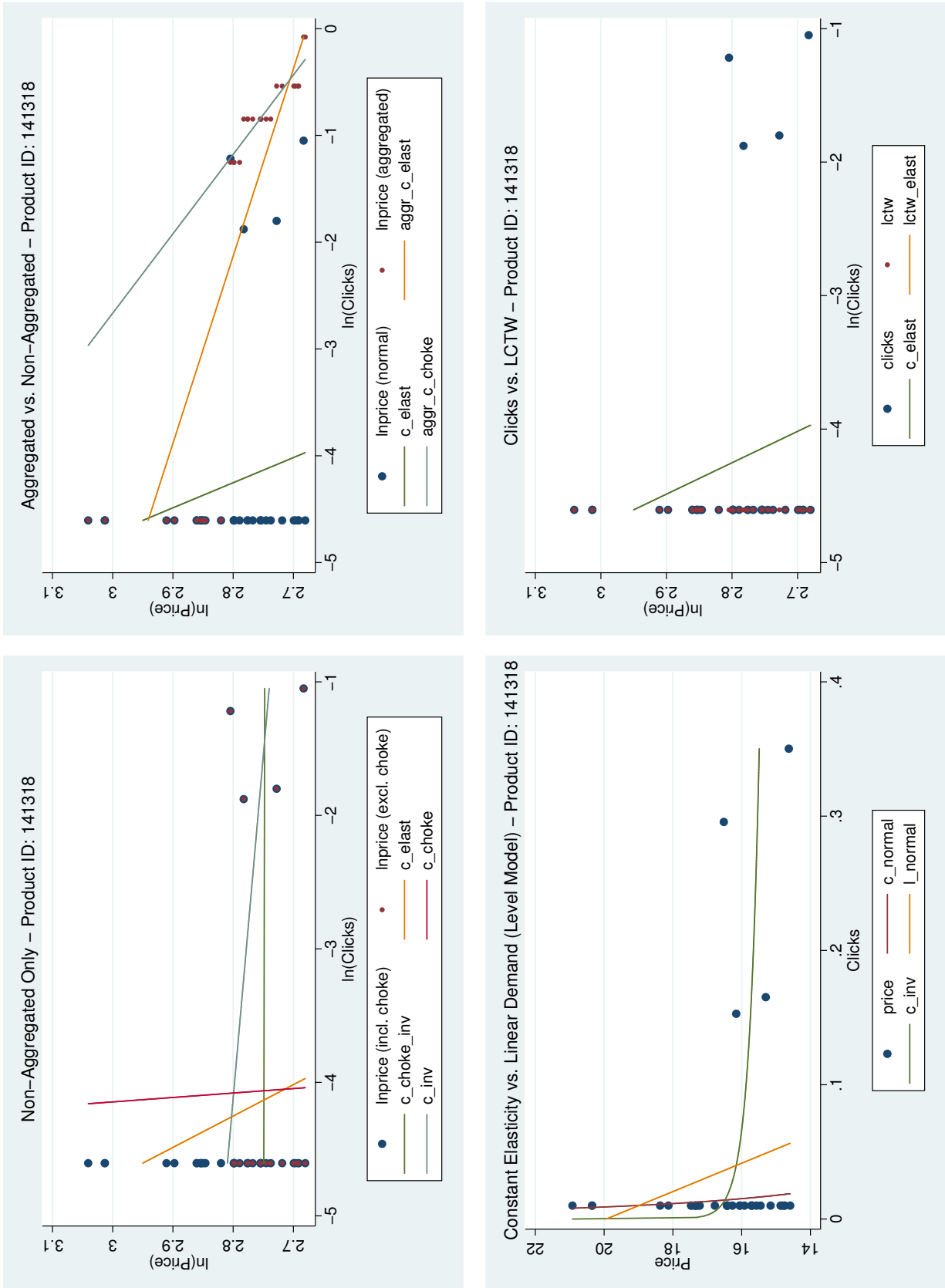


Figure 41: Product: 141318, Cherry Linux, PS/2, DE (G83-6188): Hardware \Rightarrow Eingabegeräte

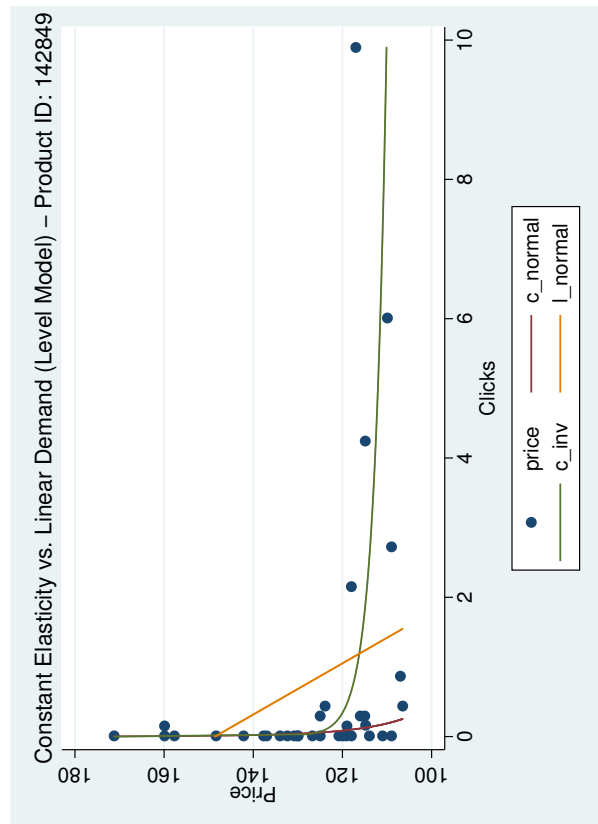
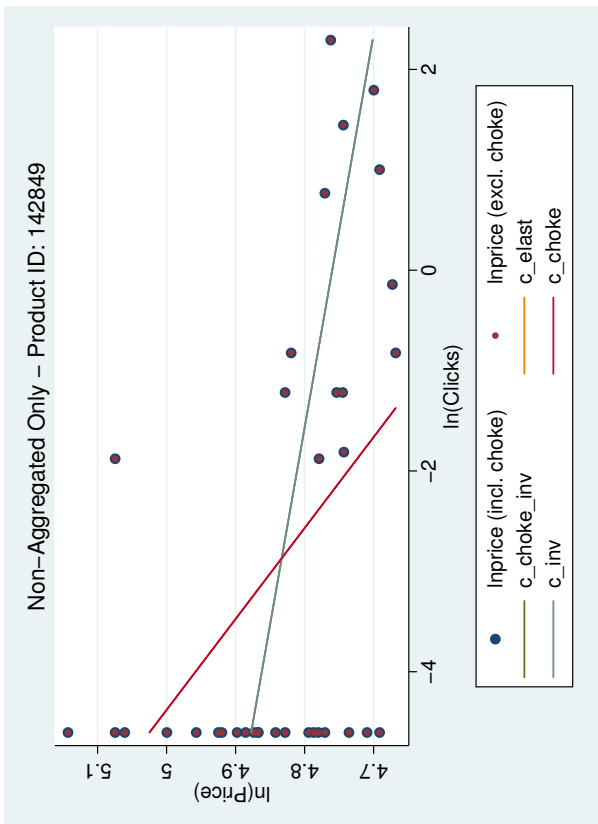
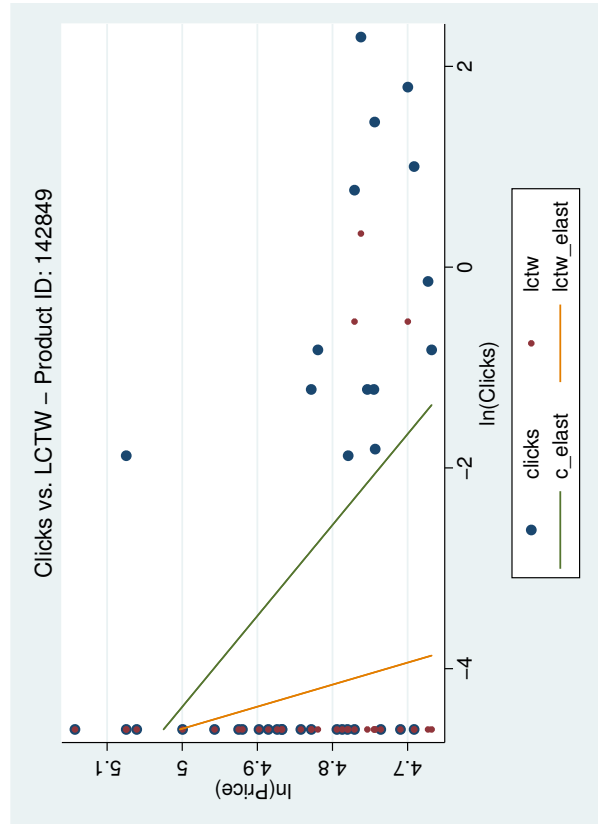
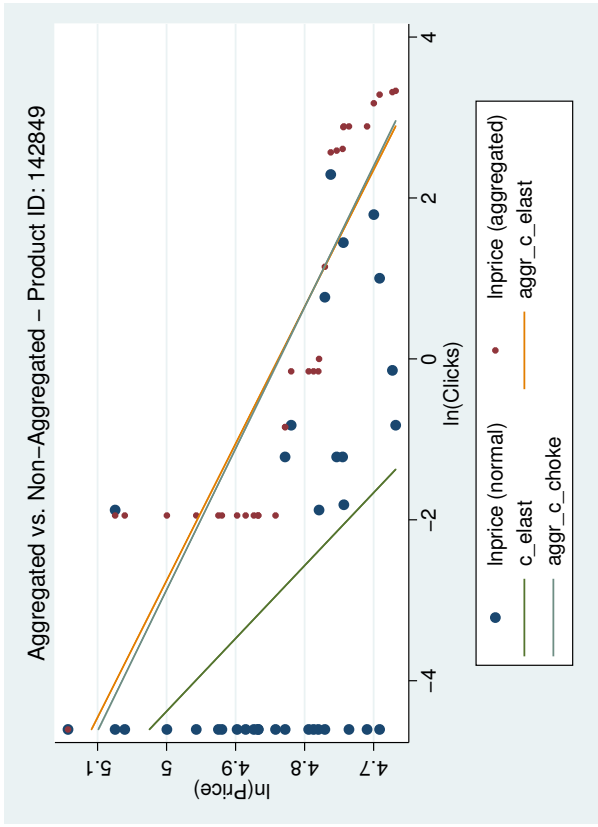


Figure 42: Product: 142849, Philips HR1861/00 Entsafter : ⇒

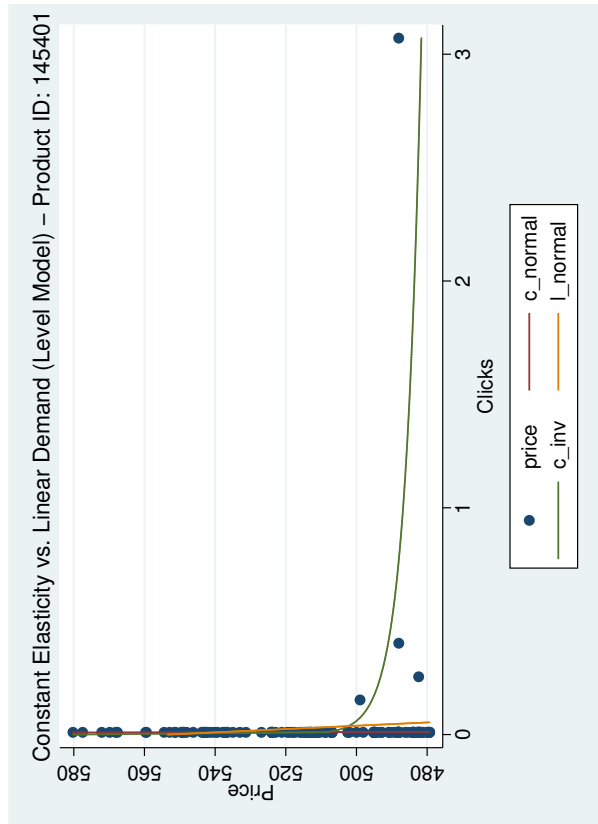
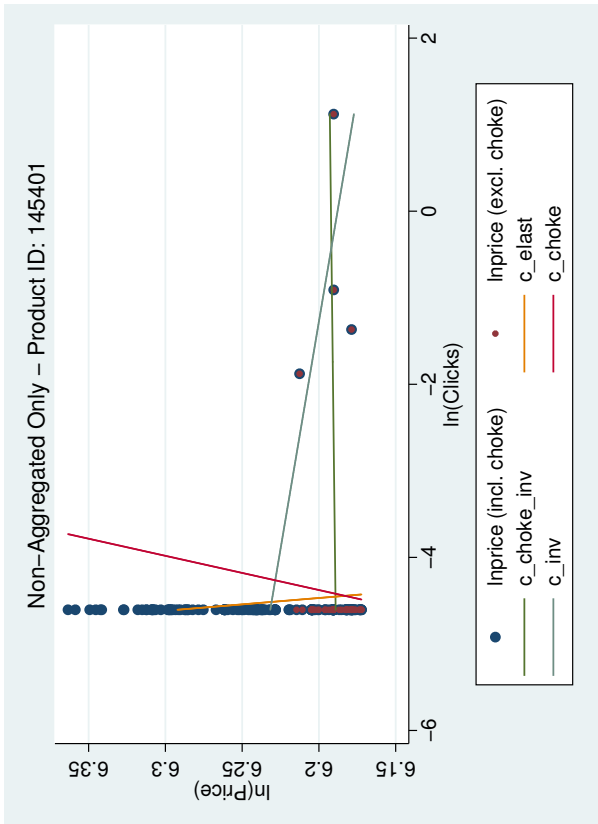
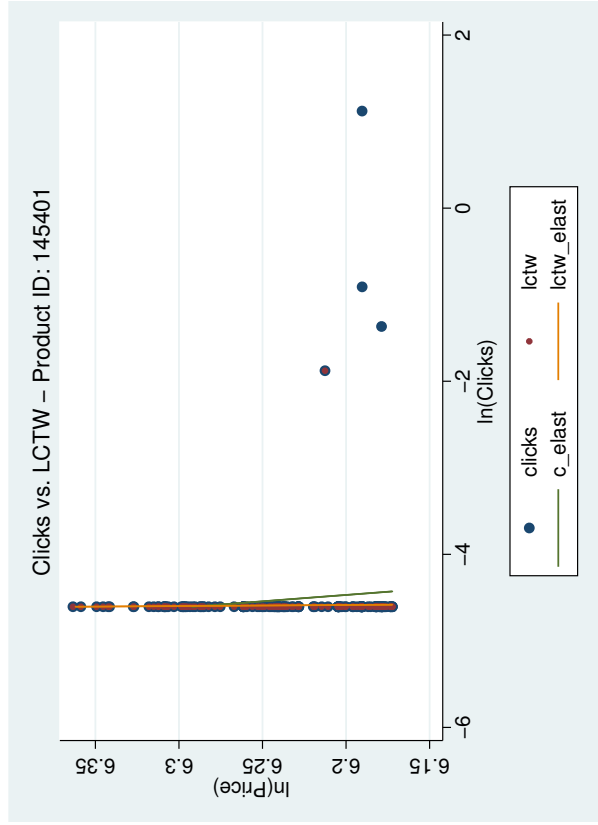
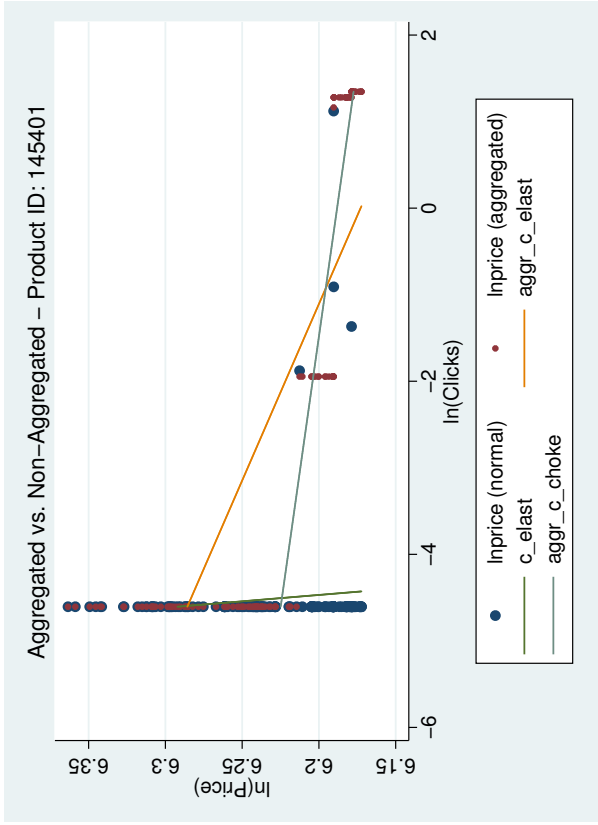


Figure 43: Product: 145401, Adobe: GoLive CS2 (deutsch) (PC) (23200457): =>

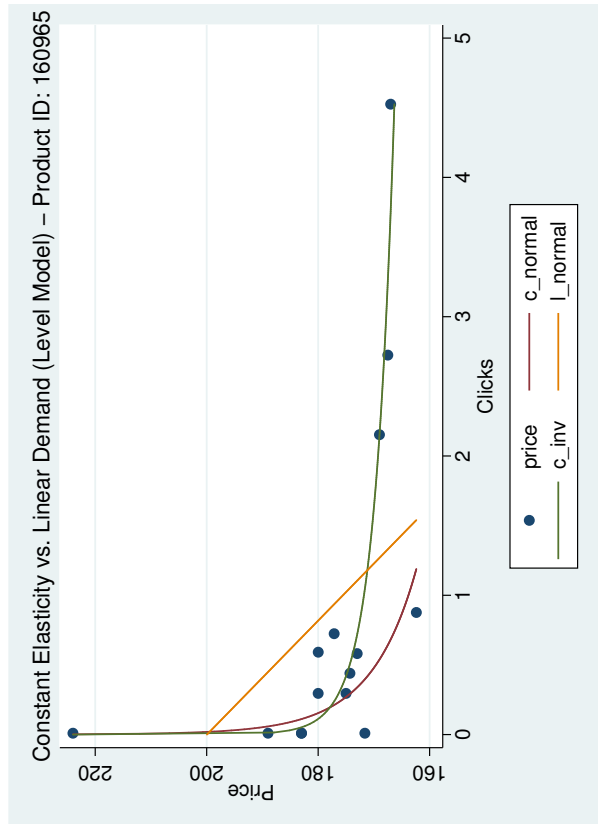
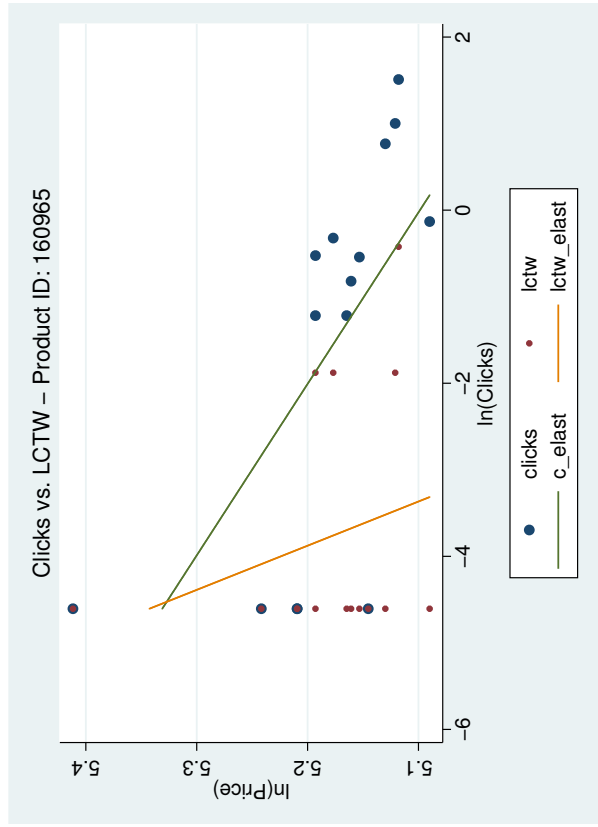
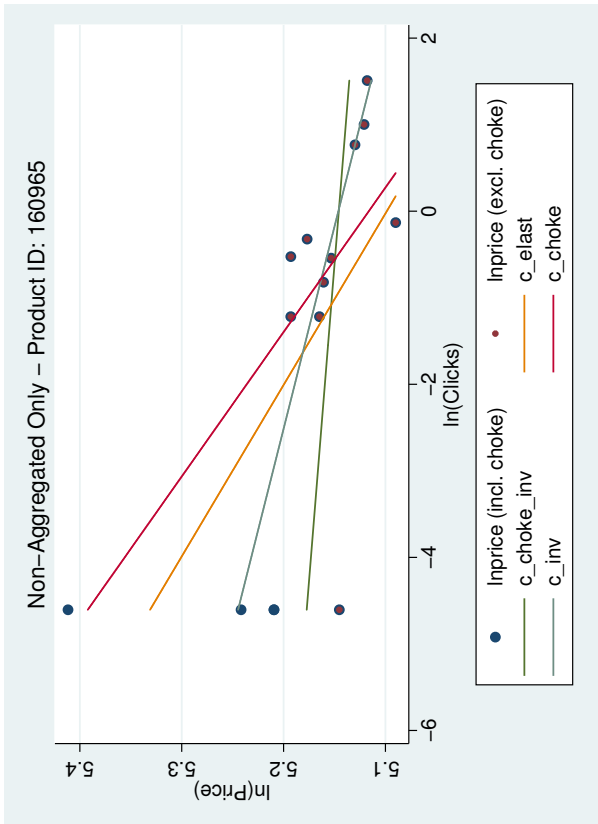
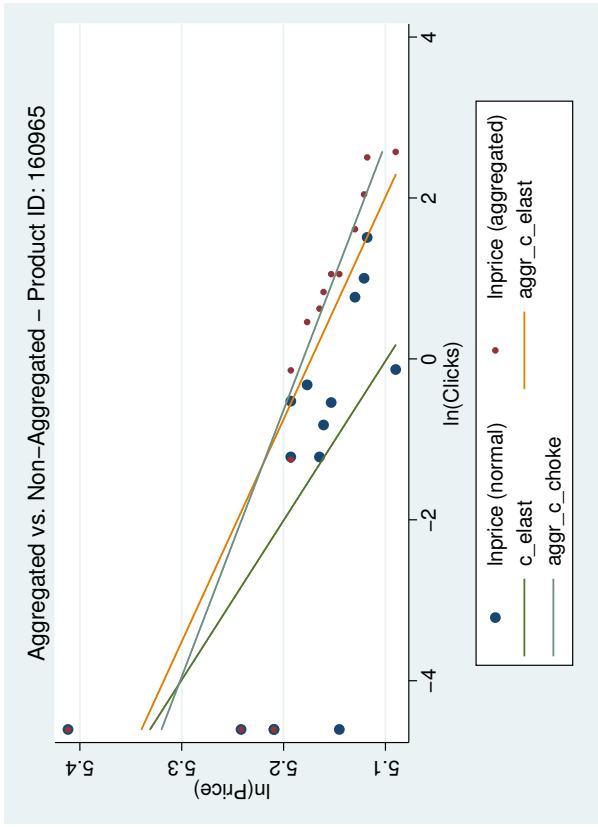


