VARYING COEFFICIENTS IN LOGISTIC REGRESSION WITH APPLICATIONS TO MARKETING RESEARCH

BY

ERIN M. CONDON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Xuming He, Chair
Professor Jeff Douglas
Professor Adam Martinsek
Professor Peiyong Qu

Abstract

In the marketing research world today, companies have access to massive amounts of data regarding the purchase behavior of consumers. Researchers study this data to understand how outside factors, such as demographics and marketing tools, affect the probability that a given consumer will make a purchase. Through the use of panel data, we tackle these questions and propose a logistic regression model in which coefficients can vary based on a consumer's purchase history. We also introduce a two-step procedure for model selection that uses a group LASSO penalty to decide which are informative and which variables need varying coefficients in the model.

To Donald E. Miller (1940 - 2008)

For challenging me everyday and igniting my passion for statistics.

Acknowledgments

I would like to thank the Statistics Department at the University of Illinois for the wonderful opportunities and supportive faculty that I have had throughout my time in the program. I would specifically like to thank my advisor, Xuming He, for all of his knowledge, patience and time. Not one step of this could have been done without him. I would also like to give a special thank you to my committee members, Jeff Douglas, Adam Martinsek, and Annie Qu, for all of their time and thorough feedback. In addition, I would like to thank Melissa Banks, Judy Whittington, and Julie Patterson who made the day to day elements of this program not only possible but enjoyable.

Secondly, I want to thank SymphonyIRI Group not only for their academic dataset but for the great network of knowledge and support they have provided me over the past five years. I specifically wish to thank Mike Kruger for all of his mentoring and support and Aaron Augustine for always investing in my future.

Finally I want to thank my friends and family who have been an amazing source of support. To my friends, you have gone above and beyond as librarians, cheerleaders, counselors, hostesses, and more. I couldn't ask for a better group. Thanks to Nana, for her unconditional love and support. Mom and Dad, thank you for always inspiring me to do more, see more, be more, and teach more. To my siblings, (Lauren, Sean, Caitlin, Bridget, and Claire) who have supported me every step of the way, thanks for always keeping me on my toes and making it impossible to be the most successful child.

Information Resources, Inc. ("IRI") has changed its name to SymphonyIRI Group, Inc. All estimates and analyses in this paper based on SymphonyIRI Group, Inc. data are by the author(s) and not by SymphonyIRI Group, Inc.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Businesses run on the simple principle of supply and demand. Companies want to produce their goods or services to meet the level at which the consumers are in demand of them and then push consumers to demand more. Thus the key to success is to understand how consumers think and behave. In recent years this phenomenon has become much simpler due to the increase in collection of purchase data. In the business world today, companies have access to nearly unlimited data regarding the purchase behavior of consumers. Whether they are produced by internet searches, bar code scanning, or internal customer data, useful data are everywhere waiting to be analyzed. Economists and marketing researchers have used those data to try and learn as much about their consumers as possible.

A common business practice is to use this data to model and predict the behavior of their consumers. While there are many existing models based on purchase behavior, many models lack the flexibility to adjust to the detailed information we have today on our consumers. In addition to the traditional modeling, with the large amount of data that is collected, there also has to be an efficient and accurate way to perform variable section. In this chapter we introduce the type of data we will analyze in §1.1 and review the marketing literature in §1.2. In §1.3 we describe the model we suggest to use and then the last two sections, §1.4 and §1.5, are devoted to the issue of model selection.

1.1 Introduction to the Data

SymphonyIRI Group is a leading global marketing research company that specializes in the collection and analysis of purchase data, specifically purchase data on consumer packaged goods (CPG). SymphonyIRI Group provides innovative solutions and services to leading CPG, retail, and health care companies across the world. In 2008, in addition to their client services, SymphonyIRI Group developed an Academic Data set for researchers to use. “These data are intended to enable academic researchers to study important research topics in marketing and economics that are of concern to practitioners, policy makers, and scholars”,(Bronnenberg and Mela (2008)). The data set provides researchers with five years of weekly store data

(2001-2005) for chain grocery and drug stores in 47 markets, store-week-UPC level for thirty large CPG categories covering forty-seven markets, five years of panel data for two markets, and advertising data for two categories.

In order to better understand the relationship between consumers and products, we focus on the collection of data from SymphonyIRI Group's panelists. A panelist is a consumer in the United States that is recruited by SymphonyIRI Group to track all of their purchases over a certain amount of time. Each panelist is given a scanning apparatus that is used to scan the barcode on every CPG he/she purchases. For example, a panelist would have to scan every barcode from a large grocery shopping trip or the one candy bar that he/she picked up at a convenience store. The data follow the weekly activity of a group of consumers and can be used to analyze how their purchase behavior is affected by relevant factors such as demographics and marketing tools. Thus the end result is a longitudinal data set that measures weekly purchase behavior of a household across many different outlets.

SymphonyIRI Group's panelists provided in the academic dataset come from two markets in the United States: Eau Claire, Wisconsin and Pittsfield, Massachusetts. With panel data we are provided with consumer tracking data in which we have an ongoing collection of every transaction a consumer has made. This type of data is invaluable to companies because they are now given the opportunity to follow each consumer and know when and how frequently an individual is purchasing their product. The panel data contain the household ID, the unique store key where the purchase was made, the week the purchase was made, the outlet that the store falls into (grocery, drug, mass, etc), the number of units purchased, the number of dollars spent, and the unique key for the product being purchased. In addition, SymphonyIRI Group's demographic data for the research data set contains over thirty variables about each panelist's household including but not limited to information on age, education, family size, occupation, and income. Using these variables, we are able to know more about the consumers who are purchasing certain products.

Along with the purchases made, we have data that correspond to the products being purchased and the marketing strategies that were used on the products at the time of purchase. Every product that is purchased can be linked up to dictionary information on each product that contains information on the company that makes the product as well as qualities about the product such as size, flavor, and texture. In addition to the physical characteristics of each product, we know if any marketing strategies were used at the time of the purchase. The marketing variables in the data set come from the stores and are simply a dummy variable that is attached to each product on a weekly basis. Thus, for each week we know if there was a price reduction, if the product was on display in the store, or if the product was featured in an ad in a paper.

Two main questions that are often at the center of marketing research in terms of consumer level data involve (1) what demographic and/or promotional factors contribute to the decision to make a purchase? and (2) How can a consumer's previous purchase behavior be used to predict what will happen in the future?

1.2 Marketing Review

Market researchers have studied consumer purchase behaviors from all different angles. Studies are constantly being done on new brands, loyalty brands, demographic influences, economic influences and many more. While the data surrounding consumer behavior can spark many different studies, we choose to focus on modeling the weekly decision to make a purchase for an existing product. There have been many factors that have been analyzed in order to see how they affect consumers' decisions. The main three factors that we will focus on are:

1. Promotional variables: Including but not limited to price reductions, in-store displays, and features in advertising.
2. Demographics: Including but not limited to household size, income, race, home-ownership and marital status.
3. Previous Purchases: Influence of previous purchase behavior in terms of frequency, recency, stability.

We begin by focusing on the the effects of promotional variables.

Promotional variables include the effect of displaying your product in a store, the effect of featuring your product in advertisements, and the effect of price reduction. These factors have been studied as main influences for whether or not a purchase will be made. Kumar and Leone (1988), discovered that price promotion, display, and feature advertising on a specific brand have a positive effect on the consumption of that product for two reasons. The first, the consumer would opt to substitute the brand that they had always been loyal to at that store with the brand that was on a price reduction. The second factor is store substitution in which advertising price reduction in features causes consumers to leave the store they are loyal to and purchase the brand at the store with the price reduction. Thus, whether a result of brand substitution or store substitution, it was confirmed that advertising did make a positive difference on sales.

Tellis (2007) published a similar study in which he focused on how advertising affects a consumer's brand loyalty. The study found that the relationship between advertising and sales is nonlinear. Instead, the relationship follows a bell-shaped curve in which the optimum amount of brand advertising was found to be 2 - 3 marketing exposures of the brand per week. Any more than that will have a reverse effect in the sense that consumers will be bothered and turned off by the product. He also found that advertising has a larger effect on loyalty customers than nonloyal customers since loyal customers tend to purchase the brand in large volume when it is on sale. He concluded that overall increased sales are most influenced by features, then displays, and lastly by price reductions alone.

Vakratsas and Ambler (1999) did a review of over 250 journal articles covering models on the effect of marketing tools on the consumption of goods. In the end they proposed a set of five generalizations on how advertising works. These generalizations are based on three key factors that contribute to the effect of advertising: experience, effect, and cognition. Experience is the consumer's history of brand usage, response to advertising, and purchase history. Cognitive is the thought process that occurs once one is presented with a marketing tool. Finally, effect is the way that a consumer reacts once they have thought about the advertisement; it is the way the advertisement makes one feel. The combination of these three ideas and how they affect purchase behavior are summarized in the generalizations below:

(1) Experience, effect, and cognition are the three key intermediate advertising effects, and the omission of any one can lead to overestimation of the effect of the others.

(2) Short-term advertising elasticities are small and decrease during the product life cycle.

(3) In mature, frequently purchased packaged goods markets, returns to advertising diminish fast. A small frequency, therefore (one to three reminders per purchase cycle), is sufficient for advertising an established brand.

(4) The concept of a space of intermediate effects is supported, but a hierarchy (sequence) is not.

(5) Cognitive bias interferes with affect measurement.

These five generalizations summarize years of studies done on marketing tools that have been published. While they emphasize how consumers react to marketing tools, they also highlight the fact that marketing variables alone do not influence purchase behavior. This is further studied by looking at how marketing variables and demographics can be used together to build predictive models for consumer purchases behavior.

Logistic regression is commonly used in predictive models based on household panel data. These models use deterministic functions in order to take qualitative variables such as demographics, marketing variables, and indications of previous purchases to generate probabilities of making purchases (Allenby (1990)). Beginning with using a conditional logit to model individual choice behavior based on the distribution of population choices (McFadden (1974)), many empirical studies have been conducted using panel level data. Guadagni and Little (1983) took these ideas and sparked a series of studies based on longitudinal scanner data when they introduced a multinomial logit brand choice model. By using data from 100 panelists collected over 32 weeks on the purchase of regular ground coffee, Guadagni and Little show the importance of brand loyalty and promotion at the individual level. This famous study lead to further work on improving this model with certain marketing mix variables and sometimes household demographic variables (Gupta (1988), L. and Raj (1988), and Lattin and Bucklin (2008)).

Allenby and Lenk (1994) sought to take the findings of their predecessors and build their own logistic normal regression with more of a focus on demographics and characteristics about the actual household. They recognized that marketing variables can have different effects on different households. In order to account for this, they added more household information to the model and included a random effect for the repeated measures on a household. Using a data set on ketchup brands, panel purchases and demographics, Allenby and Lenk include random coefficients for households and cross-sectional and serial correlation for brand preferences. In their paper they found significant differences on how households react to brands based on their incomes and response to promotional effects. Allenby and Rossi (1994) continued to push the boundary in modeling panel data with incorporating Bayesian statistics, brand positioning, and many other marketing techniques.

The final concept mentioned is the study of an individual's purchase behavior over time. One key area in which researchers study a consumer's behavior over time is known as customer lifetime value. This concept has been studied by researchers in order to develop a way of modeling consumer behavior over time. Companies are interested in modeling the probability of a consumer to purchase their product and once they purchase the product, how long they will remain a customer. Two popular models for predicting this behavior are the Pareto/NBD model and the Dirichlet Benchmark model. The first model, the Pareto/NBD model, was introduced in Schmittlein et al. (1987) in order to count customers. This model is based on the idea that customers will buy a product at a steady rate for a certain amount of time and then at some point they are no longer observed as they drop out of the market for that product. Thus the model has two parts: a NBD (Poisson-gamma mixture) model is used to model the repeated buying of the product and then a Pareto (exponential-gamma mix) model is used to model the time until a customer stops purchasing a product. The concepts behind the model were highly praised but the empirical application of the model was quite challenging.

Fader et al. (2005) revisited this model staying true to the theory behind the model, but introducing a much easier way to estimate the parameters of the model. In the new model, the NBD portion was replaced by a beta- geometric model. The new model produced similar results to the original model but is much more user friendly. Fader et al. (2005) propose using the well known RFM (recency, frequency, monetary value) paradigm in combination with the idea of CLV by way of iso-curves. They stress that regression type models are useful in predicting what will happen in the upcoming period, but not what will happen over the lifetime of a customer. Thus, they calculate iso-curves to display the tradeoff between the recency, frequency, and monetary value for each consumer and CLV. By use of iso-curves, there is now a clear picture in which we are able to model all consumers together, regardless of purchase behavior. One simply needs to input the variables: recency, frequency, and monetary values in order to capture the CLV. Glady et al. (2009) also revisited the Pareto/ NBD approach to predicting CLV and introduced a modification to the methodology. They discovered that the dependence between the future number of transactions that a customer is predicted to make and the amount of profit that the company will have as a result of these transactions can be used to increase the precision of the prediction of CLV.

The second model, the Dirichlet Benchmark model proposed by Ehrenberg et al. (2004), predicts overall brand performance measures, such as penetration, market share, and average purchases for buyers, using very little market information. This model assumes that loyalty to all brands follows a normal distribution and the sale of all brands, no matter what the product, follow a similar pattern. The Dirichlet model assumes that the market is steady and applies general laws to describe the behavior of brand performance measures for all products. The model proposes that each consumer buys in the category at a steady long-run rate over a specific amount of time. These long-run intervals are random and independent of each other. Consumers in this model are assumed to be experienced buyers that are no longer influenced by outside factors. In other words, there is no additional learning about the product. The probability of purchasing a brand varies from one person to the next but each person has a fixed likelihood of buying a brand. Although, each person has a fixed propensity to buy, the actual purchase of the product comes at random times. Furthermore, consumers' propensity to purchase a specific brand is not affected by what other brands they are purchasing at the time. Products in the model are not differentiated by their specific characteristics or if they are marketed differently. The model has two steps: a calibration step and a calculation. The calibration of the model consists of four inputs: the percent of all consumers who buy in the category, the average number of purchases each category user makes, the penetration of one of the bigger brands in the category, and the average number of purchases each of the bigger brand's buyers make. From this step there is now a model for all brands in the category and one simply needs to enter a brand's market shares or penetration in order to predict how the brand will perform across all markets.

The Dirichlet model has met many criticisms due to the lack of information used in the model, specifically brand strength, brand differentiation and outside persuasion to purchase a brand. Critics believe the model is too simple and makes big assumptions regarding the lack of influence brands have on a consumer's decision to make a purchase. In a recent paper, Bongers and Hofmeyr (2010) challenge this model and the assumptions that are used about brand performance, the psychology of consumer preference, and the role of advertising. In the paper, they use panel data to point out while the model does an accurate job of calculating brand performance measures, the assumptions on individual behavior are wrong. They believe that an individual's brand preference does change and methods must be developed in order to include this in the model.

