8-25-2016

Application of model selection techniques and measures of agreement to advertising data

Min A. Lee

Min A Lee

Candidate

Mathematics and Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

James Degnan , Chairperson

Yan Lu

Li Li

APPLICATION OF MODEL SELECTION TECHNIQUES AND MEASURES OF AGREEMENT TO ADVERTISING DATA

by

MIN A LEE

B.B.A., HANKUK UNIVERSITY OF FOREIGN STUDIES, 2014

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE
STATISTICS

The University of New Mexico
Albuquerque, New Mexico

JULY, 2016

Dedication

I would like to dedicate this work to God for leading me to this chapter in my life and always keeping me under His shelter throughout this journey.

I would like to dedicate this work to my family in Korea for their endless love and prayers. In spite of the physical distance between the United States and South Korea, we were always together in spirit through my joys and sorrows. All of my thoughts are always with them. They are my motivation and cheerleaders that help me not to give up under any circumstances.

I would like to dedicate this work to my host parents in Albuquerque, NM for their warm hospitality and unconditional love. In the love of God, they unconditionally support and encourage me to become a better person. I have learned how to become a loving person who tries to follow Jesus' example.

Acknowledgments

I would like to thank my advisor, Professor James Degnan, for his support in helping me to finish this thesis project. He always took an initiative to improve my project as well as myself as a statistician. His flexibility and responsibility as a faculty advisor contributed a lot to my academic growth during my thesis project. Our teamwork was amazingly productive and compatible enough to finish this project as planned. I would like to thank Ameritest, my employer and a client for this thesis project. I appreciate having the opportunity to work with them and their treating me as a part of their family during my internship. Particularly, I appreciate my supervisor and research director, Jesscia Sanchez, for her great support and allowing me access to internal resources in the company. I would also like to thank my friends that were always beside me. During this stressful thesis project, they were the cans of Red Bull getting me encouraged and caffeinated.

Application of model selection techniques and measures of agreement to advertising data

by

Mina Lee

B.B.A., Hankuk University of Foreign Studies, 2014

M.S., Statistics, University of New Mexico, 2016

Abstract

The primary purpose of this thesis is to explain the effectiveness of advertisement by predicting the attention score using the Flow of Attention graph and other survey responses. I address two problems: creating the algorithm to identify the peaks in the Flow of Attention graph and predicting the attention score based on predictor variables from questionnaire responses and the Flow of Attention graph. The sample data comprises a total of 141 randomly selected advertisements provided by Ameritest, a marketing research firm. The Problem 1 was addressed by two different algorithms; the first one is created manually based on moving average points with a window size of 3, and the other is an 'Edge detection function' derived from the other research. The manual moving average algorithm provided a better consistency with the reference peaks and the analysts' peaks by measurement of agreement, calculated with Cronbach's alpha. The Problem 2 was addressed by an missing imputation procedure and model selection procedure for the multiple regression model. Twenty out

of twenty three variables contained missing values and they were imputed by random regression imputation procedure. Model selection methods for the imputed data included the LASSO and all possible subsets by AIC. In order to get both a reliable and stable final model, the imputation was conducted a hundred times and found that the LASSO method provided a simpler and more stable result than all possible subsets by AIC method. Based on the final results from these two methods, the attention score increased when the audience liked the commercial, felt entertained, perceived it as different from other commercials, and felt better about the company (or the brand). The results also showed that the number of peaks, which is a variable from the Flow of Attention graph, did not indicate any significant impact to the attention score, since no model selection results contained the variable. Through the statistical analysis results in this thesis, the LASSO model selection shows a high stability of the results in the multiple random regression imputed data. Trying with various numbers of imputations and with other model selection methods can be suggested as future study to confirm the compatibility of the model selection methods in the presence of missing data.

Contents

List of Figures **xii**

List of Tables **xiii**

1 Introduction **1**

2 Data **4**

2.1 Main components used in this thesis 5

2.1.1 Attention 5

2.1.2 Brand Linkage 6

2.1.3 Motivation 7

2.1.4 Flow of attention 7

2.2 Data in terms of the problems 8

2.3 Data used in Problem 1 8

2.4 Problem 1 descriptive statistics 9

2.5 Data used in Problem 2 11

2.6 Problem 2 descriptive statistics 13

3 Research Problems **18**

3.1 Problem 1 19

3.2 Problem 2 21

4 Methodology **22**

4.1 Problem 1 22

4.1.1 Statistical methods: Measurement of agreement 23

4.1.2 Statistical methods: Bootstrapping 26

4.1.3 Algorithms 29

4.2 Problem 2 33

4.2.1 Imputation of missing data: Random regression imputation . . 33

4.2.2 Multiple regression 35

4.2.3 Variable selection 37

5 Analysis Results **44**

5.1 Problem 1: Comparison of algorithms by frames 44

5.2 Problem 1: Comparison of algorithms by number of peaks 45

5.3 Problem 1: Simulation of Cronbach's alpha 46

5.4 Problem 2: Missing Data Imputation 1 48

5.5 Problem 2: Missing Data Imputation 2 50

5.6 Problem 2: Variables selected by the model selection method 52

6 Discussion **55**

List of Figures

2.1 An example of flow-of-attention graph 8

2.2 Histogram of the number of frames per ad 10

2.3 Histogram of the number of peaks 11

4.1 An example of moving average graph with a window size of 3 (Top: the moving average graph, Bottom: the original graph) 30

4.2 Residual plots for linear model 36

4.3 Histograms of two data distribution with/without imputation 36

4.4 An example of all possible subsets plot in BIC 41

List of Tables

2.1 Survey Score Components 5

2.2 Problem 1 raw data structure 9

2.3 Problem 1 descriptive statistics 9

2.4 Frequency of Number of frames 10

2.5 Frequency of Number of peaks 11

2.6 Final samples (Reduced version of the data) 12

2.7 Variables description 13

2.8 The counts of missing values in variables (out of 136) 14

2.9 Strongest correlations between attention and other variables 14

2.10 Pairwise combinations of the highest positive correlation values . . . 16

2.11 Pairwise combinations of the highest negative correlation values . . . 17

5.1 Correspondence Analysis Results By Frames 45

5.2 Correspondence Analysis Results By Number Of Peaks 46

5.3 Confidence intervals of the Cronbach's alphas from various comparisons 1 47

5.4 Confidence intervals of the Cronbach's alphas from various comparisons 2 47

5.5 Missing Imputation: Strongest correlations between attention and other variables 48

5.6 Missing Imputation: Pairwise combinations of the highest positive/negative correlation values 49

5.7 Missing Imputation2: Model selection results by the LASSO method (Out of 100) 52

5.8 Missing Imputation2: Model selection results by the all possible subsets method by AIC (Out of 100) 52

5.9 Missing Imputation2: Selected variables from two model selection results (Out of 100) 53

Chapter 1

Introduction

These days, it is impossible to avoid advertisements in our daily lives. Advertisements have become a part of our daily conversation as well as a tool that guides our purchases. At the same time, ads are becoming more expensive, particularly, on television. For example, in 2008, national commercials produced by an advertising agency cost an average at $342,000 for a 30-second spot, according to the American Association of Advertising Agencies (Wagner, 2008). Following these trends, effectiveness of advertisements has become a measuring stick for efficiency in marketing research. In that sense, marketing research plays a big role in measuring and improving the effectiveness of advertisements.

This thesis addresses the problems of marketing research data provided by Ameritest, CY Research Group particularly about survey responses from the ad-testing process. To fully understand the meaning of these research problems and the data, it is necessary to briefly explain the basic concepts of marketing research in which this project is rooted. According to Nair (2009),

> A very popular definition of Marketing Research by the American Marketing Association is given here under : Marketing Research is the function which links the consumer, customer and public to the marketer through information—information used to identify and define marketing opportunities and problems, generate, refine and evaluate marketing actions, monitor marketing performance and improve understanding of marketing as a process. Marketing Research specifies the information required to address these issues, designs the method for collecting information, manage and implements the data collection process, analyzes the results and communicates the findings and their implications.

The research in this paper focuses on quantitative data analysis using statistical methods to identify and define the marketing opportunities and problems in an advertising film by evaluating key factors necessary for an advertisement to be effective. To measure the effectiveness of an advertisement, the ad-test process consists of a series of questions after one views five ads, including one testing ad and four control ads in a television watching environment. Each series of questions covers each kind of measurement including key performance factors, such as attention, branding, and motivation.

Young (2005) said that major marketing researchers studying advertisements today agree that leading components of advertising effectiveness are attention, branding, and motivation. However, measuring these components is a challenge. One of the most popular measures in current ad-testing systems is the attention, which indicates how efficiently the commercial catches a mass audience for the advertising message. It is commonly measured with either a degree of interest/breakthrough or recall in most ad-testing systems. Branding, or brand linkage, is also measured as a part of most ad-testing systems, and is usually measured in a proprietary procedure that can vary by individual ad-testing systems. It plays a big role to get viewers

involved in the brand after the advertising. Motivation demonstrates how persuasive the ad is at making viewers purchase a product or visit the store. It is a measurement of how much advertising affects sales.

Chapter 2

Data

The dataset in this paper comes from survey responses provided by Ameritest, a marketing research company. A total of 141 unrelated advertisements (subjects) are used in this thesis. The variables come from the questionnaires that the respondents had to complete for each ad. The variables are presented as scores, which were calculated as the average values of the responses from roughly $100 - 300$ respondents per ad. The sample size varied based on the survey design for each ad. Respondents were screened based on several characteristics such as gender, age, lifestyle, occupation, standard industry sensitivity, and previous participation in market research surveys. The demographics of people participating in the surveys varied depending on what kind of product or service that each advertisement delivered. For example, the usual 50:50 gender representativeness was allowed to vary depending on the ad.

According to Ameritest's ad-testing survey method, each respondent watched five commercials, including one test ad. After watching the commercials, the respondents participated in a 25-minute online interview that consisted of the following four steps: a) the observation of a clutter of one test and four control commercials, b) the identification of ads they perceived as interesting, c) the exposure to the test ad again

by itself, and d) the completion of a series of questionnaires. In this last step, the respondents were asked to score the brand fit/stretch, motivation, communication diagnostics, brand ratings, picture sorts®, and copy sorts™. Table 2.1 describes the components and subcomponents of the scores collected through the ad-testing survey.

Table 2.1: Survey Score Components

Components	Subcomponents
Key Performance Measures	Attention Brand Linkage Motivation
Verbal Diagnostics	Liking Entertainment Relevance Communication
Picture Sorts®	Flow of Attention Flow of Emotion
Copy Sorts™	Flow of Meaning Copy Recall and Relevance

This thesis focuses on the 'attention' and 'flow-of-attention' subcomponents in the first problem, and 'verbal diagnostics' and its subcomponents for the second problem. Later chapters present more detailed information about each subcomponent.