Figure 44: Product: 160965, Creative Sound Blaster X-Fi Platinum (705B046002000): Hardware ⇒ PCAudio

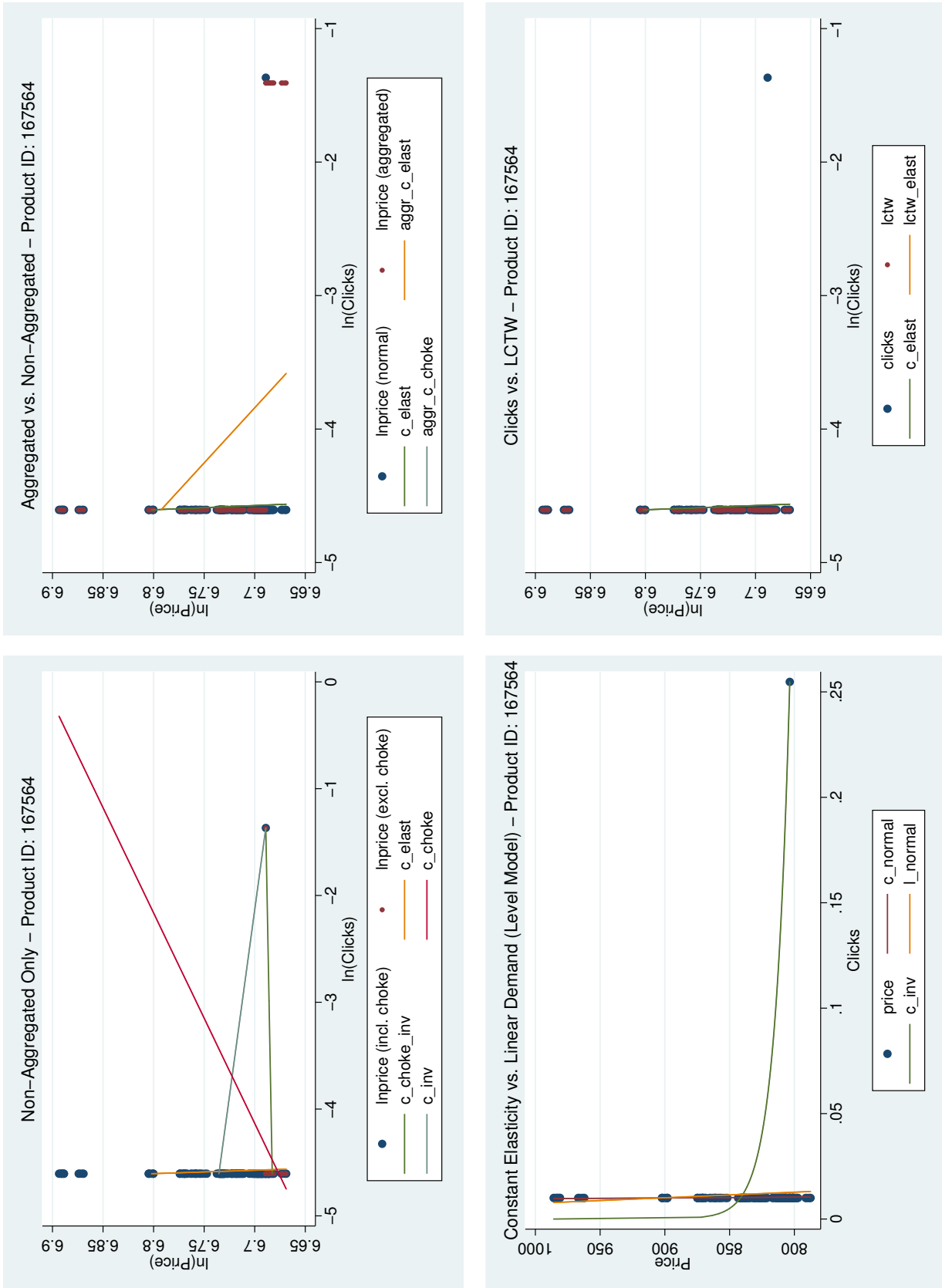


Figure 45: Product: 167564, Intel Xeon DP 3.80GHz, 200MHz FSB, 2048kB Cache, 604-pin boxed passiv (BX80546KG3800FP): Hardware \Rightarrow CPUs

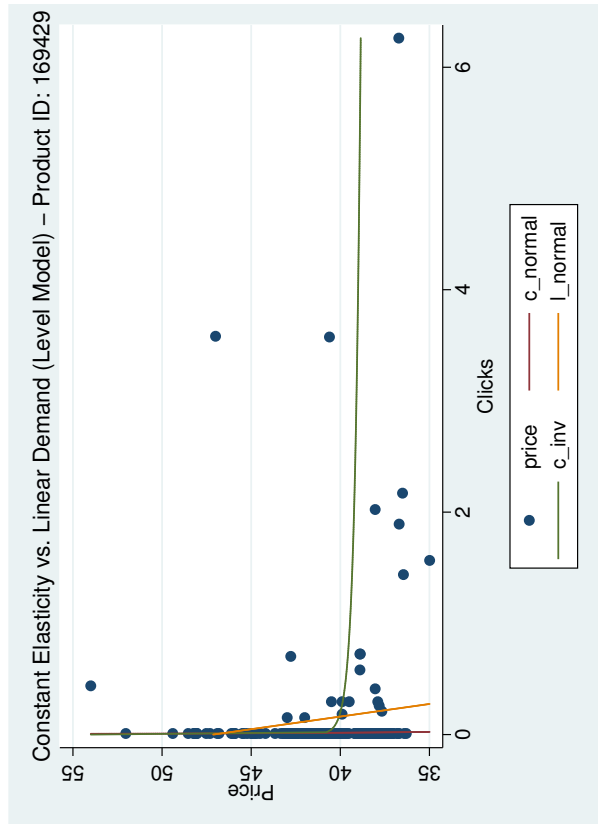
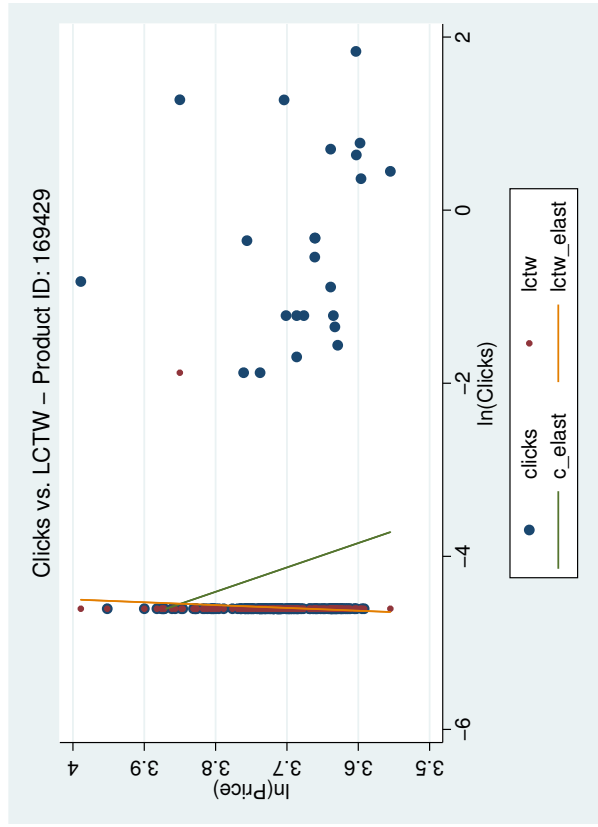
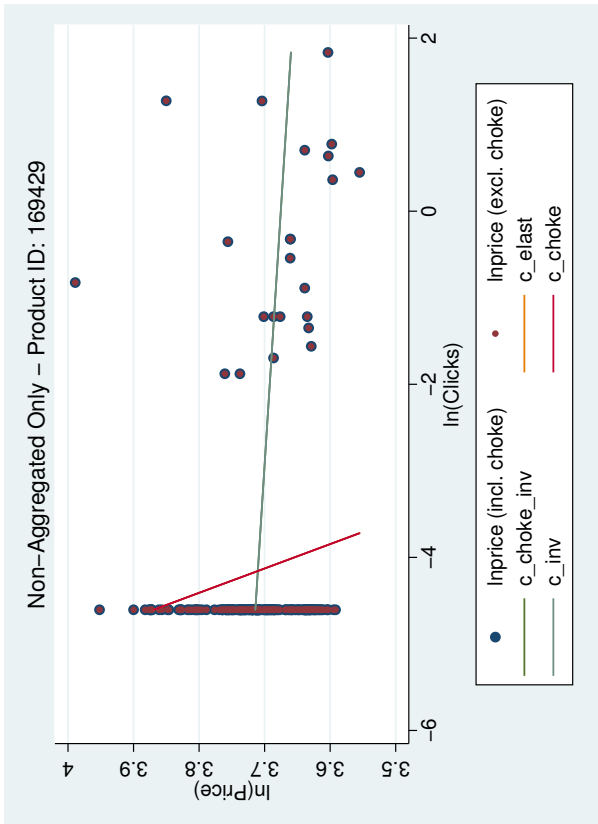
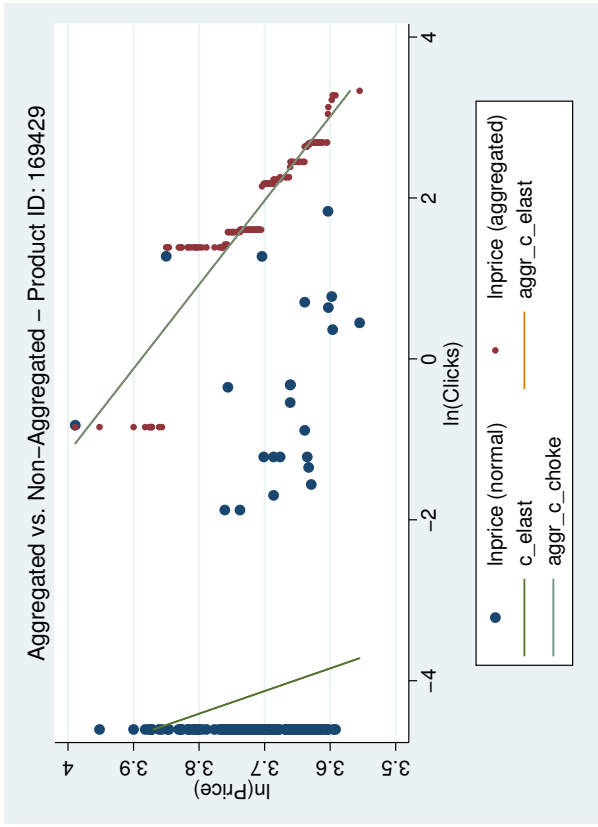


Figure 46: Product: 169429, MSI K8MM3-V, K8M800 (PC3200 DDR) (MS-7181-020R): Hardware ⇒ Mainboards

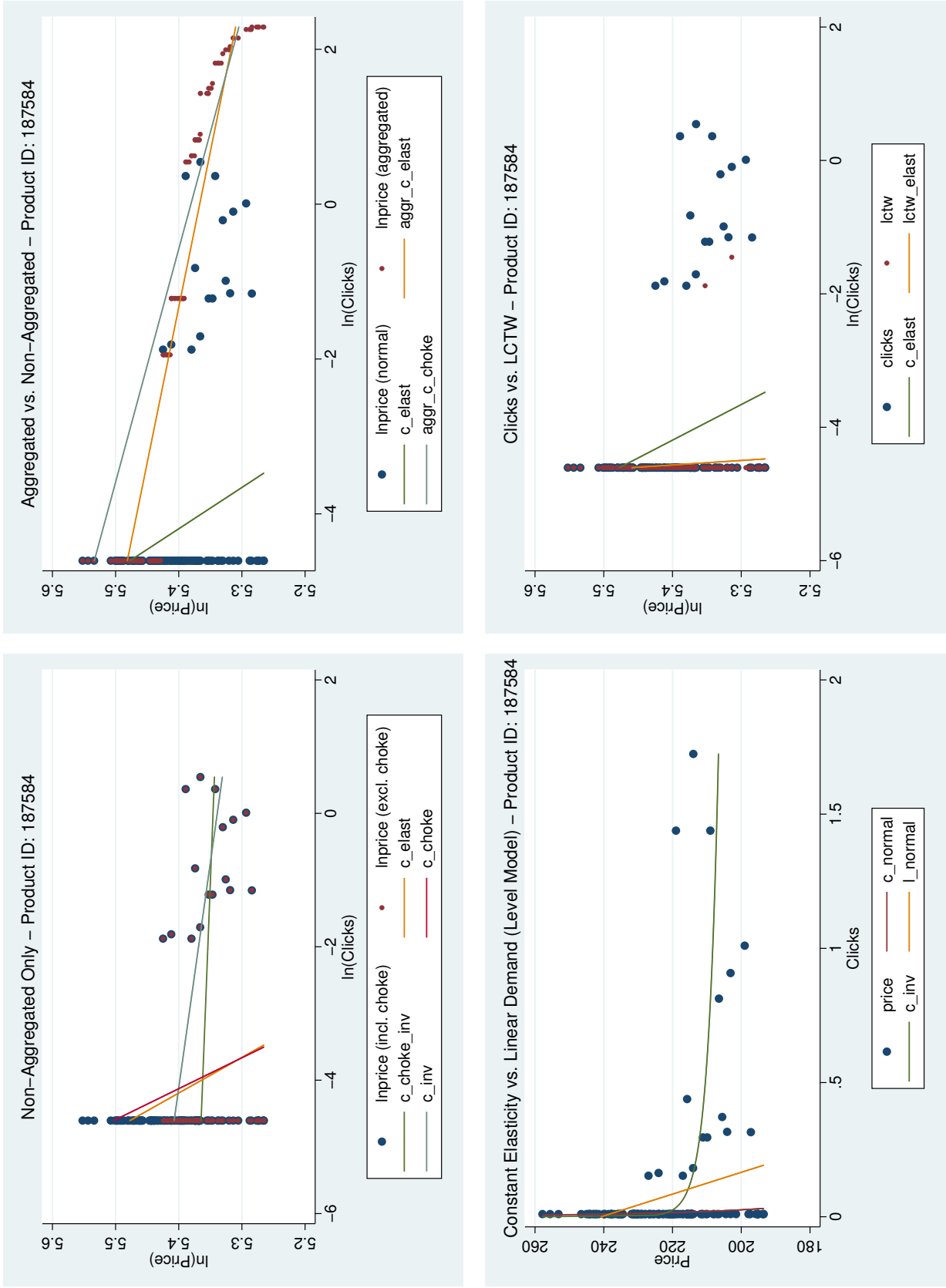


Figure 47: Product: 187584, Samsung SyncMaster 940N Pivot, 19", 1280x1024, analog (LS19HAATSB): Hardware \Rightarrow Monitore

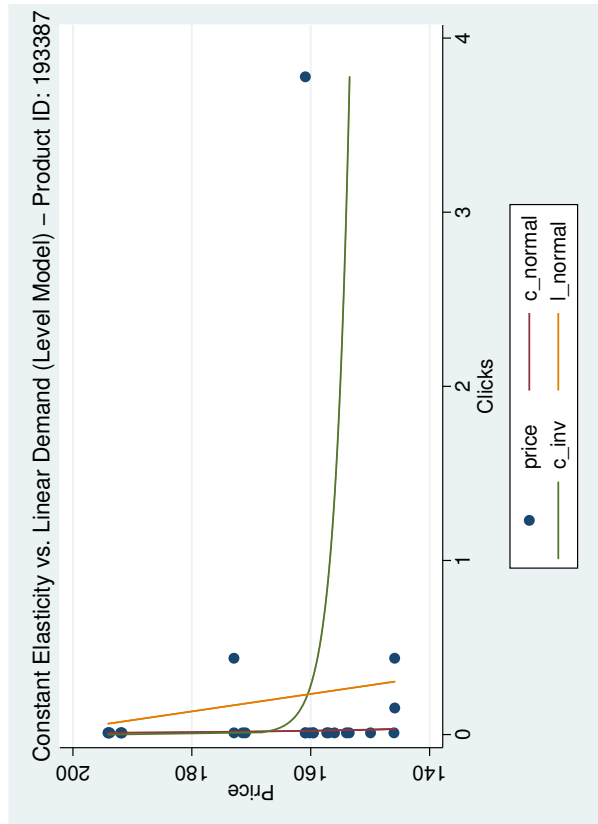
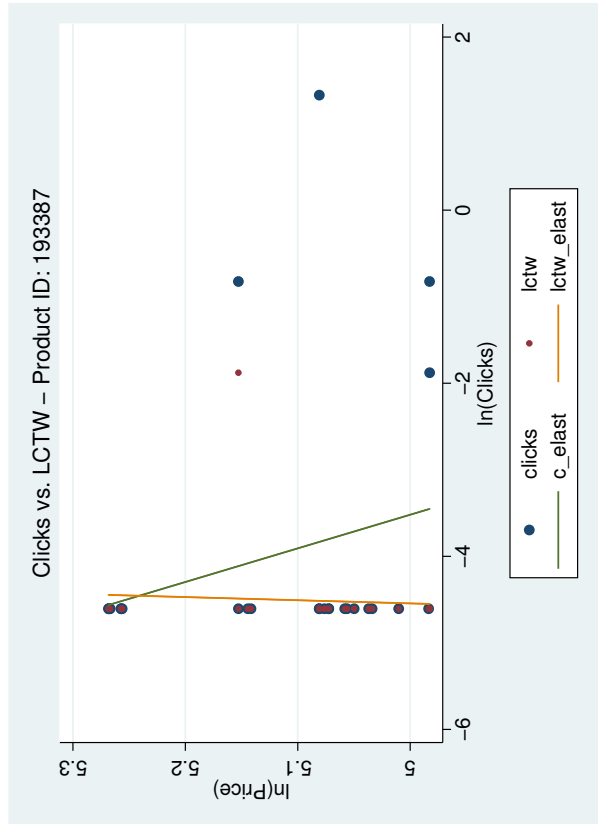
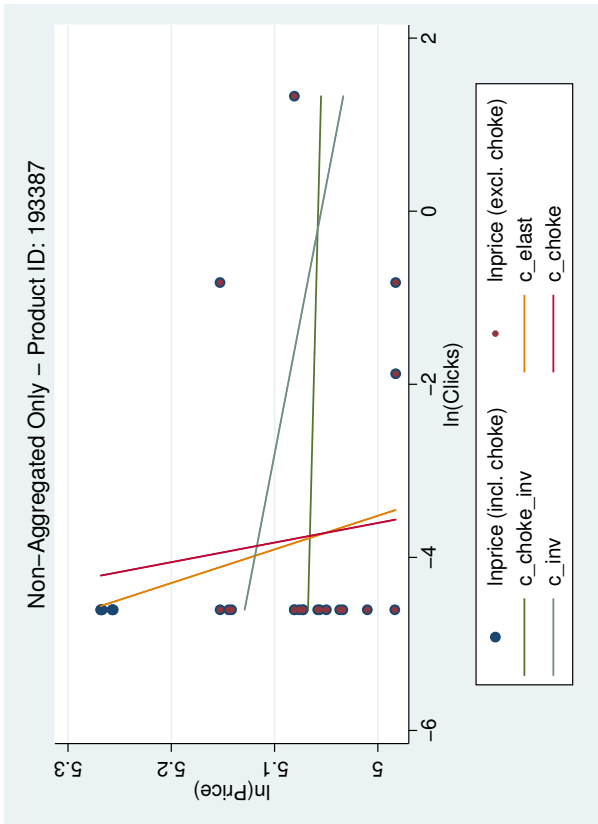
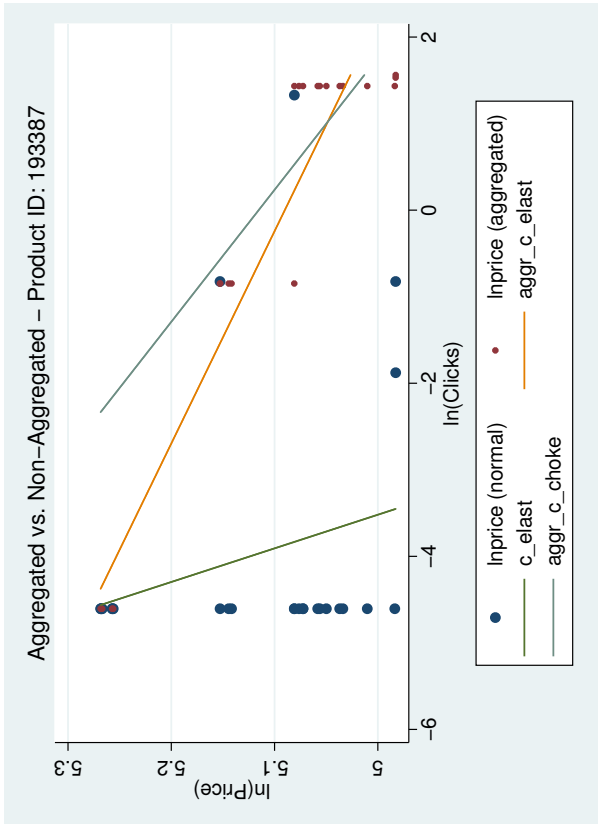