Based on the marketing research done, we know that we must develop a model that incorporates marketing influence, demographics, and past purchase behavior. We will focus on the logistic regression model that is used in order to predict the probability of a consumer making a purchase. One criticism that we wish to focus on is the fact that not every household can be captured with one coefficient to represent their behavior. Some households are more price sensitive than others and they can be quite heterogeneous in their responses to the marketing variables (Allenby and Lenk (1994)). Furthermore, demographics and behavioral choices can drive how much people are willing to buy of a certain product. For example, putting a brand of potato chips on display in a store affects the probability of a person who purchases chips all year round the same way as it affects the probability that a health conscious person who never purchases potato chips. Intuitively, this does not seem to make much sense. There may be situations in which the coefficient of a variable in logistic regression is not constant across all cases. In fact the effect may even change in response to another variable. We want to capture these differences in the interaction between the demographics of a household, the promotions of a product and the previous purchase behavior.

Therefore, we begin this study by modeling the probability of a panelist making a purchase. The basic model we begin with is the logistic regression model that models the dichotomous response variable of whether or not a purchase has been made based on a set of explanatory variables. The logistic formula allows us to assign the probability of making a purchase to a consumer by fitting data to a logit function logistic curve. Thus we will be able to produce a probability to make a purchase for every panelist. However, like all regression models there are limitations in the coefficients of the logistic regression model in the sense that it assumes that each consumer is affected by a variable in the same way, which may not always be true. As seen in the review of marketing research, knowing something about the experience of a consumer's interaction with a product can help in predicting future behavior. Given that we have years of panel data, we have the opportunity to know how the panelists have behaved in the past. In the simplest case, we know whether a person has purchased the product before. However, we wish to go deeper to see how the frequency of purchasing the product interacts with the effect of demographics and marketing tools on one's purchase behavior. In order to test this theory statistically we wish to look at the interaction between previous purchases and the standard predictor variables by way of varying coefficient models.

1.3 Varying Coefficient Models

The idea of varying coefficient models first surfaced as an adaptation to the coefficients in linear models. Linear regression is a very common tool for modeling the relationship between variables; however, linear regression models lack flexibility in modeling complex data. The solution has been to replace some or all of the parametric functions of regressors with smooth nonparametric functions (Hastie (1990)). The basic idea behind these new models is to keep the linearity in the model but allow for each coefficient to change smoothly in correspondence to a third variable. To understand this concept better, we can think about interaction terms in linear modeling. Interaction terms are a type of explanatory variable that captures the effect of the relationship between two variables, one being categorical, instead of each of the individual variables alone. Thus, their coefficients represent the effect of what happens when the variables occur at the same time. In varying coefficient models, we also wish to capture the effect of the interaction of two variables, except that we use a continuous variable instead of using a factor. Thus we are no longer just modeling the effect of the coexistence of two variables; we are modeling the relationship between the two variables and allowing the coefficient to vary accordingly. In the end, instead of one coefficient for each explanatory variable, we have a coefficient function that describes the effect of each explanatory variable as a function of third variable. We begin with the varying coefficient model for the general linear model.

Let Y be a random variable that has a distribution based on a parameter η. Let $X_1, X_2, ..., X_p$ and R be our set of predictors. Then, a varying-coefficients model for ordinary least squares regression will have the form:

$$\eta = \beta_0 + X_1\beta_1(R) + ... + X_p\beta_p(R). \tag{1.1}$$

In this model we see that R changes the coefficients of $X_1, X_2, ..., X_p$ through the functions $\beta(R)$.Thus, β_j has some dependence on R that is represented by some sort of interaction between R and X_j. The most common application of the varying - coefficient models are in generalized linear models (Hastie (1990)). Here η is related to the mean of $(Y|X, R)$ by way of the link function: $\eta = g(x)$. These models include but are not limited to the simple Gaussian model where $g(\mu) = \mu$ and Y is normal, log-linear models, and logistic models. We will focus on logistic models where

$$\mu(x) = \frac{\exp\{\alpha + X'\beta(R)\}}{1 + \exp\{\alpha + X'\beta(R)\}} \tag{1.2}$$

In our work, the function $\beta_j(R)$ will be taken to be B-splines. Hastie (1990) introduce three ways to find $\beta_j(R)$ in the linear regression case: using estimation in L_2, using penalized least squares, and using splines. We choose the regression spline approach because splines are piecewise polynomial functions that are used in mathematics to approximate smooth curves. Here, we focus on B-splines, Bickel et al. (1978) explains this specific type of splines that are based off of the Bezier curve.

One can think of B-Splines as a method of connecting n Bezier curves that join together with C^{n-1} continuity. The main components of B-splines are the degree and knots. The degree refers to the polynomial degree of the curve that will be used in the smoothing process and the knots are a vector of nondecreasing real numbers that are the key points in the domain of the data that will be used to fit a smooth curve. Knots can be designated as either boundary knots (the range) or interior knots (any points that fall in the interval of the boundary knots). Boundary knots are always included in the process of computing the curves in order to stabilize the curve, but in reality have no effect on the curve. Note that knots do not have to be actual points in the data, as long as they are in the interval. In the most basic example, we have a polynomial of degree one in which piecewise linear functions are simply connected to each other at the specified knots. As we move up to a degree n model, the knots no longer represent connecting points but rather points at which the curve is C^{n-1} continuous. In Figure 1.1 we can see plots in which B-splines were used in order to fit a smooth curve to the data. The first row contains plots that use piecewise linear functions or splines of degree 1, the second row uses quadratic splines of degree 2, and the final row contains cubic splines. The first column of graphs uses one knot, the second column uses two knots, and then third column uses three knots.

By use of B-spline functions we transform a single coefficient that would be calculated in logistic regression into a set of coefficients. Once we set up these sets of coefficients, we will need to use model selection to determine if these variables should be included in the model.

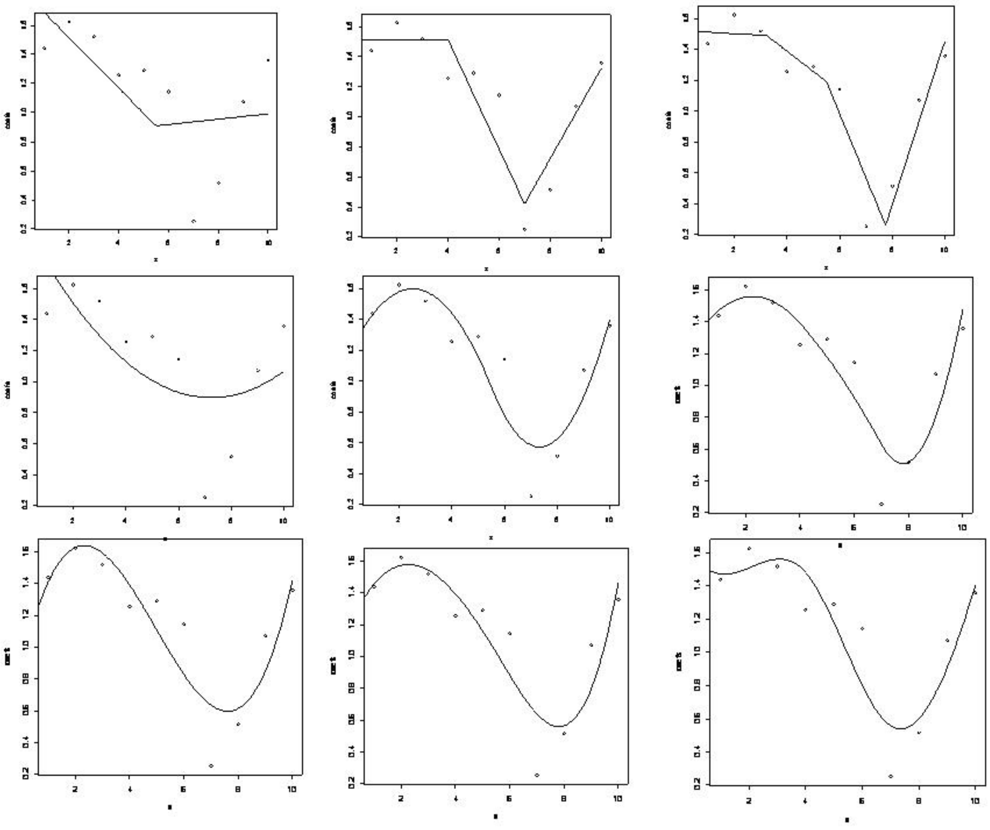


Figure 1.1: Graphs of B-splines of Different Degrees

1.4 LASSO for Logistic Regression

If we intend to find a model that predicts purchase behavior, we must also be able to confidently choose the appropriate explanatory variables for our model. In the case of a varying coefficient model we have a two-fold process for variable selection. We must first answer the question (1) does this variable belong in the model at all? and then (2) should the coefficient for this model be a constant or should it vary with R? To address these questions, we need an appropriate adaptation to the variable selection methods that are currently being used for variable selection. While we have limited our data to a subset of the variables to simplify our example, the data that we are dealing with contains over thirty demographic variables to choose from. In addition to large amount of data, we are expanding each variable by B-spline basis. The end result is over 100 of variables to choose from in our model selection. Since we are dealing with a large number of variables, we begin by looking at the popular LASSO penalty approach and its adaptations to logistic regression in order to find a solution to our problem. Tibshirani (1996) introduced the idea of LASSO as a method of model selection and shrinkage estimation for linear regression. To explain LASSO in the linear regression application, we consider a continuous response variable $Y \in R^n$, an n x p design matrix X and a parameter $\beta \in R^p$. Then the LASSO estimator is defined as:

$$\hat{\beta_\lambda} = \arg\min(\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{n} |\beta_j|). \tag{1.3}$$

Thus in the linear regression case, we minimize the least squares for the equation while adding in a penalty. As λ gets larger, the value of the coefficients, β_λ, shrink more towards zero with some elements actually being equal to zero. Thus, we are using the penalty factor to shrink that coefficients to the point that only a specific subset of coefficients are nonzero for the model, i.e; a specific subset of variables are chosen to be in the model. Note that for the LASSO penalty, the solution is not unique and varies for each λ. Thus we need to choose λ to find the optimal solution. Note that this l_1-type LASSO penalty can also be applied to logistic regression(Lokhorst (1999); Roth (2004); K and S. (2003); Genkin et al. (1997)). The LASSO penalty for logistic regression follows the same structure as that of linear regression in the sense that we have the likelihood function with an added penalty. In the case of logistic regression we use the penalized log-likelihood function. Recall the log-likelihood function for logistic regression for $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \beta^T x_i - log(1 + \exp(\beta^T x_i)]. \tag{1.4}$$

In order to ensure sparse solutions, we add a LASSO penalty to the log-likelihood:

$$\hat{\beta_\lambda} = \arg\min(\ell(\beta) + \lambda \sum_{j=1}^{n} |\beta_j|). \tag{1.5}$$

By using LASSO for logistic regression we are able to choose the variables that should be included in the model. However, that only provides a solution for half of our problem. We need a way to determine whether or not using a varying coefficient is appropriate for the variable in the model. For this problem we turn to the adaption of the group LASSO for logistic regression presented by Meier et al. (2007).

1.5 Group LASSO for Logistic Regression

Group LASSO (Yuan and Lin (2007); Bakin (1999); Cai (1997); Antoniadis and Fan (2001)) was first introduced as an extension of the LASSO for linear regression. The concept was originally developed to account for categorical variables. LASSO is limited in the sense that it can only choose whether a continuous variable or an individual dummy variable should be included in a model, not an entire factor. The group LASSO overcomes this problem by treating factor variables (multi-level categorical variables) as a group of variables in which you can conclude whether the entire group of variables should be included in the model or not. The group LASSO estimator for linear regression is defined as

$$\hat{\beta_\lambda} = \arg\min(\|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_{I_g}\|_2), \tag{1.6}$$

where g=$1, ..., G$ is an indicator of the group of variables and I_g is the index set belonging to the g^{th} group. Note that this solution is attractive not only because of its applications on grouped variables but also because it is invariant under groupwise orthogonal transformations. Meier et al.(2007) extended this concept to logistic regression.

Let (x_i, y_i) be *i.i.d.* observations $i = 1, ..., n$, of a p dimensional vector x_i with G grouped variables and a binary response variables $y_i \in \{0, 1\}$. Thus we rewrite x_i as a set of $x_{i,g} \in R^{df_g}$ vectors defined for each group g=$1, ..., G$ written as $x_i = (x_{i,1}^{'}, ..., x_{i,G}^{'})^{'}$. Note that the explanatory variables can be either categorical or continuous. For the case of a continuous variable, we treat that individual variable as its own group and have $df_g = 1$. The categorical variables have a $df_g = n_g - 1$, where n_g is the number of levels of the factor. Therefore if a factor has 5 levels, the $df_g = 4$. The logistic group estimator is given by the minimization of the convex function:

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{g=1}^{G} s(df_g)\|\beta_g\|_2, \tag{1.7}$$

where $s(\cdot)$ is a function of the degrees of freedom of each group. This function is used to scale down the penalty parameter for each group. In this model Meier et al. (2007) defines $s(df_g) = \sqrt{df_g}$ and $l(\cdot)$ as the penalized log-likelihood for logistic regression:

$$l(\beta) = \sum_{i=1}^{n}[y_i(\beta_0 + \sum_{g=1}^{G} x_{i,g}^T\beta_g) - log(1 + \exp(\beta_0 + \sum_{g=1}^{G} x_{i,g}^T\beta_g))]. \tag{1.8}$$

Just like in the standard LASSO procedure, λ is the tuning parameter that controls the amount of penalization for each group. Proposition 1 is the minimal condition needed for on the observed data in order for $S_\lambda(\cdot)$ to be minimized. Note that if the matrix X is of full rank then the solution will be unique, otherwise the solution will be a set of minimizers that corresponds to the same minimum of $S_\lambda(\cdot)$.

Proposition 1. *Assume that* $0 < \sum_{i=1}^{n} y_i < n$. *For* $\lambda > 0$ *and* $s(d) > 0$ *for all* d *in* N, *the minimum of* $S_\lambda(\beta)$ *is attained.*

Proposition 1 says for our dichotomous response variable we need to have at least 1 value to be different from the rest. When computing the groupwise l_2 - norm, we end up with a penalty that falls in between the penalty of a LASSO (l_1) and a ridge penalty function (l_2) (Meier et al. (2007)). The sparsity property of the group LASSO is one such that either $\hat{\beta}_g = 0$ or $\hat{\beta}_{g,j} \neq 0$ for all j in $(1, ..., df_g)$. In other words, for the group LASSO all of the coefficients of a group can be set to zero. Thus it doesn't choose the influential level of a factor, just whether or not an entire factor should be included in the model or not.

Let the n x df_g matrix X_g be the columns of the design matrix corresponding to the g^{th} predictor. If we assume that each X_g is a block group of full rank, we can compute a blockwise orthonormalization to obtain $X_g^T X_g = I_{df(g)}$. Thus before we minimize $S_\lambda(\cdot)$, we orthonormalize by group so that the different encoding schemes of the dummy variables do not affect the group LASSO estimator. Note that after the parameter estimation is complete, the coefficients must be transformed back into the original coding so that the coefficients may be interpreted correctly. Once the design matrix is set up as mentioned, we can begin our process of minimization.

Meier et al. (2007) propose two algorithms for parameter estimation for the group LASSO: Block Coordinate Descent and Block Coordinate Gradient descent. Block co-ordinate descent is a minimization procedure in which we cycle through the groups and minimize their objective function keeping all but the current group fixed. The first step of the algorithm is to check whether the minimizer of β_g is at 0. If it is, we set $\beta_g = 0$, and move onto the next group. If 0 is not the minimizing value, then we move to use a Newton-type algorithm in order to minimize β_g. The main drawback of this method is that the blockwise minimizations of the active groups must be performed numerically. This isn't too big of a drawback when we have small to moderate sized problems. However, for large scale applications this algorithm doesn't seem to be efficient. Thus, we turn our focus to Block Co-ordinate Gradient Descent method of Tseng and Yun (2007) which will be used in our application.