2.1 Main components used in this thesis

2.1.1 Attention

According to Ameritest, Chuck said that "the attention score is a direct measure of a commercial's ability to win in the street fight for audience attention." The way that Ameritest measures the attention score is by showing a clutter reel of five ads, in-

cluding the test ad, and asking viewers to choose the ads that they found interesting from the clutter reel. The reel simulates the environment of watching TV commercials. The test ad needs to overcome the other four control ads to gain attention. After watching the reel, the respondents answer these questions: "Which of these ads did you find interesting? Please describe the ad of which you are thinking." If respondents mention the test ad, the ad gains points.

2.1.2 Brand Linkage

The brand linkage measure from Ameritest focuses on effective branding in advertising. If the respondent's answer to the same question as above, regarding the attention score, includes references related to the brand, then the ad gains points. The responses are weighted based on how many respondents mention the brand and how clearly they explained the brand. Ameritest converts total sample responses into the ratio of respondents remembering the brand to the total number of respondents who found the test ad interesting.

> "Since the respondent only has to retrieve the brand name from short-term memory, this measure is not really a memory test but rather is a measure of brand salience or how top-of-mind the brand is in being associated with the memory of the ad. Consequently, an appropriate metaphor for understanding this branding construct is that of a handle, which the consumer uses to hold on to the experience created by well-branded advertising." (Young, 2005)

2.1.3 Motivation

Ameritest measures motivation using five point likert scale survey questions about consumer intention to buy the brand or product that the test ad is about. The questions are slightly different depending on the product or service that the ad conveys, ranging from "definitely will" to "definitely will not." For example, if the ad is for a retail brand, the question asks about the intent to visit to the store. Ameritest interprets the response of "definitely will" as the strongest weight to motivation score, which is called as top box motivation score.

2.1.4 Flow of attention

Ameritest developed the analysis tool called flow-of-attention to deconstruct and probe the ad viewer's dynamic response of frame-by-frame recognition. Basically, it is a graph showing the entire curve of recall scores frame by frame to visually demonstrate how much viewers recall each frame. Chuck Young, a creator of flow-of-attention says that, "The Flow of Attention measures how the eye preconsciously filters the visual information in an ad and serves both as a gatekeeper for human consciousness and as an interactive search engine involved in the co-creative process of constructing brand perceptions. The focus of analysis is on understanding the role of film structure and syntax in creating those powerful film experiences that can provide the basis for the consumer's emotional relationship with the brand." (Young, 2005) An example of flow-of-attention graph from Ameritest (Ameritest, Accessed Feb 2, 2015) is shown below as Figure 2.1.4 on page 8.

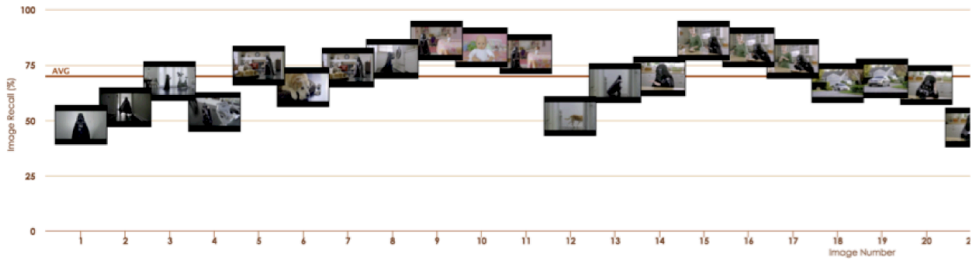


Figure 2.1: An example of flow-of-attention graph

2.2 Data in terms of the problems

There are two main problems I am working on in the thesis: one predicting peaks in the flow of attention graph, and the other on predicting attention from other variables using multiple regression. For the first problem, I have a dataset of the number of frames, recall scores from the flow of attention, and the reference peaks that were already selected. For the second problem, I have a dataset consisting of 23 variables including the attention score, the number of peaks, and the survey responses from other questionnaires.

2.3 Data used in Problem 1

A schematic of the data to be used for problem 1 in this thesis is shown in Table 2.2. Problem 1 is to create an algorithm to the problem is to determine the number of peaks and the frames for the peaks in the flow of attention graphs. The data consist of 141 observations of 50 variables including frame 1 to frame 41, number of frames, peaks, and the frames corresponding to peak 1 to peak 6. The first variable in the first column is the numbering of the ads. The 2nd to 42nd variables, called "Frame 1" up to "Frame 41" are frame numbers with recall scores of each frame per

Table 2.2: Problem 1 raw data structure

Ad Number	Frame1	···	Frame 41	Number of frames	Peaks	Peak 1	···	Peak 6
1	75	···		25	1,12,18,24	1	···	
···	···	···	···	···	···	···	···	···
141	95	···		14	2,5	2	···	

ad since maximum number of frames among 141 samples is 41. The 43rd variable, called "Number of frames", is the total number of frames per ad. The variable called "Peaks" shows the peak frame numbers per ad separated by commas. Peaks in the data are considered as reference peaks since they have been chosen by an agreement among senior analysts. From 45th to 50th variables, called "Peak 1" to "Peak 6", are each peak frame number since the maximum number of peaks was 6. Table 2.2 summarizes the scores that are measured through ad-testing survey.

2.4 Problem 1 descriptive statistics

Table 2.3 presents the descriptive statistics for the 141 ads used in problem 1. The minimum number of peaks is 1 and the maximum is 6. The most frequent number of peaks is 3, which is the same value as the mean. The average number of frames per ad is 22, the minimum number of frames is 10, the maximum is 41, and the most frequent number of frames is 21.

Table 2.3: Problem 1 descriptive statistics

	Ads	Most frequent	Min	Median	Max	Avg	Std.dev
Number of peaks	141	3	1	3	6	3.18	0.86
Number of frames	141	21	10	21	41	22.06	5.27

The distribution of the number of frames in the 141 ads is described in the histogram in Figure 2.4. It indicates that most of the ads have 15 to 30 frames per ad, and the overall pattern is a bit skewed to the right because there are two ads that have between 35 and 45 frames. Table 2.4 gives the frequencies for more accurate counts for each interval. The histogram of the number of peaks in Figure 2.4 and a frequency table of the number of peaks in Table 2.5 are provided as well, explaining that more than 90 percent of sample ads have 2, 3, or 4 peaks in a graph.

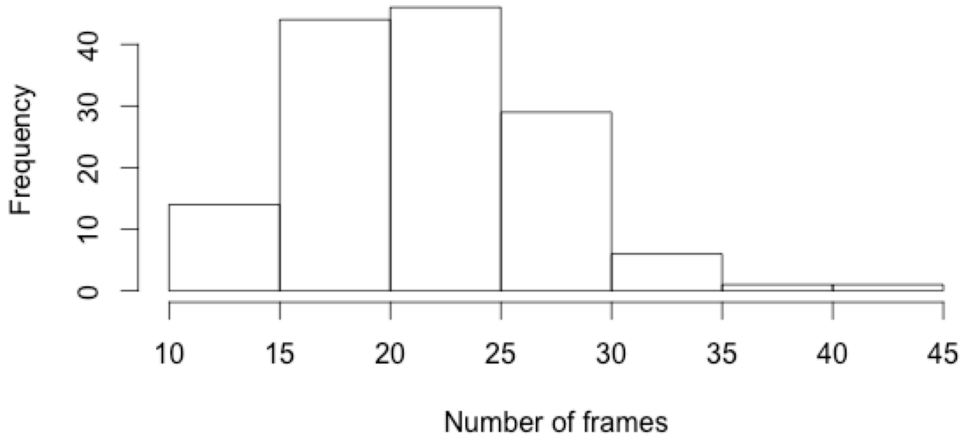


Figure 2.2: Histogram of the number of frames per ad

Table 2.4: Frequency of Number of frames

Number of frames	Counts	Percentage (%)
10-15	14	10 %
16-20	44	31%
21-25	46	33%
26-30	29	21%
31-35	6	4%
36-40	1	0.5%
41	1	0.5%

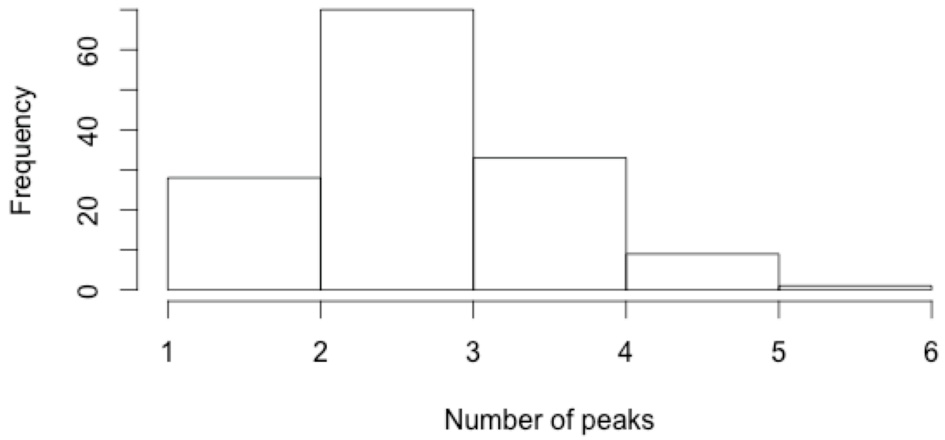


Figure 2.3: Histogram of the number of peaks

Table 2.5: Frequency of Number of peaks

Number of peaks	Counts	Percentage (%)
1	1	0.5 %
2	27	19.1 %
3	70	50.1 %
4	33	23.4 %
5	9	6.4 %
6	1	0.5 %

2.5 Data used in Problem 2

This section presents the data that I have used to address problem 2, which looks at the attention scores. Since problem 2 is an extension of problem 1, I use the same 141 observations of the ads. There are a total of 34 variables, including the number of peaks from the reference peaks in problem 1 and 33 variables from the ad-testing questionnaires. The entire dataset is composed of the actual values of 34 variables for 141 individuals. However, there is a high number of missing values in most variables, which causes a problem when fitting the prediction model. For example, two motivation variables consist of 102 missing values out of 141. Since

the model does not work appropriately with a large amount of missing values, I trimmed the variables and individuals when more than half of the values were missing. These missing values have been created because different questionnaires were used for each ad-testing, based on the product characteristics and the testing purpose. Consequently, the missing values were not introduced by dropping data nor by the interviewees' selective responses. The missing values are just the result of some questions not being asked in every ad-testing. This case is considered as 'missing at random' (MAR) since the data are missing independently of observed data. Most of the cases with MAR choose to omit the missing values since they do not affect other variables because of their independence. However, omitting observations with some missing information would lead to eliminating most observations from the dataset. To address the issue, the imputation of missing values should be conducted. A detailed description of this technique is presented in the methods section. After total 11 variables and 5 individuals were trimmed out, the final sample consists of 136 individuals with 23 variables shown in Table 2.6. Table 2.7 describes each variable. As shown in Table 2.7, most of the variables indicate the values of Top 2, which means top two boxes out of five-point scales responses. For example, top two boxes usually imply the counts of responses to 'Strongly agree' and 'Agree', whereas bottom two boxes imply the ones to 'Strongly disagree' and 'Disagree.' Except for the number of peaks, the variables take values between 0 and 100. The scores are processed to represent the interviewees' responses for each variable.