Figure 48: Product: 193387, ELO: EloOffice 7.0 Update (deutsch) (PC) (9303-70-49): Software ⇒ SicherheitBackup

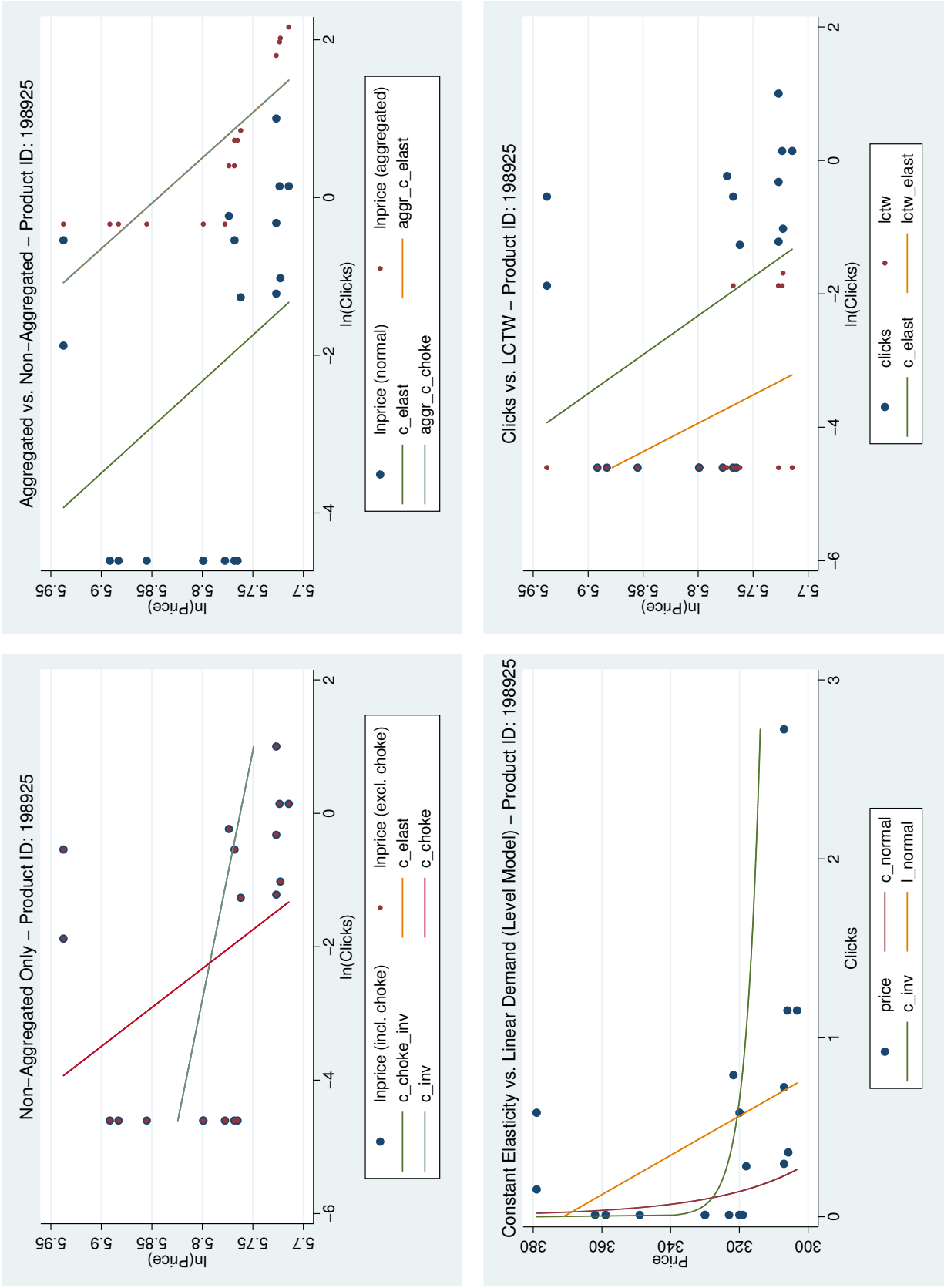


Figure 49: Product: 198925, Liebherr KTP 1810: Household \Rightarrow Kchengertegro

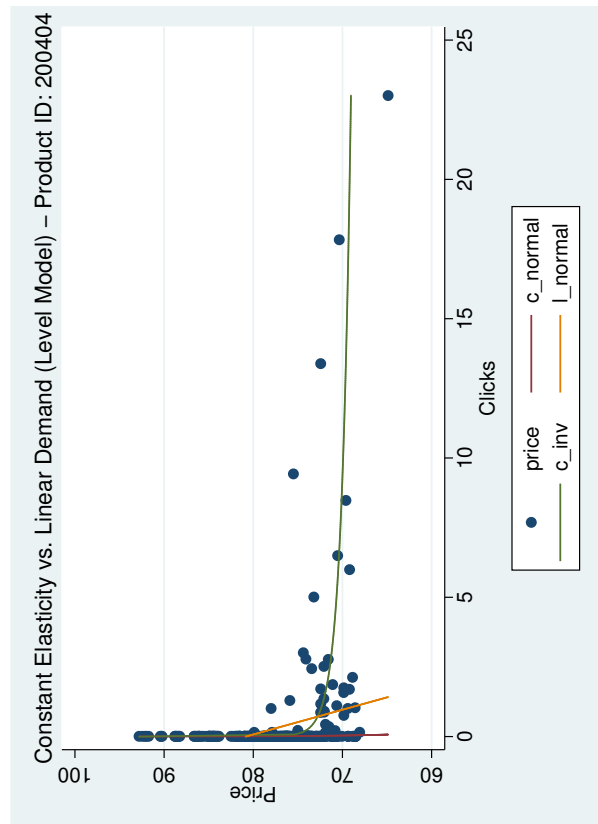
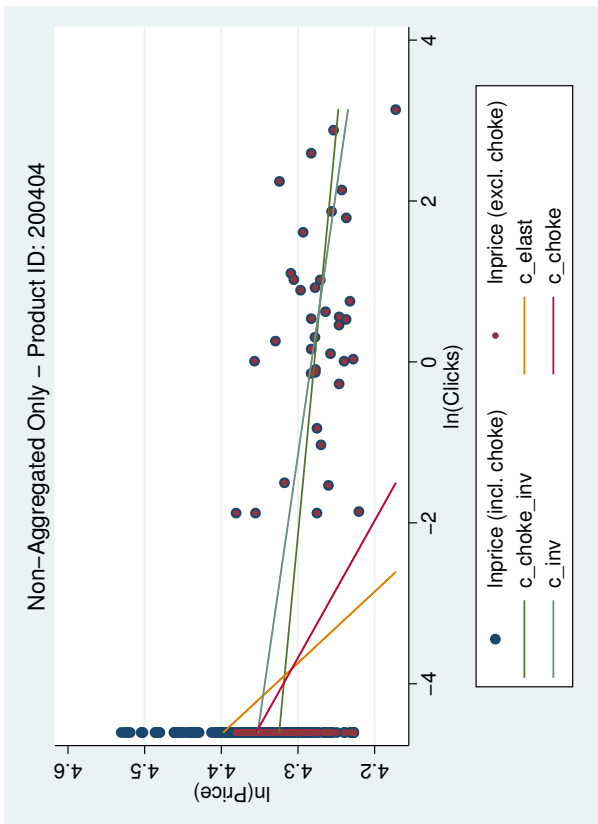
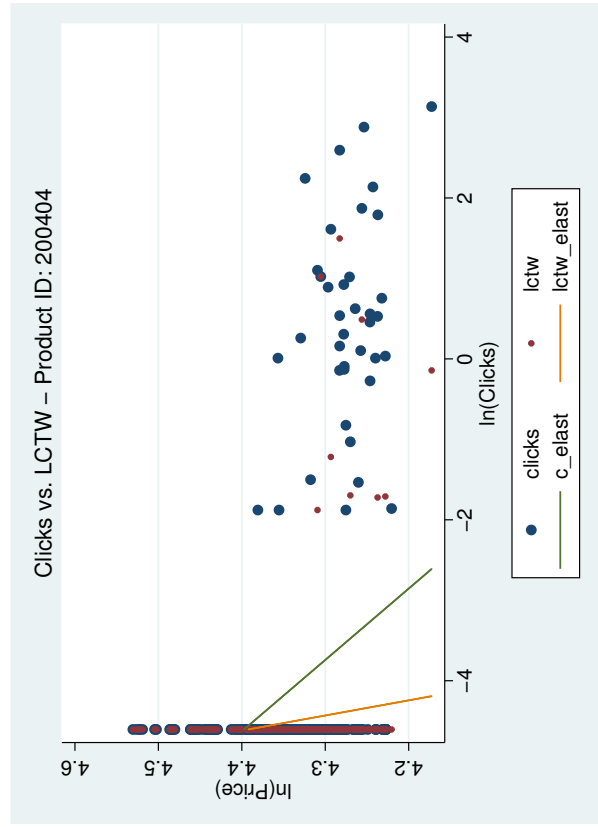
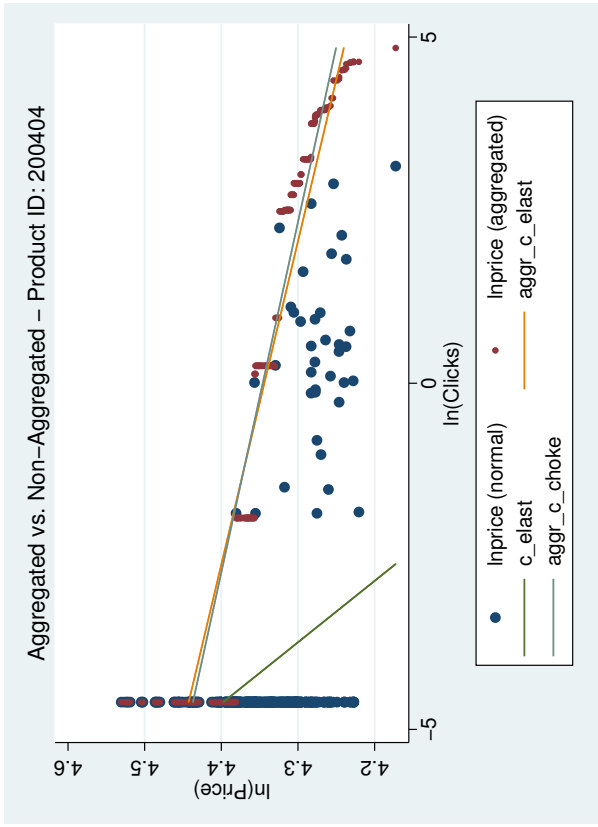


Figure 50: Product: 200404, ASUS M2N1PV-VM, GeForce 6150/MCP 430 (dual PC2-800 DDR2): Hardware ⇒ Mainboards

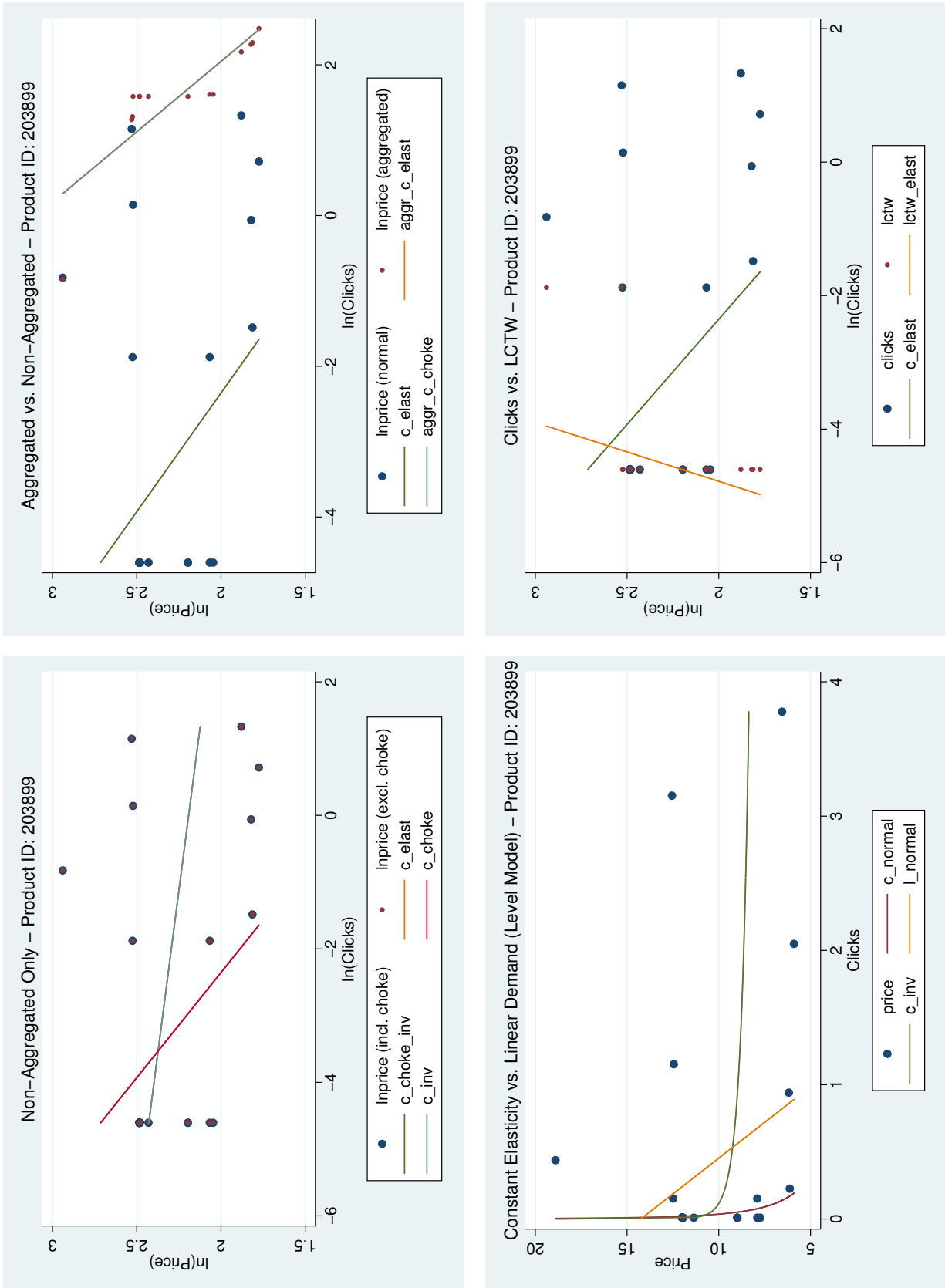


Figure 51: Product: 203899, Hama 35in1 Card Reader, USB 2.0 (55312): Hardware \Rightarrow SpeichermedienLesegeräte

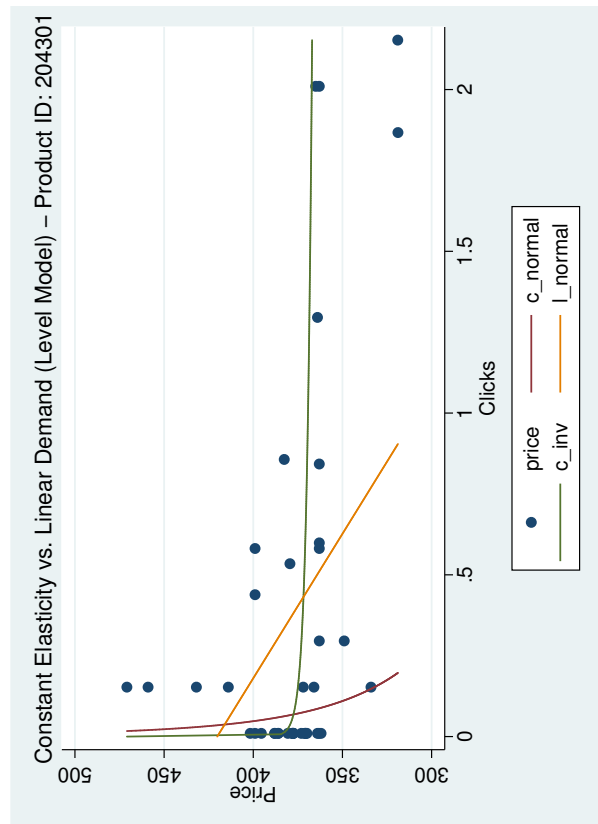
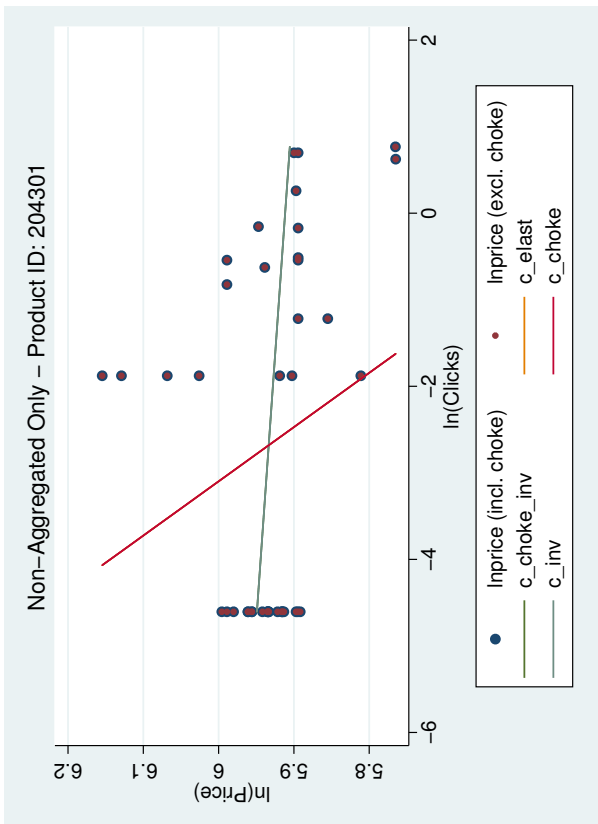
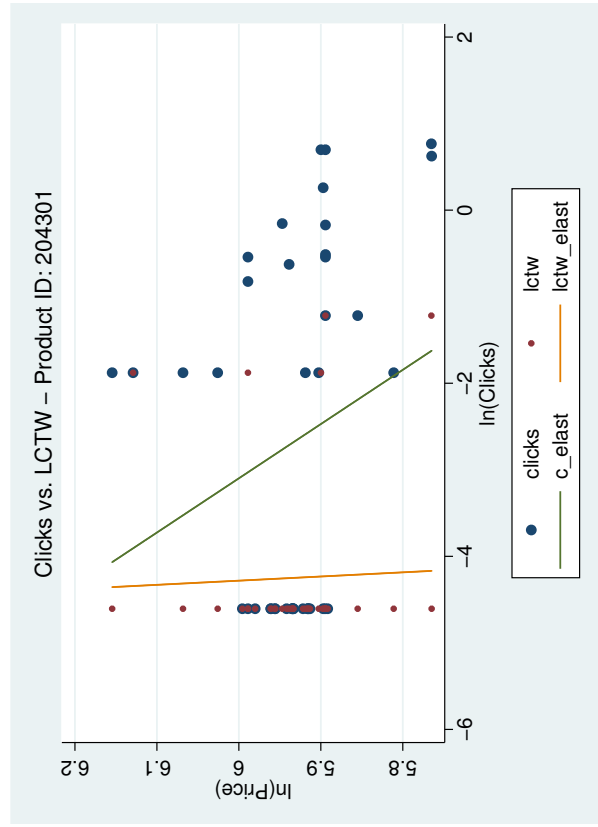
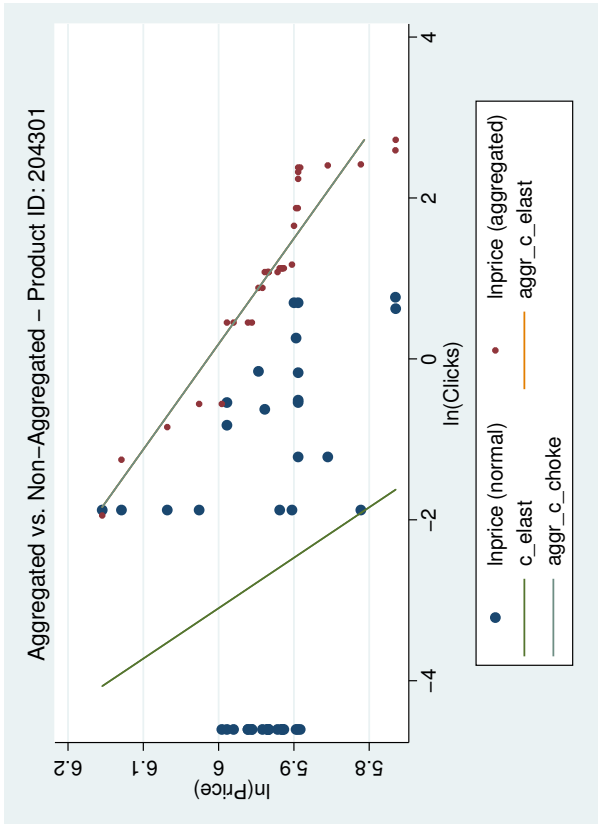


