5-2018

# From digital traces to marketing insights: Recovering consumer preferences for digital entertainment services and online shopping

Ai Phuong HOANG
*Singapore Management University*, aphoang.2013@phdis.smu.edu.sg

# FROM DIGITAL TRACES TO MARKETING INSIGHTS:

# RECOVERING CONSUMER PREFERENCES

# FOR DIGITAL ENTERTAINMENT SERVICES AND ONLINE SHOPPING

HOANG AI PHUONG

SINGAPORE MANAGEMENT UNIVERSITY

2018

# From Digital Traces to Marketing Insights:

# Recovering Consumer Preferences

# for Digital Entertainment Services and Online Shopping

by

**Hoang Ai Phuong**

Submitted to School of Information Systems in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Information Systems

**Dissertation Committee:**

Robert J. Kauffman (Chair)
Professor of Information Systems
Singapore Management University

Kapil R. Tuli
Professor of Marketing
Singapore Management University

Qian Tang
Assistant Professor of Information Systems
Singapore Management University

Ting Li (External Member)
Endowed Professor of Digital Business
Erasmus University Rotterdam

Singapore Management University

2018

# From Digital Traces to Marketing Insights:
# Recovering Consumer Preferences
# for Digital Entertainment Services and Online Shopping

Hoang Ai Phuong

## Abstract

IT innovations disrupt traditional business models and challenge conventional thinking. Thus, industry incumbents face fierce competition from start-ups with new business models and new ways of engaging customers. Digital entertainment goods and personalized services have become a lucrative market, which has undergone a transformation enabled by seamless Internet connections. Meanwhile, social networks and other online platforms have brought people and business even closer.

As consumer relationships with firms span geographic boundaries, so does their spending. It is no longer effective nor appropriate to segment consumers based on socioeconomic factors; instead, consumers can be characterized by their direct relationship with the goods and services, or their relationship with technology. Thus, firms need to develop innovative products and services, and adjust their marketing and delivery systems to address this new level of sophistication in consumer informedness.

There is one fundamental, yet intriguing question: How can firms recover consumer preferences in the digital space, to keep themselves and their consumers informed about the offerings that are suitable for the consumers? Against this backdrop, consumer analytics are key to a better understanding of the complex relationship between people

and technology. They also serve as the backbone of all key business decisions in this age of experience.

This dissertation examines how IT creates new capabilities for extracting business and consumer insights to inform traditional marketing activities, for physical products in the retail industry as well as on-demand services in the entertainment industry. It consists of two essays that employ Computational Social Science approaches involving explanatory empiricism, scientific theory, and machine learning methods to assess and evaluate different strategic marketing strategies.

Essay 1 discusses household informedness and its impact on the marketing of digital information goods, via free content samples for on-demand TV series. Different levels of household informedness influence its willingness-to-pay for *video-on-demand* (VoD), a niche class of entertainment goods that creates a high level of consumer uncertainty regarding quality and preference fit. Essay 2 proposes a methodological advance related to censored observation recovery using temporal sequences and iterative data simulation that improves statistical power for causal inference in data-driven exploratory research. The method is employed to recover traces of consumer visits to an online retailer, which give a better understanding of consumers' sources of information leading to their purchases.

The dissertation contributes to the growing body of research on consumer and business analytics by looking at the impact of consumer informedness on the sales of products and services in digital space. It also constitutes a methods innovation to handle censored data for the purpose of causal inference in explanatory research, to produce business policy-relevant findings for industry practitioners.

**Table of Contents**

## List of Figures

## List of Tables

# Acknowledgments

I would like to express my sincere gratitude and respects to my advisor, Professor Robert J. Kauffman, for his continuous support, guidance, and encouragement throughout my Ph.D. study. I feel extremely fortunate to have him as a mentor and a teacher. He has inspired me to accomplish more than I could have imagined in research that matters to people.

I would also like to thank Professors Kapil R. Tuli, Qian Tang, and Ting Li for their valuable advice, and comments that are essential in shaping my dissertation and future career. I truly appreciate the guidance from the late Professor Stephen Fienberg during my training at Carnegie Mellon University (CMU). I would also like to thank Professors Mark Kamlet and Kannan Srinivasan for being my mentors at CMU.

Professors Eric Clemons, Vladimir Zwass, Terence Saldanha, Nelson Granados, Pedro Ferreira, gave me many suggestions that helped to shape the intellectual contributions in my research. I appreciate insightful comments from Atanu Lahiri, Jennifer Zhang, Yabing Jiang, Avi Seidmann, Rong Zheng, Tuan Phan, Byungjoon Yoo the participants at workshops and conferences. Comments from research fellows Zhuolun Li, and my Ph.D. colleagues Tuan-Anh Hoang, Son Nguyen, Trong Le, Dan Geng, Kustini Lim-Wavde, Ren Jing, Deserina Sulaeman and Felicia Natali, have made my research stronger.

I would like to acknowledge and thank the Singapore National Research Foundation, under its International Research Center @ Singapore Funding Initiative and administered by the Infocomm Development Authority, the Living Analytics Research

Centre at Singapore Management University and the Post-Graduate Research Programme Office for funding my research. Many thanks go to my corporate sponsors for their sharing of data, under a binding non-disclosure agreement. I also thank Ms. Pei Huan Seow, Chew Hong Ong, Phoebe Yeo, Yong Fong Koh, Yar Ling Yeo at SMU and Barbara Diecks at CMU for all the administrative help.

And I am forever grateful for my family.

*In memory of my father,*

*To my loving family*

## Chapter 1: Introduction

Understanding human behavior has been a long-standing topic of research, and the interest in this topic has proliferated over the last decade. The relationship between people, and their interaction with the external environment are complex and hard to capture. Today, information technologies (ITs) have brought people closer together, and also enabled researchers to collect and analyze people online traces. These digital traces are hidden gems to the understanding of human behavior in the digital space. Nevertheless, the emphasis has moved away from the amount of information available, to how we choose to consume this information. The age of information has long gone, replaced by the age of experience (Jenkins 2017).

Today, information overload is the great Internet problem. In social networks, low-quality information, or fake news are disseminated quickly as information load increases (Qiu et al. 2017). Large-scale data have become more available to researchers from various sources, including public data, firm proprietary data, and user-generated data. The richness of data presents many opportunities for researchers to design insights-driven studies. Consumer analytics play a critical role in turning that data into meaningful stories, and serve as the backbone of all key business decisions.

The most noteworthy IT-driven trend in the last few years is the rise of the on-demand business model. Enabled by the seamless interaction between consumers and businesses, on-demand goods and services now are able to attract more consumer attention and spending. The three largest categories of commerce in this area include online marketplaces, transportation, and food delivery, which together account for

US$46.2 billion in annual spending (Colby and Bell 2016). The success of the technology-based business models that characterize these areas is due to their attractive offerings, as well as delivery systems that meet consumers' needs effectively.

In addition, firms no longer need to push their products out to the market, but to make their products available and accessible to the consumers. In order to do so, firms must be informed of consumer preferences and extend their marketing efforts strategically, so that the consumers can search for and purchase suitable products themselves. Nevertheless, new consumers in the on-demand market are so diverse, it is inappropriate to segment them based on socio-economic factors, when their consumption spans geographic boundaries, age groups, and other traditional descriptors. They should only be characterized by their technological readiness (Colby and Bell 2016). Today's innovative products and services are able to meet the non-tech savvy consumers halfway, and consumer benefits will be realized through time as they learn how to more effectively use new technologies.

Amid the transitions occurring in the market, industry incumbents are spending a substantial amount of effort to create innovations in their business models. The entertainment industry has quickly responded to the boost in consumption of digital entertainment content, which is fueled by new video streaming services. As a result, we have witnessed the emergence in the digital economy of rich new content with many related service innovations, as well as new content delivery platforms. This changing marketplace has created challenges and opportunities for service providers to offer effective delivery mechanisms, while protecting and monetizing the content of their products and services (Huang et al. 2009, Wu and Chen 2008).

Essay 1 examines how household consumers sample different series dramas via streaming *video-on-demand* (VoD) services to support subsequent purchases. Firms have addressed the shift in consumer behavior with innovations in their marketing strategies, so they can be more responsive to consumers' needs. Sampling strategies are effective in informing the consumers about both the quality and preference fit of the series dramas. Informed consumers are willing to pay more for content that fits their viewing preferences. The results from this work are useful in the evaluation of sampling-based strategies for experience goods.

Another industry that has also undergone a tremendous transformation is the retail industry. E-commerce platforms have disrupted the physical retail landscape, leaving the industry prospects uncertain as a result (Thomas 2017). Essay 2 assesses how online shoppers' search behavior in the digital space differs from that in the physical space. Different information sources and advertisement channels also inform shoppers in nuanced ways. Unbiased and more transparent sources of traffic, such as an unbiased comparison website, have greater influences on the consumer purchase decisions.

The theoretical lens of this work spans the IS, Economics and Marketing disciplines, with the aim to contribute new insights for the business community. In particular, it looks closer at consumer informedness in the age of experience, where the emphasis is placed on consumers' direct experiences with goods and services. With a large amount of information available, what type of information is relevant to consumer decision-making is critical to understand. Firms should realize that the market can recognize and adapt rapidly to business models that are beneficial to it.

These essays also demonstrate the use of fusion analytics, which combines machine-based methods and explanatory empiricism to overcome the limitations inherent in research designs involving digital trace data. The issue of censored observations raises a concern about an analyst's ability to make causal inferences in data-driven exploratory research. In addition, proprietary consumer data are protected under data privacy laws, making some data elements unavailable to researchers. The researcher's lack of control over a setting that generates the data shifts the research objective from *establishing true causality* to *making inferences about important relationships that come close to true causality*. So, even though a researcher may start with "big data," she may often end up only finding a "needle" of insight from multiple "haystacks" of digital traces of consumer behavior.

Essay 2 proposes a method innovation to recover censored observations using temporal sequences and iterative data stimulation. The method is used to recover households' viewing records outside the observation period in Essay 1 and consumers' visits and purchase records to an online retailer in Essay 2, so the statistical power of the empirical models can be improved. These consumer activities may have occurred outside the research period, but are essential for making inferences within it. By exploring "data needle in a large digital data haystack," my method allows insight extraction from digital trace data, in order to produce business policy-relevant findings for industry practitioners.

The next two chapters (Chapters 2 and 3) present the two essays. Section 4 shares the best research practices that I have learned and developed over the last 5 years related to the scientific research process, from the formation of research questions, and the

handling of less-than-ideal datasets to the development of a rigorous and innovative methodology framework. I also share my experiences in publication at a rank A+ journal, and how I have become a better reviewer in the research community. Section 5 concludes with contributions, limitations, and future research.

## Chapter 2: Informedness in the Consumption of Digital Information Goods

### 2.1. Introduction

Disruptive technologies, such as digital content-streaming platforms, have boosted the production and consumption of entertainment content.[1] Economies of scale now allow digital entertainment service providers to market and sell information goods directly to consumers on an on-demand, anytime, anywhere basis. Among the different types of content that are offered on-demand, *video-on-demand* (VoD) services are a key source of revenue for digital entertainment firms (Lafayette 2014). At the industry level, a consulting firm (Mordor Intelligence 2015) has estimated that the VoD market of US$47.25 billion in 2015 will grow to almost US$75 billion by 2020, representing a compound annual growth rate (CAGR) of 9.63 percent.

In the past decade, TV series have experienced a great upswing in consumer interest. Because of this surge in market demand, all of the TV studios, including industry incumbents such as ABC and CBS, and content distributors such as Netflix, Hulu, and Amazon, are competing in the race for the next "big show." They have invested heavily in original shows despite a high failure rate in the production stage, since the rewards for a successful hit come in so many different forms: more viewers, higher ad revenue, and most important perhaps, a competitive edge in sustaining the customer base (Nathanson 2013).[2] For example, the Hulu TV original series, The Handmaid's Tale, recently won eight Emmy awards. This success for the content distributor signals a whole

[1] The work is Hoang, A.P., and Kauffman, R.J. Content sampling, household informedness and the consumption of digital information goods, JMIS (in press). Thus, I will use 'we' throughout this chapter to reflect the work in the published paper.

[2] A *series drama* consists of 10, 20, 30 or more episodes. Most American TV series, packaged since the 1960s with 20 to 26 episodes a season, are like this. The economic importance of paid TV series revenue streams has increased,

new era for original on-demand content (New York Times 2017) and a growing global market.

Despite a recent audience report from Nielsen that reveals that Americans spend 70 almost eleven hours each day staring at the screen and consuming media (Howard 2016), content providers are struggling to market and sell their programming due to the high level of consumer uncertainty associated with the consumption of this class of products. A TV program's quality is known only after it has been watched, and imperfect information about its content typically decreases a consumer's willingness 75 to pay (Clemons et al. 2006, Clemons et al. 2003). In addition, entertainment products are horizontally differentiated; their value relies heavily on the subjective evaluation of consumers. With a large amount of content available, it is hard for consumers to choose what they are likely to enjoy, or what fits them best. Across different industries, various forms of sampling strategies have been used to communicate product information for experience goods to consumers. Readers of the *New York Times*, for example, can access up to ten articles each month, representing a *metered model* in the newspaper industry (Halbheer et al. 2014). In addition, software companies provide the most basic version of their software free of charge or an extended version for free during a trial period (Niculescu and Wu 2014). Online music distributors, such as Apple and Spotify, also make it possible for listeners to sample all of their songs—but for only 30 seconds each (O'Kane 2015). Production companies, meanwhile, have been making trailers and sneak peeks of shows they produce too. And firms also employ sampling strategies at

while providers have been fighting for profitability in the face of Internet delivery and digital convergence. Producing an original TV series requires a huge investment: about US$2 million to shoot a half-hour pilot and about US$5.5 million for an hour-long drama.

the service level, such as Netflix's one-month basic membership trial.

The wide implementation of sampling-based strategy for digital goods has much to do with the one-time fixed cost of content digitization and the associated cheap cost of distribution. The impact of such strategies is more profound though. The interdisciplinary literature on sampling strategies for information and experience goods has often focused on online music and software (Chellappa and Shivendu 2005, Dey at al. 2013). Such studies have investigated the determinants of consumer decision making and examined the consumption of these household purchases. We extend this literature with empirical evidence for the impact of sampled content on purchases of on-demand series dramas, a unique class of entertainment products. In this context, consumers are able to evaluate fit related to their preferences through the sampling of a series.

The theories we use are drawn from different streams of literature. The first deals with the specific characteristics of experience goods that create a high level of uncertainty (Shapiro and Varian 1999). We look at the impact of sampling strategy for physical goods (Freedman 1986, McGuinness et al. 1992), and the implications for experience goods. The second stream focuses on how sampling influences consumer buying behavior under uncertainty (Haubl and Trifts 2000, Markopoulos and Clemons 2013, Mehta et al. 2003). We examine issues related to consumer viewing behavior (McAlister and Pessemier 1982). To our knowledge, 105 this research is the first to provide empirical support for the effectiveness of sampling strategies related to the purchase of VoD series dramas, a niche product that consists of a video bundle with multiple episodes. Previously, Markopoulos (2004) examined sampling and video game purchases

with a smaller, less granular data set, as Clemons et al. (2005) later did for music sampling purchases, but not in the depth that we have.

We address two questions: (1) What are the impacts of different forms of *content samples* on a household's VoD *series purchases?* and (2) How do a household's choices of standard content and customized, add-on content affect its VoD series purchases? We also discuss the role of data analytics in effective implementation of sampling-based strategies for the marketing of digital information goods.

To answer these research questions, we designed a study to learn about the aggregate behavior related to free sampling and series purchases with an emphasis on the household level as our unit of analysis. We addressed causality and potential threats to the robustness of our main findings with additional econometric procedures. We used a blend of data analytics methods to establish evidence for causality. Our analysis work benefited from access to millions of TV viewing records, including those involving VoD content, across hundreds of thousands of households, and multiple sources of data on series dramas. The period of observation for VoD viewing records was limited though—just one month.

Without access to additional data or the ability to construct a set of formal field experiments within the operations of the sponsor of this research, we implemented an innovative approach using *propensity score matching* (PSM). It uses iterative replacement methods to pair observations across censored and noncensored data groups based on discoverable sequences over time, and patterns of observable past activities by the subjects—households, in our case. This allowed us to make inferences related to unobservable viewing records outside the study period, which caused data censoring. The

overall approach enabled us to make causal arguments about the impact of free samples, on the basis of our extensive data analysis. The findings contribute to theory and practice by highlighting the importance of an effective sampling-based strategy in marketing digital information goods, while offering new managerial knowledge about how to offer effective sampling to consumers.

## 2.2. Theoretical Background

We now turn to the relevant streams of literature: (1) product uncertainty associated with the consumption of digital information goods; (2) selling strategy for digital information goods; and (3) consumer viewing and purchase behavior for digital information goods.

### 2.2.1. Uncertainty Associated with Consumption of Digital Information Goods

*Product uncertainty* is viewed as an important construct in Marketing and IS research, as it directly affects consumers' willingness to pay for goods and services (Ba and Pavlou 2002, Rao and Monroe 1996). Hong and Pavlou (2014) distinguished between *uncertainty about product quality and uncertainty about product fit* with a consumer's taste. The product may not be in the promised condition (Pavlou et al. 2007), or the vendors may fail to communicate product information to consumers (Dimoka et al. 2012, Ghose 2009), hence *uncertainty about quality*. *Fit uncertainty* refers to the degree to which consumers are unable to assess whether a product's attributes match their preferences (Hong and Pavlou 2014). Imperfect information concerning quality and fit creates high perceived transaction costs and tends to diminish a consumer's willingness-to-pay (Liebeskind and Rumelt 1989).

In another stream of research, Nelson (1970) separated *experience goods* from

*search goods*: the *quality of search goods* can be determined simply by inspection before purchase, whereas the *quality of experience goods* is realized only after use. Thus, the assessment of digital information goods, such as music, books or movies, must involve personal experience (Jones and Mendelson 2011, Matt and Hess 2016). In fact, the actual source of quality is the experience itself, in which product fit plays a critical role (Kwark et al. 2014). A study on the craft beer industry has shown that firms with highly-differentiated products experience higher revenue growth when consumers become more informed (Clemons et al. 2006). They often are willing to pay more when the match between product characteristics and their preferences is improved. Different types and levels of informedness can also influence consumer choices (Li et al. 2014); for instance, elimination of product fit uncertainty for a digital experience good can increase the number of purchases and consumer loyalty (Matt and Hess 2016). In platforms on which entertainment is marketed and sold at the product level, the effects of consumer informedness about products and their fit become more pronounced.

### 2.2.2. Sales Strategy for Digital Information Goods

As streaming media has become affordable, and demand for content has increased, firms have had to adjust their strategies to be more effective with the selling of digital information goods. Online reviews and word-of-mouth are good sources of information on digital goods for consumers. Moretti (2011) showed that social learning and peer effects have positive impacts on the consumption of movies. Nevertheless, it is hard to describe the characteristics of an experience good, especially when consumer tastes vary significantly (Matt and Hess 2016). A TV program is better from a consumer's perspective if it fits her viewing preferences. Signaling quality and content is

achievable, while communicating fit is more complicated.

Previous studies have focused on selling strategies for digital information goods, and the market context and environment in which they are offered. Bhattacharjee et al. (2006) looked at online music sales in the presence of online piracy, and showed that effective pricing options, search tools, and licensing structures are leading strategies to mitigate the related revenue losses for the music labels and artists. The search process for digital information goods is different from that for physical goods. Each product is unique and has its own characteristics, so consumers need to repeat the search process for every purchase. As a result, the associated search cost will vary greatly and be proportional to the number of options available. As part of the transaction cost, search costs can influence consumer purchase decisions (Brynjolfsson et al. 2011, Johnson et al. 2004).

*Product sampling* lowers the search cost by effectively communicating product quality to consumers. Thus, it is a key promotional tool to stimulate sales for many products (McGuinness et al. 1992). A *sample* is a portion of a product given to consumers to try for free before making a purchase decision. Consumers like to receive free goods. Thus, *free samples* can influence their behavior at the point-of-purchase, encouraging unplanned purchases and active switching to promoted brands (Haubl and Trifts 2000, Pinsker 2014). For retailers of physical products, sampling yields a higher purchase conversion rate and return-on-investment than other direct advertising (Faugère and Kumar 2006, Freedman 1986). Nevertheless, it has mainly been used to enhance the effectiveness of traditional marketing only; the implementation of a sampling strategy is expensive, and the market reach is limited (McGuinness et al. 1992).

Considerable attention also has been given to sampling strategies for information and experience goods. Information goods are characterized by large sunk costs for development, and negligible costs of reproduction and distribution (Shapiro and Varian 1999). Digital content can be digitally broadcasted, streamed and stored at a relatively low cost. Niculescu and Wu (2014) explored the *economics of free* under perpetual licensing for two software business models. With a *feature-limited freemium,* consumers gain free access to a basic version of the software but have to pay for premium versions, while under *uniform seeding,* firms offer a full product for free to part of the market. Halbheer et al. (2014) studied the profitability of ad-supported content sampling for newspapers. In the entertainment sector, offering teasers or previews for movies and TV shows has become an industry norm; yet the implications are overlooked in the literature.

The execution of sampling strategies for digital content is not that straightforward though. Firms need to consider how individuals value the same product differently, reflecting customer heterogeneity, to design an appropriate strategy. For software products, the *rate of learning* by users determines the effectiveness of *time-locked trials* (Dey et al. 2013). Using data analytics though, firms can help buyers find their nearly "perfect" *product fit.* Netflix, for example, shows different trailers of the same series to different market segments, based on what it is able to figure out about their viewing preferences (Carr 2013). It may take longer for some consumers to reach a decision; yet offering lengthy samples is not desirable for most providers (Heiman at el. 2001). Free content may interfere with the market's consumption of programming, and free content on the Internet decreases consumer willingness-to-pay for content in other

channels (Berger et al. 2015).

### 2.2.3. Viewing and Purchase Behavior for Digital Information Goods

Research on consumer behavior has examined different aspects of TV viewing activity. Rubin (1983) looked at the interaction between viewing patterns and motivation and identified two viewer types: one watches TV out of habit to pass time; the second seeks information and watches TV to learn. Viewing activity is recognized as a *gratification-seeking process,* in which viewers search for and watch the content that matches their preferences (Lin 1993). Viewers may also modify their viewing preferences, a *variety-seeking behavior* (McAlister and Pessemier 1982). Variety-seekers respond positively to new programs, and new means of delivery across different platforms, such as their desktops, tablets, and phones.

Recently though, researchers have begun to focus more on specific types of programming, TV shows and series dramas. This has been due to the emergence of advanced content-streaming technology. A survey conducted by Harris-Netflix has shown that most viewers admitted to *binge-watching* (Newswire 2013); they get hooked and watch multiple episodes of a series in one sitting (Holloway 2015). Theoretical perspectives from multiple disciplines are helpful to explain this behavior. For example, connectedness, the relationship between a viewer and the characters, intensifies as she spends more time watching the show (Russell et al. 2004), and not having closure on how a story ends may cause *dissatisfaction* and *regret* (Bell 1982, Gilovich and Medvec 1995). The most prominent consideration is instant gratification, the desire to fulfill a need without delay (Baumeister and Bushman 2010). If the content triggers a viewer's interest, she will feel the impulse to purchase the show. On-demand services

make it easier for consumers to have access to extensive TV content, which influences consumption.

Personal experience with the viewing content is necessary, similar to other experience goods. A majority of viewers may agree on certain attractive features of a show, but they are unlikely to all enjoy watching it. A successful movie is not necessarily suitable for every member of its audience. Given a choice, consumers want to learn as much as possible about products by experiencing their content, rather than by gathering information about it from secondary sources (Mehta et al. 2003). Overall, this is a trade-off between effort and accuracy; consumers always gather risk-diminishing information when there is uncertainty. They often choose options that are satisfactory, but are suboptimal if decision costs were zero (Haubl and Trifts 2000).

### 2.3. Development of Hypotheses

This research was made possible through a partnership with a large digital entertainment firm in Singapore. There are varied kinds of programming from a number of *content clusters* (also called *genres*), such as news and children's programs, and entertainment and educational shows. Customers can specify the clusters of content as well as premium channels to be included in their subscription packages. Most channels are available in high-definition format also. Monthly subscription fees reflect the number, type, and quality of channels accessible to households.

The service provider also delivered a wide selection of movies and series dramas on demand, on top of a household's TV subscription. VoD services can be expensive though: a series with multiple episodes can cost from $3 to $60 in the market we studied. For each VoD series purchased, a household obtains immediate access over a pre-set

period – depending on the number of shows in the series. The service provider offers households the first episode of series dramas to watch for free before they make a purchase. We next develop hypotheses on content sampling, the purchase of VoD series, and the effects of subscriptions at the household level, based on different theoretical perspectives. In this study, we consider a *unitary model of the household* in which the viewing time constraint, demand, and preferences of all household members are pooled (Rode 2011).[3]

### 2.3.1. Free Sampling and Consumer Purchases

Information acquisition is known to be a costly and time-consuming, though valuable process (Demski 1980). Initially, households will be uncertain about the quality of a series and whether it fits their preferences. They actively seek fit-related, risk-diminishing product information before making purchase decisions, especially when there may be financial consequences (De Matos et al. 2016). Though they can learn about a TV series through various means – online and offline, such as through online reviews or viewership ratings – they will explore and update their evaluations of different series through the free episode samples. Samples give households direct and easy access to quality and preference fit information for a series, thus reducing the associated search cost. In addition, content sampling signals both horizontal and vertical differentiation on objective features of the series to consumers.[4] Free samples also reduce uncertainty, given that a household obtains direct experience with the content of one episode (Markopoulos at el. 2013). Thus, we offer:

[3] We observe all viewing activities at the household level only. The current technology in our setting does not permit tracking individual viewers.

[4] If the content only signals vertical differentiation, then consumers just need to know such samples are available, and they do not actually need to watch any free-sample episodes.

- **Hypothesis 1 (Household's Content Sampling).** *A household's free sampling of a series has a positive effect on its likelihood to purchase that series.*

Even when a household identifies a series that the viewers there will like, it is possible that the household members will sample a few other series to rule out the available alternatives. By sampling this way, they will be more informed in the decision to buy the VoD series. This greater involvement likely will lead to more than one purchase. First, the household members are more likely to find other acceptable entertainment goods that meet their preferences. Second, sampling also provides a way for a household to broaden its consumption. For instance, a household that normally prefers the comedy-related genre may sample a crime-related drama and find it interesting. Such variety-seeking behavior (McAlister and Pessemier 1982) may result in multiple purchases across different genres. And, because the first episodes of all series are offered for free, the perceived search cost for a household is minimal. So by increasing the household's involvement, free sampling ought to increase VoD series demand in the household. We assert:

- **Hypothesis 2 (Household's Purchase Decision Involvement).** *A household's involvement in its purchase decisions via content previews increases the number of drama series that it purchases.*

There are some drawbacks to free content. A perception that free content is available may dissuade consumers from buying programs (Kamins at el. 2009). Also, unlimited access to free content makes other programs less attractive and decreases consumers' willingness-to-pay (Berger at el. 2015). Further, some consumers may sample with no intention to purchase anything, though this is unlikely for a majority of them in the VoD setting for several reasons. Series dramas are unique, so a viewer's experience is not complete without seeing it all. So, after viewing the free sample of a series' first

episode, viewers may feel connected and want to view the rest of the content (Russell at al 2004). Those that sample a portion of the series are more likely to purchase the remainder of it. In addition, since households will have many channels in their TV subscriptions, they are unlikely to watch a free sample episode of a series if they have no prior topical interest.

### 2.3.2. Paid Sampling and Consumer Purchases

Households are likely to purchase the series that satisfy them based on their experience with free samples. This does not imply that a one-episode free sample is effective for all series though. Such a sample may not be sufficient for households to evaluate fit, as it is rarely the pilot episode that gets consumers hooked on a series. For instance, Netflix's method of releasing a series – in its entirety – has helped the company to understand customer viewing behavior for different series it offers across various market segments. This is relevant to our context, by showing that a one-episode free sample may not be sufficient for the viewers (Kastranekes 2015). After a household watches a first free episode of a series, they can purchase subsequent episodes of that series separately at the typical stated price, around $1 or $2 each, or purchase the whole series at a discount. The price of a series is fixed, regardless of how many episodes the household has already purchased. Hence, the best solution *ex post* is not the same as the best solution *ex ante*. The best option for those who like the series is to purchase it shortly after free sampling. If the household is still hesitant about buying the series, its members can also seek additional information from outside sources for further evaluation. This alternative option is not desirable though. For different series, the search costs involved can vary greatly, and yet the household will still not be able to evaluate fit.

Any episode purchased before the household has purchased the whole series is considered to be a *paid sample*, as the household pays to sample the series more. Purchasing a paid sample is preferable in this case. Continuing to watch the series is the most effective way to reduce uncertainty concerning fit; this is especially true after the household has already previewed the first episode. In addition, the consumer decision-making process involves a trade-off between effort and accuracy (Haubl and Trifts 2000), so households should be willing to pay more for direct fit over indirect fit information. As a household purchases more paid samples, it will become more informed about whether the content is suitable, and this should increase the number of series purchases. Hence, we posit:

- **Hypothesis 3 (Household's Informedness about Fit).** *A household's informedness about the fit of any drama series increases the number of drama series that it purchases.*

Pay TV and TV services represent a good source of revenue for service providers. It is useful to look at the interaction between the consumption of new service innovations, such as VoD series, and the consumption of existing services, especially when both are subject to time and budget constraints (Becker 1965). For instance, Liebowitz and Zentner (2012) showed the impact of Internet consumption as a substitute for television viewing. While the household's overall subscription package reveals its demand and preferences for TV viewing, the next two hypotheses examine a more nuanced relationship between the household's choices of content and its purchases of VoDs.

### 2.3.3. Standard Content Choices and Consumer Purchases

A household's TV subscription usually includes a selected number of standard content clusters. In our research context, the households decided a number of standard

content clusters in their TV subscriptions at the beginning of long-term service contracts. The households' content clusters were less subject to change, as they were required to wait for at least 6 months before they can change the details of their subscription. A *cluster* includes multiple channels that are similar in nature. For example, consider the News cluster, which includes local, regional and international news channels. The number of standard content clusters approximates how many channels the household has access to, as well as its monthly payment. Consequently, households with a variety of channels to choose from will be less interested in VoD content, especially because a VoD series is typically longer than other programming: a 20-episode drama, at 45 minutes per episode, takes about 15 hours to finish. A subtler implication is that even if a household likes the content of the series after the free episode, it is still less likely to purchase the series, due to time and budget constraints. The marginal utility from the consumption of a VoD series is likely to diminish. So the variety of choices in a household's subscription appears to interfere with its series purchases:

- **Hypothesis 4 (Standard Content Choice).** *The greater the number of choices of standard content in a household's TV subscription, the lower is the number of series it purchases.*

### 2.3.4. Customized Add-On Content Choices and Consumer Purchases

Households can also customize their viewing experience beyond standard content clusters by adding specific programs and niche channels, adding more channels in the same content cluster; or upgrading their subscribed channels to higher screen resolutions. These requests reveal a household's expected level of utility from TV viewing, and they reflect utility for additional paid content that goes beyond what is available in a typical household TV services subscription. The members of a household are likely

to experience different levels of utility, and not all of them will agree on the same programming content. For example, households with fewer members or those who do not have time for TV viewing are likely to be content with the basic channels; and yet households with small children may benefit from special educational programming. If TV viewing is the main form of entertainment for the household, then acquiring access to a more diversified set of channels beyond the basic subscription services is appropriate. Households with a higher level of utility are more likely to try out VoD services, and likely will have higher willingness-to-pay for more suitable content. Adding on more customized, paid services gives households more control over the content they watch, in the same manner that they were able to customize their packages when they initiated their service subscriptions. Thus, we assert:

- **Hypothesis 5 (Customized, Add-On Content Choices).** *The more customized, add-on choices a household's TV subscription service offers, the higher is the number of series it purchases.*

The household's choices for standard content versus its own customized, add-on choices have different impacts on its demand for VoD series purchases, as the service provider used different pricing structures for the standard clusters and the add-on channels.

### 2.4. Research Setting and Data

We first present our research setting and the data extraction approach that allowed us to gather information from various sources and handle the limitations that accompany it. Then, we analyze the datasets to discover the underlying causal relationships.

#### 2.4.1. Research Setting and Data Extraction Approach

The VoD and household-related data were collected through smartcards that are used in digital set-top boxes for digital cable TV and satellite entertainment systems.

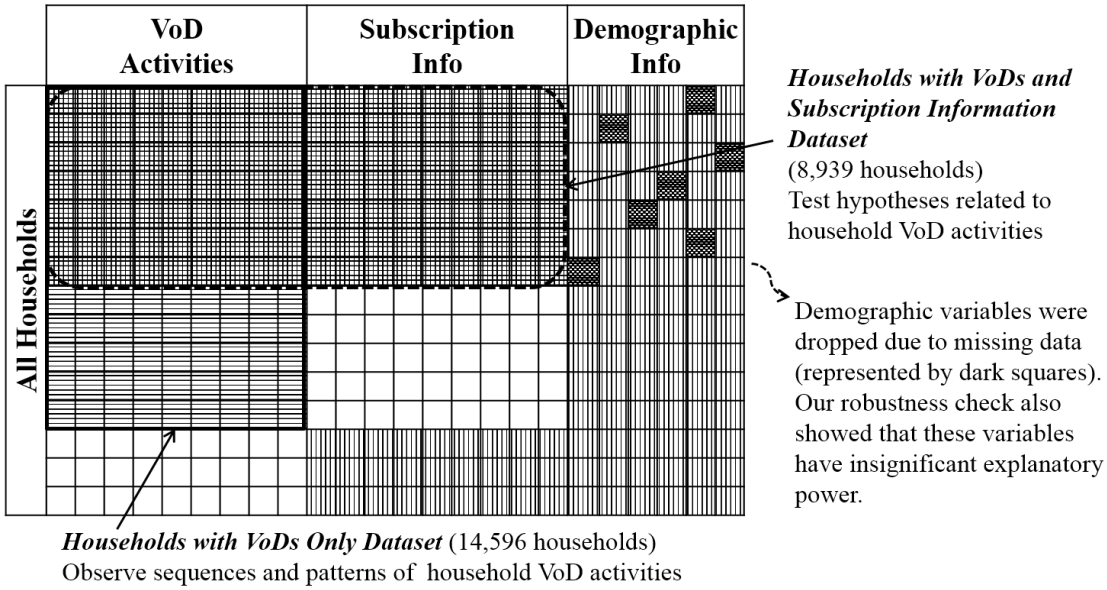
Smartcards store a household's information, the channels to which it subscribed, and all of the viewing records the smartcard captured. The technology does not identify which individual members watched the programming though. The voluminous data that we use pertain to household-level VoD viewing activities for one month between September 30 and October 30, 2011 and include 17-plus million viewing sessions. A *viewing session* for a TV program occurs when a household starts watching, and ends when it switches to another channel or turns off the TV. There are 3 categories of VoD sessions: (1) *free-sample sessions* include the viewing of first episodes of a series; (2) *paid-sample sessions* involve the viewing of purchased episodes; and (3) *series-purchase sessions* record the viewing of purchased series. Households often finish watching an episode across multiple viewing sessions, as each episode takes more than 30 minutes. So, if a household had three free-sample sessions for a series, we only admitted the earliest session to our dataset based on its timestamp and removed other duplicates. There were no holidays, promotions or special events during this period that might have influenced household viewing activities in ways which created anomalies in the data or household-level biases, to make our use of it problematic.

The large amount of set-top box data represents only one month of household viewing for the provider's market though. An important aspect of empirical research with consumer and household data-at-scale is to obtain as deep an understanding of behavior as the data will allow (Chang at el. 2014). Thus, we used multiple data sources to bring together the household information, series characteristics, and VoD activities for this study. A problem arises when there are many observations at the level of the primary

unit of analysis, but an incomplete set of variables across all the time periods or stratification. Meaningful stratification is sometimes difficult with big data research. Even though the researcher may have access to a lot of data, often it is surprisingly hard to develop research designs to support causal analysis, such as researcher-designed field experiments, and quasi-experimental designs that have "just right" conditions that can be leveraged to produce undeniably correct managerial insights. This forced us to make choices on how to construct a workable research design to support the overall research inquiry, while still yielding useful insights.

We implemented a data extraction approach, feature selection, to maximize the number of observations available for empirical testing. Feature selection refers to a process of strategically selecting a subset of variables that are relevant to address each research objective. We analyzed all VoD sessions for 14,596 different households. This set of anonymized households is called the *Households with VoDs Only Data Set*. We used it to explore the sequences and patterns of household VoD consumption. Nevertheless, it was not possible to link the full household-level information to the viewing-related variables that would have supported an ideal research design at the household level for the series-drama sampling the households did. We could only match 8,939 households with their subscription information. We call this the *Households with VoDs and Subscription Information Data Set*, and used it to test our hypotheses related to household VoD activities. (See Figure 2.1.)

**Figure 2.1. Approach Used to Extract Data for This Study**




Both datasets are representative of the entire customer population. We provide descriptive statistics for all households, and those used for empirical testing in the next section.

### 2.4.2. Analysis of Households' TV Viewing and VoD Activities

We took a closer look at the two datasets used in our research. Table 2.1 for the statistics of VoD activities for the *Households with VoDs Only Data Set* and the *Households with VoDs and Subscription Information Data Set. #SeriesPurchases* is the number of series that household *j* purchased in the study period. *#FreeSamples* is the number of one-episode free samples it watched, and *#PaidSamples* represents the number of episodes it bought. *ContentClusters* captures the number of groups of content, or groups of channels to which household *j* subscribed. *PremiumChannels* refers to the add-on channels selected when the service contract was signed. Together, they represent a household *j*'s subscription package. (See Table 2.1.)

**Table 2.1. Descriptive Stats: *Households with VoDs Only*, and *Households with VoDs and Subscription Information***

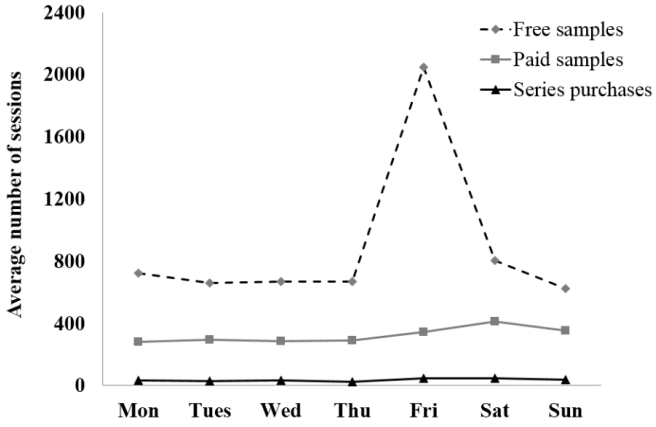
| VARIABLES | *HOUSEHOLDS WITH VODS ONLY* (14,596 OBS.) | | *HOUSEHOLDS WITH VOD AND SUBSCRIPTION INFORMATION* (8,939 OBS.) | | | | |
|---|---|---|---|---|---|---|---|
| | MEAN | SD | MEAN | SD | MIN | MEDIAN | MAX |
| $\#SeriesPurchases_j$ | 0.078 | 0.380 | 0.103 | 0.410 | 0 | 0 | 7 |
| $\#FreeSamples_j$ | 1.933 | 2.145 | 2.048 | 2.172 | 0 | 1 | 29 |
| $\#PaidSamples_j$ | 0.696 | 3.342 | 0.950 | 3.727 | 0 | 0 | 93 |
| $ContentClusters_j$ | | | 3.909 | 1.280 | 0 | 3 | 19 |
| $PremiumChannels_j$ | | | 3.201 | 3.046 | 0 | 2 | 25 |

**Notes.** The two samples were similar in terms of the mean of the main variables: $\#FreeSamples_j = 1.933 < 2.048$; $\#PaidSamples_j = 0.696 < 0.950$; $\#SeriesPurchases_j = 0.078 < 0.103$.