Table 2.6: Final samples (Reduced version of the data)

Ad number	V1	V2	...	V23
1	4	67	...	N/A
...	...	...	...	...
136	2	53.67	...	34.33

Table 2.7: Variables description

V1	Number of peaks
V2	Attention
V3	Brand Fit 1: Pretty good at making you remember
V4	Brand Fit 2: Just okay
V5	Brand Fit 3: Could be almost any kind of ad
V6	Brand Stretch 1: Fits
V7	Brand Stretch 2: New and fits
V8	Brand Stretch 3: Doesn't fit
V9	Top 2 - I can relate to the situation in the commercial
V10	Top 2 - I learned something [from the commercial] that I didn't know before
V11	Top 2 - (Liking) Overall, I like the commercial
V12	Top 2 - The commercial irritates me It is annoying
V13	Top 2 - The commercial is clever and entertaining
V14	Top 2 - The commercial is confusing
V15	Top 2 - The commercial is different from other
V16	Top 2 - The commercial is talking to people like me
V17	Top 2 - The message (in the commercial) is believable
V18	Top 2 - The message is important to me
V19	Top 2 - This is a commercial I would tell my friends or colleagues about
V20	Bottom 2 - (Liking) Overall, I like the commercial
V21	Goodwill- Better
V22	Goodwill- Same
V23	Top 2 -I like the music in the commercial (commercials with music)

2.6 Problem 2 descriptive statistics

The reduced version of the data, however, still has a high number of missing values. Table 2.8 shows the percentage of missing observations in each variable.

As the key response variable is attention (V2), I conduct a correlation analysis against each variable to see which predictor variable has the most association with the response variable. Table 2.9 shows the three strongest correlations both in a positive and negative direction. Variable V13, 'Top 2 - Entertaining', has the strongest

Table 2.8: The counts of missing values in variables (out of 136)

Variable	1	2	3	4	5	6	7	8	9	10	11	12
Missing	0	0	1	2	1	42	42	47	46	27	3	0
Variable	13	14	15	16	17	18	19	20	21	22	23	
Missing	14	12	12	33	51	15	43	12	46	49	54	

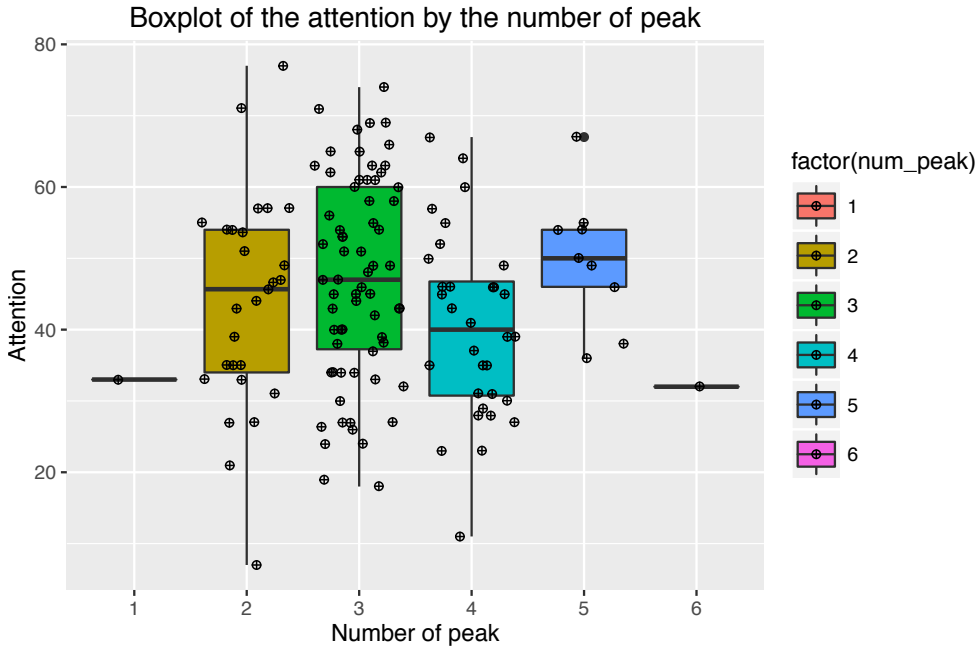
correlation with attention at 0.7333. The following variables are V11 and V16, 'Top 2 - Liking' and 'Top 2 - The commercial is talking to people like me', respectively. Regarding the correlations in a negative direction, variable V20 'Bottom 2 - Liking' has the strongest association with the attention score. Likewise, variable V22 and V12, 'Goodwill- Same' and 'Top 2 - The commercial is irritating and annoying', are also negatively correlated. The attention scores increase when people find the commercial entertaining, when they like it, and when they feel like the commercial is talking to people like them. On the other hand, attention scores decrease when people do not like the commercial, when they feel not that much of goodwill from the commercial, and when they feel irritated and annoyed by the commercial.

Table 2.9: Strongest correlations between attention and other variables

Positive			Negative		
Variable	Correlation	P-value	Variable	Correlation	P-value
V13	.7333	$<$ 2.2e-16	V20	-.4689	3.959e-08
V11	.6726	$<$ 2.2e-16	V22	-.4614	6.876e-06
V16	.5916	4.711e-11	V12	-.3864	3.389e-06

In problem 2, I mainly focus on the attention score and the number of peaks. The correlation between these two shows no significant association between them at -0.0252. The correlation seems very close to zero, which suggests that no linear relationship exists. In the boxplot in figure 2.6, most of the points are randomly scattered between 2 and 4 peak counts. No significant association is identified through numeric and graphical summaries.

I also fit a linear regression and find that there does not seem to have a linear association since the multiple R-squared came out as 0.000635, which means that only 0.06% of the variation in the attention score is accounted for by the number of peaks. The non-significance of the number of peaks ($p = 0.771$) indicates that changes in the number of peaks are not associated with changes in the attention. Therefore, I should include more variables in addition to the number of peaks to predict the attention variable.



I verify the multiple regression model assumptions by testing for multicollinearity within the explanatory variables. Table 2.10 shows nine pairs of variables have a positive correlation higher than 0.8. Table 2.11 shows five pairs that have correlations lower than -0.8. There are eight variables shown in the table 2.10, which means that there is evidence for intercorrelations between three variables. There are three combinations: a combination of V9, V16, V17, a combination of V11, V16, V17, and a combination of V9, V16, V18. For example, the first combination indicates that

the level of agreement of the respondents reacts in a similar way to the questions of 'I can relate to the situation in the commercial', 'The commercial is talking to people like me', and 'The message (in the commercial) is believable'. The other combinations are interpreted in the same way.

Table 2.10: Pairwise combinations of the highest positive correlation values

Pair	Correlation	P-values
V9 & V16	0.9268	$<$ 2.2e-16
V12 & V20	0.9152	$<$ 2.2e-16
V9 & V18	0.8785	$<$ 2.2e-16
V11 & V16	0.8755	$<$ 2.2e-16
V16 & V17	0.8698	$<$ 2.2e-16
V9 & V17	0.8676	$<$ 2.2e-16
V16 & V18	0.8516	$<$ 2.2e-16
V8 & V20	0.8321	$<$ 2.2e-16
V11 & V17	0.8195	$<$ 2.2e-16

The five strongest negatively correlated pairs are shown in table 2.11. Most of the pairs here are explained very intuitively such as the pair of V3 and V5, 'Brand Fit1: Pretty good at making you remember' and 'Brand Fit3: Could be almost any kind of ad.' An interesting pair to look at is V8 and V17, which means that many respondents answered that the ad doesn't fit to the brand well and the ad is not believable at the same time. In other words, if the ad does not fit to the brand, it also seems not believable. Based on the results with high correlation values across the explanatory variables, the model assumption of multicollinearity should be discussed in more detail in the model selection.

Table 2.11: Pairwise combinations of the highest negative correlation values

Pair	Correlation	P-values
V21 & V22	-.8673	$<$ 2.2e-16
V8 & V17	-.8551	$<$ 2.2e-16
V16 & V20	-.8513	$<$ 2.2e-16
V11 & V20	-.8406	$<$ 2.2e-16
V3 & V5	-.8015	$<$ 2.2e-16

Chapter 3

Research Problems

According to the article by Chuck, a founder of Ameritest (Young, 2009),

> Attention and memory are the alpha and the omega of advertising effectiveness in his article. The first thing an advertisement has to do is attract the attention of the consumer, or nothing else matters. If an ad does not leave some kind of lasting trace in the long-term memory of a consumer, it is difficult to argue that it had any kind of impact. But connecting the dots between attention and memory has not been easy for ad researchers.

In this thesis, a goal is to take a step closer to determining the connection between attention and memory by predicting the attention by looking at the flow-of-attention graph and other diagnostics such as 'Liking.' From the previous research in Ameritest, the attention score is a key factor that measures how memorable the ad is. An important variable of interest is the number of peaks in a flow-of-attention graph. We investigate whether a greater number of peaks in flow-of-attention graph indicates that the ad is more likely to gain attention. Furthermore, I would like

to determine the magnitude and statistical significance of the relationship between attention and memory. Therefore, the research objective is to predict the attention score of the advertisement based on multiple variables, including the number of peaks on the flow-of-attention graph, which indicates how much of the advertising delivery is kept in one's long-term memory. Two research problems need to be solved to achieve the research objective.

3.1 Problem 1

The first research problem is to create an algorithm to detect the peaks in a flow-of-attention graph which is as close as possible to research directors' manual identification of peaks. Peak moments are defined as local maxima in the flow-of-attention curve. These are images in the graph that are relatively higher than adjacent images in the neighborhood of that moment in time and are not defined relative to a norm or some absolute level of recognition (Young, 2005). The manual identification of peaks is similar to the typical computer program in that in a graph, a peak is the highest point between neighboring scores. However, Ameritest identifies moments as peaks when the following conditions are additionally met.

- A peak should reflect a visual meaning as well. Visual meaning is the image that a frame conveys about a product usage, brand logo, or some positive feeling. If it has an important visual meaning, it can be a good candidate for a peak.

- Potential peaks are analyzed in the context of the entire curve. Frames in a decreasing trend cannot be considered as peaks even when they are significantly higher than their neighbors. Without this condition, frames that are higher than their neighbors would be identified as peaks even if they are not the

highest points in the graph.

- When two neighboring points have the same scores and are potential peaks, a peak point would be determined by visual meaning. If they have similar visual meaning, the former one is chosen as a peak since the former frame is more influential to grab attention compared to the latter.

- Opening peaks and closing peaks are very tricky and controversial depending on various situations. Traditionally, the first frame has not been counted as a peak even if it scores very high because the high recall score of a first frame can be affected by the last frame of the previous advertising film. However, in the era of digital ads, attracting the viewers in the first few seconds becomes an important factor in effective advertisement since ads can be ignored or skipped if they don't grab attention at first glance. So in this thesis, peaks that exist in the first five frames are recognized as opening peaks.

