A Mixed Model Approach to Conjoint Analysis

by

Tanita Cronjé

Submitted in partial fulfilment of
the requirements for the degree

Master of Science (Mathematical Statistics)

In the Faculty of Natural & Agricultural Sciences
University of Pretoria
Pretoria

May 2014

Declaration

I, Tanita Cronjé, declare that the dissertation, which I hereby submit for the degree Master of Science (Mathematical Statistics) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE:.....

DATE:.....

Acknowledgements

I wish to express my appreciation to the following people who made this dissertation possible:

• To my three amazing supervisors, Dr. FHJ Kanfer, Prof. M. Arashi and Mr. SM Millard, who's willingness to share their knowledge and passion for research and statistics encourages many students to pursue a career in statistics, every single day.

• To my amazing Head of Department, Prof Andriëtte Bekker, who always believes in her students and who's constant encouragement and motivation is inspiring.

• To the Department of Statistics at the University of Pretoria who supported me during my postgraduate studies and STATOMET who's funding has helped me immensely to pursue my dream career in statistics.

• To my colleague, co. masters student and friend, Johan Ferreira. We both pursued this adventure in statistics, loving every moment. I could not have imagined this journey without you in the office across the hall.

• To my Mom and Dad, you taught me to follow my dreams and do my best. I am who I am because of you. Then last, but not least, Roelof Schoeman, who's shared passion for knowledge and hard work inspires me everyday.

Summary

In this decision-driven era, it has become vital for modelers to efficiently model consumer choices and preferences (from a marketing perspective for instance). Conjoint analysis is a known method which has been used to perform such analyses. A mixed effects model is proposed to perform a conjoint analysis with normal responses, illustrated by an application of modeling respondent's preferences to different industrial detergents. The proposed model allows for predicting how observed attributes (which describes a product in terms of its characteristics and features) of decision makers and choice options, influence decisions. Inference regarding the parameters of the proposed model with a normal distribution is discussed in the mixed effect conjoint setting. Extensions of this model, regarding Bayesian prior selection are also discussed.

Contents

Chapter 1

Background

"The aim of conjoint analysis is to evaluate the relative importance of attribute levels by means of a decompositional approach where only the global preference is known."
- Green and Srinivasan (1978)

1.1 Introduction

In the early 1960's, scholars of psychology started exploring the nature in which people make decisions, specifically studying consumers product preference and simulating consumer's choice. Customers will often compare several products and make subtle trade-offs, or compromises, to choose an optimal product from those that are available. The idea of conjoint analysis was introduced by Luce, a Mathematical Psychologist, and Tukey, a Statistician, in their 1964 paper "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement" which described their research on the psychology of decision-making. The consumer-oriented idea and developments in conjoint analysis was introduced in the 1970's by Green and Shrinivasan who jointly published an overview paper in 1978 called "Conjoint Analysis in Consumer Research: Issues and Outlook". It was in this paper that they reported some related issues as well as new technical developments and applications in this area, such as a number of theoretical contributions by Krantz (1964) and Tversky (1967), and algorithmic developments by Kruskal (1965), Douglas (1964) and Young (1969). By the end of the 1970's the technique became more sophisticated and well-known as conjoint analysis, the method that allows us to understand how customers compare and assess different services and products. Previously, conjoint analysis was mainly used in marketing research but today the technique has found new

applications in fields such as environmental economics, transportation research, scanner data analysis, health-care, telecommunications etc. (Gustafsonn et al., 2007). Recent developments includes models like Choice Based Conjoint Analysis (Damaraju et al., 2011), Adaptive Conjoint Analysis (Cunningham et al., 2010) and Hierarchical Bayesian methods (Wakefield, 2013).

1.2 Motivation and contributions

Conjoint modelling uses statistical techniques to measure an individual's preference on a set of options that can be described by their levels and associated parameters. This study aims to investigate the traditional methods used in conjoint analysis studies from the initial stages of designing a conjoint methodology, up to validating the obtained results. This study will also investigate how the inclusion of random components to the traditionally fixed effects model can lead to a more in-depth understanding of heterogenous relationships, as it can explain that some of the variation in the model can be due to differences in respondents' preference structures. As a wide array of fields exists where these conjoint analysis ideas can be applied, it is expected to have a range of relevant and significant applications. This will be done by investigating current methods that are employed in the actual market place and how these methods can possibly be extended to gain more information about the relevant population. In addition, a Bayesian approaches will also be considered. Concluding remarks will address future directions and developments that one can achieve in analyzing conjoint data.

1.3 Dissertation outline

- In Chapter 2 the conjoint analysis methodology is considered. Some important conjoint analysis terms are defined and related to more standard statistical terminology. Then the process, from the initial stages up to the validation of the predictive model, is discussed. Some of the main experimental designs, data collection methods and approaches are briefly discussed. It is important to understand what conjoint analysis entails, to better understand the relationship that these conjoint analysis methods have with the statistical world.

- In Chapter 3 the use of mixed effect models are introduced to conjoint analysis studies. The chapter commences by considering fixed and random effect models and the role that they play in conjoint studies. The chapter includes both theoretical concepts, practical applications, and examples.

- In Chapter 4 a Bayesian perspective is introduced to conjoint analysis. These approaches will include both theoretical concepts, practical applications and, examples.
- Chapter 5 will discuss the conclusion and the findings of the study and a short description of possible future directions that could be considered.

Chapter 2

Choosing is a Way of Life

"*We are our choices*"
- Jean-Paul Sarte

2.1 Introduction

People make thousands of decisions each day, mostly unconsciously, ranging from simple decisions to more difficult ones that require strenuous thought. Evaluating these individuals' choices can lead to a better understanding of how each individuals', as well as the group of respondents', preferences can be measured. Understanding how the changes in the characteristics of alternatives affect the preference of each individual is important, in the many fields where predicting human choice is of interest. Such fields include psychology, economics, environmental science, and many more (Damaraju et al., 2011). An important point to take into consideration, is that individuals' choices can be influenced by several factors like, habit, inertia, experience, advertising, etc. (Louviere et al., 2000). These influences reflect the turbulent nature of choice outcomes and motivates the investigation of models, which can explain individual and aggregate choice behavior, as well as predicting behavioral responses to changing opportunities. During the rest of Chapter 2 some terminology that is well known in the conjoint analysis areas will be mentioned, and the main steps to performing a conjoint analysis study, such as the general conjoint analysis procedures (section 2.2), the designing of the stimuli (section 2.3) and the different preference models that can be incorporated in a conjoint analysis study (section 2.4), will be discussed. Subsequently, in section 2.5 and 2.6 different data collection techniques will be briefly introduced as well as estimation procedures that can be implemented, in these conjoint analysis designs.

2.2 An overview of conjoint analysis

The term conjoint is based on "to join" or "become joined together" and symbolizes how people weigh up several factors at the same time (Gustafsonn et al., 2007). Conjoint analysis, also known as "trade-off" analysis, is a widely used technique that measures the trade-offs that people make when choosing between products or services. The different elements that the products consist of, as well as the responses from the different customers, are the observed variables in the conjoint study. The unobserved variables, that will need to be estimated, consists of the preference ratings (also known as the part-worth utilities) that the respondents assign to each one of the components that the product consists of. These preference levels, which indicates how much or how little a respondent likes a certain components of a product, can be estimated using statistical techniques. This method can also be used to predict consumer choices for products not yet developed and investigates the influence that this new product will have in the actual market place. Through conjoint analysis, researchers will be able to determine which factors of certain products, trigger the most interest as well as which combinations of these factors are preferred by the general population of customers.

2.2.1 Terminology

Some of the key terms that are used in conjoint analysis designs include (Hair et al., 2006):

Attributes or Factors

The attributes or factors of a certain product describes the product in terms of its characteristics and features. For instance, a tube of toothpaste can be described in terms of size, brand, flavour, price, etc. These attributes take on the form of the explanatory variables in the conjoint analysis model.

Levels

The attribute levels (or factor levels), describe the alternatives within each attribute. These levels express the possible options that the market offers for each of these attributes. When considering the toothpaste example, it is quite clear that there is more than one option that a customer will be faced with when he/she is standing in front of the toothpaste isle. The levels of the mentioned attribute include the different sizes of toothpaste, the different brands that are available, the many different flavours and then of course the different price levels. Each attribute consists of two or more levels, but typically the number of levels will not exceed 5 or 6. These levels are nonmetric values that describe the different options within an attribute. If the levels are metric (for example the price of the toothpaste) the level will be reduced to a small number of nonmetric levels.

Treatments, profiles or stimuli

A combination of one level from each attribute, which forms a full product descriptions, is what is referred to as either a treatment, profile or stimuli. By combining a different set of levels from each attribute, the researcher can create many such product profiles. It is these profile descriptions that will be represented to the respondents for evaluation.

Utility

Hair et al., 2006, defines utility as an individual's subjective preference judgement representing the holistic value or worth of a specific object.

Part-worth

The part-worth values are the preferences obtained from the conjoint analysis which represent the preference or utility associated with each level of each attribute used to define a product.

Decompositional model

A class of models that decomposes the individual's responses to estimate the part-worth values associated with each attribute level by relating the responses to the known attributes of the specific product.

Traditional conjoint model

Methodology that employs the classical principles of conjoint analysis in the conjoint task, using an additive model of consumer preferences and pairwise comparison or full-profile methods of presentation.

To further illustrate these terms, consider the following hypothetical example that describes a conjoint analysis setup that measures the consumer preference of respondents to a small set of motor vehicle configurations:

Illustration 2.2.1 Motor vehicle attributes and levels

Factors (Attributes)	**Levels**
Size	1. 2 Door 2. 4 Door 3. 5 Door
Brand	1. Toyota 2. Ford 3. Audi
Colour	1. Grey 2. Red 3. Blue
Price	1. R220 000 2. R260 000 3. R380 000

For this example there are $3 \times 3 \times 3 \times 3 = 81$ possible stimuli, one of which is a 2 door blue Audi with a price of R380 000. After these attributes and levels have been identified, the interviewers will offer a range of realistic profiles to the respondents, who will then choose between the alternatives, either by rating or ranking the possible profiles. Through this process the respondents provide data from which the key characteristics of their purchasing decisions can be identified, using statistical modelling. In recent developments these decisions have extended beyond a shopper contemplating various brands of product categories. Some of these decisions may include deciding whether to vote, and if so for whom, physicians deciding on various treatment options, or rating items in a personality questionnaire (Mayden-Olivares and Böckenholt, 2009). All of these choices can be expressed in an ordinal or continuous fashion. Since consumers evaluate the value of an object (real or hypothetical) by combining the separate amounts of part-worths provided by each attribute, conjoint analysis studies will use the same concept. By combining the individual part-worths calculated for each attribute's level the overall preference level for each stimuli, for each respondent, may be calculated. The three primary outcomes that almost all conjoint analysis procedures deliver is:

1. Conjoint part-worth utilities

For each respondent, conjoint analysis procedures deliver a set of part-worth utilities that describe the value that the respondent placed on each attribute level. These estimations on an individual level can be very valuable to the researcher and the researcher can even

use these individual level part-worth utilities to classify the different respondents into different market segments. Conjoint analysis can also determine these utilities at an aggregate level to evaluate the overall importance that the respondents placed on each attribute level.

2. Attribute importance

Since conjoint studies, in general, are more interested in the aggregate estimations that will illustrate the preferences over the whole sample of respondents, the attribute importance plays a vital role. These calculations specify the relative importance of each of the attributes over the whole sample of respondents. Through these values, the researcher can investigate which attribute plays the most important role when a respondent chooses between the different alternatives and which attribute has the largest influence on the respondents decisions. The attribute importance is calculated by dividing the distance between the best and worst level of one attribute, by the distance between the best and worst levels of all the attributes combined. These calculations will be considered in more detail in Chapter 3.

3. Market simulator

The third outcome has a variety of forms but they all share the same basic functionality. The results obtained through a conjoint analysis study can be used to predict how the market will react to certain changes. By constructing hypothetical products and services, the researcher can investigate and assess how these products and services perform in the hypothetical markets.

Over the passed few years, the use of conjoint analysis has accelerated due to the improvement of computer capabilities. Software packages can integrate the entire process, from generating the combination of independent choices to creating choice simulators, for predicting consumer choice. The general conjoint analysis procedure and steps can be represented by Illustration 2.2.2

Illustration 2.2.2 Steps to performing a conjoint analysis

Step 1:	Define the objectives of the study by describing the research problem.
Step 2:	*Choose a conjoint methodology by means of a experimental design:* • Select and define the attributes and their levels that will be in the study • Design a set of profiles by combining one level from each factor to form complete sets of profiles, which will represent true product descriptions. • Specify the basic preference model form that each attribute will follow. The respondents can convey their preference by either using rank orders, rating scales, paired comparisons or category assignments
Step 3:	*Data collection*: This step defines the way in which the profiles will be represented to the respondents for evaluation, and the way in which the preferences will be indicated by the respondents.
Step 4:	Evaluate the model assumptions in the form of an exploratory data analysis.
Step 5:	*Select the estimation method*: Some of the methods that can be used to estimate the respondents utilities include: MONANOVA, PREFMAP, LINMAP, Multiple regression, Logit and Probit.
Step 6:	Evaluate the model goodness-of-fit by assessing the reliability and the predictive accuracy.
Step 7:	Interpret the results and investigate the relative importance of the attributes.
Step 8:	Validate and apply the conjoint analysis results.

2.3 Design of the stimuli

When the objective of the study is in place, and the researcher knows what research problem needs to be investigated, the conjoint methodology can be constructed. The design of the stimuli describes the process from the beginning stages of choosing the relative attributes and determining their levels, to setting up the full profile descriptions, or stimuli. This task may be more complicated than originally expected. This very subjective step of selecting the attributes, investigating the range of each attributes' levels, constructing the stimuli and ensuring that the number of stimuli is a manageable size, is not an easy task. However, this is described by some researchers as the most important step in a conjoint analysis study as it ensures that the respondents do not report unreliable preferences, due to an information overload, while still keeping the estimation procedure reliable. The decision of which attributes and attribute levels to include in a conjoint analysis study, is certainly centered around the objective of the study, and what the researcher hopes to gain from exploring the respondents' preferences.

2.3.1 Selecting the attributes and their levels

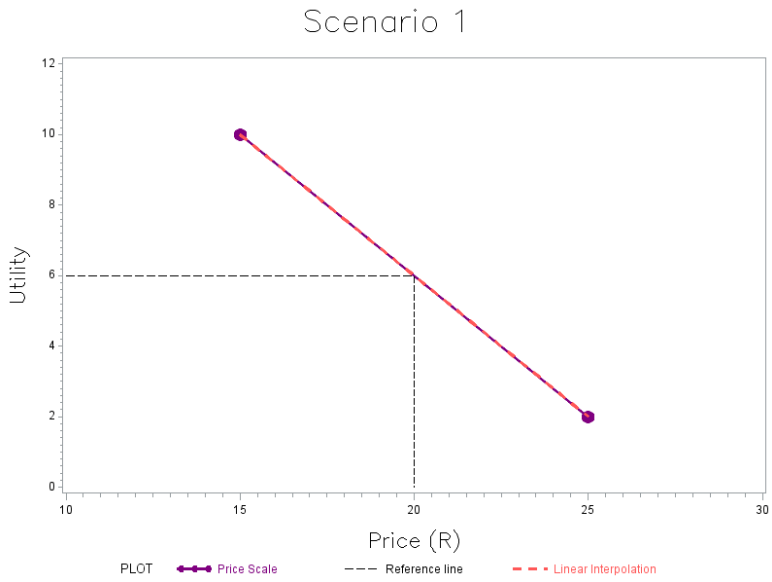
In the initial stages of a conjoint analysis, the researcher first needs to determine the number of stimuli to be included in the study, the different attributes to be included, and the range of these attribute levels. Obtaining the balance between the number of responses needed to ensure robust statistical estimation of the model, while still considering the impact that the number of profiles that will need evaluating will have on the respondents, is as much an art as a science. Some researchers believe that the "sweet spot" is between 6 to 8 attributes, but there has been studies which resulted in great success with more than 8 attributes (Grover and Vries, 2006). The realization is that as the number of attributes, and their associated levels, increase, the number of full profile descriptions that the respondents will need to evaluate, will also increase. Thus the researcher can start by eliminating unnecessary attributes and levels by just considering the following question: "Do we want respondents to carefully consider all the attributes in a study if they would not do the same in their real world behavior?"

When deciding which attributes and levels to include, it is of vital importance to consider the business decisions that will be made using the conjoint analysis data. By using the conjoint analysis procedure a researcher can also create hypothetical products and market scenarios to simulate the expected market reactions to these new products. Other studies explicitly introduce their competitions' attributes and levels into the study to simulate the impact that the business' components might have on their product. It is also of vital importance that the attribute levels appear in random combinations with

others. If a certain brand of a product is consistently seen with high prices then this will almost certainly result in low utilities for the products associated with the certain brand.

For some attributes like "price" and "speed" the attribute levels can be measured on a continuous scale rather than at discrete points or levels. This quality of these attributes can have the advantage that by only incorporating a few points on this scale into the study, the number of levels are limited, but by using linear interpolation any other value between the smallest and highest values on the scale that was tested, can be estimated. The choice of endpoints also needs great care since the researcher will be able to interpolate between these points but they will not be able to extrapolate outside of these bounds. Remember that these estimated utilities are not measured utilities, so care needs to be taken as the results will not always be stable (Grover and Vries, 2006). Consider the following example:

Example 1 *When considering the attribute price, in our simple toothpaste example the researcher should include price levels that he/she would suspect to have a great importance when the respondents evaluates the different full product descriptions. If the researcher only evaluate two price levels, say R15 and R25 then by using linear interpolation he/she can estimate the utility of say a price level of R20. If the researcher suspects that there will be a significant change in price level between these two values it will be advised to rather include the level of R20 to more accurately evaluate the price scale. Consider the following three scenario's. Suppose the researcher only tested the two price levels R15 and R25.*



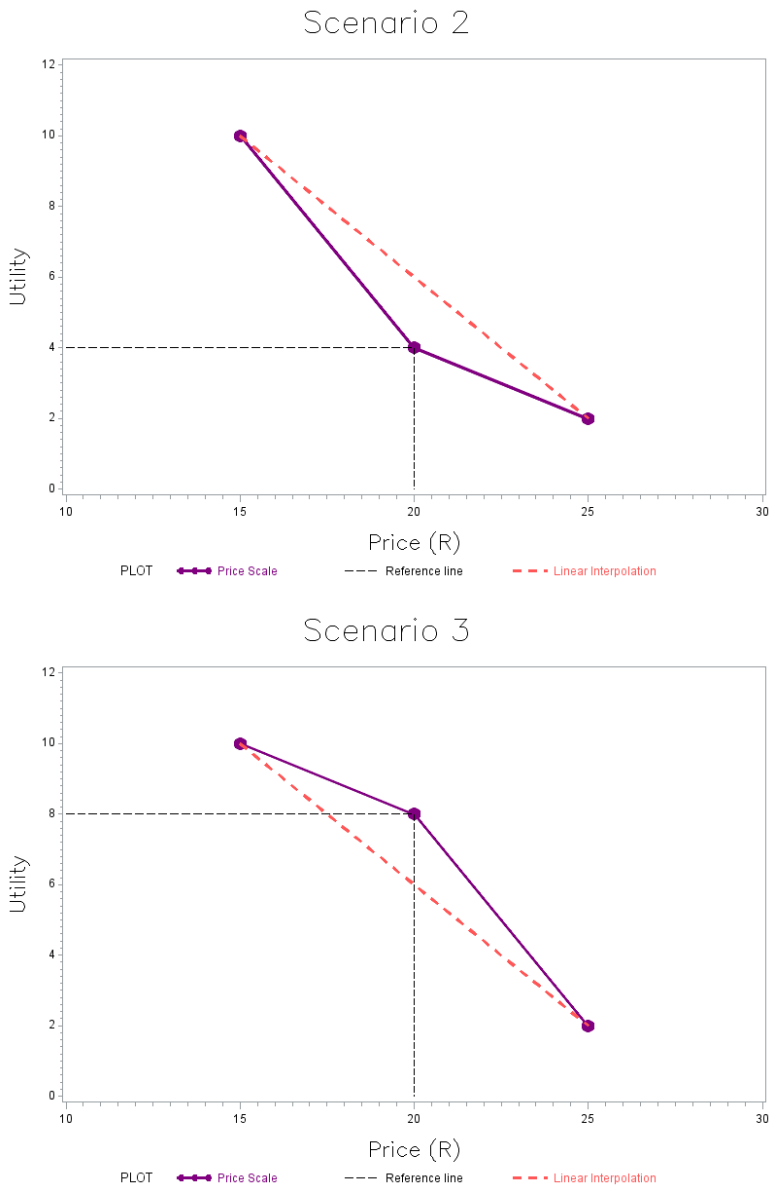


Figure 2.2.1 *Linear interpolation of the metric attribute, price.*

In Scenario 1 the estimated utility value, for the price level of R20, would then have been correctly estimated using linear interpolation. If however the utility value of the price level of R20 was more accurately represented by Scenario 2 or 3 then failing to test the attribute level at R20 would have lead to significant error in the estimation.

♦

2.3.2 Construction of the stimuli

When combining the different attributes and their levels, the researcher will obtain a collection of stimuli which usually represent a complete set of products. The main

challenge which the researchers are faced with in conjoint studies, is the large number of full profile descriptions which the respondents have to consider. Consider once again Illustration 2.2.1. If the researcher were to add another attribute with 3 levels then the number of full profile descriptions would increase to $3 \times 3 \times 3 \times 3 \times 3 = 243$. Clearly this becomes a permutation curse. This leads to experiments which take longer to complete and are more expensive to facilitate. To ensure that the estimated parameters lead to a true and accurate representation of the individuals' preferences, it is vital to ensure that the different respondents remain interested in the questionnaires and yet are not hampered by an information overload. By restricting the number of full profile descriptions, the researcher can ensure that the results reflect the individuals' true preferences.