Figure 52: Product: 204301, Sony HVL-F56AM Blitzgerät: VideoFotoTV ⇒ FotoVideozubehor

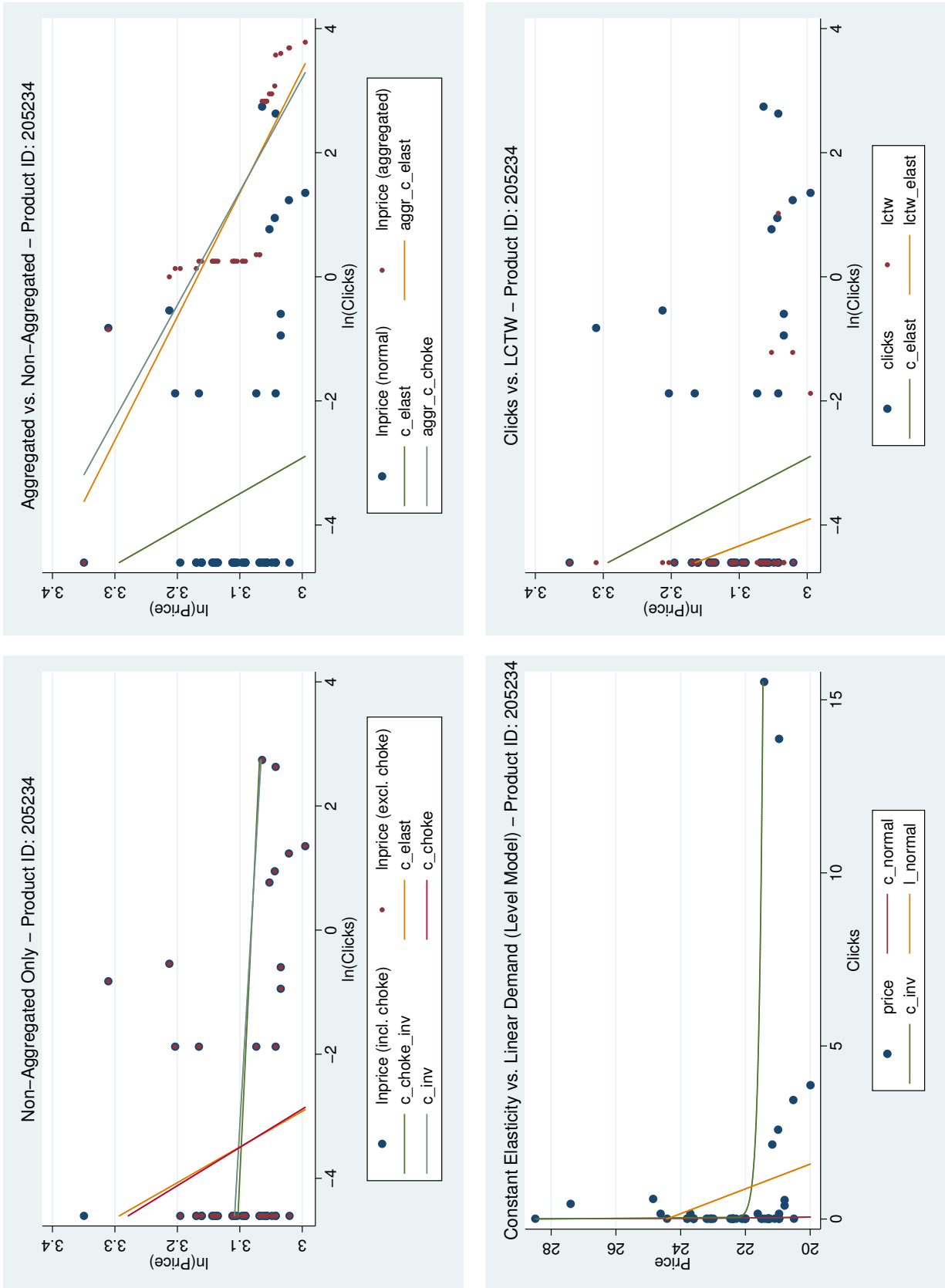


Figure 53: Product: 205234, Hähnel HL-2LHP Li-Ionen-Akku (1000 188.2): VideoFotoTV \Rightarrow FotoVideozubehr

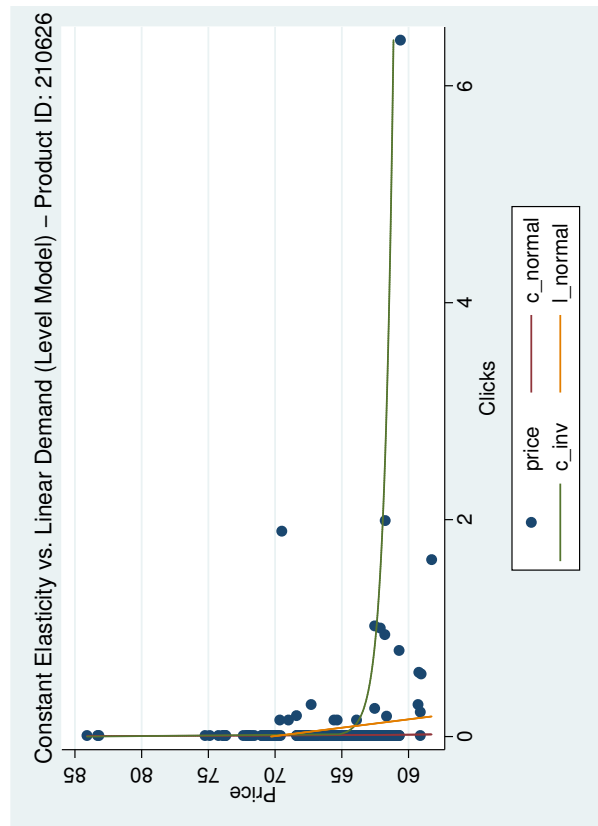
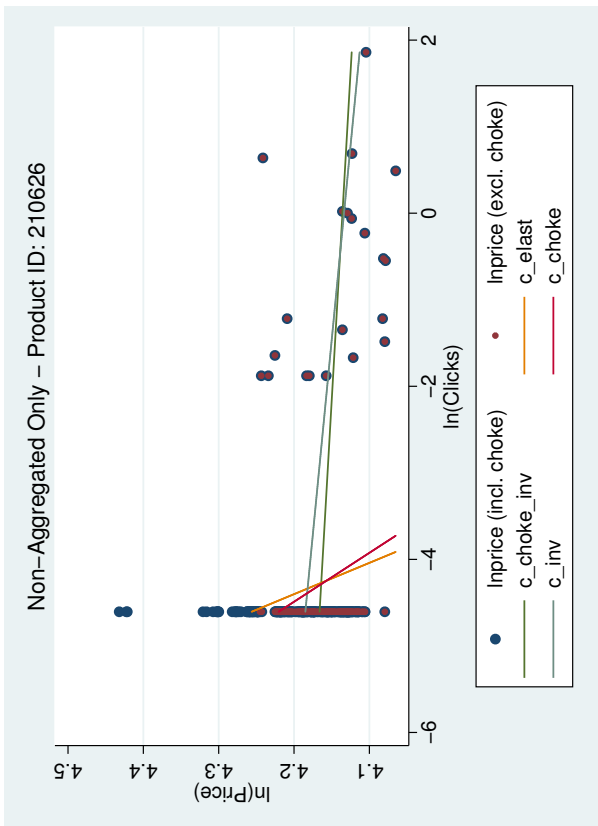
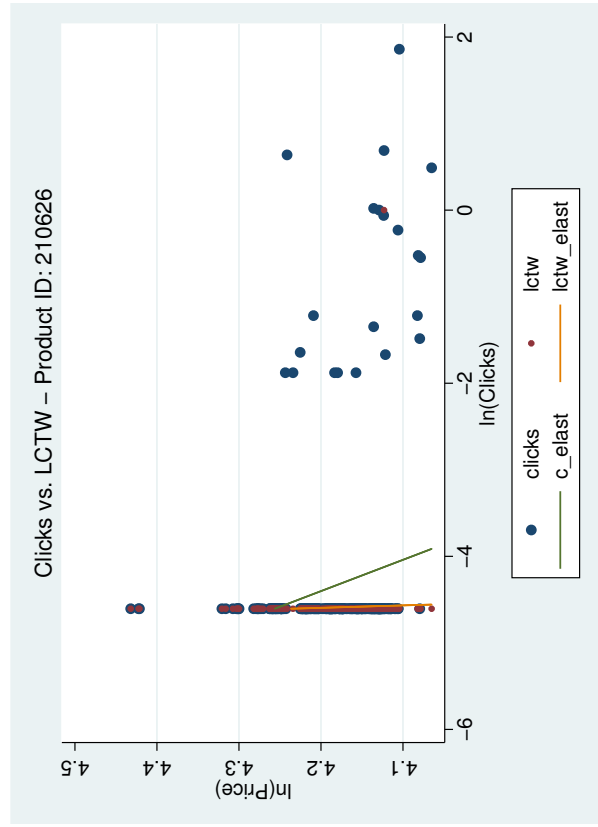
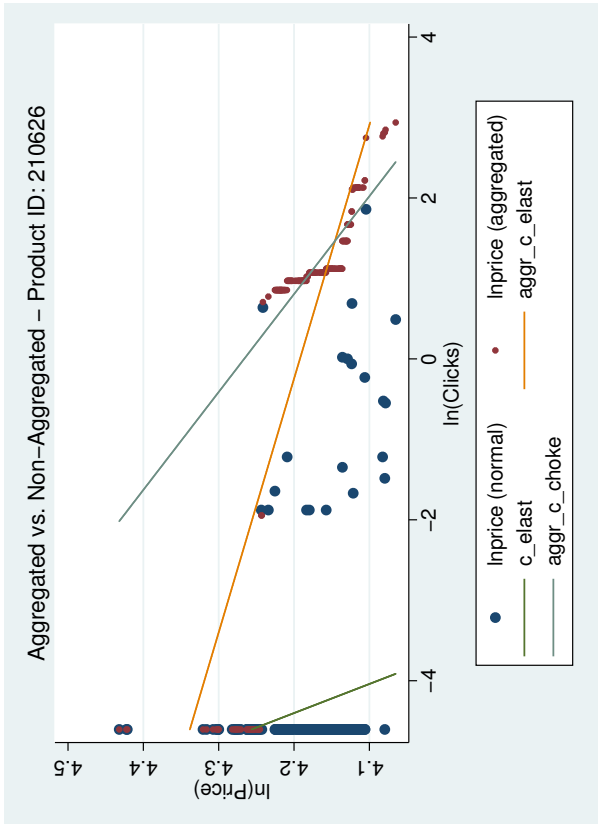


Figure 54: Product: 210626, Gigabyte GA-945GM-S2, 1945G (dual PC2-5300U DDR2): Hardware ⇒ Mainboards

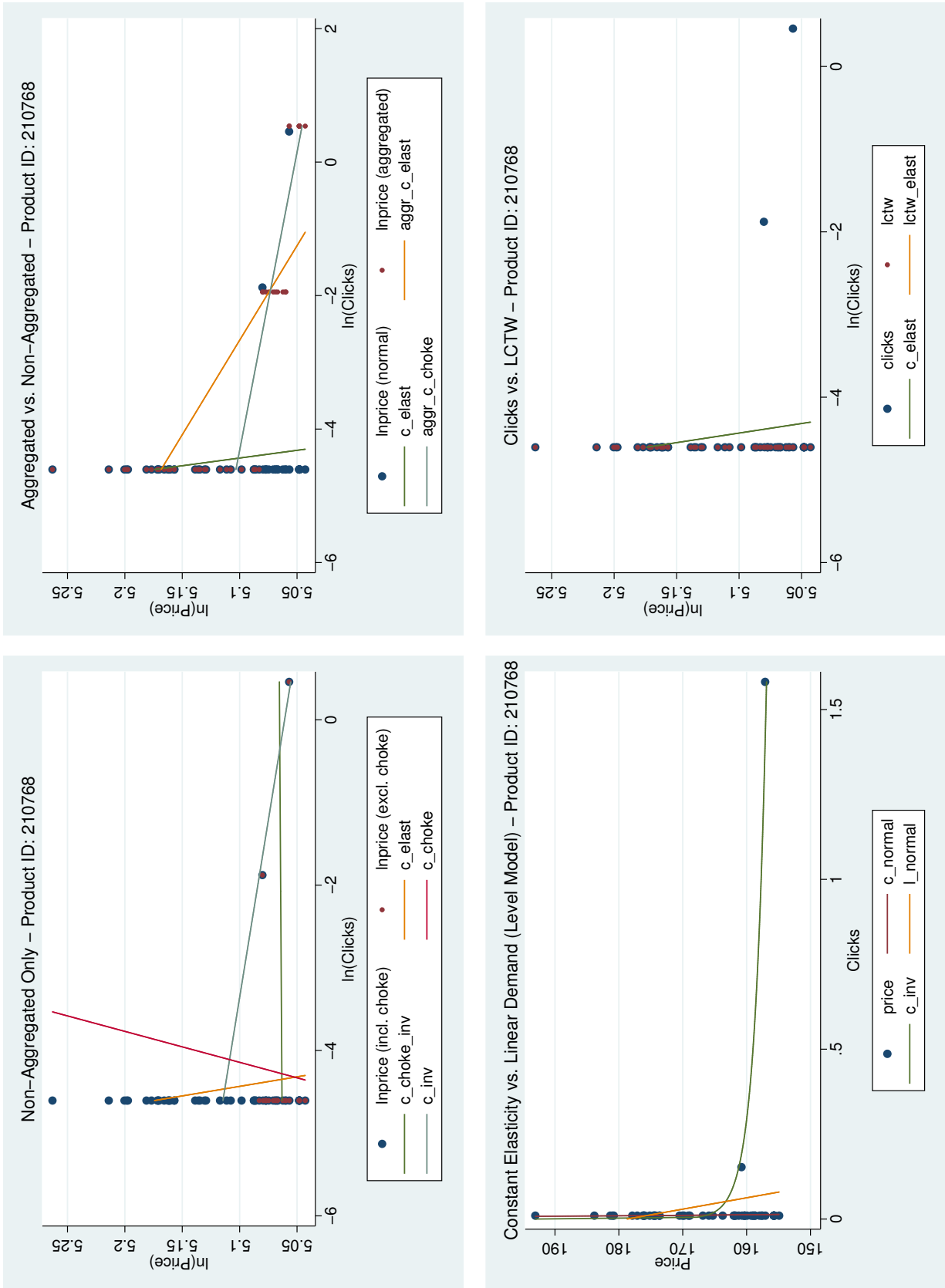


Figure 55: Product: 210768, Acer LC.08001.001 8oGB HDD: Hardware \Rightarrow Notebookzubehr

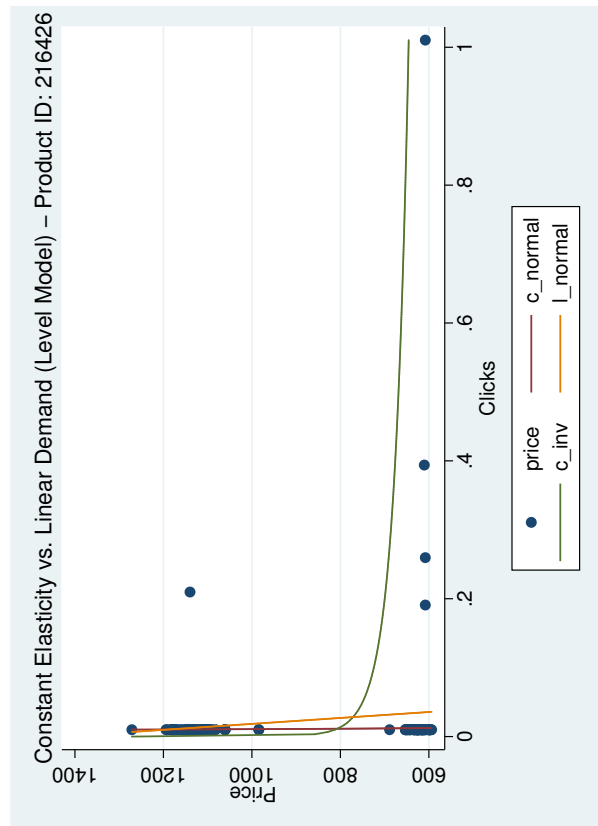
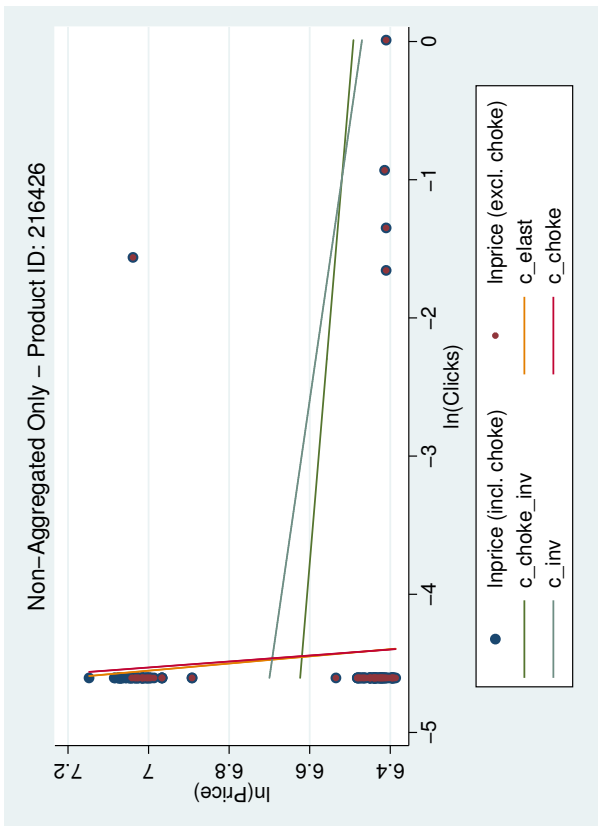
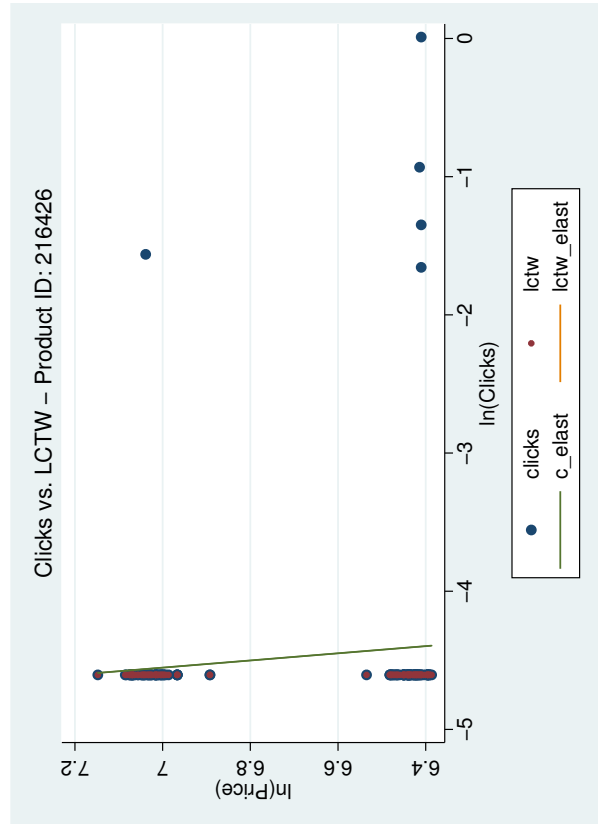
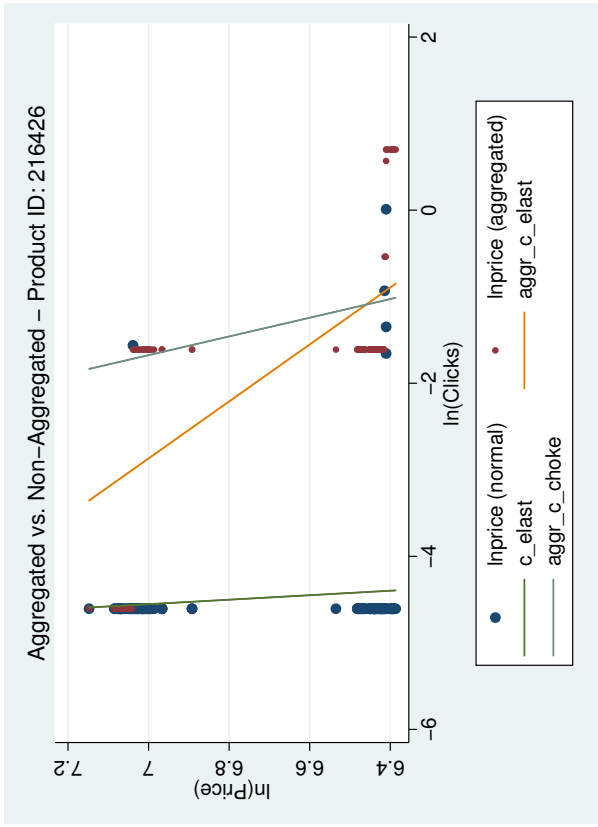