- Closing peaks are among the last three frames when the ads finish with increasing trends. They are considered important because closing frames usually carry the images of a brand logo or company name with a key message that ads want to deliver the most.

A challenge is to detect the peaks meeting all these conditions. To find out the best-fitting algorithm for these data, I compare two approaches of detecting peaks and selecte the better one between them. As a first approach, I make my own rules modifying and combining different kinds of general methods of finding peaks since the algorithm should reflect various conditions in a numerical way as well as visual. As a second approach, I refer to an existing edge detection function (Filkov et al., 2002) about "Analysis Techniques for Microarray Time-Series Data." Although his paper analyzes microarray time-series data, which is completely different data from mine, the reason why I selected it is because the process of edge detection in his

paper has a similar concept as manual identification of peaks by senior analysts in Ameritest in that it focuses on the peaks with a certain size of magnitude in an increasing trend of the curve.

3.2 Problem 2

The second problem in this thesis is to predict the attention score from the number of peaks and other covariates. The attention score is the response variable and the other 22 variables are considered as explanatory variables. Out of the 22 variables, the number of peaks is considered to be the main explanatory variable. To predict the attention score, I use a multiple regression model. The variables in the final model are selected by the all possible subsets method, stepwise method, and LASSO. Model selction techniques such as all possible subsets used criteria such as AIC, AICc, BIC, and R^2_{adj}. A comparison study between variable selection methods and model selection criteria, respectively, is reviewed in the process to finalize the final model.

However, before proceeding to the model selection for multiple regression, the concern of missing values has to be addressed. The data consist of a significant amount of missing values in most of the variables except for the attention and the number of peaks. Imputation of missing data is necessary in this case. Therefore, the solution to the problem 2 involves several steps: imputation of missing data, fitting the multiple regression model, checking the model assumptions, comparison study between model selection and variable selection procedures, and determining the final model that best fits the data.

Chapter 4

Methodology

4.1 Problem 1

In problem 1, making an algorithm to identify the peaks that correspond to the manual peaks, I have two candidates from different approaches of algorithms and end up choosing one, which has a better performance of agreement with manual peaks. One of the algorithms is created with five different rules with a combination of senior analysts' intuition and some statistical concepts such as moving averages and local maxima. The other algorithm is called edge detection function, benchmarked from existing peak detection research. For selecting the final approach, several statistical methods such as Kappa statistic, Fleiss statistic, weighted Kappa statistic, and Cronbach's alpha, are used to measure the degree of agreement among raters. To increase the reliability of evaluation of the algorithms, the manually-identified peaks by four Ameritest analysts are included in the analysis. These analysts worked independently in the identification process. To produce confidence intervals for each consistency measure, I simulate them by using the empirical bootstrapping and the bootstrap percentile methods. This chapter describes each algorithm, the statistical

methods that are used to evaluate the consensus between reference peaks, analysts, and algorithms, and two bootstrapping methods.

4.1.1 Statistical methods: Measurement of agreement

Four tools are used to compare each dataset: Kappa statistic, Fleiss statistic, Cronbach's alpha, and weighted Kappa statistic. The variable being examined determines which analytic tool will be used. The Kappa statistic is used to compare the agreement between two binary variables. In the present study, I use it to determine whether each frame was consistently classified as a peak. The Fleiss statistic allows expanding this comparison to more than two binary variables. I use this analytical tool to examine the consensus among four independent analysts in terms of their classification of each frame as a peak. I can also consider the weighted Kappa statistic to estimate the degree of consensus in the ordered ratings (number of peaks) of two raters. Since it is rooted from the Kappa statistic, this tool works best when the test is between two raters. Cronbach's alpha tests for the degree of agreement between more than two raters of discrete ratings.

4.1.1.1 Kappa statistic

The Kappa statistic, κ measures the pairwise agreement among the raters using a dichotomous classification scheme. The equation is below:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{4.1}$$

where $P(A)$ is the proportion of observations that raters agree on and $P(E)$ is the expected proportion of agreement by chance. The consistency of the raters is calculated based on the expected consistency by chance. κ is obtained by subtracting $P(E)$ from $P(A)$, $P(A) - P(E)$, and dividing by the maximum proportion that could

arise by chance, becoming an adjusted agreement (Carletta, 1996), (Randolph, 2005). κ ranges from1 to -1, and the values between 1 and 0 indicate agreement better than chance, while the values equal to and lower than 0 indicate that the agreement is lower than expected by chance. There is no general rule of measure of significance for κ and the rule can vary depending on the field of study or the type of data, but in this paper, I will follow Landis and Koch Landis and Koch (1977). The authors have suggested a standard of interpretation for the non-negative values of the kappa statistic as follows: 0 is poor agreement, 0.01- 0.20 is slight agreement, 0.21- 0.40 is fair agreement, 0.41- 0.60 is moderate agreement, 0.61- 0.80 is substantial agreement, and 0.81- 1 is almost perfect agreement.

4.1.1.2 Fleiss statistic

The Fleiss statistic, also called Fleiss Kappa statistic, is a modified version of the Kappa statistic and is used to test the agreement with more than two raters. This statistic has many similar characteristics to the of Kappa statistic in that it is a chance-adjusted index of agreement with a dichotomous classification scheme. Likewise, it has the same equation as the Kappa statistic equation 4.1 on page 23, however, the calculation of $P(A)$ and $P(E)$ is a little different and more complicated than the Kappa since it considers more raters. The equations required to estimate

these components are as follows (Carletta, 1996):

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{4.2}$$

$$P(A) = \frac{1}{N} \sum_{i=1}^{N} P_i, \tag{4.3}$$

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{4.4}$$

$$P(E) = \sum_{j=1}^{k} p_j^2, \tag{4.5}$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{4.6}$$

where N is the total number of subjects and n_{ij} represents the number of raters who assigned the ith subject to the jth category. The significance of the Fleiss statistic is assessed in the same way as the Kappa statistic.

4.1.1.3 Weighted Kappa statistic

The weighted Kappa statistic is also a modified version of the Kappa statistic. It includes weights to account for disagreements differently, which is useful when there are variables with more categories than dichotomous ratings. The weights take into consideration the distance from agreement. As the weights are added, the equation consists of three matrices as follows: the observations of agreement, the expected agreement by chance, and the weights for disagreements. The weights can be set up in various ways based on the equation.

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}} \tag{4.7}$$

where w_{ij} is the weights for disagreements, x_{ij} is the observations of agreement, m_{ij} is the expected agreement by chance when n_{ij} represents the number of raters who

assigned the ith subject to the jth category. The weighted Kappa statistic is also interpreted in the same way as the Kappa statistic and Fleiss Kappa statistic.

4.1.1.4 Cronbach's alpha

Cronbach's alpha numerically assesses internal consistency. Researchers in behavioral and social sciences use it widely to test the consistency of items, which can be questions, indicators, or raters. It calculates internal reliability using the average covariances and the total variances of the variables.

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}\right) \tag{4.8}$$

where K is the number of items, σ_i^2 is the variance of the ith item and σ_T^2 is the variance of the total score formed by the sum of all the items (Vale et al., 1997). Alpha varies between 0 and 1; when all of them are perfectly identical, alpha is 1. On the other hand, when all the items are independent, alpha is 0. George George (2003) suggests a rule of thumb to describe the internal consistency based on the value of alpha as the following: an alpha less than 0.7 indicates a poor consistency; an alpha between 0.7 and 0.8 indicates an acceptable consistency; and an alpha more than 0.8 is good. However, this statistic should be used with caution since having too many items in the test can exaggerate the alpha, while having too little items can reduce it (Cortina, 1993).

4.1.2 Statistical methods: Bootstrapping

Bootstrapping is "a data-based simulation method for statistical inference" (Efron and Tibshirani, 1994) that allows estimating the uncertainty of the sample estimates. This method is powerful since it can be applied to any test that depends on random

sampling with replacement. Similar to obtaining the statistical inference of a population through the samples, the basic concept of the bootstrapping is to perform statistical inference from samples that have been collected from the original samples. The law of large numbers applies to the bootstrapping methods in that the empirical distribution obtained with a large number of samples will approximate the true distribution. In this paper, I use bootstrapping to simulate and statistically estimate the accuracy of the agreement statistics across algorithms, analysts, and reference peaks. Particularly, I simulate Cronbach's alpha as one of the consistency estimates since it is applicable to the majority of the comparisons in this thesis. I apply non-parametric bootstrapping because parametric bootstrapping would require knowing the distribution of the data (Shalizi, 2013). I chose this method over parametric bootstrapping because, in this case, it is hard to get the data distribution estimates with a qualitative survey. To get more accurate results, I use two methods for getting bootstrap intervals: the empirical bootstrap and the percentile bootstrap method. The first is theoretically more powerful to estimate the true distribution than the bootstrap percentile method, because the empirical one bases the estimates of the statistics on the samples. The latter bases the calculation of the confidence interval on the bootstrap sample distribution that is created by bootstrapping the samples.

4.1.2.1 The empirical bootstrap

The basis of the empirical bootstrap method is using the estimates from both the original samples and the bootstrapped samples. In this process, the statistic is Cronbach's alpha. Then, for each comparison, there is one Cronbach's alpha statistic obtained with the original sample and it is based on 30 observations for analysts, algorithm1, algorithm2, and the reference peaks. I create a thousand bootstrapped Cronbach's alphas for each comparison by randomly resampling from the original samples allowing for replacement. By the law of large numbers, a thousand repli-

cations make the samples distribution approximate the true distribution more accurately. Delta indicates a variation from the original sample alpha ($\widehat{\alpha}$) and the bootstrapped alphas ($\widehat{\alpha}^*$).

$$\delta^* = \widehat{\alpha}^* - \widehat{\alpha} \tag{4.9}$$

With $\widehat{\alpha}$ from the original samples and a thousand of δ^*, I can create a confidence interval of the Cronbach's alpha by subtracting a certain percentile bootstrapping sample statistics from the statistic of the original samples. Since I determine to get a 95% confidence interval, I select the 25th and 975th bootstrapped deltas as $\delta^*_{.025}$ and $\delta^*_{.975}$ out of 1000 in ascending order. Therefore, the lower bound of the confidence interval comes from subtracting the 975th bootstrapped alpha from the original samples alpha. The upper bound of the confidence interval comes from subtracting the 25th bootstrapped alpha from the original samples alpha.

$$CI_{0.95} = [\widehat{\alpha} - \delta^*_{.975}, \widehat{\alpha} - \delta^*_{.025}] \tag{4.10}$$

4.1.2.2 The bootstrap percentile method

The bootstrap percentile method is more intuitive and simple than the empirical bootstrap. The main difference is that it bases the calculation only on the bootstrapping samples, rather than on both the original and the bootstrapped samples. The statistics of interest in this paper is the Cronbach's alpha. The original 30 samples from each group are bootstrapped for a thousand times and a thousand Cronbach's alphas are created by this procedure. I organize the one thousand alphas in an ascending order and pick the 25th and 975th bootstrapped alphas. They represent the lower bound and upper bound of the 95% confidence interval, respectively. The confidence interval is given by,

$$CI_{0.95} = [\alpha^*_{0.025}, \alpha^*_{0.975}] \tag{4.11}$$

where $\alpha^*_{0.025}$ is the 25th bootstrapped alpha and $\alpha^*_{0.975}$ is the 975th bootstrapped alpha.