The *Fractional Factorial Design* is the most commonly used procedure to minimize the number of full profile descriptions. This method conveys almost all the same information which a full profile approach would, but uses fewer trials. This allows the researcher to obtain information about a greater number of factors, while implementing a small number of runs (Halloand and Cravens, 1973). In real world scenarios a fractional factorial design is often necessary as it does not often happen that the design will be small enough to effectively use a full factorial design. The information gained by the interaction with the different variables, and the fact that the method may indicate which factors are most important, are some of the advantages of using this method. The *Orthogonal Fractional Factorial* design which reduces the number of combinations and still maintains orthogonality is preferred by most researchers since it can estimate the parameters of interest in a more efficient way. Louviere et al., 2000's Stated Choice Methods (Chapter 4) can be consulted for an overview of different experimental designs and fractional factorial methods.

2.3.3 Presentation of the stimuli

Once the stimuli are designed, the way in which they will be presented to the respondents should be considered. The main objective in presenting the different stimuli to the respondents, is to do it in the most realistic way possible. By representing these product descriptions in an easily understandable way, the respondents will not waste time trying to understand what the product entails. By using the presentation of the different profiles in an effective way, the respondents will be able to evaluate more profiles before losing interest. The order and position of the stimuli, as well as the particular attributes can also influence the measure of importance. Some studies advise that the researcher randomize the order of the stimuli presented to the respondents to eliminate this possible influence (Green and Shrinivasan, 1978). The presentation of the stimuli can be classified into

three main categories:

1. Verbal descriptions

Verbal descriptions include presenting the stimuli during face-to-face interviews or by telephonically interviewing respondents. Using this method generally leads to good results, but external influences should be considered and removed. Some of these could include, an interviewer who might knowingly or unknowingly force an opinion onto the respondents, or respondents who choose not to respond truthfully out of fear of what the interviewer might think. This problem typically occurs in sensitive questions such as asking for the respondents average income etc.

2. Paragraph descriptions

Paragraph descriptions have been known for providing more realistic and complex descriptions to the respondents. By describing the different products in such a detailed way, any uncertainties a respondents might have, can be eliminated. Some studies have also used this technique to test advertising claims (Green and Shrinivasan, 1978). Even though this process is known for its detailed descriptions, these can lead to very lengthy questionnaires or surveys, which will then result in the number of descriptions needing to be minimized.

3. Pictorial description

The use of pictorial descriptions have increased especially in the use of online questionnaires and surveys. This method minimizes the risk of an information overload since it conveys the product easily and gives a quick clear indication of the product being represented. Although this presentation of the stimuli can become costly and time consuming for the researcher to set up, with modern computer capabilities this should not be a limitation. A typical example of the effectiveness of this approach can be illustrated by the following example obtained from an online survey used to investigate the preference of respondents to different ice cream related options (Ajjan Associates (n.d.)):

Illustration 2.3.1 Ice cream conjoint analysis pictorial description

2.4 Preference utility models

Conjoint analysis is only a simple case of the broader field of preference measurement, and is used mainly for quantitative measures (Oded et al., 2008). In conjoint analysis the respondents' preference is measured by combining the preferences obtained for each of the attribute levels by using a preference utility model. This model is defined as a mathematical formulation that defines the utility level for each of the attributes. It is important to note that the form of the preference model is the same for all individuals in the sample, but the parameters that are estimated may vary across the individuals. In practice, the following three models are mostly used to evaluate the respondents preferences to different attribute levels (Qualtrics, 2011).

Assume there are $p = 1, 2, ...t$ attributes that have been chosen and $j = 1, 2, ...J$ stimuli were used in the study design. For a given respondent, let x_{jp} denote the value of the level of the p^{th} attribute in the j^{th} stimulus. This value can be either a numerical value, such as a price value or a speed measure, or in the case of a categorical attribute levels, the value of x_{jp}, will be a dummy variable which indicates the inclusion of an attribute level. Remember that each stimuli will contain one level of each attribute hence the value of x_{jp} will change as the stimuli, j, changes and will also change as the attribute, p, changes. Consider, a hypothetical example with three attributes, each with two levels. The number of possible profiles, J, will then be $2 \times 2 \times 2 = 8$.

Attribute	**Level**	**Dummy coding**
Colour	1. Blue	1
	2. Red	-1
Price	1. R20	1
	2. R50	-1
Size	1. Small	1
	2. Large	-1

The 8 different stimuli will be:

Stimuli						
$j = 1$	Blue	R20	Small	$x_{11} = 1$	$x_{12} = 1$	$x_{13} = 1$
$j = 2$	Blue	R20	Large	$x_{21} = 1$	$x_{22} = 1$	$x_{23} = -1$
$j = 3$	Blue	R50	Small	$x_{31} = 1$	$x_{32} = -1$	$x_{33} = 1$
$j = 4$	Blue	R50	Large	$x_{41} = 1$	$x_{42} = -1$	$x_{43} = -1$
$j = 5$	Red	R20	Small	$x_{51} = -1$	$x_{52} = 1$	$x_{53} = 1$
$j = 6$	Red	R20	Large	$x_{61} = -1$	$x_{62} = 1$	$x_{63} = -1$
$j = 7$	Red	R50	Small	$x_{71} = -1$	$x_{72} = -1$	$x_{73} = 1$
$j = 8$	Red	R50	Large	$x_{81} = -1$	$x_{82} = -1$	$x_{83} = -1$

The respondents' preference rating, y_j, for the j stimuli will then be used to estimate the utility. The higher the utility value, the higher the respondents preference towards stimuli j. Note that no specific index was used for the respondents as such, as separate models are usually considered for the different respondents and adhoc aggregation are usually performed over respondents to obtain global analyses.

2.4.1 The Vector model

Some attributes, like price and travel time, have a linear continuous measure from level to level and can be represented by the *Vector model*. The Vector model assumes that an individual respondent's preference y_j for the j^{th} profile is given by

$$\begin{aligned} y_j &= \sum_{p=1}^{t} \omega_p x_{jp} \\ &= \omega_1 x_{j1} + \omega_2 x_{j2} + ... + \omega_t x_{jt} \end{aligned} \tag{2.1}$$

where ω_p denotes the individual respondent's importance weight for the p attributes. These weights can be estimated by using statistical techniques. A separate vector model can be build for each individual. Note that one weight is assigned to each attribute which implies that each level of a certain attribute will have the same weight. This model is represented by a positive linear function that assumes that the preference increases as the quantity of attribute p increases (or a negative linear function which assume that the preference decreases as the quantity of attribute p increases) (Qualtrics, 2011). If, for example, an attribute has four levels then these will be represented on the straight line. Figure 2.4.1 represents the form of the vector model, for an individual respondent, for the preferences of the different levels of attribute p while holding all the other attributes constant.

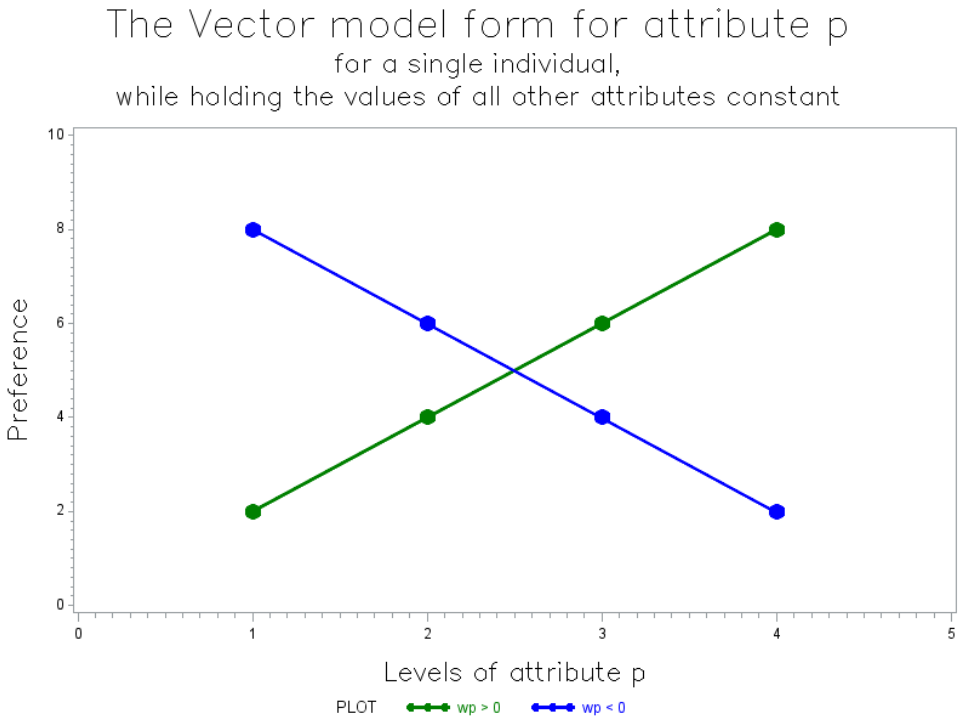


Figure 2.4.1 The Vector model form for attribute p

2.4.2 The Ideal point model

The *Ideal point model* represents an attribute's levels that follows a curvilinear function with an "ideal point" which indicates the optimum level of that attribute. This model is typically used with ordinal attributes associated with smell and taste (Qualtrics, 2011). Figure 2.4.2 illustrates this concept. Consider an attribute measuring sweetness: as the sweetness increases the preference increases until it reaches an optimum amount, whereafter the preference decreases, as too much sweetness may be less than desirable.

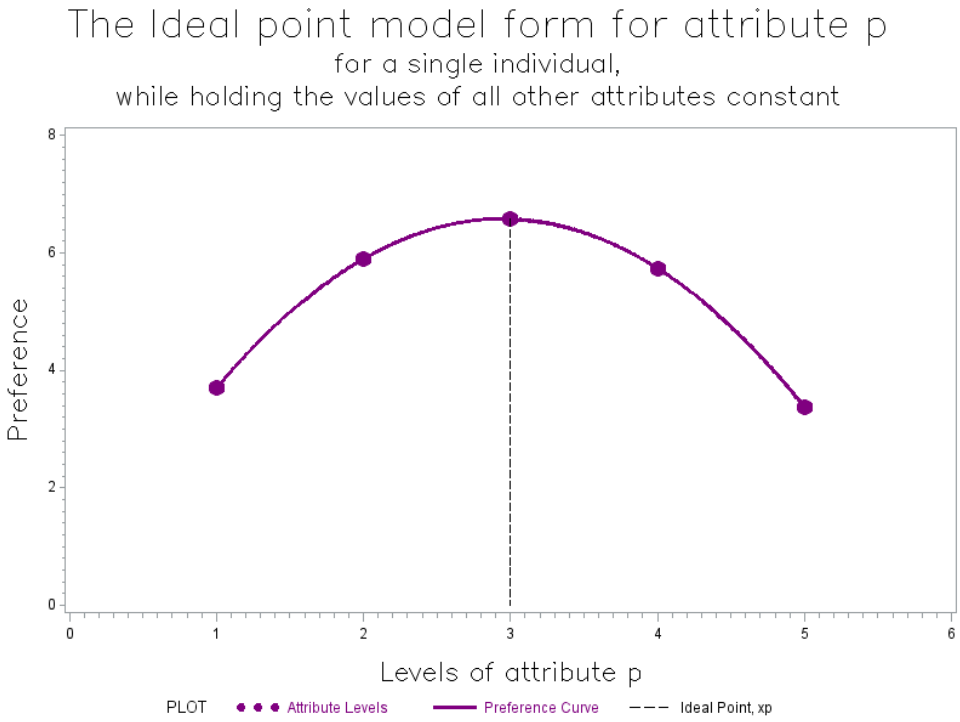


Figure 2.4.2 The Ideal point model form for attribute p

For the Ideal point model the preference of the j^{th} stimulus, y_j, is inversely related (opposite in sign) to the weighted squared distance d_j^2 of the location, x_{jp}, of the j^{th} stimulus from the individuals ideal point x_p. The weighted squared distance is denoted by

$$\begin{aligned} d_j^2 &= \sum_{p=1}^{t} \omega_p (x_{jp} - x_p)^2 \\ &= \omega_1 (x_{j1} - x_1)^2 + \omega_2 (x_{j2} - x_2)^2 + \ldots + \omega_t (x_{jt} - x_t)^2 \end{aligned} \tag{2.2}$$

and

$$y_j = -d_j^2 = -\sum_{p=1}^{t} \omega_p (x_{jp} - x_p)^2 \tag{2.3}$$

2.4.3 The Part-worth function model

The *Part-worth function model* is considered the most often used utility model and is represented by a piecewise curvilinear structure. As represented in Figure 2.4.3 the curve can be formed by a set of straight lines, that connect the point estimates of the part-worth utilities for the attribute levels (Qualtrics, 2011).

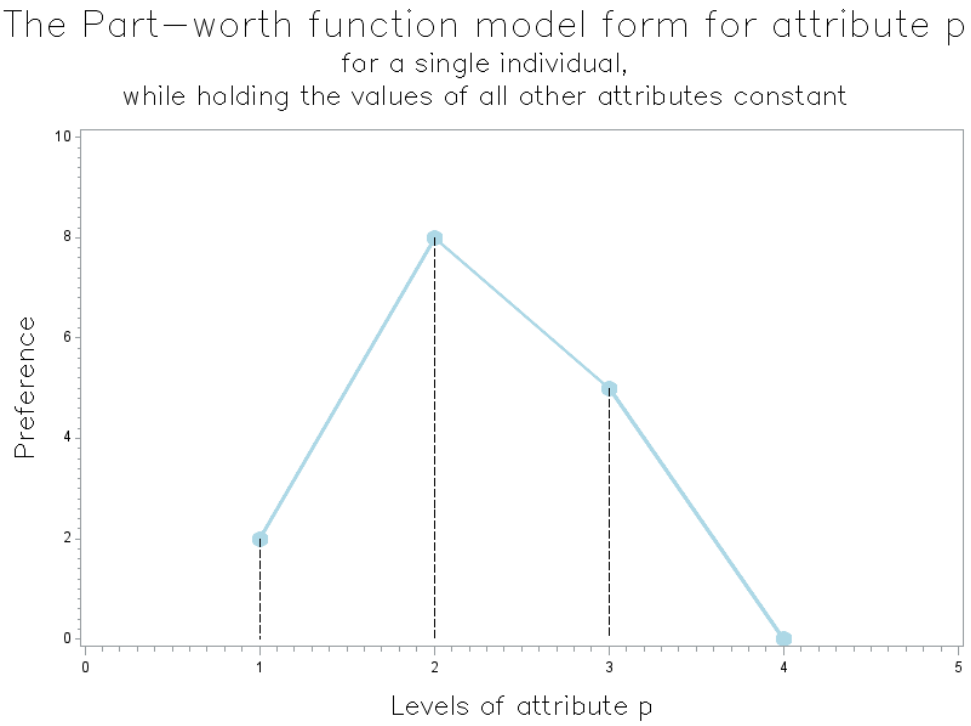


Figure 2.4.3 The Part-worth function model form for attribute p

Figure 2.4.3 clearly illustrates why the part-worth function model is so widely accepted. The visual representation is easily interpretable and clearly illustrates which levels of each attribute will be most preferred. The part-worth model is defined as

$$\begin{aligned} y_j &= \sum_{p=1}^{t} f_p(x_{jp}) \\ &= f_1(x_{j1}) + f_2(x_{j2}) + ...f_t(x_{jt}) \end{aligned} \tag{2.4}$$

where x_{jp} is the category level and f_p is a function denoting the part-worths of the different levels of, x_{jp}, for the p^{th} attribute (Green et al., 2001). Note that the part-worth function model estimates a part-worth utility value for each level of each attribute, where the vector and ideal point models only estimated one utility for each attribute. The most common way in which this model is typically used, which selects f_p to be a piecewise straight-line function for each attribute, is:

$$y_j = \sum_{p=1}^{t} \beta_{jp} x_{jp} \tag{2.5}$$

where β_{jp} is the part-worth utility for each level of each attribute for an individual respondent. The part-worth model is widely accepted, as it is easy to interpret and it allows the analyst to calculate the part-worths across each attribute level, in order to ob-

tain an overall utility of any profile compossible from these basic attribute levels. Another big advantage of the part-worth model is the flexibility of the model which allows different shapes for the preference functions along each attribute (Green and Shrinivasan, 1978). By setting $f_p(x_{jp}) = -\omega_p(x_{jp} - x_p)^2$ the model reduces to the ideal point model or by setting $f_p(x_{jp}) = \omega_p x_{jp}$ it simplifies to the vector model, a straight line for each attribute. This allows the researcher to effectively model the different levels of each attribute according to a different function.

Even though the part-worth model seems like the most attractive choice it comes at a cost. When considering these three models, regarding the number of parameters that need to be estimated in each case, the part-worth model requires the most. As the number of parameters increases, the reliability decreases. For the vector model only the p weights, ω_p, need to be estimated. For the ideal point model the number of estimations increases to $2p$ since the weights, ω_p, and the ideal points, x_p, should be estimated. When considering the part-worth model, the number of levels should now also be taken into consideration as this model estimates the utilities for each level of each attribute. Assume that there are q levels. Then the part-worth model will have $(q-1)p$ parameters that should be estimated. Figure 2.4.4 illustrates the relationship between the flexibility and the reliability of these three models.

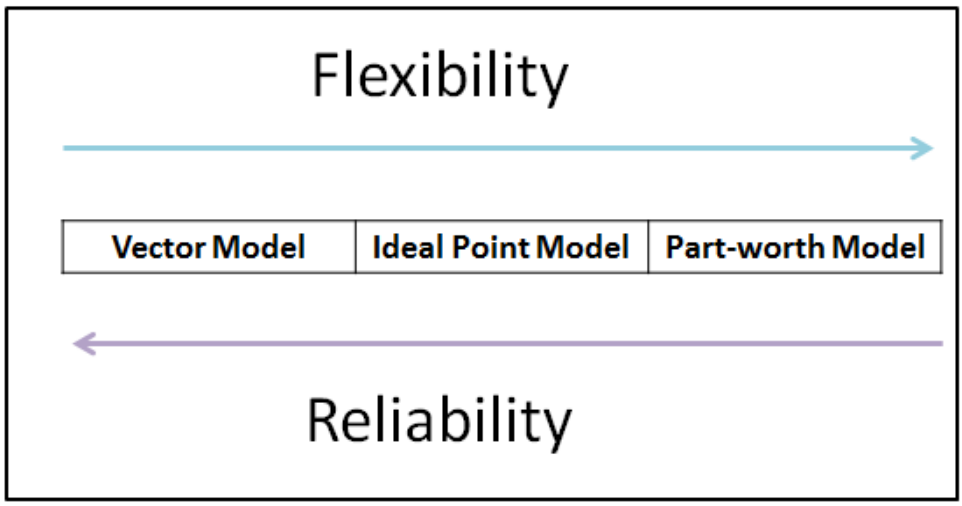


Figure 2.4.4 Flexibility and reliability of the different models

Conjoint analysis is referred to as a decomposition method, since the subjective responses are transformed into estimated parameters for each individual level. After the respondents illustrate their preference to a set of full profile descriptions, it is the analysts' job to find the part-worths for the individual attributes that is the most consistent with the respondents overall preference. Although parameters are estimated for each individual,

the conjoint analysis study is most interested in the aggregate implications of the multi-attribute utility structure and less concerned with estimation of the individual utility functions (Green and Shrinivasan, 1978). Conjoint analysis is usually carried out on an individual level because of the substantial amount of among-person variation in consumer preference, also referred to as heterogeneity in the model. Choosing the best preference model will depend on the type of product being studied, the number of relevant attributes, and the type of respondents (Green and Shrinivasan, 1978).

2.5 Data collection methods and designs

The data collection step is one of the most important steps for a successful analysis and it includes the choice of the design that will be implemented. The different methods measure preference in rank order or scaled evaluations, they can either be full profile or two factors at a time, and may include pre-assessments to better understand the respondent. With all of these differences to take into consideration it is clear that the data collection task will differ with each conjoint analysis study. To illustrate the different data collection methods, consider the following example obtained from a study of Green and Schaffer, 1991. The study considers different apartments constructed from six attributes, each with three levels. Illustration 2.5.1 represents these attributes and their corresponding levels:

Illustration 2.5.1 Apartment study's attributes and levels

Attribute	**Level**
Walking time to class	1. 10 minutes 2. 20 minutes 3. 30 minutes
Noise level of apartment	1. Very Quiet 2. Average 3. Extremely Noisy
Safety of apartment location	1. Very Safe 2. Average 3. Very Unsafe
Condition of apartment	1. Newly Renovated Throughout 2. Renovated Kitchen 3. Poor Condition
Size of living/dining area	1. 24 by 30 feet 2. 15 by 20 feet 3. 9 by 12 feet
Monthly rent	1. \$225 2. \$360 3. \$540

When considering all of these attributes with their assigned levels, the total number of different full profile descriptions of apartments that can be evaluated by the respondents, can be determined by multiplying the number of levels of each attribute. For this apart-

ment example the number of possible full profile descriptions will be $3 \times 3 \times 3 \times 3 \times 3 \times 3 = 3^6 = 729$. Clearly this is a very large number of apartment descriptions that will need to be evaluated by each respondent. The following unit will discuss methods that can be implemented to collect respondents' preferences towards certain products and services.

These different data collection procedures can also be used to reduce the number of full profile descriptions, in combination with fractional factorial designs, while still obtaining accurate results that give a true representation of the consumers' real life decisions. The two main models that originated with conjoint analysis, are the full profile approach and the two-at-a-time approach. During the development of conjoint studies over the past few years, several new models and approaches have been designed, some of which are simple variations of these two main models. Another conjoint method that is based on discrete choice modelling was later introduced and has since became one of the most widely used conjoint analysis techniques, namely Choice Based Conjoint (CBC) analysis (Damaraju et al., 2011). Some of the main data collection methods used in conjoint studies will now be discussed.