The key difference in the Block Co-ordinate Gradient Descent is the combination of a quadratic approximation of the log likelihood with an additional line search. In gradient descent the main idea is to take steps that are proportional to the negative gradient of the point in the current equation. This methodology is applied to the block co-ordinates for our groups and is the basis behind the package *grplasso* that will be discussed later (Meier et al. (2007)).

In Chapter 3 we introduce the adaptive group LASSO. The adaptive group LASSO follows the same computational methods of the group LASSO with an added weight change in order to achieve better asymptotic results. In the next chapter we introduce our model for predicting purchase behavior, develop our method for model selection, and apply it to our data. We then support the performance of our variable selection methodology with theory in Chapter 3 and through simulations in Chapter 4.

Chapter 2

Methodology

In Chapter 1 we introduced the idea of varying coefficients in logistic regression in order to capture relationships in longitudinal data that vary across some continuous variable. As previously mentioned, we start with the logistic regression formula and use a continuous random variable that has been splined in order to build a model with varying coefficients. Thus the model has the form:

$$\mu(x) = \frac{\exp\{\alpha + X'\beta(R)\}}{1 + \exp\{\alpha + X'\beta(R)\}} \tag{2.1}$$

where the function $\beta_j(R)$, j=1,...,p, will be taken to be B-splines. Here j denotes the different components of β. In this chapter we discuss the steps to build a varying coefficient model and perform model selection in order to determine not only what variables should be in the model but which coefficients should be varying and which should be constant.

2.1 Model Set-up and Model Selection

Before we begin the model selection process, we first must decide if our continuous variables is a candidate for a set of explanatory variables to vary across. Let R be a continuous random variable, $Y = (y_{i,1}, ..., y_{in_i})^T$, where $y_{i,j}$ denotes the dichotomous response of the ith subject at time t_{ij}, $i = 1, ...n$ and $j = 1, ..., m$, and X be a n x p matrix of p explanatory variables. We want to investigate whether the relationship between X and Y varies across different levels of R. We assume that the variable R is chosen partly based on the existing theory or understanding in marketing research and partly based on preliminary data analysis. A description of more concrete steps will be given in section §2.2.1. In essence we split the data up into subsets based on the values of R.

Once the data has been split into the different data sets, we can run logistic regression and record the coefficient for each variable of x_k of X. This coefficient will describe the relationship between Y and x_k at each level of R. We can plot out the coefficients at each level to see if the coefficient remains constant across all values of R or if they appear to be changing as the values of R change. If we notice that the coefficients of x_k appear to vary across R, then we will begin to begin building the model. However, if we are seeing no movement across the groups, it may not be likely that we will find any explanatory variables that have a relationship with Y that varies over R. Figure 2.1 gives an example of two variables, x_1 and x_2, that come from the same data set and each variable had a different conclusion. The dataset had been split into four subsets: A, B, C, and D based off a variable R as labeled across the x-axis. For each of these data sets, we ran the logistic regression of x_1 regressed on some response variable y and graphed the coefficient (top graph) and x_2 regressed on some response variable y and graphed the coefficient (bottom graph). We can see that across the different levels of R, the coefficients of x_1 are relatively constant while the coefficients of x_2 seem to move around a bit. Thus we would expect similar trends when we run our varying coefficient model.

Note that these preliminary investigations are not a means of assessing whether a model should have varying coefficients, but simply an initial look at the relationship that could exist between each x_k, Y and R. Once we have completed our initial data investigation, we are ready to build the model for varying coefficients.

2.1.1 B-Splines

As previously mentioned, there are many ways in which one can define the function $\beta(R)$ in varying coefficient models. In our methodology, we have chosen to use B-splines as our function. In order to produce the B-splines needed to create our varying coefficient X-matrix, we use the R package *splines* (Brown (1992)). Within this package, we are specifically interested in the function *bs()*. The default function has the following arguments:

bs(R, df = NULL, knots = NULL, degree = 3, intercept = FALSE, Boundary.knots = range(x)).(2.2)

This function generates a basis matrix for representing the family of piecewise polynomials with the specified interior knots and degree, evaluated at the values of R, our candidate variable for creating varying coefficients. The knots can be specifically chosen to be at predetermined points or determined adaptively. This default function always places knots at the boundary points of the variable,defined as the range of R.

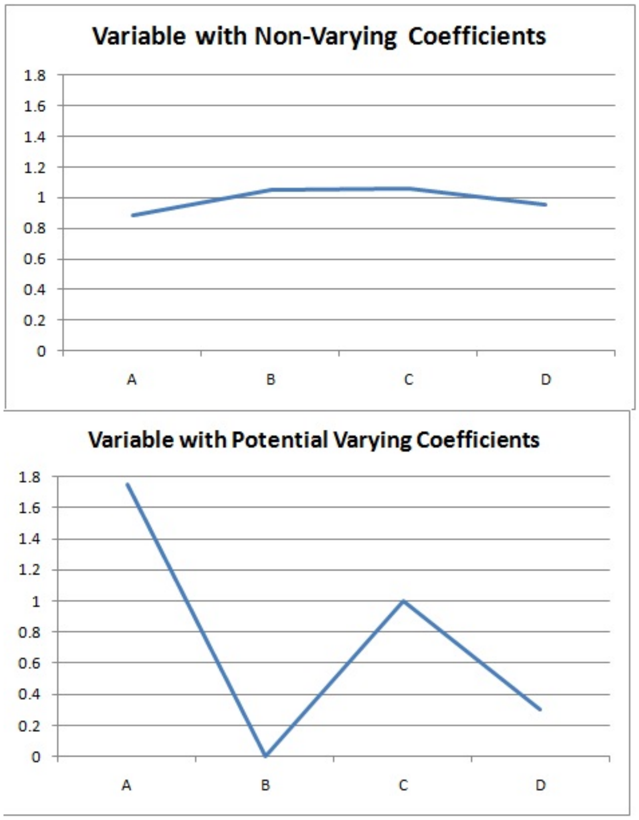


Figure 2.1: Comparing Constant Coefficients Across R to Varying Coefficients Across R. The graphs show two situations that could result from investigating the relationship between R and a coefficient. In each graph we plot the coefficient from a logistic regression of a variable x_i regressed on a response variable y for each of the four subsets of data based on values of R. In the top graph, we see little movement between the values of the coefficients, implying a constant coefficient across R. In the bottom graph we see that the coefficients move across R, implying there could be some varying coefficient based on R.

The default degree of the spline is 3, which produces cubic splines. If one wanted piecewise linear splines the degree would be 1, for quadratic the degree would be 2 and so on.

In our current methodology we will use fixed knots. The number and location of these knots can be determined through preliminary examination of the data. Once we have determined the number and placement of our knots, we are able to use the *bs()* function to produce our basis matrix. Let X_{bs} be a matrix with s columns (b_1, b_2, b_3, $\cdots$,b$_s$) where each column is a B-spline basis representing each of the s knots. Recall that s is actually equal to number of knots specified in the function plus the boundary knots if they were not removed from the default. We then use this basis matrix to expand the design matrix X. Recall our matrix of explanatory variables X that is equal to

$$\begin{pmatrix} \mathrm{x}_1 & \mathrm{x}_2 & \mathrm{x}_3 & \cdots & \mathrm{x}_p \end{pmatrix}$$

where p is the number of parameters in the full model. We then expand X by multiplying each column by the values in our basis matrix. Thus we have($x_i * b_1$, $x_i * x_2$, $x_i * b_3$, $\cdots$,x$_i * b_s$). Note that each of the p predictor variables and the intercept is expanded in this manner. Thus, our resulting design matrix for the full model is a n by $(p+1)*s$ matrix of the form:

$$\begin{pmatrix} 1^*\mathrm{b}_1, \cdots ,1^*\mathrm{b}_s & \mathrm{x}_1 * b_1,\cdots ,\mathrm{x}_1 * b_s & \mathrm{x}_2 * b_1, \cdots ,\mathrm{x}_2 * b_s & \cdots & \mathrm{x}_p * b_1,\cdots ,\mathrm{x}_p * b_s, \end{pmatrix}$$

where $(1 * b_1, ..., 1 * b_s)$ corresponds to the intercept and (x_1,...,x_p) are the p explanatory variables in the full model. By multiplying each variable by each of the s basis splines, we have expanded the matrix to account for the different levels of R. Now that we have our full model defined, we are ready to begin model selection.

2.1.2 Variable Selection

Variable selection in this model is performed in two stages. First, we decide whether a variable should be included in the model. Second, once we decide that a variable is included in the model, we answer the additional question of whether the coefficient of the variable varies over R. In our methodology we treat each x_k variable that has been expanded by the B-spline basis as a group. Thus, the intercept $(b_1, b_2, b_3, \cdots, b_s)$ is group 1, $(x_1 * b_1, x_1 * b_2, x_1 * b_3, \cdots, x_1 * b_s)$ is group 2, and so on until all g=$p+1$ groups are defined. Since we choose to look at the splined variables as grouped data, we choose to use the group LASSO penalty as our basis for model selection. In the group LASSO, one must define which variables are to be treated as a group. Then each group is penalized, and is either entirely chosen to be included in the model or all the coefficients are set to zero. Recall the group LASSO penalty:

$$l(\beta) = \sum_{i=1}^{n}[y_i(\beta_0 + \sum_{g=1}^{G} x_{i,g}^T\beta_g) - log(1 + \exp(\beta_0 + \sum_{g=1}^{G} x_{i,g}^T\beta_g))], \tag{2.3}$$

where $s(df_g) = \sqrt{df_g}$ and $l(\cdot)$ as the penalized log-likelihood for logistic regression. To run the group LASSO for logistic regression on our data set, we turn to the R package *grplasso* (Meier (2009)). In the *grplasso* package we need to define our X-matrix, our response variable, a vector of indices which defines how the variables should be grouped, and a set of λ's used for penalization. Below are the main components of the R function *grplasso()*:

$$grplasso(x, y, model = LogReg(), lambda = lambda, index = index). \tag{2.4}$$

Here x is our splined X-matrix, y is our response variable Y and *lambda* is the vector of values defined by users that will be used as the penalty in group LASSO. In the function we also need to define an index vector. The index is used to indicate which variables in the X-matrix should be treated as one group. Thus, the length of the index vector should be equal to the number of columns in X. In our example we would have a different number to represent each of the $p + 1$ parameters that would appear s times in the vector of each of the splines. As a result of *grplasso*() we have a model for each value of λ. We then need a model selection criteria in order to determine which model is the best.

Model selection criteria are sensitive to certain conditions such as correlation, number of measures recorded for each subject, and sample size that are associated with longitudinal data. While common criteria such as the Akaike information criteria (AIC) and the Bayesian information criteria (BIC) can be used for longitudinal data, adaptations to these criteria such as the corrected Akaike information criteria (AICc) and the residual information criteria (RIC) may perform better in situations with smaller samples and higher correlated data (Azari et al. (2005)). The AICc can be defined as

$$AICc = -2 * \ell(\beta) + 2(\frac{n}{n-p-2})(p+1). \tag{2.5}$$

where n is the sample size and $\ell(\beta)$ is the log likelihood. Note that as $n \to \infty$ and p is held fixed, the AICc approaches the AIC which is defined as

$$AIC = -2 * \ell(\beta) + 2(p+1). \tag{2.6}$$

The AICc outperforms the AIC in small samples and does comparably well in larger samples. The RIC can be defined as

$$RIC = -2 * \ell(\beta) + p\log(n) + \frac{(n-p)^2}{n-p-2}. \tag{2.7}$$

which approximates to the BIC as $n \to \infty$:

$$BIC = -2 * \ell(\beta) + p\log(n). \tag{2.8}$$

Through (Azari et al. (2005)), the following conclusions were made:

- AIC performs worse than AICc, BIC, and RIC in all simulated correlated samples of longitudinal data.
- BIC and RIC are the best and perform at a similar efficiency except when SNR increases, RIC does better.
- AICc performs better than AIC when the number of measures on a subject is small, but they perform similarly as n increases.
- RIC outperforms all in longitudinal cases except when there is little correlations, then performs the same as BIC.

Thus the appropriate model selection criteria can be chosen once we know something about the correlation, the number of subjects, and the number of measurements per subject in the data. Once the model selection criteria has been chosen, we work through the following stages to choose a model.

2.1.3 Preprocessing: Adaptive Group LASSO

In the next chapter we will discuss the fact that group LASSO alone does not posses the oracle properties so we must use the adaptive group LASSO. There is not currently any *R* packages for adaptive group LASSO, so we add an additional pre-processing step before we begin group LASSO on our data. The adaptive group LASSO consists of running the group LASSO for one value of λ and using the coefficients to produce weights. The following is the penalty term of the adaptive group LASSO that will be used in all of the theory to follow:

$$\sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_j\|} \|\beta_{(j)}\|, \tag{2.9}$$

where $\tilde{\beta}_j$ is a known consistent estimate of β. In the initial run, in order to keep all of the variables present, we use $\lambda = 1$. We pull the coefficients, $\tilde{\beta}_j$ from the group LASSO estimate and then use the vector of coefficients as a weight on our initial X-matrix. Thus we replace each X_i,$i = 1...p + 1$, with the weighted W_i defined as

$$W_i = \frac{1}{\tilde{\beta}_j} * X_i. \tag{2.10}$$

We build the new X-matrix in the same way defined above. Thus we define the new design-matrix W as

$$\left(\mathrm{w}_0 * b_1, ..., w_0 * b_s \quad \mathrm{w}_1 * w_1, ..., w_1 * b_s \quad \mathrm{w}_2 * b_1, ..., w_2 * b_s \quad ... \quad \mathrm{w}_{p+1} * b_1, ..., w_{p+1} * b_s \right)$$

By setting up our X-matrix in this manner, we obtain the structure of the adaptive group LASSO that can be used in the existing group LASSO package.

2.1.4 Stage 1: Variable Selection

Once we have defined our new weighted matrix, W, we are ready to begin stage 1 of our model selection. In stage 1, we start with the full model containing all g groups:

$$\left(\mathrm{w}_0 * b_1, ..., w_0 * b_s \quad \mathrm{w}_1 * w_1, ..., w_1 * b_s \quad \mathrm{w}_2 * b_1, ..., w_2 * b_s \quad ... \quad \mathrm{w}_{p+1} * b_1, ..., w_{p+1} * b_s \right)$$

and run through the group LASSO for a sequence of λ. This results in a model for each of the λ. To choose a model, we run logistic regression for each value of λ and compute the preferred model selection criteria. In the end we choose the model that performs the best based on the model selection criteria. From the model in stage 1, we know which variables should be in the model. We now move onto stage 2 of the model selection methodology to determine which variables should have varying coefficients and which variables will have constant coefficients.

2.1.5 Stage 2: Varying or Constant Coefficients

In the next stage we determine if the coefficient for each variable should be varying or constant. For this step we treat the intercept the same as in stage 1 but we treat each of the remaining predictor variables as two distinct groups. The first group contains only 1 variable and is the original variable alone (without any interaction with the B-spline basis). The second group is variable multiplied by each of the first s-1 spline bases. (Note there are originally s splines, but to avoid over-fitting we remove one).