4.1.3 Algorithms

4.1.3.1 Algorithm 1. Five rules

Algorithm 1 consists of five rules: three opening rules, one general rule, and one closing rule. The opening rule means that the rule applies to the first five frames (opening frames) to detect peaks. The general rule means that the rule applies to most of the frames with the exception of the opening frames. The closing rule means that the rule applies to last four frames (closing frames). All the rules use moving averages with a window size 3. Moving average is a useful method since it reflects the general local trends, less influenced by random noises in the trends. For example, as shown in Figure 4.1.3.1 below, the moving average graph (upper graph) delivers a much clearer picture to detect whether there is an increasing trend or decreasing trend compared to the original graph (lower graph). To distinguish the moving average values and original values, The original values are denoted by x_i, and the moving average values are denoted by y_i. Therefore, y_i is the same with a value of $(x_{i-1} + x_i + x_{i+1})/3$.

Opening rules

There are three distinct rules for opening peaks. The first rule has the same approach with the general rule, and it applies to the first five frames. A frame is identified as a peak when it satisfies the following three conditions:

- The difference of the moving average of size 3 at point i is positive

$$y_i - y_{i-1} > 0, \text{where } y_i \text{ is a moving average of size 3} \tag{4.12}$$

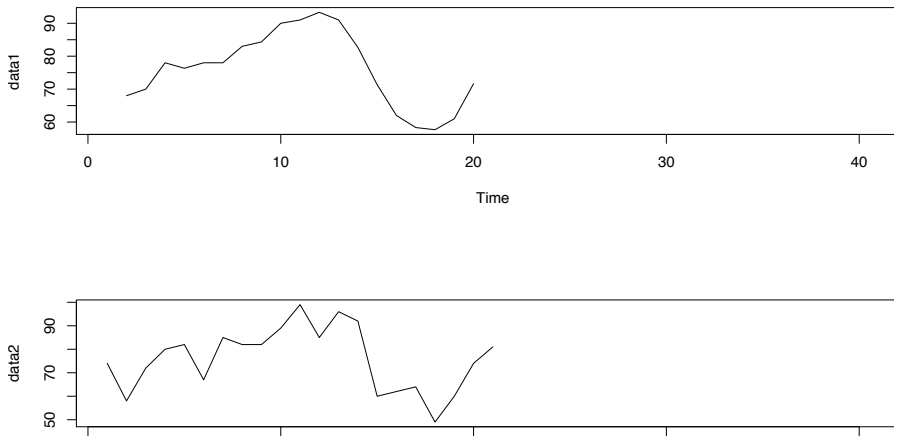


Figure 4.1: An example of moving average graph with a window size of 3 (Top: the moving average graph, Bottom: the original graph)

- The difference of the moving average of size 3 at point $i-1$ is positive

$$y_{i-1} - y_{i-2} > 0, \text{where } y_i \text{ is a moving average of size 3} \tag{4.13}$$

- The difference of the moving average of size 3 at point $i+1$ is negative

$$y_{i+1} - y_i < 0, \text{where } y_i \text{ is a moving average of size 3} \tag{4.14}$$

The second rule regarding opening peaks identifies the peak when it meets the following conditions:

- The difference of the moving average of size 3 at point i is positive

$$y_i - y_{i-1} > 0, \text{where } y_i \text{ is a moving average of size 3} \tag{4.15}$$

- The difference of the moving average of size 3 at point $i+1$ is negative

$$y_{i+1} - y_i < 0, \text{where } y_i \text{ is a moving average of size 3} \tag{4.16}$$

- The value of moving average at point i is greater than the average value of the original values.

$$y_i > \overline{x}, \text{where } y_i \text{ is a moving average of size 3 and } \overline{x} \text{ is the mean of } x \quad (4.17)$$

Since opening peaks are restricted to the first five frames, the second rule needs a shorter version of the first rule. Instead, a lower bound is set to the first rule to improve the accuracy.

The third rule of opening peaks applies specifically to the identification of the first frame as a peak. Since the first and second rules deal with the trends, the third rule considers the absolute size of the original scores. A frame is identified as a peak when its score is the highest among the first five original scores.

$$x_i \text{ is identified as a peak when } x_1 = max(x_1, x_2, x_3, x_4, x_5) \quad (4.18)$$

General rule

Under the general rule, a frame is identified as a peak when it meets the following conditions:

- The difference of the moving average of size 3 at point i is positive

$$y_i - y_{i-1} > 0, \text{where } y_i \text{ is a moving average of size 3} \quad (4.19)$$

- The difference of the moving average of size 3 at point $i-1$ is positive

$$y_{i-1} - y_{i-2} > 0, \text{where } y_i \text{ is a moving average of size 3} \quad (4.20)$$

- The difference of the moving average of size 3 at point $i+1$ is negative

$$y_{i+1} - y_i < 0, \text{where } y_i \text{ is a moving average of size 3} \quad (4.21)$$

The general rule applies to the frames located after the first five frames and before the last five frames. Among the frames satisfying the rule, I convert the moving average back into their original scores. Then, the highest original score is considered as a peak point.

Closing rule

Under the closing rule, a frame is identified as a peak when it meets the following condition:

- The difference of the moving average of size 3 at point i is positive

$$y_i - y_{i-1} > 0, \text{where } y_i \text{ is a moving average of size 3} \tag{4.22}$$

The closing rule applies to the last five frames. Among the points satisfying the rule, I convert the moving average back into original scores. Then, the highest original score is considered as a peak point.

4.1.3.2 Algorithm 2. Edge function

The edge function used in this thesis follows Filcov Filkov et al. (2002)'s "Analysis Techniques for Microarray Time-Series Data" article. Following the function set up in his research, I code the entire processes and apply it to the sample data. However, since there are important differences between our sample data and the microarray data, I add one more step to the original process to fit the sample data better. This additional step differentiates the peaks from the valleys. Therefore, I have a total of five steps to detect the peaks.

1. Primary edges: Remaining edges consisting of local maxima and minima after removing the monotone edges.

2. Secondary edges: Remaining edges after erasing insignificant edges whose (max-min)/average is smaller than a threshold, usually 30% by convention, considering that the changes below 30% are due to experimental errors.

3. Tertiary edges: Remaining edges after trimming the consecutive secondary edges of the same direction, are considered to be the same edges.

4. Quarternary edges: Remaining edges after eliminating narrow peaks or troughs that are likely resulting from the errors.

5. Final edges: Remaining edges after removing the valleys or troughs so that only peaks remain. The lower bound of all the remaining edges is the overall mean score. If the value falls below this threshold, then I drop the edge.

4.2 Problem 2

Problem 2 consists of the missing data imputation part and the multiple regression part. First, the imputation of missing data is necessary. Considering the characteristics of the missing data and the entire data, I choose one best fitting method to apply to the data. Once it becomes the complete dataset after imputation, I fit the data using the multiple regression model.

4.2.1 Imputation of missing data: Random regression imputation

Having complete and balanced data is ideal. In reality, however, data can contain missing values, which can be addressed with the use of statistical methods. The imputed data is not as good as the complete data, but still statistical imputation methods contribute to the improvement of the accuracy and concreteness in the data

analysis. In this thesis, imputation is necessary since model selection techniques did not work because of incomplete data.

There are many ways to approach the missing values, such as discarding the observations or variables with missing values, and simply filling in the missing values. The method that discards data consists of excluding the individuals with missing values and conducting the analysis only on observations with complete data. This is one of the easiest methods that you can use when you have large enough datasets with a relatively small amount of missing values so that removing the missingness does not significantly affect the entire analysis. However, this method does not work with the data in this paper since the data consists of only 141 individuals, which is not a large sample, and missing values represent a large part of the data. The other common way of addressing missingness is by filling in the missing values. In this method, all the missing values are filled with a single value, generally the mean or the last value carried forward. This method is not compatible with the data either, since this single value can bias the analysis result because some variables have several missing values. Therefore, I consider the random regression imputation method that puts the randomness back into the imputations by adding the prediction error into the regression.

Random regression imputation is based on the prediction of the missing values through the estimation of a linear regression on the observations with complete data. If the random prediction values are outside the range of the variables, it will be considered as the closest value within the range. For example, if the variable is a score between 0 and 100 and the randomly predicted value comes out as 103, I replace the 103 with 100, which is the closest value in the possible range.

I present the following example to illustrate the imputation of missing values in V3. Let V1 and V2 be the variables with complete data. Then, I run a linear regression model with V1 and V2 to predict the missing values in V3. The residual plots

are shown in figure 4.2.1. Overall, the plots show that the residuals are randomly and normally distributed and there is a potential outlier at point 88. However, I do not take any action for this outlier since it is still inside the Cook's distance. In the reduced data, V3 has one missing value. Therefore, I create one random value from the normal distribution with the predicted values ($\widehat{y}$) as a mean and a predicted standard deviation ($\widehat{\sigma}$) as a standard deviation. Before I impute the value, I check if the created value is in the feasible range and if it is compatible with our original data distribution by comparing the histograms of 'with imputation' and the one 'without imputation'. The comparison of the histogram of V3 in figure 4.2.1 shows that the two histograms look similar and compatible enough to impute the missing value in V3. This imputation process is used for each variable with missing values. Therefore, I finally get a complete dataset without any missing values. In this paper, a total of 20 prediction iterations is performed. Each prediction iteration is conducted by the order of the least number of the missing values in the variable among remaining variables that need the missing imputation procedure in order to get a better accuracy in the imputation process. However, I am still concerned about the accuracy of the missing imputation results since the prediction errors are likely to get inflated and make the imputed values inflated along through the 22 times of prediction iterations. I will discuss more details about the characteristics of this procedure based on the analysis results in the next chapters.