2.5.1 Full profile conjoint model

The full profile method is the most fundamental approach utilized for measuring attribute utilities and is still widely used today. This method gathers respondents evaluations by presenting full profile descriptions to the respondents. Each of these full profile descriptions consists of one level from each attribute included in the study. Different product descriptions are developed and presented to the respondents, which they then rank or rate to illustrate their preference or likelihood of purchase. For the rating based conjoint studies the respondents are asked to evaluate one full profile description at a time and rank each of these descriptions according to a scale from least- to most preferred. For the ranking based conjoint studies, the respondents are given a set of descriptions whereafter they are then asked to sort these descriptions from least- to most preferred.

It is important to note that large sets of product descriptions will result in many features and levels that will have to be studied. This can lead to an information overload which can easily weary respondents and lead to unreliable results. Since this method yields a large amount of information for each respondent it is best used with a moderate number of profiles. In a real world setup this will usually be the case, as it very seldom occurs that a study contains a small amount of attributes and levels. When considering the apartment example, it is clear that the 729 possible profiles will result in a tedious process for the respondents, who will have to evaluate all of these full profile descriptions. It is then easy to conclude that the preference data obtained from these respondents will

not be a true representation of the real life behavior of the individuals, as they will most likely not evaluate 729 different apartment combinations in one sitting. By controlling the attribute pairings in a fractional factorial design, the researcher can estimate the respondents' utility for each level of each attribute tested, using a reduced set of profiles (Qualtrics, 2011). Each respondent will then be faced with fewer judgements, but these judgments will be more complex. In the apartment example the authors used a fractional factorial design consisting of 18 profiles. These 18 apartments produced an orthogonal and balanced design that allowed them to estimate all levels of each of the six attributes (Qualtrics, 2011). It was these 18 profiles that were then presented to the respondents. By drastically decreasing the amount of profiles, the respondents gave more accurate preference measures, without any information of the different attributes necessarily being lost. The main advantage of the full profile approach is the fact that it gives a more realistic description of the stimuli than some of the other models that will be discussed. This method can also take into account the potential existence of environmental correlation between the factors in the real life stimuli (Green and Shrinivasan, 1978). Consider Illustration 2.5.2 which represents an example of a single full profile conjoint analysis design presented to the respondents in a rating based conjoint design.

Illustration 2.5.2 Full profile data collection method

Please look at the features present in the apartment and rate how likely you will be to rent this apartment.	
Apartment 1:	
Walking time to class	10 minutes
Noise level at apartment	Very Quiet
Safety of apartment location	Average
Condition of apartment	Renovated Kitchen
Size of living/dining area	24 by 30 feet
Monthly rent	$225
	Unlikely | ☐ ☐ ☐ ☐ ☐ | Likely

A variety of methods have been used to estimated the part-worth values for each of the attribute levels, with the ordinary least squares (OLS) regression approach the most widely used one. The other estimation methods that can also be considered are also based on a linear regression approach. These methods include latent regression and a hierarchical Bayes approach. These estimation methods will be discussed in more detail in section 2.6. For all three approaches the responses are the dependent variables and the attribute levels represents the independent variables. The different part-worth values

can then be obtained by modelling the responses as a function of the different utilities for each of the attribute levels. The three estimation methods only differ in how they deal with the preferences across respondents. This traditional conjoint analysis model will be investigated in more detail in Chapter 3 and some possible extensions will be considered.

2.5.2 Two attribute trade-off model

For this data collection method (one of the earliest methods used to measure consumer preferences) the respondent only needs to consider two attributes at a time in the form of a series of attribute by attribute trade-off tables. For each of these tables the respondent will be asked to evaluate their preference from most preferred to least preferred by ranking each of the different combinations of attribute levels in the trade-off table. These different combinations are represented by the selection of cells in the tables, each of which represents one level from each of the two attributes being considered in that trade-off table. Consider the following single attribute by attribute trade-off table for the apartment study. This is just one of a series of possible attribute by attribute trade-off tables that the respondents will need to evaluate:

Illustration 2.5.3 Two attribute trade-off data collection method

Consider the following two attributes viz. walking time from the apartment to class and the size of the apartment. Please enter your preference by ranking each cell from 1 to 9, where 1 is your most preferred choice and 9 is the least preferred choice.			
Size of the apartment / **Walking Time to class**	**24 by 30 feet**	**15 by 20 feet**	**9 by 12 feet**
10 minutes			
20 minutes			
30 minutes			

The two attributes considered here both have three levels each, hence the trade-off table will contain $3 \times 3 = 9$ cells. Since each table only considers two attributes at a time, it is easy to follow and will not be tiring to the respondents, as it reduces the information overload. If, however a study contains a lot of attributes that needs to be investigated, the process can be time consuming since each attribute needs to be evaluated against all the other attributes. It has also been found that respondents can lose their place in the table or adopt a stylized pattern just to get the job done (Qualtrics, 2011). The most important realization is that this process does not represent a real world situation. Customers' daily decisions aren't simply choices between two attributes, but rather decisions between full product descriptions. Some studies have also found that the respondents are unclear

about the role of the other $p-2$ attributes, that are not included in the attribute by attribute trade-off table. Consider for example Illustration 2.5.3. The only two attributes being considered in this trade-off table are the *size of the apartment* and the *walking distance to class*. A respondent may however argue that as the size of the apartment increases, so will the price. It is these uncertainties that will then result in estimations that will not be a true representation of respondents' preferences.

2.5.3 Choice based conjoint model

Discrete choice modelling grew from mathematical formulations developed by Luce (1959) and McFadden (1973). It was Louviere and Woodworth (1983) that integrated the mathematics of discrete choice modelling with conjoint analysis (Grover and Vries, 2006). Choice Based Conjoint analysis (CBC), also known as "discrete-choice conjoint analysis", is a form widely used in the conjoint analysis studies of today, since it replicates the way consumers would make decisions in a real world scenario. When a consumer finds himself/herself in front of the toothpaste shelf, they have to choose one of the many different toothpaste options, as opposed to ranking or rating the different options. CBC analysis is based on a full profile approach, but now the respondents are asked to indicate the product that they are most likely to purchase or choose. The respondent selects a single (or multiple) full profile stimulus from a set of stimuli (known as a choice set) instead of rating or ranking each stimulus separately (Hair et al., 2006). The fact that more than just two products can be evaluated at the same time, as well as the existence of the "none-of-these option" enables the respondents to make their decision in a realistic way. As it was stated by Qualtrics, 2011, CBC is known to produce excellent estimates, especially in regards to pricing, and the results are thought to accurately estimate the actual shopping behavior. A typical choice set for the apartment example can be represented by Illustration 2.5.4.

Illustration 2.5.4 Choice based conjoint data collection method

Please indicate which one of the following apartments you would most likely rent.			
	Apartment 1	**Apartment 2**	**Apartment 3**
Walking time to class	10 minutes	30 minutes	10 minutes
Noise level of apartment	Average	Very quiet	Extremely noisy
Safety of apartment location	Very safe	Average	Very safe
Condition of apartment	Newly renovated	Poor condition	Renovated kitchen
Size of living/dining area	9 by 12 feet	24 by 30 feet	9 by 12 feet
Monthly rent	$225	$540	$360
	☐	☐	☐

For this data collection method, different choice sets can be set up by combining either full profile descriptions, partial profile descriptions or menu-based descriptions. The menu-based description is when respondents are given a list of alternatives, much like a menu in a restaurant, and allowed to pick one, or several, items from the list. When estimating the set of coefficients or weights for each attribute level, a method is needed where the dependent variable is a probability rather than a continuous or interval scales variable, like in OLS. The statistical model that is often used for predicting and explaining choice probabilities is the logistic regression (Logit) approach. This method is conceptually similar to the OLS approach and can handle binary as well as multinomial choice situations. For the multinomial cases the model is referred to as the multinomial logistic model (MNL). A big disadvantage of the CBC approach is the inability to accurately estimate individual level models since, for the MNL and Logit approaches, the number of data points required to accurately estimate the part-worth values is very seldom available in conjoint studies.

2.5.4 Adaptive conjoint analysis model

The Adaptive Conjoint Analysis (ACA) method was designed to provide a survey process that is more engaging than the conventional approaches to conjoint analysis. By including these additional steps, the study can obtain more information than is typically available, which in turn can improve the estimation of the part-worth utilities and thus better predict the real world preferences (Cunningham et al., 2010). This method of conducting a conjoint analysis, relies on the respondents providing additional information, typically not part of the actual conjoint task (Hair et al., 2006). This pairwise rating based approach then adopts the concepts shown to the respondent, based on these previous answers. This process can be illustrated in three steps.

Step 1:

During the first step the respondents are asked to rank the attribute levels in order of importance. Illustration 2.5.5 represents how one of the attributes' levels can be ranked.

Illustration 2.5.5 Step 1 of the adaptive conjoint analysis method

Please rank the following features of the apartments from most to least preferred.
Condition of the apartment: ☐ Newly Renovated Throughout ☐ Renovated Kitchen ☐ Poor Condition

Step 2:

In the second step the respondents are then asked to indicate the importance of the differences in the attribute levels. Illustration 2.5.6 represents how the importance of one of these possible differences can be evaluated.

Illustration 2.5.6 Step 2 of the adaptive conjoint analysis method

If only these two possible conditions of the apartments where available, how important will the difference between these options be in your decision between the possible apartments?
Newly Renovated Throughout vs. Renovated Kitchen ☐ Extremely Important ☐ Very Important ☐ Somewhat Important ☐ Not at all Important

Step 3:

Once the first two steps are completed, the respondent will receive full profile descriptions in the form of paired comparisons. A paired comparison task might look like:

Illustration 2.5.7 Step 3 of the adaptive conjoint analysis method

Which apartment will you prefer?								
10 Minutes from class Average Noise Level Very Unsafe Renovated Kitchen 15 by 20 feet $540					20 Minutes from class Very Quiet Average Safety Newly Renovated 9 by 12 feet $360			
1	2	3	4	5	6	7	8	9
Strongly prefer left			No preference			Strongly prefer right		

By focussing on the respondents' most preferred factors and levels, the choice set presented to the respondents will vary based on their preferences. This leads to a relatively long introduction step where each respondent needs to complete a self-evaluation before the measurement of the respondents' preferences can start. It is also in this introduction step that respondents are asked to eliminate attributes and levels that will not be considered acceptable under any condition. The remaining attributes are then presented for evaluation with the attributes with lesser importance receiving less detailed questioning. As each full profile product is presented, the survey accounts for choices made and then makes the next question more efficient. No questions will be wasted on levels with little or no appeal. Extra care needs to be taken when choosing and designing the factors and their levels, in order to ensure reliable results. The big advantage of the adaptive conjoint method is that it reduces the survey length, without jeopardizing any accuracy (Orme, 2000).

2.5.5 Max-diff conjoint model

In some studies the researcher wants to know what product is most preferred by the customers. The previous methods discussed, included asking respondents to rate the attributes on a scale, where a value of one indicates a most preferred choice, while a higher value indicates a less compelling option. A problem that may occur with rating of data on a scale is when respondents cannot distinguish between what they want and what they must have. The responses will not give a clear indication of which features are considered more important than others. The second method included ranking the products from most important to less important. When respondents are asked to rank four to five different profiles it is relatively easy to order these products in their order of preference. If the number of profiles should increase, then this process becomes a lot more

difficult. Respondents can distinguish between their most preferred and least preferred profiles relatively easily but it is the process of distinguishing between the profiles in the middle ground that may be a lot more difficult (Qualtrics, 2011). Another consideration is that the differences between the different profiles are equally weighted which might not give a true representations of a respondents' response. A technique that overcomes these challenges and accurately identifies the most and least preferred profiles is called the Maximum Difference Scaling method (Max-diff) (Isaacson and Lesnick, 2012). Max-diff conjoint analysis presents the different products as an assortment of packages, which are then evaluated under most-preferred and least-preferred scenarios. The goal when employing this experimental design, is to ensure that each item is shown an equal amount of times, that each pair of items are shown an equal amount of times, and that the sets of items are properly represented. This is an ideal methodology when evaluating consumers preference, it's robust and easy to implement even with little statistical background, and it can accommodate a large number of items. Consider Illustration 2.5.8 which represents an example of Max-diff conjoint analysis based on the apartment study.

Illustration 2.5.8 Max-diff data collection method

Please indicate which of these apartments are you most likely, and which of the apartments are you least likely, to rent?		
Most		Least
☐	An apartment 10 minutes from class with a very quiet and average safety location. A newly renovated kitchen with a living size of 9 by 12 feet at $360 per month.	☐
☐	An apartment 10 minutes from class with a very noisy and average safety location. Newly renovated throughout with a living size of 24 by 30 feet at $560 per month.	☐
☐	An apartment 30 minutes from class with a average noise level and very safe location. A newly renovated kitchen with a living size of 9 by 12 feet at $260 per month.	☐
☐	An apartment 20 minutes from class with a very quiet and very unsafe safety location. A poor condition with a living size of 9 by 12 feet at $560 per month.	☐

2.5.6 Self-explicated conjoint analysis

Another popular method that is being described as simple to implement, robust and does not require the development of full profile concepts, is the self-explicated method

(Qualtrics, 2011). This method originated from the *multi-attribute model* that was developed by psychologists interested in the study of attitude formation. They argued that behavior is a function of the sum of all the attitudes that a person holds about a particular target, which is why the importance of each attribute is considered (Grover and Vries, 2006). Although some researchers do not include this method under strict conjoint analysis data collection methods, since this approach is compositional rather than decompositional which is a key characteristic of conjoint analysis studies, the results and data analysis methods are very similar. The self-explicated model is based on the multi-attribute attitude model that combines the attribute importance with the attributes' desirability to estimate the overall importance.

The model can be represented as (Qualtrics, 2011):

$$E_i = \sum_{j=1}^{m} \sum_{k=1}^{n} I_{ji} D_{jki} \tag{2.6}$$

where
I_{ji} represents the importance of attribute j for person i where $j = 1, 2..., m$
D_{jki} represents the desirability for level k of attribute j, for person i,where $k = 1, 2, ...n$
E_i is the evaluation of the product for person i

The process, in a nutshell, can be applied using the following 4 steps:
Step 1: Present the factors and levels to the respondents and ask them to eliminate those that will not be acceptable in any circumstances.
Step 2: The respondents are then asked to evaluate the desirability of the remaining attribute levels on a scale of 1-10 (or 1-100).
Step 3: The most desirable level of each attribute is evaluated and assigned a relative attribute importance.
Step 4: The attribute level preference is then weighted by the attribute importance, to provide utility values for each attribute level.

The self-explicit method places fewer demands on the respondent, does not require regression analysis or fractional factorial designs, and has been found to provide results equal or superior to full-profile methods (Qualtrics, 2011). Some of the disadvantages include the fact that the model assume that there are no additional unobserved factors that can contribute to the overall utility of each alternative product. The researcher must also remember that there will be measurement errors in the measure of importance of each attribute, I_{ji} , as well as in the desirability for each level of each attribute, D_{jki}. This method is also not used in general when determining the value that respondents

will place on products not yet in the market place (Grover and Vries, 2006). It is these shortcomings that are addressed in traditional conjoint analysis models.

2.6 Estimation methods

Once the data has been collected, following the specific experimental design, the estimation process can begin. This procedure refers to the method of deriving the part-worth utilities from the data that has been collected (Qualtrics, 2011). When considering which approach to use when estimating the parameters, the form of the responses (dependent variables) should be taken into consideration. Green and Shrinivasan, 1978 categorized the different estimation procedures into three categories.

2.6.1 Ordinally scaled responses

Some of the estimation methods that can be applied to data that has ordinally scaled responses include:

2.6.1.1 Monotone analysis of variance

The first approach is a monotone analysis of variance (MONANOVA) approach that is restricted to the part-worth function. In some studies this method is also referred to as the Kruskal's MONANOVA (Kruskal, 1965). A monotone transformation implies that the data is transformed by a function that is either nonincreasing or nondecreasing. This method basically entails that the use of monotonic transformations on the responses can lead to a minimization of the residual sum of squares.

2.6.1.2 Preference mapping

The preference mapping (PREFMAP) approach, aims to visually represent valuable information about each individuals' responses. When comparing the data representations of the maps compared to tables, the visual aspect aids in interpreting how product characteristics affect consumer responses. The perceptual map provides a clear presentation of the relationship among the products, and the individual differences in preferences by the consumers of these products, see Elmore et al., 1999 for more details.

2.6.1.3 LINMAP

The linear programming technique for multidimensional analysis of preference (LINMAP) estimation approach is ideal for the ideal point function, as it uses a linear programming

approach where the other techniques use classical calculus (Green and Shrinivasan, 1978). Another advantage is that the LINMAP approach can obtain global optimum parameters, where the other methods cannot guarantee to achieve these parameters. When prior knowledge exists the researcher can add certain constraints to the model, like weights being limited to nonnegative values and part-worth functions can be constrained to be more monotone. The altering of the model, by adding these prior knowledge constraints, can improve the accuracy of the estimates.

2.6.2 Interval scaled responses

When the responses are interval scaled, as with most traditional conjoint analysis studies, there are four estimation methods that are most commonly employed to estimate the unknown part-worth values. These estimation methods are: OLS regression (which is the most common estimation methods used), minimizing sum of absolute error regression (MSAE), latent class regression and hierarchical Bayes regression. The last estimation method can be applied to other models with different response forms and will be discussed in Chapter 4.

2.6.2.1 Ordinary least squares regression

Ordinary least squares (OLS) regression is one of the main estimation methods used in conjoint analysis studies is simple to apply. Additive factors to this methods popularity, is the facts that it can handle large numbers of attributes and levels, and it provides standard errors for the estimated parameters. OLS regression is a statistical technique that uses sample data to estimate the true population relationship between a dependent variable (Y) and an independent variable (X).

In the context of a conjoint study, this method derives a set of part-worth utilities for each level of the selected attributes, for each individual. When an aggregate model is obtained, by finding the average part-worth values over all the respondents, the aggregate estimates will apply equally well, or equally poorly to all the respondents. The model also takes unobserved variables into account by capturing these effects into the intercept of the model. The aggregate model has been found to be a reasonably accurate estimation model, but when there is high variability between the respondents, the model can give high errors. To investigate this case of heterogeneity between the respondents, consider the following simple example that investigates which pizza toppings consumers will prefer. Say that half of the population loves olives, while the other half of the population dislikes olives. When applying the OLS estimation method, the aggregate estimates will show that the population of respondents did not consider the inclusion or exclusion of olives as

a pizza topping as very important, which is certainly not the case. Those consumers who responded they dislike olives will definitely not purchase a pizza with olives as a topping. It is exactly this heterogeneity between individuals that will need to be considered and which will be investigated in detail in Chapter 3.

2.6.2.2 Latent class regression

Suppose that the researcher knows that there are two different sets of observations and that he/she has information that can identify these groups. By analyzing each market segment separately, the researcher can obtain more accurate estimations. The latent class regression model can be described as a simple variation of the OLS method. The only difference being that the OLS estimation will be applied conditioned on the known segment membership (DeSarbo, 2006).

2.6.2.3 Minimizing sum of absolute error regression

Minimizing sum of absolute error (MSAE) regression is a method often used when OLS regression is judged inappropriate or inadequate. In some cases the sum of absolute errors may simply be a more satisfactory approach than the quadratic function. This method of estimation is even more robust than OLS and allows the researcher to impose prior constraints on the estimated parameters (Green and Shrinivasan, 1978). Since linear programming is used, the conditions can easily be imposed (Roodman, 1974).

2.6.3 Paired comparison responses

When the responses are related to paired comparisons, the estimation methods that will be discussed assume that the paired comparisons are probabilistically independent. These models will relate these paired comparisons to choice probability models (Green and Shrinivasan, 1978).

2.6.3.1 Logit model

For conjoint analysis studies where the respondents have to evaluate two full profiles at a time, the response variable will be considered to be binary and can be classified as:

$$\text{Response} = \left\{ \begin{array}{ll} 1 & \text{if event occurred} \\ 0 & \text{if event did not occur} \end{array} \right\}$$

The occurrence of an event in the case of a conjoint study refers to the profile being chosen by the respondent over the other possible profile that was presented to that re-

spondent. Logistic regression is the method of choice when the response variable is qualitative, and follows a very similar approach to that of the OLS regression approach. Usually in OLS regression the dependent variable is assumed to follow a normal distribution, $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$, where $\mathbf{X}_{n\times p}$ represents the full rank design matrix for the explanatory variables (attribute levels) and $\boldsymbol{\beta}_j$ represents the vector containing the specific part-worth utilities for respondent j. For logistic regression the response is assumed to follow either a Bernoulli distribution if the experiment is repeated once, or a Binomial distribution if the experiment is repeated n times. The probability that an individual will select a specific alternative, p, is a function of the attributes that describe the alternatives and the individuals part-worth utilities for each of those attributes. The estimated probability can be expressed as:

$$p = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1 + \exp(\mathbf{X}\boldsymbol{\beta})} \tag{2.7}$$

2.6.3.2 Multinomial logit model

When the dependent variable has more than two possible categories then the Multinomial Logit Model (MNL) will be used. This method focusses on the probability to choose one of the possible k categories. This method is typically used in conjoint studies where the respondents are asked to select more than one option represented to them in the choice set. The analytical expression can be given by:

$$Logit\left(\frac{P(y=k|X_1)}{P(y=1|X_1)}\right) = \boldsymbol{\alpha}_k + \beta_k X_1 \tag{2.8}$$

where $k = 1, 2, ...h$ represents the number of possible categories, category 1 is called the reference (or control) category and $\boldsymbol{\alpha}_k$ represents the intercept term for category k. The probability to choose category k, can then be calculated by:

$$P(y=k|X_1) = \frac{\exp(\boldsymbol{\alpha}_k + \beta_k X_1)}{1 + \sum_{k=2}^{k} \exp(\boldsymbol{\alpha}_k + \beta_k X_1)} \tag{2.9}$$

2.6.3.3 Probit model

The Probit model is essentially the same as a logit model, the only difference being that the logistic link of the logit model is replaced by a inverse Gaussian link to produce the probit model. The link function provides the relationship between the linear predictor and the mean of the distribution function (Guidici and Figini, 2009).
The probit link function will be given by

$$g(x) = \Phi^{-1}(x) \tag{2.10}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz. \tag{2.11}$$

2.6.4 Hierarchical Bayes conjoint analysis

The last method being discussed is hierarchical Bayes (HB) conjoint analysis, and may still be considered by some as difficult to implement and demanding in both complexity and time, but it has been found to be very useful and accurate in situations where the data collection task is so large that the respondents cannot reasonably provide preference evaluations for all attribute levels (Qualtrics, 2011). Since the improvement in computer technology and capabilities, the time constraint is no longer a concern. Another advantage of the hierarchical Bayes approach is that it significantly improves the accuracy of the traditional methods when there is heterogeneity among the respondents. The term "hierarchical" refers to the two levels. The first assumes that the individual's parameters (betas/part-worths) are described by a multivariate normal distribution, while the second level assumes that, given the part-worth utilities, the probability of achieving some outcome, like choosing a certain product, is governed by a particular model, like a linear regression model or a multinomial logit model (Qualtrics, 2011). This concept will be discussed in more detail in Chapter 4.