In the *Households with VoDs Only Data Set*, the anonymized households viewed 28,214 free samples for the first episodes of the various series, and 10,164 paid samples of other episodes. There were 1,140 series purchased, which yielded a conversion rate for free samples to series purchases of 4.04%. A closer look at the volume of household sampling and purchasing activities throughout the weeks revealed an interesting pattern.

We observed similar sampling and purchasing patterns. A surge of free-sample activity on Fridays was followed by a high number of paid samples and series purchases on Saturdays. These patterns provide visual evidence for the positive relationship between sampling and purchasing and suggest that the anonymized households searched for shows so they could watch them during the weekend. All activities slowed down during the weekdays though; the households did not have as much time during the week for TV viewing. The gap between the number of free samples and series purchases points to room for service providers to improve the conversion rate for VoD content. (See Figure 2.2.)

**Figure 2.2. Average Number of Samples by Type and Series Purchased, by Day of the Week**




We used the *Households with VoDs and Subscription Information Data Set* for empirical testing to examine the underlying relationships. The conversion rate for free samples to series purchases of these households is 5.02%. The correlation matrix for the variables in this dataset is reported. (See Table 2.2.) Households with many content clusters were more likely to have more premium channels, so the correlation was 57.6%.

**Table 2.2. Correlation Matrix for the *Households with VoDs and Subscription Information Data Set* (8,939 households)**

| VARIABLES | CORRELATION MATRIX | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1. $\#SeriesPurchases_j$ | 1.000 | | | | |
| 2. $\#PaidSamples_j$ | 0.195 | 1.000 | | | |
| 3. $\#FreeSamples_j$ | 0.162 | 0.063 | 1.000 | | |
| 4. $ContentClusters_j$ | 0.084 | 0.061 | -0.031 | 1.000 | |
| 5. $PremiumChannels_j$ | 0.111 | 0.096 | -0.029 | 0.576 | 1.000 |

**Notes.** *j* denotes individual households; the least correlated variables are $\#FreeSamples_j$ and $PremiumChannels_j$ (-2.9%), and the most correlated ones are $ContentClusters_j$ and $PremiumChannels_j$ (57.6%).

Other considerations in the household VoD purchases are the nature of the service offerings and the characteristics of the series themselves. Factors such as ads, price and rental time are likely to influence household purchase decisions. In our context, the service provider advertised all series dramas under "VoD Services", thus there were no

advertisement effects for individual series. Higher-quality and more popular series from particular markets or genres may receive more attention from viewers, and thus they were sampled and purchased more. For example, romantic Korean dramas have attracted audiences worldwide in recent years. Consequently, we may over-estimate the effect of free samples on a subset of popular dramas. Due to data scarcity, however, we cannot incorporate these factors into the main models, so we conducted a series-drama level analysis separately. We also extracted outside quality information on the series, such as viewership, ratings and award nominations to assess the impact of sampling versus outside quality information on series sales.

### 2.4.3. Analysis of VoD Series and Quality-Related Information Data

There were 79 on-demand series dramas offered during the study period. We gathered additional information about them from external sources such as spcnet.tv, TVB.com and JayneStars.com. Spcnet.tv is a large Asian drama review database, with a community of 50,000+ members. TVB is one of the largest commercial Chinese program producers in Hong Kong. Its website, TVB.com, posts information such as news, events, casts, and award nomination for all programming. JayneStars.com belongs to JayneStars Media LLC, located in New York. It features the latest Asian entertainment news from Hong Kong and China, and covers current TV dramas and movies.

*SeriesPrice* refers to the amount a household pays to gain access to a particular series *i* in a given time period, or *RentalPeriod*. *FreeSamples* is the times a series' first episode was sampled, and *PaidSamples* is the number of episode purchases; *TotalSamples* is the sum total of these two variables. *TotalPurchases* refers to the number of

purchases for series *i*. Descriptive statistics and the correlation matrix for these variables are provided. (See Tables 2.3 and 2.4.)

**Table 2.3. Descriptive Statistics for Series Drama Variables**

| VARIABLES | DESCRIPTIVE STATISTICS | | | | |
|---|---|---|---|---|---|
| | MEAN | SD | MIN | MEDIAN | MAX |
| $TotalPurchases_i$ | 14.43 | 34.06 | 0 | 2 | 189 |
| $FreeSamples_i$ | 357.10 | 492.07 | 1 | 201 | 2,741 |
| $PaidSamples_i$ | 128.70 | 285.40 | 0 | 20 | 1,473 |
| $SeriesPrice_i$ | 21.90 | 11.25 | 3 | 19 | 60 |
| $RentalPeriod_i$ | 36.20 | 10.95 | 30 | 30 | 75 |

**Notes.** Obs.: 79. Each season is viewed as independent for series with multiple seasons. Origin of series: China (CHN): 13; Hong Kong (HK): 43; Indonesia (IDN): 5; Korea (KOR): 4; Malaysia (MYS): 4; Taiwan (TWN): 10.

**Table 2.4. Correlation Matrix for Series Drama Variables**

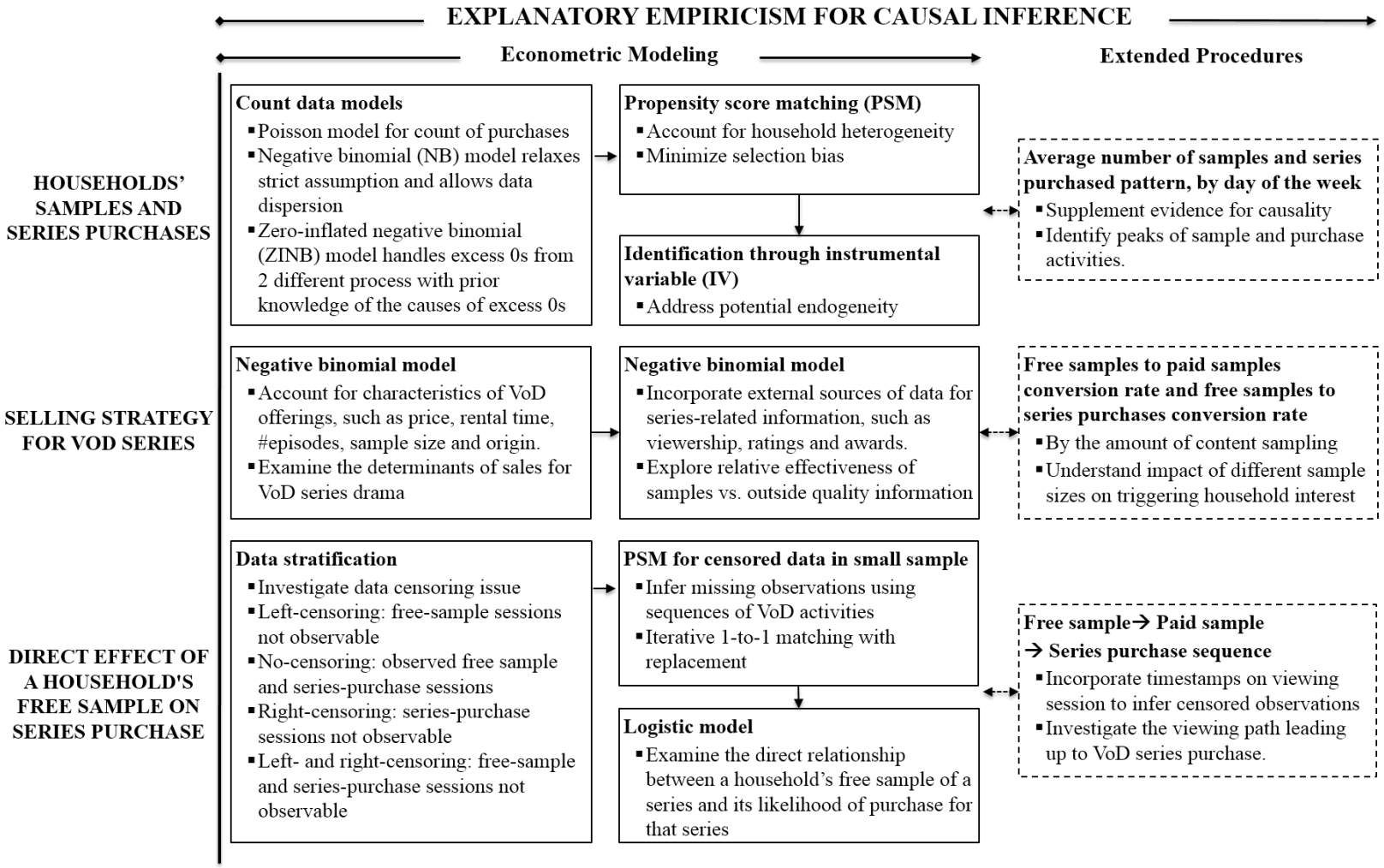
| VARIABLES | CORRELATION MATRIX | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| *1.*$TotalPurchases_i$ | 1.00 | | | | |
| *2.* $FreeSamples_i$ | 0.76 | 1.00 | | | |
| *3.* $PaidSamples_i$ | 0.72 | 0.55 | 1.00 | | |
| *4.* $SeriesPrice_i$ | 0.08 | 0.06 | 0.20 | 1.00 | |
| *5.* $RentalPeriod_i$ | 0.39 | 0.41 | 0.30 | 0.62 | 1.00 |

**Notes.** The most correlated are $TotalPurchases_i$ and $FreeSamples_i$ (76%); the least correlated are $SeriesPrice_i$ and $FreeSample_i$ (6%).

## 2.5. Research Methodology

We next present the explanatory empirical approach we used for causal inference in this study. (See Figure 2.3 for an overview of the data analytics procedures.)

**Figure 2.3. Overview of the Data Analytics Procedures in this Study**



To test the hypotheses on the overall effectiveness of sampling strategy on the consumption of series dramas, we used different count data models that can handle aggregated data at the household level over a one-month study period. We also implemented propensity score matching (PSM) to reduce selection bias due to household heterogeneity, and address the endogeneity issue, by using a suitable instrumental variable for a household's free samples. In order to test for a direct relationship between a household's free sample of a series and its likelihood of purchase for that series, we needed to handle the issue of left and right data-censoring in our dataset. Finally, we also implemented an identification strategy using heterogeneity across the VoD series.

### 2.5.1. Empirical Testing Procedures

**Count data models.** The variable of interest is the count value of VoD *#SeriesPurchases* for each household. This value is censored at 0, if a household did not purchase any series; censoring makes *ordinary least squares* (OLS) estimates inconsistent (Greene 2012).

We captured the relationship between the number of *#SeriesPurchases* and other variables via this function: *#SeriesPurchases = f (#FreeSamples, #PaidSamples, ContentClusters, PremiumChannels)* for each household *j*, and estimated:

$$\#SeriesPurchases_j = \beta_0 + \beta_1 \#FreeSamples_j + \beta_2 \#PaidSamples_j + \beta_3 ContentClusters_j + \beta_4 PremiumChannels_j + \varepsilon$$

Since most households did not make many purchases and the maximum was just 7 series, we assessed various *count data models* that are appropriate to handle these characteristics. Count models restrict the dependent variable to non-negative integer values,

and account for the mean and variance of the distribution used to characterize the dependent variable (Cameron and Trivedi 1998).

In the different count data models that we used, we did not include any household demographic characteristics as control variables. Instead, we used them in our propensity score matching approach. These variables include the demographic segmentation of the household, such as the region of the residence, age band and gender of the residents. Other specifics regarding the ethnicity of the anonymized households are not included or reported, due to our non-disclosure agreement with the research sponsor. In fact though, these variables did not add much explanatory capability for the dependent variable of interest.

**Poisson regression model.** The most well-known of the discrete regression models for count data is the *Poisson model*, which takes the form of: $y_j \sim Poisson\ (\theta_j)$ for $j$ = 1,…, $N$ and all $y_j > 0$; $\theta_j = exp\left(\sum_j^n \beta_j\, x_j\right)$ and all $\theta_j > 0$; and finally $y_j \sim Poisson\ (exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n))$. Using the Poisson distribution, the events are estimated as independent of one another, without any restrictions on the independent variables. It constrains the conditional mean and variance of the dependent variable to be the same though, which is not appropriate for our research. So, we use this model as an estimation baseline only.

**Negative binomial (NB) model.** We observed a sparse dependent variable matrix, which is common in purchase conversion research, as the majority of households did not make or made few purchases. In our data, this was a larger proportion than what we would see for a normal distribution. (See Table 2.5.)

**Table 2.5. Conversion Rates of Free Sample for Households**

| CONVERSION RATE | *HOUSEHOLDS WITH VODS AND SUBSCRIPTION INFORMATION DATA SET* |
|---|---|
| Paid samples only | 8.09% |
| Series purchases only | 3.44% |
| Paid samples and series purchases | 3.56% |
| **Notes.** Household conversion rate of free samples to purchases = (# of the household's purchases) / # of its free samples. | |

*Over-dispersion* occurs when the conditional variance of the dependent variable exceeds the conditional mean. As a result, the standard errors of the parameter estimates from the model will be underestimated (Hilbe 2011), and the estimated values of the parameters will be greater than would be predicted based on the use of the Poisson distribution for the observed event counts. We checked for over-dispersion by calculating the *over-dispersion ratio*, which is more or less than 1 if there is over-dispersion or under-dispersion, respectively. *Negative binomial regression* generalizes the Poisson model and handles this issue. It has an extra parameter, $\alpha$, to model the degree of over-dispersion: the larger $\alpha$ is, the greater the amount of over-dispersion in the data. The confidence intervals for the negative binomial model are also narrower compared to those of a Poisson regression model.

**Zero-inflated negative binomial (ZINB) model**. In addition to over-dispersion, our datasets exhibited more 0s for *no purchase decisions* than those that the Poisson model can handle. The Poisson model also assumes that the zeros and non-zeros come from the same *data-generating process* (Cragg 1971); this is not true in our setting though. The class of *zero-inflated models* relaxes this assumption (Gurmu and Trivedi 1996), by modeling the response variable as a mixture of the Bernoulli and Poisson distributions. *Hurdle models* also relax the assumption that the zeros and non-zeros in

the dataset come from the same data-generating process, by using a Bernoulli probability that governs the binary outcome for the count variable with a 0 or a positive count. Once the hurdle or threshold is crossed, and a positive number occurs, the conditional distribution is represented by a truncated-at-zero count data model. Since we had prior knowledge of the cause of the excess 0s, we chose to proceed with zero-inflated models though.

A household's *zero-purchase* decision may result from different processes. For example, if a household does not have money or time to consume the whole series, they will not purchase regardless of whether they watched the free previews. And if the household purchases a VoD series, then its decision-making process will have been a function of perceived quality and fit, in keeping with their unitary or aggregate preferences. This is a *count process* model, where the count is influenced by other variables. Based on our observation of the anonymized households' TV viewing activities, the consumption of on-demand content is bounded by several constraints. Thus, we modeled the expected count of *SeriesPurchases* as the result of a combination of two processes:

$$E(\#SeriesPurchases_j = k) = Pr(HouseholdWithConstraints) \cdot 0 + Pr(HouseholdWithoutConstraints) \cdot E(\#SeriesPurchases_j = k \mid HouseholdWithoutConstraints)$$

To account for this, we chose the *zero-inflated negative binomial* (ZINB) *regression* model, which has a *logit model* part and a *negative binomial count data model* part. The logit part models the probability of excess 0s independently; the probability of *#SeriesPurchases* = 0, due to the fact that a household's purchases are bounded by

some constraints. The covariate, *ContentClusters*, reveals some of these constraints for household *j*. The two parts do not need to use the same predictors, and the estimated parameters do not need to be the same either. Since $y_j$ below represents *#SeriesPurchases*, the number of series purchased by household *j*, the probability density function is:

$$Pr(Y_j = y_j) = \begin{cases} \Phi + (1 - \Phi)\left(1 + k\mu_j\right)^{-k^{-1}} & y_j = 0 \\ (1 - \Phi)\dfrac{\Gamma\left(y_j + k^{-1}\right)}{y_j!\,\Gamma(k^{-1})}\dfrac{\left(k\mu_j\right)^{y_j}}{\left(1 + k\mu_j\right)^{y_j + k^{-1}}} & y_j > 0 \end{cases}$$

with $E(y) = \mu_j(1 - \phi)$; and $Var(Y_j) = \mu_j(1 - \phi)\ (1 + k\mu_j + \phi\mu_j)$, where $\mu_j$ and $\phi$ depend on the covariates. Here, $\phi$ is the density function governing the binary process such that $0 \leq \phi < 1$, and the dispersion parameter $k \geq 0$ is a scalar (Lawal 2012). When $\phi$ or $k$ is greater than 0, there is over-dispersion. When $\phi = 0$, the equation reduces to a negative binomial, and for $k = 0$, it becomes a *zero-inflated Poisson* (ZIP) *model*.

### 2.5.2. Propensity Score Matching (PSM) to Address Selection Bias

Causal inference using observational studies has been a central pillar of many disciplines (Ding et al. 2017). A *causal effect* is a comparison between the potential outcome of a treatment group and a control group, averaged over a population (Rubin 1973). Without a randomized assignment, bias may arise due to systematic differences between the groups. In our business context, the households that watched free-sample episodes may be different from those that did not. The differences between these households produce bias in our estimations. Matching methods have been used effectively to address this problem (Rosenbaum and Rubin 1983, Rosenbaum and Rubin 1985); they

involve the pairing of treated and controlled observations that are similar in some observable characteristics.

In our *Households with VoD and Subscription Information Data Set,* we identified 586 households without any free-sample sessions. We used the PSM approach and found comparable matches for these anonymized households, based on two sets of covariates that are likely to have influenced the households' decisions to sample VoD content. The treatment is the household's exposure to VoD sampling, and the outcome is *SeriesPurchases.* The first set of covariates consists of *LoyalCustomer, ValueCustomer, EarlyAdopter* and *TechOptimist*, representing four different household relationships with the service provider. *TechOptimist* represents the households that typically respond promptly to new products and services, and *EarlyAdopter* represents the households that were first to subscribe to new offerings. *LoyalCustomer* refers to households that were observed to use multiple services from the provider, and *ValueCustomer* refers to those with high-value contracts with the provider.

The second set of covariates includes demographic variables such as *AgeBand, Ethnicity, HouseholdSegment, Housing*, and *Region.* The category of variables, *HouseholdSegment,* captures the diversity of the customer base, which may reflect the differences in viewing preference. *Housing* offers a way to control for household size and income, as larger and wealthier families tend to live in larger residences. We weighted the differences between the covariates for the households that were observed to have sampled VoD content and those that did not, in order to establish *statistical equivalence* between the treatment and control groups (Li 2016, Oestreicher-Singer and Zalmanson 2013). This matching method yielded 1,655 households with free samples and 394

households without free samples.

### 2.5.3. Instrumental Variable (IV) Analysis for a Household's Free Samples

Another issue in our econometric models is whether the variable, *#FreeSamples*, is exogenous. We handled this endogeneity issue by finding a suitable instrumental variable (IV) for a household's free samples. A suitable IV should be exogenously related to that household's tendency to sample VoD series, but not affect its VoD series purchases. We noticed that, at the time of the research, the service provider offered an interactive home entertainment service to households on a monthly subscription, on-demand basis. Households that subscribed to this service were able to access an extensive library of songs in various languages to watch or sing along with. It was offered on the same platform as the VoD series. Every time a household used this service, it was exposed to a variety of VoD series. Thus, households that used the service frequently were more likely to sample VoD series. Yet we did not expect to see a direct relationship between a household's usage of this service and its series purchases.

### 2.5.4. Propensity Score Matching (PSM) to Handle Data Censoring

In marketing, medical epidemiology and employment research, *data-censoring* has been a common challenge since historical data for consumers, patients and employment are rarely available in complete form. In censored-data, total observations are known but full information is not available for some (David and Johnson 1954). *Left-censoring* arises when the events of interest occurred before the study period; *right-censoring* refers to events that might or might not have occurred after the period of observation ended. Data without censoring are ideal for empirical testing.

In addition, personally-identifiable information on consumers must be masked due

to privacy regulations. In this research, we encountered left and right data-censoring for free- and paid-video sampling, as well as subsequent purchases, during the one-month time window. Thus, the number of observations in the *non-censored data* category is relatively small. This small set is also infeasible for empirical testing to gauge the effect of a household's free samples on its likelihood to purchase that series, as each free-sample session corresponds to a purchase session. Common computational and resampling approaches, such as the *partial deletion*, *multiple imputation* and *bootstrapping methods*, are not suitable to handle this issue (Efron 1981, Efron and Tibshirani 1993).

Censored-data create a roadblock for establishing a solid foundation for causal inference. We view this as an opportunity for a methodological advance, however. We propose an observation-matching method that requires the recognition of patterns and the adherence to a particular kind of ordering, or sequence in all observations, to match observations so censored records for some observations can be preserved. Our method extends the PSM and data imputation approaches to match and impute the values of the censored records from outside the observation window based on a probabilistic model (Dehejia and Wahba 2002, Gemici et al. 2012, Pirracchio 2012). This is an advance for identifying causal links, by improving the completeness of the observational data for causal inference.

### 2.6. Results

We offer the main empirical results from our econometric models, followed by analytical procedures to address concerns that a reader may raise. Last, we discuss the robustness of our identification strategy.

### 2.6.1. Household's Samples and Purchases of VoD Series

The estimation results obtained from count data models support the positive relationship between a household's samples and the number of VoD series it purchased. To strengthen this relationship, we include procedures to address selection bias and endogeneity issues arising from heterogeneity across different households and different VoD series.

**Count data models results.** At the household level of analysis, table 2.6 shows the results of the Poisson model. (See Table 2.6.)

**Table 2.6. Poisson Model Results: Household Level**

| VARIABLES | COEF. | SE | Z-VAL. | $p$ (>\|$z$\|) |
|---|---|---|---|---|
| *Intercept* | -3.320*** | 0.102 | -32.495 | < 0.001 |
| *#FreeSamples* | 0.135*** | 0.008 | 16.781 | < 0.001 |
| *#PaidSamples* | 0.041*** | 0.003 | 15.457 | < 0.001 |
| *ContentClusters* | 0.080*** | 0.027 | 2.933 | < 0.003 |
| *PremiumChannels* | 0.078*** | 0.011 | 7.202 | < 0.001 |
| **Notes.** Model: Poisson; 8,939 obs.; dep. var.: *#SeriesPurchases*. Null dev.: 4,901.6, 8,939 d.f.; resid. dev.: 4,388.3, 8,939 d.f., pseudo $R^2$: 0.080, AIC: 5,906.3. Signif.: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. | | | | |

The over-dispersion ratio of 1.289 from the Poisson model suggests over-dispersion estimation bias. The NB model, with an extra parameter that estimates the degree of over-dispersion. (See Table 2.7.)

**Table 2.7. Negative Binomial Model Results: Household Level**

| VARIABLES | COEF. | SE | Z-VAL. | $p$ (>\|$z$\|) |
|---|---|---|---|---|
| *Intercept* | -3.609*** | 0.134 | -26.973 | < 0.001 |
| *#FreeSamples* | 0.178*** | 0.013 | 13.725 | < 0.001 |
| *#PaidSamples* | 0.094*** | 0.006 | 14.957 | < 0.001 |
| *ContentClusters* | 0.090** | 0.035 | 2.536 | 0.011 |
| *PremiumChannels* | 0.085*** | 0.014 | 5.920 | < 0.001 |
| **Notes.** Model: Negative binomial; 8,939 obs.; dep. var.: *#SeriesPurchases*. Null dev.: 3,058.6; 8,939 d.f.; resid. dev.: 2,629.4, 8,939 d.f., pseudo $R^2$: 0.067, AIC: 5514. $\theta = 0.31$; degree of dispersion: $\alpha = 1/\theta = 3.27$. Signif. as above. | | | | |

The NB model produced coefficients that are slightly larger than those of the Poisson model (0.178 > 0.135, 0.094 > 0.041, 0.090 > 0.080, and 0.085 > 0.078). We justified the use of the NB model, by showing that the data are over-dispersed. The Poisson model is nested in the NB model. It relaxes the assumption that the conditional variance is equal to the conditional mean. We used a *likelihood ratio test* to assess the null hypotheses to see if this restriction is true: $\lambda = -2 \cdot (LL_{NB} - LL_{Poisson})$. We rejected the null hypothesis for it being appropriate in favor of the NB model, based on $\chi^2 = 394.29$. This exceeds 2.71 ($p < 0.001$), so overall the evidence suggested the data are over-dispersed. Next, the ZINB model deals with the excess zeros for no-purchase decisions in the dataset, by modeling "true zeros" and "inflated zeros" separately. The impact of free samples is stronger compared to the results from the prior models. (See Table 2.8.)

**Table 2.8. Zero-Inflated Negative Binomial Model Results: Household Level**

| VARIABLES | COUNT DATA PART | | | | LOGIT PART | | | |
|---|---|---|---|---|---|---|---|---|
| | COEF. | SE | Z-VAL. | P (>\|Z\|) | COEF. | SE | Z-VAL. | p (>\|z\|) |
| *Intercept* | -3.029*** | 0.259 | -11.689 | < 0.001 | 3.670 | 0.812 | 1.369 | 0.171 |
| *#FreeSamples* | 0.181*** | 0.016 | 11.532 | < 0.001 | | | | |
| *#PaidSamples* | 0.091*** | 0.009 | 9.592 | < 0.001 | | | | |
| *ContentClusters* | -0.009 | 0.049 | -0.184 | 0.854 | -1.505* | 0.805 | -1.869 | 0.062 |
| *PremiumChannels* | 0.088*** | 0.015 | 5.981 | < 0.001 | | | | |
| *Ln* ($\theta$) | -0.949*** | 0.150 | -6.316 | < 0.001 | | | | |

**Notes.** Model: Zero-infl. neg. binom.; 8,939 obs.; dep. var.: *#SeriesPurchases*. AIC: 5,506.8. $\theta = 0.387$. Signif. as above.

We show that the ZINB model fits the data better than the *null intercept-only model* does. The associated $\chi^2$ value for the difference between the model-level log likelihoods, $\lambda = -2 \cdot (LL_{ZINB} - LL_{Null})$ is 408.64. So the ZINB model is preferred over the null intercept-only model. We used a *closeness test* to check whether the two models were indistinguishable (Vuong 1989). Based on a Vuong test statistic of 1.75 ($p < 0.1$), we rejected the null hypothesis that the two models were equally close to the true data-

generating process.

We report the estimates of the ZINB model as our main results. The coefficients for *#FreeSamples, #PaidSamples* and *PremiumChannels* are positive and significant. The coefficient for *ContentClusters* is negative as we expected, but not significant though. The marginal effects of *#FreeSamples, #PaidSamples* and *PremiumChannels* are 1.198 (= $e^{0.181}$), 1.095 (= $e^{0.091}$), and 1.092 (= $e^{0.088}$), respectively. The exponential values of the coefficients represent the *incidence rate ratio,* which is the relative risk of something occurring versus not occurring (Dupont 2002). (See Table 2.9.)

**Table 2.9. Incidence Rate Ratios for Coefficients from ZINB Model and Their Confidence Intervals**

| VARIABLES | COEF. | CONFIDENCE INTERVAL | |
|---|---|---|---|
| | | 2.5% | 97.5% |
| *Intercept* | 0.048 | 0.029 | 0.080 |
| *#FreeSamples*$_j$ | 1.198 | 1.162 | 1.236 |
| *#PaidSamples*$_j$ | 1.094 | 0.075 | 1.115 |
| *ContentClusters*$_j$ | 0.991 | 0.900 | 1.092 |
| *PremiumChannels*$_j$ | 1.092 | 1.061 | 1.124 |

**Note.** 2.5% and 97.5% are lower and upper bounds of the 95% confidence intervals for coefficients.

We further leveraged them to interpret the estimation results in terms of their statistical confidence intervals. If a household were to watch one free sample more, for example, its corresponding incidence rate ratio would be expected to increase by a factor of 1.198. Thus, households with an additional free sample will purchase dramas 19.8% more of the time, supporting the Household's Purchase Decision Involvement Hypothesis (H2). Likewise, an additional paid sample caused a 9.4% increase in the number of series purchased, aligning with the Household's Informedness about Fit Hypothesis (H3). An additional premium channel predisposed a household to have a 9.2% increase in the number of series purchased, which supports the Customized, Add-On

Content Choices Hypothesis (H5). We did not find significant support for the Standard Content Choice Hypothesis (H4) though. Interestingly, the results also reveal that the log odds of the excess 0s decreased by 1.505 for each content cluster that a household subscribed to. This implied that no-purchase decisions were less likely due to time and budget constraints.

**ZINB model results after the PSM procedure.** The imbalance in the covariates may have affected the outcome of our results. Households that sampled free episodes are different from those that did not sample them, which influenced their series purchase decisions. We employed the PSM approach to match the households with and without free samples. Table 2.10 shows the ZINB model results after the PSM approach was applied. These coefficients align with our main results, which provides additional support for the impact of content sampling on the consumption of VoD series. (See Table 2.10.)

**Table 2.10. ZINB Model Results After the PSM Approach Was Applied**

| VARIABLES | COUNT DATA PART | | | | LOGIT PART | | | |
|---|---|---|---|---|---|---|---|---|
| | COEF. | SE | Z-VAL. | P (>\|Z\|) | COEF. | SE | Z-VAL. | p (>\|z\|) |
| *Intercept* | -1.945*** | 0.509 | -3.823 | < 0.001 | 2.497* | 1.409 | 1.772 | 0.076 |
| *#FreeSamples*$_j$ | 0.156*** | 0.025 | 6.355 | < 0.001 | - | - | - | - |
| *#PaidSamples*$_j$ | 0.072*** | 0.014 | 5.107 | < 0.001 | - | - | - | - |
| *ContentClusters*$_j$ | -0.121 | 0.093 | -1.303 | 0.193 | -0.911* | 0.487 | -1.873 | 0.061 |
| *PremiumChannels*$_j$ | 0.075** | 0.029 | 2.569 | 0.010 | - | - | - | - |
| *Ln* ($\theta$) | -0.149 | 0.441 | -0.337 | 0.736 | - | - | - | - |

**Notes.** ZINB model; 2,049 obs.; dep. var.: *#SeriesPurchases*. pseudo $R^2$: 0.055, AIC: 1,589, $\theta = 0.862$. Signif. as above.

**Two-stage least-squares (2SLS) IV results**. We used the number of household-level home entertainment sessions as an IV for a household's free samples. We removed all duplicate sessions on the same day. We also conducted an endogeneity test on the 479 households that subscribed to home entertainment services. The estimation

results for the OLS and 2SLS models are reported in Table 2.11, suggesting that even if the *#FreeSamples* variable is considered to be endogenous, the results are still in alignment with our earlier findings. (See Table 2.11.) The Hausman IV test result ($\chi^2$ = 0.511) for endogeneity shows that *#FreeSamples* can be treated as exogenous, however.

**Table 2.11. Linear Model Estimation Results with an Instrumental Variable (IV)**

| VARIABLES | LINEAR MODEL WITHOUT IV | | | |
|---|---|---|---|---|
| | COEF. | SE | Z-VAL. | *p* (>\|*z*\|) |
| *Intercept* | -0.058 | 0.090 | -0.647 | 0.518 |
| *#FreeSamples*$_j$ | 0.038*** | 0.009 | 4.073 | < 0.001 |
| *#PaidSamples*$_j$ | 0.036*** | 0.006 | 5.591 | < 0.001 |
| *ContentClusters*$_j$ | 0.030 | 0.025 | 1.215 | 0.225 |
| *PremiumChannels*$_j$ | 0.000 | 0.009 | 0.045 | 0.964 |
| **VARIABLES** | **2ND-STAGE ESTIMATES WITH IV** | | | |
| | **COEF.** | **SE** | **Z-VAL.** | ***p* (>\|*z*\|)** |
| *Intercept* | 0.078 | 0.211 | 0.369 | 0.713 |
| *1stStageErrors* | -0.017 | 0.077 | -0.219 | 0.827 |
| *#PaidSamples*$_j$ | 0.041*** | 0.010 | 4.253 | < 0.001 |
| *ContentClusters*$_j$ | 0.025 | 0.027 | 0.943 | 0.346 |
| *PremiumChannels*$_j$ | 0.002 | 0.010 | 0.247 | 0.805 |

**Notes**. Model: Linear without IV, estimated with OLS; 474 obs.; dep. var.: *#SeriesPurchases*; resid. SE = 0.571; 474 d.f.; $R^2$: 0.112; adj. $R^2$: 0.104; *F*-stat: 14.92 on 4 and 474 d.f.; *p* = 1.704e-11. Model: Linear with IV, estimated with 2SLS. 474 obs.; dep. var.: *#SeriesPurchases*; resid. SE = 0.581; 474 d.f.; $R^2$: 0.081; adj. $R^2$: 0.073; F-stat: 10.42 on 4 and 474 d.f.; *p* = 4.233e-8. Signif. as above.

Tables 2.8 and 2.9 offer empirical evidence of the positive impact of samples on the number of VoD series that households purchased. Those that viewed more free samples and paid samples ended up purchasing more VoD series. These results align with our main hypotheses: households that are more involved in the purchase decision, and more informed about the fit of VoD series dramas with their aggregate preferences will likely purchase more. In addition, we also wanted to see if the households' TV subscriptions influenced additional VoD purchases. As we expected, households that purchased more customized, add-on TV viewing content options beyond their basic TV

subscriptions were more likely to purchase VoDs series. These findings remained robust after we addressed the issues of heterogeneity and endogeneity. More importantly, our results offer the service provider a *directional reading on causality* between content sampling and on-demand purchases, after all of the other covariates were accounted for.

### 2.6.2. Household's Free Samples and Likelihood of Purchase for VoD Series

**Extended PSM for censored-data in a small dataset.** The non-censored data contained fewer observations than were desirable for empirical testing. At the household level, there were only 193 observations for which we had a full reading of free-sample, paid-sample and series-purchase activities, out of 30,006 observations in total. Thus, it was infeasible with this small a sample size to gauge the extent of a causal relationship between a household's decision to watch a free sample and then make a series purchase. So, we matched a censored observation to a non-censored observation based on their sequence of activities. Then, we inferred a behavior in the censored observation using the $90^{th}$ percentile of the distribution for the viewing pattern associated with all non-censored observations from that sequence of activities. The use of the 90th percentile of the distribution is appropriate based on our observation of the data. As more time goes by after watching a sample, the households were less likely to make a purchase, making the use of anything more than 90th percentile unnecessary. And yet, using anything less than 90th percentile would discard the households that needed more time to make their decision, as this data set conveys. As a result, we recovered 862 left- and 10,848 right-censored observations that were likely to have occurred just outside the study period. (Refer to Figure 2.4.)

**Figure 2.4. Overall Procedure to Recover Censored TV Viewing Observations**

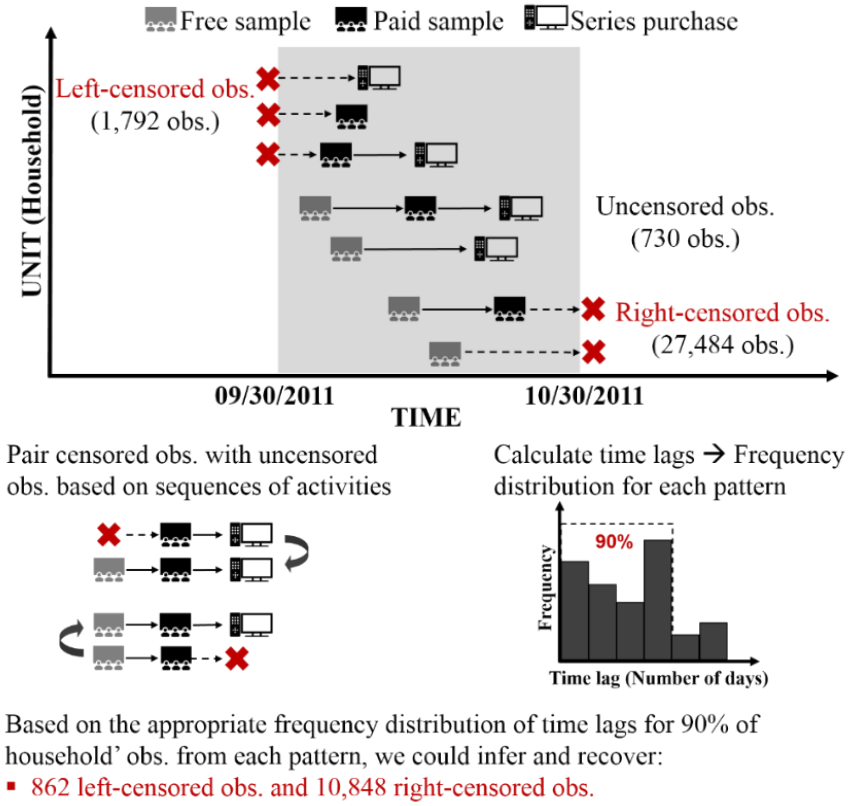



Table 2.12 reports the descriptive statistics for this new dataset. (See Table 2.12.)

**Table 2.12. Descriptive Statistics of the Dataset after Recovery of Censored-Data**

| VARIABLES | *HOUSEHOLDS WITH VODS AND SUBSCRIPTION INFORMATION* (8,939 OBS.) | | | | |
|---|---|---|---|---|---|
| | MEAN | SD | MIN | MEDIAN | MAX |
| *SeriesPurchase* (0/1) | 0.403 | 0.490 | 0 | 0 | 1 |
| $FreeSample_j$ (0/1) | 0.961 | 0.193 | 0 | 0 | 93 |
| $\#FreeSamples_j$ | 4.137 | 4.047 | 0 | 3 | 29 |
| $\#PaidSamples_j$ | 1.585 | 5.485 | 0 | 3 | 93 |
| $ContentClusters_j$ | 3.899 | 1.287 | 0 | 3 | 19 |
| $PremiumChannels_j$ | 3.198 | 3.093 | 0 | 2 | 25 |

The binary variables, $FreeSample_j$ (0/1) shows whether the household *j* had watched the free episode of series *i.* And $SeriesPurchase_j$ (0/1) indicates whether the household *j* had purchased the series *i.*

**Logit model results.** We used a logit model to estimate the effect of whether a household samples a series on the likelihood of its purchase of that series. The binary

dependent variable in this model is $SeriesPurchase_j$ (0/1). Beyond all the independent variables that are used in the count data models above, we added a binary independent variable, $FreeSample_j$ (0/1). This model tests for the direct effect of a series' free sample on the likelihood of a household's purchase of that series. The results from this model strengthened our findings above. (See Table 2.13.)