4.2.2 Multiple regression

A regression analysis is one of the most commonly used methods in statistics to estimate relationships among variables, focusing on the relationship between response and explanatory variables. In this thesis, I use the multiple regression model to predict the attention score as a function of other variables since I have more than one explanatory variable. Attention score (V2) is a response variable noted as Y in

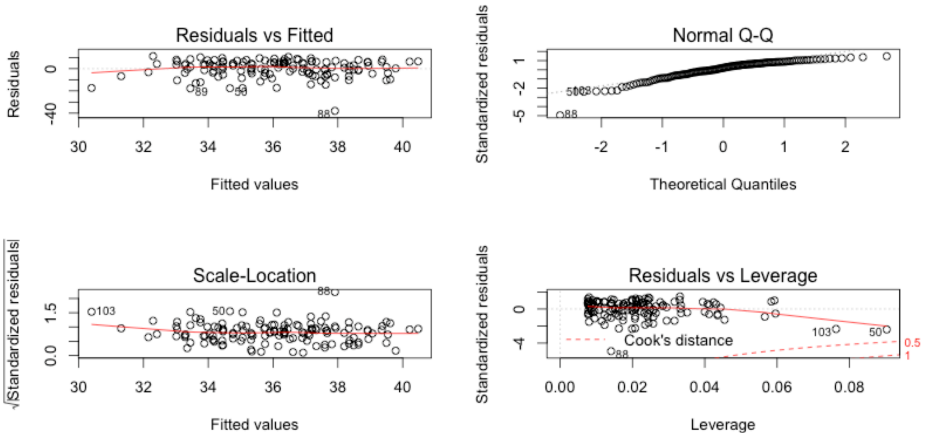


Figure 4.2: Residual plots for linear model

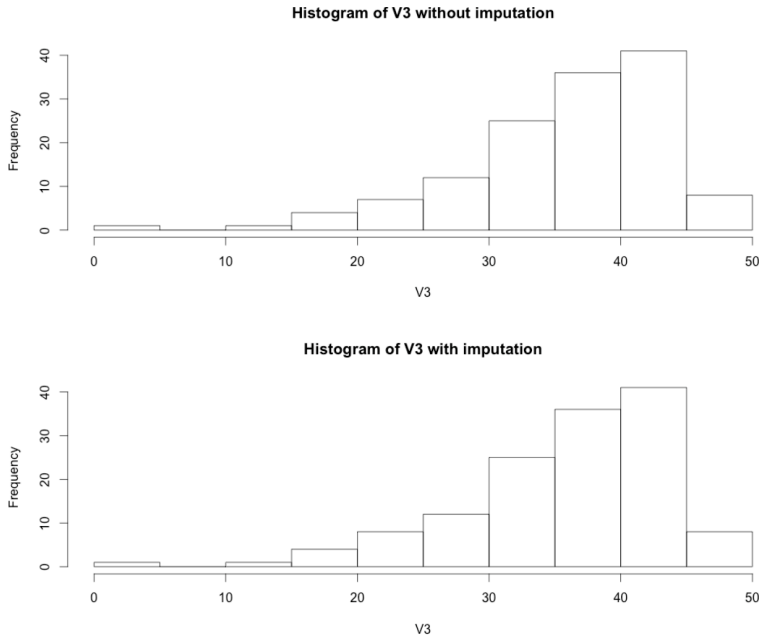


Figure 4.3: Histograms of two data distribution with/without imputation

this multiple regression and the other predictors such as V1, V3, ... , V23 are noted as $X_1, X_3, ..., X_{23}$. I choose the linear form model for this multiple regression based on the data description part in chapter 2.

Our multiple regression function consists of 22 predictors and an error term, which is statistical noise. The model is given by,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + ... + \beta_{23} X_{23i} + \varepsilon_i \tag{4.23}$$

where β_0 is an intercept, β_i is a coefficient on X_i, and ε_i is an error term. The coefficients (βs) are non random and unknown. The error terms are random and unobserved. ε is assumed to be normally distributed with mean 0.

After the multiple regression model is fitted, the fitted model is as follows:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_3 X_{3i} + ... + \widehat{\beta}_{23} X_{23i} \tag{4.24}$$

where $\widehat{Y}$ means the predicted value of Y.

4.2.3 Variable selection

The data has a total of 22 regressors to predict the attention score. To fit a multiple regression model, I should select the most fitting specification because 22 predictors are too many based on the sample size to create a simple multiple regression model. I calculate and compare R^2_{adj}, AIC, AICc, and BIC, and choose the model that minimizes AIC, AICc, BIC and maximizes R^2_{adj} at the same time. Each selection criterion is explained in detail below. Potential subsets for the best fitting multiple regression model are evaluated by two methods: all possible subsets and stepwise methods.

4.2.3.1 Evaluating predictor variables: Information criteria

R^2_{adj} is a criterion calculated by the variabilities of the regression model and the total sample model. AIC, AICc, and BIC are the criteria based on likelihood theory assuming that both the predictors and the response variable are normally distributed. They are also based on the likelihood function of the unknown parameters $\beta_0, \beta_1, \beta_2, ..., \beta_p, \sigma^2$, where p is the number of predictors given by, following Sheather's notation Sheather (2009),

$$\mathcal{L}(\beta_0, \beta_1, \beta_2, ..., \beta_p, \sigma^2|Y). \tag{4.25}$$

The log-likelihood function is then given by,

$$\log \mathcal{L}(\beta_0, \beta_1, \beta_2, ..., \beta_p, \sigma^2|Y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{RSS}{n}\right) - \frac{n}{2} \tag{4.26}$$

where $\widehat{\sigma}^2_{MLE} = \frac{RSS}{n}$.

- R^2_{adj}
  R^2_{adj} is obtained from the coefficient of determination of the regression model, R^2. R^2 is calculated as the proportion of the total sample variability in the Y's explained by the regression model (Sheather, 2009).

$$R^2 = \frac{SSreg}{SST} = 1 - \left(\frac{RSS}{SST}\right) \tag{4.27}$$

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)} \tag{4.28}$$

where p is the number of predictors in the current model and n is the number of observations.

Higher R^2 and R^2_{adj} indicate better fits. R^2_{adj} is similar to R^2 but with an adjusted version of the formula to compensate the weakness of R^2 that it often increases by adding irrelevant predictor variables. However, still R^2_{adj} sometimes tends to increase by adding more predictor variables. Therefore, a lower number of predictors in the subset is preferred when R^2_{adj} does not change significantly between two different subsets.

- AIC, AICc

  Akaike's information criterion(AIC) is used for balancing goodness of fit and a penalizing model complexity. A measure of good fit is defined as the smaller the better. It is calculated by multiplying minus one times the model likelihood, and adding a measure of complexity of the number of estimated parameters in the fitted model, K. AIC is calculated with the following formula:

$$AIC = 2[-\log \mathcal{L}(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p, \widehat{\sigma^2}|Y) + K] \tag{4.29}$$

  where $K = p + 2$. A MLE calculated version of AIC is the following one:

$$AIC = n \log\left(\frac{RSS}{n}\right) + 2p \tag{4.30}$$

  AICc, a corrected version of AIC, is better used when the sample size is small, or when the number of parameters estimated is a moderate to large fraction of the sample size. Conventionally, AICc is recommended when $n/K > 40$ where $K = p + 2$. The AICc formula is shown below:

$$AICc = AIC + \frac{2(p+2)(p+3)}{n-p-1} \tag{4.31}$$

  AICc has a larger penalty for model complexity when the sample size is small, or when the number of parameters estimated is a moderate to large fraction

of the sample size. AIC's penalty for model complexity is not enough under those circumstances.

- BIC

  Bayesian Information Criterion(BIC) is very similar to AIC but there is a difference in the second term as shown below.

$$BIC = -2\log \mathcal{L}(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p, \widehat{\sigma^2}|Y) + K\log(n) \tag{4.32}$$

  where $K = p + 2$. BIC has a heavier penalty for model complexity than AIC when n is bigger than 8 since $\log(n) > 2$. Therefore, BIC favors simpler models than AIC.

4.2.3.2 Deciding Potential Subsets of Predictor Variables: All possible subsets method

There are two different subset selection methods to choose potential subsets of predictor variables that are commonly used in many research projects: All possible subsets method and Stepwise methods. In the data, the 'all possible subsets' method considers all possible combinations out of the 22 predictor variables. The potential subsets of predictor variables are evaluated according to the selected information criteria that can be any of the criteria above. The 'stepwise methods' consist of adding or eliminating the predictor variables based on the information criteria from each iteration until the model reaches the best possible value of the information criterion. There are three ways of selecting the subsets through the stepwise methods: forward selection, backward elimination, and stepwise regression, which is a combination of the first two. In this paper, I use only the all possible subsets method since it is more accurate in that it considers all possible subsets out of all variables in the data, whereas the stepwise methods do not consider the entire possible subsets.

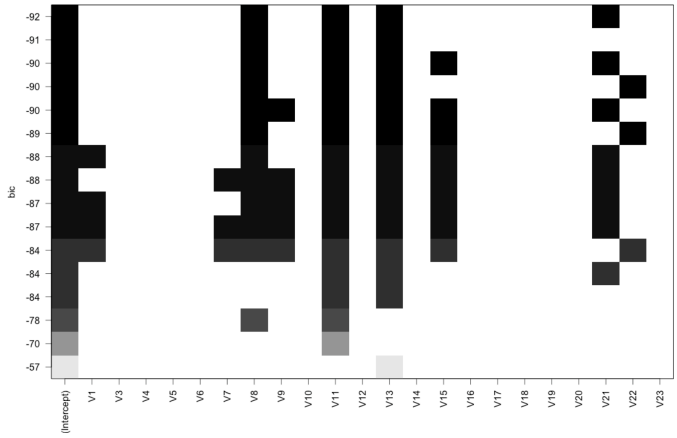


Figure 4.4: An example of all possible subsets plot in BIC

The all possible subsets method considers all possible regression subsets and chooses the best working subset of predictors that minimizes information criterion or maximizes a measure of fit. In this paper, I use the AIC as an information criterion for this method since the AIC would be the most reliable and traditional information criterion. BIC tends to under-fit the model because of too much penalty to the model complexity, at the same time, R^2_{adj} is not a reliable measure when the data has many variables. I think the AIC would be reliable enough since the it is equivalent to Mallows's C_p in a special case of Gaussian linear regression (Boisbunon et al., 2013). The color intensity indicates the degree of the significance of the results in the all possible subsets method plot. An example of this plot is shown in figure 4.2.3.2, and it tells that the best subset with the lowest BIC value is composed of V8, V11, V13, and V21.

4.2.3.3 Deciding Potential Subsets of Predictor Variables: Least Absolute Shrinkage and Selection Operator (LASSO)

Statistical prediction procedures face a dilemma between the variance and prediction. Higher variance leads to better accuracy since it covers the broader range of prediction values, but often comes out as not useful to predict the precise values. In the same context, the subset selection method presented above is a discrete process of selecting subsets with a high variance since the model retains only a significant subset of the variables and drops the rest. It means that even though the subset selection methods get a pretty significant prediction model, it tends to present a less useful model in terms of the model interpretation. To compensate for the high variance, another option can be shrinkage methods in that they are more continuous and have a lower variability. In this section, I present the least absolute shrinkage and selection operator (LASSO) given by the following formula (Oyeyemi et al., 2015).

$$\begin{aligned} \widehat{\beta}^{lasso} &= \arg\min_{x} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta j)^2 \\ &\quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{4.33}$$

The LASSO is a relatively new method that is currently popular both in academic fields and industry. Because of its lower variability, it tends to present a clear prediction interpretation for a response variable. Also, it works well even though there exists a high level of multicollinearity among model predictors since it minimizes the residual sum of squares subject to the sum of the absolute values of coefficient (β) being less than a constant (t in Equation 4.33), and just selects one most influential predictor out of the highly correlated group of the variables (Oyeyemi et al., 2015). This method consists of two parts: the least squared term and the coefficients constraint term. The constraint in coefficients helps to eventually choose a parsimonious

model in the model selection procedure. Due to the nature of the constraint, the LASSO may produce coefficients that are equal to zero that bring a simpler model with a reduced number of the variables.