2.7 Model evaluation

Since conjoint methods are mostly just concerned with the product that can be described as the "winner" among a set of possible potential products, not many conjoint studies investigate the model fit, or test the reliability or validity of the models. This is however an important step to include in any study where parameters are estimated, to ensure that the model the researcher will apply will indeed supply accurate results. Some measures that can be used to test the reliability and validity of different conjoint analysis models include the following.

2.7.1 Reliability

The reliability of a model refers to the consistency of the results that are obtained. This means that if a variable has no variation when observed using the prediction model,

then there are no random errors present and the model can be deemed a good fit. The reliability of a conjoint analysis study can be measured at the level of the input of the different respondents (Green and Shrinivasan, 1978). When considering predicting the preferences of people, it is important to note that there will always be demographical and social effects that will influence the way in which people will respond to different evaluations. It is safe to argue that the researcher will need to evaluate the reliability of the different respondents' evaluations. One way to approach this, is to ask a sample of the judges to illustrate their preference on a second set of stimuli which consists of only a subset of the original sample. These repeated evaluations can then be used to test the reliability of these inputs. Another test that can be done, is to approach a sample of respondents after a certain period of time and ask them to re-evaluate a subset of the stimuli. The correlation of the estimated parameters for the two tasks can be determined, to measure the reliability.

2.7.2 Validity

When evaluating the estimated parameters the researcher should be concerned with the accuracy of the predicted utilities. The degree to which the predicted choices align with the actual choices needs to be investigated. The closer these two measures are aligned the better the predicted model. This might sound easy in theory but this may not always be the case. In most conjoint analysis studies there will be a number of differences between the research environment and the actual market place. Some of these differences include (Grover and Vries, 2006):

- The conjoint analysis data collection procedures ensure that each respondent is fully aware of all the alternatives and has easy access to all of these alternatives. This will almost certainly not be the case in the actual market place.

- In the hypothetical market place each respondent will usually receive a full description of each product that they have to evaluate. This product description may contain a lot more information that will not necessarily be so easily available to the respondents in the actual market place. This can then lead to predicted utilities which will not represent the actual respondents behavior.

- Another difference that needs to be considered is the time lag between the conjoint study and the implementation into the market place. There are many factors that can change during the time from when the study was done to the time that it can be implemented.

Despite all of these difficulties, conjoint analysis is still considered to achieve reasonable predictive accuracy, when compared to the actual market place.
Two methods that are commonly used to asses the validity of conjoint analysis models will be briefly discussed.

Method 1

The first method is a simple but effective approach that most researchers use to validate their results. The *holdout validation method* has many variations of which it is always beneficial to apply more than one to evaluate the quality of the conjoint model. This procedure simply removes a proportion of the observations during the estimation process. Once the estimation process is done, the researcher can evaluate how well the model predicts the observations that were held out. In conjoint analysis the researcher can either set aside some of the profiles or some of the respondents, or both.

▶ Holdout profiles

By using holdout profiles the researcher will be able to evaluate the *internal consistency* of the model. The data from randomly chosen profiles are removed from the estimation process. Once the part-worth values have been estimated for those profiles that where included in the estimation process, the responses for those profiles that were not included in the estimation process, can be predicted. Typically the results for the holdout validation methods can be expressed in terms of the *Mean Squared Error* (MSE) or the *Mean Absolute Error* (MAE). The smaller the distance between the predicted values and the observed values the better. In the case of individual level models the *Hit ratio's* are calculated. This ratio is determined by calculating the proportion of cases that where correctly predicted. The higher the hit ratio, the better. It is however important to note that it is possible for a good-fitting model to do a poor job in predicting the responses of the holdout task. This can be caused by factors such as the simulation method and the composition of the holdout task (Grover and Vries, 2006). It is hence recommended to use two or more holdout tasks.

▶ Holdout respondents

This method can be used to evaluate the generalizability of the model to other consumers (Grover and Vries, 2006). The process starts by randomly selecting a subset of the responses, usually 20% or more depending on the available sample, and excluding this subset from the estimation step. Once the part-worth utilities are calculated they can then be used to predict the responses that were not included in estimation procedure. It is also recommended to incorporate a *k-fold* cross validation approach which consists of the exact same process but with a different subset of respondents kept out of the estimation process and evaluated after the part-worth utilities were estimated. This process is then

repeated k times (Hastie et al., 2008). In general the researcher will be most confident in the method that repeatedly generates results that are similar to those obtained in the market place.

Method 2

The second model validation method validates conjoint models by predicting the current market shares for the existing products. This method requires all the attribute levels of the product in the currant market place. The data about the current markets can be obtained from industry research, self-reported surveys or client supplied estimates. When this method results in poor results for the chosen model, it can be due to respondent confusion or poor definition of one or more of the attribute levels (Grover and Vries, 2006).

Chapter 3

Conjoint Analysis Design

"Design is not just what it looks like and feels like. Design is how it works."

- Steve Jobs

In the well-known article "Thirty Years of Conjoint Analysis: Reflections and Prospects" (Green et al., 2001) it is stated that conjoint analysis is by far the most used marketing research method for analyzing consumer trade-offs. Surveys conducted by Wittink & Cattin,1989 and Wittink et al., 1994, attests to its popularity. This popularity can be attributed to the fact that conjoint analysis evaluates the central management question: why do consumers choose one brand or supplier over another? This chapter will discuss the traditional conjoint analysis design often used in marketing, and possible extensions of this model will be investigated.

3.1 The traditional conjoint analysis model

Since consumers are faced with many day-to-day decisions, marketing researchers can use these day-to-day concerns to collect trade-off information from tens, hundreds or even thousands of respondents. This information, in turn, can then be used to measure preference, intention to buy, how consumers might react to changes in current products or how they feel about the possibility of introducing new products

This chapter will start by investigating the traditional conjoint analysis design, also commonly referred to as the metric conjoint analysis approach. This approach uses a part-worth preference model: the respondents indicate their preference through ratings and OLS estimation is used to estimate the part-worth utilities. One great advantage of OLS estimation is the fact that it accounts for the arbitrary parameter heterogeneity at the individual consumers' level (Frühwirth-Schatter and Otter, 1999). Because of the quantitative nature of the responses, the qualitative nature of the dummy variables featuring in the design matrix, and the simple reading of the part-worths, this approach to conjoint analysis is the most used (Fichet et al., 2011). This model considers a well defined data structure that can be analyzed with a classical regression model. To better understand this model and why it is so widely used, the fixed effects linear regression model, which essentially is the approach that the traditional conjoint analysis uses, will be discussed in detail in section 3.2.

Extending the traditional model to improve the predictive capability will be investigated during the rest of the chapter. By introducing a random component to the usually fixed effects of the traditional conjoint analysis method, a model can be produced that can explain the variation between the different preference structures of the respondents. Since the different respondents can be classified on the basis of several socio-demographical characteristics (characteristics of certain groups within a society), it is important to consider these differences and measure their influences on the model. To illustrate the reason for possibly extending the traditional conjoint analysis model, consider the following example based on research done by Allison and Christakis, 1994, where they investigate the way in which physicians make decisions about the withdrawal of life support. The questionnaire that was sent to 475 physicians included the following question:

Illustration 3.1.1 Physician questionnaire

Some physicians may feel differently about withdrawing life-sustaining therapy, depending on what the therapy is. In general, what kind of life-sustaining therapy are you likely or unlikely to withdraw if the circumstances presented themselves?
Rank your responses from 1 (most likely to withdraw) to 8 (least likely to withdraw). Use the same number for any therapies you believe deserve the same rank.
- antibiotics ☐
- blood products ☐
- intravenous fluids ☐
- intravenous vasopressors ☐
- mechanical ventilation ☐
- renal dialysis ☐
- total parenteral nutrition ☐
- tube feeding and fluids ☐
If relevant, please also check none, one, or both of these two boxes:
☐ I would not withdraw any of these medical therapies
☐ I consider all of these equally easy or difficult to withdraw

When estimating the utilities for each of these 8 types of therapies in the traditional conjoint analysis way, the model takes the differences across the types of therapies into consideration, but not the differences among the physicians. Consider the possibility that the different physicians will have different opinions as explained in Böckenholt, 2001, based on factors such as:

- Physicians can differ in their importance weighting of therapy expenses;
- Physicians can also differ in their assessment of therapy attributes that are not explicitly listed or included in the study. For example, the above mentioned questionnaire does not include a measure of effectiveness of the therapy;
- The training of the physicians such as traditional or holistic medicine, may also lead to differences between their responses.

These differences in opinions can explain the latent heterogeneity in the regression weights and the evaluation of the profiles. The addition of extra information in the model should

result in more accurate predictions. Firstly it is of vital importance to understand the differences between the concept of fixed effects and random effects. Once these two effects are combined the resulting model, called the linear mixed effect model, will be investigated.

3.2 Traditional approach - fixed effects model

Typically fixed effect models consist of parameters that are purposefully chosen and fixed by the researcher. In classical linear regression the inputs are considered linear and fixed by the researcher. Although linear models were almost entirely designed and developed in the precomputer era they are still widely used today due to their simplicity and their accurate and interpretable description of how the inputs affects the output (Hair et al., 2006). Studies have also shown that their prediction models often outperforms fancier non-linear methods. This traditional approach to conjoint analysis fits a linear regression model for each individual that does not allow correlations between the different profiles. The aggregate model can then be determined by taking the global average of the parameters, over all the respondents. The linear regression model will briefly be discussed to ensure a comprehensive understanding of the traditional conjoint analysis model. Although this is a well-documented topic it is included for completeness and to facilitate the flow of the discussion.

3.2.1 Classical linear regression model

The classical linear regression model is used to predict the value of a response (*dependent variable*) given a set of predictors (*independent variables*). It can also be used to analyze the effect that the predictor has on the response. Given that $\mathbf{y}$ *(dependent variable)* consists of n independent observations, the model can be represented in matrix notation, as follows:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times(p+1)}{\mathbf{X}} \underset{(p+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}} \tag{3.1}$$

where $\mathbf{y}$ is the vector of n responses, $\mathbf{X}$ is the full rank, deterministic, design matrix, $\boldsymbol{\beta}$ is the vector of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is the vector of errors. In this model

the regression coefficients are considered fixed since each coefficient is linked to an actual group or individual. The error terms are assumed to have the following properties:

- The errors have zero means, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, and constant variances, $Var(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}_n$.
- The error terms are uncorrelated, $Cov(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_k) = 0$ if $i \neq k$
- The errors terms are normally distributed $\boldsymbol{\varepsilon} \backsim N(0, \sigma^2)$.

3.2.1.1 Parameter estimation

The ordinary least squares (OLS) and maximum likelihood (ML) methods can be used to estimate the unknown parameters, $\boldsymbol{\beta}$ and σ^2. It will be shown that for the linear model case the ML method will produce the exact same results as the OLS method if the assumptions of the error terms are met.

▶ OLS method

To determine the values of the $\beta's$ the Residual Sum of Squares (RSS) need to be minimized. The *RSS* can be expressed as:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_p X_{ip})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' = \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \tag{3.2}$$

which leads to the estimated parameters being given by (Johnson and Wichern, 2007)

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3.3}$$

Under the classical linear model the estimated parameter will follow a normal distribution (*From Result 1*):

$$\widehat{\boldsymbol{\beta}} \backsim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \tag{3.4}$$

In order to determine the estimate for the unknown parameter, σ^2, the residuals will need to be investigated further. If $\widehat{\sigma}^2$ is the least squares estimate of σ^2, then the estimated variance can be defined as follows:

$$\widehat{\sigma}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-(p+1)} = \frac{\mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]\mathbf{y}}{n-(p+1)} = \frac{\mathbf{y}'[\mathbf{I} - \mathbf{H}]\mathbf{y}}{n-(p+1)} \tag{3.5}$$

where the hat matrix $\mathbf{H}$ is an indempotent matrix (*From Result 2*). This will result in an unbiased estimate of the variance

$$E(\hat{\sigma}^2) = E\left(\frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-(p+1)}\right) = \sigma^2.$$

Using *Result 3* it can be proven that $n\widehat{\sigma}^2$ will then have a chi-square distribution with $n-p-1$ degrees of freedom, $\sigma^2\chi^2_{n-p-1}$. It is also important to note that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ is statistically independent.

Finally, the response variable $\mathbf{y}$ will then be normally distributed (*from Result 4*):

$$y \backsim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n). \tag{3.6}$$

▶ Maximum likelihood method

The maximum likelihood method chooses the estimates who's values of the parameters are most consistent with the sample data. If

$$y_i \backsim F(\underline{\boldsymbol{\theta}}), i = 1, 2, ...n$$

where $\underline{\boldsymbol{\theta}}$ is the vector of parameters of interest, then the likelihood function, based on a random sample, will be

$$L(\underline{\boldsymbol{\theta}}, \{y_i\}_{i=1}^n) = \Pi_{i=1}^n F(y_i, \underline{\boldsymbol{\theta}})$$

where $\underline{\boldsymbol{\theta}}$ is the vector containing the parameters that need to be estimated. In order to pursue the theory of the ML method, the derivative with respect to $\underline{\boldsymbol{\theta}}$ should be determined, to find the value for the parameter that will maximize the function (using *Result 5*). The results of the estimated values are exactly the same as those obtained using the OLS method in (3.3):

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3.7}$$

When calculating the estimated variance the same process can be followed, only in terms of σ^2. The estimate for the variance, determined using the ML method, is slightly different than the estimate obtained using the OLS method given by (3.5):

$$\widehat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n}. \tag{3.8}$$

However, this difference becomes negligibly small as $n \to \infty$.

3.2.1.2 Tests for regression parameters

A large part of regression analysis is concerned with assessing the effects that particular predictor variables may have on the response variable. A brief discussion of a few tests that can be conducted will follow. These tests use a Analysis of Variance (ANOVA) approach which investigates the differences between the groups represented by the independent variables. The ANOVA table, for the linear regression models, is given by Table 3.1 (Kutner et al., 2005):

Table 3.1 Linear regression ANOVA table

Source of Variation	Sum of Squares	df	MS
Regression	$SSR = \mathbf{b}'\mathbf{X}'\mathbf{y} - \left(\frac{1}{n}\right)\mathbf{y}'\mathbf{J}\mathbf{y}$	$p-1$	$MSR = \frac{SSR}{p-l}$
Error	$SSE = \epsilon'\epsilon = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	$SST = \mathbf{y}'\mathbf{y} - \left(\frac{1}{n}\right)\mathbf{y}'\mathbf{J}\mathbf{y}$		

where $\mathbf{b}$ denotes the $(p+1) \times 1$ vector of least squares estimated regression coefficients and $\mathbf{J}$ is a $n \times n$ matrix of one's. If the between group variation (SSR) is significantly greater than the within group variation (SSE), then it is likely that there is a statistically significant difference between the groups.

▶ The overall test:

The overall test, tests whether the entire set of predictor variables contribute significantly to the response variable. Consider the following hypothesis:

$$\begin{aligned} H_0 &: \beta_0 = \beta_1 = ... = \beta_p = 0 \\ H_A &: \beta_j \neq 0 \qquad \text{for} \quad j = 0, 1, ...p. \end{aligned}$$

If H_0 is rejected it can be concluded that at least one coefficient, β_j, will contribute significantly to the model, since it will have a non zero value. The F-test can be used to test this hypothesis. Under H_0 the sources of variation will have the following distributions *(From Result 3)*:

$$\frac{SSR}{\sigma^2} \backsim \chi^2_p \quad \text{and} \quad \frac{SSE}{\sigma^2} \backsim \chi^2_{n-(p+1)}. \tag{3.9}$$

Since these distributions are independent the test statistic will have a F-distribution with p and $n-(p+1)$ degrees of freedom *(From Result 6)*:

$$F^* = \frac{MSR}{MSE} = \frac{SSE/(n-(p+1))}{SSR/p} \backsim F_{p,n-(p+1)}(\alpha). \tag{3.10}$$

The null hypothesis will be rejected if the test statistic is greater than the critical value,

$F^* > F_{p,n-(p+1)}(\alpha)$ where $F_{p,n-(p+1)}(\alpha)$ is the upper alpha-level critical value of the F-distribution with p and $n-p-1$ degrees of freedom.

▶ Testing the individual regression coefficients:

If it is known that at least one β_j contributes to the model then the t-test can be used to test the different individual coefficients' significant contribution to the model. Consider the following hypothesis:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_A &: \beta_j \neq 0 \qquad \text{for} \quad j = 0, 1, ...p \end{aligned}$$

Under H_0 the t-test statistic will have a t-distribution with $n-(p+1)$ degrees of freedom

$$t = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \backsim t_{n-(p+1)} \tag{3.11}$$

where $se(\widehat{\beta}_j)$ is the standard error. The null hypothesis will be rejected if the absolute value of the test statistic is greater than the critical value, $|t| > t_{n-(p+1),1-\alpha/2}$ where α is the level of significance. This is a partial test since $\widehat{\beta}_j$ depends on all the other predictors in the model. Thus, this test, tests the contribution of x_j given all the other predictors in the model.

▶ Testing a subset of regressors:

In certain cases it may be necessary to test whether a certain subset of predictors do influence the response variable or not. This can be done by using a test called the Likelihood Ratio Test. The predictors that will be tested to see if they influence the response will be labelled as $\mathbf{x}_{q+1}, \mathbf{x}_{q+2,}...\mathbf{x}_p$ (Draper and Smith, 1998). Consider the following hypothesis:

$$\begin{aligned} H_0 &: \beta_{q+1} = \beta_{q+2} = ... = \beta_p = 0 \text{ or } H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0} \\ H_A &: \boldsymbol{\beta}_{(2)} \neq \mathbf{0} \end{aligned}$$

where $\boldsymbol{\beta}_{(2)} = [\boldsymbol{\beta}_{q+1}\ \boldsymbol{\beta}_{q+2}...\boldsymbol{\beta}_p]$. If H_0 cannot be rejected then the predictors will not influence the response variable, since all the coefficients will be zero. By splitting the design matrix and the vector of coefficients into two parts, one containing the variables we are testing and the other part containing the rest of the variables,

$$\mathbf{X} = \underset{n\times(q+1)}{\left[\mathbf{X}_1\right.} \Big| \underset{n\times(p-q)}{\left.\mathbf{X}_2\right]} \quad \text{and} \quad \boldsymbol{\beta} = \left[\begin{array}{c} \underset{(q+1)\times 1}{\boldsymbol{\beta}_{(1)}} \\ \hline \underset{(p-q)\times 1}{\boldsymbol{\beta}_{(2)}} \end{array}\right]$$

the general linear model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = [\mathbf{X}_1|\mathbf{X}_2] \begin{bmatrix} \boldsymbol{\beta}_{(1)} \\ \hline \boldsymbol{\beta}_{(2)} \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_{(1)} + \mathbf{X}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}.$$

Under H_0 the model will reduce to

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}$$

since $\boldsymbol{\beta}_{(2)} = 0$ and the SSR will reduce to $SSR_{(1)} = \widehat{\beta}'_{(1)}\mathbf{X}'_1\mathbf{y}$. The likelihood ratio test under the null hypothesis is based on the extra sum of squares calculated by

$$Extra\ Sum\ of\ Squares = SSR_{(1)} - SSR$$

The likelihood ratio test will then reject H_0 if the test statistic is greater than the critical value:

$$\frac{(SSR_{(1)} - SSR)/(p-q)}{MSE} > F_{p-q,n-(p+1)}(\alpha)$$

where $F_{p,n-(p+1)}(\alpha)$ is the upper alpha-level critical value of the F-distribution with.p and $n-p-1$ degrees of freedom. For applications of these test, refer to Johnson and Wichern, 2007. Since the concept of a fixed effects model has been summarized, the conjoint analysis application, which is based on this model, will now be discussed in detail.