**Table 2.13. Logit Model Results**

| VARIABLES | COEF. | SE | Z-VAL. | $p$ (>\|$z$\|) |
|---|---|---|---|---|
| *Intercept* | -1.713*** | 0.114 | -14.970 | < 0.001 |
| $FreeSample_j$ (0/1) | 1.366*** | 0.104 | 13.092 | < 0.001 |
| $\#FreeSamples_j$ | -0.007* | 0.004 | -1.927 | 0.054 |
| $\#PaidSamples_j$ | -1.017*** | 0.003 | -5.632 | < 0.001 |
| $ContentClusters_j$ | 0.010 | 0.014 | 0.729 | 0.466 |
| $PremiumChannels_j$ | 0.002 | 0.006 | 0.401 | 0.688 |

**Notes.** Model: logit; 19,815 obs.; dep. var.: *SeriesPurchase* (0/1). Null dev.: 26,712; 19,814 d.f.; resid. dev.: 26,422; 19,809 d.f., pseudo $R^2$: 0.011, AIC: 26,434. Signif. as above.

The coefficient of *FreeSample* is positive and significant; so a household that sampled a series was more likely to purchase that series. This supports the direct relationship between a household's sampling and purchase for each series, which is our Household's Content Sampling Hypothesis (H1). Overall, a free sample of a series directly influenced a household's purchase decision of that series. It also positively influenced the household's decision to purchase other VoD series.

### 2.6.3. Sampling-based Strategy versus Outside Sources of Quality Information

To emphasize how content sampling may stimulate demand for series dramas, we examined the relationship with series-level analysis and considered all of the factors that we have mentioned. At the series-level, the coefficient for *TotalSamples* (i.e., the sum of *FreeSamples* and *Paid Samples*, to avoid high pair-wise correlation) was still positive and significant. *RentalPeriod* and *SeriesPrice* were not significant. (See Table 2.14.)

**Table 2.14. Negative Binomial Model Results: Series Level**

| VARIABLES | COEF. | SE | *z*-VAL. | *p* (>\|*z*\|) |
|---|---|---|---|---|
| *Intercept* | -0.842** | 0.751 | -1.122 | 0.262 |
| *TotalSamples*$_i$ | 0.001*** | 0.000 | 5.857 | < 0.001 |
| *RentalPeriod*$_i$ | 0.003 | 0.018 | 0.179 | 0.858 |
| *SeriesPrice*$_i$ | 0.031 | 0.022 | 1.421 | 0.155 |
| *Origin_HK* | 1.583*** | 0.438 | 3.611 | < 0.001 |
| *Origin_IDN* | -1.808** | 1.045 | -1.730 | 0.084 |
| *Origin_KOR* | 2.683*** | 0.756 | 3.546 | < 0.001 |
| *Origin_MYS* | 0.391 | 0.849 | 0.461 | 0.645 |
| *Origin_TWN* | 0.326 | 0.619 | 0.527 | 0.598 |

**Notes.** Model: Negative binomial; 79 obs.; dep. var.: *TotalPurchases;* baseline: *Origin_CHN*. Null dev.: 204.2; 78 d.f.; resid. dev.: 85.6, 70 d.f., pseudo-$R^2$: 0.136, AIC: 451. $\theta = 0.76$; degree of dispersion: $\alpha = 1/\theta = 1.32$. Signif. as above.

We also observed that Hong Kong and Korean dramas attracted more attention from Singaporean households. Thus, we looked at the Hong Kong series to explore the impact of outside quality information on the number of series purchases.

**Outside sources of quality information.** Among the series, there were 27 Television Broadcasts Ltd. (TVB) dramas with viewership and ratings from Hong Kong. The dramas aired in the years 2009-2011 in Hong Kong, before their availability in Singapore. Those more than 5 years old were excluded. *#NominAward* represents the number of nominations and awards that the series had received in the Hong Kong market; this is an indicator of drama series quality and the likelihood of success in the Singapore market. Popular series in Hong Kong were likely to have had a *spillover effect* due to the popularity of Hong Kong entertainment news and magazines in Singapore, and the interest that digital entertainment firms cultivated among Singaporean residents for Chinese-language content.

We note the difference in ethnic composition between the two markets; for example, Hong Kong's population is over 90% Chinese, while Singapore's is less than 75% Chi-

nese; and the education levels and income distributions are different. Thus, it is a reasonable, but not a perfect proxy for outside quality information. *1stEpiRating* is the observed ratings of the 1st episode of the series by Hong Kong viewers. We expected to see that *FreeSamples* and *1stEpiRating* had a positive effect on *TotalPurchases*, which would have provided additional evidence for the causal relationship between sampling and purchase. We employed a negative binomial model because the dependent variable *TotalPurchases* was dispersed: the number of purchases varied from one series to another. (See Table 2.15.)

**Table 2.15. Negative Binomial Model Results: Hong Kong Series Dramas**

| VARIABLES | COEF. | SE | *z*-VAL. | *p* (>\|*z*\|) |
|---|---|---|---|---|
| *Intercept* | 1.416 | 2.007 | 0.706 | 0.480 |
| $FreeSamples_i$ | $0.001^{***}$ | 0.000 | 3.458 | < 0.001 |
| $1stEpiRaing_i$ | 0.030 | 0.074 | 0.403 | 0.687 |
| $\#NominAwards_i$ | 0.054 | 0.045 | 1.202 | 0.229 |
| **Notes.** Negative binomial; 27 obs.; dep. var.: *TotalPurchases*. Null dev.: 55.9; 26 d.f.; resid. dev.: 29.5, 23 d.f., pseudo-$R^2$: 0.078, AIC: 235.7. $\theta = 1.467$; degree of dispersion: $\alpha = 1/\theta = 0.68$. Signif. as above. | | | | |

We assessed the effects of *FreeSamples*, *1stEpiRating* and *#NominAwards* for series *i* on *TotalPurchases*. The coefficient for *FreeSamples* for individual series was positive and significant, as content sampling played an important role in reducing household uncertainty concerning series fit. The coefficients for *1stEpiRating* and *#NominAwards* for individual series were not significant. Consumers seem likely to have purchased the series dramas that fit their viewing preferences and expectations, rather than those that they perceived as being of generally good quality.

### 2.6.4. Robustness Check Analysis for the Empirical Research Design

The main objective of this research has been to extend our understanding of entertainment content-service providers' sampling-based strategy in the context of digital

information goods. Causal inference with observational data still remained a challenge though we were able to access more than 17 million digital traces of households' viewing activities. This entertainment service provider and this dataset did not permit us to conduct a full test to infer causality in the manner we wished, since we had no control over the business setting. So, we took a divide-and-conquer approach to understand more deeply the causality relationship between content sampling and purchases in a scientific manner. First, the count data models were useful for understanding this dataset, allowing us to reach a general conclusion: over the one-month study period, the more samples a household watched, the more series dramas it purchased. We conducted a matching procedure to address selection bias due to household heterogeneity. And we addressed potential endogeneity with a Hausman test and a suitable instrument, as well as to increase our ability to claim the presence of a causal relationship.

An intriguing question remains: Did the households really purchase the same series that they had sampled? To address this question, we repurposed PSM to impute censored observations for a smaller dataset, but still the one that was entirely representative of our study's setting overall. This innovation provided us with a sufficient number of observations to analyze the direct impact of sampling on series purchases. We also accounted for series drama heterogeneity, and examined the relative effectiveness of content sampling versus outside quality information. Our findings indicate that the impact of series samples on purchases remained significant. Households were likely to purchase series dramas that fit their viewing preferences and expectations, rather than those that they perceived as being a generally good quality.

## 2.7. Discussion and Limitations

Our findings suggest that there is not just an association, but also a causal link between episode samples and series purchases. A household's free sample increases its likelihood to purchase the series. This suggests that sample content signals both vertical and horizontal differentiation on objective features. In addition, free-episode samples are effective in increasing the purchase conversion rate not only because they were made available to the customers; the customers actually watched the content to evaluate its fit related to their preferences. An additional free sample for a household caused a 19.8% increase in the number of series it purchased. This indicates that, for entertainment goods, customers also search and evaluate different alternatives before making a purchase. Watching free-episode samples is a faster and cheaper way for them. Thus, this action had a positive impact on series purchases in our study.

An important finding from this research is that an additional paid-sample episode led to a 9.4% increase in the number of series a household purchased. This seems counter-intuitive because purchasing individual episodes of a series will increase the transaction cost of buying the remaining content of that series. Yet this result aligns with our overarching theory in this research: customers are willing to pay to be more well-informed about the content they like to watch, and informed households will end up purchasing more series dramas. Several aspects of this research deserve further discussion, especially in terms of the business insights that they have to offer. Next, we discuss the implications for service providers for their use of sampling-based strategies.

### 2.7.1. Implications for Service Providers

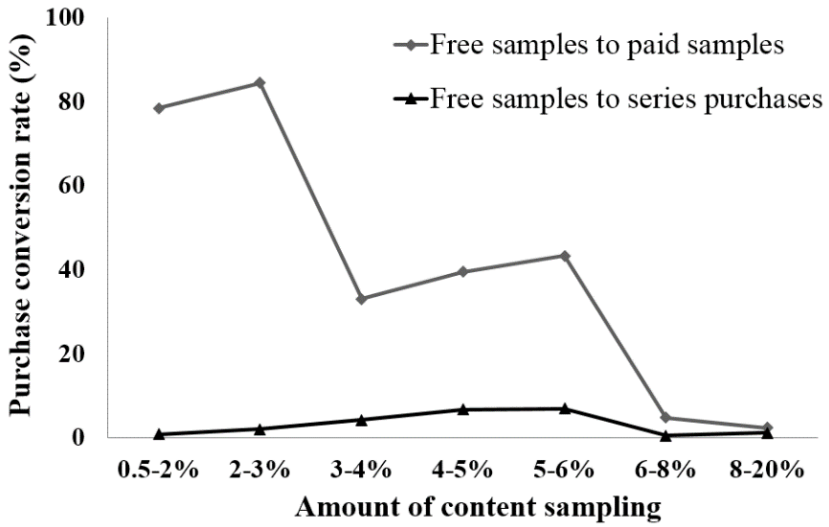
Omni-platform consumption and binge-watching of digital content have become

the new norms. Analytics with big data on consumers' digital traces also play a salient role in guiding business strategic planning. Our research contributes to the understanding of content sampling as a strategic marketing tool. It also raises important questions regarding more effective implementation of sampling-based strategy: (1) Is there an appropriate amount of content sampling that stimulates series purchases by households? Would it be easier to convince household's viewers to purchase a cheaper, shorter series after a single free episode? (2) Can a service provider influence consumer conversion rates for different types of TV series? If the service provider has limited screen space to advertise free-TV series episodes, should it promote a cheaper, shorter series or a longer, more expensive one?

To the best of our knowledge, none of the prior studies addressed the issue of how much free content is enough in the context of series dramas, largely due to other authors' limited access to data; thus, the most important problems have remained unsolved. We attempted to provide a sneak peek of some answers in this study. Across the households, in many cases, it was evident that one free-sample episode for a series was not enough for a purchase to occur. Service providers gain an additional stream of revenue from paid-sample episodes, however, it is not a desirable approach for everyone involved. Paid samples impose additional transaction costs for households, making VoD content more expensive. For example, even if a household sampled Episode 1 for free and then purchased Episodes 2 and 3 of a 10-episode drama, it still would have had to pay a fixed price for the seven remaining episodes. This may dissuade households from purchasing the series, creating a potential opportunity for the provider that would be missed.

The diverse nature of the series dramas in the datasets allowed us to examine the effect of the amount of sampling on household purchases, when the provider offered one free-sample episode for each series. The number of episodes in a series is a proxy for its price: the longer the series, the more expensive it is, and vice versa. (See Figure 2.5.)

**Figure 2.5. Conversion Rates by Amount of Content Sampled**




**Note.** Series dramas were sorted and aggregated based on their length in episodes term. So, a one-episode preview for a 20-episode series is 5.0%, for a 30-episode series it is 3.3%, and for just 6 episodes it is 16.7%. The *x*-axis values represent the average conversion rate of all dramas within a given range of the amount of content sampling in percentages.

For longer series in episodes terms, the conversion rate for paid samples also was high, while the conversion rate for series purchases was low. This suggests that a small portion of the sample content was not effective to stimulate series purchases, as households ended up purchasing many paid samples for additional viewing. So service providers, as a result, may wish to offer more episodes as free samples for longer series dramas. When one episode represents around 5% to 6% of the episode-length of a drama series, the conversion rate of the free samples to series purchases was at its maximum, suggesting that this amount may be sufficient to spark a household's interest in a series. Service providers apparently will not benefit from simply increasing the

number of free episodes, as our results suggest that the conversion rates for purchases quickly diminished for short series with a larger percentage of free content.

There are many possible explanations for this. When a household has watched a substantial portion of a drama series via free samples, the remaining portion will have become relatively more expensive, and a series purchase may be less attractive. Though our results only provide a glimpse into what really happened, the practical implications are important. Service providers should consider customizing their offerings of free samples and paid samples for different series dramas. An *appropriate amount of free sampling* is that amount that sparks a household's interest in a drama. An even more direct strategy is to offer a decreasing price scheme for the remainder of the series, encouraging more sampling and purchasing.

Another concern worth mentioning is that online piracy has taken a new form via illegal streaming services. It was estimated that there were over 141 billion visits across 200 million devices to the 14,000 largest piracy sites. According to the same source, music and TV series are at the top of all illegally-streamed content; streaming websites made up 73.7% of 78.5 billion visits to access pirated TV content in 2015 (BI Intelligence 2016). Offering content on an on-demand basis via legal streaming services has not been sufficient though: the rise of music streaming services has not killed music piracy (Dunn 2017). This poses a major challenge and, at the same time, presents a new opportunity for content producers and service providers. Firms must leverage new technology and proprietary data for understanding consumer behavior more deeply to improve their market offerings, and to do so in a way that consumers cannot benefit from when they obtain programming from other illegal streaming sources.

### 2.7.2. Research Design Issues

Even with an innovative research methodology coupled with a strong theoretical foundation across different disciplines, the limited coverage of our one-month of observational data hindered causal testing. This led us to shift our objective to *making inferences about important relationships that come close to true causality*, and at the same time, providing managerially important results. We formulated empirical testing models that worked well with the available data to make reasonable inferences about causality, based on appropriate theoretical background. The different count data models that we used, with one improving on another, addressed the specific characteristics of set-top box viewing data. In addition, the key variables that we selected for these models relate directly to the VoD business. Next, we adapted the PSM approach to handle selection bias. We also conducted a Hausman test and used an instrumental variable estimation to address endogeneity. Finally, our use of PSM to impute censored observations for the datasets allowed us to utilize more observations and achieve more convincing empirical test results.

Our research is unique in that we studied a specific area of digital goods, on-demand series dramas, very closely. Thus, our results may not be generalizable to other types of digital entertainment products. In addition, the study was done in Singapore, so it would be interesting to conduct a similar study in other markets, such as the U.S., where TV series play a major role in media consumption. An extension of this work also should consider a non-unitary model of the household to account for the differences among households whose average consumption preferences are similar, but whose individual members express different preferences (Rode 2011). How much free content

is appropriate to make available for sampling remains a question for researchers and managers alike, and open up new empirical research opportunities. We call for future studies that explore new marketing strategies for digital information goods, and to assess causality more thoroughly, by building on our method.

## 2.8. Conclusion

This research provides an empirical validation for the common wisdom that *information goods are experience goods* too, and giving the consumer a glimpse of the experience will be the most effective way to stimulate more purchases. Series dramas represent a major source of revenue for digital entertainment service providers, and the market for VoD drama series is unique for the application of sampling strategies to the consumption of digital information goods. This research is the first to provide empirical support for how episode sampling works in the context VoD drama series purchases. A free-sample episode of a series has a beneficial effect, by reducing a household's fit uncertainty for that series.

Even when a household's members know what they want to watch, they may need to sample other dramas to rule out any alternatives. Thus, a free sample of a series serves as a point of comparison for other series. Households with more customized content in their TV services are more likely to purchase VoD content, yet the number of content clusters that a household subscribed to apparently interferes with its VoD purchase intention. In addition, recognizing that a one-episode free sample will have different implications for dramas with various lengths in episode terms permitted us to gain insights on the appropriate amount of sampling that needs to be supported. Although households were willing to acquire paid samples to ensure that a series fit their

tastes, service providers should offer free samples more strategically, than on a common market-wide basis.

We emphasize that the main message is that a personal experience – Experience me! – is more influential than second-hand information for digital information goods sales to household consumers. With this in mind, service providers should invest more in marketing strategies that provide useful information about the fit of their digital goods with household preferences, since such strategies will help firms to reduce their marketing costs and increase sales and revenue performance in the long run. Another possibility is a decision support system that offers specific recommendations based on household viewing pattern matches on the households' TV screens. The will allow like-minded viewers to share their comments about their choices of VoD series with others. Equally important, digital entertainment service providers should implement incentive schemes that encourage viewers to watch more episodes and eventually make purchases, instead of looking for alternate sources of entertainment.

## Chapter 3: Censored Observation Recovery for Causal Inference

### 3.1. Introduction

Digital traces from consumer online activities present IS researchers with opportunities for research on the interplay between people and processes in the presence of IT when the insights extracted from such large-scale data would not be possible in traditional experimental research designs (Chang et al. 2014, Müller 2016). Data representing the digital traces of fine-grained consumer behavior differ from other types in several ways though (Martens et al. 2016). Some aspects of the required data are prone to not being entirely observable, even when the events of theoretical interest recur. In addition, personally-identifiable information for consumers must be masked and protected in compliance with privacy regulations (Chen et al. 2012). When this is the case, it may be impossible to identify and match individuals across data sources, when the same consumers, for example, are involved during the same period of time.

Data limitations often undermine researchers' efforts to explore meaningful relationships in the data, as a result. For instance, what can be learned about how online users interact with one another to form friendships in social networks? Or to what extent do shoppers search for multiple products before making a purchase decision? These research questions can be answered more effectively when relatively complete consumer data are available. This frequently is not the case though (Bapna and Umyarov 2015, Bapna et al. 2016).

In the last decade, large datasets have become more accessible to researchers, including proprietary consumer data collected by firms and public data collected by government agencies (Chang et al. 2014). There is also a large amount of sensor data now

being produced by Internet of things (IoT) devices (Galer 2017). This has prompted IS researchers to expand their capability to work with very large, but less-than-ideal datasets, and to integrate digital trace data into research designs to support causal inferences in new and wickedly complex contexts (Bareinboim and Pearl 2016, Howison et al. 2011, Ketter et al. 2016).

An observation is *censored* when one or more records on a subject are not observable, because they occur before or after the research timeframe, but are necessary to complete logical causal sequences that lead to some theorized outcomes. This poses a statistical challenge for econometric modeling-based empirical research designs, when unobservable data limit the analytical construction of a causal relationship. A subtler implication is that, for studies that employ some form of *aggregated data* over time, the parameter estimates produced by empirical models that use big data approaches still will suffer from a common lack of "necessary data" to support the desired causality tests. We refer to this problem as "lost needles in a digital haystack of data."

In this research, we explore whether the empirical regularities found in a dataset can be used to complete the unobservable logical sequences of consumer behavior and improve its informedness for causal inferences? And will a more informative dataset support research designs to extract meaningful relationships? We propose a context-specific probabilistic inference method that can enhance the statistical power of a censored dataset for causal inference in data-driven exploratory research. It improves the sequential completeness of the dataset via econometric imputation of data outside the study period. Our iterative data simulation design allows us to evaluate and examine the performance of the proposed method. The variability of the estimates produced

from the simulated datasets allows us to assess the generalizability and robustness of the method in computing censored customer-level observations to support causal inference. [5]

This method is motivated by our early work in Hoang and Kauffman (2016), in which we examined the effectiveness of household VoD series samples on its subsequence purchase. Nevertheless, the events of interest were censored, before and after the 1-month observation period. As these censored observations were essential to our causal testing framework, we extended the propensity score matching (PSM) procedure to match similar observations, and then imputed censored records based on a probabilistic model.

In more detail, we paired observations across censored and uncensored data groups based on discoverable sequences and patterns of observable past activities (Read at el. 1989), in order to establish similarity between observations from both groups. Then, some values of variables for the censored observations were able to be inferred, by leveraging the richness of the entire dataset. The notion here was that, for settings in which the event of interest recurs, observations can be matched and computed based on relevant behaviors related to them. In contrast to typical PSM methods that are used with large samples (Dehejia and Wahba 2002), our approach involved the repeated estimation of a censored dataset, using an *exact one-to-one matching algorithm with iterative replacement* (Pirracchio 2012). This process yielded more observations with

[5] The readers should note that the word *observation* refers to the customer level, which includes all records of a customer. Thus, a *censored observation* refers to a customer-level observation, in which some records are not observable as they lie outside the observational period of study. This is similar to the naming convention that has been used across healthcare and epidemiology research. An observation of a patient is censored when some of her records are not observed outside the study period. Our method focuses on the recovery of censored records for customer-level observation, to complete her sequence of activities.

appropriate statistical properties for empirical testing to support causal arguments.

In the second application of this method, we analyzed a 4-month non-censored dataset that includes the complete online journeys of more than 30,000 consumers on a European e-retailer's website. All consumers visited and made at least one purchase within the study period, and thus all of them were converted customers from searching to purchasing. We explored the causal relationship between a customer's sequence of visits to a website and her purchases. Different traffic sources can drive the consumer to a website, but some of are more effective in informing her about the retailer's products. The full information on consumer behavior in the dataset gave us the flexibility to showcase the use of our method. First, we simulated multiple censored subsets from the original non-censored data, by using its data-generating process (Davidson and Mackinnon 2006). Different levels of data censoring in each subset affect the ability to construct and make causal inferences on a causal relationship.

Next, we established evidence for the presence of empirical regularities in the data and used these values to impute censored records for each data subset. By assessing the variations in our econometric estimates from different censored subsets and those with some imputed values, we showed how our proposed method improved the statistical power from empirical testing. This process provides a better understanding of the value of information gained based on recovering data on the sequences of consumer actions.

Section 3.2 discusses the cross-disciplinary methodological background relating to data censoring and data recovery for causal inference, including the matching methods, dataset comparison techniques, and information gain and value. Section 3.3 introduces

our context-specific probabilistic inference method and the evaluation approach. Sections 3.4 and 3.5 showcase two applications of the proposed method in imputing censored viewing activities for household consumers and censored website visits and purchases for online consumers. Section 3.6 concludes and discusses future development of the method.

## 3.2. Methodological Background

### 3.2.1. Data Censoring and Causal Inference

Missing data often occur due to three mechanisms. *Missing completely at random* (MCAR) data are those for which there is no relationship between the missing data and any other values in the dataset, whether that data is actually missing or just unobserved. *Missing at random* (MAR) data refers to cases where the propensity for data to be missing is unrelated to other missing data but is correlated with some other observable data. Finally, missing data that do not share the characteristics of MCAR or MAR are referred to as *missing not at random* (MNAR). Incomplete data create a roadblock for establishing a solid foundation for causal inference, especially in cases of MNAR.

IS researchers often encounter missing data due to unmatched records from multiple sources. Even though they may start with "big data," the data relevant for answering meaningful research questions are often limited (Bareinboim and Pearl 2016). Data sample size has important implications for statistical significance testing, though the specific size required for adequate statistical power depends on the research objectives and the complexity of the empirical model (Maloney et al. 2010).

A form of missing data, *censoring*, refers to when the observed values for a variable

are only partially known. Censored data refer to observations with records that lie outside the observation period and cannot be observed. Studies in healthcare and medical epidemiology that dealt with disease refer to the end date of the observation period as a *point of data censoring* or *data censoring date* as the patients' conditions cannot be observed after that date (Holmes et al. 2008, Prentice and Gloeckler 1978); this is referred to as *right censoring* (Chintagunta and Dong 2006). And in longitudinal studies of developmental and disease processes, *left censoring* occurs if a participant joined but the *event of interest* occurred prior to study entry but its timing is unknown (Cain et al. 2011). In clinical trials, the withdrawal of patients and non-responding customers often cause observations of the primary variable to be lost (Wu and Carroll 1988).

Censoring poses a statistical challenge for econometric modeling-based empirical research design, though all datasets often have some degree of left and right censoring. It reduces the external validity and statistical power of the related empirical model and creates biased hypothesis test results. For studies that employ some form of aggregated data over time, the parameter estimates produced by empirical models that use big data approaches also suffer from a common lack of data that are necessary to support the desired causality tests. (Newman 2009, Newman et al. 2009).

### 3.2.2. Methods for Tackling Censored Data Within the Observation Period

A common approach in dealing with missing observations in duration modeling in Labor Economics, for example, is to use the duration of the observation period to approximate the time length for the duration until an event occurs (Chintagunta and Dong 2006). In Marketing research, censored observations typically are discarded; researchers only account for events of interest that occur during the study period ended (and

sometime before it). Researchers in different disciplines have used various computational approaches (Schafer 1999, Wei and Shih 2001) to deal with missing data. Their objective is to fill in missing data with values based on a model with assumptions. Elaborate procedures, such as bootstrapping (iterative resampling), are used to handle right-censored data (Efron 1981, Gross and Lai 1996). These methods help to complete the dataset, so all of the statistical tools for complete data can be applied (Shih 2002).

There are some concerns though. Discarding censored data substantially reduces dataset size. And using imputed values or generating new data via a resampling procedure often produces biased values in a dataset, which may influence the researcher's ability to assess the treatment effect in empirical testing. These approaches do not address left-censored observations at all, and they are less effective in handling small datasets. Still, this should not be viewed as a shortcoming but as an opportunity for a methodological advance to recover censored observations for causal inference. Table 3.1 provides a summary of current approaches to handle censored data, along with their implications and limitations. (See Table 3.1.)

**Table 3.1. Approaches to Handle Censored Data**

| APPROACH / DESCRIPTION | IMPLICATIONS | LIMITATIONS |
|---|---|---|
| **Partial deletion.** Censored or incomplete data discarded | • Reduces sample size; biased sample with discarded data | • Left-censored data not addressed<br>• Less useful, small data<br>• Reduces usable data for empirical testing<br>• Produces biased sample, influences treatment effect<br>• Ignores meaningful censored data |
| **Last observation carried forward.** Events recur; ending values are true | • Conservative; underestimates treatment effect in right-censored data | |
| **Proper multiple imputation.** Events recur; regression imputes data; missing data still | • Works with missing at random data; biased if don't have all missing data | |
| **Partial imputation.** Identifies data that balance patterns in treatments | • Produces balanced biased data, but underestimates treatment effect | |
| **Bootstrapping.** Resampling with uncensored data; inference possible | • Only uses uncensored data, but loses insights from censored data | |

### 3.2.3. Matching Methods and Causal Inference in Empirical Studies

Causal inference is the central goal of many empirical investigations. A *causal effect* is defined as a comparison between the potential outcomes for a *randomized treatment group* and a *control group*, averaged over a population (Rosenbaum 1989). Bias can arise due to systematic differences between the groups, for example, due to *selection bias,* when the study population is not randomly selected from the targeted sample. *Propensity score matching* (PSM) addresses this bias by pairing the treated and controlled observations that are similar in some observable characteristics (Rosenbaum and Rubin 1983, Rubin 1973).

The objective of all matching procedures is to reduce the bias in the association between the treatment and the outcome, due to the imbalance in the covariates that may affect the outcome (Li 2016, Oestreicher-Singer and Zalmanson 2013). Other standalone, automated, data-driven methods (e.g., the *tree-based approach*), may outperform PSM in adjusting for observable self-selection bias, particularly for large sample sizes with many variables (Yahav et al. 2016).

### 3.2.4. Information Gain and Value

Bayesian approaches have achieved popularity across various research disciplines. The exponential growth of data and the development of computational methods have drawn attention to potential applications of Bayesian methods, especially how newly-arriving information may change an analyst's estimates (Ibrahim et al. 2001). There are two components in Bayesian inference approach: a *gain function* (or a *loss function*, as it is often referred to in the Bayesian methods literature) and a *posterior distribution*. A *gain function* associates an outcome with a state of nature and an action, $G(a, \theta)$,

where $a$ is the action and $\theta$ is the state of nature. Thus, a rational decision-maker chooses an action to maximize the expected gain, where the expectation is taken with respect to the posterior distribution (adapted from Rossi and Allenby 2003):

$$\min_{a} \bar{G}(a) = \int G(a,\theta) Pr(\theta|data) d\theta$$

This approach supports making inferences, or posterior statements about what is still unobservable based on what had been observed. The likelihood-based approach relies heavily on prior information, which can be subjective or objective based on data. The ability to make inferences that support the recovery of information from censored data in a meaningful manner offers important implications for the kind of research we are doing based on the Bayesian framework (Yap at el. 2008). Empirical regularities extracted from the observed data can be used to make inferences on the posterior, or the conditional distribution of the unobserved data. The goal here is to maximize the information gain from obtaining additional data, which is also associated with an improvement in the ability to make causal inferences.

In decision-making, the *value of information* is measured by the difference in the outcome between the decision made with and without that information (Demski 1980). This provides a benchmark to determine whether it is worthwhile to obtain additional information. It is widely acknowledged that *perfect information* is typically worth no less and possibly more than *imperfect information*; it is not clear though, how to assess the marginal gain from the acquisition of additional information when data censoring is considered. In empirical studies with observational data, the value of additional observations is often overlooked, due to the availability of the required data (Bareinboim and Pearl 2016).

### 3.2.5. Quantitative Conceptual Distances between Datasets

*Distance* is a very important concept across many fields in Computer Science, from information retrieval, data mining, pattern recognition to machine learning. It provides the basis for evaluating the similarity or assessing the difference between data objects. Many distance functions have been proposed in the literature to measure the similarity between two finite sets of points in a metric space. We summarize the common distance measures and their implications (Eiter and Mannila 1997). (See Table 3.2.)

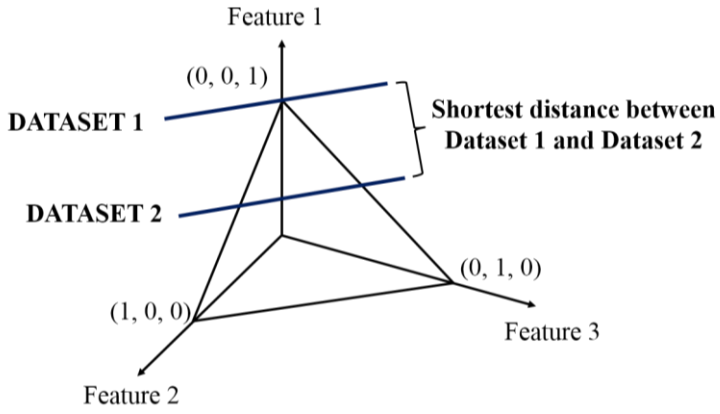
**Table 3.2. Common Distance and Similarity Measures for Point Sets**

| DISTANCE FUNCTION | DESCRIPTION AND IMPLICATION |
|---|---|
| **Hausdorff distance** | • Distance between two subsets in metric space<br>• Calculate distance by taking only the most distant objects of each set into account: two subsets are close if every point of either set is close to some point of the other set |
| **Sum of minimum distance** | • Take into account distances between each element and the other set<br>• For every element in the first set, the closest element in the other set is considered |
| **Surjection distance** | • Measure distance between two sets using surjections that map the larger set to the smaller set: every object in the larger set is mapped to some object in the smaller set<br>• Can be minimized using pairwise distance values computed by comparing individual objects |
| **Fair surjection distance** | • Variant of the surjective distance measure<br>• Overcome unsatisfactory behavior of the surjection measure<br>• Impose that admissible surjections must be fair: map the larger set evenly onto the smaller set |

A holistic approach to compute the distance between data subsets can simplify the task of understanding complex data collection, as industry data are often incomplete and come in many forms. For instance, researchers in Marketing and IS often deal with multiple subsets of consumer data from different markets, and different timeline. In many research settings, data collections may be naturally divided into several data sets (Tatti 2007).

The *constrained minimum* (CM) *distance* (Tatti 2007) between datasets $D_1$ and $D_2$ with feature function, *S* is defined as: $d_{CM}(D_1, D_2 | S) = \sqrt{|\Omega|}\, \|u_1 - u_2\|_2$, for a set of samples in a finite *sample space* Ω. With the feature function *S*, then $C(S, D_i)$ is the constraint set for $D_i$, and $u_i = \underset{u \in C(S,D_i)}{\operatorname{argmin}} \|u\|_i, \ i = 1, 2$ is a vector from each constrained space with the shortest norm. This distance is based on the observed frequencies of the feature function *S* from its overall distribution. This specification of distance allows us to compare data sets based on a user-selected set of features, or the summaries statistics computed from the data sets. Figure 3.1 provides a simple Euclidean geometrical representation of the shortest distance between datasets when three features are considered. (See Figure 3.1.)

**Figure 3.1. Representation of Distance Between Datasets**

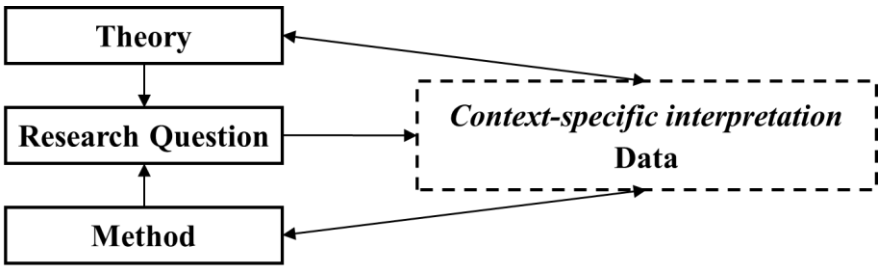



## 3.3. Context-Specific Probabilistic Inference Method

### 3.3.1. Motivation

A better understanding of the data, combined with a strong theoretical background and a suitable methodology framework enables researchers to answer their research question thoroughly. Thus, there are bi-directional relationships between data and theory, and between data and method. (See Figure 3.2.)

**Figure 3.2. Data-driven Explanatory Research Framework**




This methods innovation aims to complete logical sequences of data and improve their informedness to support causal inferences in theory-focused empirical research. In particular, it enables us to address data censoring – a crucial problem in empirical investigations across different fields. We recover records of activities that may have occurred before and after the data observation period but are needed to complete the logical sequences of consumer or other human behavior. This is different from other methods that try to impute missing data within the observation period.

### 3.3.2. Overview of the Context-Specific Probabilistic Inference Method

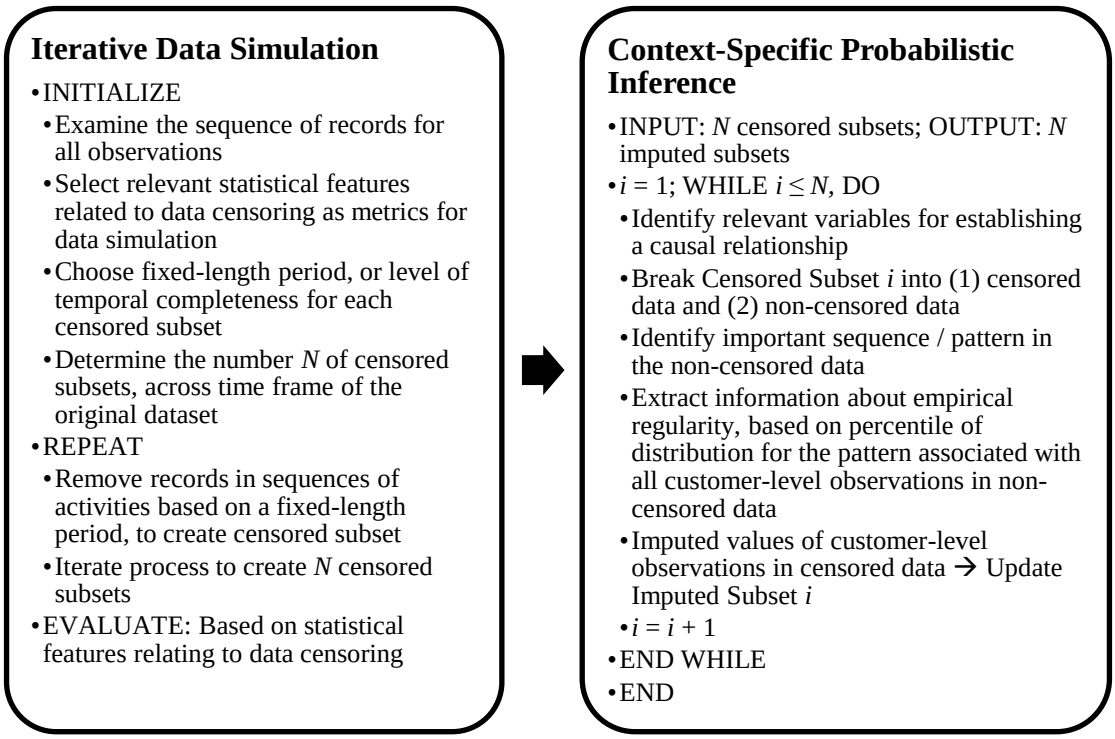
A purchase history of a given customer is a good example for the problem of censored customer-level observations in *data with logical sequences*. Complete observation of the customer is not available, as we cannot observe her activities before and after the study period. Nevertheless, those records are essential for making inferences about her purchase decision (Shih 2002, Wu and Carroll 1988).

We address this problem via *econometric imputation* of the *logical values of censored records* outside the observation period, based on the statistical matching of observable patterns and sequences for all observations (Davies and Bouldin 1979, Lenis et al. 2017) with those observations for which some records are censored. This is similar to sequential pattern mining: researchers focus on finding statistically-relevant patterns in a data sample where the values are arrayed in a temporal sequence. We propose

that the *empirical regularities* – derived by the analyst as the related facts associated with the data that can be observed – in the patterns of activities are used as a basis for the imputation. Prior research in the IS literature explored the underlying regularities in different empirical research settings, such as Ayal and Seidman (2009) and Bergen et al. (2005). Other works in Marketing, Platzer and Reutterer (2016) and Uncles et al. (1995), examined empirical regularities in customer activity and purchase patterns.

Our method is context-specific, as the relevant patterns and sequences depend on the context in which the data are generated. The *temporal completeness* of the dataset (Cooper 1990, Ross 1988) provides a basis for testing *causal sequences*, which will strengthen empirical evidence on the causal relationships over time. It also supports *causal prediction* over real-time event data. (See Figure 3.3)

**Figure 3.3. Pseudocode for Iterative Data Simulation and Probabilistic Inference Method**



### 3.3.3. Evaluation of the Method

Observational and descriptive methods of evaluation are suitable for our proposed method, as other forms of evaluation are not feasible (Hevner et al. 2004). Our method yields utility for addressing the problem of data censoring, as it supports the imputation of censored records outside the study's observation period. In addition, it provides clear contributions not only in the e-commerce business context, but also in other IS contexts to infer causal relationships from datasets with incomplete behavior sequences.

In the second application, we evaluate the performance of our method and show its robustness via a series of simulated data subsets. As each *censored subset* has different proportions of left-censored, right-censored and left- and right-censored customer-level observations, we assess their level of sequential completeness by looking at context-specific features relating to data censoring. Then, to evaluate the method's performance with respect to the information value improvement for causal inference, we construct a hypothesis testing framework, which allows us to extract a reading on the *incremental statistical information gain* for asserting the presence of causality (Cooper 1990, Yap at el. 2008).

## 3.4. Application 1: Recover Censored Records for Household-level TV Viewing Data

We have been examining the effectiveness of sampling strategy on household VoD series drama purchases (Hoang and Kauffman 2016). The event of interest recurs (Schaubel and Cai 2006); a household is associated with more than one viewing or purchasing activities during the one-month study period. The provider offers single free-episode samples of all dramas. After previewing the free sample, a household can

purchase subsequent episodes, including paid samples for additional previewing. Alternatively, it can purchase the discounted series.