Thus for each predictor variable X that was chosen in stage 1, we have the expanded weighted design matrix (x_i, $w_i * b_1$, $w_i * b_2$, $\cdots$,w$_i * b_{s-1}$) where x_i is considered one group and ($w_i * b_1$, $w_i * b_2$, $\cdots$,w$_i * b_{s-1}$) is a second group. In this step we do not penalize the first group since we decided in stage 1 that it should be included in the model. However, we do penalize the second group where each variable is multiplied by splines. If the coefficients of the group of splined variables does not shrink to zero, then we conclude that the coefficient of the variable should be varying. Likewise, if the coefficients shrink to zero then they will not be varying in the model, but instead they will have a constant coefficient that comes from including the original variable alone. We take the same approach as in stage 1 and run through the group LASSO for a set of λ and record the value of the model selection criteria. The model for which the criteria points to the final model and model selection is complete.

2.1.6 Post Processing: Check Variables for Significance

Once stage 2 of the variable selection is complete, we have one final step to determine the final model. In this step, we run logistic regression using the model produced through the two stage variable selection. We use backward selection to remove one variable at a time that is not statistically significant at the $\alpha = .05$ level. For the variables that were determined to have varying coefficients, at least one of the five elements of each group must be statistically significant at the $\alpha = .05$. level. In the next section we work through our methodology with some real world data.

2.2 Application to Marketing Research Data

Before we begin our analysis we describe the specific dataset that we use. For this part of the analysis we subset the panelist data to a set of panelists from Pittsfield, MA. More specifically, we choose a set of panelists who have shopped at a specific grocery store in the year 2005. We then limit this group of panelist to those who also have panel data from 2004 so that we have history on these panelists. In addition to subsetting to panelist activity at one store, we also decided to limit our purchase data to a specific consumer

Table 2.1: Variables Included in Data

Variable	Definition	Value
Purchase (y)	The product was purchased by the panelist	Yes=1,No=0
Prev Purchase (z)	The number of weeks the product was purchased the previous year	Yes=1,No=0
Display (x_1)	The product was on display in the store this week.	Yes=1,No=0
Price (x_2)	The product had a price reduction of 5% or more this week.	Yes=1,No=0
Income1 (x_3)	Household income is $0-$19,000.	Yes=1,No=0
Income2 (x_4)	Household income is $20,000-$34,999.	Yes=1,No=0
Income3 (x_5)	Household income is $35,000-$64,999.	Yes=1,No=0
Income4 (NA)	Household income is $65,000 +	Yes=1,No=0
Hispanic (x_6)	Household is Hispanic.	Yes=1,No=0
Owner (x_7)	Residence is owned not rented	Yes=1,No=0
Family1 (x_8)	Size of family is 1.	Yes=1,No=0
Family2 (x_9)	Size of family is 2 or less	Yes=1,No=0
Family3 (x_{10})	Size of family is 3 or less	Yes=1,No=0
Family4 (x_{11})	Size of family is 4 or less.	Yes=1,No=0
Family5 (x_{12})	Size of family is 5 or less.	Yes=1,No=0
Family6 (NA)	Size of family is 6+.	Yes=1,No=0
Single (x_{13})	Head of Household is single.	Yes=1,No=0
Married (x_{14})	Head of Household is married.	Yes=1,No=0
NewSingle (NA)	Head of Household is divorced, separated, or widowed.	Yes=1,No=0

packaged good from the salty snack category that was purchased frequently by panelists. Note that we do not remove panelists that did not purchase the product because we need to analyze their behavior as well.

We keep the 2005 data at the weekly level and create a dummy variable indicating whether this product is purchased on the given week by the given panelists for each panelist for each of the 52 weeks of data. Thus, each panelist has a data entry of 1 (purchase made) or 0 (no purchase made) for each of the 52 weeks that are in the 2005 data. Note that we are only interested if a purchase was made or not, we do not take into account the quantity of the product that was purchased. We also need to develop a variable that represents the previous purchase behavior of this product for these panelists in the prior year. We summarize this behavior as the count of the weeks in the previous year in which the product was purchased by a panelist. This count ranged from 0 purchases in the previous year to 31 weeks in which the product was purchased. Once again this is simply a count of weeks the item was purchased not the quantity that was purchased.

Finally, panel demographics and marketing variables were attached to the data. The panel demographics received by SymphonyIRI were reorganized and translated from categorical variables to dummy variables. For example, marital status (1 for single, 2 for married, 3 for separated, 4 for divorce and 5 for widow) is represented by three different dummy variables for married, single, or newsingle which includes all divorced, separated and widowed head of households. Table 2.1 contains the variables that are used in the investigation.

We now apply our methodology to our marketing data. In the varying coefficient logistic model:

$$\mu(x) = \frac{\exp\{\alpha + X'\beta(z)\}}{1 + \exp\{\alpha + X'\beta(z)\}}, \tag{2.11}$$

we define z as the number of previous purchases that a panelist has made and X is a matrix containing the explanatory marketing and demographic variables defined above. Thus, varying coefficients will be based on the number of purchases in the past year. These variables can either be removed from the model, be included in the model and have constant coefficients, or be included in the model and have varying coefficients. In the following subsections we obtain our B-spline basis, describe our process for model selection, and present our results.

2.2.1 Calculating B-Splines for Varying Coefficients

Before we build our model, we do a preliminary investigation of the data. In order to asses this relationship, we subset the data based on the values of z. The breaks in the data will be dependent on the distribution of the sample across z. For example if z would be the variable time expressed as days of the year, the data could be split into the measurements taken each month, or week or quarter. If z is a measure of frequency, the groups can be divided into quartiles, into high medium low values, or any other logical split based on the data. The analysis conducted to use the varying coefficient model began as an exploratory look at the relationship between the previous number of purchases and all of the other explanatory variables. The initial step was to break our panelists into four groups based on their number of purchases and to run individual logistic regressions on each group for each explanatory variable as a predictor of purchasing the product. The goal of this exercise was to see if the explanatory variables affected each group differently. Thus, we would be able to see if the coefficients of the explanatory variables varied across the groups.

The panelists were split up based on their previous purchase behavior: (A) those didn't purchase the product in the previous year, (B) those who purchased the product one week in the previous year, (C) those who purchased the product two, three, or four weeks in the previous year, and (D) those that purchased the product 5 to 15 weeks in the past year. A different model was run for each of the four groups for each of the explanatory variables. The coefficient was stripped from the model and the odds ratio was calculated and plotted out for each group to see what the overall pattern was across groups; See Figure 2.2. Since clear variation was found in the graphs, we moved onto finding what the actual functions were by way of B-splines.

In order to calculate the splines in the data we used the R package *splines* and the function $bs()$. We chose our knots based on the probability of making a purchase at each level of previous purchase. In order to do this the panelists were grouped into 10 groups:

1. Those who did not purchase the product in the previous year.

2. Those who bought the product 1 time in the previous year.

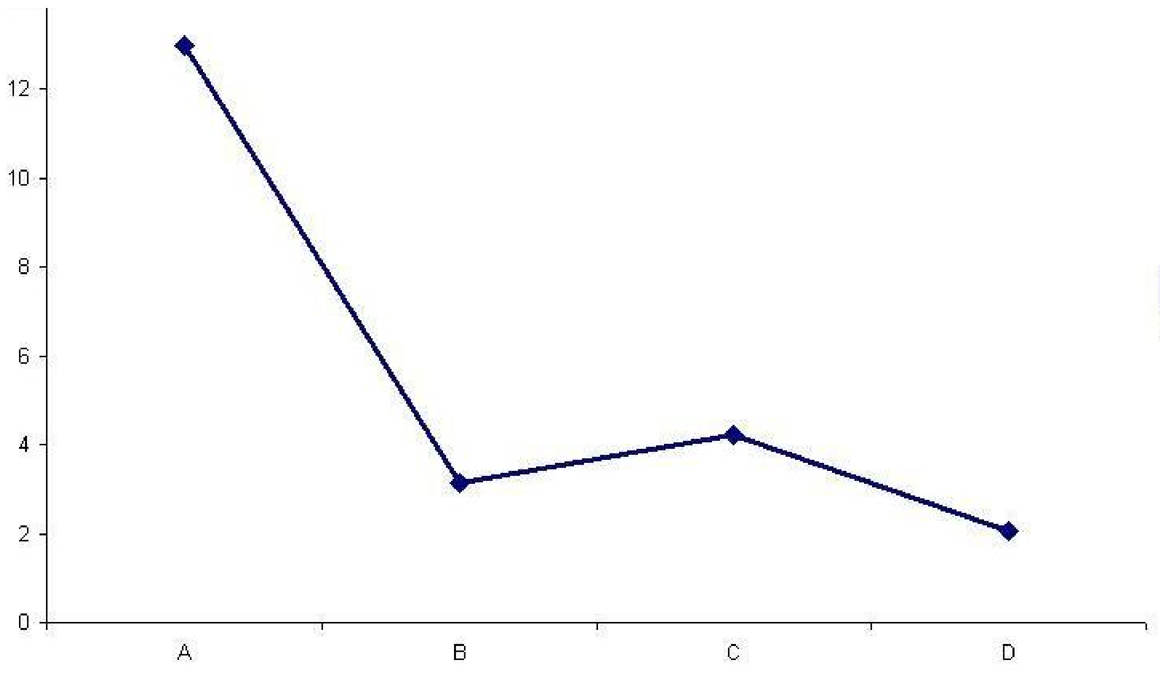


Figure 2.2: The Coefficients for Display modeled across different levels of previous purchases

3. Those who bought the product 2 times in the previous year.
4. Those who bought the product 3 times in the previous year.
5. Those who bought the product 4 times in the previous year.
6. Those who bought the product 5 times in the previous year.
7. Those who bought the product 6 times in the previous year.
8. Those who bought the product 7 times in the previous year.
9. Those who bought the product 8 times in the previous year.
10. Those who bought the product 9 or more times in the previous year.

We ran logistic regression using the different levels and plotted the probability of making a purchase for each group. The knots were chosen to be points where there appeared to be a change in the curve. Figure 2.3 is the plot of the coefficients for the variable *display* across these 10 groups. From this graph, we decided to put knots at z=(0, 3, 7). Once the knots were chosen, we use the $bs()$ function to create our basis matrix. The basis matrix is computed using the variable purchase. Thus we are able to form a set of spline basis functions that represent the effect of the previous years purchase.

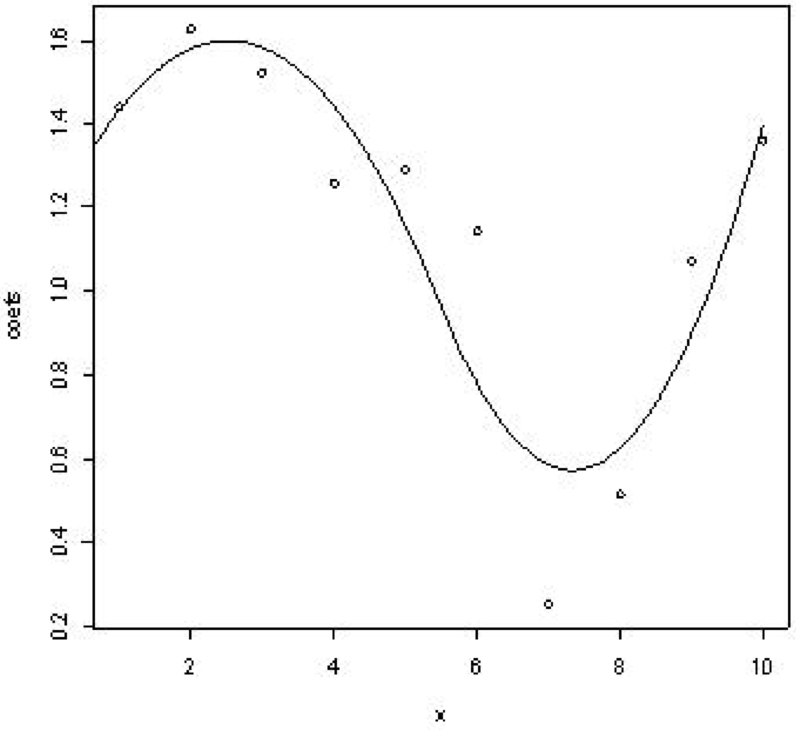


Figure 2.3: Graph of Coefficients for Display across 10 Groups

2.2.2 Variable Selection Methodology

Stage 1. We begin stage 1 of model selection by first building the full model. The full model begins with every variable in our data set multiplied by our fixed-knot B-spline matrix. As previously mentioned we will be computing our spline basis off the variable *previous purchase* at the fixed knots: 0, 3, and 7. In addition to these three knots, the *bspline* package also includes boundary knots at the minimum and the maximum values of the variable that has been splined. In our case, the previous purchase of our panel ran from 0 (never purchasing the product in the previous year) to 15 (purchasing the product 15 weeks in the previous year). Thus, for the B-spline matrix, we have five columns that correspond to each of the basis. We now use this matrix, in order to create the matrix for our full model.

Consider a predictor variable x, which could be any of the fourteen explanatory variables we have in our full model. Note that Income4, Family6, and Newsingle, are not in the full model to avoid overfitting. Let X_{bs} be a matrix of five columns (b_1, b_2, b_3, b_4, b_5) where each column is a B-spline basis. For example, the expanded design matrix for *display* is ($display * b_1$, $display * b_2$, $display * b_3$, $display * b_4$, $display * b_5$). Note that each of the fourteen predictor variables and the intercept is expanded in this manner. Thus, our resulting design matrix for the full model is n by 15*5:

$$\begin{pmatrix} 1\text{*b}_1, ..., 1 * b_5 & \text{x}_1 * b_1, ..., x_1 * b_5 & \text{x}_2 * b_1, ..., x_2 * b_5 & \text{x}_3 * b_1, ..., x_3 * b_5 & ... & \text{x}_{14} * b_1, ..., x_{14} * b_5 \end{pmatrix}$$

where $(1 * b_1, ..., 1 * b_5)$ is the intercept and (x_1,...,x_{14}) are the variables listed in Table 2.1.

Now that we have our full model defined, we need to begin our process for group LASSO for logistic regression. In stage 1, we treat the x variable that has been expanded by the B-spline basis as a group. Thus, the intercept $(b_1, b_2, b_3, b_4, b_5)$ is group 1, $(x_1 * b_1, x_1 * b_2, x_1 * b_3, x_1 * b_4, x_1 * b_5)$ is group 2, and so on until all fifteen groups are defined.