Figure 56: Product: 216426, Samsung ML-4551ND: Hardware ⇒ DruckerScanner

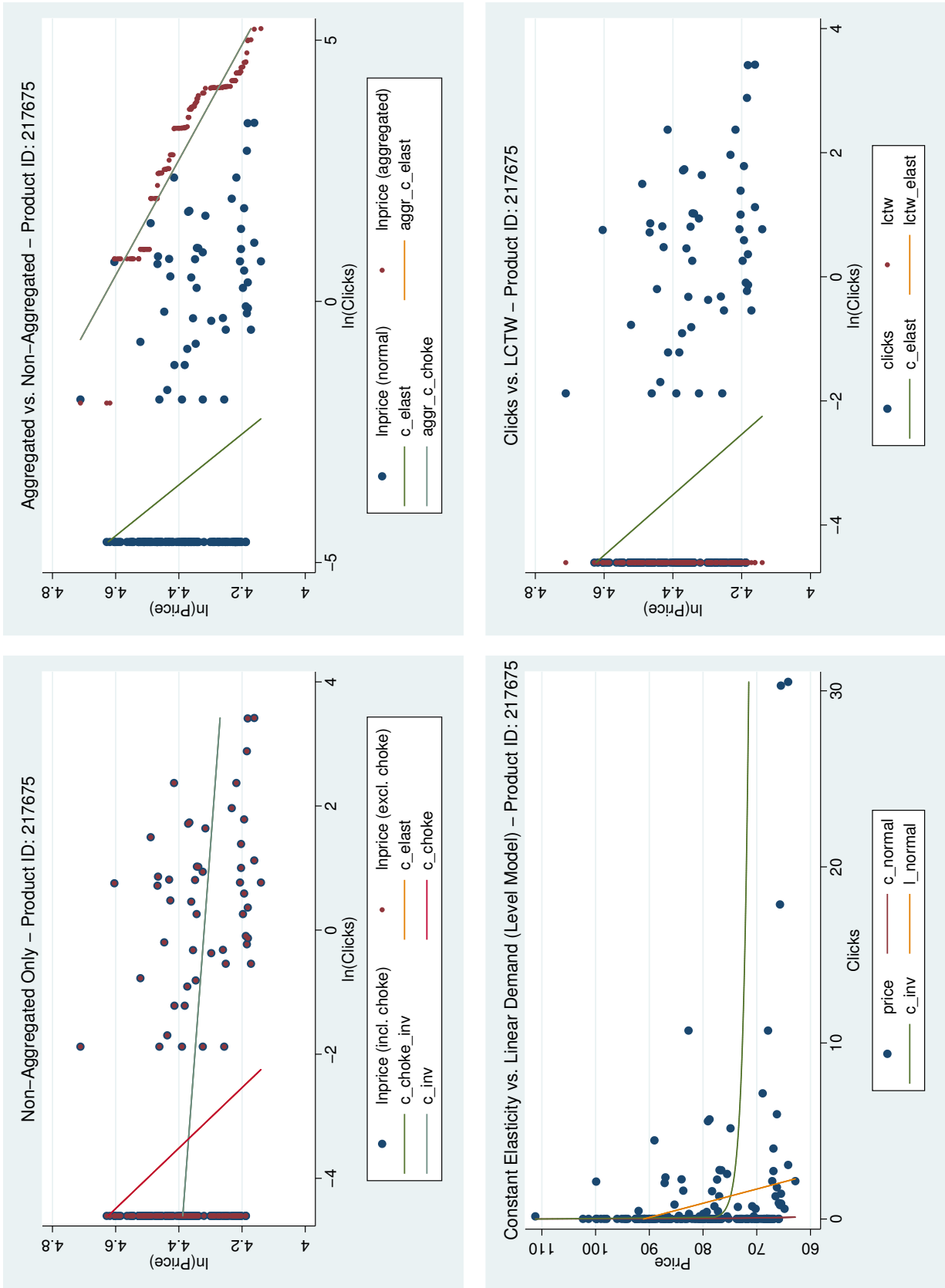


Figure 57: Product: 217675, Adobe: Photoshop Elements 5.0 (deutsch) (PC) (29230441): =>

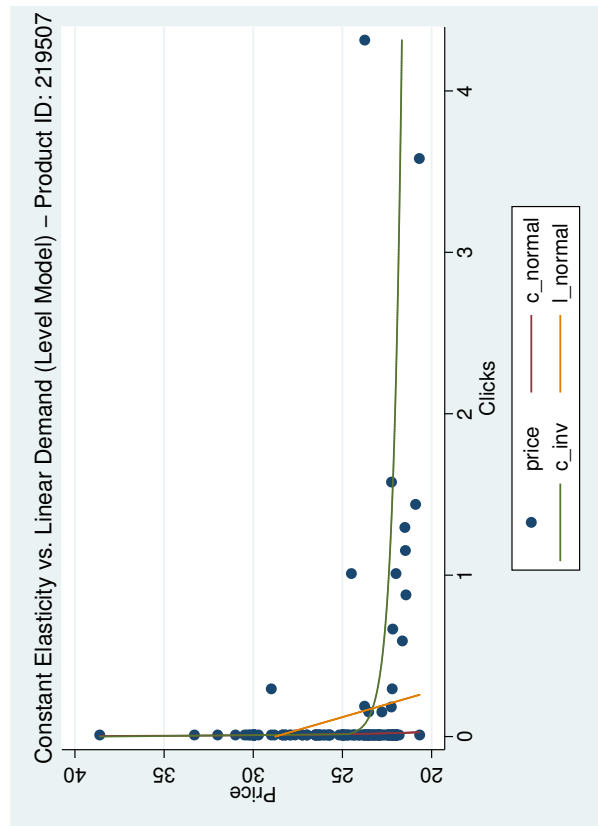
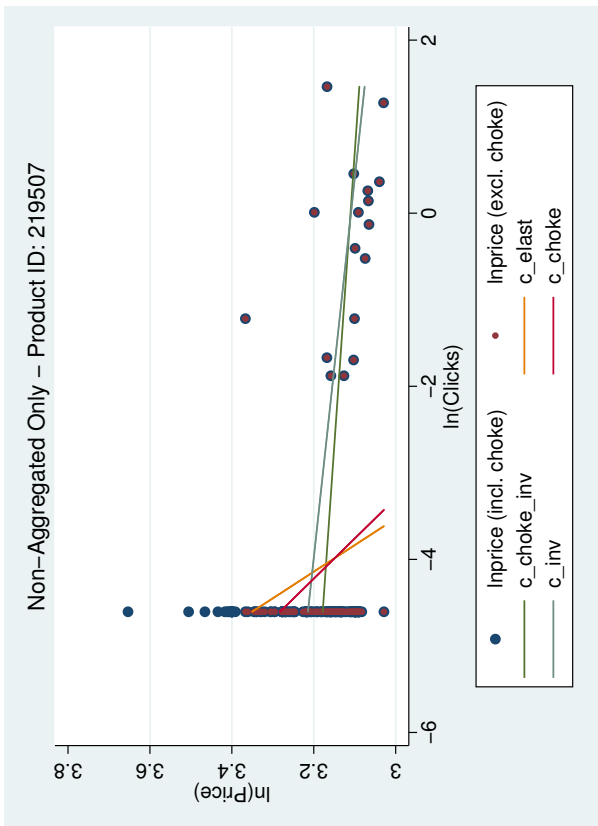
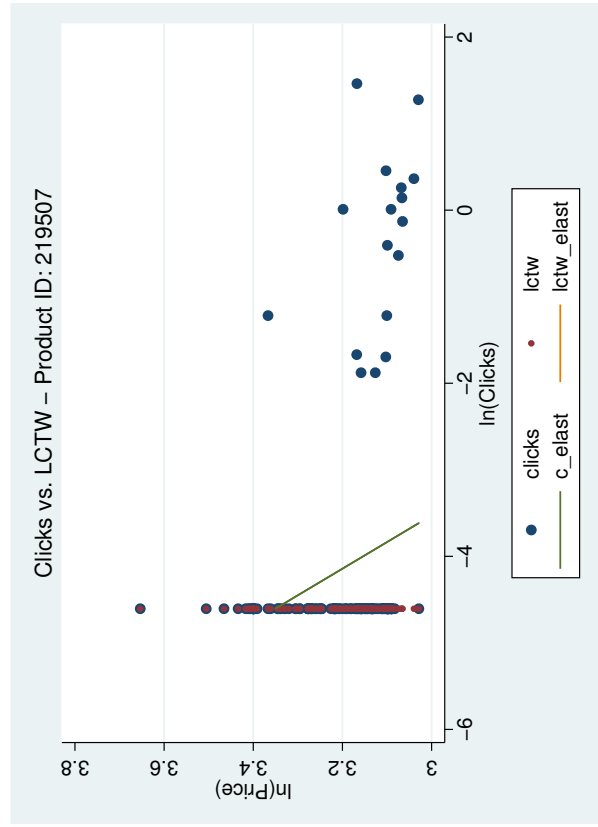
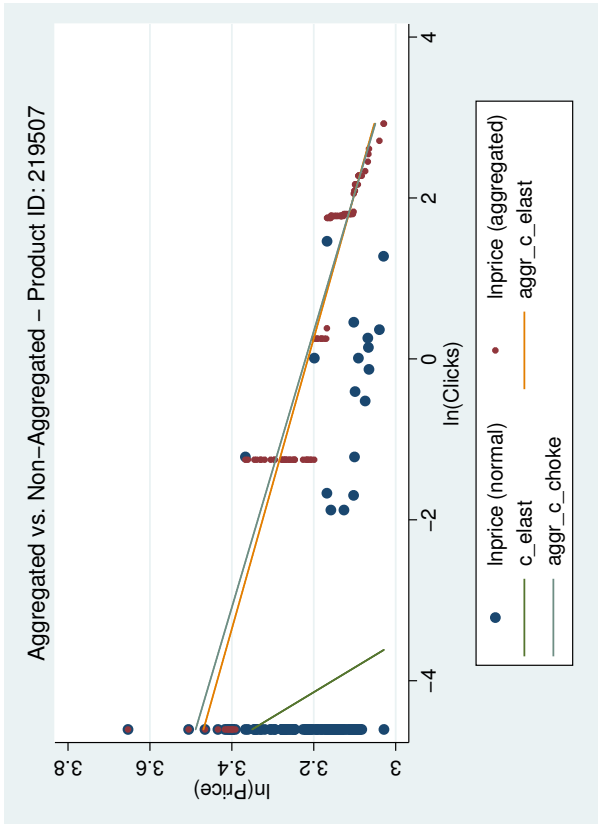


Figure 58: Product: 219507, Canon BCI-6 Multipack Color (S800/820D/BJC 8200/9000/ig100/19950) (4706A022): Hardware ⇒ Verbrauchsmaterial

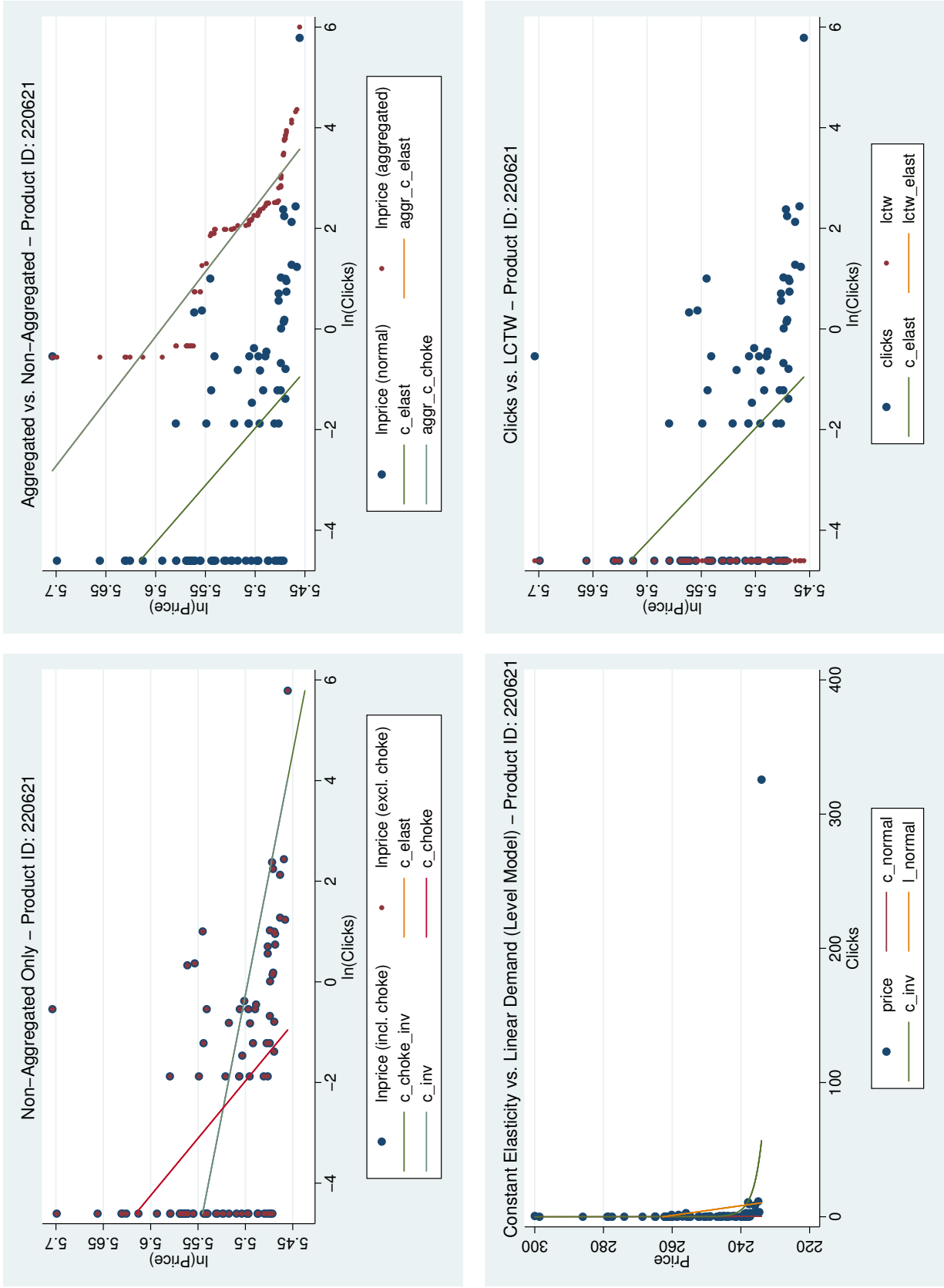


Figure 59: Product: 220621, V7 Videoseven L22WD, 22", 1680x1050, analog/digital, Audio: Hardware => Monitore

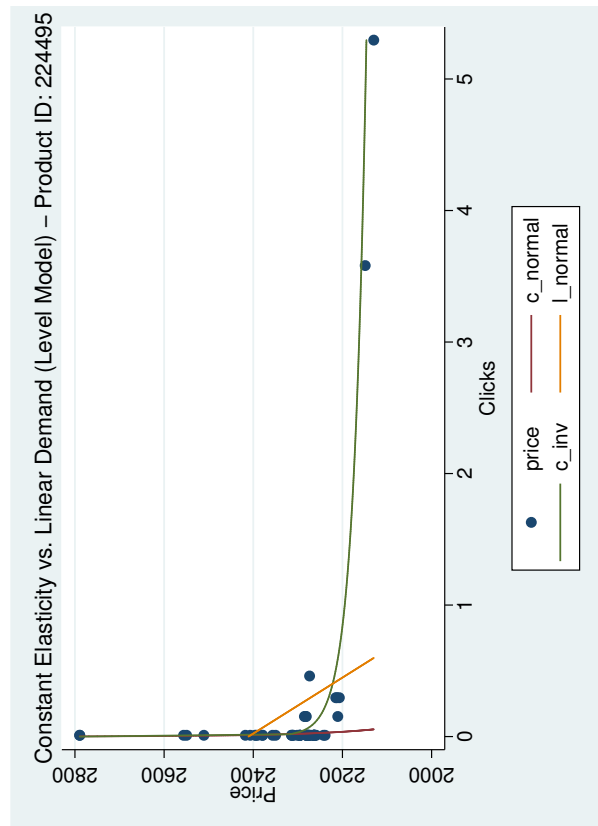
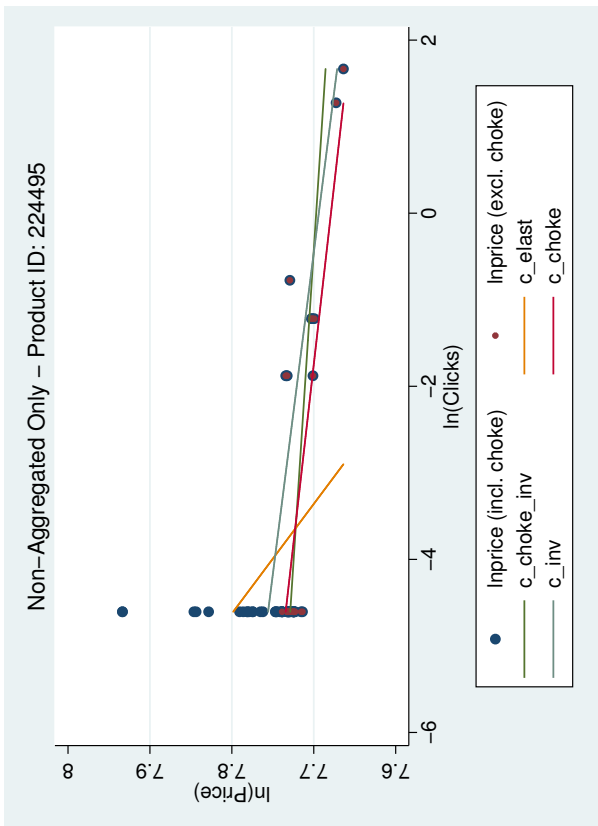
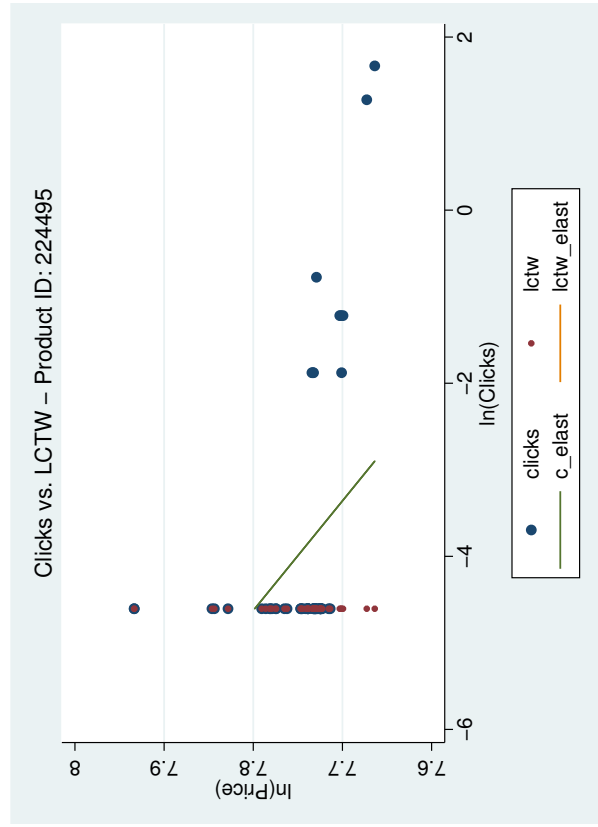
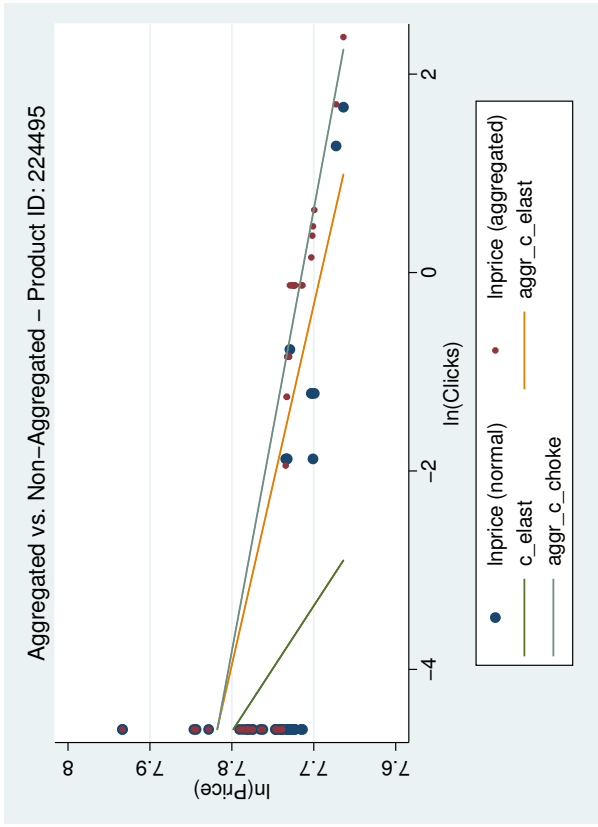