To examine the extent to which households purchase the same series that they sample, we looked at households' viewing activities for each series. Nevertheless, we did not have sufficient data to conduct a full analysis focusing on household behavior, nor did we have enough months in a time-series of observations. We also encountered a *data-censoring issue* for free-sample and series-purchase sessions during the time window. Thus, we used *propensity score matching* (PSM) to address data censoring in a small dataset, when the events of interest recur during the study period.

Our approach involves pairing observations across censored and non-censored-data groups based on their discoverable sequences and patterns of observable past activities, so that censored records in some observations can be imputed. In contrast to typical PSM use with large samples (Dehejia and Wahba 2002, Gemici et al. 2012), our approach uses an exact one-to-one iterative replacement matching algorithm (Pirracchio et al. 2012). Thus, we can address both left- and right-censored data in a small dataset, by utilizing the richness of insights from the censored-data.

First, censored observations were matched to suitable sequences of viewing records represented by different non-censored observations. Using exact timestamps available in the dataset, we calculated the time lags, or the difference in dates for free-sample, paid-sample and series-purchase sessions for each sequence in the non-censored data category. Then, we identified the $90^{th}$ percentile of distribution for the viewing patterns associated with each of the sequences, and used this value to make inferences about the unobserved values of the censored data. In the next section, we explain in detail how

we used this method.

### 3.4.1. Evaluation of Data Censoring in the VoD Dataset

A total of 39,518 viewing sessions can be classified into four categories. The *left-censored data* category includes household-level observations of VoD viewing for which we observed either a combined paid-sample and series-purchase session, or a series-purchase session alone, but when a free-sample session was not observed. The *right-censored data* category includes household-level observations of VoD viewing for which a free sample occurred, with or without a paid sample being purchased; but a series purchase was not observable. The *left- and right-censored data* category only includes household paid-sample sessions. It is not known whether a household sampled or purchased, since these actions would have occurred outside the study period. In the *non-censored data* category, we observed that a free sample and a series purchase occurred, with or without a paid sample having been purchased; the data in this category are very small compared to the censored categories though. Table 3.3 presents the breakdown of the four data categories. (See Table 3.3.)

**Table 3.3. Data Censoring Issues for Household Observations**

| Censored-Data Category | # Obs. | VoD Session Observations | | | Explanations for Unobserved Behavior |
|---|---|---|---|---|---|
| | | FS | PS | SP | |
| **Left-Censored Data** | 337 | × | × | ✓ | Free sample (FS) may have occurred before study period; though no paid sample (PS) was observed, a series purchase (SP) occurred. |
| | 18 | × | ✓ 1 / series | ✓ | FS may have occurred before study period; then a PS and a SP were made. |
| | 54 | × | ✓ Multiple / series | ✓ | FS may have occurred before study period; multiple PS and a SP were made. |
| **Non-Censored Data** | 537 | ✓ | × | ✓ | FS occurred; no PS, and a SP was made. |
| | 50 | ✓ | ✓ 1 / series | ✓ | FS occurred; PS and SP were made. |
| | 143 | ✓ | ✓ Multiple / series | ✓ | FS occurred; multiple PS and a SP were made. |
| **Right-Censored Data** | 26,659 | ✓ | × | × | FS occurred; no PS, a SP may have occurred after study period. |
| | 315 | ✓ | ✓ 1 / series | × | FS occurred, PS was purchased; a SP may have occurred after study period. |
| | 510 | ✓ | ✓ Multiple / series | × | FS occurred, multiple PS purchased; a SP may have occurred afterward. |
| **Left- and Right-Censored Data** | 616 | × | ✓ 1 / series | × | FS may have occurred before study period; PS was purchased; SP may have occurred after study period. |
| | 767 | × | ✓ Multiple / series | × | FS may have occurred before study period; multiple PS were purchased; SP may have occurred afterward. |

**Notes**. SP denotes Free Sample, PS denotes Paid Sample, SP denotes Series Purchase. A household can have different kinds of censored observations, with respect to different series. For instance, a household that belongs to *left-censored data* for Series A may also have *non-censored data* for Series B.

We assessed different approaches regarding their abilities to handle this issue and establish causal linkage. (See Table 3.4.)

**Table 3.4. Ability to Establish Causal Linkages for Household-level TV Viewing Records**

| APPROACH | IMPLICATION ON THE DATASET | ABILITY TO ESTABLISH CAUSAL LINKAGE? |
|---|---|---|
| Partial deletion | Discard left-, right-censored obs.; 730 uncensored obs. | Bias, over-estimated effect of free samples; all households in uncensored category made purchase |
| Last obs. carried forward | Discard left-censored obs.; 730 uncensored; 27,484 right-censored. | Bias, under-estimated effect of free samples; all households in right-censored category did not purchase |
| Proper multiple imputation and partial imputation | No basis to compute unobserved viewing records | Infeasible |
| Bootstrapping, by resampling data with replacement | Can't resample w/ uncensored obs. | Infeasible; more obs. generated from uncensored obs. do not have desirable characteristics for causal test though |

In the *non-censored data* category, all free-sample sessions correspond with series-purchase sessions, thus it is infeasible to gauge the impact of the free samples. Furthermore, out of the voluminous original dataset, there are only 730 observations in the *non-censored data* category; this is too small for empirical testing. This forced us to use the new approach to capture censored observations. First, we looked at all sequences of activities for each household observation in the *non-censored data* category, and identified three main sequences: (1) free sample (FS) → series purchase (SP), (2) free sample (FS) → paid sample (PS) → series purchase (SP), (3) free sample (FS) → multiple paid samples (PS) → series purchase (SP). Table 3.5 shows sequences of VoD viewing records. (See Table 3.5.)

## Table 3.5. Examples of Different Sequences of Household Observations in the Dataset

| Household ID | Sep 30 | Oct 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Left-censored data: series purchase (SP) only** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #434 | | | | | | | | | | | | | SP | | | | | | | | | | | | | | | | | | |
| #055 | SP | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #699 | | | | | | | | | | | | | | | | SP | | | | | | | | | | | | | | | |
| **Left-censored data: paid sample (PS) → series purchase (SP)** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #541 | | | | | | | | | | | | | | **PS** | **PS** | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | SP | | | | | | | | | | | | | | | | |
| #109 | | | | | | | | | | | | | | | | | **PS** | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | SP | | | | | | | | | | | | | | |
| #369 | | | | | **PS** | SP | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Non-censored data: free sample (FS) → series purchase (SP)** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #033 | | | | | | | | | | | | | | | FS | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | SP | | | | | | | | | | | | | | | | |
| #394 | FS | SP | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #644 | | | | | | | | | | | | | | FS | | SP | | | | | | | | | | | | | | | |
| **Non-censored data: free sample (FS) → paid sample (PS) → series purchase (SP)** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #712 | | | | | | | | | | | | | | | | | | | | | | | | | | | FS | PS | SP | | |
| #339 | | | | | | | | FS | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | PS | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | SP | | | | | | | | | | | | | | | | | | | | | | | |
| #786 | | | | | | | | | | | FS | PS | | | SP | | | | | | | | | | | | | | | | |
| **Right-censored data: free sample (FS) → paid sample (PS)** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #406 | | | | | | | | | | | | | | | FS | PS | | | | | | | | | | | | | | | |
| #726 | FS | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **PS** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #570 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | FS | **PS** | |
| **Right-censored data: free sample (FS) only** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #051 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | FS | |
| #112 | | | | | | | | | | | | | | | | | | | | | | | | FS | | | | | | | |
| #218 | | | | | | | | | | FS | | | | | | | | | | | | | | | | | | | | | |
| **Left- and right-censored data: paid sample (PS) only** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| #112 | | | | | | | | | | | | | | | | | | | | | | | | | | | | PS | | | |
| #076 | | | | | | | | | | | | | | | | | | | | | | | | PS | | | | | | | |
| #394 | | | | | | | | | | | | | | | | | | | | PS | PS | | | | | | | | | | |

**Notes.** Households purchased multiple paid samples of the same series on the same day. This is denoted by the bold **PS**.

### 3.4.2. Imputation of Censored Records in the VoD Dataset

Each observation in the *censored-data* category was mapped to the most likely sequence in the *non-censored data* category, based on sessions that were observed. For instance, an observed series purchase in the *left-censored data* category was matched to the free sample (FS) → series purchase (SP) sequence in the *non-censored data* category. Finally, we used the dominant pattern in each *non-censored data* sequence as a threshold to infer the unobserved activities. Let say the dominant pattern in the free sample (FS) → series purchase (SP) sequence is within 3 days. So, if we only could have observed a household's series purchase, but nothing before within that 3-day threshold, we inferred that the household did not watch any free sample. However, if the 3-day threshold lies outside the study period, we inferred that the household had watched a free sample; but that activity was not observable nonetheless.

The assumption we made here was that a household which followed a viewing sequence of a category was likely to lie within the $90^{th}$ percentile of the distribution for viewing pattern associated with that category. Table 3.6 illustrates our matching approach using discoverable viewing sequences and patterns to impute censored records. With the *left- and right-censored data* category, we only tried to infer the unobserved free-sample sessions, as the households were much more likely to have watched the first episode of a series before the $2^{nd}$, $3^{rd}$ or any other subsequent episode. We recovered 862 free-sample sessions and 10,848 series-purchase sessions. (See Table 3.6.)

**Table 3.6. Data Censoring Issue for Household-Level VoD Series Viewing Observations**

| Data Category | Sequences of VoD Sessions for Household-Level Viewing Obs. | Patterns of VoD Sessions for Household-Level Viewing Obs. | Matching Approach |
|---|---|---|---|
| Left-censored Data 1 (LD1) | → SP | | Match with ND1 → Infer behavior in censored FS session at 90th percentile of distribution for viewing pattern associated with ND1 |
| Left-censored Data 2 (LD2) | PS → SP | | Match with ND2 → Infer behavior in censored FS session at 90th percentile of distribution for viewing pattern associated with ND2 |
| Left-censored Data 3 (LD3) | Multiple PS → SP | | Match with ND3 → Infer behavior in censored FS session at 90th percentile of distribution for viewing pattern associated with ND3 |
| Non-censored Data 1 (ND1) | FS → SP | Time lag between FS and SP → Note period when 90% of households moved from watching FS to a SP. | |
| Non-censored Data 2 (ND2) | FS → PS → SP | Time lag between FS and PS → Note period when 90% of households moved from watching FS to purchasing PS.<br>Time lag between PS and SP → Note period when 90% of households moved from purchasing PS to SP. | |
| Non-censored Data 3 (ND3) | FS → Multiple PS → SP | Time lag between FS and earliest PS → Note period when 90% of households moved from watching FS to purchasing PS.<br>Time lag between latest PS and SP → Note period when 90% of households moved from purchasing PS to SP. | |
| Right-censored Data 1 (RD1) | FS | | Match with ND1 → Infer behavior in censored SP session at 90th percentile of distribution for viewing pattern associated with ND1 |
| Right-censored Data 2 (RD2) | FS → PS | | Match with ND2 → Infer behavior in censored SP session at 90th percentile of distribution for viewing pattern associated with ND2 |
| Right-censored Data 3 (RD3) | FS → Multiple PS | | Match with ND3 → Infer behavior in censored SP session at 90th percentile of distribution for viewing pattern associated with ND3 |
| Left- and Right-censored Data | PS | We only tried to impute the left-censored data. We did not try to recover the right-censored data as there was not enough supporting evidence for inferences. | Match with ND2 → Infer behavior in censored FS session at 90th percentile of distribution for viewing pattern associated with ND2 |
| | Multiple PS | | Match with ND3 → Infer behavior in censored FS session at 90th percentile of distribution for viewing pattern associated with ND3 |

**Evaluation of performance.** The performance of this method is based on the extent to which it can recover unobserved data to support causal testing. The additional observations are used as a basis to strengthen the relationship between the free samples and the corresponding series purchases; this was not feasible before. Nevertheless, the underlying assumption is that censored data mimics the distribution of the non-censored data, since we used 90% of the observations in the censored-data category as thresholds to infer censored observations. In the next application, we explored a more rigorous framework to evaluate the strengths and weaknesses of this method.

## 3.5. Application 2: Recover Censored Records for Customer-level Online Shopping Data

### 3.5.1. Research Setting and Data

In this section, we show how our method works with a less-than-ideal dataset and enables more insights on meaningful relationships from data. We used the clickstream data of a European e-retailer representing a period of four months in 2013. The dataset includes all customer journals leading up to a purchase on the retailer's over a 4-month period in 2013. All customers made at least one purchase in this time period. We refer to this as a *non-censored dataset*: all visit records and purchase records of the visitors are included. This dataset allows us to design a data simulation strategy that creates a series of censored subsets, in order to assess our data recovery method.

The company collects all of its customers' online activities and uses them to develop a conversion attribution algorithm for its client retailers. There are many online traffic sources that lead a customer to the retailer's website, each of which provides different types of product information to the customer. In this work, we examine how

these sources influence a customer's likelihood of making an online purchase.

### 3.5.2. Customers' Visits and Purchase Activities in the Non-Censored Dataset

We provide definitions of all relevant variables related to customers' visits and purchase activities, as well as the characteristics of these visits at the dataset level and customer level. (See Table 3.7.) We used the variables at the dataset level to assess how the simulated subsets are different from others, relating to the degrees of data censoring.

**Table 3.7. Variable Descriptions at the Dataset Level and Customer Level**

| VARIABLES | DEFINITIONS |
|---|---|
| **At the Dataset Level** | |
| *#Cust* | # customers in dataset |
| *#AllVis* | # visits of all customers to retailers' site |
| *#Purch* | # purchases that all tracked customers made |
| *AvgVis* | Average # visits by each customer |
| *AvgPurch* | Average # purchases by each customer |
| *Conver* | Conversion rate of customers' site visits to purchases |
| *#NonCensor* | # customers with complete visits and purchase records |
| *#VisCensor* | …… censored visit records |
| *#VisPurchCensor* | …… censored visit and purchase records |
| *%VisCensor* | % visit-censored customer observations |
| *%VisPurchCensor* | % purchase-censored customer observations |
| *#RecovPurch* | # customers with purchase records recovered |
| **At the Customer Level** | |
| *Purch* | Whether customer made purchase (1 = one purchase or more; else 0) |
| *#Vis* | # times customer visited site |
| *#Direct* | # times customer visited site by typing in retailer's domain; traffic generated directly |
| *#SEngVis* | # times customer visited site via organic and advertising search engine search |
| *#AdsVis* | …… via affiliated marketing, display ads, and other sources |
| *#PersonVis* | …… via unbiased comparison site, email advertising and social advertising |
| **Variables to Impute Censored Records for Customer *i*** | |
| $t_i^{FirstVis}$ | Date of a customer *i*'s first visit |
| $t_i^{LastVis}$ | …….. last visit |
| $t_i^{Purch}$ | …….. purchase |
| $t_i^{FirstObs}$ | …….. first observed visit in censored subset |
| $t_i^{LastObs}$ | …….. last observed visit in censored subset |
| $t^{End}$ | End of observation period |
| $VisPurchLag_i$ | # days between customer *i*'s first visit and purchase |
| $90^{th}PurchLag$ | # days 90% of customers in dataset took to go from first visit to purchase |
| $95^{th}PurchLag$ | # days 95% of customers in dataset took to go from first visit to purchase |
| $ObsVisLag_i$ | # days between a customer *i*'s first observed visit and last date observation period |

From June 1 to September 30, 2013, there were 33,303 customers who accounted for 92,217 online visits to the retailer's websites. These visits led to 33,632 purchases, as all customers made at least 1 purchase. 90% of all customers in this dataset made their purchases within 10 days after their first visits to the website. Table 3.8 presents summary statistics of the non-censored dataset. (See Table 3.8.)

**Table 3.8. Customer Site Visits, Purchases in a Non-Censored Dataset**

| VARIABLES | |
|---|---|
| *#Cust* | 33,305 |
| *#AllVis* | 92,217 |
| *#Purch* | 33,632 |
| *AvgVis* | 2.769 |
| *AvgPurch* | 1.010 |
| *Conver* | 0.365 |
| **Notes.** Conversion rate is inflated for this dataset: all customers converted. | |

Next, we aggregated all website visits and purchases to the customer level. [6] At the customer level, the minimum number of visits and purchases are 1. On average, each customer visited the website almost 3 times. The descriptive statistics of this dataset at the customer level are shown in Table 3.9. (See Table 3.9.)

**Table 3.9. Descriptive Statistics at Customer Level - Non-Censored Dataset**

| VARIABLES | NON-CENSORED DATASET (33,303 OBS.) | | | | |
|---|---|---|---|---|---|
| | MEAN | SD | MIN | MEDIAN | MAX |
| *Purch* | 1.01 | 0.99 | 1 | 1 | 2 |
| *#Vis* | 2.77 | 3.76 | 1 | 2 | 71 |
| *#Direct* | 0.93 | 2.03 | 0 | 0 | 69 |
| *#SEngVis* | 0.47 | 1.23 | 0 | 0 | 29 |
| *#AdsVis* | 0.66 | 1.48 | 0 | 0 | 48 |
| *#PersonVis* | 0.71 | 1.20 | 0 | 0 | 49 |
| **Notes.** Only 33,303 customers were able to be matched. | | | | | |

[6] Name of company is masked under an NDA; identities of its individual customers cannot be traced through the data.

### 3.5.3. Data Simulation Design to Create Temporal Censored Subsets

#### Metrics for Data Simulation: Data-Censoring Features

Closely examining the activities of customers across time, we saw that the customer visits are essential to the understanding of their eventual purchases. These records depict how a customer reached her purchase. If some of these records are not observable or not available for empirical testing, that will hinder our ability to explore the causal relationship between the customers' visits and their purchase.

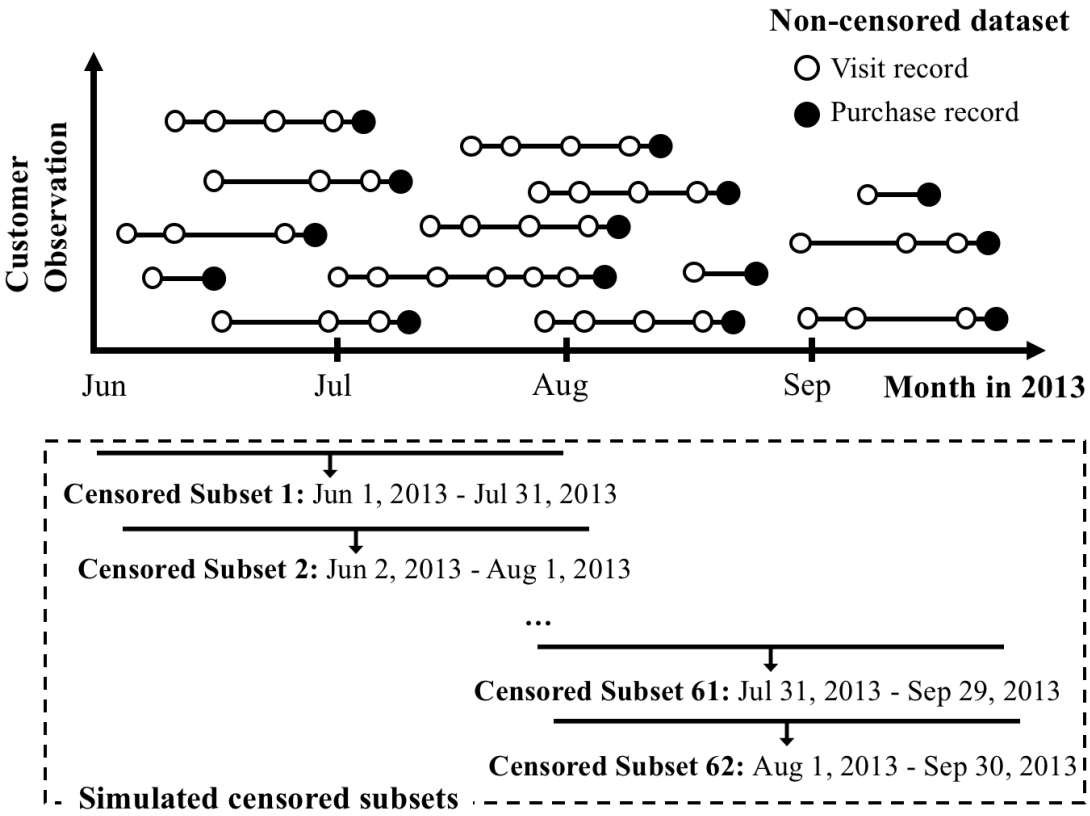
Thus, we looked at three statistical features relating to data censoring as our metrics for data simulation: the proportion of visit-censored customer-level observations in which some visit records are not available; the proportion of purchase-censored customer-level observations in which some purchase records cannot be observed; and the proportion of visit- and purchase-censored customer-level observations in which some visit records and purchase records are not available. In this context, the purchase-censored observations are the right-censored observations. And the visit-censored observations can be left- or right-censored.

#### Iterative Data Simulation to Create a Series of Fixed-Length, Censored Subsets

Our approach uses an iterative process of removing records from different observed sequences of customers' online journeys. This is done via temporal selection of observations across the 4-month observation period. Figure 3.4 illustrates our simulation design, in which we extracted 62 censored subsets across the original time period. We chose a 2-month, 61-day duration for the data subset, and shifted this period across the 4-month original timeline one day at a time. This duration ensured that each subset gave us a sufficient number of left- and right-censored customer-level observations.

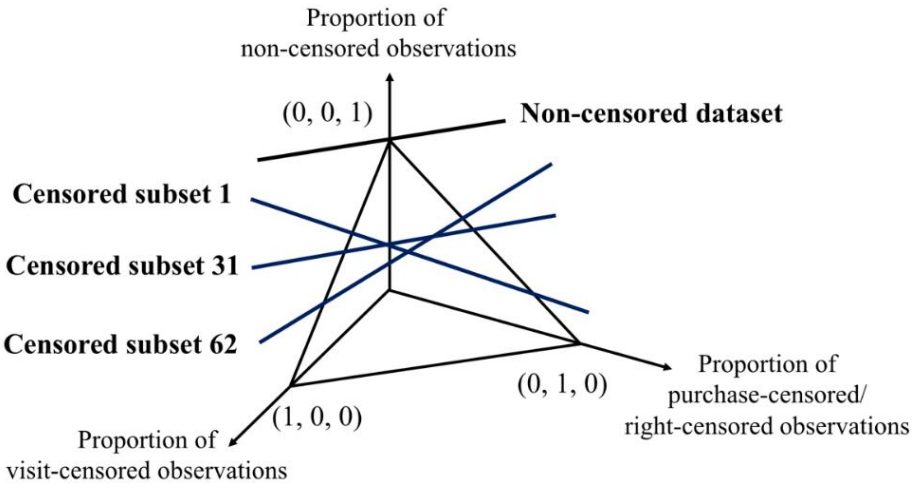
And by shifting this duration one day at a time, we obtained the largest number of censored subsets possible. This allowed us to assess the changes in their features related to data censoring. A more elaborate data-simulation design would be to vary the time duration of the subsets iteratively across the timeframe of the original dataset. For instance, we can simulate a series of 2-week subsets across this timeframe. (See Figure 3.4.)

**Figure 3.4. Data Simulation Strategy to Create Censored Subsets**




Censored Subset 1 includes records from June 1 to July 31. the majority of its customers are purchase-censored. Censored Subset 2 shifts the 61-day time period by one day from June 2 to August 1. This process was repeated iteratively until we extracted the last Censored Subset 62. We further noted that the extent of data censoring differs from one subset to another. (See Figure 3.5.)

**Figure 3.5. Representation of Distance between Censored Subsets**




**Notes.** For illustration purpose only; the actual level of data-censoring varies depends on how each subset is simulated.

Our design produced 62 censored subsets, each with a different degree of data-censoring. As expected, the average number of purchases in Censored Subset 1 is lower than that in the non-censored dataset, and the average number of visits in this subset is higher. Table 3.10 shows the summary statistics for relevant variables of the Censored Subset 1. The customer-level observations regarding online activities of all customers in each censored subset can be naturally classified into: (1) the non-censored data include all customer-level observations for which we can observe all visit and purchase records, (2) the visit-censored data, and (3) the visit-censored and purchase-censored data. (See Table 3.10.)

**Table 3.10. Summary: Customer Site Visits, Purchases in Censored Subset 1**

| SUMMARY STATISTICS | | FEATURE RELATED TO DATA CENSORING | |
|---|---|---|---|
| *#Cust* | 12,034 | *#VisCensor* | 205 |
| *#AllVis* | 33,730 | *#VisPurchCensor* | 925 |
| *#Purch* | 11,122 | *%VisCensor* | 1.70% |
| *AvgVis* | 2.80 | *%VisPurchCensor* | 7.69% |
| *AvgPurch* | 0.92 | *%NonCensor* | 90.61% |
| *Conver* | 0.33 | | |
| **Notes.** There are no left-censored data in this subset; all visit-censored observations in this subset are right-censored. | | | |

Alternatively, Censored Subset 31 contains all records in July and most records in August (up to August 30th), the middle two months of the original dataset. We would see high proportions of both visit-censored and purchase-censored customer-level observations. Thus, the average number of purchases and the average number of visits in this subset are lower than those in the non-censored dataset. We compared all 62 simulated subsets with the non-censored dataset using relevant descriptive statistics as well as the set of features related to data-censoring. These features are essential to our causal inference task. Next, we used our method to impute the censored records for each subset. (Refer to Appendix A for additional descriptive statistics of all censored subsets.)

### 3.5.4. Imputation of Censored Records for Censored Subsets

Our method utilized the temporal sequences of records in the dataset to match and classify all observations into the non-censored data to the visit-censored and purchase-censored data. The goal here is similar to that of the optimal matching in sequence analysis. It is to identify similarities across sequences, which can be used for pattern identification (Biemann and Datta 2014, Rosenbaum 1989). In data mining, Chang and Lee (2005) proposed a mining method for sequential patterns to retrieve embedded knowledge in a continuous data stream. We then identified the empirical regularities in patterns of customer visiting and purchasing activities, and used these values to impute censored records in the censored data. (See Table 3.11.)

**Table 3.11. Imputation Approach for Censored Records**

| SEQUENCES OF RECORDS FOR CUSTOMER-LEVEL OBSERVATIONS | PATTERNS OF RECORDS FOR CUSTOMER-LEVEL OBSERVATIONS | PRESENCE OF EMPIRICAL REGULARITIES |
|---|---|---|
| Non-censored data $t_i^{FirstVis} \ldots t_i^{LastVis} \rightarrow t_i^{Purch}$ | Number of days between customer *i*'s first visit $(t_i^{FirstVis})$ and purchase $(t_i^{Purch})$. $VisPurchLag_i = t_i^{Purch} - t_i^{FirstVis}$ | $VisPurchLag_i$ value distribution for customers gives evidence for existence of empirical regularities in dataset for time period in which the majority of customers made purchases. For instance, $90^{th}PurchLag$ refers to number of days associated with the 90th percentile of distribution associated with customer purchases. This is used to infer and impute censored records. |
| Visit-censored and purchase-censored data $t_i^{FirstObs} \ldots t_i^{LastObs}$ | Number of days between customer *i*'s first observed visit $(t_i^{FirstObs})$ and end of observation period $(t^{End})$. $ObsVisLag_i = t^{End} - t_i^{FirstObs}$ | |
| **Notes.** *t* denotes date, and *i* an individual customer. The percentile of distribution choice depends on the data settings and analyst's objectives for establishing empirical regularities. | | |

**Imputation of Censored Records for Censored Subset 1.** We next demonstrate our imputation approach in detail for Censored Subset 1. First, we examined the sequences of customer activities that gave rise to important patterns. Customers usually visited the websites multiples time before making a purchase. As this subset includes all customer-level observations in the first two months of the non-censored dataset, no observation is censored on the left. Thus, all 925 observations in the visit-censored and purchase-censored data are censored on the right; these customers' visit records and purchase records occurred after the observation period concluded. As some customers made more than 1 purchase, there are 205 customer observations in the visit-censored only data: the customers who already bought the products one time. These customer-level observations are not relevant to our research objective though: we want to recover records from the visit-censored and purchase-censored data only.

The underlying assumption of our method is that the customers that we cannot observe fully in the visit-censored and purchase-censored data do not deviate far from those we observe fully in the non-censored dataset, relating to their visits and purchase

activities. In statistical terms, we looked at the distribution of $VisPurchLag_i$ values and identified the empirical regularities for the time period (in days) in which 90% of all customers made purchases after their first visits. This time period is referred to as the $90^{th}PurchLag$. From the Censored Subset 1, we observed that the majority of customers completed their online shopping journey within 9 days.

For each censored observation at the customer level, if the *ObsVisLag* between the customer's first observed visit and the last day of the observation period is less than 9 days, we inferred that the customer has not completed the journey, and there is a high probability that she would make a purchase. Thus, we imputed a purchase record for that customer. Vice versa, if the *ObsVisLag* between the customer's first observed visit and the last day of the observation period is 9 days or more, we also inferred that the customer has completed the journey and decided not to purchase. This method allowed us to recover 562 censored purchase records. If we had chosen the empirical regularity associated with the purchase patterns for the 95$^{th}$ percentile of the distribution associated with customer purchases instead of the 90$^{th}$ percentile, we would have recovered 775 censored purchase records.[7]

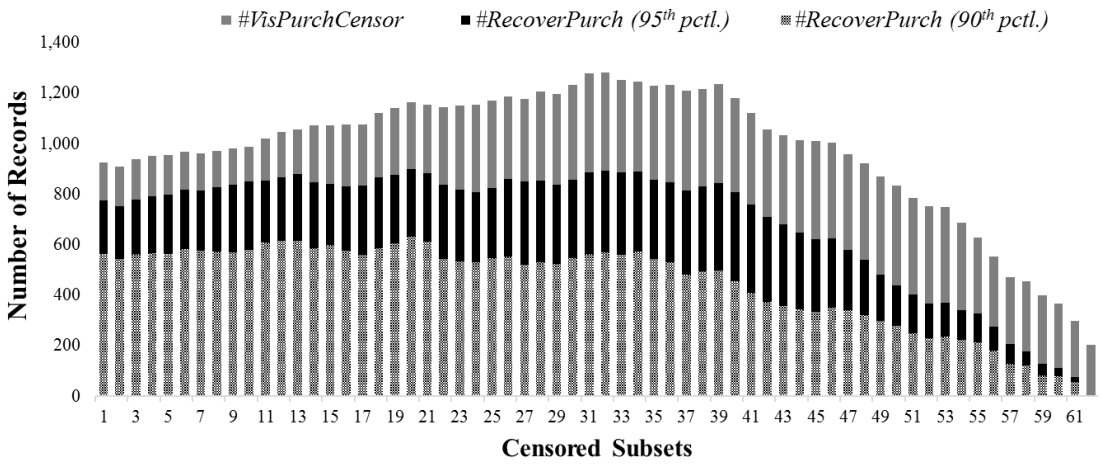
We updated the Censored Subset 1 with these imputed values, using the 90$^{th}$ and 95$^{th}$ percentiles of the distribution, and refer to them as Subset 1 with Imputed Values (90$^{th}$ pctl.) and Subset 1 with Imputed Values (95$^{th}$ pctl.).

[7] The choice of the percentile of distribution depends on the assumption for the fixed-length time period that we chose for the simulated subsets. With this dataset, the fixed-length of the simulated subsets is 61 days. Thus, we conducted our analysis at the 90$^{th}$ percentile, with the associated empirical regularity of around 9 days in lag time between site visit and the customer's purchase, which is the *ObsVisLag*. We also reported the results at the 95$^{th}$ percentile to illustrate the method's flexibility if the empirical regularity was to represent a long lag from visit to purchase on average.

**Imputation of Censored Records for 62 Censored Subset.** Based on our assessment across all 62 censored subsets, the 90th percentile of the distribution of purchases occurring within 7 to 9 days of the shopping, suggesting an empirical regularity in the patterns of customers' visiting and purchasing activities. And for the 95th percentile, the purchases occurred within 14 to 17 days. We applied this method to infer customers' activities and impute censored records for all 62 simulated subsets. (See Figure 3.6.)

These results show that the empirical regularities found in the non-censored data can be used to complete the logical sequences for censored customer-level data. The assumption about the appropriate percentile depends on the research objectives and the causal relationships the analyst wants to explore.

**Figure 3.6. Recovering Censored Records for 62 Censored Data Subsets**




### 3.5.5. Method Performance in Improving Causal Inference

We focus on the effectiveness of a method in recovering censored data to support empirical testing, as well as the computation of the information gained from the recovered data for causal inference. This evaluation design is both observational and descriptive within a specific business content (Hevner et al. 2004). First, we compared the

estimates produced from the censored-dataset and those produced from our dataset with recovered customer-level observations. Then, we assessed the generalizability and robustness of the method in computing those censored observations, with respect to the specific context.

In our research setting, we have the complete online journeys of more than 30,000 customers. Each journey includes all website visits leading to a customer's purchase. Thus, the characteristics and the nature of these visits can be used to explain the customer's decision to purchase. After each visit, she learns more about the product availability on the retailer's website. Depending on the source of traffic, she is more informed about the products' quality and fit information. For instance, comparison sites contain comparison information of similar products, based on price, features and reviews. And social media ads are more personalized as people in social networks may know and communicate with each other via different means. Word-of-mouth, social media and viral marketing are considered the most important channels for marketing; in fact, a report from Nielsen shows that 92% of consumers believe recommendations from friends and family over all other forms of advertising (Whitler 2014).

We explored different relationships between the characteristics of customer visits and their subsequent purchases, to demonstrate how additional information on censored data affects the ability to infer causal relationships. For example, do direct and more personalized advertisements increase the customers' likelihood to purchase? This can have important managerial implications for marketing executives.

From our non-censored dataset, we know that all customers ultimately converted. This was not the case in the censored subsets though: some purchase records are not

observable. Thus, the best research design ex ante is not a suitable design ex post. Researchers working with real-time datasets often face this challenge. For a 2-month dataset, like what we obtained for Censored Subset 1, our method can support the recovery of censored records, so that these records can be used in empirical tests.

We compared the conditional probability of a customer making a purchase, given that the customer's record was right-censored after the observation period for (1) Censored Subset 1 and Subset 1 with Imputed Values (90$^{th}$ pctl.), and (2) Subset 1 with Imputed Values (95$^{th}$ pctl.), via the following Bayes's theorem equation:

$$Pr\ (Purch \mid Censored) = \frac{Pr\ (Censored \mid Purch) \cdot Pr\ (Purch)}{Pr(Censored)}$$

Using our method, the conditional probability of the customer making a purchase after the observation period ended increased from 0 to 0.61 in the 90$^{th}$ percentile case, and to 0.84 in the 95$^{th}$ percentile case. (See Table 3.12.)

**Table 3.12. Conditional Probability of Purchase, Given the Censored Record**

| | Censored Subset 1 | | Subset 1 with Imputed Values (90th Pctl.) | | Subset 1 with Imputed Values (95th Pctl.) | |
|---|---|---|---|---|---|---|
| | **Censored** | **Non-Censored** | **Censored** | **Non-Censored** | **Censored** | **Non-Censored** |
| No Purchase | 925 | 0 | 363 | 0 | 150 | 0 |
| Purchase | 0 | 11,109 | 562 | 11,109 | 775 | 11,109 |
| *Pr(Purch\|Censored)* | 0 | | 0.61 | | 0.84 | |

Thus, the number of purchases, *#Purch*, in the Censored Subset is updated, and the number of imputed purchases represents the information gain in statistical terms from applying our method. Based on the Bayesian inference framework, we recovered more information from recovering censored data records on the basis of prior information, via empirical regularities observed in the non-censored data.

$$\text{Updated } \#Purch = \#Purch + [\#VisPurchCensor \cdot Pr(Purch|Censored)]$$

Next, we assessed how these recovered records affected our causal empiricism. We examined the effect of different types of traffic sources on the customer's likelihood of purchase. The objective is to explore the relevant causal relationship between these variables. Establishing these causal relationships required a more rigorous empirical framework though. So we used a logit model for our empirical testing, as if we only had access to the censored subset, in which some records were not available. Our empirical framework captured the relationship between a customer's likelihood of purchase and other variables via this function:

$$Pr\ (Purch_i = 1) = \frac{e^{\beta_0 + \beta_1 \#Direct_i + \beta_2\ \#SEngVis_i + \beta_3 \#AdsVis_i + \beta_4 \#PersonVis_i}}{1 + \ e^{\beta_0 + \beta_1 \#Direct_i + \beta_2\ \#SEngVis_i + \beta_3 \#AdsVis_i + \beta_4 \#PersonVis_i}}$$

The dependent variable in this model is a binary variable, *Purch*, which represents whether a customer made a purchase. We are interested in the number of customer visits representing the 4 main types of traffic sources. *#Direct* refers to customer visits to the retailer's website directly. A higher number of direct searches indicates that the customer has a special interest in the products offered on the website. *#SEngVis* refers to customer visits that are acquired through organic and paid search with a search engine. The customers are likely looking for a product category that the retailer happens to offer. *#AdsVis* refers to customer visits acquired through online advertisements, such as display ads, affiliate marketing, and Google Adwords. The last variable, *#PersonVis*, refers to the number of visits that originated from more personalized sources, such as an unbiased comparison site, or email and social ads. We expect to see that the number of visits generated from more informative sources, such as *#AdsVis* and *#PersonVis*, have greater influences on a customer's likelihood of purchase. For each customer *i*, we estimated the log form:

$$\ln\left(\frac{Pr\ (Purch_i = 1)}{1 - Pr\ (Purch_i = 1)}\right) = \beta_0 + \beta_1 \#Direct_i + \beta_2 \#SEngVis_i + \beta_3 \#AdsVis_i + \beta_4 \#PersonVis_i$$

To assess the extent to which the recovered records improve the statistical inference for causal explanation, we assessed the variability of our estimates from the simulated subset and its corresponding subset with some imputed values. We ran our logit model using the Censored Subset 1, the Subset 1 with Imputed Values (90th pctl.) and the Subset 1 with Imputed Values (95th pctl.). The results are shown below. (See Table 3.13.) [8]

**Table 3.13. Logit Model Results: Censored Subset 1 vs. Subsets 1 with Imputed Values (90th pctl. and 95th pctl.)**

| | Censored Subset 1 | | Subset 1 With Imputed Values (90TH pctl.) | | Subset 1 With Imputed Values (95TH pctl.) | |
|---|---|---|---|---|---|---|
| Variables | Coef. | *p* (>\|*z*\|) | Coef. | *p* (>\|*z*\|) | Coef. | *p* (>\|*z*\|) |
| *Intercept* | 2.262*** | < 0.001 | 3.446*** | < 0.001 | 4.416*** | < 0.001 |
| *#Direct* | -0.035** | 0.021 | -0.082*** | < 0.001 | -0.104*** | < 0.001 |
| *#SEngVis* | -0.036 | 0.262 | -0.028 | 0.512 | -0.026 | 0.718 |
| *#AdsVis* | 0.093*** | 0.004 | 0.020 | 0.625 | 0.012 | 0.842 |
| *#PersonVis* | 0.460*** | < 0.001 | 0.217*** | 0.001 | 0.107 | 0.211 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 1; 12,034 obs. Null dev.: 6,524; 12,033 d.f.; resid. dev.: 6,404; 12,029 d.f., AIC: 6,414. Subset 1 with Imputed Values (90th pctl.); 12,034 obs. Null dev.: 3,257; 12,033 d.f.; resid. dev.: 3,224; 12,029 d.f., AIC: 3,234. Subset 1 with Imputed Values (95th pctl.); 12,034 obs. Null dev.: 1,614; 12,033 d.f.; resid. dev.: 1,593; 12,029 d.f., AIC: 1,603. Signif.: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

These results from the logit model using the Censored Subset 1 align with our reasoning. The coefficients for *#AdsVis* and *#PersonVis* are positive and significant, suggesting that online ads are effective in influencing customers' likelihood of purchase.