Recall that the pre-processing step is to run one round of *grplasso*() with λ=1 so that we compute weights without heavy penalization. After our initial run of the group LASSO, we take the coefficients and use them as weights. Once we have defined our new weighted matrix, we are ready to begin stage 1 of our model selection. In stage 1, we start with the full weighted model containing 15 groups, and run through the group LASSO for a sequence of $\lambda = (1, 10, 20, 30, ..., 290, 300)$. This results in a model for each of the λ. To choose a model, we run logistic regression for each value of λ and compute the BIC. The results are given in Table 2.2. Based on this table, we choose the model at which λ=180 and BIC=7127.255. Here BIC is

Table 2.2: Table of BIC vs. Lambda in Step 1 of Model Selection

Lambda	Number Nonzero Coefficients	BIC	Variable Dropped
1	75	7415.937	None
10	75	7415.937	None
20	65	7370.276	Family5, Married
30	55	7335.682	Income2, Single
40	50	7293.49	Income3
50	45	7255.205	Hispanic
60	40	7222.978	Family1
70	35	7184.382	Price
80	35	7184.382	None
90	35	7184.382	None
100	35	7184.382	None
110	35	7184.382	None
120	35	7184.382	None
130	35	7184.382	None
140	30	7154.026	Income1
150	25	7127.255	Family2
160	25	7127.255	None
170	25	7127.255	None
180	25	7127.255	None
190	20	7397.463	Display
200	20	7397.463	None
210	20	7397.463	None
220	20	7397.463	None
230	20	7397.463	None
240	20	7397.463	None
250	20	7397.463	None
260	20	7397.463	None
270	15	7356.22	Family4
280	10	7388.128	Family3
290	10	7388.128	None
300	10	7388.128	None

at its minimum. Thus, we have determined that model will include the Intercept, Display, Owner, Family3, and Family4.

Stage 2. In the next stage we determine if the coefficient for each variable should be varying or constant. For this step we treat the intercept the same as in stage 1 but we treat each of the remaining 4 predictor variables as two distinct groups. The first group contains only 1 variable and is the original variable alone (without any interaction with the B-spline basis). The second group is variable multiplied by each of the first 4 spline bases. (Note there are originally 5 splines, but to avoid over-fitting we remove one). If the coefficients of the group of splined variables does not shrink to zero, then we conclude that the coefficient of the variable should be varying. Likewise, if the coefficients shrink to zero then they will not be varying in the model, but instead they will have a constant coefficient that comes from including the original variable alone. We take the same approach as in stage 1 and run through the group LASSO for a set of λ and record the BIC. Table 2.3 contains the results of stage 2 of the model selection. Based on the table we choose the model where $\lambda = 190$ and the BIC=7225.71. We now have the final model where variables family3, family4 and owner have constant coefficients, while display has a varying coefficient.

Once stage 2 of the variable selection is complete, we have one final stage to determine the final model. In this step, we run logistic regression using the model produced through the two stage variable selection and throw away any variables that are not significant. Once we removed these variables, we found the final model to include the variables: [intercept, display multiplied by the B-spline basis, family3]. These variables can be used to predict the probability that a panelist will make a purchase.

2.2.3 Examination of Possible Random Effects

A major assumption we made in this model selection was that the the weekly purchases are independent given z. If this assumption is severely violated, the use of the BIC criterion becomes questionable. To check for correlation we consider a model with random effects. To account for this dependence in the data, we added a random effect to our model to represent each panelist. This addition to our varying coefficient model will account for the dependence between purchases and variation that can be explained by panelist. To incorporate the random effect into our model, we decided to expand our design matrix in a way similar to what is proposed in Ibrahim et al. (2011). This consisted of a matrix that had a column for each panelist in which 52 of the rows contained an indicator function of 1 each indicating a week of the panelist's purchases and the rest of the rows contained 0's. By building this matrix, we were able to append it to our design matrix and treat the panelists as another grouped variable. Thus we were able to proceed with the *grplasso*() function in R on our new expanded matrix.

Once the new matrix was built, we began running through the group LASSO across a series of λ's to determine if and when the grouped variable of the panelist indicators would fall out of the model. Recall that in our process we found λ=180 to produce the model with the lowest BIC. With these large values of λ in mind, we decided to start with λ=1 and work our way up until the panelists indicators dropped from the model. In the end, the panelist indicators dropped from the model at λ=25. We computed the BIC for the models leading up to this point and for the models right after this point. We see a large jump at the point λ=25 where the individual effect parameters falls out. Based on these results, we concluded that we did not need to add a random effect in the model. Whatever dependence between purchases and variation that comes from the panelists appears to be captured given the past purchase behavior z.

Since we found no evidence of week to week correlation, we conclude that we do not need to adjust our

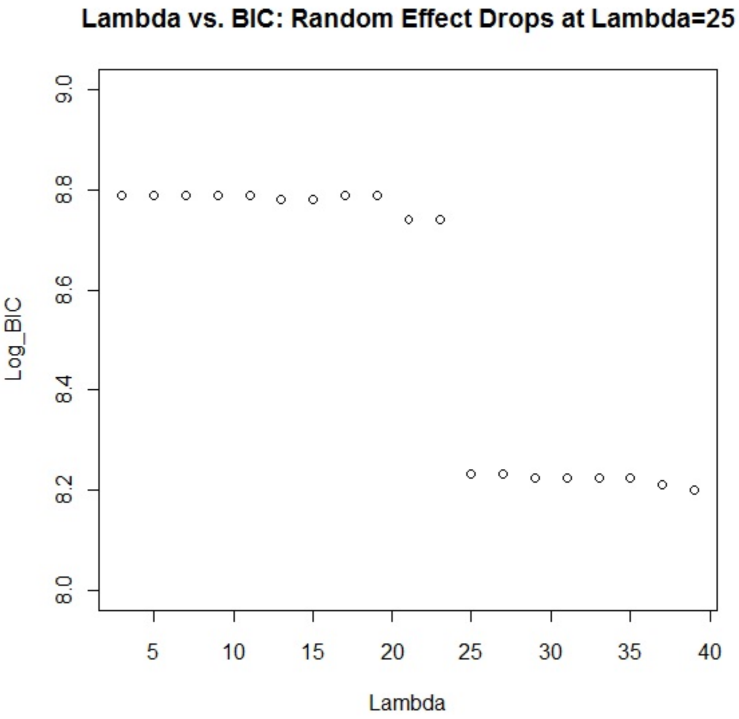


Figure 2.4: BIC vs. Lambda for Model with Random Effect

model selection criteria. However, if there was significance in the random effect component, we would need to look at the adjusted model selection criteria such as the AICc or the RIC mentioned in §2.1.

2.2.4 Final Model and Interpretation

Through our model setup and variable selection methodology, we have chosen a varying coefficient model for our dataset. Table 2.4 contains the coefficients of the final model. We now wish to assess the model that we have chosen. In our model display is a special case in which the probability of making a purchase is a function of the number of previous purchases that a panelist has made. Because display has a varying coefficient, we now wish to graph the function $\beta(z)$. By graphing the function we are able to understand the relationship between previous purchases and the probability of making a purchase when the variable at question is present. In other words, since we see that display should have a varying coefficient, we know that a customer will react differently to a product when it is on display based on the number of times they have previously purchased the product.

We start with the curve in Figure 2.5 which contains the coefficient curve for the variable display. In this curve we see that if a consumer has never bought a product before they are likely to try the product when it is on display. In marketing research this is called the trial period. As we move along the curve, we see that people who have tried the product before a small amount of times are less likely to be affected by the display. These consumers may have tried the product, but didn't develop a loyalty to the product. Thus they are not as affected by the product on display. Then, as we move to more loyal consumers of the product, we see that they enjoy the product and when it is on display they will more likely purchase the product. Finally as we move to the most loyal customers we see that the effect of the display begins to decrease. Note that this does not mean that they are less likely to purchase the product, they are just less likely to purchase the product more when it is on display. This can be explained by the fact that the most loyal customers will purchase the product regardless of whether it is on display or not. The second element of the model is family3, which indicates a family of size 3 or less. We see that this coefficient has a negative coefficient suggesting that larger families are more likely to purchase this salty snack. Now that we have chosen our model, we assess the performance.

2.2.5 Model Prediction Assessment

After choosing the model, we wish to test the predictive power of our model. We found a model that fits the year of data that we chose, but how does that model work on predicting other years of data? In order to analyze the performance of the model, we pulled a year of data for 2006 and replicated the original data

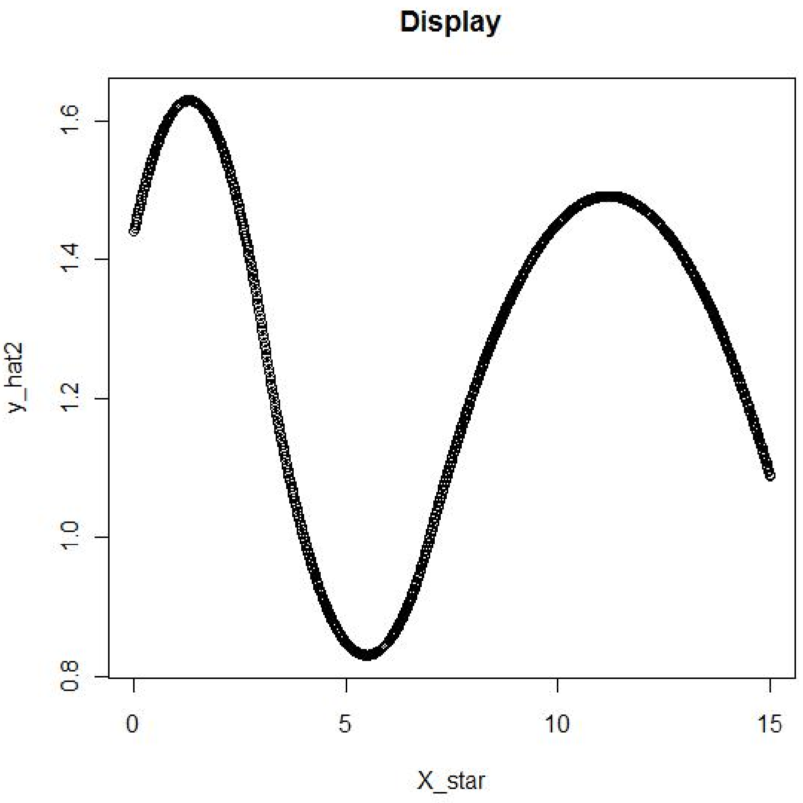


Figure 2.5: Varying Coefficient of Display

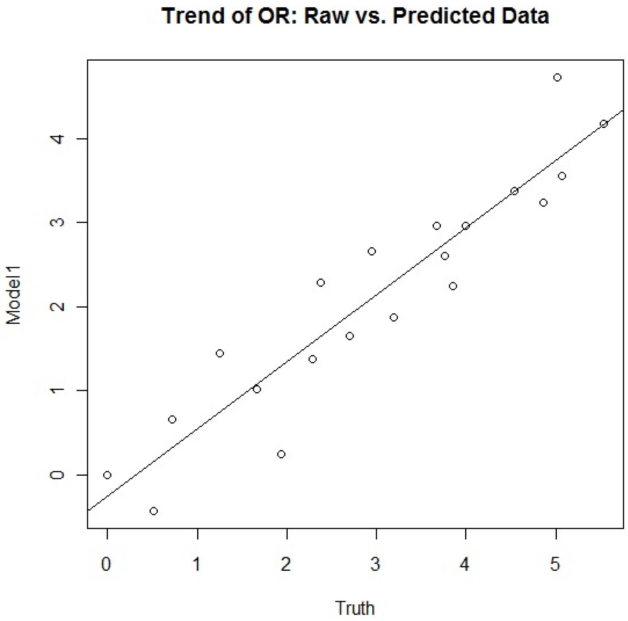


Figure 2.6: Trend of Modeled vs. Raw Data. This graph shows the relationship between the OR of making a purchase for a set X and z for the raw data vs. the OR for the modeled data.

set. Thus, the previous year of panel purchases were based on the 2005 data. For each week of the new 2006 data, we use our model to come up with the probability of a specific panelist making a purchase on a given week. We then take these probabilities and compare them to what the results would be if we just took the average purchases. This will tell us if our model provides any more information.

In our model, we found that family size and display in relation to the number of previous purchases played a role in predicting the probability of making a purchase. Thus, we begin by separating the panelists into groups based on their previous purchases, the size of their family, and whether a display was present in the given week. For previous purchase, we create buckets based on the curve for display from our model. We ended up with five values for previous purchase: 0 purchases, 1 purchase, 2-5 purchases, 5-12 purchases, and 13 or more purchases. Thus we ended up looking at predicted purchase vs. average purchase from the raw data for each of the 20 situations.

The effect of being in each of the groups on purchase behavior was captured by computing the odds ratio against the most basic situation: no previous purchase, no display, small family size. After computing the odds ratio for the modeled points and the raw data, we graph them to see if they follow the same trend. Graph 2.6 shows that our model is capturing the trend of the raw data. Thus, when we use the model to predict a new year of data, we find that our results validate to a reasonable extent.

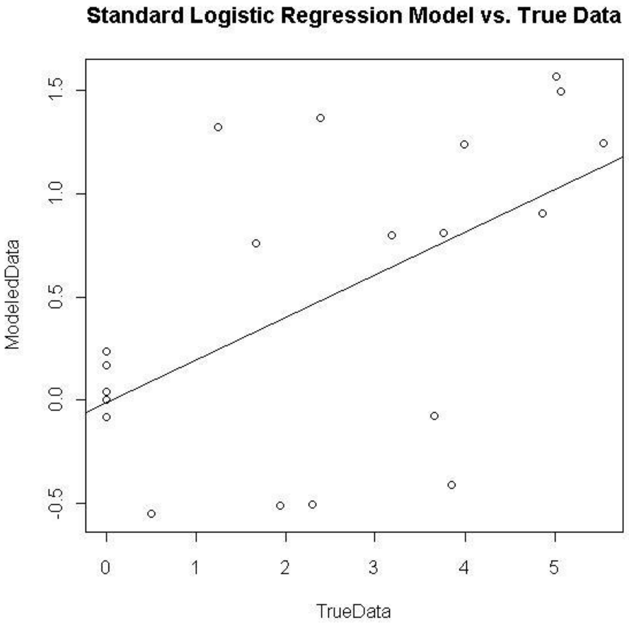


Figure 2.7: Trend of Modeled vs. Raw Data. This graph shows the relationship between the OR of making a purchase for a set X and z for the raw data vs. the OR for the modeled standard regression data.

We now wish to compare the predictive power of our model with a standard logistic regression model. By going through the model selection process, we showed that varying coefficients were chosen over constant coefficients to better model the data at hand. In order to show that the varying coefficient model outperforms the standard logistic regression model, we must also show our model out predicts the model chosen for non varying coefficients.

Therefore, we start again with the original full data set and include all the original variables in Table 2.1 aside from the variable "Prev Purchase". We put all of the variables into the logistic regression and use stepwise model selection in order to pick the best model. Through model selection, the chosen model included *display, income1, single, married, owner, family2, family3, family4,* and *family5.*

Using this model, we now attempt to predict the 2006 data. We run through the same exercise as we did with our varying coefficient model in which we split the data into the 20 situations and compare the OR's of the actual and modeled data. Figure 2.7 shows the results of the standard logistic regression model. We can see that the varying coefficient model in Figure 2.6 does a much better job of following the true trends in the data. The fit of the data in Figure 2.6 is $R^2 = 0.86$, while Figure 2.7 has an R^2= 0.30. Based on these graphs, we conclude that our varying coefficient model outperforms the standard logistic regression model in terms of prediction.

2.3 Modifications to the Model

In the current model, previous purchases were the number of weeks the product was purchased in the previous static calendar year. While this does provide an idea of the frequency of the purchases made by each consumer, we are missing the element of recency. By using a static year, every week the panelist is being compared to the same static 52 weeks. Thus, week 53 is predicted based on purchases in weeks 1-52 the same way that week 100 is predicted based on what was purchased week 1-52. In some situations, this static year may be a good indication of one's behavior. However, in the grocery industry, specifically with a product that comes from heavily consumed category like salty snacks, we would want to look at more recent weeks of purchase information.