Figure 60: Product: 224495, Apple MacBook Pro Core 2 Duo, 15.4", T7400 2x 2.16GHz, 2048MB, 160GB: Hardware ⇒ Notebooks

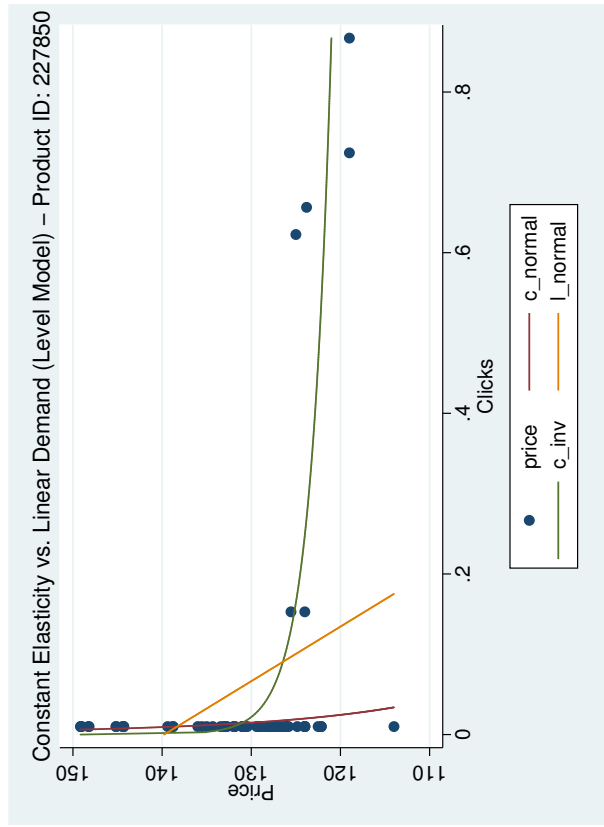
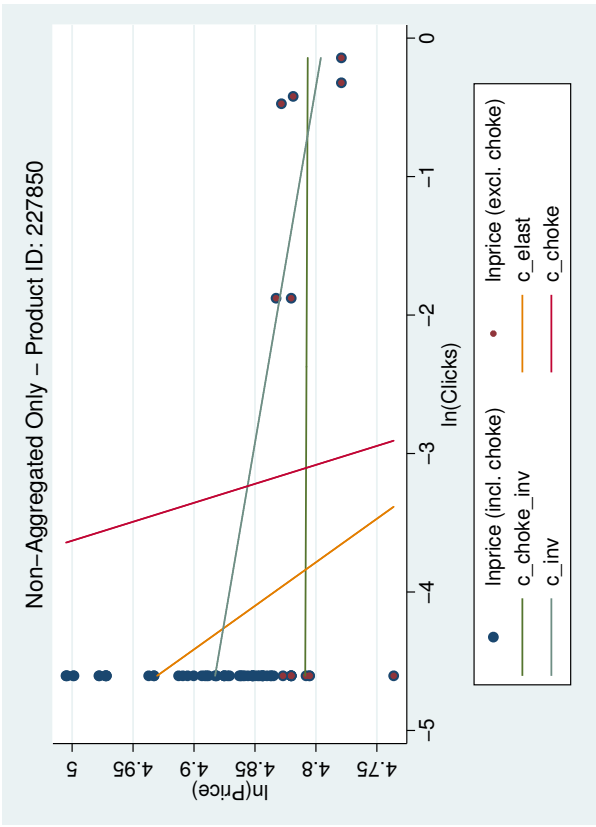
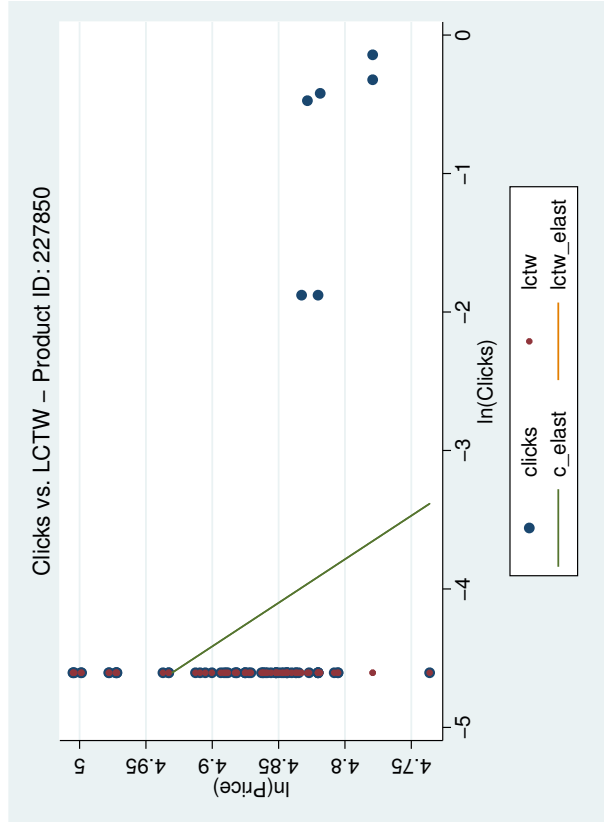
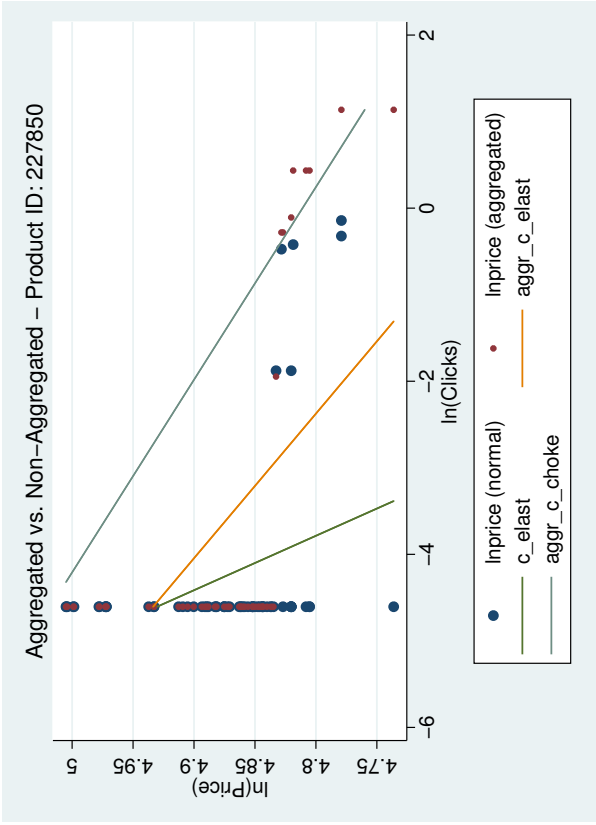


Figure 61: Product: 227850, Sony Walkman NW-S703FV 1GB violet: AudioHIFI => PortableAudio

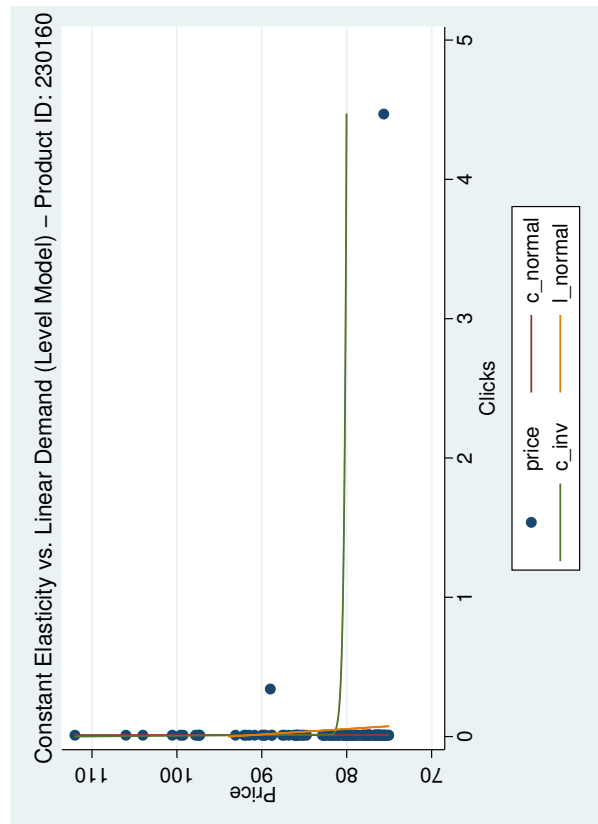
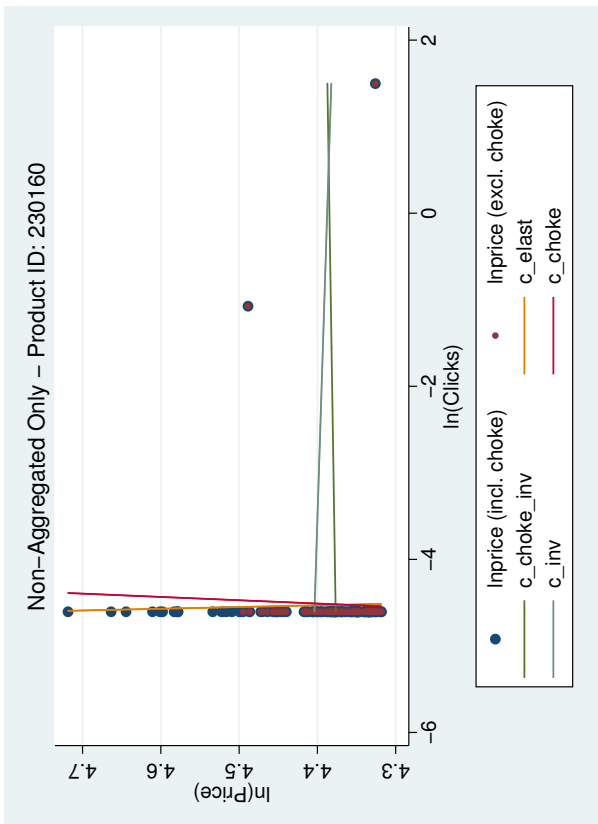
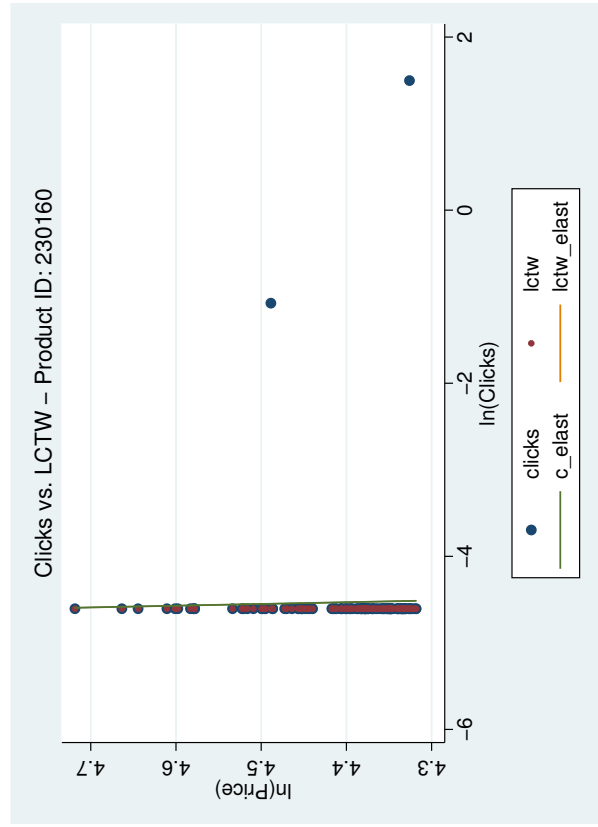
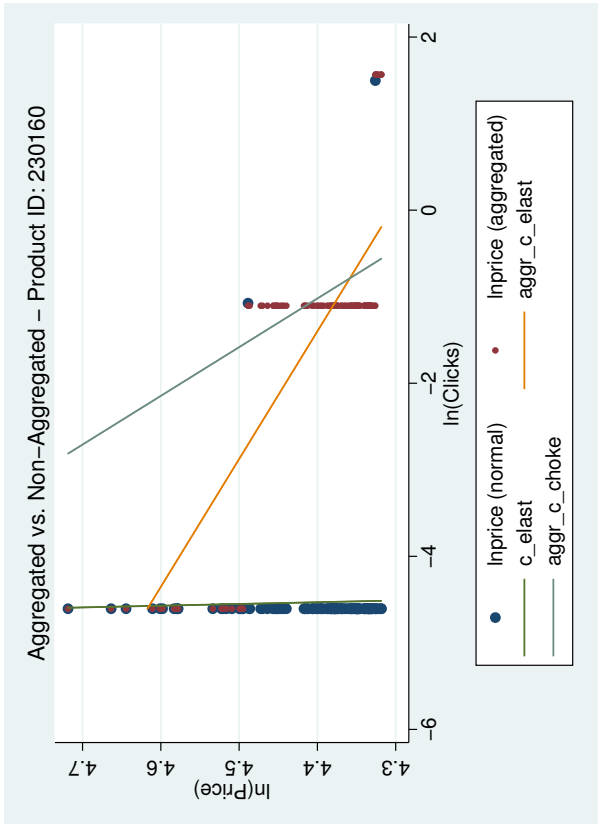


Figure 62: Product: 230160, Seagate SV35 320GB (ST3320620AV): Hardware ⇒ Festplatten

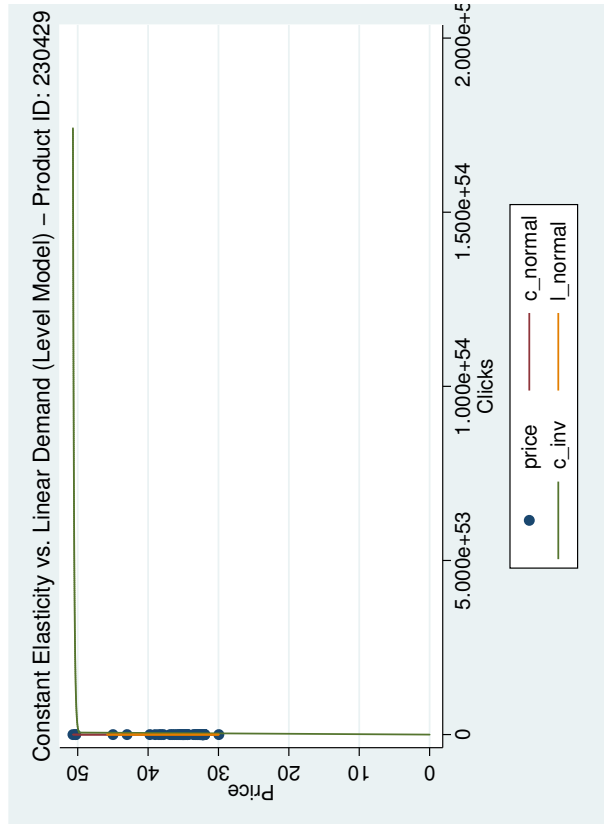
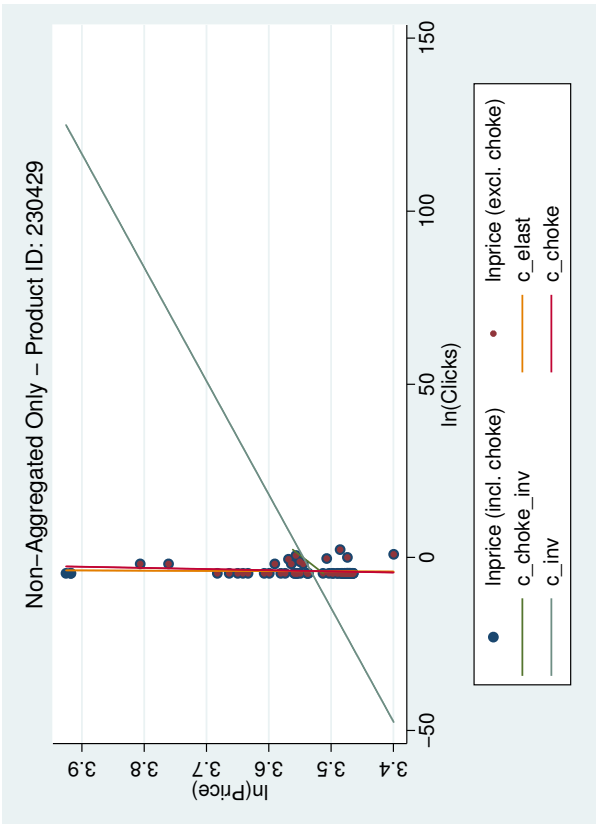
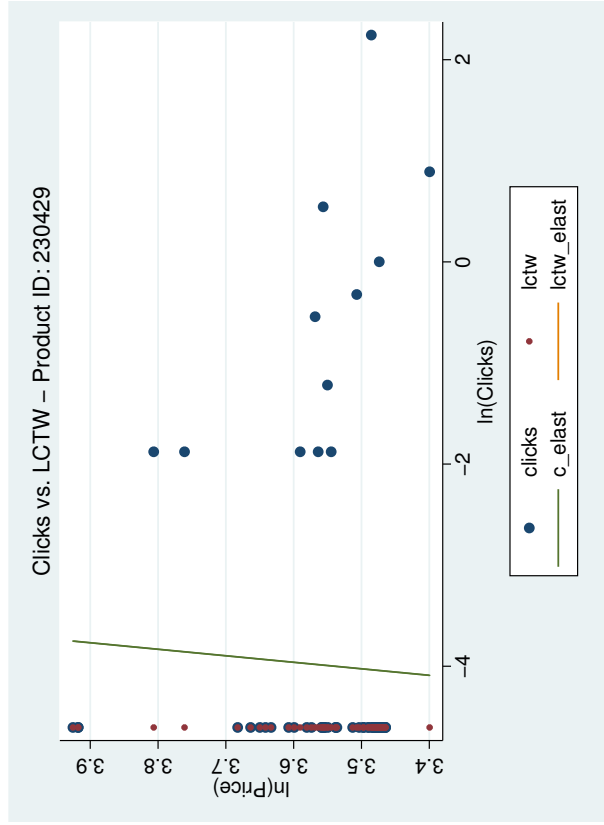
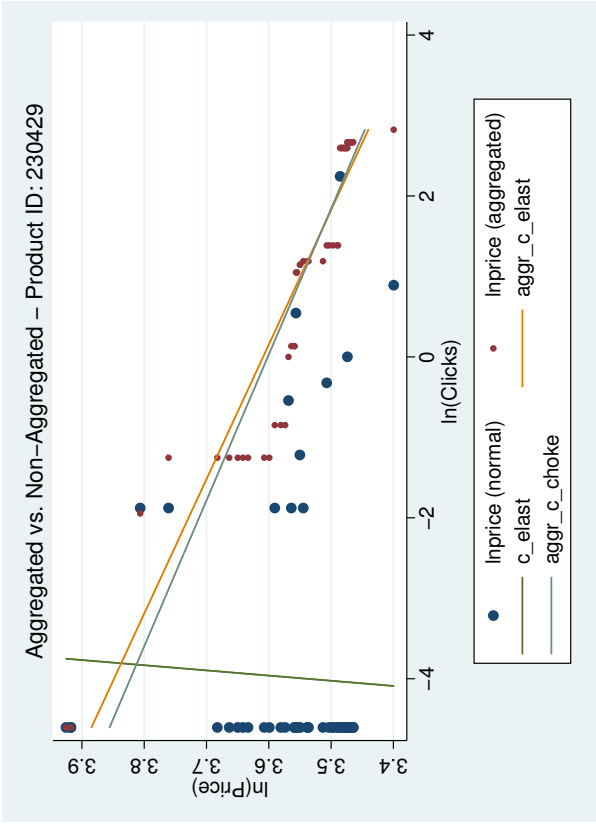


Figure 63: Product: 230429, Trust WB-5400 4 Megapixel Webcam USB 2.0 (15007): Hardware \Rightarrow PCVideo

APPENDIX - LISTINGS

C.1 CREATION OF PRODUCT-SPECIFIC DATA FOR THE SECOND STAGE REGRESSIONS

The PHP-script described in this section has been used to generate the product-specific variables for the second stage regression. The variables generated in this script are used as *RHS*-variables in order to explain differences in the price elasticity of demand between products.