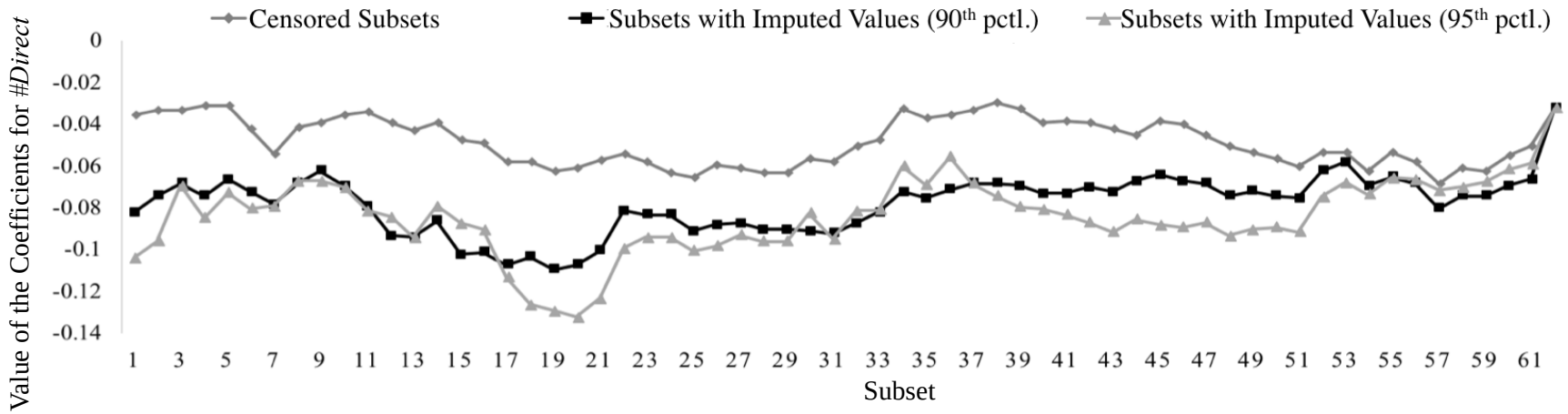
[8] In our research context, if the researchers were given a 2-month dataset only, it would not have been feasible to compare the estimates produced from the 4-month, non-censored dataset and those produced from the 2-month, censored subsets. To the best of our knowledge, there are no other methods handling censored data for causal inferences that we can use as benchmarks for evaluation. Thus, we followed Hevner et al.'s (2004) guideline for evaluating new artefacts. We showed the information value of the method for enhancing the collection of logical sequences of data – including those outside the observation period, which led to an improvement in statistical inference.

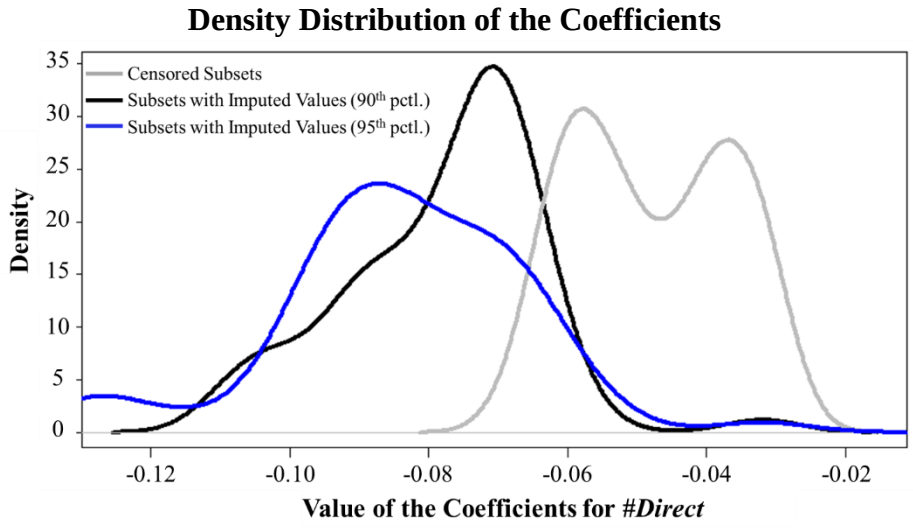
Surprisingly though, the coefficients for *#Direct* are negative; this may imply that a general interest in a retailer's wide range of product offerings does not necessarily lead to a purchase of a particular product.

In the subsets with imputed values, the variable *Purch* is updated with the imputed records. The results obtained from these subsets are consistent with what we have seen for Censored Subset 1. The directions of the relationships, based on the signs of the coefficients, remain the same, though the coefficients and their significance levels have changed. Thus, the subsets with some imputed values allow us to obtain a more reliable reading on the relationships between different acquisition channels and the customer's likelihood of purchase.

We repeated this process and ran the logit model using all 62 censored subsets and their respective subsets with imputed values. Then we compared and evaluated the variations and distributions in our econometric estimates produced using different subsets. Figures 3.7 to 3.10 below provides the direct comparison between the coefficients produced from all pairs of subsets. We also look closely at the distributions of the coefficients from censored subsets and from those subsets with imputed values. (Refer to Appendix B for more detailed result tables.)

**Figure 3.7. Comparison of the Coefficients for *#Direct* from Censored Subsets and Subsets with Imputed Values**



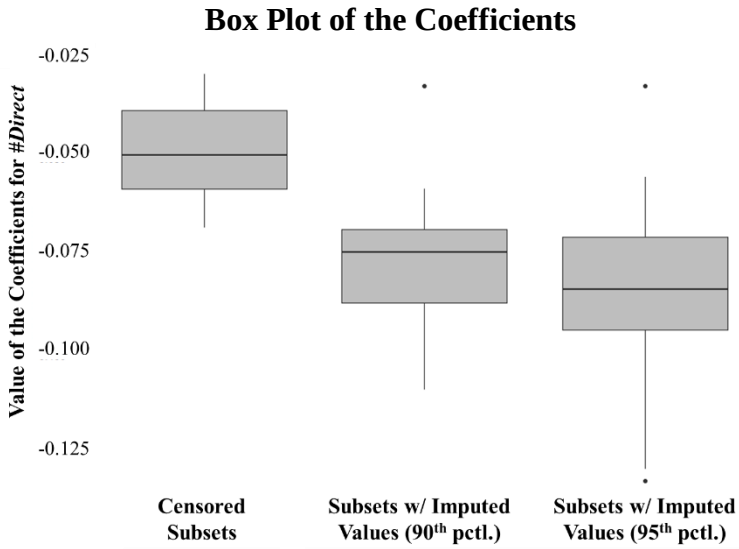



**Figure 3.8. Comparison of the Coefficients for *#SEngVis* from Censored Subsets and Subsets with Imputed Values**

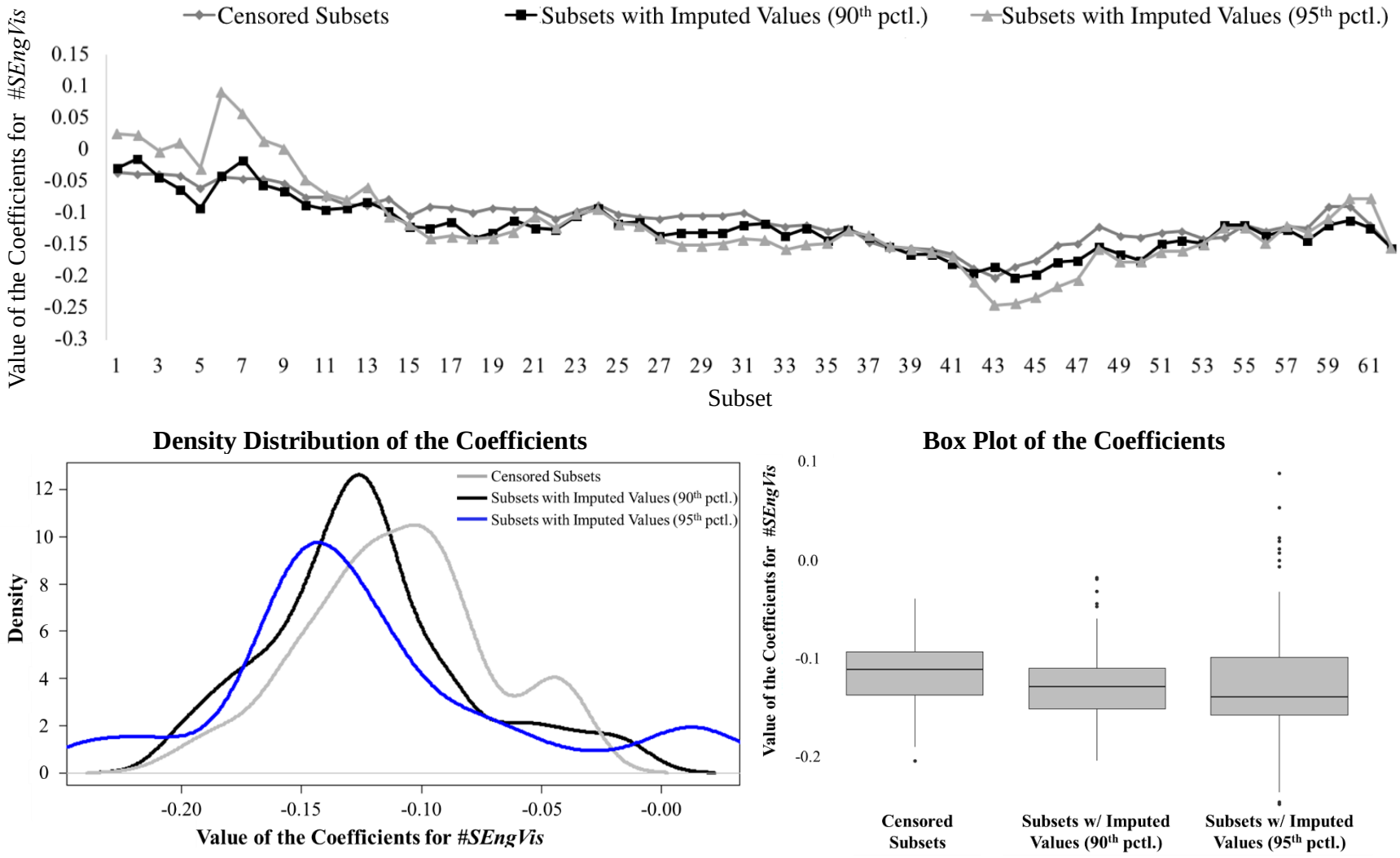


**Figure 3.9. Comparison of the Coefficients for *#AdsVis* from Censored Subsets and Subsets with Imputed Values**

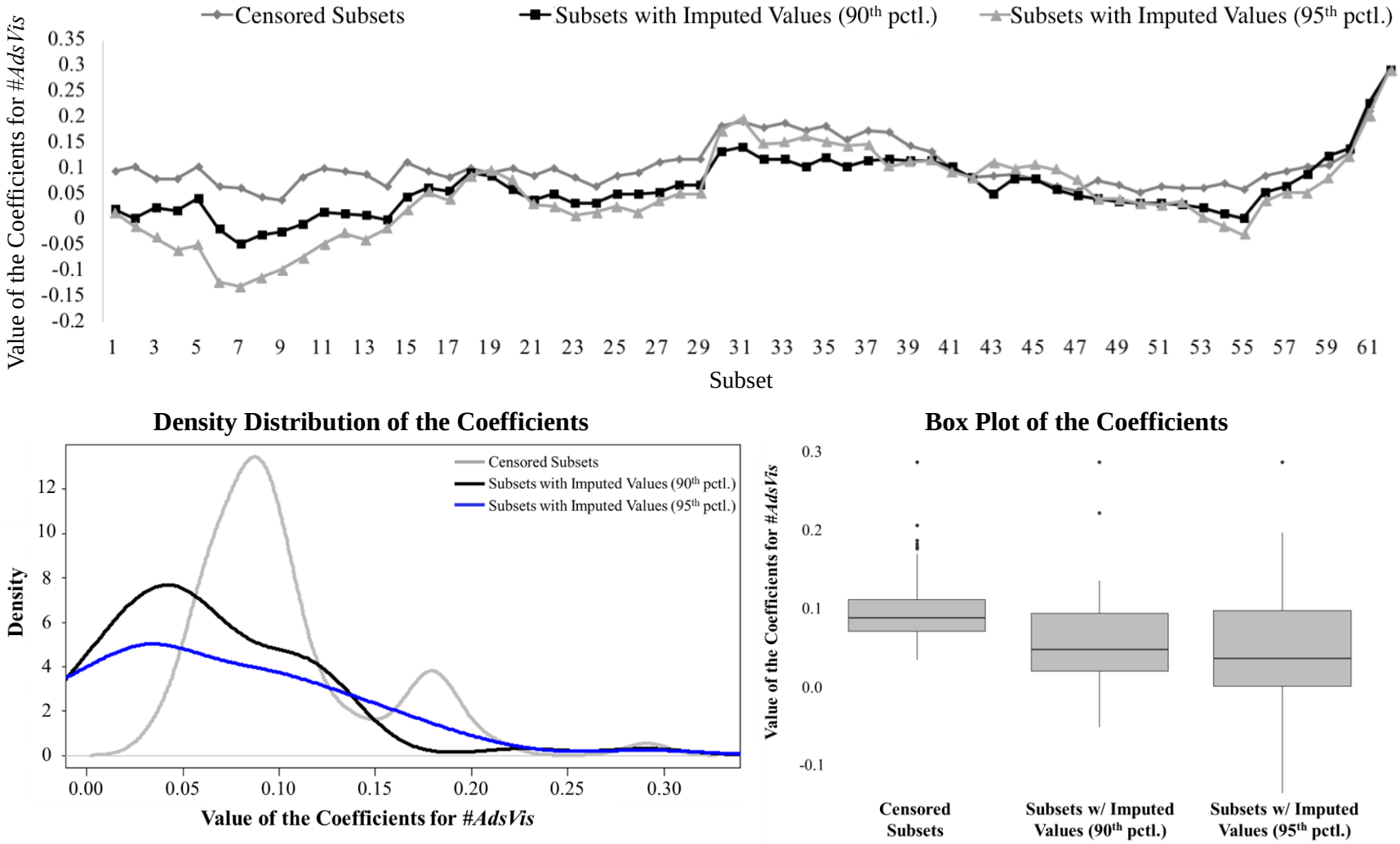
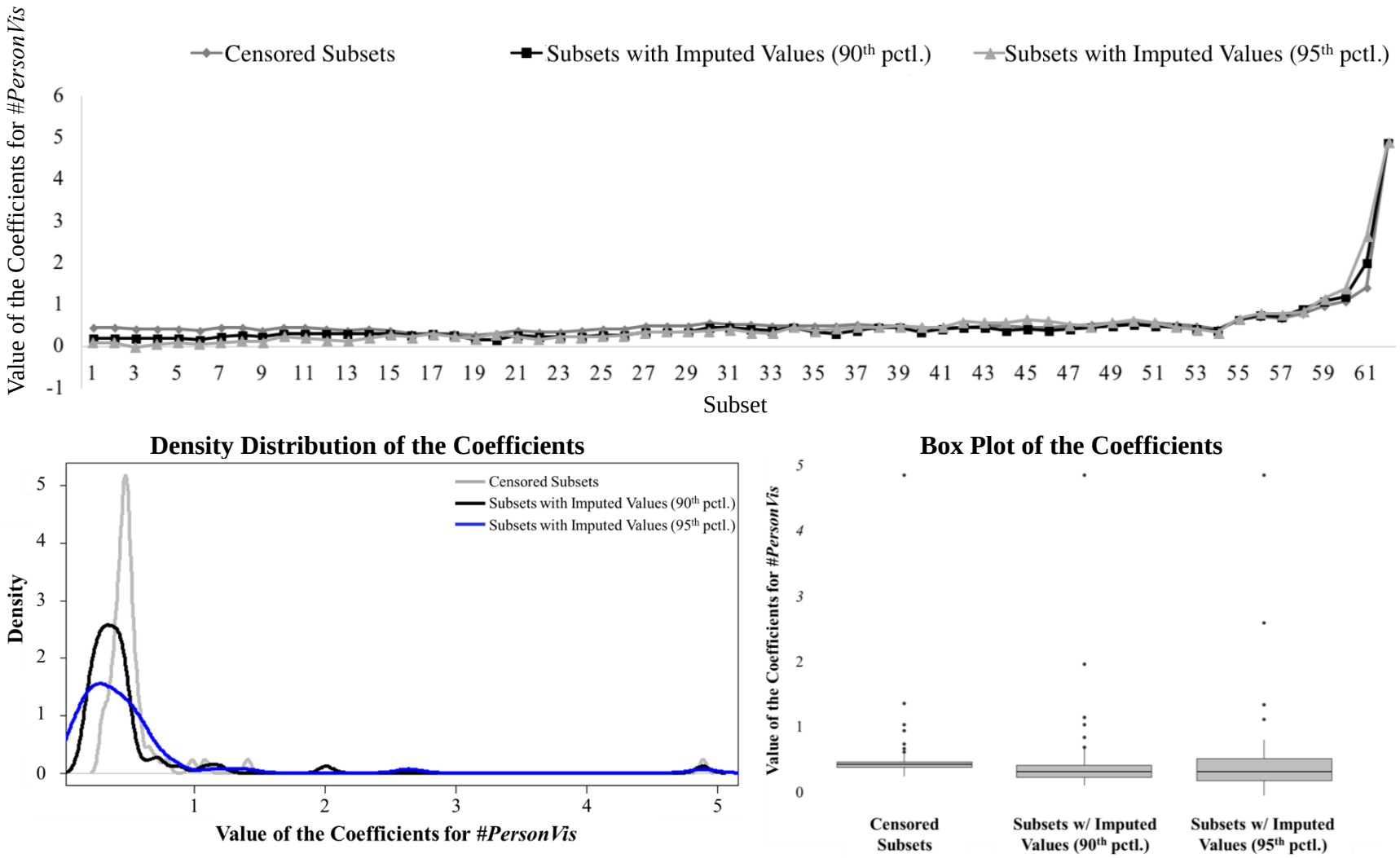


**Figure 3.10. Comparison of the Coefficients for *#PersonVis* from Censored Subsets and Subsets with Imputed Values**



For each variable, we conducted a paired *t*-test to compare the mean of coefficients produced from the censored subsets and the mean of the coefficients produced from their corresponding subsets with imputed values using the $90^{th}$ percentile. The results indicate that these coefficients are significantly different from the others, which provides evidence for an information gain from implementing our method.

We obtained a stronger reading on the negative relationship between *#Direct* and the customer's likelihood of purchase. The coefficients of *#Direct* from the subsets with imputed values are lower (on the negative side) than those from the censored subsets (paired *t*-value: 17.87, $p < 0.001$). If we had used the censored data, we would have under-estimated the relationship between *#Direct* and *Purch.* Likewise, the coefficients of *#SEngVis* from the subsets with imputed values are still negative, and they are slightly lower (on the negative side) than those from the censored subsets (paired *t*-value: 6.71, $p < 0.001$). This suggests that the retailer should not rely too much on search engine traffic alone to predict sales, as customers only explore the range of product offerings without having a real interest in any particular product.

For *#AdsVis* and *#PersonVis,* the coefficients of the censored subsets are slightly higher (on the positive side) than those from the subsets with imputed values (paired *t*-value: 11.66, $p < 0.001$; paired *t*-value: 5.27, $p < 0.001$). This suggests that the positive effect of general and personalized advertisements traffic sources on the likelihood of purchase was over-estimated. These results show that the differences between the estimates produced using the censored data subsets with imputed values are statistically significant. Next, we looked closer into how we achieved an information gain using the subsets with imputed values, via the box plots for the coefficients of each variable.

The box plots provide visual evidence for the information gain from the subset with imputed values. The line inside each box represents the median, or 50$^{th}$ percentile, whereas the box itself represents the range of all the coefficient estimates from using different series of subsets, including the Censored Subsets, Subsets with Imputed Values (90$^{th}$ pctl.) and Subsets with Imputed Values (95$^{th}$ pctl.). For all variables, we observed that the medians of the coefficient values produced using the subsets with our imputed values are smaller than those using the censored subsets. This suggests that with the censored subsets, we are likely to underestimate the negative effects of the variables *#Direct* and *#SEngVis* on *Purch*, and overestimate the positive effects of *#AdsVis* and *#PersonVis* on *Purch.*

The differences between the upper and lower quantiles, and the levels of skewness also show the variation of the estimates and provide more insight into the underlying causal relationships we want to examine. For instance, we are able to gauge the variations in the econometric estimates from a larger number of simulated datasets and assess the point at which the information gain reaches its maximum level. There, the marginal benefit from obtaining more data is minimal, as we already have sufficient data for causal inference.

The information gain from obtaining a more sequentially-complete dataset for theory-focused empirical research is associated with a better estimate of the coefficients that provide evidence for the causal relationships of interest in a model. The method we proposed enables researchers to establish causality in findings that can be obtained from less-than-ideal datasets.

### 3.5.6. Discussion and Limitations

We assessed our proposed method using a 4-month, non-censored dataset. In a data simulation, we created 61-day censored subsets with different degrees of censoring based on a complete dataset. Incomplete sequences of records for customer-level observations in each subset hinder a researcher's ability to conduct to support causal inferences for theory evaluation. This is a common issue with big industry data.

We also showed that our method can impute censored records for customer-level observations. As a result, complete sequences of records can be used to obtain a stronger reading on relevant causal relationships. In particular, subsets with imputed values yielded better estimates of the relationship we wanted to study and offered a fuller picture of how different traffic sources influenced the likelihood of purchase or the decisions of those who purchased. This would not have been possible if we had used the censored subsets only.

Regarding the empirical regularities associated with the distribution of the customer visit-to-purchase time-lags, we used the $90^{th}$ percentile as well as the $95^{th}$ percentile to show that our method is both flexible and context-specific enough to support the acquisition of useful results. Researchers can have different assumptions of the appropriate percentile to use, based on the context of their data. In addition, it will be interesting to try out different fixed-length periods in the data simulation design. They can also choose to shift this time period across the original dataset's time frame by different amounts of time. There are some constraints though. The sample size, or the size of the simulated subsets required for adequate statistical power depends on the researcher's objectives as well as the complexity of her empirical model.

Another possibility for the evaluation of the method is via data simulation. For instance, we can follow these high-level steps: (1) simulate a dataset with true parameters; (2) extract a censored subset from the simulated dataset; (3) apply the method to get the subset with imputed values; and (4) compare the parameters produced from the censored dataset and the subset with imputed values to the true parameters. Nevertheless, the implementation is not as straightforward as we described above. In order to evaluate the method with a synthetic dataset, we also must design a context-specific empirical testing approach for it.

### 3.6. Conclusion

Our methods innovation in this research is motivated by prior work with an industry dataset that involves 17+ million household TV viewing and purchasing records (Hoang and Kauffman 2017). We faced the issue of only having access to a much smaller dataset ("a needle in a digital haystack") due to left- and right-censored data at the household-level for a relatively short timeline of observations. We overcame this data limitation by using the logical sequences and patterns of all households' observations to make inferences about the behavior of censored records. The early version of our method innovation is a modified propensity score matching method. It allowed us to show the relationship between households' VoD sampling and purchases. This had been infeasible to do with a censored dataset, while retaining sufficient statistical power.

The availability of large datasets is prompting new forms of research inquiry across different disciplines. Academic researchers and their industry partners must figure new ways to extract relevant insights from the data more effectively though (Müller et al. 2016). In this work, we provided a methodology contribution to the IS literature, in

which causal inference plays an important role. The objective of our method – *iterative data simulation for context-specific probabilistic inference* – is to utilize the logical sequences of records in a dataset to recover censored data, and so to inform causal inferences in theory-based explanatory research. To the best of our knowledge, our method is the first of its kind that recovers censored data records outside the observation period of a study, so that the data can be used to improve the explanatory statistical power of econometric estimation results that are obtained from a less-than-ideal dataset.

We have addressed challenges that data analytics researchers face in practice, not only in IS but also across different disciplines, as Phillips-Wren et al. (2016) have called for. Our method relies on the implementation of data simulation based on an industry dataset that iteratively compares model performance in the presence of somewhat different experimental treatments that vary the extent of the censored data. This approach, we believe, can improve both the explanatory as well as the predictive power of the dataset.

This has important implications on different major IS research context areas, such as auctions, customer churn management, online selling, P2P lending, and crowdfunding. In these contexts, consumer data are limited, and researchers must work under regulatory constraints. In addition, researchers and managers often do not have control over the data-generating process nor the ability to conduct experiments that would allow them to infer importance causal relationships. Hence, our method offers a workaround solution that allows researchers to look beyond the observation period which makes it possible to establish the sequence of causal activities that lead to different outcomes for consumers and people in other applied contexts.

## Chapter 4: Business and Consumer Analytics Research Practice

In this age of information and transformation, the relationship between people and technology, or between the creators and their creations have become more complex. Business executives have a trendy acronym to describe the external environment, VUCA, which stands for Volatility, Uncertainty, Complexity and Ambiguity. Everything is changing rapidly and unpredictably. Through the lens of a scientist, this presents many opportunities – and many questions worth exploring. Depending on where one stands, a person may have a unique perspective on these relationships. As a result, researchers and industry experts should be able to take a pioneering role in exploring these new territories of knowledge.

My interdisciplinary research lies in the niche area of IT and Marketing, where I have an opportunity to create new insights by looking closely at how people interact with new technologies and systems. I also explore new methodologies that allow researchers to extract causal relationships from the digital traces of consumer behavior. These topics express my interest in trying to understand consumers to the extent that the data allow. My research journey started with a comprehensive review of current literature across IS, Marketing and Consumer Behavior; and this has allowed me to have a holistic view of different business phenomena. The advanced training in data analytics, econometrics, and machine learning also have enabled me to have a sneak peek at what has happened in the world based on the data, so that I am able to more effectively explore the business issues in depth. The exposure to datasets from different industries also has fueled my passion for consumer and business analytics.

### 4.1. How Can We Formulate Novel Research Questions?

Impactful research requires scientifically-valid research questions. They must involve a process of discovering empirical facts in a specific context, an assessment of theory, and development of research methodology to solve a specific problem. The first lesson that I learned from Professor Robert Kauffman, my advisor at SMU, had to do with "walkabout empiricism" and observational science, and how to form research questions for industry settings. These questions may be stated for a specific context, but also can be generalized beyond it.

My two essays make theoretical contributions to existing knowledge related to the discovery of hidden consumer preferences, especially for on-demand digital entertainment goods and online products. This is especially interesting since the data-sponsor firms may not be fully aware of the new knowledge that researchers have been able to discover through relevant theory and rigorous analytics.

### 4.2. Theory in Theory? Or Theory in Practice?

I have always considered myself a "business" researcher, due to my background and interest in entrepreneurship. Before undertaking any research project, I often ask whether it is relevant. Nevertheless, it is easier to lose sight of the practical aspects, as I dive deeper into a literature review and explore the value of testing different models.

The richness and availability of data have allowed me to avoid the loophole of "it works in theory but not in practice." Online digital traces can be put together to create interesting stories about people, just like their shadows, but in the digital world. Thus, they have the potential to play a critical role in interdisciplinary research. In Essay 1, I

examined households' TV viewing activities with a fine-grained dataset, which allowed me to gain an in-depth understanding of their sampling and purchasing behavior for video-on-demand series dramas. The combination of industry data with a strong theoretical foundation produces meaningful research that offers useful managerial insights. With a well-supported theory narrative, I was able to not only understand what these households would do, but also explain what they actually decided to do. I view the work on this essay as having created an extension to existing theory.

In Essay 2, I looked at how consumers shop online, using a much larger dataset. I can observe consumers' online searches leading to their purchases, similar to how they would visit a brick-and-mortar store in the physical world. The search cost associated with online shopping is significantly less than that of physical store shopping. Nevertheless, with online channels, consumers don't have a chance to examine the products directly, so they will be less informed about these products. I expect to see that these subtle differences will have great impacts on consumer purchase decisions. First, consumers' product searches will be more intensive in the absence of search cost. Second, the source of product information and channel of advertisement will have a greater influence on the customers' decision to purchase. This context presented great opportunities for me to test existing theories and develop new theories that explain the interaction between consumers and business in the digital world.

### 4.3. What Is an Ideal Dataset?

All datasets are valuable for those who seek to understand them, just like all keys will open something. But there is no ideal dataset. There are many critical issues with my dataset used in Essay 1. It was not small with 17 million records. It was also rich

and informative in terms of the households' subscription information and their video-on-demand viewing activities. And yet, it was not an ideal dataset. The final dataset that I was able to use for the research turned out to be relatively small, and certain aspects of the relevant internal corporate data were not available to me. My effort in extracting valuable information from this dataset was an iterative and continuous process.

Also, the binding non-disclosure agreement that I worked under required some creativity on my part to be successful. This is a common challenge for practitioners and researchers alike, when data, resources and time are limited. Researchers have to accept the limitations of their data and figure out new ways to work with them in a scientific manner to establish insightful causal explanations. And only by doing so, can they find the answers to their questions.

### 4.4. What Is a Useful Toolbox of Methodologies for Effective Research?

In my 10-month residential training at Carnegie Mellon University, the other Ph.D. students and I got a chance to work under the mentorship of the late Professor Stephen E. Fienberg. He shared with us the basis for field experimental design through the stories of plant and crop planning in agriculture. It turns out that the A/B testing set-ups that all of the high-tech companies such as Google, Facebook or Netflix use had humble beginnings in the real "field." What I have learned is that all methodologies are tools to researchers, whom must learn how to use them effectively. And as the nature of the research objective and of the data change, researchers must replenish and customize their toolbox of techniques and methodologies accordingly.

In Essay 1, the original dataset that I obtained from the service provider included

more than 17 million household TV viewing and transactional records over a one-month period in 2011. By all reasonable metrics, this dataset could be considered as large-scale and rich. Nevertheless, I still faced the issue of it actually being a "small dataset" for my chosen research objectives. In order to produce managerially meaningful results related to TV VoD sampling and purchases, I needed to be able to observe the households for a longer time period. In addition, I did not have any ability to effect control over how the data were generated, nor was I involved in the firm's business decision-making. It was also not possible to conduct field experiments to obtain an ideal dataset for empirical testing.

This motivated me to come up with a workaround solution that would work with the existing dataset and still obtain as deep an understanding of household behaviour as the data had to offer. Since I could not obtain additional data points, it was necessary for me to look closer at those that I already had. By doing so, I discovered that the sequences and the patterns of household viewing activities offered fresh insights into what might have happened before and after the observation period. Thus, I was able to recover censored-data observations, based on a logical process of statistical inference, which later could be used for empirical testing.

In Essay 2, I extended this solution to propose a practical scientific method for working with large-scale observational data subject to data censoring. This is highly relevant to researchers as well as industry practitioners in this digital age. I have been able to demonstrate this proposed method more thoroughly, using another large dataset that was provided by my external Committee Member from the Rotterdam School of Management, Erasmus University in the Netherlands, Professor Ting Li. This dataset

included more than 90,000 records of consumer visits to an online retailer over a 4-month period, which gave me more flexibility to conduct iterative data simulations. My objective was to improve the explanatory statistical power of my modeling approach for econometric estimation using a simulated "imperfect dataset."

From this methods innovation process, I have learned that there is no universal method to address all research questions. Instead, it is up to the researcher to be flexible enough in designing a methodological framework to support stronger scientific analysis of the data and the setting. Furthermore, it is the researcher who has the most knowledge of the dataset. Thus, she must be creative in finding or creating a suitable method that works well with the dataset.

### 4.5. How Can We Engage an Academic-Industry Audience for Constructive Developmental Feedback?

For both of my essays, I received a lot of constructive feedback and developmental comments in the early stages of the work. It was intimidating at first, especially when some comments were challenging, even when they were developmental. Over time though, I have learned to appreciate the wisdom of the crowd and am able to work more closely with academic colleagues and industry experts.

I happen to be working in a research area that is accessible to a lot of people who know about TV programming and cable TV services. I have shared my research ideas and designs, and actively asked other colleagues for their input. My study area is highly attractive, and it seems to attract people who care about consumer behavior, digital goods, and entertainment services. They always have more knowledge than I do on different topics across multiple relevant disciplines. As a result, these interactions have

brought new perspectives into my work, which helped me deepen my research inquiry and make my research designs more effective. As a result, I befriended a group of very supportive academic brothers and sisters, which has led to many beneficial encounters and collaboration opportunities for my future research.

In addition, in January 2017 I received the Young Scholar Travel and Participation Fellowship for the 2017 Pacific Telecommunications Council Conference in Honolulu, Hawaii. There I met many executives from the telecom, information security, and high-tech industries. The conference allowed me to bring my research closer to the industry, and it blurred the line between academia and industry that I originally had in my mind. For data-driven empirical research, I learned that recognition from industry professionals is as important as that from the members of the university research community.

It is equally important that I offer my review service and offer insightful reviews to other authors. I have improved this skill set over time, especially for the identification of the strengths and weaknesses of each manuscript that I review. Like myself, other authors will benefit greatly from my developmental comments. I sharpened this skill set based on my participation in the 2017 ICIS Doctoral Consortium in Seoul, Korea. At the Consortium, all 40 Ph.D. student participants were assigned to smaller groups of 8 with 2 faculty mentors. This setting enabled us to interact more closely with each other about our research, and we were able to share our feedback freely.

### 4.6. How Can We Achieve Publication in a Leading Journal?

Three words: persistence, patience, publication! Essay 1 is the product of 5 years of work on digital entertainment research with support from my corporate sponsor. It started off with a small data exploration project when I was a Master's student in 2013.

The earliest version of this academic research, entitled "Experience Me! The Impact of Content Sampling Strategies on the Marketing of Digital Entertainment Goods," was presented at the 2016 Hawaii International Conference on System Science (HICSS), where it was selected as the Best Research Paper in the E-Marketing Minitrack in the Digital Economy Track. I believe that this was early proof of novel research ideas for my work, and that I had something people would care about.

Even though I received a lot of advice on theory development from HICSS, this work hit a roadblock right after that. The data limitations prevented me from conducting a more thorough empirical test of the causal relationships I wanted to look at, even though I tried many different methods. It took me several months to move from the mindset of "There is nothing to be done with this dataset" to "What can I actually learn from the data that I have?" After numerous brainstorming sessions with my advisor, we were able to come up with a new method to overcome the data limitations – essentially creating a *synthesized dataset*, so that causal inference became possible rather than unattainable. Some aspects of this new method related to censored-data and extending the propensity score matching (PSM) procedure.

I also shared my views on a new way of thinking about how to leverage the "logical sequences" that are present in most consumer shopping settings, involving searching, sampling, decision-making, and purchasing a chosen product. I presented and discussed these ideas at the 2017 Statistical Challenges in Electronic Commerce Workshop (SCECR), held in Ho Chi Minh City, Vietnam in July 2017. Through this research work, I was able to "check the box" for doing rigorous and innovative methods work involving IS and Marketing research. I was recently invited to return to SCECR 2018,

for another research presentation. This time, the work is about my effort with the further development of the censored-data recapture method and the use of statistical methods to understand the cost and benefit relationship of using the new approach in terms of additional information for supporting evidence for causal inference in my e-commerce field study in Essay 2.

The development of this project has been challenging, yet the most rewarding research experience that I have had over the past few years. In each round of review with the leading IS journal I targeted, the Review Team brought up various issues that I had not addressed thoroughly enough in my manuscript related to theory, causality and exposition issues. It was a long, yet fruitful intellectual dialogue with the Review Team members, even when it was not easy for me to come up with a solution. From this, I learned from my advisors the art of drafting a review response, and doing so in a way that all the reviewers' concerns are addressed properly, yet my research objectives and contributions still are preserved. The process was similar to having a debate with the reviewers, but I really learned something useful along the way.

I also learned to be mindful about timing. As new ideas for projects like mine can be generated every day, it is important to remember that research published on an important subject in a timely way can make a great contribution to academia as well as to industry. My hard work eventually paid off, because my extension of the Essay 1 work was accepted for publication by the *Journal of Management Information Systems* (JMIS). This is an "A+" journal in the IS discipline, similar to *MIS Quarterly* and INFORMS *Information Systems Research*, and one of the top 50 business journals ranked

by the *Financial Times* (London). More importantly, this experience has given me confidence for undertaking future projects.

### 4.7. Is There a Scientific Process in This Research?

After my publication experience with the JMIS article, I recognize that there is indeed a scientific process that can be followed for conducting strong empirical research. It starts with defining novel research questions, and includes developing a rigorous methodological approach that allows a researcher to draw statistical conclusions, which also are based on a strong theoretical background. But there is a catch: three essential elements in this process that are necessary go beyond the basic scientific method: determination, persistence and creativity.

Having a complete experience with this process after 5 years, I ask myself: What has brought me here? It has been my passionate search for a better understanding of myself and of others. Like most people, I often overlook these three deep questions, even though we ask them every day: (1) Who are you? (2) Where are you from? and (3) Where are you going? In my research, I want to answer these questions for all the people that the data represent. It is fascinating to see how technology brings people closer and strengthens their ties, and how I can discover these nuanced relationships. While doing my consumer research, I gained a deeper understanding of people whom I have never met. With the skillsets and experiences that I have gained from this Ph.D. program, I want to follow my passion for discovery and become a successful multidisciplinary scientist.

## Chapter 5: Conclusion

This dissertation has demonstrated the use of fusion analytics in uncovering business and consumer insights for a better understanding of how consumers interact with goods and services in IT-enabled platforms, for physical products in the retail industry as well as on-demand services in the entertainment industry. These insights are valuable to business executives in evaluating the effectiveness of current strategies and assessing the potential of new marketing strategies. Firms can make more informed business decisions about experience goods, in which the consumer's focus has shifted from the amount of information available to the specific type of information that is highly relevant.

Technology and media are delivering content that is transforming society today. People spend a substantial amount of time consuming media content, and binge-watching has become a new norm. As the consumption of content evolves, so must marketing strategies. Providers must compete for consumer attention to sell their digital information goods effectively. This is especially challenging, since there is a high level of uncertainty associated with the consumption of such goods. Service providers often use *free programming,* a sampling strategy to share product information.

Essay 1 contributes consumer insights from empirical research on household informedness that influences the households' purchases of on-demand content. In this research, I examined the effectiveness of content sampling strategy used for *video-on-demand* (VoD) series dramas, a unique class of entertainment goods. I extracted data from a large set of household VoD viewing records, provided by a digital entertainment firm, and combined it with external data sources. I also extended a *propensity score*

*matching* (PSM) approach to handle censored-data, which permitted me to explore the main causal relationships. Relevant theories in the Marketing and IS disciplines informed my research on consumer involvement and informedness for decision-making under uncertainty, the consumption of information goods, and seller strategies for digital content.

The results show that content sampling stimulates higher demand for series dramas, but in a more nuanced way than was expected. Samples of the series reveal quality information to consumers, and allow them to assess preference fit directly. As a result, they become more informed about their purchase decisions. Also, households seem to be willing to pay more to be better informed, and informed households tend to purchase more. This suggests that content providers should invest in strategies that help consumers to understand the preference fit of information goods.

The data limitation that I faced in Essay 1 motivated a workaround solution related to working with censored data. In that essay, I employed an extended propensity score matching (PSM) procedure to match censored and non-censored observations based on households' records of sampling and purchasing that could be observed. Using a probabilistic model, I inferred the censored records of households' activities that may have occurred outside the one-month study period. The idea was valuable for the research: logical sequence provides stronger support for causal inferences.