In the equivalent Lagrangian form of the LASSO as shown in Equation 4.34, λ is a tuning parameter that controls the amount of shrinkage. The larger λ, the larger shrinkage in the model. The shrinkage level of the coefficients is determined based on the λ as one of the parameters in the LASSO model. Either t, λ, or s can be used as the tuning parameters, where s, sometimes called fraction in the codes, is a standardized version of t ($s = t/t_0, 0 \leq s \leq 1$). The LASSO model selection analyses in this paper mainly perform through the R package, LARS. In the package, the most proper value for a tuning parameter is determined by the cross-validated value that minimizes the mean square error of cross validation (cv-MSE). The best fitted model by LASSO is represented by the coefficients for each variable based on the chosen 's'.

$$\widehat{\beta}^{lasso} = \arg\min_{x} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,. \tag{4.34}$$

Chapter 5

Analysis Results

5.1 Problem 1: Comparison of algorithms by frames

Table 5.1 contains the results of the analysis by frames. It presents the results obtained through both the Kappa statistic and the Fleiss statistic. Both methods produce essentially the same results for the pairwise comparisons between the reference peaks, algorithm 1, and algorithm2. However, regarding the analysts' data, the results show differences between the methods because slightly different data have been used for each method. The data from only one analyst, out of four, has been selected for the Kappa statistic tests since this method only allows for bivariate data. I picked analyst 2 since he had the highest correlation with the reference peaks among all the analysts. For the Fleiss statistic, all four analysts were used since this statistic can handle more than two variables. Therefore, I focus on Fleiss because it considers the full dataset. As a result, algorithm 1 works better than algorithm 2 because the first shows a higher agreement between the reference peaks and the analysts' peaks than algorithm 2. Interestingly, the Fleiss statistic shows that the agreement between algorithm 1 and the analysts is very similar to the agreement

between the reference and the analysts' peaks.

Table 5.1: Correspondence Analysis Results By Frames

	Kappa	Fleiss
Reference & Algorithm1	0.59	0.59
Reference & Algorithm2	0.11	0.11
Reference & Analysts	0.58	0.67
Algorithm1 & Algorithm2	0.12	0.12
Algorithm1 & Analysts	0.58	0.67
Algorithm2 & Analysts	0.11	0.52

5.2 Problem 1: Comparison of algorithms by number of peaks

The results of the analysis of the number of peaks are shown in table 5.2 below. The results appear to be quite different from the frames analysis. I use two different methods to test the agreement between the reference peaks, algorithm 1, algorithm 2, and the analysts. Like the Kappa statistic, the weighted Kappa statistic has a systematic limitation in that it only accepts two variables. Similar to the analysis by frames in the previous section, the weighted Kappa uses analyst 2 data for "analysts" and the Cronbach alpha uses four analysts. The Cronbach alpha results suggest that the agreement between reference and algorithm 1 is stronger than the agreement between reference and algorithm 2. However, the agreement between analysts and algorithm 2 is stronger than the agreement between analysts and algorithm 1. In addition, the indicators of correspondence between analysts and both algorithm 1 and algorithm 2 are higher than the indicator of correspondence between analysts and reference. This gap, however, does not seem to be large.

Table 5.2: Correspondence Analysis Results By Number Of Peaks

	Weighted Kappa	Cronbach Alpha
Reference & Algorithm1	0.42	0.59
Reference & Algorithm2	0.28	0.44
Reference & Analysts	0.12	0.89
Algorithm1 & Algorithm2	0.60	0.75
Algorithm1 & Analysts	0.20	0.89
Algorithm2 & Analysts	0.28	0.92

5.3 Problem 1: Simulation of Cronbach's alpha

I simulate the Cronbach's alphas of 10 different comparisons using two different bootstrapping methods and obtained a total of 20 confidence intervals. Table 5.3 and 5.4 show all the comparisons and the corresponding confidence intervals. The results can vary slightly on each time you run the bootstrapping codes in R since the sampling procedure generates different random samples. Based on these results, I present two separate analyses: the comparison between the bootstrapping methods, and the comparison between algorithms.

Regarding the comparison analysis between the bootstrapping methods, the confidence intervals generated by the empirical bootstrap appear to be shifted higher since their lower and upper bounds are generally larger than the ones generated through the percentile method. The confidence intervals are very similar in size except for a few cases. The similarity between the two does not allow determining which one is a better method.

The interpretation of the comparisons analysis is also similar since we interpret the consistency based on a certain interval changing by a tenth (0.1). For example, both intervals of a comparison between reference and analysts in Table 5.3 are [.8893, .9838] and [.8548, .9550]. Both are interpreted in the same way as the interval is between good to excellent according to the interpretation rule, which states that

values between 0.8 and 0.9 indicate good consistency and values between 0.9 and 1.0 indicate excellent consistency. Despite little difference between the results, in this case, I refer to the percentile bootstrap method since it estimates more conservative confidence intervals with lower bounds.

Table 5.3: Confidence intervals of the Cronbach's alphas from various comparisons 1

	Cronbach's alpha	Empirical BS Confidence Interval	BS percentile Confidence Interval
Among Analysts	.922	[.8893 , .9838]	[.8548 , .9550]
Reference & Analysts	.890	[.8520 , .9509]	[.8267 , .9254]
Reference & Algorithm1 & 2	.718	[.6152 , .8879]	[.5438 , .8142]
Algorithm1 & Algorithm2	.752	[.6307 , .9811]	[.5372 , .8737]

Table 5.4: Confidence intervals of the Cronbach's alphas from various comparisons 2

		Cronbach's alpha	Empirical BS Confidence Interval	BS percentile Confidence Interval
Alg 1	analysts	.894	[.8557 , .9614]	[.8294 , .9313]
	reference	.593	[.4143 , 1.006]	[.2336 , .7796]
	reference & analysts	.880	[.8430 , .9477]	[.8135 , .9181]
Alg 2	analysts	.916	[.8875 , .9734]	[.8711 , .9440]
	reference	.443	[.1726 , .8984]	[-.0334 , .7519]
	reference & analysts	.897	[.8662 , .9516]	[.8465 , .9303]

Table 5.4 clearly shows the comparison analysis results of two algorithms. There are two values that are outside the regular Cronbach's alpha interval [0.0, 1.0]. The values below 0 mean that the test consistency is very poor or not applicable. In this case, however, the absolute size of the deviation is very small so it can be explained as an error from the estimation process. In the comparisons with analysts, and with reference & analysts, algorithm 2 has slightly larger values for the confidence intervals. However, the difference in the gaps between them are not so significant since the size of the differences are roughly 0.02 to 0.03. Also, the comparisons with

both reference and analysts should not be considered as a main reason to choose the one as a better algorithm since the high consistency between reference and analysts is highly likely to exaggerate the consistency among reference, analysts, and algorithm. The alphas must reflect a high consistency of the analysts and the reference peaks, which possibly affects the results. The lower bound of the percentile method of each algorithm with the reference peaks differs as algorithm 1 has 0.2336 and algorithm 2 has -.0334. Through this comparison of the intervals, I still think that algorithm 1 has more consistency with reference peaks because it has a significantly higher lower bound.

5.4 Problem 2: Missing Data Imputation 1

Since the original data contain a lot of missing values, I perform a random regression imputation as introduced in Chapter 4. In this section, I focus on consistency of the interpretation of the data based on the comparison of descriptive statistics between the original data and processed data. According to Table 5.5, V11 and V13 have the strongest correlations between attention as same as they have in the original data. However, V16 is replaced with V20 after missing values imputation. Overall, the correlations from the imputed data are lower than in the original data. The top three variables with the strongest negative correlations with attention are the same in the original data but just in a different order.

Table 5.5: Missing Imputation: Strongest correlations between attention and other variables

Positive			Negative		
Variable	Correlation	P-value	Variable	Correlation	P-value
V11	.6689	< 2.2e-16	V20	-.4555	2.247e-08
V13	.6244	1.332e-15	V12	-.3864	7.455e-08
V21	.4617	8.152e-11	V22	-.3149	3.389e-06

In Table 5.6, I present the pairwise combinations of the highest positive correlation values from the imputed data to check for multicollinearity. The combinations are identified as highly correlated if their correlation value is higher than 0.8, which is the same threshold used in the original data analysis in Table 2.10. Unlike the original data correlation analysis, there is only one correlation value over 0.8 from a pair of V12 and V20. The correlation value is again relatively lower than the original values, which shows that the imputed values weaken the strength of the correlations between variables. In particular, the correlation between V9 and V16 significantly decreased since it is the highest correlation combination in the original data but it disappears in Table 5.6. The original V9 and V16 includes 46 and 33 missing values out of 136, respectively, which means that more than 30% of the observations in this variable have been replaced through imputation procedures. Regarding the highest negative correlation values, there are two pairs whose absolute value of the correlations are above 0.8: V11-V20 and V3-V5. In a comparison with the original ones, their correlation values have not changed significantly. However, three other pairs whose correlation values are stronger disappear after the missing values imputation procedure. Most of the variables that disappear after missing imputation procedure contain a large number of missing values.

Table 5.6: Missing Imputation: Pairwise combinations of the highest positive/negative correlation values

	Pair	Correlation	P-values
Positive	V12 & V20	0.8825	$<$ 2.2e-16
Negative	V11 & V20	- 0.8140	$<$ 2.2e-16
Negative	V3 & V5	- 0.8014	$<$ 2.2e-16

Based on the comparison analysis between the descriptive statistics for the original data and the imputed data, in general, the imputation procedure presents a limitation of the data with a large amount of the missing values. Most of the variables that have only few missing values show great prediction in correspondence with

the original data, whereas, the variables with a lot of missing values show poor prediction compared to the original data. However, I still need to verify if the variables that change significantly affect the prediction of our final model.

5.5 Problem 2: Missing Data Imputation 2

The missing imputation procedure used in this paper is a random regression imputation. In this process, normal random variables are created and then used to impute missing values. To confirm if this missing data imputation procedure is reliable and compatible with the data, I need to check if the randomness from the imputation procedure is consistent with the multiple regression model results. In order to verify the stability of the model results, the imputation procedure is performed a hundred times. I saved the results from two chosen model selection methods: all possible subsets method by AIC and LASSO.

After the one hundred imputations for the model selection results from the all possible subsets method by AIC and the LASSO method, the results are presented in Table 5.7 and Table 5.8. The LASSO method selected fewer variables than the all possible subsets method by AIC. The LASSO method selected a total of 403 variables and each model selection subset varies in a range between 2 and 12. The all possible subsets method selected a total of 800 variables, excluding the intercept (β_0), and each model consistently contains 8 variables including β_0.