Therefore, we decided to include rolling 52 week periods to depict the previous purchase behavior of each panelist. We reconstructed our data so that for every panelist at each of the 52 weeks, we had the number of weeks for which the product was purchased over the past 52 weeks. Thus week 53 was based on the previous purchase of week 1-52, but now week 100 was based off of the purchases that occurred in weeks 48-99. We ran through the process of creating splines based on the new variable for previous purchase, building the model, and running through the two stages of model selection. In the end, we found that the best model was found at λ=180, display was varying, and family size of 4 or less, were in the final model. Thus, the static 52 weeks captured the same trends in the data.

A second modification to the model was to investigate a transition model similar to Albert and Follmann (2003) in which we would use one model if we knew that a purchase was made recently and a second different model if we knew that no purchase had been made in that time frame. This model also aims to bring in the element of recency of purchase into the model. This variable of recent purchase is not what we wish to vary across, currently prior year, but instead an indicator that says whether or not the panelist has recently made a purchase. Knowing that someone recently made a purchase could effect the probability of buying the product in the current week. This is more often the case with shelf stable products like salty snacks that wouldn't necessarily need to be stocked up on week to week. Thus, we rebuilt our original design matrix by adding a variable *prior* which is used as an interaction term. Our new design matrix has the form:

$$\begin{pmatrix} 1 & \mathrm{x}_1 & \mathrm{x}_2 & ... & \mathrm{x}_{14} & \mathrm{x}_1 * (1 - prior) & \mathrm{x}_2 * (1 - prior) & ... & \mathrm{x}_{14} * (1 - prior) \end{pmatrix}$$

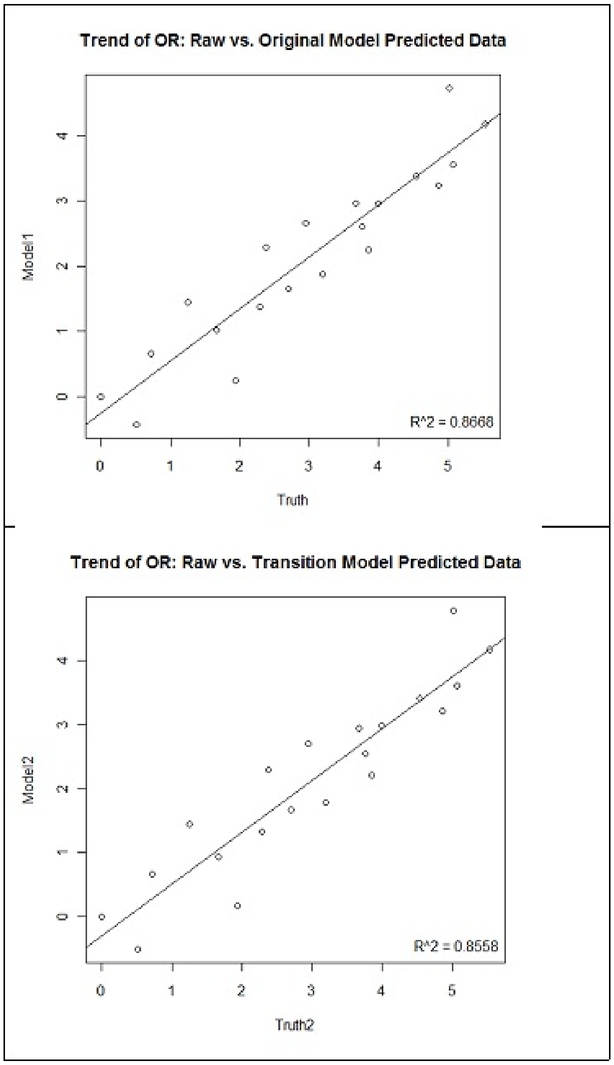
where *prior* is an dichotomous indicator if a purchase as been made in the past two weeks. Note that for our product we believed two weeks was a sufficient amount of time but this number can be chosen to be any number of weeks, as long as it is binary. After our model is built, we went through the process of expanding the matrix by the B-splines and then running model selection. Our final transition model ending up including [intercept, display multiplied by the B-spline basis, family3*(1-prior)]. Note that this is the same model that we previous had, except we found that if we knew that the panelist purchased the product in the past two weeks then the variable *family3* would fall out. So the final model would look like:

$$\left(\begin{array}{ll} purchase \sim \beta_0(z) + \beta_1(z) * display + \beta_3 * family3 * (1 - prior), & \text{prior=0;} \\ purchase \sim \beta_0(z) + \beta_1(z) * display, & \text{prior=1.} \end{array} \right.$$

Here z is our count of weeks a purchase was made in the prior year. In order to compare this model against our original data, we looked at the predictive power of this model on the 2006 data vs. the predictive power of the original model. In the Figure 2.7, we look at the OR between the actual data and the prediction data for both models as previously done with the original model. Based on these graphs, we see that the results are very similar and the original model even has a slightly higher correlation. However, we may see in some cases with certain products that knowing what was done in the prior weeks may have an effect, so it is important to look into the use of a transition model.

2.4 Conclusions

In this section we were able to develop a model selection method to not only determine if the explanatory variables should be included in the model, but if they should be included, should their coefficients be varying or constant. In our test data we found that display had a varying coefficient, which supports the claim of many marketing researchers that different people react differently to marketing variables. In our case, the difference came across the frequency of purchasing a product. We also found that the demographic variable of having a larger family was significant as a constant coefficient. We then took the model a step further to test claims that recency of purchases plays an important role. By considering the rolling 52 weeks and the transition model based on previous purchase, we found that the recency of the purchases were not as significant as the frequency of the purchases that was modeled through varying coefficients. Further research could be done on different products to see if any of these variations of the model are significant.


Trend of OR: Raw vs. Original Model Predicted Data

Trend of OR: Raw vs. Transition Model Predicted Data

Figure 2.8: Trend of Modeled vs. Raw Data for Original vs. Transition Model. Here we compare Figure 2.6 which was built off our original model with the same graph built off of using the Transition model.

Table 2.3: Table of BIC vs. Lambda in Step 2 of Model Selection

Lambda	Number Nonzero Coefficients	BIC	Variable No Longer Varying
1	25	7127.255	None
10	25	7127.255	None
20	25	7127.255	None
30	25	7127.255	None
40	25	7127.255	None
50	25	7127.255	None
60	25	7127.255	None
70	21	7119.31	Owner
80	21	7119.31	None
90	21	7119.31	None
100	21	7119.31	None
110	21	7119.31	None
120	21	7119.31	None
130	21	7119.31	None
140	21	7119.31	None
150	17	7083.323	Family4,Family3
160	17	7083.323	None
170	17	7083.323	None
180	17	7083.323	None
190	17	7083.323	None
200	13	7225.71	Display
210	13	7225.71	None
220	13	7225.71	None
230	13	7225.71	None
240	13	7225.71	None
250	13	7225.71	None
260	13	7225.71	None
270	13	7225.71	None
280	13	7225.71	None
290	13	7225.71	None
300	13	7225.71	None
310	13	7225.71	None
320	13	7225.71	None
330	13	7225.71	None
340	13	7225.71	None
350	13	7225.71	None
360	9	7225.71	None
370	9	7225.71	None
380	9	7225.71	None
390	9	7225.71	None
400	9	7225.71	None

Table 2.4: Coefficients from Final Model

Variable	Coefficient
intercept.1	-5.3190
intercept.2	-4.2625
intercept.3	-2.5194
intercept.4	-2.852
intercept.5	-1.2244
display.1	1.4415
display.2	1.8771
display.3	0.5411
display.4	1.9349
display.5	1.0877
family3	-0.5025

Chapter 3

Theory

In this chapter we study the asymptotic results of the variable selection method used in our work. The group LASSO itself does not possess the oracle property because the penalty is continuous in β, which does not allow for efficient discrimination between zero coefficients and nonzero coefficients (Liu (2010)). In order to obtain an efficient penalty that possess the oracle property, we turn to the adaptive group LASSO. The ideal adaptive group LASSO in has the following penalty:

$$\lambda \sum_{j=1}^{K} \sqrt{p_j} I\{\beta_{(j)} \neq 0\}, \tag{3.1}$$

where j=1,..,K refers to the group number. The adaptive LASSO penalty is discontinuous at zero and encourages coefficients to be zero. In the sections that follow we will prove that unlike the group LASSO, the adaptive group LASSO penalty possesses the oracle property.

It is important to note that for our model and all the theory to follow, we use a logistic model with a linear index. While there has been plenty of research on the use of splines as nonparametric approximation to smooth functions (Speckman (1985); Eilers and Marx (1995); Verbyla et al. (1999)), we limit our model to fixed knots. In some literature, people don't fix their knots because they the want to adapt the model to whatever shape the data may have by allowing for the number of knots to increase. When the true function is not a spline, and the number of knots increases with the sample size, the theory becomes more complicated, because there is an additional source of error: approximation error. Because in most applications the number of knots is small and fixed, we will not treat the more complicated case for the asymptotic theory and instead focus on this fixed-knot parametric model.

3.1 Approximating the Adaptive Group LASSO

The key feature of the ideal adaptive group LASSO is the discontinuity at zero. However, this greatly increases the complexity of the computation of the penalty. In order to preserve the oracle property of the ideal adaptive penalty while improving the computational time, we use an approximation of the ideal adaptive group LASSO (Liu (2010)). The main idea behind this approximation is to use a known consistent estimate of β as the weight. The following is the approximation of the adaptive group LASSO that will be used in all of the theory to follow:

$$\sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_j\|}\|\beta_{(j)}\|, \tag{3.2}$$

where $\tilde{\beta}_j$ is a known consistent estimate of β. This approximation method falls somewhere in between the group LASSO penalty and the LASSO penalty that we introduced at the beginning of this chapter and is asymptotically equivalent to the ideal adaptive group LASSO (Liu (2010)). In this approximation we define $\tilde{\beta}_j$ as the estimate from the group LASSO.

3.2 KKT Conditions

Like all variations of the LASSO penalty, the goal is to minimize the log-likelihood function plus some penalty. For the adaptive group LASSO, we wish to minimize the following function:

$$S(\beta;\lambda) = -\frac{1}{n}\ell_n(\beta) + \sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_j\|}\|\beta_{(j)}\|, \tag{3.3}$$

where $\ell_n(\beta)$ is the likelihood function for logistic regression as previously defined. Thus we define the adaptive group LASSO estimate to be:

$$\hat{\beta}_n(\lambda) = \arg\min(S(\beta;\lambda)). \tag{3.4}$$

In order to prove the oracle property of this estimate, we need to use the KKT conditions. Let β_j be the vector of coefficients that corresponds to group j. Define $\dot{\ell}_{n(j)}$ to be the first order derivative on the log-likelihood function with respect to β_j. Liu (2010) lays out these following KKT conditions for the adaptive group LASSO:

$$\begin{cases} \beta_j = 0, & \forall \tilde{\beta}_j = 0; \\ \mathrm{n}^{-1}\dot{\ell}_{n(j)}(\beta) + \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_j\|}\frac{\beta_{(j)}}{\|\beta_{(j)}\|} = 0, & \forall \beta_{(j)} \neq 0 \text{ and } \tilde{\beta}_{(j)} \neq 0; \\ \|\frac{1}{n}\dot{\ell}_{n(j)}(\beta)\| \leq \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|}, & \forall \beta_{(j)} = 0 \text{ and } \tilde{\beta}_{(j)} \neq 0. \end{cases}$$

The KKT set up the conditions for finding the minimizer of our objective function $S(\beta;\lambda)$ over β. Note that the first condition is a special case of the third condition since when $\tilde{\beta}_j = 0$, we have $\dot{\ell}_{n(j)}(\beta) = 0$. The third condition forces some of the components of β to zero. The second condition lays out that function that must be minimized when β is not zero. These conditions are not only necessary but sufficient for the implementation of the adaptive group LASSO. They are important for proving the asymptotic results in this chapter.

3.3 Asymptotic Results

We now prove that the adaptive group LASSO possesses the oracle property. In order to reach this result, we first must prove some statements about the convexity of our objective function and convergence of our estimate. We begin our proof of the oracle property with a proof that our objective function is convex.

Theorem 1. *The objective function (3.3) of the adaptive group LASSO is convex.*

Proof. Recall the objective function that we wish to minimize in the adaptive group LASSO:

$$S(\beta;\lambda) = -\ell_n(\beta) + \sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_j\|}\|\beta_{(j)}\|, \tag{3.5}$$

where, $\ell_n(\beta)$ is the log-likelihood function of the logistic regression with varying coefficients and can be defined as:

$$\ell_n = \sum_{i=1}^{n}[y_i(\beta_0 + \sum_{j=1}^{K}(x_{i,j})^T\beta_j) - \log(1 + \exp(\beta_0 + \sum_{j=1}^{K}(x_{i,j})^T\beta_j))] \tag{3.6}$$

In order to show that (3.5) is convex, we first show that the penalty is convex and then that the objective function of logistic regression with varying coefficients is convex. Note that the penalty in adaptive group LASSO consists of weights (λ and p) and the L_2 norm of β_j and $\tilde{\beta}_j$. We know by the axioms of a normed space (Vlaslov,1973), that for any normed space X, the function $x \to \|x\|$ is convex on X. Thus, since our penalty is compromised of constants multiplied by a function of L_2 norms, we know that the adaptive group LASSO penalty is convex.

We must now show that the $\ell_n(\beta)$ is convex. Let $g(z) = \frac{1}{1+e^{(z)}}$. Note that $1 - g(z) = \frac{e^{-z}}{1+e^{-z}}$ and $\frac{\partial(g(z))}{\partial(z)} = -g(z)(1-g(z))$. Thus,

$$\frac{\partial \ell_n(\beta)}{\partial \beta_j} = -\sum x_{ij}(y_i - g(\beta^T x_i)) \tag{3.7}$$

and

$$\frac{\partial^2 \ell_n(\beta)}{\partial \beta_j \partial \beta_k} = \sum x_{ij}x_{ik}g(\beta^T x_i)(1 - g(\beta^T x_i)). \tag{3.8}$$

In order for the objective function to be convex, we need to show that the Hessian matrix of second derivatives is positive semi-definite. Let ∇^2 be the Hessian matrix for our objective function. We need to show that for all vectors a, that $a^T \nabla a \geq 0$. For our problem, let us define $P_i = g(y_i\beta^T x_i)(1 - g(y_i\beta^T x_i))$ and $\rho_{ij} = x_{ij}\sqrt{P_i}$. Then we see

$$\begin{aligned} a^T\nabla a &= \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} a_i a_k x_{ij} x_{ik} P_i \\ &= \sum_{i=1}^{n} a^T \rho_i \rho_i^T a \\ &= \sum_{i=1}^{n} (a^T \rho_i)^2 \geq 0. \end{aligned}$$

Therefore, we have shown that for all vectors a, our Hessian matrix is PSD. Thus the objective function is convex. Since the adaptive group LASSO objective function plus penalty is the sum of two convex functions, we know that (3.5) is a convex function. □

Now that we have shown that our objective function is convex, we use this to show our estimate is consistent. We now define λ_n to be the optimal tuning parameter which changes with the sample size n. Thus we define $\hat{\beta}_n(\lambda_n)$=$\arg\min S(\beta; \lambda_n)$. We now show that $\hat{\beta}_n(\lambda_n)$ is consistent.