In Essay 2, I asked: To what extent can empirical regularities found in a dataset be used to enhance the collection of logical sequences of data in theory-focused empirical research? How can a more informative dataset with fewer censored customer-level observations be acquired for research designs that support causal inferences? I propose a

probabilistic inference method that improves the sequential completeness of the dataset via econometric imputation of data outside the observational period of study. I also offer an iterative data simulation approach to assess the generalizability and robustness of the method in recovering censored records to support causal inference. To demonstrate its use, I explored a large dataset with more than 90,000 customer visit and purchase records to a European electronic retailer's website. I examined the causal relationship between different traffic sources – with some more informative than others – and the likelihood of customer purchases. I found that empirical regularities in the pattern of customer activities can be leveraged to recover censored records outside a study's observation period associated with logical sequences of consumer behavior that enable more effective theoretical conclusions to be drawn. I further showed that such additions to an empirical dataset yield measurable information gains in the capability to accomplish statistical inference for causal explanation.

The two essays are examples of how fusion analytics can support the extraction of business and consumer insights from large-scale datasets. I used machine-based methods, such as feature selection and data simulation to obtain datasets with desirable characteristics for empirical testing. In addition, my empirical models were developed with respect to the theoretical background and the business context from which the data were generated.

There are some limitations to the research presented in this dissertation. These essays are dependent on their business contexts. Essay 1 looked at a specific area of digital goods, on-demand series dramas, and thus the results may not be generalizable to

other types of digital entertainment products. As the current technology does not support the tracking of individual views, I was only able to consider a unitary model of household: all of the family members as a single unit, not as utility-expressing individuals. Due to data limitations and data privacy issues, I also could not examine how households' subscription information and demographic characteristics influenced my results.

In contrast, Essay 2 looked at an e-commerce website in Europe, and the insights extracted from that market are probably not totally applicable to business planning in other markets. In addition, I did not account for external factors that might have influenced consumer decision-making on purchases, such as word-of-mouth effects. Consumers are heterogeneous with diverse product and service needs, and they also have access to different sources of information about the products and services they are interested. Thus, firms must take these kinds of information into account for even more effective strategic planning.

Future studies should explore different aspects of consumer preferences, and demand and consumption patterns. There is much more that we don't know than we do know in this time of digital transformation. Thus, the research potential of data analytics for business and consumer insights is limitless. This dissertation is my first attempt as a multidisciplinary scientist to contribute relevant insights that address the complex relationships between the consumer and business in online platforms.

## References


Ayal, M., and Seidman, A. An Empirical Investigation of the Value of Integrating Enterprise Information Systems: Case of Medical Imaging Informatics. *Journal of Management Information Systems 26,* 2 (September 2009), 43-68.

Ba, S., and Pavlou, P.A. Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior. *MIS Quarterly 26*, 3 (September 2002), 243–268.

Bapna, R. and Umyarov, A. Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks. *Management Science 61*, 8 (April 2015), 1902–1920.

Bapna, R., Ramaprasad, J., Shmueli, G., and Umyarov, A. One-Way Mirrors in Online Dating: A Randomized Field Experiment. *Management Science 62,* 11 (February 2016), 3100–3122.

Bareinboim, E., and Pearl, J. Causal Inference and The Data-Fusion Problem. *Proceedings of the National Academy of Sciences of the United States of America 113*, 27, (June 2016), 7345-7352.

Baumeister, R.F., and Bushman, B.J. *Social Psychology and Human Nature,* 2nd ed., Boston, MA: Cengage Learning, 2010.

Becker, G.S. A Theory of the Allocation of Time. *Economic Journal 75*, 299 (September 1965), 493-517.

Bell, D. E. Regret in Decision Making Under Uncertainty. *Operations Research 30,* 5 (October 1982), 961–981.

Bergen, M., Kauffman, R., and Lee, D. Beyond the Hype of Frictionless Markets: Evidence of Heterogeneity in Price Rigidity on the Internet. *Journal of Management Information Systems 22,* 2 (Fall 2005) , 57-89.

Berger, B., Matt, C., Steininger, D., and Hess, T. It Is Not Just about Competition with "Free": Differences Between Content Formats in Consumer Preferences and Willingness to Pay. *Journal of Management Information Systems 32,* 3 (July 2015), 105-128.

Bhattacharjee, S., Gopal, R., Lertwachara, K., and Marsden, J.R. Consumer Search and Retailer Strategies in the Presence of Online Music Sharing. *Journal of Management Information Systems 23,* 2 (Summer 2006), 129-159.

BI Intelligence. Illegal Streaming Is Dominating Online Piracy. *Business Insider* (August 1, 2016).

Biemann, T., and Datta, D. Analyzing Sequence Data: Optimal Matching in Management Research. *Organizational Research Methods 17,* 1 (January 2014), 51-76.

Brynjolfsson, E., Hu, Y.J., and Simester, D. Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales. *Management Science 57*, 8 (June 2011), 1373–1386.

Cain, K., Harlow, S., Little, R., Nan, B., Yosef, M., Taffe, J., and Elliott, M. Bias Due To Left Truncation and Left Censoring in Longitudinal Studies of Developmental and Disease Processes. *American Journal of Epidemiology 173*, 9 (May 2011), 1078-1084.

Cameron, A.C., and Trivedi, P.K. *Regression Analysis of Count Data*, 2nd Ed. Econometric Society Monograph No. 53, Cambridge, UK: Cambridge University Press, 1998.

Carr, D. Giving Viewers What They Want. *New York Times* (February 24, 2013).

Chang, J., and Lee, W. Efficient Mining Method for Retrieving Sequential Patterns Over Online Data Streams. *Journal of Information Science 31*, 5 (October 2005), 420-432.

Chang, M.R., Kauffman, R.J., and Kwon, Y. Understanding the Paradigm Shift to Computational Social Science in The Presence Of Big Data. *Decision Support Systems 63* (2014), 67-80.

Chellappa, R., and Shivendu, S. Managing Piracy: Pricing and Sampling Strategies for Digital Experience Goods in Vertically Segmented Markets. *Information Systems Research 16*, 4 (December 2005), 400-417.

Chen, H., Chiang, R., and Storey, V. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly 36*, 4 (December 2012), 1165–1188.

Chintagunta, P.K., and Dong, X. Hazard/ Survival Models in Marketing. In R. Grover and M. Vriens (eds.), *The Handbook of Marketing Research.* Thousand Oaks: Sage, 2006, pp. 441-454.

Clemons, E.K., Gao, G., and Hitt, L. When Online Reviews Meet Hyper Differentiation: A Study of the Craft Beer Industry. *Journal of Management Information Systems 23*, 2 (Fall 2006), 149-171.

Clemons, E.K., Gu, B., and Spitler, R. Hyper-Differentiation Strategies: Delivering Value, Retaining Profits. In R.H. Sprague, Jr. (ed.), *Proceedings of the Thirty-Sixth Hawaii International Conference on System Science*, Los Alamitos, CA: IEEE Computer Society Press, 2003.

Clemons, E.K., Spitler, R., Gu, B., and Markopoulos, P. Information, Hyperdifferentiation, and Delight: The Value of Being Different. In S. Bradley and R. Austin (eds.), *The Broadband Explosion: Leading Thinkers on the Promise of a Truly Interactive World.* Boston, MA: Harvard Business School Press, 2005, pp. 137–164.

Colby, C., and Bell, K. The On-Demand Economy Is Growing, and Not Just for the Young and Wealthy. *Hbr.org* (April 14, 2016).

Cooper, G. The Computational Complexity of Probabilistic Inference using Bayesian Belief Networks. *Artificial Intelligence 42*, 2 (March 1990), 393-405.

Cragg, J.G. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica 39*, 5 (September 1971), 829-844.

David, F.N., and Johnson, N.L. Statistical Treatment Of Censored Data I: Fundamental Formulae. *Biometrika 41*, 1 (June 1954), 228-240.

Davidson, R., and Mackinnon, J.G. The Power of Bootstrap and Asymptotic Tests. *Journal of Econometrics 133*, 2 (2006), 421-441.

Davies, D.L., Bouldin, D.W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence 1*, 2 (February 1979), 224-227.

De Matos, M., Ferreira, P., Smith, M.D., and Telang, R. Culling the Herd: Using Real-World Randomized Experiments to Measure Social Bias with Known Costly Goods. *Management Science 62*, 9 (February 2016), 2563-2580.

Dehejia, R.H., and Wahba, S. Propensity Score-Matching Methods for Non-Experimental Causal Studies. *Review of Economics and Statistics 84*, 1 (February 2002), 151-161.

Demski, J. *Information Analysis.* Reading, MA: Addison-Wesley, 1980.

Dey, D., Lahiri, A., and Liu, D. Consumer Learning and Time-Locked Trials of Software Products. *Journal of Management Information Systems 30*, 2 (Fall 2013), 239-268.

Dimoka, A., Hong, Y., and Pavlou, P.A. On Product Uncertainty in Online Markets: Theory and Evidence. *MIS Quarterly 36*, 2 (June 2012), 395-426.

Ding, P., VanderWeele, T., and Robins, J. Instrumental Variables as Bias Amplifiers with General Outcome and Confounding. *Biometrika 104*, 2 (June 2017), 291-302.

Dunn, J. The Rise of Music Streaming Services Hasn't Killed Music Piracy. *Business Insider* (April 17, 2017).

Dupont, W.D. Multiple Poisson Regression. Chapter 9 in *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data.* Cambridge, UK: Cambridge University Press, 2002, pp. 295-318.

Efron, B. Censored Data And The Bootstrap. *Journal of American Statistical Association 76*, 374 (June 1981), 312-319.

Efron, B., and Tibshirani, R. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

Eiter, T. and Mannila, H. Distance Measures for Point Sets and Their Computation. *Acta Informatica 34*, 2 (February 1997), 109-133.

Faugère, C., and Kumar, G.T. Designing Free Software Samples: A Game Theoretic Approach. *Information Technology and Management 8*, 4 (September 2006), 263-278.

Freedman, A.M. Use of Free Product Samples Wins New Favor as Sales Tool. *Wall Street Journal* (August 28, 1986).

Galer, S. What To Do with All the IOT Sensor Data Your Business Is Collecting. *Forbes.com* (October 10, 2017).

Gemici, S., Rojewski, J.W., and In-Heok, L. Use of Propensity Score Matching for

Training Research with Observational Data. *International Journal of Training Research 10*, 3 (June 2012), 219-232.

Ghose, A. Internet Exchanges for Used Goods: An Empirical Analysis of Trade Patterns and Adverse Selection. *MIS Quarterly 33*, 2 (June 2009), 163–291.

Gilovich, T., and Medvec, V.H. The Experience of Regret: What, When, and Why. *Psychology Review 102*, 2 (April 1995), 379–395.

Greene, W.H. *Econometric Analysis*. London, UK: Pearson, 2012.

Gross, S., Lai, T. Bootstrap Methods for Truncated and Censored Data. *Statistical Sinica 6*, 3 (July 1996), 509-530.

Gurmu, S., and Trivedi, P.K. Excess Zeros in Count Models for Recreational Trips. *Journal of Business and Economic Statistics 14*, 4 (October 1996), 469-477.

Halbheer, D., Stahl, F., Koenigsberg, O., and Lehmann, D.R. Choosing a Digital Content Strategy: How Much Should Be Free? *International Journal of Research in Marketing 31*, 2 (June 2014), 196-206.

Haubl, G., and Trifts, V. Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids. *Management Science 19*, 1 (February 2000), 4-21.

Heiman, A., McWilliams, B., Shen, Z., and Zilberman, D. Learning and Forgetting: Modeling Optimal Product Sampling Over Time. *Management Science 47*, 4 (April 2001), 532-546.

Hevner, A., March, S. , Park, J., and Ram, S. Design Science in Information Systems Research. *MIS Quarterly 28*, 1 (March 2004), 75–105.

Hilbe, J.M. *Negative Binomial Regression*, 2nd ed. Cambridge, UK: Cambridge Univ. Press, 2011.

Hoang, A.P., Kauffman, R.J. Experience Me! The Impact of Content Sampling Strategies on the Marketing of Digital Entertainment Goods. In T. Bui and R. Sprague (eds.), HICSS, IEEE Comp. Soc. Press, Washington, DC, 2016.

Hoang, A.P., Kauffman, R.J. Extending Propensity Score Matching to Capture Censored Observations for Causal Explanation. *Workshop on Information Systems and Economics*, Seoul, Korea, Dec. 2017.

Holloway, K. The Reasons You Can't Stop Binge Watching. *Alternet.com* (December 30, 2015).

Holmes, C., Boche, D., Wilkinson, D., Yadegarfar, G., Hopkins, V., Bayer, A., Jones, R.W., Bullock, R., Love, S., Neal, J.W., Zotova, E., and Nicoll, J.A. Long-Term Effects of Aβ 42 Immunisation in Alzheimer's Disease: Follow-Up of a Randomised, Placebo-Controlled Phase I Trial. *The Lancet 372*, 9634 (July 2008), 216-223.

Hong, Y., and Pavlou, P. A. Product Fit Uncertainty in Online Markets: Nature, Effects, and Antecedents. *Information Systems Research 25*, 2 (April 2014), 328–344.

Howard, J. Americans Devote More Than 10 Hours a Day to Screen Time, and

Growing. CNN.com (July 29, 2016).

Howison, J., Wiggins, A., and Crowston, K. Validity Issues in the Use of SNA with Digital Trace Data. *Journal of the Association of Information Systems 12*, 12 (January 2011), 767-797.

Huang, P., Lurie, N,. and Mitra, S. Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods. *Journal Of Marketing 73*, 2 (March 2009), 55-69.

Ibrahim, J.G., and Chen, M.H., and Sinha, D. *Bayesian Survival Analysis*, New York: Springer, 2001.

Jenkins, S. We're Over the Digital Revolution – This Is The Age of Experience. *TheGuardian.com* (February 2, 2017).

Johnson, E.J., Moe, W.W., Fader, P.S., Bellman, S., and Lohse, G.L. On the Depth and Dynamics of Online Search Behavior. *Management Science 50*, 3 (March 2004), 299–308.

Jones, R., and Mendelson, H. Information Goods vs. Industrial Goods: Cost Structure and Competition. *Management Science 57*, 1 (January 2011), 164-176.

Kamins, M.A., Folkes, V.S., and Fedorikhin, A. Promotional Bundles and Consumers' Price Judgments: When the Best Things in Life Are Not Free. *Journal of Consumer Research 36*, 4 (November 2009), 660-670.

Kastranekes, J. Netflix knows the exact episode of a TV show that gets you hooked. *TheVerge.com* (September 23, 2015).

Ketter, W., Peters, M., Collins, J., and Gupta. Competitive Benchmarking: An IS Research Approach to Address Wicked Problems with Big Data Analytics. *MIS Quarterly 40*, 4 (December 2016), 1057-1080.

Kwark, Y., Chen, J., and Raghunathan, S. Online Product Reviews: Implications for Retailers and Competing Manufacturers. *Information Systems Research 25*, 1 (March 2014), 93–110.

Lafayette, J. Threat becomes profit center as TV leverages technology. BroadcastingCable.com (January 6, 2014).

Lawal, B. Zero-Inflated Count Regression Models with Applications to Some Examples. *Quality and Quantity 46*, 1 (January 2012), 19-38.

Lenis, D., Nguyen, T., Dong, N., Stuart, E. It's All about Balance: PSM in the Context of Complex Survey Data. Biostatistics, 2018, forthcoming.

Li, T., Kauffman, R.J., Van Heck, E., Vervest, P., and Dellaert, B.G.C. Consumer Informedness And Firm Information Strategy. *Information Systems Research 25*, 2 (May 2014), 345-363.

Li, X. Could Deal Promotion Improve Merchants' Online Reputations? The Moderating Role of Prior Reviews. *Journal of Management Information Systems 33*, 1 (June 2016), 171-201.

Liebeskind, J., and Rumelt, R.P. Markets for Experience Goods with Performance

Uncertainty. *The RAND Journal of Economics 20*, 4 (Winter 1989), 601-621.

Liebowitz, S.J., and Zentner, A. Clash of the Titans: Does Internet Use Reduce Television Viewing? *Review of Economics and Statistics 94*, 1 (February 2012), 234-245.

Lin, C.A. Modeling The Gratification-Seeking Process of TV Viewing. *Human Communication Research 20*, 2 (December 1993), 224-244.

Maloney, M., Johnson, S., and Zellmer-Bruhn, M. Assessing Group-Level Constructs Under Missing Data Conditions: A Monte Carlo Simulation. *Small Group Research 41*, 3 (May 2010), 281-307.

Markopoulos, P.M. Product Information Dissemination in Internet Markets and Markets for Product Information. Unpublished doctoral thesis, Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 2004.

Markopoulos, P.M., and Clemons, E.K. Reducing Buyers' Uncertainty about Taste-Related Product Attributes. *Journal of Management Information Systems 30*, 2 (Fall 2013), 269-299.

Martens, D., Provost, F., Clark, J., and Junqué de Fortuny, E. Mining Massive Fine-Grained Behavior to Improve Predictive Analytics. *MIS Quarterly 40*, 4, (December 2016), 869-888.

Matt, C., and Hess, T. Product Fit Uncertainty and Its Effects on Vendor Choice: An Experimental Study. *Electronic Markets 26*, 1 (February 2016), 83-93.

McAlister, L., and Pessemier, E. Variety Seeking Behavior: An Interdisciplinary Review. *Journal of Consumer Research 9*, 3 (December 1982), 311-322.

McGuinness, D., Gendall, P., and Mathew, S. The Effect Of Product Sampling On Product Trial, Purchase And Conversion. *International Journal of Advertising 11*, 1 (January 1992), 83-92.

Mehta, N., Rajiv, S., and Srinivasan, K. Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation. *Marketing Science 22*, 1 (Winter 2003), 58-84.

Mordor Intelligence. Global Video On Demand Market by Revenue Model, Platform, Applications, Industry Verticals, Geography and Vendors: Market Shares, Forecasts and Trends (2015-2020). Hyderabad, India, October 2015.

Moretti, E. Social Learning and Peer Effects in Consumption: Evidence from Movie Sales. *Review of Economic Studies 78*, 1 (January 2011), 356-393.

Müller, O., Junglas, I., vom Brocke, J and Debortoli, S. Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines. *European Journal of Information Systems 25*, 4, (July 2016), 289-302.

Nathanson, J. The Economics of a Hit TV Show. *Priceconomics.com* (October 17, 2013).

Nelson, P. Information and Consumer Behavior. *Journal of Political Economy 78*, 2 (March-April 1970), 311-329.

*New York Times*. The Best and Worst Moments of the 2017 Emmys. (September 18, 2017).

Newman, D. A. Missing Data Techniques and Low Response Rates: The Role of Systematic Nonresponse Parameters. In Lance, C. E., Vandenberg, R. J. (eds.), Statistical and Methodological Myths and Urban Legends: Doctrine, Verity, and Fable in the Organizational and Social Sciences. New York, NY: Routledge, 2009, pp. 7-36.

Newman, D. A., and Sin, H. P. How Do Missing Data Bias Estimates of Within Group Agreement? *Organizational Research Methods 12*, 1 (2009), 113-147.

*Newswire*. Netflix Declares Binge Watching is the New Normal: Study Finds 73% of TV Streamers Feel Good About It. (December 13, 2013).

Niculescu, M., and Wu, D. J. Economics of Free under Perpetual Licensing Implications for the Software Industry. *Information Systems Research 25*, 1, (March 2014), 173-199.

O'Kane, S. Spotify Unveils Touch Preview, A Beautiful New Music Discovery Tool. *TheVerge.com* (January 22, 2015).

Oestreicher-Singer, G., and Zalmanson, L. Content or Community? A Digital Business Strategy for Content Providers in the Social Age. *MIS Quarterly 37*, 2 (June 2013), 591-616.

Pavlou, P., Liang, H., and Xue, Y. Understanding and Mitigating Uncertainty in Online Exchange Relationships: A Principal-Agent Perspective. *MIS Quarterly 31*, 1 (March 2007), 105–136.

Phillips-Wren, G., Iyer, L., Kulkarni, U., and Ariyachandra, T. 2015. Business Analytics in the Context of Big Data: A Roadmap for Research. *Communications of the Association for Information Systems* 37, (August 2015), 23.

Pinsker, J. The Psychology Behind Costco's Free Samples. *The Atlantic* (October 2014), 1-6.

Pirracchio, R., Resche-Rigon, M., and Chevret S. Evaluation of the Propensity Score Methods for Estimating Marginal Odds Ratios in Case of Small Sample Size. *BMC Medical Research Methodology 12*, 1 (May 2012), 70-79.

Platzer, M., and Reutterer, T. Ticking Away the Moments: Timing Regularity Helps To Better Predict Customer Activity. *Marketing Science 35*, 5 (September 2016), 779-799.

Prentice, R., and Gloeckler, L. Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics 34*, 1 (March 1978), 57-67.

Qiu, X., Oliveira, D., Shirazi, A., Flammini, A., and Menczer, F. Limited Individual Attention and Online Virality of Low-Quality Information. *Nature Human Behavior*, 2017, forthcoming.

Rao, A.R., and Monroe, K.B. Causes and Consequences of Price Premiums. *Journal of Business 69*, 4 (October 1996), 511–535.

Read, S. J., Druian, P. R., and Miller, L. C. The Role of Causal Sequence in the Meaning Of Actions. *British Journal of Social Psychology 28*, 4 (December 1989), 341-351.

Rode, A. Literature review: Non-Unitary Models of The Household (Theory And Evidence). Working paper, U. California, Santa Barbara, CA, 2011.

Rosenbaum, P. Optimal Matching for Observational Studies. *Journal of the American Statistical Association 84*, 408 (December 1989), 1024-1032.

Rosenbaum, P., and Rubin, D. Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity. *American Statistician 39*, 1 (February 1985), 33-38.

Rosenbaum, P., and Rubin, D. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika 70*, 1 (April 1983), 41-55.

Ross, D.S. Probabilistic Inference and Influence Diagrams. *Operation Research 36*, 4, (August 1988), 589-604.

Rossi, P.E., and Allenby, G.M. Bayesian Statistics and Marketing. *Marketing Science 22*, 3 (August 2003), 304-328.

Rubin, A.M. Television Uses And Gratifications: The Interactions of Viewing Patterns and Motivations. *Journal of Broadcasting 27*, 1 (January 1983), 37-51.

Rubin, D. Matching to Remove Bias in Observational Studies. *Biometrics 29*, 1 (March 1973), 159-183.

Russell, C.A., Norman, A.T., and Heckler, S.E. The Consumption of Television Programming: Development and Validation of the Connectedness Scale. *Journal of Consumer Research 31*, 1 (June 2004), 150-161.

Schafer, J.L. *Analysis of Incomplete Multivariate Data*. Bota Raton, FL: Chapman and Hall, 1999.

Schaubel, D., Cai, J. Multiple Imputation Methods for Recurrent Event Data with Missing Event Category. *The Canadian Journal of Statistics 34*, 4 (December 2006), 677-692.

Shapiro, C., and Varian, H.R. *Information Rules: A Strategic Guide to the Network Economy*. Boston, MA: Harvard Business School Press, 1999.

Shih W. J. Problems in Dealing with Missing Data and Informative Censoring in Clinical Trials. *Current Controlled Trials in Cardiovascular Medicine 3*, 1 (January 2002), 4.

Tatti, N. Distances Between Data Sets Based on Summary Statistics. *Journal of Machine Learning Research 8* (Decemember 2007), 131-154.

Thomas, L. Retail's Not Dead, and Physical Stores Still Matter, Goldman Says. *CNBC.com* (November 21, 2017).

Uncles, M., Ehrenberg, A., and Hammond, K. Patterns of Buyer Behavior: Regularities, Models, and Extensions. *Marketing Science 14*, 3 Supplement (August 1995), G71-G78.

Vuong, Q. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica 57*, 2 (March 1989), 307–333.

Wei L., and Shih, W.J. Partial Imputation Approach to Analysis of Repeated Measurements with Dependent Dropouts. *Statistics in Medicine 20*, 8 (April 2001), 1197-1214.

Whitler, K. Why Word of Mouth Marketing is the Most Important Social Media. *Forbes.com* (June 17, 2014).

Wu, M., Carroll, R. Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics 44*, 1 (March 1988), 175-188.

Wu, S., and Chen, P. Versioning and Piracy Control for Digital Information Goods. *Operations Research 56*, 1 (January 2008), 157–172.

Yahav, I., Shmueli, G., and Mani, D. A Tree-based Approach for Addressing Self-Selection in Impact Studies with Big Data. *MIS Quarterly 40*, 4 (Deember 2016), 819-848.

Yap, G.E., Tan, A.H., Pang, H.H. Explaining Inferences in Bayesian Networks. *Applied Intelligence 29*, 3 (December 2008), 263-278.

## APPENDIX A. SUMMARY STATISTICS FOR ALL CENSORED DATA SUBSETS

### Table A1. Summary Statistics Censored Subsets 1-8

| VARIABLES | CENSORED SUBSET | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| *#Cust* | 12,034 | 12,310 | 12,625 | 12,946 | 13,208 | 13,577 | 13,964 | 14,369 |
| *#AllVis* | 33,730 | 34,481 | 35,279 | 36,049 | 36,708 | 37,793 | 38,889 | 39,990 |
| *#Purch* | 11,122 | 11,413 | 11,699 | 12,009 | 12,266 | 12,623 | 13,015 | 13,411 |
| *AvgVis* | 2.803 | 2.801 | 2.794 | 2.785 | 2.779 | 2.784 | 2.785 | 2.783 |
| *AvgPur* | 0.924 | 0.927 | 0.927 | 0.928 | 0.929 | 0.930 | 0.932 | 0.933 |
| *Conver* | 0.330 | 0.331 | 0.332 | 0.333 | 0.334 | 0.334 | 0.335 | 0.335 |
| **Traffic source of customers' visits** | | | | | | | | |
| *Direct* | 11,480 | 11,754 | 12,013 | 12,271 | 12,515 | 12,898 | 13,291 | 13,656 |
| *SEngin* | 06,062 | 6,204 | 6,336 | 6,452 | 6,543 | 6,755 | 6,936 | 7,104 |
| *SEnginAd* | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 |
| *ComparSite* | 7,054 | 7,231 | 7,408 | 7,605 | 7,777 | 8,002 | 8,270 | 8,536 |
| *AffilMktg* | 1,086 | 1,104 | 1,128 | 1,161 | 1,182 | 1,245 | 1,285 | 1,329 |
| *Adwords* | 6,117 | 6,225 | 6,351 | 6,491 | 6,602 | 6,770 | 6,962 | 7,183 |
| *Display* | 135 | 138 | 143 | 146 | 149 | 153 | 153 | 155 |
| *Email* | 673 | 675 | 712 | 719 | 720 | 721 | 721 | 718 |
| *SocMedAdv* | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| *RSSFeed* | 62 | 63 | 64 | 66 | 66 | 69 | 72 | 73 |
| *OnlFolder* | 93 | 93 | 93 | 93 | 93 | 94 | 97 | 98 |
| *Other* | 921 | 947 | 984 | 998 | 1,014 | 1,039 | 1,055 | 1,091 |
| **Data Censoring** | | | | | | | | |
| *#VisCensor* | 205 | 210 | 222 | 239 | 246 | 239 | 242 | 254 |
| *#VisPurchCensor* | 925 | 911 | 940 | 951 | 956 | 968 | 963 | 972 |
| *%VisCensor* | 1.70% | 1.71% | 1.76% | 1.85% | 1.86% | 1.76% | 1.73% | 1.77% |
| *%VisPurchCensor* | 7.69% | 7.40% | 7.45% | 7.35% | 7.24% | 7.13% | 6.90% | 6.76% |
| *%NonCensor* | 90.61% | 90.89% | 90.80% | 90.81% | 90.90% | 91.11% | 91.37% | 91.47% |

## Table A2. Summary Statistics Censored Subsets 9-16

| VARIABLES | CENSORED SUBSET | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| *#Cust* | 14,720 | 15,027 | 15,333 | 15,680 | 16,147 | 16,589 | 16,972 | 17,373 |
| *#AllVis* | 40,915 | 41,798 | 42,489 | 43,331 | 44,586 | 45,738 | 46,785 | 47,820 |
| *#Purch* | 13,754 | 14,054 | 14,326 | 14,647 | 15,107 | 15,530 | 15,914 | 16,312 |
| *AvgVis* | 2.780 | 2.782 | 2.771 | 2.763 | 2.761 | 2.757 | 2.757 | 2.753 |
| *AvgPur* | 0.934 | 0.935 | 0.934 | 0.934 | 0.936 | 0.936 | 0.938 | 0.939 |
| *Conver* | 0.336 | 0.336 | 0.337 | 0.338 | 0.339 | 0.340 | 0.340 | 0.341 |
| **Traffic source of customers' visits** | | | | | | | | |
| *Direct* | 13,935 | 14,189 | 14,394 | 14,650 | 15,059 | 15,449 | 15,757 | 16,062 |
| *SEngin* | 7,245 | 7,415 | 7,519 | 7,649 | 7,859 | 8,046 | 8,242 | 8,393 |
| *SEnginAd* | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 30 |
| *ComparSite* | 8,741 | 8,953 | 9,140 | 9,369 | 9,678 | 9,958 | 10,222 | 10,486 |
| *AffilMktg* | 1,374 | 1,394 | 1,427 | 1,452 | 1,501 | 1,551 | 1,585 | 1,620 |
| *Adwords* | 7,340 | 7,501 | 7,634 | 7,804 | 8,017 | 8,229 | 8,399 | 8,619 |
| *Display* | 157 | 162 | 164 | 166 | 177 | 179 | 183 | 190 |
| *Email* | 777 | 810 | 815 | 815 | 824 | 825 | 865 | 881 |
| *SocMedAdv* | 18 | 19 | 20 | 21 | 25 | 25 | 25 | 25 |
| *RSSFeed* | 77 | 79 | 79 | 82 | 84 | 88 | 89 | 93 |
| *OnlFolder* | 99 | 100 | 101 | 109 | 116 | 128 | 141 | 148 |
| *Other* | 1,123 | 1,147 | 1,167 | 1,185 | 1,217 | 1,231 | 1,248 | 1,273 |
| **Data Censoring** | | | | | | | | |
| *#VisCensor* | 249 | 283 | 318 | 361 | 354 | 392 | 415 | 441 |
| *#VisPurchCensor* | 980 | 988 | 1,022 | 1,048 | 1,055 | 1,074 | 1,073 | 1,077 |
| *%VisCensor* | 1.69% | 1.88% | 2.07% | 2.30% | 2.19% | 2.36% | 2.45% | 2.54% |
| *%VisPurchCensor* | 6.66% | 6.57% | 6.67% | 6.68% | 6.53% | 6.47% | 6.32% | 6.20% |
| *%NonCensor* | 91.65% | 91.54% | 91.26% | 91.01% | 91.27% | 91.16% | 91.23% | 91.26% |

## Table A3. Summary Statistics Censored Subsets 17-24

| Variables | Censored Subset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| *#Cust* | 17,696 | 18,024 | 18,437 | 18,949 | 19,361 | 19,752 | 20,146 | 20,458 |
| *#AllVis* | 48,675 | 49,450 | 50,373 | 51,662 | 52,712 | 53,698 | 54,687 | 55,480 |
| *#Purch* | 16,637 | 16,921 | 17,314 | 17,804 | 18,229 | 18,630 | 19,021 | 19,328 |
| *AvgVis* | 2.751 | 2.744 | 2.732 | 2.726 | 2.723 | 2.719 | 2.715 | 2.712 |
| *AvgPur* | 0.940 | 0.939 | 0.939 | 0.940 | 0.942 | 0.943 | 0.944 | 0.945 |
| *Conver* | 0.342 | 0.342 | 0.344 | 0.345 | 0.346 | 0.347 | 0.348 | 0.348 |
| **Traffic source of customers' visits** | | | | | | | | |
| *Direct* | 16,298 | 16,543 | 16,830 | 17,221 | 17,566 | 17,902 | 18,240 | 18,514 |
| *SEngin* | 8,549 | 8,658 | 8,777 | 9,013 | 9,171 | 9,327 | 9,475 | 9,590 |
| *SEnginAd* | 31 | 32 | 32 | 33 | 33 | 35 | 35 | 35 |
| *ComparSite* | 10,694 | 10,916 | 11,199 | 11,534 | 11,814 | 12,069 | 12,301 | 12,491 |
| *AffilMktg* | 1,659 | 1,679 | 1,710 | 1,754 | 1,785 | 1,803 | 1,827 | 1,848 |
| *Adwords* | 8,785 | 8,917 | 9,090 | 9,314 | 9,514 | 9,701 | 9,874 | 10,013 |
| *Display* | 194 | 199 | 203 | 211 | 217 | 226 | 234 | 240 |
| *Email* | 904 | 919 | 927 | 924 | 927 | 913 | 950 | 972 |
| *SocMedAdv* | 25 | 25 | 25 | 26 | 26 | 28 | 29 | 29 |
| *RSSFeed* | 99 | 103 | 104 | 105 | 108 | 108 | 109 | 109 |
| *OnlFolder* | 152 | 154 | 159 | 180 | 187 | 197 | 209 | 215 |
| *Other* | 1,285 | 1,305 | 1,317 | 1,347 | 1,364 | 1,389 | 1,404 | 1,424 |
| **Data Censoring** | | | | | | | | |
| *#VisCensor* | 468 | 484 | 519 | 535 | 565 | 587 | 624 | 657 |
| *#VisPurchCensor* | 1,077 | 1,121 | 1,142 | 1,164 | 1,154 | 1,146 | 1,150 | 1,155 |
| *%VisCensor* | 2.64% | 2.69% | 2.81% | 2.82% | 2.92% | 2.97% | 3.10% | 3.21% |
| *%VisPurchCensor* | 6.09% | 6.22% | 6.19% | 6.14% | 5.96% | 5.80% | 5.71% | 5.65% |
| *%NonCensor* | 91.27% | 91.10% | 90.99% | 91.03% | 91.12% | 91.23% | 91.19% | 91.14% |

## Table A4. Summary Statistics Censored Subsets 25-32

| VARIABLES | CENSORED SUBSET | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| *#Cust* | 20,794 | 21,202 | 21,717 | 22,175 | 22,644 | 23,053 | 23,415 | 23,356 |
| *#AllVis* | 56,245 | 57,015 | 58,180 | 59,143 | 60,230 | 61,225 | 62,037 | 61,994 |
| *#Purch* | 19,650 | 20,044 | 20,570 | 21,013 | 21,535 | 21,951 | 22,375 | 22,402 |
| *AvgVis* | 2.705 | 2.689 | 2.679 | 2.667 | 2.660 | 2.656 | 2.649 | 2.654 |
| *AvgPur* | 0.945 | 0.945 | 0.947 | 0.948 | 0.951 | 0.952 | 0.956 | 0.959 |
| *Conver* | 0.349 | 0.352 | 0.354 | 0.355 | 0.358 | 0.359 | 0.361 | 0.361 |
| **Traffic source of customers' visits** | | | | | | | | |
| *Direct* | 18,775 | 19,021 | 19,376 | 19,670 | 20,028 | 20,373 | 20,650 | 20,653 |
| *SEngin* | 9,704 | 9,826 | 10,012 | 10,160 | 10,333 | 10,502 | 10,607 | 10,579 |
| *SEnginAd* | 35 | 35 | 35 | 34 | 32 | 31 | 30 | 30 |
| *ComparSite* | 12,703 | 12,945 | 13,299 | 13,604 | 13,892 | 14,163 | 14,378 | 14,442 |
| *AffilMktg* | 1,886 | 1,924 | 1,962 | 1,979 | 2,013 | 2,034 | 2,054 | 2,048 |
| *Adwords* | 10,110 | 10,204 | 10,390 | 10,557 | 10,749 | 10,890 | 11,044 | 10,979 |
| *Display* | 241 | 244 | 247 | 253 | 255 | 259 | 266 | 269 |
| *Email* | 981 | 984 | 983 | 983 | 1,006 | 1,016 | 1,032 | 1,031 |
| *SocMedAdv* | 29 | 29 | 30 | 29 | 30 | 30 | 32 | 32 |
| *RSSFeed* | 111 | 118 | 124 | 127 | 128 | 131 | 129 | 129 |
| *OnlFolder* | 221 | 227 | 230 | 240 | 244 | 246 | 250 | 251 |
| *Other* | 1,449 | 1,458 | 1,492 | 1,507 | 1,520 | 1,550 | 1,565 | 1,551 |
| **Data Censoring** | | | | | | | | |
| *#VisCensor* | 712 | 746 | 778 | 797 | 830 | 854 | 903 | 964 |
| *#VisPurchCensor* | 1,171 | 1,186 | 1,176 | 1,207 | 1,196 | 1,232 | 1,277 | 1,282 |
| *%VisCensor* | 3.42% | 3.52% | 3.58% | 3.59% | 3.67% | 3.70% | 3.86% | 4.13% |
| *%VisPurchCensor* | 5.63% | 5.59% | 5.42% | 5.44% | 5.28% | 5.34% | 5.45% | 5.49% |
| *%NonCensor* | 90.94% | 90.89% | 91.00% | 90.96% | 91.05% | 90.95% | 90.69% | 90.38% |

## Table A5. Summary Statistics Censored Subsets 33-40

| Variables | Censored Subset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| *#Cust* | 23,380 | 23,446 | 23,483 | 23,625 | 23,651 | 23,718 | 23,598 | 23,465 |
| *#AllVis* | 62,074 | 62,183 | 62,296 | 62,554 | 62,671 | 62,850 | 62,490 | 62,096 |
| *#Purch* | 22,455 | 22,527 | 22,580 | 22,721 | 22,768 | 22,830 | 22,691 | 22,613 |
| *AvgVis* | 2.655 | 2.652 | 2.653 | 2.648 | 2.650 | 2.650 | 2.648 | 2.646 |
| *AvgPur* | 0.960 | 0.961 | 0.962 | 0.962 | 0.963 | 0.963 | 0.962 | 0.964 |
| *Conver* | 0.362 | 0.362 | 0.362 | 0.363 | 0.363 | 0.363 | 0.363 | 0.364 |
| **Traffic source of customers' visits** | | | | | | | | |
| *Direct* | 20,715 | 20,781 | 20,822 | 20,849 | 20,938 | 20,982 | 20,837 | 20,714 |
| *SEngin* | 10,574 | 10,587 | 10,635 | 10,687 | 10,712 | 10,745 | 10,647 | 10,537 |
| *SEnginAd* | 28 | 27 | 25 | 23 | 23 | 23 | 21 | 19 |
| *ComparSite* | 14,481 | 14,521 | 14,539 | 14,672 | 14,688 | 14,729 | 14,688 | 14,690 |
| *AffilMktg* | 2,032 | 2,026 | 2,029 | 2,026 | 2,012 | 2,030 | 2,031 | 2,029 |
| *Adwords* | 10,999 | 11,009 | 11,023 | 11,077 | 11,089 | 11,060 | 10,991 | 10,884 |
| *Display* | 271 | 272 | 278 | 283 | 288 | 293 | 295 | 286 |
| *Email* | 1,028 | 1,007 | 993 | 967 | 960 | 1,018 | 1,030 | 1,031 |
| *SocMedAdv* | 31 | 31 | 31 | 31 | 31 | 31 | 29 | 29 |
| *RSSFeed* | 130 | 131 | 132 | 132 | 133 | 132 | 128 | 126 |
| *OnlFolder* | 252 | 253 | 251 | 249 | 249 | 249 | 241 | 227 |
| *Other* | 1,533 | 1,538 | 1,538 | 1,558 | 1,548 | 1,558 | 1,552 | 1,524 |
| **Data Censoring** | | | | | | | | |
| *#VisCensor* | 955 | 984 | 987 | 996 | 1,021 | 1,069 | 1,051 | 1,026 |
| *#VisPurchCensor* | 1,253 | 1,247 | 1,231 | 1,232 | 1,211 | 1,216 | 1,235 | 1,180 |
| *%VisCensor* | 4.08% | 4.20% | 4.20% | 4.22% | 4.32% | 4.51% | 4.45% | 4.37% |
| *%VisPurchCensor* | 5.36% | 5.32% | 5.24% | 5.21% | 5.12% | 5.13% | 5.23% | 5.03% |
| *%NonCensor* | 90.56% | 90.48% | 90.55% | 90.57% | 90.56% | 90.37% | 90.31% | 90.60% |