3.2.2 Industrial cleaner conjoint analysis example

Consider the following example of a conjoint analysis study done by a company that wanted to evaluate the preferences of respondents to different attributes in an industrial cleaner (Hair et al., 2006). 86 Respondents were asked to rate 18 detergents on a 7 point scale ranging from "not at all likely to buy" to "certain to buy". Before the actual conjoint study was performed, internal marketing research teams, in consultation with the product development group, identified five attributes as the determinant factors in the targeted segment of the industrial cleaner market. These five attributes are represented in Illustration 3.2.1. The 18 different detergents profiles that each respondents had to evaluate, were obtained using a orthogonal fractional design involving these 5 factors:

Illustration 3.2.1 Industrial cleaner attributes and levels

Attribute	**Levels**
Form of the product	1. Premixed liquid 2. Concentrated liquid 3. Powder
Number of applications per bottle	1. 50 applications 2. 100 applications 3. 200 applications
Addition of disinfectant	1. Yes 2. No
Biodegradable	1. No 2. Yes
Price per application	1. \$0.35 2. \$0.49 3. \$0.79

The original setup of each of the 18 stimuli is shown in Illustration 3.2.2. On investigation of Illustration 3.2.2 it can be seen that the first profile will be an industrial detergent in concentrated liquid form, with a bottle size that can assure 200 applications, it has an added disinfectant, it is not considered biodegradable, and the price per application is \$0.35.

Illustration 3.2.2 The 18 profiles evaluated by the 86 respondents

Profile	Product Form	Number of Application	Disinfectant	Biodegradable	Price
1	2	3	1	1	1
2	3	3	1	1	1
3	1	2	1	2	2
4	3	3	1	2	2
5	3	1	1	1	3
6	2	3	2	2	3
7	1	2	1	1	3
8	1	3	1	1	2
9	3	2	2	1	2
10	2	1	1	1	2
11	3	2	2	1	1
12	2	2	1	1	3
13	1	3	2	1	3
14	1	1	1	1	1
15	2	2	1	2	1
16	1	1	2	2	1
17	2	1	2	1	2
18	3	1	1	2	3

For this experiment the design matrix was set up using dummy variables, which can be considered as artificial variables created to represent an attribute with two or more distinct levels. These dummy variables indicated the presence or absence of a specific level in each profile. For each attribute containing k levels, $k-1$ levels will be mathematically independent and it will only be those that will be used in estimating the model. For example, the attribute, product form, has three levels so only two levels will be used in the design matrix $\mathbf{X}$. The way in which the dummy variables where set up is represented in Illustration 3.2.3.

Illustration 3.2.3 Rules used for setting up die dummy variables

Product form	$F1 = \left\{ \begin{array}{ll} 1 & \text{if Premixed} \\ 0 & \text{if Concentrate} \\ -1 & \text{if Powder} \end{array} \right\}$	$F2 = \left\{ \begin{array}{ll} 1 & \text{if Concentrate} \\ 0 & \text{if Premixed} \\ -1 & \text{if Powder} \end{array} \right\}$
Number of applications	$A1 = \left\{ \begin{array}{ll} 1 & \text{if 50 App.} \\ 0 & \text{if 100 App.} \\ -1 & \text{if 200 App.} \end{array} \right\}$	$A2 = \left\{ \begin{array}{ll} 1 & \text{if 100 App.} \\ 0 & \text{if 50 App.} \\ -1 & \text{if 200 App.} \end{array} \right\}$
Disinfectant and Biodegradability	$D1 = \left\{ \begin{array}{ll} 1 & \text{Yes} \\ -1 & \text{No} \end{array} \right\}$	$B1 = \left\{ \begin{array}{ll} 1 & \text{No} \\ -1 & \text{Yes} \end{array} \right\}$
Price per application	$P1 = \left\{ \begin{array}{ll} 1 & \$0.35 \\ 0 & \$0.49 \\ -1 & \$0.79 \end{array} \right\}$	$P2 = \left\{ \begin{array}{ll} 1 & \$0.49 \\ 0 & \$0.35 \\ -1 & \$0.79 \end{array} \right\}$

The design matrix, $\mathbf{X}$, containing these dummy variables of each of the 18 profiles was represented by Illustration 3.2.4.

Illustration 3.2.4 Design matrix

Profile	Intercept	F1	F2	A1	A2	D1	B1	P1	P2
1	1	0	1	-1	-1	1	1	1	0
2	1	-1	-1	-1	-1	1	1	1	0
3	1	1	0	0	1	1	-1	0	1
4	1	-1	-1	-1	-1	1	-1	0	1
5	1	-1	-1	1	0	1	1	-1	-1
6	1	0	1	-1	-1	-1	-1	-1	-1
7	1	1	0	0	1	1	1	-1	-1
8	1	1	0	-1	-1	1	1	0	1
9	1	-1	-1	0	1	-1	1	0	1
10	1	0	1	1	0	1	1	0	1
11	1	-1	-1	0	1	-1	1	1	0
12	1	0	1	0	1	1	1	-1	-1
13	1	1	0	-1	-1	-1	1	-1	-1
14	1	1	0	1	0	1	1	1	0
15	1	0	1	0	1	1	-1	1	0
16	1	1	0	1	0	-1	-1	1	0
17	1	0	1	1	0	-1	1	0	1
18	1	-1	-1	1	0	1	-1	-1	-1

This linear regression model will then take on the following form:

• Let $i = 1, 2, ...m$ represent the number of respondents, where $m = 86$.

• Let $j = 1, 2, ...n$ represent the number of profiles that where evaluated by each respondent, where $n = 18$.

The traditional conjoint analysis model estimates different part-worth utilities for each of the attribute levels, for each respondent. All of these effects are considered to be fixed. The model can be represented as follows:

$$y_{ij}=\beta_I x_{j1}+\beta_{F1} x_{j2}+\beta_{F2} x_{j3}+\beta_{A1} x_{j4}+\beta_{A2} x_{j5}+\beta_{D1} x_{j6}+\beta_{B1} x_{j7}+\beta_{P1} x_{j8}+\beta_{P2} x_{j9}+\varepsilon_{ij} \tag{3.12}$$

or the model can be represented in matrix form as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}\boldsymbol{\beta}_i+\boldsymbol{\varepsilon}_i \text{ for } i = 1, 2, ..., m \\ \boldsymbol{\varepsilon}_i &\backsim N(\mathbf{0}, \sigma^2) \end{aligned}$$

where
$\mathbf{y}_i$: Vector of respondent $i's$ 18 preference ratings
$\boldsymbol{\beta}_i$: Vector containing respondent $i's$ part-worth utilities
$\mathbf{X}$: Design matrix containing the observed attributes in the form of dummy variables
$\boldsymbol{\varepsilon}_i$: Vector of respondent $i's$ error terms

By using OLS to estimate the linear regression model a set of part-worth utilities can be estimated for each respondent. This results in the estimation of 9 parameters per individual. (Consider SAS program 1 for the code). For completeness the calculations of respondent 1 is given.

3.2.2.1 Respondent 1 (i=1)

When the researcher just considers the responses obtained on an individual level (for the first respondent), the estimated part-worth values are given by Table 3.2.

Table 3.2 Part-worth utilities obtained for respondent 1

β_0	β_{F1}	β_{F2}	β_{A1}	β_{A2}	β_{D1}	β_{B1}	β_{P1}	β_{P2}
4.11	-0.06	0.61	0.44	0.61	-0.21	0.54	1.44	0.94

The goal of a fixed effect model is to estimate the mean effect that the explanatory variables have on the responses. These part-worth values obtained for the first respondent illustrates how much, or how little, the respondent likes each level of the 5 attributes. To obtain the complete set of part-worth values that will describe the utility of respondent 1 towards each level of all the attributes, the researcher can use the property of dummy variable design matrices. The property is such that the parameters, of each level, for one attribute should add up to 0:

$$\begin{aligned} \beta_{\text{Premixed}} + \beta_{\text{Concentrate}} + \beta_{\text{Powder}} &= 0 \\ \beta_{\text{F}_1} + \beta_{\text{F}_2} + \beta_{\text{F}_3} &= 0 \\ -0.06 + 0.61 + \beta_{\text{F}_3} &= 0 \\ \text{thus } \beta_{\text{F}_3} &= -0.55 \end{aligned} \tag{3.13}$$

Following the same process for all 5 attributes the complete set of part-worth utilities can be determined for respondent 1 and was tabulated in Table 3.3.

Table 3.3 Complete set of part-worth utility values for respondent 1

β_I	β_{F1}	β_{F2}	β_{F3}	β_{A1}	β_{A2}	β_{A3}	β_{D1}	β_{D2}	β_{B1}	β_{B2}	β_{P1}	β_{P2}	β_{P3}
4.11	-0.06	0.61	-0.55	0.44	0.61	-1.05	-0.21	0.21	0.54	-0.54	1.44	0.94	-2.38

The intercept term, β_I, illustrates the overall likelihood of purchase that respondent 1 has for all 18 profiles. When considering the part-worth values obtained for the attribute product form (F1, F2 and F3) this respondent preferred the concentrated liquid product form over the premixed and powder form. The highest part-worth value in each attribute illustrates which level of each attribute the respondent preferred. The attribute levels associated with the lowest part-worth values represents the attribute levels which respondent 1 least preferred.

The part-worth values represented in Table 3.3, can in turn be used to predict the utility scores for respondent 1, by applying the following prediction formula:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}. \tag{3.14}$$

The following predicted ratings where obtained for the 18 profiles evaluated by respondent 1. The higher the predicted ratings, the greater the preference of respondent 1 towards the specific profile.

Table 3.4 Predicted responses for respondent 1

Profile	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
y	6	3	5	2	3	1	1	6	6
$\widehat{\mathbf{y}}$	5.44	4.28	4.86	2.69	1.94	0.94	2.61	4.28	5.86
Profile	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
y	6	7	4	1	6	6	6	6	1
$\widehat{\mathbf{y}}$	6.44	6.36	3.28	1.36	6.28	6.03	5.61	6.86	0.86

The accuracy of the model can then also be determined by evaluating the predicted utilities against the actual rating the respondent gave each of the profiles. A common way to evaluate the model is by calculating the coefficient of determination, denoted by R^2. The closer the value is to 1, the better the model fit. The coefficient of determination is given by:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

The accuracy of the model when predicting the first respondents ratings was relatively good: with an $R^2 = 0.8630$, the researcher will be able to conclude that the model is a

good fit for this respondent. Another measure that a researcher will want to determine during a conjoint analysis study is the *overall importance* that the respondent unknowingly assigned to each attribute. These attribute importance values reflect the relative impact each attribute has on the calculation of the overall preference. The higher the value, the higher the impact. These values can be determined by measuring the range of each attributes' part-worth values compared to the range over all attributes. This range is determined by measuring the distance between the smallest and highest part-worth utility for each attribute. For the first respondent the importance measures were:

Table 3.5 Importance measure for respondent 1

Attribute	**Min**	**Max**	**Range**	**Importance**
Product form	-0.55	0.61	1.16	14.24%
Number of applications	-1.05	0.61	1.66	20.40%
Disinfectant	-0.21	0.21	0.42	5.15%
Biodegradable	-0.54	0.54	1.08	13.27%
Price per application	-2.38	1.44	3.82	46.93%
Total			8.14	100%

Considering these results it can be seen that the attribute that is considered the most important is the price per application, with the second most important attribute, the number of applications per product. This respondent is clearly most concerned with cost and the amount of applications he/she can get before having to buy a new product. These are general concerns that consumers are faced with in this era. The business then has to find the balance, where the respondents are happy and willing to pay, while the business still makes a sustainable profit. This process is then typically repeated for all 86 respondents.

3.2.2.2 Aggregate model

Conjoint analysis studies are, however, more interested in aggregate part-worth results (the average rating for the separate part-worth utilities across all respondents), than simply considering the individual part-worth values. The individual part-worth estimations can be of vital importance for a conjoint study as they can be used to classify and group the respondents into different market segments. The means and variances of the parameter estimates averaged across all the parameter of all 86 respondents, resulted in the values obtained in Table 3.6

Table 3.6 Estimated global mean and variance of fixed part-worth values

Level	**Global Estimated Means ($\beta'.s$)**	**Global Estimated Variances**
Residual		2.62
Intercept	3.74	0.63
Premixed Liquid	-0.22	0.23
Concentrate Liquid	0.17	0.15
50 Applications	-0.35	0.32
100 Applications	0.02	0.19
Disinfectant	0.51	0.38
Biodegradable	-0.15	0.17
Price - \$0.35	1.13	0.63
Price - \$0.49	0.08	0.25

Table 3.6 contains the estimated part-worth values for each attribute level present in the dummy design matrix, **X**. In order to obtain the part-worth utilities for all the attribute levels, the same property of regression analysis using dummy variables, as used in (3.13), can be applied to the aggregate estimates of the part-worth values. Table 3.7 shows the part-worth values for all the attribute levels and Figure 3.2.1 visually represents the part-worth utilities for each level of each attribute. This will graphically indicate which attribute levels were most preferred by the respondents and which had an overall lesser importance. This visualization purely illustrates the preference value for each attribute level. The dashed lines does not necessarily indicate continuity but rather empathizes the preference for each level. From Figure 3.2.1 it is clear why this method is preferred by many researchers, since they can quickly and easily determine which attribute levels will be more preferable by the respondents. By simply choosing the level of each attribute that has the highest preference value, the researcher can, with relative accuracy, estimate the preferred level of each attribute.

Table 3.7 Part-worth values for all attribute levels

Attribute level	**Part-worth utilities** ($\beta's$)
Premixed liquid	-0.22
Concentrate liquid	0.17
Powder	0.05
50 Applications	-0.35
100 Applications	0.02
200 Applications	0.33
Disinfectant - Yes	0.51
Disinfectant - No	-0.51
Biodegradable -No	-0.15
Biodegradable - Yes	0.15
Price per application - $0.35	1.13
Price per application - $0.49	0.08
Price per application - $0.79	-1.21

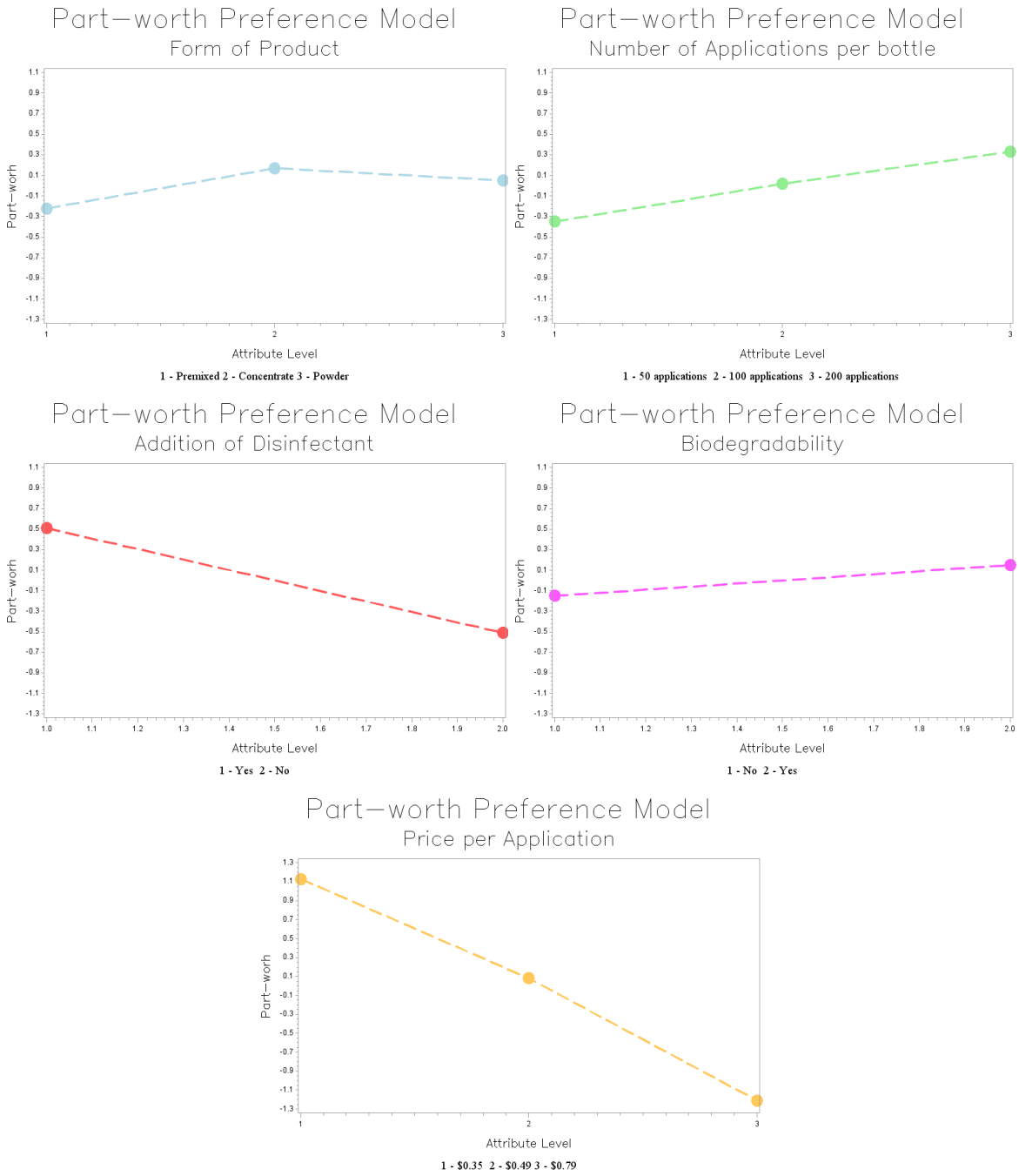


Figure 3.2.1 Graphical representation of the part-worth utilities

When the price per application is considered (which is the only continuous attribute), it should be noted that linear interpolation could be used, with relative efficiency, to estimate the part-worth values of other price levels, since the part-worth utilities are relatively linear (as discussed in Chapter 2.3.1). From all the displays in Figure 3.2.1 it can be seen that the attribute levels that were preferred by the population of respondents are:

Illustration 3.2.5 Most preferred attribute levels

Concentrated product form
200 Applications
Must contain disinfectant
Biodegradable
$0.35 per application

Although these part-worth utilities are very important indications of the preferences of the respondents, it is nesseasary to evaluate the overall importance that the group of respondents places on each of these attributes. If the biodegradable attribute is considered it can be seen that the part-worth utility associated with this attribute is not very large in comparison to the other part-worth utilities. This means that although the largest subset of the population of respondents did prefer an industrial cleaner that is biodegradable, the biodegradability of the product did not have a big influence on the respondents preference of the overall product descriptions. The overall importance of the attributes, for the aggregate model, can be calculated in the exact same way as when the study considered only respondent 1. The ranges and importance values are illustrated in Table 3.8 and is visually represented in Figure 3.2.2.

Table 3.8 Overall importance measure for all respondents

Attribute	**Min**	**Max**	**Range**	**Importance**
Product form	-0.22	0.17	0.39	8.25%
Number of applications	-0.35	0.33	0.68	14.38%
Disinfectant	-0.51	0.51	1.02	21.56%
Biodegradable	-0.15	0.15	0.3	6.34%
Price per application	-1.21	1.13	2.34	49.47%
Total			4.73	100%

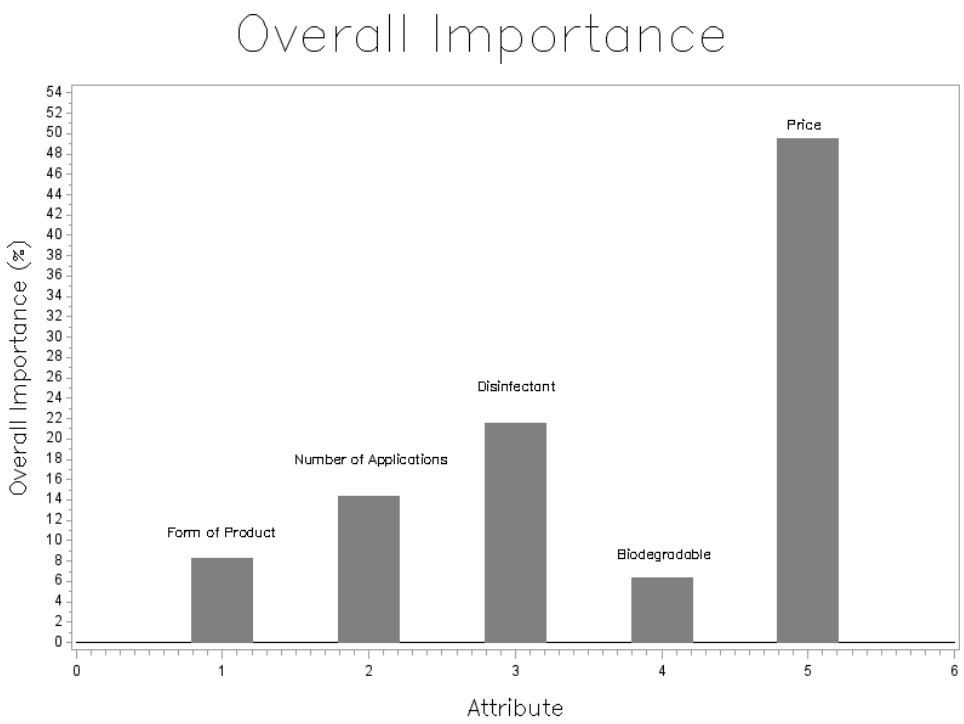


Figure 3.2.2 Overall importance values of the industrial cleaner attributes

When considering the overall importance of each attribute, it is quite clear that the price greatly affects the choices that the respondents make. This means that if the price of the product would increase, the probability that the respondents will still buy the product, will drastically decrease, as can be seen in the price attribute graph in Figure 3.2.1. In a lot of studies, researchers will choose to exclude price from the study as it will overpower the other attributes and lead to less accurate estimates. The attribute that had the second biggest importance value was whether or not the industrial cleaner contains disinfectant. Clearly most respondents feel that the product should include disinfectant. They feel so strongly about it that most respondents will consider not purchasing products that do not contain disinfectant, as the preference decreases to a negative utility which implies that the respondents will not prefer to buy the product that does not contain disinfectant. When the overall importance values of each attribute is known, the researcher can re-evaluate the most preferred attribute levels of Illustration 3.2.5. The most important attribute levels to include in their product will be the addition of disinfectant at a price as low as possible with as much as possible applications per contained for that price. The product does not necessarily have to be biodegradable since it will not influence the respondents preferences that much but biodegradability can have other advantages such as environmental safety grants etc. Finally the most respondents preferred a concentrated

liquid form.