Theorem 2. *(Consistency) Assume that for some constant $0 < \varepsilon \leq \frac{1}{2}$, we have $\varepsilon \leq p_{\beta^0}(x) \leq 1 - \varepsilon$ for all x where $p_{\beta^0}(x) = E[y|x]$. Also assume that the matrix $E[xx^T]$ is nonsingular. Let $\tilde{\beta}$ be the minimizer found by group LASSO and β^0 be the true model parameter. If $\lambda_n \to 0$ as $n \to \infty$, then there exists a minimizer $\hat{\beta}_n(\lambda_n)$ of $S(\beta; \lambda_n)$ such that $\hat{\beta}_n(\lambda_n) \to \beta^0$ as $n \to \infty$.*

Proof. In order to show that the adaptive group LASSO results in a consistent estimator, we need to show $\forall a > 0$,

$$Pr\{\sup_{\beta:\|\beta-\beta^0\|=a} S(\beta; \lambda_n) > S(\beta^0, \lambda_n)\} \to 1. \tag{3.9}$$

Meier et al. (2007) showed that $\tilde{\beta}$ is a consistent estimate of β^0. Without loss of generality, assume the coefficients of the first S covariant groups, out of the $K \geq S$ groups are nonzero and the rest are zero. Since $\tilde{\beta}$ is consistent we know that $\|\tilde{\beta}_{(j)}\| I\{\beta^0_{(j)} \neq 0\} > \frac{1}{2}\|\beta^0_{(j)}\|$ with probability tending to 1 and there exists $c > 0$ such that

$$\min_{1\leq j\leq S} \|\tilde{\beta}_{(j)}\| > \frac{1}{2}\min_{j=1}^{S} \|\beta^0_{(j)}\| > c. \tag{3.10}$$

Now, if we look at the difference between $S(\beta, \lambda_n)$ and $S(\beta^0, \lambda_n)$, we see that

$$\begin{aligned}
S(\beta, \lambda_n) - S(\beta^0, \lambda_n) &= \frac{-1}{n}(\ell_n(\beta) - \ell_n(\beta^0)) + \sum_{j=1}^{K} \frac{\lambda_n \sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|}(\|\beta_{(j)}\| - \|\beta^0_{(j)}\|) \\
&= \frac{-1}{n}(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta})'(\beta - \beta^0) + (\beta - \beta_0)'(n^{-1}\frac{\partial^2 \ell_n(\beta^0)}{\partial \beta^2})(\beta - \beta^0) \\
&\quad + n^{-1}o_p(\|\beta - \beta^0\|^2) + \sum_{j=1}^{S} \frac{\lambda_n \sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|}(\|\beta_{(j)}\| - \|\beta^0_{(j)}\|) \\
&\geq -n^{-1/2}O_p(1)\|\beta - \beta^0\| + (\beta - \beta^0)'(\Sigma + o_p(1))(\beta - \beta^0) \\
&\quad + n^{-1}o_p(\|\beta - \beta^0\|^2) - \lambda_n \sum_{j=1}^{S} \frac{\sqrt{p_j}}{c}\|\beta_{(j)} - \beta^0_{(j)}\|,
\end{aligned}$$

where $\Sigma \triangleq E\{-\ddot{\ell}(\beta^0)\}$ is the positive definite Fisher information matrix. Note that since $\lambda_n \to 0$ as $n \to \infty$, all of the terms go to zero except $(\beta - \beta^0)'(\Sigma + o_p(1))(\beta - \beta^0)$. We showed in the consistency proof that Σ is positive. Thus, equation (5) holds. □

We have just shown that we have a consistent estimate. We now want to show the rate at which our estimate converges. Before we can prove that our estimate possesses the oracle property, we must show that our estimate is root-n consistent.

Theorem 3. *(Convergence Rate) Assume that for some constant $0 < \varepsilon \leq \frac{1}{2}$, we have $\varepsilon \leq p_{\beta^0}(x) \leq 1 - \varepsilon$ for all x where $p_{\beta^0}(x) = E[y|x]$. Also assume that the matrix $E[xx^T]$ is nonsingular. Let $\tilde{\beta}$ be the minimizer found by group LASSO and β^0 be the true model parameter. If $\lambda_n = O(n^{-1/2})$, then $\|\hat{\beta}(\lambda_n) - \beta^0\| = O_p(n^{-1/2})$.*

Proof. We need to show that $\forall \varepsilon > 0$, $\exists M > 0$, such that

$$P\{\sqrt{n}\|\hat{\beta}(\lambda_n) - \beta^0\| > 2^M\} < \varepsilon, \tag{3.11}$$

as $n \to \infty$. We begin by partitioning the parameter space defined as $\{\beta; \sqrt{n}\|\beta - \beta^0\| \geq 2^M\}$ into many different sections defined as $R_{j,n} = \{\beta; 2^{j-1} < \sqrt{n}\|\beta - \beta^0\| \leq 2^j\}$ where $j = M+1, M+2, \ldots$. Then for every $\eta > 0$, we have

$$\begin{aligned}
P\{\sqrt{n}\|\hat{\beta}(\lambda_n) - \beta^0\| > 2^M\} &= \sum_{j>M, 2^j \leq \eta\sqrt{n}} P\{2^{j-1} < \sqrt{n}\|\hat{\beta}_n(\lambda_n) - \beta^0\| \leq 2^j\} && (3.12)\\
&+ P\{\sqrt{n}\|\hat{\beta}_n(\lambda_n) - \beta^0\| > \eta\sqrt(n)\} && (3.13)\\
&= \sum_{j>M, 2^j \leq \eta\sqrt{n}} P\{\hat{\beta}_n(\lambda_n) \in R_{j,n}\} + P\{\|\hat{\beta}_n(\lambda_n) - \beta^0\| > \eta\}. && (3.14)
\end{aligned}$$

Recall in Theorem 2, we showed that $\hat{\beta}_n(\lambda_n)$ is a consistent estimator of β_0. Thus, $P\{\|\hat{\beta}_n(\lambda_n) - \beta^0\| > \eta\} \to 0$. Note that for $\hat{\beta}_n(\lambda_n) \in R_{j,n}$, we have

$$\inf_{\beta \in R_{j,n}} S(\beta; \lambda_n) \leq S(\beta^0, \lambda_n) \tag{3.15}$$

Thus,

$$P\{\hat{\beta}_n(\lambda_n) \in R_{j,n}\} \leq P\{\inf_{\beta \in R_{j,n}} \{S(\beta; \lambda_n) - S(\beta^0, \lambda_n)\} \leq 0\}. \tag{3.16}$$

We now define $h_n = \sqrt{n}(\beta - \beta^n)$ and r_1 to be the smallest eigenvalue of Σ. From our convergence theorem we see:

$$\begin{aligned} S(\beta;\lambda_n) - S(\beta^0,\lambda_n) &= -n^{-1}(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta}' h_n + n^{-1}h_n'(n^{-1}\frac{\partial^2 \ell_n(\beta^0)}{\partial \beta^2}))h_n \\ &\quad -n^{-1/2}\lambda_n \sum_{j=1}^{S} \frac{\sqrt{p_j}}{c}\|h_n\| + n^{-2}o_p(\|h_n\|^2) \\ &\geq -n^{-1}(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta})' h_n + n^{-1}r_1\|h_n\|^2 - n^{-1/2}\lambda_n \sum_{j=1}^{S} \frac{\sqrt{p_j}}{c}\|h_n\| + n^{-2}o_p(\|h_n\|^2) \end{aligned}$$

Because $\lambda_n = O(n^{-1/2})$, we know that

$$\lambda_n\sqrt{n}\sum_{j=1}^{S} \frac{\sqrt{p_j}}{c}\|h_n\| \leq 0.5 r_1 \|h_n\|^2 \tag{3.17}$$

for large enough n and 2^M, when $\|h_n\| > 2^M$. Therefore we have:

$$S(\beta,\lambda_n) - S(\beta^0,\lambda_n) \geq -n^{-1}(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta})' h_n + 0.5n^{-1}r_1\|h_n\|^2 + n^{-2}o_p(\|h_n\|^2). \tag{3.18}$$

We now define r_2 as the largest eigenvalue of $\sum$ and let $r = 8\sqrt{r_2}/r_1$ then from (3.15),(3.17), and the definition of h_n we have

$$P\{\hat{\beta}_n(\lambda_n) \in R_{j,n}\} \leq P\{\inf_{2^{j-1}<\|h_n\|\leq 2^j} -(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta})' h_n + 0.5 r_1\|h_n\|^2 \leq 0\} + o(1) \tag{3.19}$$

$$\leq P\{\sup_{2^{j-1}<\|h_n\|\leq 2^j} (n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta})' h_n \geq r_1 2^{2j-3}\} + o(1) \tag{3.20}$$

By applying Markov's Inequality we know that equation (3.20)

$$
\begin{aligned}
&\leq E\frac{|\sup_{2^{j-1}<\|h_n\|\leq 2^j}(\frac{\partial \ell_n(\beta^0)}{\partial \beta})' h_n|}{r_1 2^{2j-3}} + o(1) && (3.21)\\
&\leq \frac{(E\{\sup_{2^{j-1}<\|h_n\|\leq 2^j}(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta})^{\bigotimes 2} h_n\})^{1/2}}{r_1 2^{2j-3}} + o(1) && (3.22)\\
&\leq \frac{(E\{\sup_{2^{j-1}<\|h_n\|<2^j} h_n'(\Sigma + o_p(1))h_n\})^{1/2}}{r_1 2^{2j-3}} + o(1) && (3.23)\\
&\leq \frac{2^j\sqrt{r_2}}{r_1 2^{2j-3}} + o(1) && (3.24)\\
&\leq r2^{-j} + o(1), && (3.25)
\end{aligned}
$$

for some constant r. So we have shown,

$$
P\{\sqrt{n}\|\hat{\beta}_n(\lambda_n) - \beta^0\| > 2^M\} \leq o(1) + \sum_{j=M+1}^{\infty} r2^{-j} \leq o(1) + r2^{-M} \tag{3.26}
$$

Thus, $\sqrt{n}\|\hat{\beta}_n(\lambda_n) - \beta^0\| = O_p(1)$. □

The final step is to prove that the group LASSO possesses the oracle property.

Theorem 4. *(Oracle Property) Let $\hat{\beta}_{n(j)}(\lambda_n)$, a sub vector of $\hat{\beta}_n(\lambda)$, be the estimator of all β_j whose true values are 0, then $\hat{\beta}_{n(j)}(\lambda_n) = 0$ with probability tending to 1.*

Proof. From the KKT conditions we know that in order to minimize the objective function, it is necessary and sufficient that for all $\hat{\beta}_{n(j)}(\lambda_n) = 0$, we have:

$$\frac{1}{n}\|\dot{\ell}_{n(j)}(\hat{\beta}_n(\lambda_n))\| \leq \frac{\lambda_n\sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|} \tag{3.27}$$

Thus, we need to show that

$$Pr\{\frac{1}{\sqrt{n}}\|\dot{\ell}_{n(j)}(\hat{\beta}_n(\lambda_n))\| \geq \frac{\lambda_n\sqrt{n}\sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|}\} \to 0, \tag{3.28}$$

as $n \to 0$. We begin by determining the convergence rate of the left side of the equation. Using Taylor's expansion, we can see that

$$\begin{aligned}
n^{-1/2}\dot{\ell}_{n(j)}(\hat{\beta}_n(\lambda_n)) &\triangleq n^{-1/2}\frac{\partial\ell_n(\hat{\beta_n}(\lambda_n))}{\delta(\beta_{(j)})} \\
&= n^{-1/2}\dot{\ell}_{n(j)}(\beta^0) + \frac{1}{n}(\frac{\delta\dot{\ell}_{n(j)}(\beta^*)}{\partial\beta})'\sqrt{n}(\hat{\beta}_n(\lambda_n) - \beta^0) \\
&= O_p(1) - O_p(1)O_p(1) \\
&= O_p(1)
\end{aligned}$$

where β^* is a value that falls between $\hat{\beta_n}(\lambda_n)$ and β^0. Now if we turn to the right side of the equation, by Theorem 3, we know that $\|\tilde{\beta}_{(j)}\| = O(n^{-1/2})$. So,

$$\frac{\lambda_n\sqrt{n}\sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|} = O(\lambda_n n\sqrt{p_j})) \to \infty,$$

as $n \to 0$. Thus,

$$Pr\{\frac{1}{\sqrt{n}}\|\dot{\ell}_{n(j)}(\hat{\beta}_n(\lambda_n))\| \geq \frac{\lambda_n\sqrt{n}\sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|}\} \to 0, \tag{3.29}$$

which implies that $\hat{\beta}_{n(j)}(\lambda_n) = 0$ with probability tending to 1.

□

Since the group LASSO alone would not have been consistent in model selection, we needed to introduce the adaptive group LASSO. In this section we provided the theory to show that the adaptive group LASSO possesses the oracle property. We were able to show this by first showing out estimate is consistent and then further showing it was root-n consistent. Through these proofs, we can now support our model selection methodology.

Chapter 4

Simulations

In this next section, we wish to assess the validity of our model selection methodology. In order to assess the performance of our model selection methodology, we run a series of simulations. In the first set of simulations, we generate a basic example in which one coefficient is set to be constant, one moves along a curve and the rest are zero. We use this simulation study as a basic check that our method is performing correctly. In the next set of simulation we want to make the purchase data look as real as possible, so we use our model in order to simulate weekly purchases based on actual panel data. The goal of this exercise is to produce many data sets, perform the model selection, and record the frequency in which the correct model is chosen. The next sections lay out the steps in order to simulate data and run through simulations of choosing the best model.

4.1 General Simulation Outline

In the next two sections we describe in detail two series of simulations that we ran in order to test our methodology. Before we delve into the details of each simulation, we first set up a general outline of the algorithm we will use:

- Step 1: Generate a matrix of response variables, X, and a continuous variable, z, that is believed to produce varying coefficients. Using X, z, and a vector of coefficients, β, create a response variable y. The combination of X, z, and y will be the data that is used to test our model selection methodology.

- Step 2: Use group LASSO to run through some simulated data sets in order to find the optimal λ. This is the value which produces the model with the lowest BIC. Once this λ is determined, it will be fixed in the model selection methodology. This allows for automated simulations to be easily executed. Note that each stage of our model selection may have a unique λ.

- Step 3: Compute the b-splined matrix of z that will be used to expand our X matrix. This expanded matrix, X_2, will be the input into our group LASSO.

- Step 4: Using λ=1, run group LASSO in order to produce weights for adaptive group LASSO. Once the weights have been incorporated into the matrix, run stage 1 of model selection with the optimal λ chosen in step 2. Any grouped variable for which the norm of the group LASSO coefficients is nonzero, moves onto stage 2 of model selection.

- Step 5: Using only the variables selected in step 4, run stage 2 of model selection with the optimal λ from step 2. Any grouped variable for which the norm of the group LASSO coefficients is nonzero is determined to have a varying coefficient, all other variables that are still present will have a constant coefficient.

- Step 6: Run logistic regression on the final model and remove any variables that are not significant at the .05 level.

4.2 Simulations: Toy Example

In order to test our model selection methodology, we simulate our own data and run through our model selection methodology. The details of this simulation and results are as follows.

4.2.1 Toy Example Simulation Algorithm

Step 1: In this first set of simulations we test our methodology by defining our data in the following manner:

- Let X be 10,000 by 5 matrix of random N(0,1) values.