## Table A6. Summary Statistics Censored Subsets 41-48

| VARIABLES | CENSORED SUBSET | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| *#Cust* | 23,370 | 23,315 | 23,351 | 23,388 | 23,397 | 23,251 | 23,250 | 23,361 |
| *#AllVis* | 61,973 | 61,901 | 62,029 | 62,150 | 62,162 | 61,762 | 61,635 | 61,848 |
| *#Purch* | 22,577 | 22,585 | 22,645 | 22,702 | 22,715 | 22,573 | 22,620 | 22,767 |
| *AvgVis* | 2.652 | 2.655 | 2.656 | 2.657 | 2.657 | 2.656 | 2.651 | 2.647 |
| *AvgPur* | 0.966 | 0.969 | 0.970 | 0.971 | 0.971 | 0.971 | 0.973 | 0.975 |
| *Conver* | 0.364 | 0.365 | 0.365 | 0.365 | 0.365 | 0.365 | 0.367 | 0.368 |
| **Traffic source of customers' visits** | | | | | | | | |
| *Direct* | 20,645 | 20,586 | 20,651 | 20,670 | 20,633 | 20,435 | 20,391 | 20,423 |
| *SEngin* | 10,490 | 10,470 | 10,473 | 10,459 | 10,476 | 10,395 | 10,330 | 10,388 |
| *SEnginAd* | 19 | 18 | 17 | 17 | 15 | 11 | 10 | 7 |
| *ComparSite* | 14,777 | 14,858 | 14,923 | 14,981 | 14,988 | 14,960 | 14,991 | 15,091 |
| *AffilMktg* | 2,026 | 2,011 | 2,015 | 2,044 | 2,040 | 2,033 | 2,046 | 2,058 |
| *Adwords* | 10,862 | 10,836 | 10,821 | 10,846 | 10,852 | 10,795 | 10,741 | 10,758 |
| *Display* | 285 | 290 | 293 | 297 | 300 | 306 | 311 | 314 |
| *Email* | 968 | 940 | 934 | 930 | 975 | 967 | 978 | 963 |
| *SocMedAdv* | 30 | 29 | 30 | 31 | 31 | 31 | 30 | 30 |
| *RSSFeed* | 128 | 128 | 130 | 130 | 131 | 131 | 128 | 128 |
| *OnlFolder* | 217 | 207 | 204 | 203 | 191 | 185 | 181 | 180 |
| *Other* | 1,526 | 1,528 | 1,538 | 1,542 | 1,530 | 1,513 | 1,498 | 1,508 |
| **Data Censoring** | | | | | | | | |
| *#VisCensor* | 1,020 | 982 | 970 | 1,019 | 1,048 | 1,071 | 1,074 | 1,053 |
| *#VisPurchCensor* | 1,121 | 1,058 | 1,034 | 1,014 | 1,010 | 1,005 | 957 | 921 |
| *%VisCensor* | 4.36% | 4.21% | 4.15% | 4.36% | 4.48% | 4.61% | 4.62% | 4.51% |
| *%VisPurchCensor* | 4.80% | 4.54% | 4.43% | 4.34% | 4.32% | 4.32% | 4.12% | 3.94% |
| *%NonCensor* | 90.84% | 91.25% | 91.42% | 91.31% | 91.20% | 91.07% | 91.26% | 91.55% |

## Table A7. Summary Statistics Censored Subsets 49-56

| Variables | Censored Subset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 |
| *#Cust* | 23,418 | 23,461 | 23,464 | 23,472 | 23,255 | 23,101 | 23,018 | 23,109 |
| *#AllVis* | 61,988 | 62,082 | 62,030 | 62,046 | 61,356 | 61,027 | 60,800 | 60,876 |
| *#Purch* | 22,875 | 22,953 | 23,006 | 23,048 | 22,833 | 22,742 | 22,717 | 22,883 |
| *AvgVis* | 2.647 | 2.646 | 2.644 | 2.643 | 2.638 | 2.642 | 2.641 | 2.634 |
| *AvgPur* | 0.977 | 0.978 | 0.980 | 0.982 | 0.982 | 0.984 | 0.987 | 0.990 |
| *Conver* | 0.369 | 0.370 | 0.371 | 0.371 | 0.372 | 0.373 | 0.374 | 0.376 |
| **Traffic source of customers' visits** | | | | | | | | |
| *Direct* | 20,432 | 20,461 | 20,478 | 20,493 | 20,256 | 20,182 | 20,121 | 20,120 |
| *SEngin* | 10,400 | 10,425 | 10,421 | 10,437 | 10,247 | 10,136 | 10,097 | 10,066 |
| *SEnginAd* | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| *ComparSite* | 15,194 | 15,293 | 15,306 | 15,311 | 15,220 | 15,217 | 15,206 | 15,325 |
| *AffilMktg* | 2,063 | 2,035 | 2,009 | 1,994 | 1,975 | 1,943 | 1,937 | 1,950 |
| *Adwords* | 10,794 | 10,859 | 10,834 | 10,841 | 10,711 | 10,614 | 10,578 | 10,585 |
| *Display* | 318 | 319 | 320 | 323 | 323 | 325 | 328 | 332 |
| *Email* | 954 | 890 | 869 | 868 | 867 | 866 | 794 | 767 |
| *SocMedAdv* | 30 | 25 | 25 | 23 | 23 | 23 | 24 | 24 |
| *RSSFeed* | 126 | 122 | 120 | 117 | 116 | 112 | 114 | 115 |
| *OnlFolder* | 178 | 174 | 173 | 172 | 171 | 171 | 171 | 171 |
| *Other* | 1,493 | 1,473 | 1,469 | 1,461 | 1,441 | 1,432 | 1,424 | 1,415 |
| **Data Censoring** | | | | | | | | |
| *#VisCensor* | 1,036 | 1,033 | 1,054 | 1,054 | 1,063 | 1,035 | 1,007 | 1,010 |
| *#VisPurchCensor* | 870 | 835 | 785 | 751 | 749 | 686 | 628 | 553 |
| *%VisCensor* | 4.42% | 4.40% | 4.49% | 4.49% | 4.57% | 4.48% | 4.37% | 4.37% |
| *%VisPurchCensor* | 3.72% | 3.56% | 3.35% | 3.20% | 3.22% | 2.97% | 2.73% | 2.39% |
| *%NonCensor* | 91.86% | 92.04% | 92.16% | 92.31% | 92.21% | 92.55% | 92.90% | 93.24% |

## Table A8. Summary Statistics Censored Subsets 57-62

| VARIABLES | CENSORED SUBSET | | | | | |
|---|---|---|---|---|---|---|
| | 57 | 58 | 59 | 60 | 61 | 62 |
| *#Cust* | 23,137 | 23,063 | 22,978 | 22,722 | 22,519 | 22,401 |
| *#AllVis* | 60,789 | 60,527 | 60,254 | 59,482 | 58,873 | 58,487 |
| *#Purch* | 22,978 | 22,923 | 22,893 | 22,670 | 22,535 | 22,510 |
| *AvgVis* | 2.627 | 2.624 | 2.622 | 2.618 | 2.614 | 2.611 |
| *AvgPur* | 0.993 | 0.994 | 0.996 | 0.998 | 1.001 | 1.005 |
| *Conver* | 0.378 | 0.379 | 0.380 | 0.381 | 0.383 | 0.385 |
| **Traffic source of customers' visits** | | | | | | |
| *Direct* | 20,114 | 20,055 | 19,987 | 19,733 | 19,504 | 19,389 |
| *SEngin* | 10,034 | 9,995 | 9,915 | 9,773 | 9,659 | 9,589 |
| *SEnginAd* | 7 | 7 | 7 | 7 | 7 | 7 |
| *ComparSite* | 15,351 | 15,313 | 15,256 | 15,110 | 14,982 | 14,910 |
| *AffilMktg* | 1,934 | 1,929 | 1,924 | 1,903 | 1,881 | 1,864 |
| *Adwords* | 10,555 | 10,461 | 10,411 | 10,243 | 10,141 | 10,031 |
| *Display* | 335 | 335 | 332 | 332 | 333 | 334 |
| *Email* | 731 | 717 | 720 | 710 | 711 | 711 |
| *SocMedAdv* | 25 | 25 | 25 | 25 | 25 | 25 |
| *RSSFeed* | 118 | 115 | 116 | 117 | 118 | 118 |
| *OnlFolder* | 171 | 171 | 171 | 171 | 171 | 171 |
| *Other* | 1,414 | 1,404 | 1,390 | 1,358 | 1,341 | 1,338 |
| **Data Censoring** | | | | | | |
| *#VisCensor* | 1,036 | 1,033 | 1,054 | 1,054 | 1,063 | 1,035 |
| *#VisPurchCensor* | 870 | 835 | 785 | 751 | 749 | 686 |
| *%VisCensor* | 4.42% | 4.40% | 4.49% | 4.49% | 4.57% | 4.48% |
| *%VisPurchCensor* | 3.72% | 3.56% | 3.35% | 3.20% | 3.22% | 2.97% |
| *%NonCensor* | 91.86% | 92.04% | 92.16% | 92.31% | 92.21% | 92.55% |

## APPENDIX B. EMPIRICAL RESULTS: CENSORED SUBSET VS. SUBSETS WITH IMPUTED VALUES

### Table B1. Logit Model Results:

### Censored Subset 1 vs. Subsets 1 with Imputed Values

| | Censored Subset 1 | Subset 1 w. Imputed Values (90th pctl.) | Subset 1 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.262*** | 3.446*** | 4.416*** |
| *#Direct* | -0.035** | -0.082*** | -0.104*** |
| *#SEngVis* | -0.036 | -0.028 | 0.026 |
| *#AdsVis* | 0.093*** | 0.020 | 0.012 |
| *#PersonVis* | 0.460*** | 0.217*** | 0.107 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 1; 12,034 obs. Null dev.: 6,524; 12,033 d.f.; resid. dev.: 6,404; 12,029 d.f., AIC: 6,414. Subset 1 with Imputed Values (90th pctl.); 12,034 obs. Null dev.: 3,257; 12,033 d.f.; resid. dev.: 3,224; 12,029 d.f., AIC: 3,234. Subset 1 with Imputed Values (95th pctl.); 12,034 obs. Null dev.: 1,614; 12,033 d.f.; resid. dev.: 1,593; 12,029 d.f., AIC: 1,603. Signif. as above.

### Table B2. Logit Model Results:

### Censored Subset 2 vs. Subsets 2 with Imputed Values

| | Censored Subset 2 | Subset 2 w. Imputed Values (90th pctl.) | Subset 2 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.298*** | 3.465*** | 4.403*** |
| *#Direct* | -0.033** | -0.074*** | -0.096*** |
| *#SEngVis* | -0.039 | -0.014 | 0.023 |
| *#AdsVis* | 0.101*** | 0.003 | -0.015 |
| *#PersonVis* | 0.465*** | 0.184*** | 0.093 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 2; 12,310 obs. Null dev.: 6,492; 12,309 d.f.; resid. dev.: 6,372; 12,305 d.f., AIC: 6,382. Subset 2 with Imputed Values (90th pctl.); 12,310 obs. Null dev.: 3,308; 12,309 d.f.; resid. dev.: 3,283; 12,305 d.f., AIC: 3,293. Subset 2 with Imputed Values (95th pctl.); 12,310 obs. Null dev.: 1,690; 12,309 d.f.; resid. dev.: 1,672; 12,305 d.f., AIC: 1,682. Signif. as above.

### Table B3. Logit Model Results:

### Censored Subset 3 vs. Subsets 3 with Imputed Values

| | Censored Subset 3 | Subset 3 w. Imputed Values (90th pctl.) | Subset 3 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.317*** | 3.459*** | 4.4580*** |
| *#Direct* | -0.033** | -0.068*** | -0.069*** |
| *#SEngVis* | -0.038 | -0.044 | -0.003 |
| *#AdsVis* | 0.079** | 0.022 | -0.037 |
| *#PersonVis* | 0.426*** | 0.184*** | -0.003 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 3; 12,625 obs. Null dev.: 6,687; 12,624 d.f.; resid. dev.: 6,579; 12,620 d.f., AIC: 6,589. Subset 3 with Imputed Values (90th pctl.); 12,625 obs. Null dev.: 3,397; 12,624 d.f.; resid. dev.: 3,372; 12,620 d.f., AIC: 3,382. Subset 3 with Imputed Values (95th pctl.); 12,625 obs. Null dev.: 1,733; 12,624 d.f.; resid. dev.: 1721; 12,620 d.f., AIC: 1,731. Signif. as above.

### Table B4. Logit Model Results:

### Censored Subset 4 vs. Subsets 4 with Imputed Values

| | Censored Subset 4 | Subset 4 w. Imputed Values (90th pctl.) | Subset 4 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.327*** | 3.474*** | 4.497*** |
| *#Direct* | -0.031** | -0.074*** | -0.085*** |
| *#SEngVis* | -0.042 | -0.063* | 0.011 |
| *#AdsVis* | 0.078 | 0.016 | -0.061 |
| *#PersonVis* | 0.441*** | 0.214*** | 0.071 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 4; 12,946 obs. Null dev.: 6,791; 12,945 d.f.; resid. dev.: 6,679; 12,941 d.f., AIC: 6,689. Subset 4 with Imputed Values (90th pctl.); 12,946 obs. Null dev.: 3,465; 12,945 d.f.; resid. dev.: 3,432; 12,941 d.f., AIC: 3,442. Subset 4 with Imputed Values (95th pctl.); 12,946 obs. Null dev.: 1,706; 12,945 d.f.; resid. dev.: 1,688; 12,941 d.f., AIC: 1,698. Signif. as above.

**Table B5. Logit Model Results:**

**Censored Subset 5 vs. Subsets 5 with Imputed Values**

| | Censored Subset 5 | Subset 5 w. Imputed Values (90th pctl.) | Subset 5 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.342*** | 3.472*** | 4.515*** |
| *#Direct* | -0.031** | -0.066*** | -0.073*** |
| *#SEngVis* | -0.060** | -0.091** | -0.029 |
| *#AdsVis* | 0.102*** | 0.039 | -0.052 |
| *#PersonVis* | 0.430*** | 0.190*** | 0.094 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 5; 13,208 obs. Null dev.: 6,857; 13,207 d.f.; resid. dev.: 6,742; 13,203 d.f., AIC: 6,752. Subset 5 with Imputed Values (90th pctl.); 13,208 obs. Null dev.: 3,544; 13,207 d.f.; resid. dev.: 3,513; 13,203 d.f., AIC: 3,523. Subset 5 with Imputed Values (95th pctl.); 13,208 obs. Null dev.: 1,695; 13,207 d.f.; resid. dev.: 1,679; 13,203 d.f., AIC: 1,689. Signif. as above.

**Table B6. Logit Model Results:**

**Censored Subset 6 vs. Subsets 6 with Imputed Values**

| | Censored Subset 6 | Subset 6 w. Imputed Values (90th pctl.) | Subset 6 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.394*** | 3.546*** | 4.606*** |
| *#Direct* | -0.042*** | -0.072*** | -0.080*** |
| *#SEngVis* | -0.044 | -0.041 | 0.092 |
| *#AdsVis* | 0.065** | -0.018 | -0.123*** |
| *#PersonVis* | 0.399*** | 0.180*** | 0.057 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 6; 13,577 obs. Null dev.: 6,973; 13,576 d.f.; resid. dev.: 6,869; 13,572 d.f., AIC: 6,879. Subset 6 with Imputed Values (90th pctl.); 13,577 obs. Null dev.: 3,509; 13,576 d.f.; resid. dev.: 3,480; 13,572 d.f., AIC: 3,490. Subset 6 with Imputed Values (95th pctl.); 13,577 obs. Null dev.: 1,659; 13,576 d.f.; resid. dev.: 1,637; 13,572 d.f., AIC: 1,647. Signif. as above.

**Table B7. Logit Model Results:**

**Censored Subset 7 vs. Subsets 7 with Imputed Values**

| | Censored Subset 7 | Subset 7 w. Imputed Values (90th pctl.) | Subset 7 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.426*** | 3.563*** | 4.651*** |
| *#Direct* | -0.054*** | -0.078*** | -0.079*** |
| *#SEngVis* | -0.045 | -0.016 | 0.057 |
| *#AdsVis* | 0.062** | -0.048 | -0.132*** |
| *#PersonVis* | 0.455*** | 0.238*** | 0.103 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 7; 13,964 obs. Null dev.: 7,003; 13,963 d.f.; resid. dev.: 6,873; 13,959 d.f., AIC: 6,883. Subset 7 with Imputed Values (90th pctl.); 13,964 obs. Null dev.: 3,539; 13,963 d.f.; resid. dev.: 3,499; 13,959 d.f., AIC: 3,509. Subset 7 with Imputed Values (95th pctl.); 13,964 obs. Null dev.: 1,649; 13,963 d.f.; resid. dev.: 1,623; 13,959 d.f., AIC: 1,633. Signif. as above.

**Table B8. Logit Model Results:**

**Censored Subset 8 vs. Subsets 8 with Imputed Values**

| | Censored Subset 8 | Subset 8 w. Imputed Values (90th pctl.) | Subset 8 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.438*** | 3.550*** | 4.682*** |
| *#Direct* | -0.041*** | -0.068*** | -0.067*** |
| *#SEngVis* | -0.045 | -0.056 | 0.015 |
| *#AdsVis* | 0.042 | -0.031 | -0.114*** |
| *#PersonVis* | 0.474*** | 0.261*** | 0.129 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 8; 14,369 obs. Null dev.: 7,108; 14,368 d.f.; resid. dev.: 6,978; 14,364 d.f., AIC: 6,988. Subset 8 with Imputed Values (90th pctl.); 14,369 obs. Null dev.: 3,640; 14,368 d.f.; resid. dev.: 3,599; 14,364 d.f., AIC: 3,609. Subset 8 with Imputed Values (95th pctl.); 14,369 obs. Null dev.: 1,621; 14,368 d.f.; resid. dev.: 1,600; 14,364 d.f., AIC: 1,610. Signif. as above.

**Table B9. Logit Model Results:**

**Censored Subset 9 vs. Subsets 9 with Imputed Values**

| | Censored Subset 9 | Subset 9 w. Imputed Values (90th pctl.) | Subset 9 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.490*** | 3.544*** | 4.730*** |
| *#Direct* | -0.039*** | -0.062*** | -0.067*** |
| *#SEngVis* | -0.052** | -0.064* | 0.003 |
| *#AdsVis* | 0.038 | -0.025 | -0.099** |
| *#PersonVis* | 0.382*** | 0.249*** | 0.119 |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 9; 14,720 obs. Null dev.: 7,198; 14,719 d.f.; resid. dev.: 7,101; 14,715 d.f., AIC: 7,111. Subset 9 with Imputed Values (90th pctl.); 14,720 obs. Null dev.: 3,738; 14,719 d.f.; resid. dev.: 3,700; 14,715 d.f., AIC: 3,710. Subset 9 with Imputed Values (95th pctl.); 14,720 obs. Null dev.: 1,591; 14,719 d.f.; resid. dev.: 1,574; 14,715 d.f., AIC: 1,584. Signif. as above.

**Table B10. Logit Model Results:**

**Censored Subset 10 vs. Subsets 10 with Imputed Values**

| | Censored Subset 10 | Subset 10 w. Imputed Values (90th pctl.) | Subset 10 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.460*** | 3.543*** | 4.727*** |
| *#Direct* | -0.035** | -0.069*** | -0.070*** |
| *#SEngVis* | -0.075*** | -0.087** | -0.048 |
| *#AdsVis* | 0.082*** | -0.011 | -0.076* |
| *#PersonVis* | 0.448*** | 0.320*** | 0.251** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 10; 15,027 obs. Null dev.: 7,277; 15,026 d.f.; resid. dev.: 7.151; 15,022 d.f., AIC: 7,161. Subset 10 with Imputed Values (90th pctl.); 15,027 obs. Null dev.: 3,762; 15,026 d.f.; resid. dev.: 3,709; 15,022 d.f., AIC: 3,719. Subset 10 with Imputed Values (95th pctl.); 15,027 obs. Null dev.: 1,560; 15,026 d.f.; resid. dev.: 1,538; 15,022 d.f., AIC: 1,548. Signif. as above.

**Table B11. Logit Model Results:**

**Censored Subset 11 vs. Subsets 11 with Imputed Values**

| | Censored Subset 11 | Subset 11 w. Imputed Values (90th pctl.) | Subset 11 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.429*** | 3.556*** | 4.561*** |
| *#Direct* | -0.034** | -0.079*** | -0.081*** |
| *#SEngVis* | -0.075*** | -0.095*** | -0.071 |
| *#AdsVis* | 0.099*** | 0.012 | -0.049 |
| *#PersonVis* | 0.464*** | 0.328*** | 0.212** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 11; 15,333 obs. Null dev.: 7,499; 15,332 d.f.; resid. dev.: 7,360; 15,328 d.f., AIC: 7,370. Subset 11 with Imputed Values (90th pctl.); 15,333 obs. Null dev.: 3,793; 15,322 d.f.; resid. dev.: 3,734; 15,328 d.f., AIC: 3,744. Subset 11 with Imputed Values (95th pctl.); 15,333 obs. Null dev.: 1,860; 15,322 d.f.; resid. dev.: 1,831; 15,328 d.f., AIC: 1,841. Signif. as above.

**Table B12. Logit Model Results:**

**Censored Subset 12 vs. Subsets 12 with Imputed Values**

| | Censored Subset 12 | Subset 12 w. Imputed Values (90th pctl.) | Subset 12 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.451*** | 3.562*** | 4.528*** |
| *#Direct* | -0.039*** | -0.093*** | -0.084*** |
| *#SEngVis* | -0.088*** | -0.091*** | -0.079* |
| *#AdsVis* | 0.092*** | 0.009 | -0.027 |
| *#PersonVis* | 0.436*** | 0.301*** | 0.153** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 12; 15,680 obs. Null dev.: 7,685; 15,679 d.f.; resid. dev.: 7,550; 15,675 d.f., AIC: 7,560. Subset 12 with Imputed Values (90th pctl.); 15,680 obs. Null dev.: 3,948; 15,679 d.f.; resid. dev.: 3,880; 15,675 d.f., AIC: 3,890. Subset 12 with Imputed Values (95th pctl.); 15,680 obs. Null dev.: 1,984; 15,679 d.f.; resid. dev.: 1,956; 15,675 d.f., AIC: 1,966. Signif. as above.

**Table B13. Logit Model Results:**

**Censored Subset 13 vs. Subsets 13 with Imputed Values**

| | Censored Subset 13 | Subset 13 w. Imputed Values (90th pctl.) | Subset 13 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.494*** | 3.572*** | 4.614*** |
| *#Direct* | -0.043*** | -0.094*** | -0.094*** |
| *#SEngVis* | -0.086*** | -0.083** | -0.059 |
| *#AdsVis* | 0.088*** | 0.007 | -0.041 |
| *#PersonVis* | 0.400*** | 0.300*** | 0.141* |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 13; 16,147 obs. Null dev.: 7,785; 16,146 d.f.; resid. dev.: 7,663; 16,142 d.f., AIC: 7,673. Subset 13 with Imputed Values (90th pctl.); 16,147 obs. Null dev.: 4,031; 16,146 d.f.; resid. dev.: 3,964; 16,142 d.f., AIC: 3,974. Subset 13 with Imputed Values (95th pctl.); 16,147 obs. Null dev.: 1,941; 16,146 d.f.; resid. dev.: 1,909; 16,142 d.f., AIC: 1,919. Signif. as above.

**Table B14. Logit Model Results:**

**Censored Subset 14 vs. Subsets 14 with Imputed Values**

| | Censored Subset 14 | Subset 14 w. Imputed Values (90th pctl.) | Subset 14 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.504*** | 3.485*** | 4.317*** |
| *#Direct* | -0.039*** | -0.086*** | -0.079*** |
| *#SEngVis* | -0.077*** | -0.097*** | -0.105*** |
| *#AdsVis* | 0.065** | 0.000 | -0.018 |
| *#PersonVis* | 0.412*** | 0.317*** | 0.220*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 14; 16,589 obs. Null dev.: 7,946; 16,588 d.f.; resid. dev.: 7,824; 16,584 d.f., AIC: 7,834. Subset 14 with Imputed Values (90th pctl.); 16,589 obs. Null dev.: 4,403; 16,588 d.f.; resid. dev.: 4,331; 16,584 d.f., AIC: 4,341. Subset 14 with Imputed Values (95th pctl.); 16,589 obs. Null dev.: 2,416; 16,588 d.f.; resid. dev.: 2,383; 16,584 d.f., AIC: 2,393. Signif. as above.

**Table B15. Logit Model Results:**

**Censored Subset 15 vs. Subsets 15 with Imputed Values**

| | Censored Subset 15 | Subset 15 w. Imputed Values (90th pctl.) | Subset 15 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.541*** | 3.541*** | 4.287*** |
| *#Direct* | -0.047*** | -0.102*** | -0.087*** |
| *#SEngVis* | -0.104*** | -0.122*** | -0.118*** |
| *#AdsVis* | 0.111*** | 0.044 | 0.017 |
| *#PersonVis* | 0.370*** | 0.323*** | 0.288*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 15; 16,972 obs. Null dev.: 7,991; 16,971 d.f.; resid. dev.: 7,870; 16,967 d.f., AIC: 7,880. Subset 15 with Imputed Values (90th pctl.); 16,972 obs. Null dev.: 4,327; 16,971 d.f.; resid. dev.: 4,239; 16,967 d.f., AIC: 4,249. Subset 15 with Imputed Values (95th pctl.); 16,972 obs. Null dev.: 2,453; 16,971 d.f.; resid. dev.: 2,414; 16,967 d.f., AIC: 2,424. Signif. as above.

**Table B16. Logit Model Results:**

**Censored Subset 16 vs. Subsets 16 with Imputed Values**

| | Censored Subset 16 | Subset 16 w. Imputed Values (90th pctl.) | Subset 16 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.586*** | 3.518*** | 4.264*** |
| *#Direct* | -0.049*** | -0.101*** | -0.090*** |
| *#SEngVis* | -0.090*** | -0.123*** | -0.140*** |
| *#AdsVis* | 0.092*** | 0.060* | 0.052 |
| *#PersonVis* | 0.324*** | 0.275*** | 0.234*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 16; 17,373 obs. Null dev.: 8,065; 17,372 d.f.; resid. dev.: 7,964; 17,368 d.f., AIC: 7,974.
Subset 16 with Imputed Values (90th pctl.); 17,273 obs. Null dev.: 4,534; 17,372 d.f.; resid. dev.: 4,451; 17,368 d.f., AIC: 4,461. Subset 16 with Imputed Values (95th pctl.); 17,273 obs. Null dev.: 2,592; 17,372 d.f.; resid. dev.: 2,550; 17,368 d.f., AIC: 2,560. Signif. as above.

**Table B17. Logit Model Results:**

**Censored Subset 17 vs. Subsets 17 with Imputed Values**

| | Censored Subset 17 | Subset 17 w. Imputed Values (90th pctl.) | Subset 17 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.635*** | 3.494*** | 4.312*** |
| *#Direct* | -0.058*** | -0.107*** | -0.113*** |
| *#SEngVis* | -0.092*** | -0.115*** | -0.137*** |
| *#AdsVis* | 0.082*** | 0.056* | 0.038 |
| *#PersonVis* | 0.296*** | 0.307*** | 0.304*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 17; 17,696 obs. Null dev.: 8,100; 17,695 d.f.; resid. dev.: 8,004; 17,691 d.f., AIC: 8,014. Subset 17 with Imputed Values (90th pctl.); 17,696 obs. Null dev.: 4,672; 17,695 d.f.; resid. dev.: 4,577; 17,691 d.f., AIC: 4,587. Subset 17 with Imputed Values (95th pctl.); 17,696 obs. Null dev.: 2,567; 17,695 d.f.; resid. dev.: 2,503; 17,691 d.f., AIC: 2,513. Signif. as above.

**Table B18. Logit Model Results:**

**Censored Subset 18 vs. Subsets 18 with Imputed Values**

| | Censored Subset 18 | Subset 18 w. Imputed Values (90th pctl.) | Subset 18 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.606*** | 3.477*** | 4.304*** |
| *#Direct* | -0.058*** | -0.103*** | -0.126*** |
| *#SEngVis* | -0.099*** | -0.141*** | -0.14*** |
| *#AdsVis* | 0.098*** | 0.089** | 0.083* |
| *#PersonVis* | 0.294*** | 0.284*** | 0.245*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 18; 18,024 obs. Null dev.: 8,382; 18,023 d.f.; resid. dev.: 8,279; 18,019 d.f., AIC: 8,289. Subset 18 with Imputed Values (90th pctl.); 18,024 obs. Null dev.: 4,810; 18,023 d.f.; resid. dev.: 4,715; 18,019 d.f., AIC: 4,725. Subset 18 with Imputed Values (95th pctl.); 18,024 obs. Null dev.: 2670; 18,023 d.f.; resid. dev.: 2,598; 18,019 d.f., AIC: 2,608. Signif. as above.

**Table B19. Logit Model Results:**

**Censored Subset 19 vs. Subsets 19 with Imputed Values**

| | Censored Subset 19 | Subset 19 w. Imputed Values (90th pctl.) | Subset 19 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.615*** | 3.569*** | 4.303*** |
| *#Direct* | -0.062*** | -0.109*** | -0.129*** |
| *#SEngVis* | -0.091*** | -0.130*** | -0.139*** |
| *#AdsVis* | 0.091*** | 0.085** | 0.096** |
| *#PersonVis* | 0.292*** | 0.151*** | 0.201*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 19; 18,437 obs. Null dev.: 8,543; 18,436 d.f.; resid. dev.: 8,440; 18,432 d.f., AIC: 8,450. Subset 19 with Imputed Values (90th pctl.); 18,437 obs. Null dev.: 4,835; 18,436 d.f.; resid. dev.: 4,752; 18,432 d.f., AIC: 4,762. Subset 19 with Imputed Values (95th pctl.); 18,437 obs. Null dev.: 2,766; 18,436 d.f.; resid. dev.: 2,694; 18,432 d.f., AIC: 2,704. Signif. as above.

**Table B20. Logit Model Results:**

**Censored Subset 20 vs. Subsets 20 with Imputed Values**

| | Censored Subset 20 | Subset 20 w. Imputed Values (90th pctl.) | Subset 20 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.604*** | 3.592*** | 4.300*** |
| *#Direct* | -0.061*** | -0.107*** | -0.132*** |
| *#SEngVis* | -0.094*** | -0.111*** | -0.129*** |
| *#AdsVis* | 0.099*** | 0.058* | 0.076 |
| *#PersonVis* | 0.328*** | 0.182*** | 0.297*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 20; 18,949 obs. Null dev.: 8,728; 18,948 d.f.; resid. dev.: 8,608; 18,944 d.f., AIC: 8,618. Subset 20 with Imputed Values (90th pctl.); 18,949 obs. Null dev.: 4,836; 18,948 d.f.; resid. dev.: 4,757; 18,944 d.f., AIC: 4,767. Subset 20 with Imputed Values (95th pctl.); 18,949 obs. Null dev.: 2,772; 18,948 d.f.; resid. dev.: 2,692; 18,944 d.f., AIC: 2,702. Signif. as above.

**Table B21. Logit Model Results:**

**Censored Subset 21 vs. Subsets 21 with Imputed Values**

| | Censored Subset 21 | Subset 21 w. Imputed Values (90th pctl.) | Subset 21 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.612*** | 3.559*** | 4.332*** |
| *#Direct* | -0.057*** | -0.100*** | -0.123*** |
| *#SEngVis* | -0.094*** | -0.125*** | -0.105*** |
| *#AdsVis* | 0.085*** | 0.038 | 0.029 |
| *#PersonVis* | 0.398*** | 0.282*** | 0.222*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 21; 19,361 obs. Null dev.: 8,719; 19,360 d.f.; resid. dev.: 8,577; 19,356 d.f., AIC: 8,587. Subset 21 with Imputed Values (90th pctl.); 19,361 obs. Null dev.: 4,923; 19,360 d.f.; resid. dev.: 4,834; 19,356 d.f., AIC: 4,844. Subset 21 with Imputed Values (95th pctl.); 19,361 obs. Null dev.: 2,843; 19,360 d.f.; resid. dev.: 2,776; 19,356 d.f., AIC: 2,786. Signif. as above.

**Table B22. Logit Model Results:**

**Censored Subset 22 vs. Subsets 22 with Imputed Values**

| | Censored Subset 22 | Subset 22 w. Imputed Values (90th pctl.) | Subset 22 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.656*** | 3.464*** | 4.231*** |
| *#Direct* | -0.054*** | -0.081*** | -0.099*** |
| *#SEngVis* | -0.110*** | -0.126*** | -0.123*** |
| *#AdsVis* | 0.099*** | 0.049 | 0.024 |
| *#PersonVis* | 0.358*** | 0.227*** | 0.166*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 22; 19,752 obs. Null dev.: 8,710; 19,751 d.f.; resid. dev.: 8,583; 19,747 d.f., AIC: 8,593. Subset 22 with Imputed Values (90th pctl.); 19,752 obs. Null dev.: 5,368; 19,751 d.f.; resid. dev.: 5,298; 19,747 d.f., AIC: 5,308. Subset 22 with Imputed Values (95th pctl.); 19,752 obs. Null dev.: 3,158; 19,751 d.f.; resid. dev.: 3,103; 19,747 d.f., AIC: 3,113. Signif. as above.

**Table B23. Logit Model Results:**

**Censored Subset 23 vs. Subsets 23 with Imputed Values**

| | Censored Subset 23 | Subset 23 w. Imputed Values (90th pctl.) | Subset 23 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.683*** | 3.462*** | 4.134*** |
| *#Direct* | -0.058*** | -0.083*** | -0.094*** |
| *#SEngVis* | -0.097*** | -0.105*** | -0.101*** |
| *#AdsVis* | 0.082*** | 0.032 | 0.006 |
| *#PersonVis* | 0.353*** | 0.236*** | 0.229*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 23; 20,146 obs. Null dev.: 8,774; 20,145 d.f.; resid. dev.: 8,650; 20,141 d.f., AIC: 8,660. Subset 23 with Imputed Values (90th pctl.); 20,146 obs. Null dev.: 5,468; 20,145 d.f.; resid. dev.: 5,399; 20,141 d.f., AIC: 5,409. Subset 23 with Imputed Values (95th pctl.); 20,146 obs. Null dev.: 3,360; 20,145 d.f.; resid. dev.: 3,308; 20,141 d.f., AIC: 3,318. Signif. as above.

**Table B24. Logit Model Results:**

**Censored Subset 24 vs. Subsets 24 with Imputed Values**

| | Censored Subset 24 | Subset 24 w. Imputed Values (90th pctl.) | Subset 24 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.692*** | 3.451*** | 4.090*** |
| *#Direct* | -0.063*** | -0.083*** | -0.094*** |
| *#SEngVis* | -0.087*** | -0.092*** | -0.094** |
| *#AdsVis* | 0.065** | 0.031 | 0.012 |
| *#PersonVis* | 0.385*** | 0.236*** | 0.235*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 24; 20,458 obs. Null dev.: 8,838; 20,457 d.f.; resid. dev.: 8,700; 20,453 d.f., AIC: 8,710. Subset 24 with Imputed Values (90th pctl.); 20,458 obs. Null dev.: 5,563; 20,457 d.f.; resid. dev.: 5,498; 20,453 d.f., AIC: 5,508. Subset 24 with Imputed Values (95th pctl.); 20,458 obs. Null dev.: 3,493; 20,457 d.f.; resid. dev.: 3,442; 20,453 d.f., AIC: 3,452. Signif. as above.

**Table B25. Logit Model Results:**

**Censored Subset 25 vs. Subsets 25 with Imputed Values**

| | Censored Subset 25 | Subset 25 w. Imputed Values (90th pctl.) | Subset 25 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.672*** | 3.460*** | 4.125*** |
| *#Direct* | -0.065*** | -0.091*** | -0.100*** |
| *#SEngVis* | -0.101*** | -0.116*** | -0.117*** |
| *#AdsVis* | 0.084*** | 0.050 | 0.024 |
| *#PersonVis* | 0.440*** | 0.292*** | 0.250*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 25; 20,794 obs. Null dev.: 8,961; 20,793 d.f.; resid. dev.: 8,791; 20,789 d.f., AIC: 8,801. Subset 25 with Imputed Values (90th pctl.); 20,794 obs. Null dev.: 5,563; 20,793 d.f.; resid. dev.: 5,473; 20,789 d.f., AIC: 5,483. Subset 25 with Imputed Values (95th pctl.); 20,794 obs. Null dev.: 3,480; 20,793 d.f.; resid. dev.: 3,416; 20,789 d.f., AIC: 3,426. Signif. as above.

**Table B26. Logit Model Results:**

**Censored Subset 26 vs. Subsets 26 with Imputed Values**

| | Censored Subset 26 | Subset 26 w. Imputed Values (90th pctl.) | Subset 26 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.676*** | 3.475*** | 4.204*** |
| *#Direct* | -0.059*** | -0.088*** | -0.098*** |
| *#SEngVis* | -0.106*** | -0.113*** | -0.120*** |
| *#AdsVis* | 0.091*** | 0.048 | 0.012 |
| *#PersonVis* | 0.427*** | 0.262*** | 0.258*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 26; 20,202 obs. Null dev.: 9,093; 20,201 d.f.; resid. dev.: 8,930; 20,197 d.f., AIC: 8,940. Subset 26 with Imputed Values (90th pctl.); 20,202 obs. Null dev.: 5,644; 20,201 d.f.; resid. dev.: 5,562; 20,197 d.f., AIC: 5,572. Subset 26 with Imputed Values (95th pctl.); 20,202 obs. Null dev.: 3,344; 20,201 d.f.; resid. dev.: 3,282; 20,197 d.f., AIC: 3,292. Signif. as above.

**Table B27. Logit Model Results:**

**Censored Subset 27 vs. Subsets 27 with Imputed Values**

| | Censored Subset 27 | Subset 27 w. Imputed Values (90th pctl.) | Subset 27 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.676*** | 3.434*** | 4.193*** |
| *#Direct* | -0.061*** | -0.087*** | -0.093*** |
| *#SEngVis* | -0.110*** | -0.136*** | -0.141*** |
| *#AdsVis* | 0.110*** | 0.053* | 0.034 |
| *#PersonVis* | 0.509*** | 0.347*** | 0.332*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 27; 21,717 obs. Null dev.: 9,088; 21,716 d.f.; resid. dev.: 8,885; 21,712 d.f., AIC: 8,895. Subset 27 with Imputed Values (90th pctl.); 21,717 obs. Null dev.: 5,835; 21,716 d.f.; resid. dev.: 5,729; 21,712 d.f., AIC: 5,739. Subset 27 with Imputed Values (95th pctl.); 21,717 obs. Null dev.: 3,334; 21,716 d.f.; resid. dev.: 3,270; 21,712 d.f., AIC: 3,280. Signif. as above.