The overall part-worth utility values, as represented in Table 3.9, can then be used to calculate the predicted utilities, of the 18 profiles, for the population of respondents. Consider the following prediction equation:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \tag{3.15}$$

or equivalently

$$\widehat{y}_j = \widehat{\beta}_I.x_{j1} + \widehat{\beta}_{F1}.x_{j2} + \widehat{\beta}_{F2}.x_{j3} + \widehat{\beta}_{A1}.x_{j4} + \widehat{\beta}_{A2}.x_{j5} + \widehat{\beta}_{D1}.x_{j6} + \widehat{\beta}_{B1}.x_{j7} + \widehat{\beta}_{P1}.x_{j8} + \widehat{\beta}_{P2}.x_{j9} + \varepsilon_j$$

for $j = 1, 2, ...18$
The predicted values for the 18 profiles were:

Table 3.9 Predicted utilities for the 18 profiles

Profile	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
$\widehat{\mathbf{y}}$	6	6	4	5	3	3	3	4	3	4	4	3	2	5	6	4	3	3

The three predicted profiles that obtained the highest ratings (and hence will be the products that are thought to ensure the highest number of sales), are:

Table 3.10 Highest predicted preferred profiles

Profile	**Product Form**	**Applications**	**Disinfectant**	**Biodegradable**	**Price**
1	Concentrated Liquid	200	Yes	Yes	\$0.35
2	Powder	200	Yes	Yes	\$0.35
15	Concentrated Liquid	100	Yes	No	\$0.35

Once again it is not strange to see that the highest rated profiles are all products with the lowest prices, and two of the three ideal products have the highest number of applications per container. It is however not always possible for a business to produce these products with all the best qualities and then sell the product at the lowest possible cost. A business needs to find the perfect balance between the best product and the best price. This problem has been described as the "managers headache". With the high competition between companies and the easy accessibility of the different products, companies need to distinguish their product above other companies' products. These insights, that conjoint studies can give companies, will assist the producers to see which other attributes influence the consumer in general. This can show them how they can change or improve current products or guide them in designing new products.

When the accuracy of this aggregate level model was considered it was found that R^2 value decreased to 0.33. This is due to the error that occurs when the aggregate part-worth utilities are used to estimate all the respondents preferences toward the 18 profiles. In a case where the general populations share the same preference over all attributes, the R^2 value will show that the aggregate model will be a much better fit. For this example the general preferences of the population do vary so it will be obvious that the aggregate model will lead to greater errors. It is because of this reason that researchers are encouraged to evaluate these differences between the respondents preferences. If these differences can be determined they can be incorporated into the model and should allow the researcher to better predict the different respondents preferences.

The traditional approach, with it's simplified covariance structure can be extended by applying a multivariate approach which will allow for more complex covariance structures. While the aggregate model may be seen by some as statistically insufficient, since it just evaluates the global averages of the part-worth utilities, it has been found to deliver good results. The importance values that are typically determined in a conjoint study can be considered, at first glance, as an effective nonparametric approach. Another approach that can possibly be implemented and may lead to a better understanding of the respondents preferences towards the different types of industrial cleaners, is to include a random component to the fixed effect model. The next unit will briefly introduce the random effects model, and explain how this effect can possibly improve the traditional conjoint analysis method. This random component will then be added to the fixed effects model to form a linear mixed effects model.

3.3 Mixed effects model

A mixed effects model allows for more complex covariance structures. Section 3.3.1 will briefly introduce the random effects model after which the mixed effects model will be discussed. This technique will then be applied to the industrial cleaner conjoint analysis example.

3.3.1 Introduction to random components

The capabilities of the random regression model are far reaching and allows the researchers to model data that may have missing values, data that has time-varying or invariant covariates, and data of subjects measured on different occasions (Hedeker et al., 1991). When modeling some effects in a model as random, the researcher wants to draw a conclusion about the population from which the observed units were drawn, rather than about these particular units themselves, as is the case with fixed effect models (Snijders, 2005). When measuring consumer preference it is of vital importance to measure fluctuations over the population of respondents. As different respondents will most probably have different preference structures, the inclusion of random effects can contribute to the overall accuracy of the model by revealing more information about the sample of the dependent variables.

3.3.1.1 One-way random effects model

Consider the dependent variables y_{ij}, $i = 1, 2, ...k$; $j = 1, 2, ...n_i$ where k is the number of groups and n_i is the sample sizes for group i. A one-way random effects model, assuming normality, can be expressed as (Madsen and Thyregod, 2011):

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{3.16}$$

$$\varepsilon_{ij} \backsim N(0, \sigma^2)$$

$$\tau_i \backsim N(0, \sigma_\tau^2)$$

where

$\varepsilon'_{ij}s$ are mutually independent

τ_i's are mutually independent

τ_i's are independent of the $\varepsilon'_{ij}s$

σ^2 is the error variance

σ_τ^2 is the variance associated with the random components.

Let $\sum_{i=1}^{k} n_i = N$ be the total number of observations. When the sample sizes are equal for all the groups, $n_i = n$, the model will be referred to as a balanced model. For this model the parameters that need to be estimated are the random parameters, τ_i, (which will be different for each group) as well as the fixed parameters μ, σ^2, and σ_τ^2.

It is assumed that y_{ij} will then be normally distributed with an expected value $E(y_{ij}) = \mu$ and a covariance structure of

$$Cov(y_{ij}, y_{hl}) = \begin{cases} \sigma_\tau^2 + \sigma^2 & (ij) = (hl) \\ \sigma_\tau^2 & i = h, j \neq l \\ 0 & i \neq l \end{cases}$$

If $\mathbf{J}_{n_i}$ is an $n_i \times n_i$ matrix containing only 1's then the dependent variables will be normally distributed

$$\mathbf{y}_i \backsim \mathbf{N}_{n_i}(\boldsymbol{\mu}, \mathbf{V}_i) \tag{3.17}$$

where

$$\begin{aligned} \mathbf{V}_i &= \sigma^2 \mathbf{I}_{n_i} + \sigma_\tau^2 \mathbf{J}_{n_i} \\ &= \begin{pmatrix} \sigma_\tau^2 + \sigma^2 & \sigma_\tau^2 & \cdots & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma^2 & \cdots & \sigma_\tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\tau^2 & \sigma_\tau^2 & \cdots & \sigma_\tau^2 + \sigma^2 \end{pmatrix}. \end{aligned}$$

To estimate the unknown parameters the ANOVA approach can again be used. Given that the group means and the overall sample mean are given by

$$\overline{y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \text{ and } \overline{y} = \frac{1}{N} \sum_{i=1}^{k} \overline{y_i} \tag{3.18}$$

the sum of squares can be calculated by using Table 3.11.

Table 3.11 ANOVA table for one-way random effects model

Variation	Sum of Squares
Between Groups	$SSB = \sum_{i=1}^{k} n_i(\overline{y_i} - \overline{y})^2$
Within a Group	$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$
Total	$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$

Some important results that will be used when estimating the unknown parameters include the *signal to noise ratio* which expresses the heterogeneity between the groups in relation to the internal variation (Madsen and Thyregod (2011)):

$$\gamma = \frac{\sigma_\tau^2}{\sigma^2}$$

and the *shrinkage factor* which illustrates how much the grand mean contributes to the unit specific random effects estimates*:*

$$w_i(\gamma) = \frac{1}{1 + n_i\gamma} < 1.$$

In the balanced case the distributions of the sum of squares are (*from Results 3*):

$$\begin{aligned} \frac{SSE}{\sigma^2} &\backsim \chi^2_{N-k} \\ \frac{w(\gamma)SSB}{\sigma^2} &\backsim \chi^2_{k-1} \end{aligned}$$

and the resulting expected values of the mean squared estimates will be:

$$\begin{aligned} E(MSE) &= E\left(\frac{SSE}{N-k}\right) = \sigma^2 \\ E(MSB) &= E\left(\frac{SSB}{k-1}\right) = \sigma^2 + n_o\sigma_\tau^2 \end{aligned}$$

where

$$n_o = \frac{N - \sum_i \frac{n_i^2}{N}}{k-1}. \tag{3.19}$$

The estimates for the unknown parameters can then be determined by:

$$\widehat{\mu} = \overline{y}$$

$$\widehat{\sigma}^2 = MSE = \frac{SSE}{N-k}$$

$$\widehat{\sigma}_\tau^2 = \frac{MSB - \widehat{\sigma}^2}{n_o} = \frac{SSB/(k-1) - SSE/(N-k)}{n_o}.$$

► **Test for random components**

When considering modelling effects as random, the variances of these random effects should be significantly greater than zero in order to contribute to the model. When testing the significance of the variance component of the random effect, σ_τ^2, the following hypothesis should be considered:

$$H_0 \quad : \quad \sigma_\tau^2 = 0$$
$$H_A \quad : \quad \sigma_\tau^2 > 0$$

If the null hypothesis can be rejected then the variance of the random effect will be significantly greater than zero and the random effect will be a significant contributor to the model. This means that the variability measured among the randomly selected groups is significantly larger than the variability that can be explained by the error term alone. If $\sigma_\tau^2 = 0$, then all the random effects will be identically zero.

Under this null hypothesis the test statistics will have an F-distribution with $k-1$ and $N-k$ degrees of freedom (*From Result 6*):

$$F^* = \frac{MSB}{MSE} = \frac{SSB/k-1}{SSE/N-k} \backsim F_{k-1,N-k} \tag{3.20}$$

For a balanced design the distribution under the alternative hypothesis will be:

$$F^* = \frac{MSB}{MSE} = \frac{SSB/k-1}{SSE/N-k} \backsim (1+n\gamma)F_{k-1,N-k} \tag{3.21}$$

Example 2 *Consider the following example obtained from Neter et al., 1996. The study was interested in quantifying and accounting for the effect of different interviewers in the evaluation of applicants applying for a job. 5 different interviewers where chosen to*

randomly interview the 4 applicants. The model can be represented by:

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \varepsilon_{ij} \\ \alpha &\backsim N(0, \sigma_\alpha^2) \\ \varepsilon &\backsim N(0, \sigma^2) \\ \text{then } Y &\backsim N(\mu, \sigma_\alpha^2 + \sigma^2) \end{aligned}$$

where $n = 4$ (different applicants), thus there will be 20 observations and Y will be the rating (dependent variable) given to each applicant. It is important to remember that there are two sources of variation. The data collected is displayed in Table 3.12.

Table 3.12 *Interviewer data*

Rating	*Officer*	*Rating*	*Officer*	*Rating*	*Officer*	*Rating*	*Officer*	*Rating*	*Officer*
76	*1*	*59*	*2*	*49*	*3*	*74*	*4*	*66*	*5*
65	*1*	*75*	*2*	*63*	*3*	*71*	*4*	*84*	*5*
85	*1*	*81*	*2*	*61*	*3*	*85*	*4*	*80*	*5*
74	*1*	*67*	*2*	*46*	*3*	*89*	*4*	*79*	*5*

The data is visually represented by Figure 3.3.1 and Figure 3.3.2:

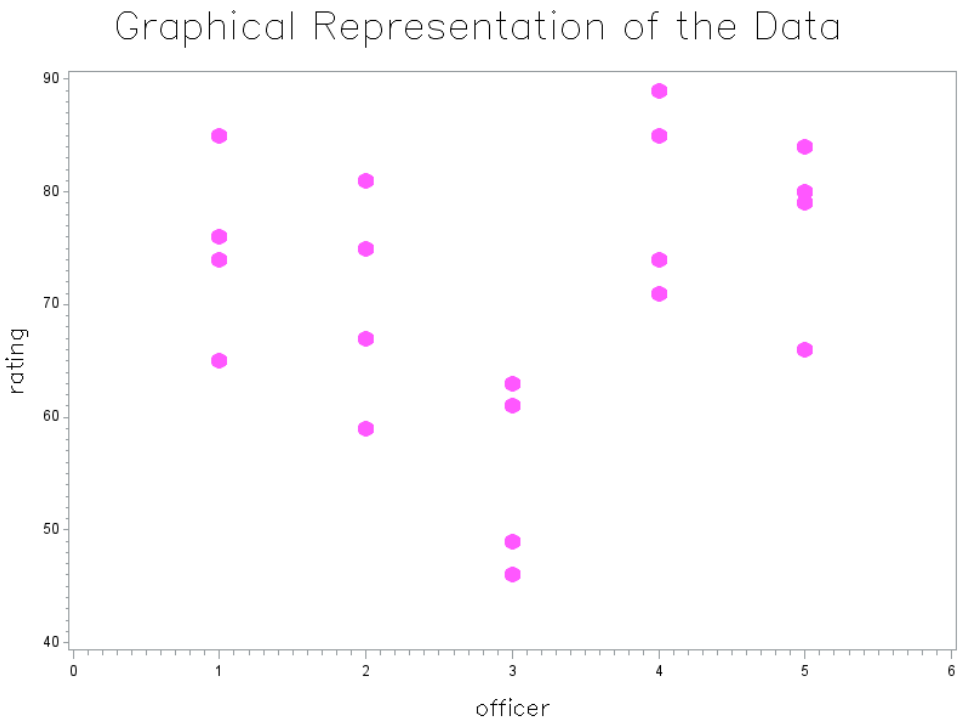


Figure 3.3.1 *Officer ratings*

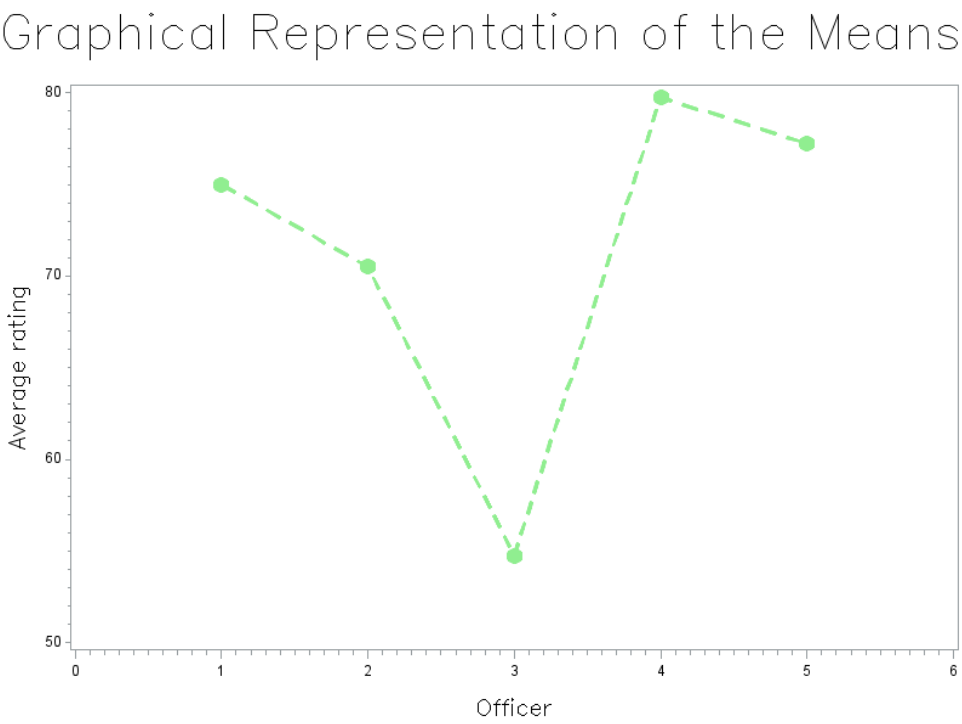


Figure 3.3.2 *Overall mean rating for each officer*

Figure 3.3.1 is a representation of the ratings that each of the officers gave the four applicants who they interviewed. Figure 3.3.2 represents the overall mean rating each officer gave the applicants. It is clear that the different officers have very different overall means, for instance, it seems like officer 3 is very strict and rated the four applicants very poorly. The question however, is if the differences between the officers ratings can be explained by σ^2 alone or if a second variance needs to be considered.

The following hypotheses were tested and the results were obtained using SAS 9.3 (Program 2):

$$\begin{aligned} H_0 &: \sigma_\alpha^2 = 0 \\ H_A &: \sigma_\alpha^2 > 0. \end{aligned}$$

The calculated parameter estimates where $\widehat{\mu} = 71.45, \widehat{\sigma}^2 = 73.28$ and $\widehat{\sigma}_\alpha^2 = 80.41$. To test the given hypothesis the F-statistic and the critical value were compared. With $F = 5.3890$ and $F_\alpha = 0.3273$ the null hypothesis was rejected and the researcher could then conclude that, at a 5% level of significance, the variability measured among the officers is significantly larger than the variability that can be explained by σ^2 alone. Computing the influence of σ_α^2 the interclass correlation can be calculated by:

$$\frac{\sigma_\alpha^2}{\sigma^2 + \sigma_\alpha^2} = \frac{80.41}{73.28 + 80.41} = 0.52$$

Thus 52% of the variability in the data is explained by variance of the random effects, σ^2_α, which represents the variability between the interviewers.

♦

This example shows the power within a random effect model. If the traditional conjoint analysis model, as discussed in chapter 3.2, can be extended to include these measures of differences between the respondents, the researcher can gain more insight regarding the variability within the model.

3.3.2 Basic properties of the linear mixed model

Mixed effects models are used to describe the relationships between a response variable and some covariates in the data that are grouped according to one or more classification factors. Some examples of these types of grouped data include longitudinal data, repeated measures data, multilevel data and block designs (Pinheiro and Bates, 2004).

The linear mixed model is an extension of the classical linear model, the only difference being the inclusion of a random effect in the model already containing fixed effects. Deciding which effects to assign fixed and random depends on the context of the problem, the question of interest and how the data was gathered.

The main use of fixed effect models is modeling the mean of the dependent variable as well as modelling the effect that different independent variables may have on the dependent variable. The random effects are used to model the variance-covariance structure of the dependent variable. Random effects simplify the otherwise difficult task of specifying the distinct elements of the variance of the dependent variable.

Since it cannot be assumed that the preference structure will be the same for all respondents in a conjoint study, a model needs to be built that takes these differences into consideration. The basic properties of a mixed model will be investigated first, followed by an application of modelling preference data using a mixed model approach.

3.3.3 Linear mixed model

The linear mixed model (LMM) has the following form (Jensen et al., 2007):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \text{ where } i = 1, 2, ...m \tag{3.22}$$

with n the number of observations, p the number of fixed effects, q the number of random effects and m the number of independent responses. Let

$\mathbf{y}_i : n \times 1$: response vector;

$\mathbf{X}_i : n \times p$: known design matrix for the fixed effects;

$\boldsymbol{\beta} : p \times 1$: vector of unknown fixed effects;

$\mathbf{Z}_i : n \times q$: known design matrix for the random effects;

$\boldsymbol{\varepsilon}_i : n \times 1$: vector of error terms;

$\mathbf{b}_i : q \times 1$vector of unknown random effects.

Also note that the random effects and the error terms are assumed to be normally distributed as:

$$\begin{aligned} \mathbf{b}_i &\backsim N_q(\mathbf{0}, \mathbf{D}_i) \\ \boldsymbol{\varepsilon}_i &\backsim N_n(\mathbf{0}, \mathbf{R}_i) \end{aligned}$$

where $\mathbf{D}_i : q \times q$ positive definite diagonal variance-covariance matrix of $\mathbf{b}_i$ (since the random effects are assumed uncorrelated) and $\mathbf{R}_i$: $n \times n$ positive definite variance-covariance matrix of residual terms. Since the errors are assumed independent, the error variance matrix simplifies to $\mathbf{R}_i = \sigma^2 \mathbf{I}_n$ and the variance-covariance structure, $\mathbf{D}_i$, is considered fixed for all the respondents, hence $\mathbf{D}_i = \mathbf{D}$. Also note that since it is assumed that the error terms and the random effects are uncorrelated.

The linear mixed model is very flexible in the sense that it can allow the errors to be independent or correlated to each other. In the case where a study wants to consider the errors to be correlated it is advised to keep the structure simple in order to reduce the number of parameters that need to be estimated. The random effects are also assumed to be independent, since the variance-covariance matrix, $\mathbf{D}$, is considered diagonal, but the model also allows for an extension in this regard.

It can be seen that the mixed effect model allows for two levels of correlation for a measurement in a profile. The random effects causes all measurements among the profiles to be correlated to each other while the second order of correlation results from the within-profile variance-covariance matrix $\mathbf{R}_i$. The dependent variable will then be normally distributed as (Jensen et al., 2007):

$$\mathbf{y}_i \backsim \mathbf{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$$

where

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$$

is an $n \times n$ positive definite variance-covariance matrix.

3.3.3.1 Stacked structure

Another way of representing this model in matrix notation is through the *Stacked form*. This model literally stacks all the respondents matrices on top of the others to form one big matrix that contains the information of all m respondents. This method simplifies the calculations that will follow and will be illustrated as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{3.23}$$

where

$\mathbf{y}$ is a $N \times 1$ stacked vector containing the responses for all profiles with $N = nm$;

$\mathbf{X}$ is a $N \times (p+1)$ stacked matrix of $\mathbf{X}_i's$;

$\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects;

$\mathbf{Z}$ is a $N \times mq$ block design such that $\mathbf{Z} = diag(\mathbf{Z}_i)$;

$\mathbf{b}$ is a $mq \times 1$ matrix containing the vectors of random effects for each response and is assumed to be normally distributed as $\mathbf{b} \backsim \mathbf{N}(\mathbf{0}, \mathbf{D})$ where $\mathbf{D} = \mathbf{Diag}(\mathbf{D}_i)$;

$\boldsymbol{\varepsilon}$ is a $N \times 1$ matrix containing the residuals for all m responses and is assumed to be normally distributed as $\boldsymbol{\varepsilon} \backsim \mathbf{N}(\mathbf{0}, \mathbf{R})$ where $\mathbf{R} = \mathbf{Diag}(\mathbf{R}_i)$.