I consider the variables that were selected more than 50 times out of 100 times as reliable and significant variables for predicting the response variable. In other words, I consider the variables that are selected less than 50 times to be noise in the model selection procedures. I present two measures to indicate the noise levels in the model selection results. One is an intuitive way to calculate the noise percentage out of a total number of variables selected. The other one is a similarity probability, which

is the probability that the same variable selection happens twice in a row.

First, the noise is measured in an intuitive way, summing the number of variables selected over the 100 simulations, but only counting those variables selected less than 50 times. The LASSO method presents very stable results out of each randomly imputed dataset. The noise variables in the LASSO are only 73, which concludes that the noise rate is 18.1%. However, V10 is a confusing variable, which was selected 46 times. Although, it does not record more than 50 times, there is still a possibility that it would be selected more than 50% of the time with a larger number of imputations. If I do not count the V10 as noise, the noise rate decreases to 6.7%. Regarding the results from the all possible subsets method, it has more variables in each model as well as more noises compared to the LASSO method. Out of a total of 900 variables selected including the β_0, 309 selected variables were selected less than 50 times and are considered as the noise, concluding that the noise of the all possible subsets method is 34.4%. Based on the noise percentage, the LASSO is a more reliable and stable model selection method.

Second, the stability of the model selection is measured by the similarity probability. Two models agree for a given variable if they either both include the variable or both do not include the variable. The equation is given as the following:

$$p = \prod_{i=1}^{22} p_i^2 + (1 - p_i)^2 \tag{5.1}$$

where p = the counts of the variable has been selected over 100. According to the equation, the LASSO has $p = 0.0927$ and the all possible subsets method has $p = 0.00071$. It shows that the LASSO has much bigger probability to have the same variable selection twice in a row than the all possible subsets method, which means the LASSO's variable selection results are more stable with less noise.

Table 5.7: Missing Imputation: Model selection results by the LASSO method (Out of 100)

Variable	1	3	4	5	6	7	8	9	10	11	12
Counts	1	0	0	0	1	1	4	8	46	100	0
Variable	13	14	15	16	17	18	19	20	21	22	23
Counts	100	0	55	4	0	0	1	0	75	6	1

Table 5.8: Missing Imputation: Model selection results by the all possible subsets method by AIC (Out of 100)

Variable	1	3	4	5	6	7	8	9	10	11	12
Counts	46	54	36	35	16	28	89	17	36	100	16
Variable	13	14	15	16	17	18	19	20	21	22	23
Counts	100	0	89	35	8	23	0	0	59	13	0

5.6 Problem 2: Variables selected by the model selection method

Table 5.9 shows the variables that are selected by the all possible subsets method and the LASSO method. There are some variables in common such as V11 and V13 that both model selection methods agree on. V15 and V21 are significant and reliable variables that both methods agree on. However, V8 and V3 are significant only in the all possible subsets method. V10 is not considered to be a significant variable based on both methods since the all possible subsets method clearly shows it as a noise variable. As a conclusion, each method presents a difference in variable selections, particularly in terms of the number of variables in the subset. According to the nature of the shrinkage method, the LASSO provides a smaller subsets size as a predicting model with four variables, whereas the all possible subsets method contains more variables in the selected subsets with seven variables including the intercept.

Table 5.9: Missing Imputation2: Selected variables from two model selection results (Out of 100)

Variable	All possible subsets	LASSO
V11	100	100
V13	100	100
V15	89	55
V21	59	75
V8	89	4
V3	54	0
V10	36	46
V1	46	1

Regarding the interpretation of the model analysis for the final model, it can be summarized by the most significant predictors. By a unanimous consent, V11 and V13 are the most significant variables that specifically represent the following questionnaires: V11 is 'Top 2 - (Liking) Overall, I like the commercial' and V13 is 'Top 2 - The commercial is clever and entertaining.' Both models have V11, V13, V15, and V21. According to the meanings of V15 (Top 2 - The commercial is different from others) and V21 (Top 2 - Goodwill - Better), the attention score increases when the audience likes the commercial, feels entertained, perceives it as different from other commercials and better about the company (or the brand). It is very intuitive to understand that the attention score is attained when the audience likes the ad, but it is interesting to know that the attention score is directly related to whether this ad is considered different from other ads. Since people are already exposed to the ads for a long time, they get bored very quickly and lose interest in the typical ads styles that they were familiar with. To make a better advertisement, one that grabs the audience's attention, a fresh and creative idea for the contents of the ad would be essential. The ad should have a delightful and humorous atmosphere to attain higher attention from the audience. Something positive that can bring a better image or information about the company (or the brand of the ad) would also

be very helpful. In addition, the connection to the brand is important as shown as the all possible subsets choose V3 and V8, which both are brand-related variables. A high attention scored advertisement has to be good at reminding about the company (or the brand) through the advertisement.

In this thesis problem, we wanted to investigate whether the number of peaks (V1) performs well as the main explanatory variable in this prediction model or not. Unfortunately, neither of the model selection results contain the number of peaks as an explanatory variable. It concludes that the number of peaks does not have a significant impact to the attention score. For future research, I suggest the other variables from the flow of attention graph such as the average recall scores and the number of frames to verify if the flow of attention graph has any significant impact to the attention score.

Chapter 6

Discussion

This paper addressed the first problem by suggesting two detection algorithms. One is manually created based on the moving average and the other is from another research paper. Since the peak identification was, in general, very tricky and little details of the identification process were more based on the images rather than the score itself, I could not use statistical methods that are commonly used in peak detection processes such as time series methods or control charts. According to each condition that is required for a sensitive peak, the rules are created by hand to satisfy the conditions. While creating the rule, I considered not only the detection of the local maxima by numeric values from each score, but also the overall trend of each score so that the peak can be identified only when it is in a rising trend.

As a result, the manual identification algorithm, mainly called 'Algorithm 1' in the paper, achieved a moderate agreement with the reference peaks (Cronbach's alpha = .593) and a better agreement with the analysts' peaks (Cronbach's alpha = .894). 'Algorithm 2' was also very consistent. However, these two algorithms are still only based on the numeric values of the scores, not on the images. To make a better algorithm to detect the peaks as manual identification from the analysts, I

suggest machine learning techniques for future studies. Machine learning techniques include image detection techniques such as Naive Bayes classifier and Support Vector Machines. In this problem, adding the image detection to Algorithm 1 would be much more accurate and appropriate to identify the same peaks as the human analysts do.

The second problem was to predict the attention score based on twenty two regressors containing a lot of missing values. To address the problem, the missing imputation procedure and the model selection for multiple regression model were conducted. For the missing imputation procedure, a random regression imputation was conducted in order to keep the uniqueness of each variable, which improves the prediction at the end. Then, the all possible subsets method by AIC and the LASSO were performed to choose the most fitting subsets to predict the attention score. The LASSO method provided us the simpler model only with four consistently significant variables out of a hundred imputations. In the other hand, the all possible subsets method by AIC consistently suggested nine variables every hundred times, but with a different selection of variables each time. Before using the LASSO method, I had a few questions about the capability of the LASSO method since there is a controversy between LASSO advocates and opponents. For example, in practice, the variable selection results by LASSO often provide different results compared to other methods such as all possible subsets and stepwise methods.

Through the results from this study, I learned that the LASSO method obviously favors the simpler model and tends to get rid of the highly correlated variables except for the one variable that has the biggest impact in the correlated group. In addition, LASSO provides us pretty stable results for variable selection, even with missing values in the data. In conclusion, even though the datasets were imputed with random variables, the variables selection for the regression model was relatively stable in both model selection methods. However, it is still hard to determine the methods' stability since this research only had 100 iterations to simulate the random

imputed datasets in this paper.

The limitation of the research in the paper is presented in several ways. The data contained a few limitations for conducting the analysis. First, the size of the sample was small. Second, since the data were coded by analysts, it is relatively subjective, which means it is not unbiased. Regarding the second problem, there were twenty two explanatory variables while the sample included 141 individuals. The twenty two variables had a high degree of multicollinearity, which was not avoidable since each variable represents each questionnaire in the survey interview. Multicollinearity interrupts the model selection procedure and the entire process of the regression analysis.

For future research, more than 100 iterations should be conducted to see how stable the model selection results are based on the random regression imputation datasets. Since the random regression imputation procedure required twenty imputations per each variable, I could not run more than a hundred times to see the results using R program in a regular computer. In future research, the number of iterations and the missing imputation variables in the model can vary across cases to see how the stability changes by the number of iterations and the missing imputation variables. Probably, these comparison analyses on stability of the missing imputation data can be conducted by other model selection methods as well.

Multiple imputation by random regression with LASSO model selection method performed in this paper can be possibly used in other research as well when the data contains a large amount of missing values. Based on the results that this paper provided, a combination of multiple random regression imputation procedure and the LASSO model selection method gives a pretty stable and reliable result. Particularly, I highly recommend this combination to future survey responses research since the data often contains some missing values and multicollinearity between the variables in the survey responses data. This multiple imputation by random regression is very

applicable to many other datasets as well since it is pretty flexible with the number of variables or missing values.

References

Ameritest. Tv testing. `http://ameritest.com/products/tv_test`, Accessed Feb 2, 2015.

Aurélie Boisbunon, Stephane Canu, Dominique Fourdrinier, William Strawderman, and Martin T Wells. Aic, cp and estimators of loss for elliptically symmetric distributions. *arXiv preprint arXiv:1308.2766*, 2013.

Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

Jose M Cortina. What is coefficient alpha? an examination of theory and applications. *Journal of applied psychology*, 78(1):98, 1993.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Vladimir Filkov, Steven Skiena, and Jizu Zhi. Analysis techniques for microarray time-series data. *Journal of Computational Biology*, 9(2):317–330, 2002.

Darren George. *SPSS for windows step by step: A simple study guide and reference, 17.0 update, 10/e*. Pearson Education India, 2003.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Suja R Nair. *Marketing research*. Himalaya Publishing House, 2009.

Gafar Matanmi Oyeyemi, Eyitayo Oluwole Ogunjobi, and Adeyinka Idowu Folorunsho. On performance of shrinkage methods–a monte carlo study. *International Journal of Statistics and Applications*, 5(2):72–76, 2015.

Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online Submission*, 2005.

Cosma Rohilla Shalizi. Advanced data analysis from an elementary point of view. *URL: http://www. stat. cmu. edu/cshalizi/ADAfaEPoV/13*, 24, 2013.

Simon Sheather. *A modern Approach to Regression with R*. Springer Science & Business Media, 2009.

Luke Vale, Jonathan Silcock, and John Rawles. An economic evaluation of thrombolysis in a remote rural community. *BMJ*, 314(7080):570, 1997.

Nancy Wagner. How much does television advertising really cost?, 2008. URL `http://smallbusiness.chron.com/much-television-advertising-really-cost-58718.html`.

Charles E Young. *The advertising research handbook*. Ideas in Flight, 2005.

Charles E Young. Ad response tests show how attention connects to memory. *ADMAP magazine*, 1, 2009.