**Table B28. Logit Model Results:**

**Censored Subset 28 vs. Subsets 28 with Imputed Values**

| | Censored Subset 28 | Subset 28 w. Imputed Values (90th pctl.) | Subset 28 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.678*** | 3.413*** | 4.120*** |
| *#Direct* | -0.063*** | -0.090*** | -0.096*** |
| *#SEngVis* | -0.105*** | -0.131*** | -0.152*** |
| *#AdsVis* | 0.117*** | 0.067** | 0.050 |
| *#PersonVis* | 0.488*** | 0.357*** | 0.360*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 28; 22,175 obs. Null dev.: 9,276; 22,174 d.f.; resid. dev.: 9,077; 22,170 d.f., AIC: 9,087. Subset 28 with Imputed Values (90th pctl.); 22,175 obs. Null dev.: 5,988; 22,174 d.f.; resid. dev.: 5,877; 22,170 d.f., AIC: 5,887. Subset 28 with Imputed Values (95th pctl.); 22,175 obs. Null dev.: 3,574; 22,174 d.f.; resid. dev.: 3,498; 22,170 d.f., AIC: 3,508. Signif. as above.

## Table B29. Logit Model Results:

## Censored Subset 29 vs. Subsets 29 with Imputed Values

| | Censored Subset 29 | Subset 29 w. Imputed Values (90TH pctl.) | Subset 29 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 2.678*** | 3.413*** | 4.120*** |
| *#Direct* | -0.063*** | -0.090*** | -0.096*** |
| *#SEngVis* | -0.105*** | -0.131*** | -0.152*** |
| *#AdsVis* | 0.117*** | 0.067** | 0.050 |
| *#PersonVis* | 0.488*** | 0.357*** | 0.360*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 29; 22,644 obs. Null dev.: 9,276; 22,643 d.f.; resid. dev.: 9,077; 22,639 d.f., AIC: 9,087. Subset 29 with Imputed Values (90th pctl.); 22,644 obs. Null dev.: 5,988; 22,643 d.f.; resid. dev.: 5,877; 22,639 d.f., AIC: 5,887. Subset 29 with Imputed Values (95th pctl.); 22,644 obs. Null dev.: 3,574; 22,643 d.f.; resid. dev.: 3,498; 22,639 d.f., AIC: 3,508. Signif. as above.

## Table B30. Logit Model Results:

## Censored Subset 30 vs. Subsets 30 with Imputed Values

| | Censored Subset 30 | Subset 30 w. Imputed Values (90TH pctl.) | Subset 30 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 2.642*** | 3.371*** | 4.005*** |
| *#Direct* | -0.056*** | -0.091*** | -0.082*** |
| *#SEngVis* | -0.104*** | -0.131*** | -0.149*** |
| *#AdsVis* | 0.183*** | 0.133*** | 0.172*** |
| *#PersonVis* | 0.571*** | 0.455*** | 0.371*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 30; 23,053 obs. Null dev.: 9,406; 23,052 d.f.; resid. dev.: 9,161; 22,048 d.f., AIC: 9,171. Subset 30 with Imputed Values (90th pctl.); 23,053 obs. Null dev.: 6,069; 23,052 d.f.; resid. dev.: 5,931; 22,048 d.f., AIC: 5,941. Subset 30 with Imputed Values (95th pctl.); 23,053 obs. Null dev.: 3,759; 23,052 d.f.; resid. dev.: 3,692; 22,048 d.f., AIC: 3,702. Signif. as above.

## Table B31. Logit Model Results:

## Censored Subset 31 vs. Subsets 31 with Imputed Values

| | Censored Subset 31 | Subset 31 w. Imputed Values (90TH pctl.) | Subset 31 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 2.634*** | 3.328*** | 3.958*** |
| *#Direct* | -0.058*** | -0.092*** | -0.095*** |
| *#SEngVis* | -0.099*** | -0.118*** | -0.141*** |
| *#AdsVis* | 0.191*** | 0.139*** | 0.196*** |
| *#PersonVis* | 0.547*** | 0.462*** | 0.415*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 31; 23,415 obs. Null dev.: 9,636; 23,414 d.f.; resid. dev.: 9,392; 23,410 d.f., AIC: 9,402. Subset 31 with Imputed Values (90th pctl.); 23,415 obs. Null dev.: 6,307; 23,414 d.f.; resid. dev.: 6,162; 23,410 d.f., AIC: 6,172. Subset 31 with Imputed Values (95th pctl.); 23,415 obs. Null dev.: 3,910; 23,414 d.f.; resid. dev.: 3,825; 23,410 d.f., AIC: 3,835. Signif. as above.

## Table B32. Logit Model Results:

## Censored Subset 32 vs. Subsets 32 with Imputed Values

| | Censored Subset 32 | Subset 32 w. Imputed Values (90TH pctl.) | Subset 32 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 2.637*** | 3.349*** | 4.001*** |
| *#Direct* | -0.050*** | -0.087*** | -0.081*** |
| *#SEngVis* | -0.117*** | -0.117*** | -0.142*** |
| *#AdsVis* | 0.180*** | 0.117*** | 0.147*** |
| *#PersonVis* | 0.526*** | 0.427*** | 0.342*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 32; 23,356 obs. Null dev.: 9,671; 23,355 d.f.; resid. dev.: 9,438; 23,351 d.f., AIC: 9,448. Subset 32 with Imputed Values (90th pctl.); 23,356 obs. Null dev.: 6,275; 23,355 d.f.; resid. dev.: 6,148; 23,351 d.f., AIC: 6,158. Subset 32 with Imputed Values (95th pctl.); 23,356 obs. Null dev.: 3,876; 23,355 d.f.; resid. dev.: 3,813; 23,351 d.f., AIC: 3,823. Signif. as above.

**Table B33. Logit Model Results:**

**Censored Subset 33 vs. Subsets 33 with Imputed Values**

| | Censored Subset 33 | Subset 33 w. Imputed Values (90th pctl.) | Subset 33 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.655*** | 3.388*** | 4.064*** |
| *#Direct* | -0.047*** | -0.082*** | -0.081*** |
| *#SEngVis* | -0.122*** | -0.135*** | -0.157*** |
| *#AdsVis* | 0.187*** | 0.118*** | 0.149*** |
| *#PersonVis* | 0.510*** | 0.402*** | 0.330*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 33; 23,380 obs. Null dev.: 9,575; 23,379 d.f.; resid. dev.: 9,352; 23,375 d.f., AIC: 9,362. Subset 33 with Imputed Values (90th pctl.); 23,380 obs. Null dev.: 6,158; 23,379 d.f.; resid. dev.: 6,040; 23,375 d.f., AIC: 6,050. Subset 33 with Imputed Values (95th pctl.); 23,380 obs. Null dev.: 3,728; 23,379 d.f.; resid. dev.: 3,667; 23,375 d.f., AIC: 3,677. Signif. as above.

**Table B34. Logit Model Results:**

**Censored Subset 34 vs. Subsets 34 with Imputed Values**

| | Censored Subset 34 | Subset 34 w. Imputed Values (90th pctl.) | Subset 34 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.650*** | 3.388*** | 4.004*** |
| *#Direct* | -0.032** | -0.072*** | -0.060*** |
| *#SEngVis* | -0.120*** | -0.125*** | -0.150*** |
| *#AdsVis* | 0.173*** | 0.101*** | 0.162*** |
| *#PersonVis* | 0.504*** | 0.467*** | 0.478*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 34; 23,446 obs. Null dev.: 9,588; 23,445 d.f.; resid. dev.: 9,381; 23,441 d.f., AIC: 9,391. Subset 34 with Imputed Values (90th pctl.); 23,446 obs. Null dev.: 6,028; 23,445 d.f.; resid. dev.: 5,910; 23,441 d.f., AIC: 5,920. Subset 34 with Imputed Values (95th pctl.); 23,446 obs. Null dev.: 3,638; 23,445 d.f.; resid. dev.: 3,571; 23,441 d.f., AIC: 3,581. Signif. as above.

**Table B35. Logit Model Results:**

**Censored Subset 35 vs. Subsets 35 with Imputed Values**

| | Censored Subset 35 | Subset 35 w. Imputed Values (90th pctl.) | Subset 35 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.666*** | 3.413*** | 4.004*** |
| *#Direct* | -0.037*** | -0.075*** | -0.069*** |
| *#SEngVis* | -0.129*** | -0.144*** | -0.147*** |
| *#AdsVis* | 0.182*** | 0.120*** | 0.151*** |
| *#PersonVis* | 0.494*** | 0.351*** | 0.379*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 35; 23,483 obs. Null dev.: 9,557; 23,482 d.f.; resid. dev.: 9,348; 23,478 d.f., AIC: 9,358. Subset 35 with Imputed Values (90th pctl.); 23,483 obs. Null dev.: 6,143; 23,482 d.f.; resid. dev.: 6,042; 23,478 d.f., AIC: 6.052. Subset 35 with Imputed Values (95th pctl.); 23,483 obs. Null dev.: 3,797; 23,482 d.f.; resid. dev.: 3,737; 23,478 d.f., AIC: 3,747. Signif. as above.

**Table B36. Logit Model Results:**

**Censored Subset 36 vs. Subsets 36 with Imputed Values**

| | Censored Subset 36 | Subset 36 w. Imputed Values (90th pctl.) | Subset 36 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.675*** | 3.401*** | 3.938*** |
| *#Direct* | -0.035** | -0.071*** | -0.055*** |
| *#SEngVis* | -0.125*** | -0.127*** | -0.128*** |
| *#AdsVis* | 0.155*** | 0.101*** | 0.143*** |
| *#PersonVis* | 0.492*** | 0.328*** | 0.440*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 36; 23,625 obs. Null dev.: 9,613; 23,624 d.f.; resid. dev.: 9,413; 23,620 d.f., AIC: 9,423. Subset 36 with Imputed Values (90th pctl.); 23,625 obs. Null dev.: 6,264; 23,624 d.f.; resid. dev.: 6,175; 23,620 d.f., AIC: 6,185. Subset 36 with Imputed Values (95th pctl.); 23,625 obs. Null dev.: 3,876; 23,624 d.f.; resid. dev.: 3,816; 23,620 d.f., AIC: 3,826. Signif. as above.

## Table B37. Logit Model Results:

## Censored Subset 37 vs. Subsets 37 with Imputed Values

| | Censored Subset 37 | Subset 37 w. Imputed Values (90TH pctl.) | Subset 37 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 2.681*** | 3.324*** | 3.902*** |
| *#Direct* | -0.033** | -0.068*** | -0.068*** |
| *#SEngVis* | -0.146*** | -0.139*** | -0.135*** |
| *#AdsVis* | 0.173*** | 0.115*** | 0.145*** |
| *#PersonVis* | 0.516*** | 0.398*** | 0.466*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 37; 23,651 obs. Null dev.: 9,534; 23,650 d.f.; resid. dev.: 9,314; 23,646 d.f., AIC: 9,324. Subset 37 with Imputed Values (90th pctl.); 23,651 obs. Null dev.: 6,488; 23,650 d.f.; resid. dev.: 6,375; 23,646 d.f., AIC: 6,385. Subset 37 with Imputed Values (95th pctl.); 23,651 obs. Null dev.: 4,024; 23,650 d.f.; resid. dev.: 3,950; 23,646 d.f., AIC: 3,960. Signif. as above.

## Table B38. Logit Model Results:

## Censored Subset 38 vs. Subsets 38 with Imputed Values

| | Censored Subset 38 | Subset 38 w. Imputed Values (90TH pctl.) | Subset 38 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 2.677*** | 3.306*** | 3.968*** |
| *#Direct* | -0.029** | -0.068*** | -0.074*** |
| *#SEngVis* | -0.155*** | -0.153*** | -0.152*** |
| *#AdsVis* | 0.171*** | 0.116*** | 0.102** |
| *#PersonVis* | 0.514*** | 0.478*** | 0.467*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 38; 23,718 obs. Null dev.: 9,599; 23,717 d.f.; resid. dev.: 9,377; 23,713 d.f., AIC: 9,387. Subset 38 with Imputed Values (90th pctl.); 23,718 obs. Null dev.: 6,478; 23,717 d.f.; resid. dev.: 6,340; 23,713 d.f., AIC: 6,350. Subset 38 with Imputed Values (95th pctl.); 23,718 obs. Null dev.: 3,953; 23,717 d.f.; resid. dev.: 3,876; 23,713 d.f., AIC: 3,886. Signif. as above.

## Table B39. Logit Model Results:

## Censored Subset 39 vs. Subsets 39 with Imputed Values

| | Censored Subset 39 | Subset 39 w. Imputed Values (90TH pctl.) | Subset 39 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 2.678*** | 3.291*** | 3.948*** |
| *#Direct* | -0.032** | -0.069*** | -0.079*** |
| *#SEngVis* | -0.155*** | -0.165*** | -0.157*** |
| *#AdsVis* | 0.143*** | 0.115*** | 0.111** |
| *#PersonVis* | 0.494*** | 0.466*** | 0.491*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 3; 23,598 obs. Null dev.: 9,697; 23,597 d.f.; resid. dev.: 9,485; 23,593 d.f., AIC: 9,495. Subset 39 with Imputed Values (90th pctl.); 23,598 obs. Null dev.: 6,574; 23,597 d.f.; resid. dev.: 6,431; 23,593 d.f., AIC: 6,441. Subset 39 with Imputed Values (95th pctl.); 23,598 obs. Null dev.: 3,982; 23,597 d.f.; resid. dev.: 3,895; 23,593 d.f., AIC: 3,905. Signif. as above.

## Table B40. Logit Model Results:

## Censored Subset 40 vs. Subsets 40 with Imputed Values

| | Censored Subset 40 | Subset 40 w. Imputed Values (90TH pctl.) | Subset 40 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 2.749*** | 3.346*** | 3.996*** |
| *#Direct* | -0.039*** | -0.073*** | -0.080*** |
| *#SEngVis* | -0.158*** | -0.165*** | -0.162*** |
| *#AdsVis* | 0.132*** | 0.114*** | 0.115** |
| *#PersonVis* | 0.454*** | 0.366*** | 0.477*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 40; 23,465 obs. Null dev.: 9,362; 23,464 d.f.; resid. dev.: 9,174; 23,460 d.f., AIC: 9,184. Subset 40 with Imputed Values (90th pctl.); 23,465 obs. Null dev.: 6,483; 23,464 d.f.; resid. dev.: 6,364; 23,460 d.f., AIC: 6,374. Subset 40 with Imputed Values (95th pctl.); 23,465 obs. Null dev.: 3,830; 23,464 d.f.; resid. dev.: 3,747; 23,460 d.f., AIC: 3,757. Signif. as above.

## Table B41. Logit Model Results:

## Censored Subset 41 vs. Subsets 41 with Imputed Values

| | Censored Subset 41 | Subset 41 w. Imputed Values (90th pctl.) | Subset 41 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.816*** | 3.361*** | 4.045*** |
| *#Direct* | -0.038*** | -0.073*** | -0.083*** |
| *#SEngVis* | -0.166*** | -0.180*** | -0.169*** |
| *#AdsVis* | 0.100*** | 0.102*** | 0.092** |
| *#PersonVis* | 0.459*** | 0.408*** | 0.462*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 41; 23,370 obs. Null dev.: 9,003; 23,369 d.f.; resid. dev.: 8,822; 23,365 d.f., AIC: 8,832. Subset 41 with Imputed Values (90th pctl.); 23,370 obs. Null dev.: 6,373; 23,369 d.f.; resid. dev.: 6,242; 23,365 d.f., AIC: 6,252.Subset 41 with Imputed Values (95th pctl.); 23,370 obs. Null dev.: 3,744; 23,369 d.f.; resid. dev.: 3,662; 23,365 d.f., AIC: 3,672. Signif. as above.

## Table B42. Logit Model Results:

## Censored Subset 42 vs. Subsets 42 with Imputed Values

| | Censored Subset 42 | Subset 42 w. Imputed Values (90th pctl.) | Subset 42 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.897*** | 3.401*** | 4.080*** |
| *#Direct* | -0.039*** | -0.070*** | -0.087*** |
| *#SEngVis* | -0.187*** | -0.195*** | -0.209*** |
| *#AdsVis* | 0.081*** | 0.080** | 0.081* |
| *#PersonVis* | 0.464*** | 0.452*** | 0.590*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 42; 23,315 obs. Null dev.: 8,618; 23,314 d.f.; resid. dev.: 8,433; 23,310 d.f., AIC: 8,443. Subset 42 with Imputed Values (90th pctl.); 23,315 obs. Null dev.: 6,182; 23,314 d.f.; resid. dev.: 6,037; 23,310 d.f., AIC: 6,047. Subset 42 with Imputed Values (95th pctl.); 23,315 obs. Null dev.: 3,617; 23,314 d.f.; resid. dev.: 3,506; 23,310 d.f., AIC: 3,516. Signif. as above.

## Table B43. Logit Model Results:

## Censored Subset 43 vs. Subsets 43 with Imputed Values

| | Censored Subset 43 | Subset 43 w. Imputed Values (90th pctl.) | Subset 43 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.923*** | 3.425*** | 4.081*** |
| *#Direct* | -0.042*** | -0.072*** | -0.091*** |
| *#SEngVis* | -0.201*** | -0.184*** | -0.245*** |
| *#AdsVis* | 0.083*** | 0.049 | 0.110** |
| *#PersonVis* | 0.490*** | 0.453*** | 0.584*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 43; 23,351 obs. Null dev.: 8,474; 23,350 d.f.; resid. dev.: 8,274; 23,346 d.f., AIC: 8,284. Subset 43 with Imputed Values (90th pctl.); 23,351 obs. Null dev.: 6,142; 23,350 d.f.; resid. dev.: 6,000; 23,346 d.f., AIC: 6,010. Subset 43 with Imputed Values (95th pctl.); 23,351 obs. Null dev.: 3,668; 23,350 d.f.; resid. dev.: 3,540; 23,346 d.f., AIC: 3,550. Signif. as above.

## Table B44. Logit Model Results:

## Censored Subset 44 vs. Subsets 44 with Imputed Values

| | Censored Subset 44 | Subset 44 w. Imputed Values (90th pctl.) | Subset 44 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.937*** | 3.445*** | 4.043*** |
| *#Direct* | -0.045*** | -0.067*** | -0.085*** |
| *#SEngVis* | -0.184*** | -0.201*** | -0.243*** |
| *#AdsVis* | 0.087*** | 0.078** | 0.099** |
| *#PersonVis* | 0.482*** | 0.399*** | 0.579*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 44; 23,388 obs. Null dev.: 8,354; 23,387 d.f.; resid. dev.: 8,170; 23,383 d.f., AIC: 8,180. Subset 44 with Imputed Values (90th pctl.); 23,388 obs. Null dev.: 6,095; 23,387 d.f.; resid. dev.: 5,965; 23,383 d.f., AIC: 5,975. Subset 44 with Imputed Values (95th pctl.); 23,388 obs. Null dev.: 3,778; 23,387 d.f.; resid. dev.: 3,652; 23,383 d.f., AIC: 3,662. Signif. as above.

## Table B45. Logit Model Results:

## Censored Subset 45 vs. Subsets 45 with Imputed Values

| | Censored Subset 45 | Subset 45 w. Imputed Values (90th pctl.) | Subset 45 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.938*** | 3.425*** | 3.949*** |
| *#Direct* | -0.038*** | -0.064*** | -0.088*** |
| *#SEngVis* | -0.175*** | -0.197*** | -0.233*** |
| *#AdsVis* | 0.079*** | 0.078** | 0.106** |
| *#PersonVis* | 0.470*** | 0.412*** | 0.646*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 45; 23,397 obs. Null dev.: 8,330; 23,396 d.f.; resid. dev.: 8,160; 23,392 d.f., AIC: 8,170. Subset 45 with Imputed Values (90th pctl.); 23,397 obs. Null dev.: 6,131; 23,396 d.f.; resid. dev.: 6,003; 23,392 d.f., AIC: 6,013. Subset 45 with Imputed Values (95th pctl.); 23,397 obs. Null dev.: 3,967; 23,396 d.f.; resid. dev.: 3,827; 23,392 d.f., AIC: 3,837. Signif. as above.

## Table B46. Logit Model Results:

## Censored Subset 46 vs. Subsets 46 with Imputed Values

| | Censored Subset 46 | Subset 46 w. Imputed Values (90th pctl.) | Subset 46 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.935*** | 3.459*** | 3.970*** |
| *#Direct* | -0.040*** | -0.067*** | -0.089*** |
| *#SEngVis* | -0.150*** | -0.177*** | -0.216*** |
| *#AdsVis* | 0.065** | 0.058* | 0.097** |
| *#PersonVis* | 0.456*** | 0.393*** | 0.616*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 46; 23,251 obs. Null dev.: 8,292; 23,250 d.f.; resid. dev.: 8,138; 23,246 d.f., AIC: 8,148. Subset 46 with Imputed Values (90th pctl.); 23,251 obs. Null dev.: 5,982; 23,250 d.f.; resid. dev.: 5,866; 23,246 d.f., AIC: 5,876. Subset 46 with Imputed Values (95th pctl.); 23,251 obs. Null dev.: 3,889; 23,250 d.f.; resid. dev.: 3,760; 23,246 d.f., AIC: 3,770. Signif. as above.

## Table B47. Logit Model Results:

## Censored Subset 47 vs. Subsets 47 with Imputed Values

| | Censored Subset 47 | Subset 47 w. Imputed Values (90th pctl.) | Subset 47 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.981*** | 3.520*** | 4.005*** |
| *#Direct* | -0.045*** | -0.068*** | -0.087*** |
| *#SEngVis* | -0.147*** | -0.176*** | -0.205*** |
| *#AdsVis* | 0.055** | 0.046 | 0.076* |
| *#PersonVis* | 0.491*** | 0.416*** | 0.529*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 47; 23,250 obs. Null dev.: 7,993; 23,249 d.f.; resid. dev.: 7,830; 23,245 d.f., AIC: 7,840. Subset 47 with Imputed Values (90th pctl.); 23,250 obs. Null dev.: 5,703; 23,249 d.f.; resid. dev.: 5,587; 23,245 d.f., AIC: 5,597. Subset 47 with Imputed Values (95th pctl.); 23,250 obs. Null dev.: 3,872; 23,249 d.f.; resid. dev.: 3,761; 23,245 d.f., AIC: 3,771. Signif. as above.

## Table B48. Logit Model Results:

## Censored Subset 48 vs. Subsets 48 with Imputed Values

| | Censored Subset 48 | Subset 48 w. Imputed Values (90th pctl.) | Subset 48 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 2.988*** | 3.532*** | 4.014*** |
| *#Direct* | -0.050*** | -0.074*** | -0.093*** |
| *#SEngVis* | -0.121*** | -0.154*** | -0.156*** |
| *#AdsVis* | 0.074** | 0.040 | 0.040 |
| *#PersonVis* | 0.540*** | 0.460*** | 0.489*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 48; 23,361 obs. Null dev.: 7,774; 23,360 d.f.; resid. dev.: 7,604; 23,356 d.f., AIC: 7,614. Subset 48 with Imputed Values (90th pctl.); 23,261 obs. Null dev.: 5,586; 23,360 d.f.; resid. dev.: 5,467; 23,356 d.f., AIC: 5,477. Subset 48 with Imputed Values (95th pctl.); 23,261 obs. Null dev.: 3,909; 23,360 d.f.; resid. dev.: 3,808; 23,356 d.f., AIC: 3,818. Signif. as above.

## Table B49. Logit Model Results:

## Censored Subset 49 vs. Subsets 49 with Imputed Values

| | Censored Subset 51 | Subset 9 w. Imputed Values (90th pctl.) | Subset 49 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 3.0498*** | 3.5724*** | 3.969*** |
| *#Direct* | -0.0532*** | -0.0718*** | -0.090*** |
| *#SEngVis* | -0.1350*** | -0.1652*** | -0.176*** |
| *#AdsVis* | 0.0664** | 0.0331 | 0.039 |
| *#PersonVis* | 0.5812*** | 0.5096*** | 0.587*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 49; 23,418 obs. Null dev.: 7,450; 23,417 d.f.; resid. dev.: 7,267; 23,413 d.f., AIC: 7,277. Subset 49 with Imputed Values (90th pctl.); 23,418 obs. Null dev.: 5,391; 23,417 d.f.; resid. dev.: 5,261; 23,413 d.f., AIC: 5,271. Subset 49 with Imputed Values (95th pctl.); 23,418 obs. Null dev.: 3,976; 23,417 d.f.; resid. dev.: 3,854; 23,413 d.f., AIC: 3,864. Signif. as above.

## Table B50. Logit Model Results:

## Censored Subset 50 vs. Subsets 50 with Imputed Values

| | Censored Subset 50 | Subset 50 w. Imputed Values (90th pctl.) | Subset 50 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 3.113*** | 3.599*** | 3.939*** |
| *#Direct* | -0.056*** | -0.074*** | -0.089*** |
| *#SEngVis* | -0.138*** | -0.174*** | -0.176*** |
| *#AdsVis* | 0.051* | 0.032 | 0.030 |
| *#PersonVis* | 0.569*** | 0.546*** | 0.644*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 50; 23,461 obs. Null dev.: 7,224; 23,460 d.f.; resid. dev.: 7,048; 23,456 d.f., AIC: 7,058. Subset 50 with Imputed Values (90th pctl.); 23,641 obs. Null dev.: 5,275; 23,460 d.f.; resid. dev.: 5,134; 23,456 d.f., AIC: 5,144.Subset 50 with Imputed Values (95th pctl.); 23,641 obs. Null dev.: 4,034; 23,460 d.f.; resid. dev.: 3,901; 23,456 d.f., AIC: 3,911. Signif. as above.

## Table B51. Logit Model Results:

## Censored Subset 51 vs. Subsets 51 with Imputed Values

| | Censored Subset 51 | Subset 51 w. Imputed Values (90th pctl.) | Subset 51 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 3.168*** | 3.641*** | 3.990*** |
| *#Direct* | -0.060*** | -0.075*** | -0.091*** |
| *#SEngVis* | -0.130*** | -0.149*** | -0.161*** |
| *#AdsVis* | 0.064** | 0.030 | 0.027 |
| *#PersonVis* | 0.573*** | 0.508*** | 0.576*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 51; 23,464 obs. Null dev.: 6,891; 23,463 d.f.; resid. dev.: 6,724; 23,459 d.f., AIC: 6,734. Subset 51 with Imputed Values (90th pctl.); 13,464 obs. Null dev.: 5,111; 23,463 d.f.; resid. dev.: 4,991; 23,495 d.f., AIC: 5,001. Subset 51 with Imputed Values (95th pctl.); 13,464 obs. Null dev.: 3,928; 23,463 d.f.; resid. dev.: 3,810; 23,495 d.f., AIC: 3,820. Signif. as above.

## Table B52. Logit Model Results:

## Censored Subset 52 vs. Subsets 52 with Imputed Values

| | Censored Subset 52 | Subset 52 w. Imputed Values (90th pctl.) | Subset 52 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 3.228*** | 3.668*** | 3.997*** |
| *#Direct* | -0.053*** | -0.062*** | -0.075*** |
| *#SEngVis* | -0.128*** | -0.143*** | -0.159*** |
| *#AdsVis* | 0.060* | 0.029 | 0.034 |
| *#PersonVis* | 0.517*** | 0.448*** | 0.476*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 52; 23,472 obs. Null dev.: 6,661; 23,471 d.f.; resid. dev.: 6,524; 23,467 d.f., AIC: 6,534. Subset 52 with Imputed Values (90th pctl.); 23,472 obs. Null dev.: 5,021; 23,471 d.f.; resid. dev.: 4,927; 23,467 d.f., AIC: 4,937. Subset 52 with Imputed Values (95th pctl.); 23,472 obs. Null dev.: 3,929; 23,471 d.f.; resid. dev.: 3,841; 23,467 d.f., AIC: 3,851. Signif. as above.

## Table B53. Logit Model Results:

## Censored Subset 53 vs. Subsets 53 with Imputed Values

| | Censored Subset 53 | Subset 53 w. Imputed Values (90th pctl.) | Subset 53 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 3.237*** | 3.690*** | 4.037*** |
| *#Direct* | -0.053*** | -0.058*** | -0.068*** |
| *#SEngVis* | -0.142*** | -0.149*** | -0.149*** |
| *#AdsVis* | 0.062** | 0.021 | 0.004 |
| *#PersonVis* | 0.487*** | 0.431*** | 0.405*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 53; 23,255 obs. Null dev.: 6,634; 23,254 d.f.; resid. dev.: 6,502; 23,250 d.f., AIC: 6,512. Subset 53 with Imputed Values (90th pctl.); 23,255 obs. Null dev.: 4,928; 23,254 d.f.; resid. dev.: 4,839; 23,250 d.f., AIC: 4,849. Subset 53 with Imputed Values (95th pctl.); 23,255 obs. Null dev.: 3,872; 23,254 d.f.; resid. dev.: 3,799; 23,250 d.f., AIC: 3,809. Signif. as above.

## Table B54. Logit Model Results:

## Censored Subset 54 vs. Subsets 54 with Imputed Values

| | Censored Subset 54 | Subset 54 w. Imputed Values (90th pctl.) | Subset 54 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 3.354*** | 3.813*** | 4.152*** |
| *#Direct* | -0.062*** | -0.069*** | -0.073*** |
| *#SEngVis* | -0.139*** | -0.119*** | -0.123*** |
| *#AdsVis* | 0.069** | 0.010 | -0.014 |
| *#PersonVis* | 0.411*** | 0.371*** | 0.345*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 54; 23,101 obs. Null dev.: 6,190; 23,100 d.f.; resid. dev.: 6,083; 23,096 d.f., AIC: 6,093. Subset 54 with Imputed Values (90th pctl.); 23,101 obs. Null dev.: 4,545; 23,100 d.f.; resid. dev.: 4,473; 23,096 d.f., AIC: 4,483. Subset 54 with Imputed Values (95th pctl.); 23,101 obs. Null dev.: 3,594; 23,100 d.f.; resid. dev.: 3,533; 23,096 d.f., AIC: 3,543. Signif. as above.

## Table B55. Logit Model Results:

## Censored Subset 55 vs. Subsets 55 with Imputed Values

| | Censored Subset 55 | Subset 55 w. Imputed Values (90th pctl.) | Subset 55 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 3.344*** | 3.824*** | 4.168*** |
| *#Direct* | -0.053*** | -0.065*** | -0.065*** |
| *#SEngVis* | -0.120*** | -0.118*** | -0.123*** |
| *#AdsVis* | 0.058* | 0.002 | -0.031 |
| *#PersonVis* | 0.640*** | 0.640*** | 0.669*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 55; 23,018 obs. Null dev.: 5,776; 23,017 d.f.; resid. dev.: 5,629; 23,013 d.f., AIC: 5,639. Subset 55 with Imputed Values (90th pctl.); 23,018 obs. Null dev.: 4,156; 23,017 d.f.; resid. dev.: 4,048; 23,013 d.f., AIC: 4,058. Subset 55 with Imputed Values (95th pctl.); 23,018 obs. Null dev.: 3,218; 23,017 d.f.; resid. dev.: 3,129; 23,013 d.f., AIC: 3,139. Signif. as above.

## Table B56. Logit Model Results:

## Censored Subset 56 vs. Subsets 56 with Imputed Values

| | Censored Subset 56 | Subset 56 w. Imputed Values (90th pctl.) | Subset 56 w. Imputed Values (95th pctl.) |
|---|---|---|---|
| *Intercept* | 3.448*** | 3.886*** | 4.186*** |
| *#Direct* | -0.058*** | -0.068*** | -0.066*** |
| *#SEngVis* | -0.128*** | -0.137*** | -0.148*** |
| *#AdsVis* | 0.085** | 0.052 | 0.034 |
| *#PersonVis* | 0.715*** | 0.739*** | 0.802*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 56; 23,109 obs. Null dev.: 5,236; 23,108 d.f.; resid. dev.: 5,082; 23,104 d.f., AIC: 5,092. Subset 56 with Imputed Values (90th pctl.); 23,109 obs. Null dev.: 3,818; 23,108 d.f.; resid. dev.: 3,703; 23,104 d.f., AIC: 3,713. Subset 56 with Imputed Values (95th pctl.); 23,109 obs. Null dev.: 3,010; 23,108 d.f.; resid. dev.: 2,914; 23,104 d.f., AIC: 2,924. Signif. as above.

## Table B57. Logit Model Results:

## Censored Subset 57 vs. Subsets 57 with Imputed Values

| | Censored Subset 57 | Subset 57 w. Imputed Values (90TH pctl.) | Subset 57 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 3.630*** | 3.976*** | 4.213*** |
| *#Direct* | -0.068*** | -0.080*** | -0.071*** |
| *#SEngVis* | -0.121*** | -0.126*** | -0.120** |
| *#AdsVis* | 0.092** | 0.065 | 0.051 |
| *#PersonVis* | 0.664*** | 0.720*** | 0.776*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 57; 23,137 obs. Null dev.: 4,632; 23,136 d.f.; resid. dev.: 4,505; 23,132 d.f., AIC: 4,515. Subset 57 with Imputed Values (90th pctl.); 23,137 obs. Null dev.: 3,587; 23,136 d.f.; resid. dev.: 3,477; 23,132 d.f., AIC: 3,487. Subset 57 with Imputed Values (95th pctl.); 23,137 obs. Null dev.: 2,914; 23,136 d.f.; resid. dev.: 2,826; 23,132 d.f., AIC: 2,836. Signif. as above.

## Table B58. Logit Model Results:

## Censored Subset 58 vs. Subsets 58 with Imputed Values

| | Censored Subset 58 | Subset 58 w. Imputed Values (90TH pctl.) | Subset 58 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 3.620*** | 3.946*** | 4.152*** |
| *#Direct* | -0.061*** | -0.074*** | -0.070*** |
| *#SEngVis* | -0.125*** | -0.143*** | -0.130*** |
| *#AdsVis* | 0.101** | 0.087* | 0.051 |
| *#PersonVis* | 0.791*** | 0.885*** | 0.846*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 58; 23,063 obs. Null dev.: 4,481; 23,062 d.f.; resid. dev.: 4,337; 23,058 d.f., AIC: 4,347. Subset 58 with Imputed Values (90th pctl.); 23,063 obs. Null dev.: 3,500; 23,062 d.f.; resid. dev.: 3,371; 23,058 d.f., AIC: 3,381. Subset 58 with Imputed Values (95th pctl.); 23,063 obs. Null dev.: 3,018; 23,062 d.f.; resid. dev.: 2,918; 23,058 d.f., AIC: 2,928. Signif. as above.

## Table B59. Logit Model Results:

## Censored Subset 59 vs. Subsets 59 with Imputed Values

| | Censored Subset 59 | Subset 59 w. Imputed Values (90TH pctl.) | Subset 59 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 3.680*** | 3.920*** | 4.068*** |
| *#Direct* | -0.062*** | -0.074*** | -0.067*** |
| *#SEngVis* | -0.090* | -0.119** | -0.108* |
| *#AdsVis* | 0.105** | 0.123** | 0.079 |
| *#PersonVis* | 0.985*** | 1.084*** | 1.158*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 59; 22,978 obs. Null dev.: 4,042; 22,977 d.f.; resid. dev.: 3,883; 22,973 d.f., AIC: 3,893. Subset 59 with Imputed Values (90th pctl.); 22,978 obs. Null dev.: 3,345; 22,977 d.f.; resid. dev.: 3,196; 22,973 d.f., AIC: 3,206. Subset 59 with Imputed Values (95th pctl.); 22,978 obs. Null dev.: 2,963; 22,977 d.f.; resid. dev.: 2,830; 22,973 d.f., AIC: 2,840. Signif. as above.

## Table B60. Logit Model Results:

## Censored Subset 60 vs. Subsets 60 with Imputed Values

| | Censored Subset 60 | Subset 60 w. Imputed Values (90TH pctl.) | Subset 60 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 3.712*** | 3.959*** | 4.038*** |
| *#Direct* | -0.055*** | -0.069*** | -0.061*** |
| *#SEngVis* | -0.090* | -0.112** | -0.077 |
| *#AdsVis* | 0.130** | 0.137** | 0.122* |
| *#PersonVis* | 1.082*** | 1.196*** | 1.387*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 60; 22,722 obs. Null dev.: 3,765; 22,721 d.f.; resid. dev.: 3,604; 22,717 d.f., AIC: 3,614. Subset 60 with Imputed Values (90th pctl.); 22,722 obs. Null dev.: 3,106; 22,721 d.f.; resid. dev.: 2,957; 22,717 d.f., AIC: 2,967. Subset 60 with Imputed Values (95th pctl.); 22,722 obs. Null dev.: 2,797; 22,721 d.f.; resid. dev.: 2,651; 22,717 d.f., AIC: 2,661. Signif. as above.

**Table B61. Logit Model Results:**

**Censored Subset 61 vs. Subsets 61 with Imputed Values**

| | Censored Subset 61 | Subset 61 w. Imputed Values (90TH pctl.) | Subset 61 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 3.820*** | 3.980*** | 3.999*** |
| *#Direct* | -0.050** | -0.066*** | -0.059*** |
| *#SEngVis* | -0.119** | -0.125** | -0.076 |
| *#AdsVis* | 0.210*** | 0.226*** | 0.201*** |
| *#PersonVis* | 1.409*** | 2.010*** | 2.639*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 61; 22,519 obs. Null dev.: 3187; 22,518 d.f.; resid. dev.: 3,009; 22,514 d.f., AIC: 3,019. Subset 61 with Imputed Values (90th pctl.); 22,519 obs. Null dev.: 2,694; 22,518 d.f.; resid. dev.: 2,487; 22,514 d.f., AIC: 2,497. Subset 61 with Imputed Values (95th pctl.); 22,519 obs. Null dev.: 2,520; 22,518 d.f.; resid. dev.: 2,297; 22,514 d.f., AIC: 2,307. Signif. as above.

**Table B62. Logit Model Results:**

**Censored Subset 62 vs. Subsets 62 with Imputed Values**

| | Censored Subset 62 | Subset 62 w. Imputed Values (90TH pctl.) | Subset 62 w. Imputed Values (95TH pctl.) |
|---|---|---|---|
| *Intercept* | 4.003*** | 4.003*** | 4.003*** |
| *#Direct* | -0.032 | -0.032 | -0.032 |
| *#SEngVis* | -0.155** | -0.155** | -0.155** |
| *#AdsVis* | 0.291*** | 0.291*** | 0.291*** |
| *#PersonVis* | 4.890*** | 4.890*** | 4.890*** |

**Notes.** Model: logit; dep. var.: *Purch* (0/1). Censored Subset 62; 22,401 obs. Null dev.: 2,351; 22,400 d.f.; resid. dev.: 2,083; 22,396 d.f., AIC: 2,093. Subset 62 with Imputed Values (90th pctl.); 22,401 obs. Null dev.: 2,351; 22,400 d.f.; resid. dev.: 2,083; 22,396 d.f., AIC: 2,093. Subset 62 with Imputed Values (95th pctl.); 22,401 obs. Null dev.: 2,351; 22,400 d.f.; resid. dev.: 2,083; 22,396 d.f., AIC: 2,093. Signif. as above.