Visually this model can be represented as:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{Z}_m \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{pmatrix}.$$

Note that $\mathbf{R}$ and $\mathbf{D}$ are both symmetric block diagonal matrices and $cov(\boldsymbol{\varepsilon}, \mathbf{b}) = \mathbf{0}$. The corresponding marginal model will then be given by (Jensen et al., 2007):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

where $\boldsymbol{\varepsilon}^* \backsim \mathbf{N}(\mathbf{0}, \mathbf{V})$. Then

$$\mathbf{y} \backsim \mathbf{N}_N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \tag{3.24}$$

where

$$\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R} \tag{3.25}$$

which is an $N \times N$ positive definite matrix. The conditional model form can then be determined by using Theorem 3.1:

$$\mathbf{y}|\mathbf{b} \backsim \mathbf{N}_N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R}) \tag{3.26}$$

Theorem 3.1 *If* $\mathbf{y} : N \times 1$ *is a stacked vector of responses and* $\mathbf{b}$ *:*$mq \times 1$ *is the matrix of random effects, then the conditional distribution of* $\mathbf{y}$ *conditioned on* $\mathbf{b}$*, will be given by*

$$\mathbf{y}|\mathbf{b} \backsim \mathbf{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R}). \tag{3.27}$$

where $\mathbf{X} : N \times p$ *stacked matrix of design matrices* $\mathbf{X}_i's$ *for the fixed effects,* $\mathbf{Z} : N \times mq$ *diagonally stacked matrix of design matrices* $\mathbf{Z}_i's$ *and* $\mathbf{D}$ *and* $\mathbf{R}$ *are variance-covariance*

matrices.

Proof: Given the distribution of $\mathbf{y}$ and $\mathbf{b}$

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{b}\end{pmatrix} = \left\{\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}; \begin{pmatrix}\mathbf{V} & \mathbf{ZD}\\ \mathbf{DZ}' & \mathbf{D}\end{pmatrix}\right\}$$

using *Result 8* the distribution of $\mathbf{y}$, conditional on $\mathbf{b}$ will be normally distributed with a mean

$$\begin{aligned}\mu_{\mathbf{y}|\mathbf{b}} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{ZDD}^{-1}(\mathbf{b}-\mathbf{0})\\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb}\end{aligned}$$

and a variance-covariance structure of

$$\begin{aligned}\Sigma_{\mathbf{y}|\mathbf{b}} &= \mathbf{V} - \mathbf{ZDD}^{-1}\mathbf{DZ}'\\ &= \mathbf{ZDZ}'+\mathbf{R} - \mathbf{ZDZ}'\\ &= \mathbf{R}.\end{aligned}$$

■

3.3.4 Parameter estimation

The different parameters that will need estimation in a linear mixed effects model includes the fixed effects as well as the random effects. Many researchers do not like the term "estimation" when working with random effects, but would rather call the process "the prediction" of random effects as it does suit the procedure better. Hence the estimation of the fixed effects will be given by the Best Linear Unbiased Estimates (BLUE) and the prediction of the random effects will be given by the Best Linear Unbiased Predictors (BLUP). The parameter estimation in mixed effects models can follow two directions. The first one that will be considered is the situation where the variance components are known. The variance components are usually unknown which will results in a more complicated estimation and prediction process; but by illustrating the case where the components are known, a better understanding of the process can be obtained.

3.3.4.1 Estimation when the variance components are known

When $\mathbf{V}$ is known it can be assumed that $\mathbf{D}$ and $\mathbf{R}$ will also be known (3.25). The only parameters that needs estimation is the fixed effects, $\boldsymbol{\beta}$, and the random effects, $\mathbf{b}$. The estimation of the fixed effects will be considered first.

► **Fixed parameter estimation**

By simplifying the distribution of the mixed effect model to a marginal model the estimates can easily be obtained using the maximum likelihood (ML) approach. The marginal model will be:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \tag{3.28}$$

where

$$\varepsilon^* \backsim N(\mathbf{0}, \mathbf{V}).$$

Since $\mathbf{y}$ is normally distributed the likelihood function can be represented as:

$$L = \frac{1}{(2\pi)^{n/2}|\mathbf{V}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{3.29}$$

In order to find the estimates, the parameters will be evaluated against the loglikelihood which is expressed as:

$$\begin{aligned} LogL &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad (3.30)\\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}[\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - \\ &\qquad \mathbf{X}'\boldsymbol{\beta}'\mathbf{V}^{-1}\mathbf{y} + \mathbf{X}'\boldsymbol{\beta}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}]. \end{aligned}$$

The parameter estimates, $\widehat{\boldsymbol{\beta}}$,will then be determined by setting the first order derivative of the loglikelihood in terms of $\boldsymbol{\beta}$, equal to zero, as follows:

$$\begin{aligned} \frac{\partial LogL}{\partial \boldsymbol{\beta}} &= -\frac{1}{2}\left[0 - \mathbf{y}'\mathbf{V}^{-1}\mathbf{X} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}\right] \\ &= -\frac{1}{2}\left[-2\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}\right] \\ &= 0 \end{aligned}$$

which results in the closed form expression:

$$\therefore \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{3.31}$$

► **Random parameter prediction**

Since the assumptions for the random effects differ from those of fixed effects, the method of estimation also differs. The fixed effects are considered to be constant, while the random effects are considered as effects coming from a population of effects. As Mc Culloch et al., 2008 stated, it is the population that is the extra assumption, compared to fixed effects, and it is hoped to lead to an estimation method for random effects being

an improvement over that of fixed effects.

To obtain the prediction of the random effect, the expected value of the conditional distribution of $\mathbf{b}$, conditioned on $\mathbf{y}$, needs to be determined:

Theorem 3.2 *If* $\mathbf{y} : N \times 1$ *is a stacked vector of responses and* $\mathbf{b} : mq \times 1$ *is the matrix of random effects, then the conditional distribution of* $\mathbf{b}$ *conditioned on* $\mathbf{y}$,*will be given by*

$$\mathbf{b}|\mathbf{y} \backsim \mathbf{N}(\mathbf{DZ'V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{D} - \mathbf{DZ'V}^{-1}\mathbf{ZD})$$

where $\mathbf{X} : N \times (p+1)$ *stacked matrix of design matrices* $\mathbf{X}_i's$ *for the fixed effects,* $\mathbf{Z} : N \times mq$ *diagonally stacked matrix of design matrices* $\mathbf{Z}_i's$ *and* $\mathbf{D}$ *and* $\mathbf{R}$ *are variance-covariance matrices.*

Proof: Given the distribution of $\mathbf{y}$ and $\mathbf{b}$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{b} \end{pmatrix} = \left\{ \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} ; \begin{pmatrix} \mathbf{V} & \mathbf{ZD} \\ \mathbf{DZ'} & \mathbf{D} \end{pmatrix} \right\}$$

using *Result 8* the distribution of $\mathbf{b}$, conditional on $\mathbf{y}$ will be normally distributed with a mean

$$\begin{aligned} \mu_{b|y} &= \mathbf{0} + \mathbf{DZ'V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{DZ'V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

and a variance-covariance structure of

$$\Sigma_{b|\mathbf{y}} = \mathbf{D} - \mathbf{DZ'V}^{-1}\mathbf{ZD}. \tag{3.32}$$

■

The prediction of the random effect will then be:

$$\widehat{\mathbf{b}} = E[\mathbf{b}|\mathbf{y}] = \mathbf{DZ'V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{3.33}$$

3.3.4.2 Estimation when the variance components are unknown

When $\mathbf{V}$ is unknown, the variance-covariance structures $\mathbf{D}$ and $\mathbf{R}$ must first be determined before the unknown parameters $\boldsymbol{\beta}$ and $\mathbf{b}$ can be estimated. This is no longer a straight forward calculation as the estimates will no longer have closed forms. A solution to this potential problem is applying an iterative approach to find the optimal solution.

One method that can be used to calculate these parameters is the Expectation Maximization (EM) algorithm. The EM algorithm is an iterative algorithm typically used to compute the ML estimates or the restricted maximum likelihood (REML) estimates in the presence of missing data or unobservables. The term 'expectation maximization' algorithm was coined by Dempster, Laird, and Rubin in their fundamental paper published in 1977. The use of the EM algorithm has increased drastically over the last few decades and the development of computer capabilities extended the growth even more. Wu, 2010 list some advantages of this algorithm:

• The convergence is stable and with each iteration the likelihood is maximized.
• The M-Step (which is the second step in the process and will be discussed in more detail shortly) involves only complete data maximization so it is often compositionally straight forward; and
• The algorithm is very general and can be used in almost any maximum likelihood estimation problem with unknown quantities. In the mixed model case, the random effects are considered to be the missing values.

The disadvantage of the EM algorithm is that it may be slow to converge, although there are many suggestions available in the literature to speed up the process. Meng and Van Dyk (1997) proposes strategies that aim to make the EM algorithm converge faster while maintaining its simplicity and stability.

▶ The theoretical EM framework

The EM algorithm is an iterative method used to find the MLE's of a statistical model where there is dependency on unobserved latent variables. The theoretical framework of the EM algorithm follows in two basic steps. The first step is the E-Step. This step creates a function Q, which represents the expectation of the loglikelihood. This step then further evaluates the function Q at the current parameter estimates. The second step is called the M-Step. This step then maximizes the simplified likelihoods. Since the random effects are treated as the "missing-data", and the responses are referred to as the "observed-data", the "complete-data" can be represented by:

$$(\mathbf{y}, \mathbf{b}) = \{(\mathbf{y}_i, \mathbf{b}_i); i = 1, 2, ...m\}. \tag{3.34}$$

Let k $(k = 0, 1, 2...)$ denote the number of iterations, that will continue until convergence is met, then after choosing a initial set of starting values, the EM process continues as follows (Wu, 2010):

E-step:

At iteration k, this step computes the conditional expectation of the complete-data log-likelihood given the observed data and the current parameter estimates. The conditional expectation is represented in the form of the function Q:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E\left[\log L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{b})|\mathbf{y}, \boldsymbol{\theta}^{(k)}\right].$$

M-step:
This step maximizes the conditional expectation, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, found in the E-Step with respect to the unknown parameters, to produce updated estimates of the parameters

$$Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}).$$

This process is iterated repeatedly until convergence is met. It is important to note that the EM algorithm only converges to local maxima. Hence, when the likelihood has multiple nodes a good starting value should be chosen and different starting values should be considered to ensure that the algorithm converges to a global maxima. It is no surprise that the EM algorithm converges very slowly if poor choices of starting values are made.

▶ EM algorithm for the ML approach
Since the E-step starts off by computing the conditional expectation of the complete-data loglikelihood, the first step will be setting up this complete-data loglikelihood. Please note that model (3.22) will be used in these calculations, and not the stacked model that was considered up to now. Given that complete-data represented by:

$$(\mathbf{y}, \mathbf{b}) = \{(\mathbf{y}_i, \mathbf{b}_i); i = 1, 2, ...m\}.$$

The likelihood function of the observed data linear mixed model with multiple responses, (3.22) can be represented as,

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{m} f(\mathbf{y}_i|\boldsymbol{\theta})$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})$ represents a vector containing all the parameters that will need to be estimated, and $\boldsymbol{\eta}$ represents the vector containing the distinct parameters of the variance-covariance matrices $\mathbf{R}_i$ and $\mathbf{D}_i$. By using introductory concepts of probability for the mixed model the likelihood can be partitioned into two subsets:

$$f(\mathbf{y}_i|\boldsymbol{\theta}) = f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \mathbf{R}_i) f(\mathbf{b}_i|\mathbf{D}_i).$$

Since it is known from (3.27) that $\mathbf{y}_i|\mathbf{b}_i \backsim \mathbf{N}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i)$ and from the assumptions about a linear mixed model that $\mathbf{b}_i \backsim N_q(\mathbf{0}, \mathbf{D}_i)$, where $\mathbf{D}_i$ remains the same for all

individuals, it follows that:

$$\begin{aligned} f(\mathbf{y}_i|\mathbf{b}_i,\boldsymbol{\beta},\mathbf{R}_i) &= \frac{1}{(2\pi)^{-n/2}|\mathbf{R}_i|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i)'\mathbf{R}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i)\right\} \\ &= \frac{1}{(2\pi)^{-n/2}|\mathbf{R}_i|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{e}_i)'\mathbf{R}_i^{-1}(\mathbf{e}_i\right\} \end{aligned} \tag{3.35}$$

where $\mathbf{e}_i = \mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i$ and

$$f(\mathbf{b}_i|\mathbf{D}) = \frac{1}{(2\pi)^{q/2}|\mathbf{D}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathbf{b}_i\mathbf{D}^{-1}\mathbf{b}_i\right\}. \tag{3.36}$$

The loglikelihood of respondent i will then follow as:

$$LogL(\boldsymbol{\theta}|\mathbf{y}_i,\mathbf{b}_i) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{R}_i| - \frac{1}{2}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)'\mathbf{R}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)$$

$$-\frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\mathbf{b}_i\mathbf{D}^{-1}\mathbf{b}_i \tag{3.37}$$

Once the complete-data loglikelihood is obtained, the E- and M-steps can commence.

The E-step:

The conditional expectation of the complete-data loglikelihood will be represented by the function Q as:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= E\left[\log L(\boldsymbol{\theta}|\mathbf{y},\mathbf{b})|\mathbf{y},\boldsymbol{\theta}^{(k)}\right] \\ &= E\left[\sum_{i=1}^{m}\left\{\log f(\mathbf{y}_i|\mathbf{b}_i,\boldsymbol{\beta},\mathbf{R}_i) + \log f(\mathbf{b}_i|\mathbf{D})|\mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right\}\right] \end{aligned}$$

where $\log L(\boldsymbol{\theta}|\mathbf{y}_i,\mathbf{b})$ is represented by (3.37). The E-step can then be simplified to the computation of the following sufficient statistics of the variance-covariance parameters, $\boldsymbol{\eta}$, in the covariance matrices $\mathbf{R}_i$ and $\mathbf{D}$ (Wu, 2010):

$$\sum_{i=1}^{m} E\left(\mathbf{e}_i'\mathbf{e}_i|\mathbf{y}_i,\widehat{\boldsymbol{\theta}}^{(k)}\right) = \sum_{i=1}^{m}\left[\widehat{\mathbf{e}}_i'^{(k)}\widehat{\mathbf{e}}_i^{(k)} + tr\left[cov(\mathbf{e}_i|\mathbf{y}_i,\widehat{\boldsymbol{\theta}}^{k})\right]\right]$$

and

$$\sum_{i=1}^{m} E\left[\mathbf{b}_i\mathbf{b}_i'|\mathbf{y}_i,\widehat{\boldsymbol{\theta}}^{(k)}\right] = \sum_{i=1}^{m}\left[\widehat{\mathbf{b}}_i^{(k)}\widehat{\mathbf{b}}_i'^{(k)} + cov(\mathbf{b}_i|\mathbf{y}_i,\widehat{\boldsymbol{\theta}}^{k})\right]$$

where

$$\widehat{\mathbf{e}}_i = \mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}-\mathbf{Z}_i\widehat{\mathbf{b}}_i^{(k)}$$

$$\widehat{\mathbf{b}}_i^{(k)} = D(\widehat{\boldsymbol{\eta}}^{(k)})\mathbf{Z}_i'\mathbf{V}_i^{-1}(\widehat{\boldsymbol{\eta}}^{(k)})(\mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)})$$

$$\mathbf{V}_i^{-1}(\widehat{\boldsymbol{\eta}}^{(k)}) = \mathbf{Z}_i D(\widehat{\boldsymbol{\eta}}^{(k)})\mathbf{Z}_i' + \mathbf{R}_i(\widehat{\boldsymbol{\eta}}^{(k)})$$

The M-step:

For this step the parameters that will maximize the conditional expectation, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, must be calculated to produce the updated parameters.. The updated estimates of the parameters can be determined as follows. Firstly, the estimated fixed effects will be:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left[\sum_{i=1}^{m} \mathbf{X}_i'\widehat{\mathbf{V}}_i^{-1}(\widehat{\boldsymbol{\eta}}^{(k)})\mathbf{X}_i\right]^{-1} \sum_{i=1}^{m} \mathbf{X}_i'\mathbf{V}_i^{-1}(\widehat{\boldsymbol{\eta}}^{(k)})\mathbf{y}_i \tag{3.38}$$

This follows in the exact same way as the parameter estimation when $\mathbf{V}$ was known. Estimated variance-covariance values are used to determine the estimated fixed effects. The iterations will result in more accurate results with each iteration. The prediction of the variance components can then be determined by the calculation of the conditional expectations within $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$. Firstly it is important to consider the following theorem.

Theorem 3.3 *Given than* $\mathbf{y}$ *:*$N \times 1$ *is a stacked matrix of responses and* $\boldsymbol{\varepsilon}$ *:*$N \times 1$ *is a stacked matrix of residuals, then the conditional distribution of* $\boldsymbol{\varepsilon}$*, conditioned on* $\mathbf{y}$ *will be given by*

$$\boldsymbol{\varepsilon}|\mathbf{y} \backsim \mathbf{N}(\mathbf{R}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{R} - \mathbf{R}'\mathbf{V}^{-1}\mathbf{R})$$

where $\mathbf{X} : N \times p$ *stacked matrix of design matrices* $\mathbf{X}_i's$ *for the fixed effects,* $\mathbf{Z} : N \times mq$ *diagonally stacked matrix of design matrices* $\mathbf{Z}_i's$ *and* $\mathbf{D}$ *and* $\mathbf{R}$ *are variance-covariance matrices.*

Proof: Given the distribution of $\mathbf{y}$ and $\boldsymbol{\varepsilon}$

$$\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\varepsilon}\end{pmatrix} = \left\{\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}; \begin{pmatrix}\mathbf{V} & \mathbf{R}\\ \mathbf{R}' & \mathbf{R}\end{pmatrix}\right\}$$

using result 8 the distribution of $\boldsymbol{\varepsilon}$, conditional on $\mathbf{y}$ will be normally distributed with a mean

$$\begin{aligned}\mu_{\boldsymbol{\varepsilon}|\mathbf{y}} &= \mathbf{0} + \mathbf{R}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\\ &= \mathbf{R}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

and a variance-covariance structure of

$$\Sigma_{\boldsymbol{\varepsilon}|\mathbf{y}} = \mathbf{R} - \mathbf{R}'\mathbf{V}^{-1}\mathbf{R}$$

■

Thus for the LMM the parameter estimations simplifies to the following computations (Wu, 2010):

$$\widehat{\sigma}^{2(k+1)} = \frac{\sum_{i=1}^{m} E\left[\mathbf{e}_i'^{(k)}\mathbf{e}_i^{(k)}|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^k\right]}{\sum_{i=1}^{m} n} = \frac{\sum_{i=1}^{m} E\left[\mathbf{e}_i'^{(k)}\mathbf{e}_i^{(k)}|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^k\right]}{N} \tag{3.39}$$

where using *Result 9:*

$$\begin{aligned}
&\sum_{i=1}^{m} E\left[\mathbf{e}_i'^{(k)}\mathbf{e}_i^{(k)}|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^k\right] \\
&= \sum_{i=1}^{m}\left[\widehat{\mathbf{e}}_i'^{(k)}\widehat{\mathbf{e}}_i^{(k)} + tr\left[cov(\mathbf{e}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^k)\right]\right] \\
&= \sum_{i=1}^{m}\left[\widehat{\mathbf{e}}_i'^{(k)}\widehat{\mathbf{e}}_i^{(k)} + tr\left[\widehat{\sigma}^{2(k)}\mathbf{I}_n + \widehat{\sigma}^{4(k)}\mathbf{I}_n'\widehat{\mathbf{V}}_i^{-1}(\widehat{\eta}^{(k)})\mathbf{I}_n\right]\right] \\
&= \sum_{i=1}^{m}\left[\widehat{\mathbf{e}}_i'^{(k)}\widehat{\mathbf{e}}_i^{(k)} + tr\left[\widehat{\mathbf{R}}_i^{(k)} - \widehat{\mathbf{R}}_i^{(k)}\mathbf{I}_n'\widehat{\mathbf{V}}_i^{-1}(\widehat{\eta}^{(k)})\mathbf{I}_n\widehat{\mathbf{R}}_i^{(k)}\right]\right]
\end{aligned}$$

The variance components of the random variables can be determined in a similar way by using *Result 9* and (3.32):

$$\begin{aligned}
\widehat{\mathbf{D}}^{(k+1)} &= \frac{\sum_{i=1}^{m} E\left[\mathbf{b}_i\mathbf{b}_i'|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right]}{m} \\
&= \frac{\sum_{i=1}^{m}\left[\widehat{\mathbf{b}}_i^{(k)}\widehat{\mathbf{b}}_i'^{(k)} + cov(\mathbf{b}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^k)\right]}{m} \\
&= \frac{\sum_{i=1}^{m}\left[\widehat{\mathbf{b}}_i\widehat{\mathbf{b}}_i' + (\widehat{\mathbf{D}}^{(k)} - \widehat{\mathbf{D}}^{(k)}\mathbf{Z}_i'\widehat{\mathbf{V}}_i^{-1(k)}\mathbf{Z}_i\widehat{\mathbf{D}}^{(k)})\right]}{m}
\end{aligned} \tag{3.40}$$

The final step:
If convergence is reached at iteration k, then set $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(k+1)}$, $\widehat{\mathbf{D}} = \widehat{\mathbf{D}}^{(k+1)}$, $\widehat{\mathbf{R}} = \widehat{\sigma}^{2(k+1)}\mathbf{I}_n$ and $\widehat{\mathbf{b}} = \widehat{\mathbf{b}}^{(k+1)}$, otherwise increase k by 1 and return to the E-Step.