- Let z be a vector of 10,000 values drawn from a Uniform(0,15) distribution.

- Let β_0 =1.

- Let β_1=β_2=0.

- Let β_3=0.75.

- Let β_4=$\cos\frac{(z-2)*\pi}{4}$.

Figure 4.1 shows the curve of β_4 that we wish to capture in our model selection. By the definition of β_4, we see that the coefficient is a function of z. Therefore, in our model selection methodology we expect to see this chosen as a varying coefficient. We use this initial matrix and the coefficients to generate a response variable y by first calculating the probability vector:

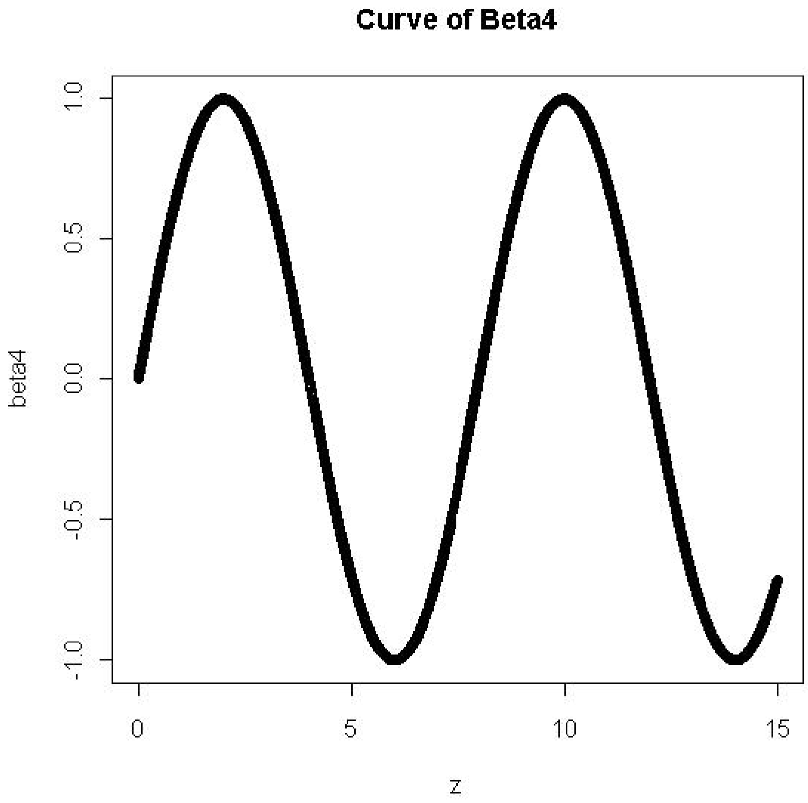


Figure 4.1: Plot of β_4 Curve

$$p = \frac{1}{1 + \exp\{-X'\beta\}}. \tag{4.1}$$

We then produce $Y \sim \text{Bin}(n,p)$. This model was specifically chosen to test that our model can:

1. Remove the columns of X that correspond to β values of zero.

2. Determine that there should be 1 varying coefficient and 1 constant coefficient in the model.

Step 2: Now that we have specified our model that we will use to generate data, we must determine what is the optimal value for λ that we will use in our group LASSO at each stage. We test out 10 simulated data sets and found the model with the lowest BIC was reached at λ=60 for stage 1 and λ=35 for stage 2.

Step 3: We now set up our B-splined X-matrix that will be used for model selection. For this data set we have set our knots at z=3,6,9 in addition to the boundary knots at z=0 and z=15. The matrix based on these values in then used to expand our X defined in step 1.

Step 4: We run an initial round of group LASSO with λ=1 to get our weights. We then run group LASSO on all of the variables with λ=60 (from step 2). We keep any variables where the norm of the group LASSO coefficients is non zero.

Step 5: Taking only the variables that were chosen in step 4, we run through stage 2 of model selection with λ=35. Any grouped variable for which the norm of the group LASSO coefficients is nonzero is determined to have a varying coefficient, all other variables that are still present will have a constant coefficient.

Step 6: Run logistic regression on the final model and determine which variables are significant and should remain in the model.

4.2.2 Toy Example Results

For this initial exploration of our model selection capabilities, we ran 1,000 simulations and each time recorded the model that was chosen. The results from this process can be found in Table 4.1. We see a success rate of 91.5%. Most of the issues came from variables being classified as varying when then should be constant and vice versa. While these runs gave us some support in our model selection methodology, we turn to a more advanced simulation study to test the performance of our methodology.

Table 4.1: Simulation Results for Toy Data Example after 1,000 simulations.

Result	Count	Percent
Correct Model Selected	915	91.5%
Variable Missing	8	0.8%
Extra Variable - Varying	22	2.2%
Extra Variable - Constant	55	5.5%

4.3 Marketing Data Simulations

In our next set of simulations,we chose to make the simulated data as similar to the data as possible. Thus, we decided to take a subset of the actual data, use the final model to generate probabilities, and then use these probabilities to randomly generate response variables. By using the final model to generate the response variable, we know what model should be selected each time because we know the exact relationship of the data. The following sections provide the detailed steps to produce these simulations and the results.

4.3.1 Marketing Data Example Simulation Algorithm

Step 1: Using the coefficients of the final model from Chapter 2, we let β be the vector of coefficients as defined in Table 4.2. Recall in the final model we have the intercept that has expanded by the B-spline matrix, the variable display that has been expanded by the B-spline matrix, and the variable family3 which is an indicator of all families size three or less. We define z to be the actual previous purchases in our data. We then let X be defined by the variables that were chosen for our final model in Chapter 2: splined intercept, splined display, and family3. We generate $p = \frac{1}{1+\exp\{-X'\beta\}}$ and the response variable $y \sim Bin$(n,p) where 1 denotes a purchase and 0 no purchase. We now combine this response vector y with the complete set of explanatory variables to form the dataset that we will run through the model selection process.

Step 2: Before we begin testing our model selection methodology and simulating all of our data, we first fix some parameters of group LASSO in order to optimize our simulation time. After 10 rounds, we found the value of lambda that produced the model with the optimal BIC to be λ=110. In the second step, we also found that λ=110 is the optimal value. We are now ready to run through the automated algorithm.

Step 3: Before we run the group LASSO, we need to set up the the B-splined X matrix that includes all of the variables and will be used in the model selection process. The matrix will be defined in the same manner as in Chapter 2 where X_{bs} is a matrix of five columns (b_1, b_2, b_3, b_4, b_5) where each column is a B-spline basis. Then our final expanded matrix is: (1*b_1, ..., 1*b_5, x_1*b_1, ..., x_1* b_5, x_2* b_1, ..., x_2* b_5, x_3*b_1, ...,x_3*b_5 ... x_{14}*b_1, ..., x_{14}*b_5) where (1*b_1, ..., 1*b_5) is the intercept and (x_1,...,x_{14}) are the variables listed

Table 4.2: Coefficients from Final Model

Variable	Coefficient	Value
intercept.1	β_{0_1}	-5.3190
intercept.2	β_{0_2}	-4.2625
intercept.3	β_{0_3}	-2.5194
intercept.4	β_{0_4}	-2.852
intercept.5	β_{0_5}	-1.2244
display.1	β_{1_1}	1.4415
display.2	β_{1_2}	1.8771
display.3	β_{1_3}	0.5411
display.4	β_{1_4}	1.9349
display.5	β_{1_5}	1.0877
family3	β_2	-0.5025

in Table 2.1.

Step 4: Using *grplasso*(), run a group LASSO with λ =1 in order to obtain weights for adaptive group LASSO. Using the coefficients from this initial group LASSO to weight the X matrix and form the new matrix

$$W\!= (w_0 * b_1, ..., w_0 * b_5 \;\; w_1 * b_1, ..., w_1 * b_5 \;\; w_2 * b_1, ..., w_2 * b_5 \;\; ... \;\; w_{14} * b_1, ..., w_{14} * b_5).$$

Once the weighted matrix W is built, run stage 1 of the model selection with $lambda$ = 110 and W as our matrix, to determine which variables should be in the model. Recall in stage one we treat each variable expanded by the B-spline basis as a group. So for every data set we use, the input to stage 1 will always be the full matrix,since at this point we will always have the 15 groups (intercept plus 14 variables) and the five spline basis. After stage 1, we pull the coefficients from the group LASSO to see which coefficients are non-zero. If the norm of the coefficients is 0 for all of the 5 coefficients in the group, then the variable is not in the model. If the norm of the coefficients is non-zero, then the variable remains in the model, and advances to the second stage of model selection.

Step 5: Using only the variables which were chosen to move on to stage 2, we next set up the W matrix for Stage 2 of model selection. Recall in this stage we much decide if a variable should have a varying coefficient or should be constant. Thus, we treat each variable as two distinct groups: the variable as is and the variable expanded by four of the five spline basis. So we now have (x_i, $w_i * b_1$, $w_i * b_2$, $\cdots$,w$_i * b_{s-1}$) where x_i is considered one group and ($w_i * b_1$, $w_i * b_2$, $\cdots$,w$_i * b_{s-1}$) is a second group for each x_i that was chosen in stage 1. Using *grplasso*() with $lambda$ = 110, run stage 2 of the model selection to determine which coefficients should vary. After stage 2, we once again pull the coefficients from the group LASSO to see which coefficients are non-zero. If the norm of the coefficient is 0 for all of the four coefficients in the second group for each variable, then the variable does not vary. If the norm of the coefficients are non-zero,

Table 4.3: Simulation Results for Marketing Research Data after 200 simulations.

Result	Count	Percent
Correct Model Selected	175	87.5%
Variable Missing	0	0%
Extra Variable - Varying	14	7%
Extra Variable - Constant	11	5.5%

then the variable is determined to have a varying coefficient.

Step 6: Based on the results from stage 1 and stage 2 of model selection, build the final X matrix for our model. Now, run the logistic regression on the final X-matrix and check to see that all variables are significant at the .05 for each constant coefficients and for at least one of the splines for the varying coefficients.

4.3.2 Results

We ran the algorithm on 200 data sets and recorded the results from each of the $\vec{r}_i$ vectors where $i = 1, ..., 200$. By looking at these simulations, we wanted to check the success rate of our data selection process where a success was defined as picking the exact model and a failure of the process was either including extra variables or excluding a variable. The results are summarized in table 4.3. We see that 7/8 of the time we choose the correct model. We never exclude a variable and 1/8 of the time we add an additional variable. Based on these results, we see that we have shown some validity in our model selection process.

References

Albert, P. S. and D. A. Follmann (2003). A random effects transition model for longitudinal binary data with informative missingness.

Allenby, G. M. (1990). Hypothesis testing with scanner data: The advantage of bayesian methods. *Journal of Marketing Research 27*, 379–389.

Allenby, G. M. and P. J. Lenk (1994). Modeling household purchase behavior with logistic normal regression. *Journal of American Statistical Association 89*, 1218–1229.

Allenby, G. M. and P. E. Rossi (1994). A bayesian approach to estimating household parameters. *Journal of Marketing Research 30*, 171–182.

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations (with discussion). *Journal of American Statist. Ass 96*, 939–967.

Azari, R., L. Li, and C.-L. Tsai (2005). Longitudinal data model selection. *Computational Statistics & Data Analysis 50*(11), 3053–3066.

Bakin, S. (1999). Adaptive regression and model selection in data mining problems. *Australian NationalUniversity, Canberra PhD Thesis*.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1978). *A Practical Guide to Splines*. Springer-Verlag.

Bongers, M. and J. Hofmeyr (2010). Why modeling averages is not good enough a critique of the law of double jeopardy. *JOURNAL OF ADVERTISING RESEARCH*.

Bronnenberg, Bart J.and Kruger, M. W. and C. F. Mela (2008). The iri marketing dataset. *Marketing Science 27*(4).

Brown, T. A. (1992). *Statistical Models in S*.

Cai, T. T. (1997). Discussion of regularization of wavelet approximations. *Journal of American Statist. Ass 96*, 960–962.

Ehrenberg, A., M. Uncles, and G. Goodhardt (2004). Understanding brand performance measures: using dirichlet benchmarks. *Journal of Business Research 57*.

Eilers, P. H. C. and B. D. Marx (1995). Flexible smoothing with b-splines and penalties. *Statist. Sci. 11*(2), 89–121.

Fader, P. S., B. G. S. Hardie, and K. L. Lee (2005). "counting your customers" the easy way: An alternative to the pareto/nbd model. *Marketing Science 24*(2).

Genkin, A., D. Lewis, and D. Madigan (1997). Large-scale bayesian logistic regression for text categorization. *Technometrics 49*, 291–304.

Glady, N., B. Baesens, and C. Croux (2009). A modified pareto/nbd approach for predicting customer lifetime value. *Expert Systems with Applications 36*, 2062–2071.

Guadagni, P. and J. Little (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science 2*(3), 203–238.

Gupta, S. (1988). Impact of sales promotion when, what, and how much to buy. *Journal of Marketing Research 25*(4), 342–355.

Hastie, T. Tibshirani, R. (1990). *Generalized Additive Models.* London: CHAPMAN & HALL/CRC.

Ibrahim, M., A. Sallau, A. Salihu, and O. K.C. (2011). Partial characterization of phospholipase a2 from the erythrocytic stage of plasmodium berghei. *Asian J. Biochem. 6*, 208–213.

K, S. and K. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics 19*, 2246–2253.

Kumar, V. and R. P. Leone (1988). Measuring the effect of retail store promotions on brand and store substitution. *Journal of Marketing Research 25*(2), 178–185.

L., K. and S. Raj (1988). A model of brand choice and purchase quantity price sensitivities. *Marketing Science 6*(1), 1–20.

Lattin, J. and R. Bucklin (2008). Reference effects of price and promotion on brand choice behavior. *Journal of Marketing Research 26*, 299–310.

Liu, L. (2010). Grouped variable selection in high dimensional partially linear additive cox model. *University of Iowa dissertation.*

Lokhorst, J. (1999). The lasso and generalised linear models. *University of Adelaide, Adelaide. Honors Project*.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *FRONTIERS IN ECONOMETRICS*, pp. 105–142. New York: Academic Press.

Meier, L. (2009). Fitting user specified models with group lasso penalty. *R Package*.

Meier, L., S. Van de Geer, and P. Buhlmann (2007). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B 70*, 53–71.

Roth, V. (2004). The generalized lasso. *IEEE Trans. Neur. Netwrks 15*, 16–28.

Schmittlein, D. C., D. G. Morrison, and R. Colombo (1987). Counting your customers: Who are they and what will they do next? *Management Science 33*, 1–24.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics 13*(3), 970–983.

Tellis, G. J. (2007). Advertising exposure, loyalty, and brand purchase: A two-stage model of choice. *Journal of Marketing Research 25*, 134–144.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B 58.*

Tseng, P. and S. Yun (2007). Acoordinate gradient descent method for nonsmooth separable minimization. *Math.Programmng B*, 1–28.

Vakratsas, D. and T. Ambler (1999). How advertising works: What do we really know? *The Journal of Marketing 63*, 26–43.

Verbyla, P., B. Cullis, M. Kenward, and S. Welham (1999). The analysis of designed experiments and longitudinal data using smoothing spline. *Applied Statistics 48*, 269–312.

Yuan, M. and Y. Lin (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B 68*(1), 49–67.