Listing 2 shows the code used to connect to the Geizhals database. This code is used in every script that uses any sort of database connectivity and will therefore be explained only once.

```

1 // MySQL database service
2 // Falls das PHP auf der Origin gestartet wird, müsste der Port
   42000 anstatt 42001 sein
3 $server = "127.0.0.1:47000";
4
5 // database instance (schema)
6 $database="geizhals_k122";
7
8 // user credentials
9 $user="*****";
10 $password="*****";
11
12 mysql_connect($server,$user,$password);
13
14 @mysql_select_db($database) or die( "Unable to select database");

```

Listing 2: Creation of destination table

The next part of the script gathers the list of products and creates a table in the Geizhals database which will be used to store the generated data.

```

1 $start = 1180051200; // Fri, 25.05.2007
2 $end = 1180656000; // Fir, 01.06.2007
3
4 // Open log files
5 $prodsq = fopen("create_product.sql","w+");
6 $prodlog = fopen("create_product.log","w+");
7
8 $command = "";
9 if($argc > 1) $command = $argv[1];
10
11 // Get full list of subsubcategories
12 $sql_prd = "select distinct produkt_id from gz_kaufclick_250607
   where dtime between $start and $end";
13 $res_prd = mysql_query($sql_prd);
14 $num_prd = mysql_num_rows($res_prd);
15
16 $prods = array();
17 for($i=0;$i<$num_prd;$i++) {
18     $row = mysql_fetch_array($res_prd);
19     array_push($prods,$row[0]);

```

```

20 }
21
22 // Create Table for Product Data
23 printlog($prodlog,"Deleteing old mh_product_250409 ...");
24 $sql_tbl = "drop table if exists mh_product_250409";
25 if(mysql_query($sql_tbl))
26     printlog($prodlog,"Ok\n");
27 else
28     printlog($prodlog,"Error\n");
29
30 printlog($prodlog,"Creating new mh_product table ... ");
31 $sql_tbl =
32 "create table mh_product_250409(
33     produkt_id        int(6) primary key,
34     subsubkat          varchar(45),
35     prod_avgquali     decimal(7,5),
36     prod_avgsupport   decimal(7,5),
37     prod_avgfeature   decimal(7,5),
38     prod_avgvalue     decimal(7,5),
39     prod_quality      decimal(7,5),
40     prod_recommendation decimal(6,5),
41     prod_avgmissing   tinyint(1),
42     prod_numclicks    int(7),
43     prod_numretailer  int(6),
44     prod_avgprice     decimal(17,10),
45     prod_numratings   int(7),
46     prod_marke        varchar(255),
47     brand1to10        tinyint(1) default 0,
48     brand11to20       tinyint(1) default 0,
49     brand21to30       tinyint(1) default 0,
50     brand31to40       tinyint(1) default 0,
51     brand41to50       tinyint(1) default 0,
52     brand51to70       tinyint(1) default 0,
53     brand71to100     tinyint(1) default 0,
54     brand101toEnd    tinyint(1) default 0,
55     nobrandatall     tinyint(1) default 0
56 );
57 if(mysql_query($sql_tbl))
58     printlog($prodlog,"Ok\n");
59 else
60     printlog($prodlog,"Error\n");

```

Listing 3: Database connection

The next part in the script is the main body. The loop iterates through the complete list of products. For each product a set of functions which compute the required data is executed. Furthermore the computed product data is used to create an SQL insert statement and the dataset is inserted into the database. Finally the script checks whether there is any data on the quality criteria missing. If this is the case the script computes average values and updates the inserted records in the table.

```

1 $updatereq=0;
2 $i = 0;
3 foreach($prods as $prod)
4 {
5     $i++;
6     printlog($prodlog, "Processing $prod\n");
7

```

C.1 Creation Of Product-Specific Data For The Second Stage Regressions

```

 8  printlog($prodlog, "\t ... getting subsubkat");
 9  // Gather kat and subkat
10  $subsubkat = get_prod_ssk($prod);
11  // Get string for subcat and cat dummies
12  printlog($prodlog, "\t ... Ok\n");
13
14
15  // Gather the number of clicks the product received during the
    period of observation
16  printlog($prodlog, "\t ... generating prod_numclicks");
17  $prod_numclicks = get_prod_numclicks($prod, $start, $end);
18  printlog($prodlog, " ... Ok\n");
19
20  // Gather the average product ratings
21  printlog($prodlog, "\t ... generating prod_avgratings");
22  $prod_avgratings = get_prod_avgratings($prod, $end);
23  $prod_avgfeature = $prod_avgratings["avgfeature"];
24  $prod_avgquali = $prod_avgratings["avgquality"];
25  $prod_avgvalue = $prod_avgratings["avgvalue"];
26  $prod_avgsupport = $prod_avgratings["avgsupport"];
27  $prod_quality = $prod_avgratings["quality"];
28  $prod_recommendation = $prod_avgratings["recommendation"];
29  $prod_avgmissing = $prod_avgratings["avgmissing"];
30  if($prod_avgmissing == 1) $updatereq = 1;
31  printlog($prodlog, " ... Ok\n");
32
33  // Gather the number of retailers offering that specific
    product during the period of observation
34  printlog($prodlog, "\t ... generating prod_numretailer");
35  $prod_numretailer = get_prod_numretailer($prod, $start, $end);
36  printlog($prodlog, " ... Ok\n");
37
38  // Gather the average product price during the period of
    observation
39  printlog($prodlog, "\t ... generating prod_avgprice");
40  $prod_avgprice = get_prod_avgprice($prod, $start, $end);
41  printlog($prodlog, " ... Ok\n");
42
43  // Gather the brand of the product
44  printlog($prodlog, "\t ... generating prod_marke");
45  $prod_marke = get_prod_marke($prod);
46  printlog($prodlog, " ... Ok\n");
47
48  // Gather the number of ratings the product received during the
    period of observation and during 90 days before that week
49  printlog($prodlog, "\t ... generating prod_numratings");
50  $prod_numratings = get_prod_numratings($prod, $start, $end);
51  printlog($prodlog, " ... Ok\n");
52
53  // Creation of data string for insert statement
54  $data = "$prod, '$subsubkat', $prod_avgquali, $prod_avgsupport,
    $prod_avgfeature, $prod_avgvalue, $prod_quality,
    $prod_recommendation, $prod_avgmissing,
55  $prod_numclicks, $prod_numretailer, $prod_avgprice,
    $prod_numratings, '$prod_marke', 0, 0, 0, 0, 0, 0, 0, 0";
56
57  // Inserting subsubkat into database
58  printlog($prodlog, "\t ... inserting");
59  $sql_ins = "insert into mh_product_250409 values($data)";

```

```

60     fwrite($prodsqL,$sql_ins."\n");
61     if(mysql_query($sql_ins)) {
62         printlog($prodlog, " ... Ok\n");
63     } else {
64         printlog($prodlog, " ... Error\n $sql_ins \n\n");
65     }
66
67
68 }
69
70 // If avgqual is missing for an ssc calculate total avgquali and
71 // update missing values
72 if($updatereq == 1) {
73     printlog($prodlog,"Updating Missing Rating Values...");
74     $sql_avg = "select avg(prod_avgquali),
75                 avg(prod_avgsupport),
76                 avg(prod_avgvalue),
77                 avg(prod_avgfeature),
78                 avg(prod_quality),
79                 avg(prod_recommendation)
80                 from mh_product_250409 where prod_avgmissing = 0";
81     $res_avg = mysql_query($sql_avg);
82     $num_avg = mysql_num_rows($res_avg);
83     if($num_avg > 0) {
84         $qual = mysql_fetch_array($res_avg);
85         $sql_upd = "update mh_product_250409 set prod_avgquali = ".
86                 $qual[0].
87                 ",prod_avgsupport = ".$qual[1].
88                 ",prod_avgvalue = ".$qual[2].
89                 ",prod_avgfeature = ".$qual[3].
90                 ",prod_quality = ".$qual[4].
91                 ",prod_recommendation = ".$qual[5].
92                 "where prod_avgmissing = 1";
93         if(mysql_query($sql_upd)) {
94             printlog($prodlog,"avg computed ... Update Ok\n");
95         } else {
96             printlog($prodlog,"avg computed ... Update Error\n
97             $sql_upd \n\n");
98         }
99     } else {
100         printlog($prodlog,"avg not computed ... no Update\n");
101     }
102 }
103 printlog($prodlog,"Finished!");
104 fclose($prodsqL);
105 fclose($prodlog);
106 mysql_close();

```

Listing 4: Main body

The last listing in this section lists the functions needed to compute the product-specific data. The functions themselves are simple and straightforward and should therefore be self-explanatory.

```

1 function get_prod_($prod) {
2     $sql_tmp = "";
3     $res_tmp = mysql_query($sql_tmp);
4     $tmprow = mysql_fetch_array($res_tmp);
5     mysql_free_result($res_tmp);

```

```

6     return $tmprow[0];
7 }
8
9 function get_prod_marke($prod) {
10    $sql_tmp = "select name from (select * from de_produk_marke
11              where produkt = $prod) pm, de_genMarke gm where pm.genMarke
12              = gm.id";
13    $res_tmp = mysql_query($sql_tmp);
14    $tmprow = mysql_fetch_array($res_tmp);
15    mysql_free_result($res_tmp);
16    return $tmprow[0];
17 }
18
19 function get_prod_numratings($prod,$start,$end) {
20    $sql_tmp = "select count(*) from gz_produkbewertung_250607
21              where produkt_id=$prod and dtime between ($start
22              -60*60*24*90) and $end";
23    $res_tmp = mysql_query($sql_tmp);
24    $tmprow = mysql_fetch_array($res_tmp);
25    mysql_free_result($res_tmp);
26    return $tmprow[0];
27 }
28
29 function get_prod_numclicks($prod,$start,$end) {
30    $sql_tmp = "select count(*) from gz_kaufclick_250607
31              where produkt_id = $prod
32              and dtime >= $start
33              and dtime <= $end";
34    $res_tmp = mysql_query($sql_tmp);
35    $tmprow = mysql_fetch_array($res_tmp);
36    mysql_free_result($res_tmp);
37    return $tmprow[0];
38 }
39
40 function get_prod_numretailer($prod,$start,$end) {
41    $sql_tmp = "select count(distinct haendler_bez) from
42              gz_angebot_250607
43              where produkt_id = $prod
44              and dtimeBegin < $end
45              and dtimeEnd > $start";
46    $res_tmp = mysql_query($sql_tmp);
47    $tmprow = mysql_fetch_array($res_tmp);
48    mysql_free_result($res_tmp);
49    return $tmprow[0];
50 }
51
52 function get_prod_avgprice($prod,$start,$end) {
53    $sql_tmp = "select ifnull(avg(preis_avg),-1) from
54              gz_angebot_250607
55              where produkt_id = $prod
56              and dtimeBegin < $end
57              and dtimeEnd > $start";
58    $res_tmp = mysql_query($sql_tmp);
59    $tmprow = mysql_fetch_array($res_tmp);
60    mysql_free_result($res_tmp);
61    return $tmprow[0];
62 }
63
64 function get_prod_ssk($prod) {

```

```

60     $sql_tmp = "select subsubkat from gz_produk_250607 where
        produkt_id = $prod";
61     $res_tmp = mysql_query($sql_tmp);
62     $tmprow = mysql_fetch_array($res_tmp);
63     mysql_free_result($res_tmp);
64     return $tmprow[0];
65 }
66
67 function get_prod_avgratings($prod, $end) {
68     $sql_tmp = "select ifnull(avg(features),-1),
69                 ifnull(avg(quality),-1),
70                 ifnull(avg(value),-1),
71                 ifnull(avg(support),-1),
72                 ifnull(avg(empfehlung),-1)
73                 from gz_produkbewertung_250607 where produkt_id =
                    $prod and dtime < $end";
74     $res_tmp = mysql_query($sql_tmp);
75     $tmprow = mysql_fetch_array($res_tmp);
76     mysql_free_result($res_tmp);
77     $ret = array();
78     $ret["avgfeature"] = $tmprow[0];
79     $ret["avgquality"] = $tmprow[1];
80     $ret["avgvalue"] = $tmprow[2];
81     $ret["avgsupport"] = $tmprow[3];
82     $ret["quality"] = ($tmprow[0]+$tmprow[1]+$tmprow[2]+$tmprow[3])
        /4;
83     $ret["recommendation"] = $tmprow[4];
84     if($tmprow[0] == -1 || $tmprow[1] == -1 || $tmprow[2] == -1 ||
        $tmprow[3] == -1 || $tmprow[4] == -1) $ret["avgmissing"] =
        1;
85     else $ret["avgmissing"] = 0;
86     return $ret;
87 }
88
89
90
91 function printlog($log,$msg) {
92     fwrite($log,$msg);
93     echo $msg;
94 }
95
96
97 ?>

```

Listing 5: Computation of product-specific data

C.2 CREATION OF SUBSUBCATEGORY-SPECIFIC DATA

The creation of the subsubcategory-specific data works in the same way as the creation of the product-specific data. For this reason this section will only show the exact functions used to create the subsubcategory-specific data. The main purpose of this section is the documentation of the exact computation of the subsubcategory-specific data.

```

1 function get_ssk_numprods($ssc) {
2     $sql_tmp = "select count(*) from gz_produk_250607 where
        subsubkat = '$ssc'";
3     $res_tmp = mysql_query($sql_tmp);

```



```

4     $tmprow = mysql_fetch_array($res_tmp);
5     mysql_free_result($res_tmp);
6     return $tmprow[0];
7 }
8
9 function get_ssk_numclickedprods($ssc) {
10    $sql_tmp = "select count(*) from mh_tmp_clickedprods_250409
11              where subsubkat = '$ssc'";
12    $res_tmp = mysql_query($sql_tmp);
13    $tmprow = mysql_fetch_array($res_tmp);
14    mysql_free_result($res_tmp);
15    return $tmprow[0];
16 }
17
18 function get_ssk_numofferedprods($ssc) {
19    $sql_tmp = "select count(*) from mh_tmp_offeredprods_250409
20              where subsubkat = '$ssc'";
21    $res_tmp = mysql_query($sql_tmp);
22    $tmprow = mysql_fetch_array($res_tmp);
23    mysql_free_result($res_tmp);
24    return $tmprow[0];
25 }
26
27 function get_ssk_numtotclicks($ssc) {
28    $sql_tmp = "select ifnull(sum(prod_numclicks),0) from
29              mh_product_250409 where subsubkat = '$ssc'";
30    $res_tmp = mysql_query($sql_tmp);
31    $tmprow = mysql_fetch_array($res_tmp);
32    mysql_free_result($res_tmp);
33    return $tmprow[0];
34 }
35
36 function get_ssk_numretailer($ssc) {
37    $sql_tmp = "select count(distinct haendler_bez) from
38              mh_tmp_angebot_250409 a ".
39              "where a.produkt_id in (select produkt_id from
40              gz_produk_250607 where subsubkat = '$ssc') ";
41    $res_tmp = mysql_query($sql_tmp);
42    $tmprow = mysql_fetch_array($res_tmp);
43    mysql_free_result($res_tmp);
44    return $tmprow[0];
45 }
46
47 function get_ssk_numratings($ssc,$start) {
48    $sql_tmp = "select count(*) from mh_tmp_produkbewertung_250409
49              b ".
50              "where dtime >= ($start-60*60*24*90) and b.
51              produkt_id in (select produkt_id from
52              gz_produk_250607 where subsubkat = '$ssc') ";
53    $res_tmp = mysql_query($sql_tmp);
54    $tmprow = mysql_fetch_array($res_tmp);
55    mysql_free_result($res_tmp);
56    return $tmprow[0];
57 }
58
59 function get_ssk_qualityvars($ssc) {
60    $sql_tmp = "select ifnull(avg(quality),-1),
61              ifnull(avg(features),-1),
62              ifnull(avg(support),-1),

```

```

55         ifnull(avg(value),-1),
56         ifnull(avg(empfehlung),-1),
57         count(*)
58         from mh_tmp_produktbewertung_250409 b
59         where b.produkt_id in (select produkt_id
                                from gz_produk250607 where
                                subsubkat = '$ssc' );
60     $res_tmp = mysql_query($sql_tmp);
61     $tmprow = mysql_fetch_array($res_tmp);
62     $ret = array();
63     $ret["quality"] = ($tmprow[0]+$tmprow[1]+$tmprow[2]+$tmprow[3])
64                       /4;
65     $ret["recommendation"] = $tmprow[4];
66     $ret["qualisamplesize"] = $tmprow[5];
67     mysql_free_result($res_tmp);
68     return $ret;
69 }
70 function printlog($log,$msg,$tofile = false) {
71     if($tofile) fwrite($log,$msg);
72     echo $msg;
73 }
74
75 function get_parent_cats($ssc) {
76     $sql_tmp = "select sk.subkat,k.kat
77               from gz_subsubkategorie_250607 ssk,
78               gz_subkategorie_250607 sk,
79               gz_kategorie_250607 k
80               where ssk.subkat = sk.id
81               and sk.kat = k.id
82               and ssk.subsubkat = '$ssc'";
83     $res_tmp = mysql_query($sql_tmp);
84     $tmprow = mysql_fetch_array($res_tmp);
85     $ret = array();
86     if($tmprow) {
87         $ret["subkat"] = ereg_replace("[^A-Za-z]", "", $tmprow[0]);
88         $ret["kat"] = ereg_replace("[^A-Za-z]", "", $tmprow[1]);
89     } else {
90         $ret["subkat"] = "";
91         $ret["kat"] = "";
92     }
93     mysql_free_result($res_tmp);
94     return $ret;
95 }
96
97 function get_dummy_string($sscinfo, $subkats, $kats) {
98     $missing=1;
99     $retarr = array();
100
101     foreach($subkats as $sk) {
102         if($sscinfo["subkat"]==$sk) {
103             array_push($retarr,1);
104             $missing = 0;
105         } else
106             array_push($retarr,0);
107     }
108
109     foreach($kats as $kat) {
110         if($sscinfo["kat"]==$kat) {

```

```

111     array_push($retarr,1);
112     $missing = 0;
113     } else
114     array_push($retarr,0);
115     }
116     return implode(",",$retarr) . ",$missing";
117 }

```

Listing 6: Computation of subcategory-specific data

C.3 OUTLIER DETECTION WITH GRUBBS' TEST

The listing shows the script used to execute Grubbs' test. The first part of this script is the connection to the database which can be seen in listing 2 and which is therefore omitted in this listing.

```

1
2 <?php
3 $tbl = "mh_elasticities";
4
5 $log = fopen("grubbs_250409.log","w+");
6
7 testTable($tbl,false);
8 testTable($tbl,true);
9
10
11 function testTable($tbl,$filter) {
12     grubbs($tbl,"clicks_c",$filter);
13     grubbs($tbl,"clicks_l",$filter);
14     grubbs($tbl,"lctw_c",$filter);
15     grubbs($tbl,"lctw_l",$filter);
16     grubbs($tbl,"aggr_clicks_c",$filter);
17     grubbs($tbl,"aggr_clicks_l",$filter);
18     grubbs($tbl,"aggr_lctw_c",$filter);
19     grubbs($tbl,"aggr_lctw_l",$filter);
20 }
21
22 function grubbs($tbl,$type,$filter) {
23     global $log;
24
25     fwrite($log, "Table $tbl type=$type filter=".(($filter?"true":"false")."... ");
26     $error = fopen("grubbserror.log","w+");
27     if($filter == false) {
28         $sql_elast = "select ".$type."_elast from $tbl where ".
29             $type."_elast < 0 and ".$type."_rmse != 0";
30     } else {
31         $sql_elast = "select ".$type."_elast from $tbl where ".
32             $type."_elast < 0 and ".$type."_rmse != 0 and ".$type."_numobs > 30";
33     }
34     $res_elast = mysql_query($sql_elast);
35     $num_elast = mysql_num_rows($res_elast);
36
37     echo "\nTable $tbl type=$type filter=".(($filter?"true":"false").
38         ".\n";
39     echo "\taching elast ...";
40     // Load elasticities

```

```

38     $arrElast = array();
39
40     for($i=0;$i<$num_elast;$i++) {
41         $row = mysql_fetch_array($res_elast);
42         $arrElast[count($arrElast)] = $row[0];
43     }
44 }
45
46 echo "Done\n\tcalculating outliers ... ";
47 $outlier = outlierElast($arrElast, 0.05,$filter);
48 echo "Done\n\twriting outlier to file ... ";
49 $pstderr = 0;
50 for($i=0;$i<count($outlier);$i++){
51     if($filter == false)
52         $sql_upd = "update $tbl set " . $type . "_elast_outlier
                    = " . ($i+1) . " where " . $type . "_elast = " .
                    $outlier[$i];
53     else
54         $sql_upd = "update $tbl set " . $type . "
                    _elast_outlier_prefiltered = " . ($i+1) . " where "
                    . $type . "_elast = " . $outlier[$i];
55     if(!mysql_query($sql_upd)) {
56         $pstderr++;
57         fwrite($error,"$sql_upd\n");
58     }
59 }
60 echo "Done\n";
61 fwrite($log, "Done\n");
62 }
63
64 function outlierElast ($x, $perror, $filtered)
65 {
66     $log;
67     sort($x);
68     $n = count($x);
69
70     $result = array();
71     for($i = $n-1; $i>=0; $i--) {
72         if(count($x) < 3) return $result;
73         $mean = stat_mean($x);
74         if( ($x[count($x)-1]-$mean) > ($mean-$x[0]) && $filtered ==
            false)
75             $ix = count($x)-1;
76         else
77             $ix = 0;
78         $o = $x[$ix];
79         $g = abs($o - $mean)/stat_stdev($x);
80         $pval = 1 - stat_pgrubbs($g, $n);
81         if($filtered == false) {
82             $pval = $pval * 2;
83             if($pval > 1) $pval = 2 - $pval;
84         }
85         if($pval > $perror) return $result;
86         //echo "found outlier: $o\n";
87         $result[count($result)] = $o;
88         array_splice($x, $ix, 1);
89     }
90     return $result;
91 }

```

```

92
93 function get_outlier_click ($x, $perror, $min_intervall)
94 {
95
96     sort($x);
97     $n = count($x);
98
99     //$o = $x[$n-1];
100    //$d = array_slice($x, 0, $n-1);
101    //$u = stat_var($d)/stat_var($x) * ($n - 2)/($n - 1);
102
103
104    $result = array();
105    for($i = $n-1; $i>=0; $i--) {
106        if(count($x) < 3) return $result;
107        $o = $x[count($x)-1];
108        $g = abs($o - stat_mean($x))/stat_stdev($x);
109        $pval = 1 - stat_pgrubbs($g, $n);
110        if($pval > $perror || $o < $min_intervall) return $result;
111        $result[count($result)] = $o;
112        $x = array_slice($x, 0, count($x)-1);
113    }
114    return $result;
115 }
116
117 function stat_pgrubbs($p, $n) {
118     $s = (pow($p,2) * $n * (2 - $n))/(pow($p,2) * $n - pow(($n - 1)
119         ,2));
120     $t = sqrt($s);
121     if ($t==null) {
122         $res = 0;
123     }
124     else {
125         $res = $n * (1 - stats_cdf_t($t, $n - 2,1));
126     }
127     return (1 - $res);
128 }
129
130
131 function stat_mean ($data) {
132     // calculates mean
133     return (array_sum($data) / count($data));
134 }
135
136 function stat_median ($data) {
137     // calculates median
138     sort ($data);
139     $elements = count ($data);
140     if (($elements % 2) == 0) {
141         $i = $elements / 2;
142         return (($data[$i - 1] + $data[$i]) / 2);
143     } else {
144         $i = ($elements - 1) / 2;
145         return $data[$i];
146     }
147 }
148
149 function stat_range ($data) {

```

```
150 // calculates range
151 return (max($data) - min($data));
152 }
153
154 function stat_var ($data) {
155 // calculates sample variance
156 $n = count ($data);
157 $mean = stat_mean ($data);
158 $sum = 0;
159 foreach ($data as $element) {
160 $sum += pow (($element - $mean), 2);
161 }
162 return ($sum / ($n - 1));
163 }
164
165 function stat_varp ($data) {
166 // calculates population variance
167 $n = count ($data);
168 $mean = stat_mean ($data);
169 $sum = 0;
170 foreach ($data as $element) {
171 $sum += pow (($element - $mean), 2);
172 }
173 return ($sum / $n);
174 }
175
176 function stat_stdev ($data) {
177 // calculates sample standard deviation
178 return sqrt (stat_var($data));
179 }
180
181 function stat_stdevp ($data) {
182 // calculates population standard deviation
183 return sqrt (stat_varp($data));
184 }
185 ?>
```

Listing 7: Outlier detection using Grubbs' test

D.1 EXPLAINING THE BRANDRANK

Brands offer an interesting factor with a possible influence on the value of the price elasticity of demand. For this reason this thesis tries to incorporate information on brands into the second stage regressions. However, this topic causes some problems which one must deal with in order to include brand-data into the regressions. First and foremost Geizhals does not display the brand of a product in an explicit data-field. Therefore one has to try to obtain the brand names from the list of products. This is possible because the product names usually contain the name of the brand, like e.g. Sapphire Radeon HD 5850 which is a graphics accelerator from the firm *Sapphire*. The extraction of the brand names is part of the diploma thesis from David Egger and will not be explained in this thesis.

In a next step the list of brand names has been used to generate the brandrank. To compute the brandrank, a script has been executed which counts the total number of clicks for a brand. The total number of clicks for a brand is the sum of all clicks on products which belong to the specific brand. The brands have then been sorted in accordance to the number of clicks received and the rank of 1 has been assigned to the brand with the highest number of clicks. The brand with the second highest number of clicks has received the rank of 2, etc. Brands with an equal amount of clicks have received the same brandrank. From a conceptual point of view the rank itself is pretty simple, the technical implementation however turns out to be rather tricky, because MySQL does not offer a `rank()` function like it is the case for Oracle¹.

The listing below shows the pseudo-code for an SQL statement which should retrieve a list of brands ordered by the number of clicks per brand.

```

1 CREATE TABLE brands(
2 select name, sum(clicks) AS 'sumclicks'
3 from
4 (select name,product, (select count(*) from clicks k where k.
   product_id = d.product_id) AS 'clicks'
5  from (select name, product_id from brandlist where name != "
   Komplettsystem" and name != "No-Name") d) mc
6 GROUP BY name
7 ORDER BY sum(clicks) DESC)

```

Listing 8: Computation of clicks for a brand

What one can notice from the pseudo-code is that the query explicitly excludes the "brands" *Komplettsystem* and *No-Name*. The extraction of brand names from product names is rather complicated and unfortunately the algorithm also yields nonsensical results. Geizhals lists products which are explicitly titled as *no-name* products. The same

¹ For a guide on how to emulate a ranking function in MySQL consult <http://www.oreillynet.com/pub/a/mysql/2007/03/01/optimize-mysql-rank-data.html?page=1> (retrieved on October 27, 2009).

applies to PCs where Geizhals features a list of complete PC systems which do not belong to a specific brand. If one leaves these records in the brandrank it will be disturbed. To prevent this the query ensures that meaningless brand names are excluded from the list of brand names.

AN ALTERNATIVE MEASURE OF BRAND STRENGTH A possible alternative approach to determine the strength of a brand would be to compute the number of clicks which the average product of a brand receives. However, it turns out that a ranking of brands according to the number of clicks on the average product favors brands with only a very small range of products. In most instances the top ranked brands in this case consist of only a single product and feature rather unknown brand names like *Sonnettech*. For this reason it seems more adequate to use the total number of received clicks as a measure to rank the brands.

D.2 THE MYSQL INFORMATION_SCHEMA

This section represents a short note on the so called `INFORMATION_SCHEMA` of MySQL. It is included into this thesis, because the schema is a useful feature which has been used in several occasions throughout the scripts used for data-preparation.

The `INFORMATION_SCHEMA` enables a user to retrieve meta data information of a MySQL database.

Meta data is data about the data, such as the name of a database or table, the data type of a column, or access privileges. Other terms that sometimes are used for this information are data dictionary and system catalog.²

Therefore the `INFORMATION_SCHEMA` is a set of tables which contains information about all databases of a MySQL installation. In fact the `INFORMATION_SCHEMA` does not contain real tables, but rather views which are therefore marked as *read-only*. From our point of view the most important views in the `INFORMATION_SCHEMA` are the `COLUMNS`-view and the `TABLES`-view.

COLUMNS-VIEW: This view contains data on the columns of the tables stored in the MySQL database. This view can be used to retrieve e.g. the name or the data types of the columns of a MySQL table. The name of a column can be retrieved by using `column_name`, the data type by `data_type`. One has to note that this view contains the column-meta-data for *all* tables of a MySQL database. Therefore if one only wants the columns of the specific table, one has to build in a `where`-clause on the field `table_name`. An example of the usage of this view is given in the script for Grubbs' test. This script has to retrieve a list of columns which contain the different elasticities for a specific elasticity table. Since all elasticity related columns in those tables contain the word "elast" one can easily use the `COLUMNS`-view of the `INFORMATION_SCHEMA` to obtain the desired results.

```
1 select column_name from INFORMATION_SCHEMA.COLUMNS
2 where table_name = "mh_elast_final"
```

² <http://dev.mysql.com/doc/refman/5.0/en/information-schema.html>, retrieved on Nov 2, 2009.


```
3 and column_name like "%elast%"
```

Listing 9: Retrieving the columns of a table with the INFORMATION_SCHEMA

Listing 9 shows how one can retrieve all columns of table `mh_elast_final` which contain "elast"³.

TABLES-VIEW: Like the name already suggests, this view contains data on the tables of a MySQL database. This view has mainly been used to retrieve a certain list of table names of the database. The field containing the table name is called `table_name`. A script might e.g. need to retrieve a list of all product-related tables. Firstly, the name of the tables start with the prefix "mh_". Secondly, they will somewhere contain the string "prod". Therefore the SQL statement to retrieve the list of products-tables would look like listing 10.

```
1 select distinct table_name from INFORMATION_SCHEMA.COLUMNS
2 where table_schema = "geizhals_k122"
3 and table_name like "mh\_%prod%"
```

Listing 10: Retrieving a list of table names with the INFORMATION_SCHEMA

Note that the "_" has to be escaped with a "\", because otherwise the MySQL pattern-matching procedure would interpret it as a wildcard character. Apart from regular expressions MySQL supports "_" and "%" as wildcard-characters. The former matches any single character, the latter matches any arbitrary number of characters, also including zero characters. A table with the name `mh_prod` would therefore also match the pattern given in listing 10.

D.3 AN ALTERNATIVE CRITERION FOR DATA QUALITY

This section shortly introduces an alternative criterion or measure for data quality. Recall that besides Grubbs' test the share of zero-click-offers and the number of observations used to estimate an elasticity were the only measurements for data quality. The latter is based on the idea that a larger number of offers (hence observations in the context of the first stage regressions) yields better elasticity estimates. The former criterion stemmed from the detailed analysis which made clear that the quality of the elasticity estimates drastically suffer from a larger number of zero-click-offers.

The approach introduced in this section uses the notion that the huge differences between `c_elast` and `c_elast_inv` are an indicator for potential problems like endogeneity and heteroskedasticity⁴. In the case of perfect data it should make no difference if one regresses clicks on price or price on clicks. The coefficient of the one regression should exactly be the inverse of the coefficient of the other regression. However, if the data suffers from impurities there will be a wedge between the

³ In MySQL the %-symbol denotes a wildcard-character which represents zero or more arbitrary symbols or characters.

⁴ Recall that `c_elast` is the elasticity estimate for an isoelastic demand which is retrieved by regressing `clicks` on `price` (both in logarithmic form). `c_elast_inv`, on the other hand, is the elasticity computed as the inverse of the coefficient of `clicks` when regressing `price` on `clicks`.

two demand curves which are built upon the estimation results. A logical conclusion of this observation would be to measure the size of this wedge and use it as a quality or filter criterion. This idea can be illustrated by looking at figures 64 and 65.

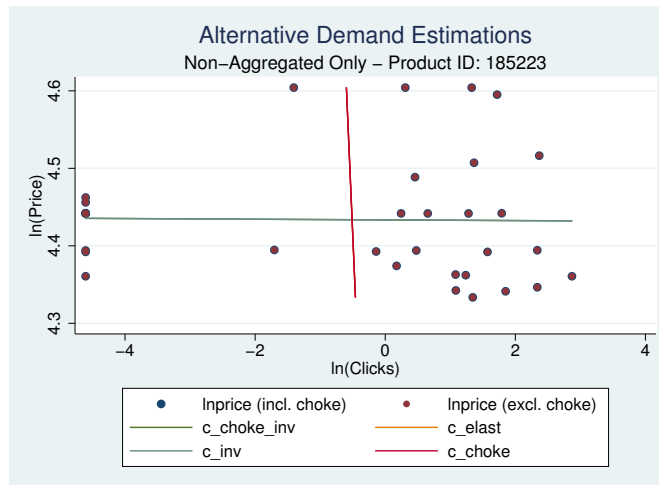


Figure 64: Large wedge due to inappropriate data

The scatterplot in figure 64 does not reveal any clear demand structure. In fact the scatterplot does not assume any specific shape at all. Furthermore it seems that the zero-click-offers do have a significant impact on the estimated curves and hence the estimated elasticities. Both the shape of the scatterplot and the zero-click-offers suggest that the demand data on this product seems to be rather poor and it will be troublesome or even unpromising to retrieve any significant and meaningful results from it. The two demand curves depicted in figure 64 confirm aforementioned misgivings. The two demand structures are completely different and hence feature a large wedge between them.

As opposed to figure 64, figure 65 shows a completely different picture. The scatterplot clearly features a negatively sloped shape. The zero-click-offers reside in the regions with a rather high price, meaning that they could be justified from a theoretical point of view. As a result, the two demand structures have a rather similar slope, yielding only a small wedge between them.

So in order to use this wedge as a quality criterion, one has to measure its size. This can easily be done by computing the angle of intersection of the estimated curves, which is depicted in figure 66. However, one has to notice that the formula presented in this section will always compute the smaller part of the intersecting angle, where *smaller* means the angle which is less than 90° . The notion of the angle of intersection is shown in figure 66, whereas α is the *small* part of the angle of intersection and β is the *large* part.

As a demand curve will have a negative slope in the vast majority of the cases to decide on the data quality one can always safely use the small part of the angle of intersection. In this setup an angle of 90° would constitute the worst case which would indicate that the data suffers from serious impurities or data problems. On the other hand an angle of 0° would be the best case, because this angle would mean that the two demand structures are identical.

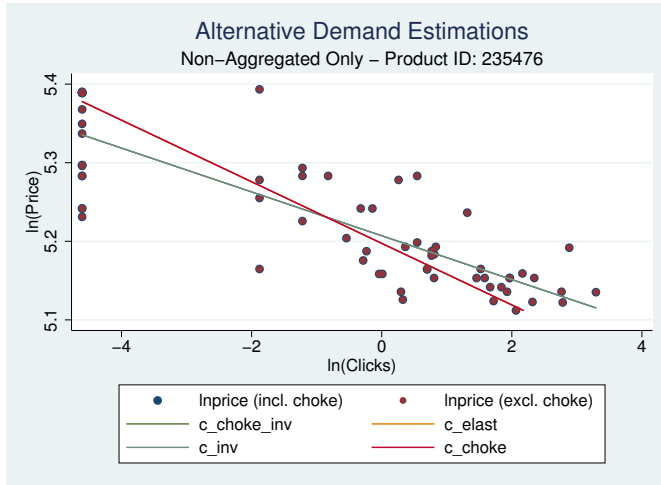


Figure 65: Acceptable data, which yields only a small wedge

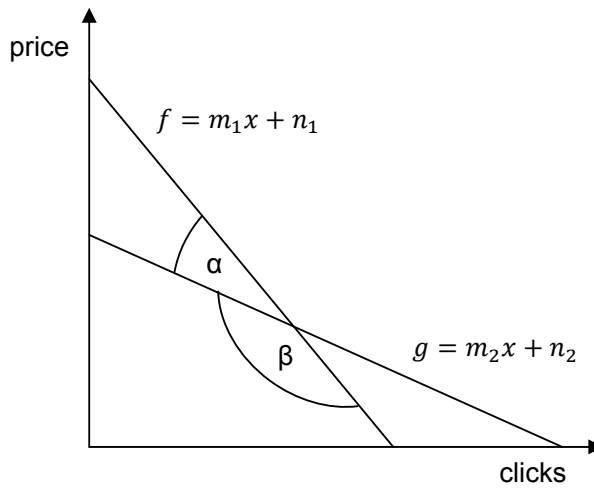


Figure 66: Angle of intersection of two lines

The remaining question is how to compute α , the angle of intersection. Given the two lines $f = m_1x + n_1$ and $g = m_2x + n_2$ from figure 66 one can simply compute the angle of intersection α by using the slope parameters m_i as given in the following equation:

$$\alpha = \arctan \left(\left| \frac{m_1 - m_2}{1 + m_1 m_2} \right| \right) \quad (D.1)$$

Before applying equation D.1, one has to check whether $1 + m_1 m_2 = 0$. In this case one cannot solve equation D.1. This, however, only happens in the case where $\alpha = 90^\circ$.

Once the angle of intersection has been computed for every product, it can be used as a filter criterion by setting a threshold. For example, one could say that one excludes all products for which the wedge is 30° or above. Furthermore one could also use the angle of intersection as a weight in the second stage regressions. Compared to \bar{R}^2 this approach offers the advantage that the angle of intersection will never be negative. Thus this implies that no observations would be dropped, as it happens in the case of a negative \bar{R}^2 . If one wants to use the angle of intersection

as a weight one would have to build its inverse, since Stata favors observations with larger weights. In order to circumvent a division by zero when building the inverse of the angle of intersection one could add a negligibly small value ϵ , like e.g. 0.000001.

BIBLIOGRAPHY

- [1] Michael R. Baye, J. Rupert J. Gatti, John Morgan, and Paul A. Kattuman. Estimating Firm-Level Demand at a Price Comparison Site: Accounting for Shoppers and the Number of Competitors. *SSRN eLibrary*, 2004. (Cited on pages 20 and 21.)
- [2] Richard E. Caves. *American Industry: Structure, Code, Performance*. Prentice-Hall, Englewood Cliffs, USA, 7th edition, 1964. (Cited on page 71.)
- [3] Alpha C. Chiang and Kevin Wainwright. *Fundamental Methods of Mathematical Economics*. McGraw-Hill, New York, NY, USA, 4th edition, 2005. (Cited on page 3.)
- [4] J. A. Cornell and R. D. Berger. Factors that influence the value of the coefficient of determination in simple linear and non-linear regression models. The American Phytopathological Society, 1986. URL http://www.apsnet.org/phyto/SEARCH/1987/Phyto77_63.asp. (Cited on page 64.)
- [5] Russel Davidson and James G. MacKinnon. *Econometric Theory and Methods*. Oxford University Press, New York, NY, USA, 1st edition, 2004. (Cited on pages 24, 52, 53, and 64.)
- [6] Frank E. Grubbs. Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21(1):27–58, March 1950. (Cited on page 30.)
- [7] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, 1st edition, September 2000. ISBN 1558604898. URL <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1558604898>. (Cited on pages 10, 28, and 29.)
- [8] A. C. Johnson Jr. and Peter Helmberger. Price elasticity of demand as an element of market structure. *The American Economic Review*, 57(5):1218–1221, 1967. ISSN 00028282. URL <http://www.jstor.org/stable/1814404>. (Cited on page 71.)
- [9] Tjalling C. Koopmans. Identification problems in economic model construction. *Econometrica*, 17(2):125–144, 1949. ISSN 00129682. URL <http://www.jstor.org/stable/1905689>. (Cited on page 12.)
- [10] Emilio Pagoulatos and Robert Sorensen. What determines the elasticity of industry demand? *International Journal of Industrial Organization*, 4(3):237–250, September 1986. URL <http://ideas.repec.org/a/eee/indorg/v4y1986i3p237-250.html>. (Cited on pages 45, 46, 47, 54, and 55.)
- [11] Hal R. Varian. *Grundzüge der Mikroökonomik*. Oldenbourg, Munich, Germany, 5th edition, 2001. (Cited on pages 1, 5, and 62.)
- [12] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, MA, USA, 1st edition, 2002. (Cited on pages 21, 78, and 79.)

BIBLIOGRAPHY

- [13] Jeffrey M. Wooldridge. *Introductory Econometrics - A Modern Approach*. South-Western, Ohio, USA, 4th edition, 2003. (Cited on pages [12](#), [13](#), [52](#), and [79](#).)

DECLARATION

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Linz, Februar 2010

Mario Hofer