Example 3 *(SAS Program 3) Consider the following example obtained from Lasaffre and Lawson, 2012. A study was done to investigate the cholesterol intake between subsidiaries corrected for age and gender. They wanted to find out how much of the subsidiary*

variability in cholesterol intake could be explained by age and gender. The model was represented as follows:

$$y_{ij} = \beta_o + \beta_1 age_{ij} + \beta_2 gender_{ij} + b_{oi} + \varepsilon_{ij}$$

where y_{ij} *is the cholesterol intake of the* i^{th} *subject in the* j^{th} *subsidiary. Further it was assumed that* $b_{oi} \backsim N(\mathbf{0}, \sigma^2_{b_o})$ *and* $\varepsilon_{ij} \backsim N(0, \sigma^2)$. *Using the EM algorithm the following results where obtained:*

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 421.71 \\ -0.744 \\ -55.16 \end{bmatrix}$$

$$\widehat{\boldsymbol{\sigma}^2_b} = 94.23.$$

The estimated parameters indicate that age does not have a great influence on the cholesterol intake since it does not contribute greatly to the model, but gender definitely did.

♦

For more complex data structures, methods that can be used to determine the estimated values include the Restricted Maximum Likelihood (REML) approach and the Newton-Rhapson (NR) method.

▶ EM algorithm for the REML approach

When using the REML approach to estimate the unknown parameters the same steps can be followed as when using the ML approach. The difference however is that the REML approach partitions the likelihood into two parts where one will be free of fixed effects. The part of the likelihood function that does not contain the fixed effects will then be maximized to obtain the unknown variance components (Corbeil and Searle, 1979). Using these estimates, the part of the likelihood that was not used in the first step of the procedure will be maximized with respect to the obtained variance estimates to determine the estimates for the fixed effects. It is advisable to use REML estimates for more complex data since the variances are generally non-negative and the REML method is uninfluenced by outliers.

▶ Computing estimates using the NR approach

Another popular method used to estimate the unknown estimates is the Newton-Rhapson (NR) method. This is also an iterative approach, often employed in statistical software packages, which uses a root-finding technique, and can be summarized as follows (McCulloch et al., 2008). Starting with a function $\mathbf{f}(\boldsymbol{\theta})$, this method aims to find the

root of

$$\frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{3.41}$$

which will hopefully be a maximum. (3.41) can then be expanded about $\boldsymbol{\theta}_o$ as

$$\frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{f}'(\boldsymbol{\theta}) \approx \mathbf{f}'(\boldsymbol{\theta}_0) + \frac{\partial^2 \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{3.42}$$

Equating (3.42) to zero

$$\mathbf{f}'(\boldsymbol{\theta}_0) + \frac{\partial^2 \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \mathbf{0} \tag{3.43}$$

and solving for the roots of (3.43), results in

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 - \left[\frac{\partial^2 \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} \mathbf{f}'(\boldsymbol{\theta}_0).$$

The root can then be calculated iteratively to refine the estimates

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - \left[\frac{\partial^2 \mathbf{f}(\boldsymbol{\theta}^{(\mathbf{m})})}{\partial \boldsymbol{\theta}^{(m)} \partial \boldsymbol{\theta}^{(m)\prime}}\right]^{-1} \mathbf{f}'(\boldsymbol{\theta}^{(\mathbf{m})}).$$

Lindstrom and Bates (1988) discussed the differences between the EM algorithm and the NR approach. Although both approaches are relatively quick in estimating the unknown parameter it is the NR method that converges in less iterations than the EM algorithm and is consistent in it's convergences. Another big advantage of the NR approach is the objective of the convergence criteria. While the convergence criteria implemented for the EM algorithm is the size in the change of the parameter estimates from one iteration to the next, this is not considered by some as a pure indication of actual convergences but rather of a lack of progress. The NR approach uses a orthogonality criteria that can easily be calculated from iteration to iteration. The NR method can also be easily adapted to handle most common extensions of mixed models.

Example 4 *(SAS Program 3) Consider the previously discussed example based on the study to investigate the cholesterol intake between subsidiaries corrected for age and gender. The model was represented as follows:*

$$y_{ij} = \beta_o + \beta_1 age_{ij} + \beta_2 gender_{ij} + b_{oi} + \varepsilon_{ij}$$

where y_{ij} is the cholesterol intake of the i^{th} subject in the j^{th} subsidiary. Further it was assumed that $b_{oi} \backsim N(\mathbf{0}, \sigma^2_{b_o})$ and $\varepsilon_{ij} \backsim N(0, \sigma^2)$. Using the NR algorithm the following

results where obtained:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 423.98 \\ -0.5719 \\ -54.2712 \end{bmatrix}$$

$$\widehat{\boldsymbol{\sigma}_b^2} = 110.66$$

These estimates are very close to those obtained by using the EM algorithm, in example 3, and the results of the study was exactly the same.

♦

3.3.5 Hypothesis test

When modelling data as a mixed effects model the estimates that will need to be determined include the fixed effects, $\boldsymbol{\beta}'s$, as well as the variance components of the random effects, $\boldsymbol{\eta}$. It is necessary to evaluate these parameters to ensure that they are significantly different than zero. If this is not the case then these estimates will not contribute significantly to the model. Tests can be applied to the fixed effects as well as the variance components of the random effects to identify whether they can be included in the model or if they do not contribute to the overall study.

3.3.5.1 Test for fixed effects

There are just a few tests mentioned in the literature used to test the significance of the estimated fixed effects parameters. The first step to testing the significance is choosing a contrast matrix $\mathbf{L}$, which is a vector defining an estimable linear combination of $\boldsymbol{\beta}'s$, to test the hypothesis:

$$\begin{aligned} H_o &: \mathbf{L}\boldsymbol{\beta} = 0 \\ H_a &: \mathbf{L}\boldsymbol{\beta} \neq 0 \end{aligned}$$

The t-test, Wald test and the F-test's can be used to tests these hypotheses. The particular test statistics can be calculated as follows:

t-test statistic $\quad t = \frac{\mathbf{L}\widehat{\boldsymbol{\beta}}}{\sqrt{\mathbf{L}(\mathbf{X}'\widehat{V}^{-1}\mathbf{X})^{-1}\mathbf{L}'}}$

Wald test statistic $\quad Wald = \widehat{\boldsymbol{\beta}}'\mathbf{L}'\left\{\mathbf{L}(\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{L}'\right\}\mathbf{L}\widehat{\boldsymbol{\beta}}$

F-test statistic $\quad F = \frac{\widehat{\boldsymbol{\beta}}'\mathbf{L}'\left\{\mathbf{L}(\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{L}'\right\}\mathbf{L}\widehat{\boldsymbol{\beta}}}{rank(\mathbf{L}(\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{L}')}.$

The Wald test statistic will have an approximate χ^2 distribution with $rank(\mathbf{L}(\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{L}')$ degrees of freedom. It has been mentioned that the Wald test statistic does not produce accurate results when the number of data points are small. The F-test statistic will have a F-distribution with numerator degrees of freedom of $rank(\mathbf{L}(\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{L}')$ and denominator degrees of freedom estimated from the data (Khattree and Naik, 2003).

3.3.5.2 Tests for random components

When evaluating the significance of the random effects the variance components of these effects must be considered. If a variance component is not significantly greater than zero then the associated random effect will not be included in the model. The researcher will only prefer a mixed model approach to a fixed model if the random effects can significantly contributes to the model (if the variances are significantly greater than zero). The hypotheses that will be tested are:

$$\begin{aligned} H_o &: \sigma_b^2 = 0 \\ H_A &: \sigma_b^2 > 0. \end{aligned}$$

If the null hypothesis can be rejected, the researcher will include the associated variable as a random effect in the model. One way to test this hypothesis is by determining the approximate Wald Z statistic. The test statistic can be calculated by (Kleinbaum et al., 2008):

$$Z = \frac{\widehat{\sigma}_b^2}{S_{\widehat{\sigma}_b^2}}. \tag{3.44}$$

Since this test is only accurate for examples with a large set of groups it is advisable to include the likelihood ratio test to compare the results obtained using the Wald Z test. The likelihood ratio test compares two models (one with and one without the random components) using a Chi-square statistic with r degrees of freedom, where r is the number of variance and covariance parameters that will be estimated in $\mathbf{D}$.

3.3.6 Conjoint analysis industrial cleaner example

OLS estimation at the individual consumer's level accounts for arbitrary parameter heterogeneity between consumers. In conjoint analysis studies, these parameter estimates and the consumers' preference heterogeneity can be explored and possible market segmentations can be obtained. This has been proved to be a major advantage to companies that use conjoint analysis studies to gain these insights. Unfortunately the number of data points available at the individual consumers' level is generally very close to the number of unknown model parameters, due to the limited time and attention span of interviewed consumers. This results in statistically based model comparisons at an individual consumers' level which is nearly impossible (Frühwirth-Schatter and Otter, 1999). A possible statistical solution will be to use the mixed effect model approach.

When measuring the preferences of 86 individuals it is necessary to consider that not all of these individuals will share the same preferences and ideas. It is these differences in respondents' preferences that excite the idea of including random effects in the study. A researcher cannot explicitly model these differences but by adding a random component to the fixed effect model the researcher can evaluate the effect that the individual variation has on the model. In a conjoint setting the mixed effects model can be represented in a similar way as the models in the fixed effect environment, but now random effects are also included in the model.

• Let $i = 1, 2, ...m$ represents the number of respondents, $(m = 86)$, who took part in the study.

• Let $j = 1, 2, ...n$ represents the number of profiles that each individual respondent will evaluate, $(n = 18)$.

• Let p represent the number of fixed effects and q the number of random effects in the model. For this example, $p = q = 9$, including the intercept term since the model will include a random component for each of the fixed effects in the model.

Then the linear mixed model will be represented by:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \tag{3.45}$$

where $\mathbf{X}_i : n \times p$ design matrix set up using the specific attribute levels present in each profile, $\mathbf{Z}_i : n \times q$ is the design matrix for the random effects, $\boldsymbol{\beta} : p \times 1$ is the vector containing the unknown fixed effects (these effects stays fixed for all respondents), $\mathbf{b}_i : q \times 1$ is the vector of random effects (which differs for each respondent) and $\boldsymbol{\varepsilon}_i : n \times 1$ is the vector of error terms. It was also assumed that

$$\mathbf{b}_i \backsim \mathbf{N}(\mathbf{0}, \mathbf{D}_i) \text{ and } \boldsymbol{\varepsilon}_i \backsim \mathbf{N}(\mathbf{0}, \mathbf{R}_i).$$

For the industrial cleaner example each respondent will have the same attribute level design matrix, since all of them evaluated the same set of profiles. This means that $\mathbf{X}_i$ will be the same for all respondents. The matrix $\mathbf{X}_i$ was represented by Illustration 3.2.4. A special case of the linear mixed effect model is called the *Random Coefficient Model*. For this model, $\mathbf{Z}_i = \mathbf{X}_i$ since the model will investigate the inclusion of random effects for each of the attribute levels as well as the inclusion of a random effect for the intercept term which will allow the researcher to evaluate the effects at an individual consumer's level as well. Consider once again the 5 attributes' levels as presented in the design matrix $\mathbf{X}_i$, were the shorter notation was once again used. The mixed effect model will be:

$$\begin{aligned} y_{ij} &= \beta_I x_{j1} + \beta_{F1} x_{j2} + \beta_{F2} x_{j3} + \beta_{A1} x_{j4} + \beta_{A2} x_{j5} + \beta_{D1} x_{j6} + \beta_{B1} x_{j7} + \beta_{P1} x_{j8} + \\ &\quad \beta_{P2} x_{j9} + b_{Ii} z_{j1} + b_{F1i} z_{j2} + b_{F2i} z_{j3} + b_{A1i} z_{j4} + b_{A2i} z_{j5} + b_{D1i} z_{j6} + b_{B1i} z_{j7} + \\ &\quad b_{P1i} z_{j8} + b_{P2i} z_{j9} + \varepsilon_{ij}. \end{aligned}$$

Since $\mathbf{Z}_i = \mathbf{X}_i$ for this model, the fixed and random components can be grouped together:

$$\begin{aligned} y_{ij} &= (\beta_I + b_{Ii}) x_{j1} + (\beta_{F1} + b_{F1i}) x_{j2} + (\beta_{F2} + b_{F2i}) x_{j3} + (\beta_{A1} + b_{A1i}) x_{j4} + \\ &\quad (\beta_{A2} + b_{A2i}) x_{j5} + (\beta_{D1} + b_{D1i}) x_{j6} + (\beta_{B1} + b_{B1i}) x_{j7} + (\beta_{P1} + b_{P1i}) x_{j8} \\ &\quad + (\beta_{P2} + b_{P2i}) x_{j9} + \varepsilon_{ij} \end{aligned}$$

which can then be simplified to the model with form:

$$\begin{aligned} y_{ij} &= (\beta_{Ii}) x_{j1} + (\beta_{F1i}) x_{j2} + (\beta_{F2i}) x_{j3} + (\beta_{A1i}) x_{j4} + (\beta_{A2i}) x_{j5} + (\beta_{D1i}) x_{j6} \\ &\quad + (\beta_{B1i}) x_{j7} + (\beta_{P1i}) x_{j8} + (\beta_{P2i}) x_{j9} + \varepsilon_{ij} \end{aligned}$$

or

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$$

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{b}_i$ and

$$\boldsymbol{\beta}_i \backsim \mathbf{N}(\boldsymbol{\beta}, \mathbf{D}_i) \text{ and } \boldsymbol{\varepsilon}_i \backsim \mathbf{N}(\mathbf{0}, \mathbf{R}_i).$$

For this model y_{ij} will be the utility measurement for respondent i and profile j. $\boldsymbol{\beta}_I$ is considered the overall population grand mean, and d_{Ii} is the overall effect due to respondent i. This model allows the researcher to asses the effects on a respondent level as well as the effects measured on a profile level. For this example assume that each

individual will have the same variance-covariance structures, $\mathbf{D}_i = \mathbf{D}$, and $\mathbf{R}_i = \sigma^2\mathbf{I}_n$.

The inclusion of the random effect in the model allows the researcher to model the distribution of the responses for all the particular attribute levels. By calculating the variance components of each of these attribute levels the researcher can see how wide the spread, over all the individuals' responses are for each attribute level. By evaluating the variance component of the intercept term, the researcher can investigate how the variation among the respondents preferences influences the model. When the spread is wide (large variance) the preferences will be more difficult to determine as there was not an overall group agreement between the different respondents about this particular attribute. When the variance is small for a given attribute level it means that the overall population of respondents did agree on whether or not a particular attribute level will contribute to the overall appeal of a product.

The study investigated three different variance-covariance structures for this mixed effect model and evaluated the significance of each of the estimated parameters at a 5% level of significance. Although some of the parameters will not significantly contribute to the model, they will not be removed from the conjoint analysis study. This insignificance simply illustrate that those specific part-worth utilities will not influence the respondents overall preference towards the profiles. In *SAS, PROC MIXED* was used to estimate the unknown parameters using a ML approach (SAS Program 4).

3.3.6.1 Mixed effects model 1

The first structure considered is the *compound symmetry structure* which has the following form:

$$D = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_1^2 & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_1^2 \end{pmatrix}.$$

This structure allows only one variance component for all of the random effects. This means the model will have less parameters that will need to be estimated, than the other models that will be investigated, but this method does not allow the researcher to investigate different variances for the different attribute levels. Applying this model to the industrial cleaner example the estimated values together with the significance of the parameters were indicated in Table 3.13 and Table 3.14.

Table 3.13 Estimated fixed effects (Model 1)

Fixed Effect		Estimated Value	p-value	Significant
Intercept	$\widehat{\beta}_I$	3.7398	< 0.0001	$\surd$
F1	$\widehat{\beta}_{F1}$	-0.2171	0.0008	$\surd$
F2	$\widehat{\beta}_{F2}$	0.1667	0.0094	$\surd$
A1	$\widehat{\beta}_{A1}$	-0.3450	< 0.0001	$\surd$
A2	$\widehat{\beta}_{A2}$	0.0233	0.7118	$\times$
D1	$\widehat{\beta}_{D1}$	0.5102	< 0.0001	$\surd$
B1	$\widehat{\beta}_{B1}$	-0.1541	0.0084	$\surd$
P1	$\widehat{\beta}_{P1}$	1.1318	< 0.0001	$\surd$
P2	$\widehat{\beta}_{P2}$	0.0814	0.1980	$\times$

Table 3.14 Estimated variance components (Model 1)

Effect		Estimated Variances	p-value	Significant
Residual	$\widehat{\sigma}^2$	1.2000	< 0.0001	$\surd$
Intercept	$\widehat{\sigma}^2_I$	0.2052	< 0.0001	$\surd$
F1	$\widehat{\sigma}^2_{F1}$	0.2052	< 0.0001	$\surd$
F2	$\widehat{\sigma}^2_{F2}$	0.2052	< 0.0001	$\surd$
A1	$\widehat{\sigma}^2_{A1}$	0.2052	< 0.0001	$\surd$
A2	$\widehat{\sigma}^2_{A2}$	0.2052	< 0.0001	$\surd$
D1	$\widehat{\sigma}^2_{D1}$	0.2052	< 0.0001	$\surd$
B1	$\widehat{\sigma}^2_{B1}$	0.2052	< 0.0001	$\surd$
P1	$\widehat{\sigma}^2_{P1}$	0.2052	< 0.0001	$\surd$
P2	$\widehat{\sigma}^2_{P2}$	0.2052	< 0.0001	$\surd$

3.3.6.2 Mixed effects model 2

The second structure that was considered is the *full variance components structure* which has the following form:

$$D = \begin{pmatrix} \sigma_I^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_{F1}^2 & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_{P2}^2 \end{pmatrix}.$$

This structure ensures that each variable will have their own assigned variance components and implies that the random effects are uncorrelated, as assumed in the original mixed model assumptions. Together with the estimated values, the significance of the parameters were also indicated

Table 3.15 Estimated fixed effects (Model 2)

Fixed Effect		**Estimated Value**	**p-value**	**Significant**
Intercept	$\widehat{\beta}_I$	3.7398	< 0.0001	$\surd$
F1	$\widehat{\beta}_{F1}$	-0.2171	< 0.0001	$\surd$
F2	$\widehat{\beta}_{F2}$	0.1667	< 0.0001	$\surd$
A1	$\widehat{\beta}_{A1}$	-0.3450	< 0.0001	$\surd$
A2	$\widehat{\beta}_{A2}$	0.0233	0.6024	$\times$
D1	$\widehat{\beta}_{D1}$	0.5102	< 0.0001	$\surd$
B1	$\widehat{\beta}_{B1}$	-0.1541	0.0009	$\surd$
P1	$\widehat{\beta}_{P1}$	1.1318	< 0.0001	$\surd$
P2	$\widehat{\beta}_{P2}$	0.0814	0.1168	$\times$

Table 3.16 Estimated variance components (Model 2)

Effect		Estimated Variances	p-value	Significant
Residual	$\widehat{\sigma}^2$	1.2346	< 0.0001	$\surd$
Intercept	$\widehat{\sigma}^2_I$	0.5202	< 0.0001	$\surd$
F1	$\widehat{\sigma}^2_{F1}$	0.0661	0.0059	$\surd$
F2	$\widehat{\sigma}^2_{F2}$	0	.	$\times$
A1	$\widehat{\sigma}^2_{A1}$	0.1471	< 0.0001	$\surd$
A2	$\widehat{\sigma}^2_{A2}$	0.0330	0.0086	$\times$
D1	$\widehat{\sigma}^2_{D1}$	0.2978	< 0.0001	$\surd$
B1	$\widehat{\sigma}^2_{B1}$	0.0940	0.0002	$\surd$
P1	$\widehat{\sigma}^2_{P1}$	0.4394	< 0.0001	$\surd$
P2	$\widehat{\sigma}^2_{P2}$	0.0898	0.0042	$\surd$

3.3.6.3 Mixed effects model 3

The third model considered an *unstructured* variance-covariance matrix. This matrix drops the assumption of uncorrelated random effects and takes on the following form:

$$D = \begin{pmatrix} \sigma^2_I & \sigma_{IF1} & \cdots & \cdots & \sigma_{IP2} \\ \sigma_{IF1} & \sigma^2_{F1} & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \\ \sigma_{IP2} & \cdots & \cdots & \sigma_{P1P2} & \sigma^2_{P2} \end{pmatrix}$$

This results in more than just the 9 variances of the random effects, including the intercept, that needs estimation, now the related covariance structures also need to be estimated. The fixed effects and the variances of the variance-covariance matrix that was obtained, with the evaluation of whether they are significant or not, are represented in Table 3.17 and Table 3.18:

Table 3.17 Estimated fixed effects (Model 3)

Fixed Effect		Estimated Value	p-value	Significant
Intercept	$\widehat{\beta}_I$	3.7398	< 0.0001	$\surd$
F1	$\widehat{\beta}_{F1}$	-0.2171	< 0.0001	$\surd$
F2	$\widehat{\beta}_{F2}$	0.1667	0.0001	$\surd$
A1	$\widehat{\beta}_{A1}$	-0.3450	< 0.0001	$\surd$
A2	$\widehat{\beta}_{A2}$	0.0233	0.6277	$\times$
D1	$\widehat{\beta}_{D1}$	0.5102	< 0.0001	$\surd$
B1	$\widehat{\beta}_{B1}$	-0.1541	0.0009	$\surd$
P1	$\widehat{\beta}_{P1}$	1.1318	< 0.0001	$\surd$
P2	$\widehat{\beta}_{P2}$	0.0814	0.1337	$\times$

Table 3.18 Estimated variance components (Model 3)

Effect		Estimated Variances	p-value	Significant
Residual	$\widehat{\sigma}^2$	1	< 0.0001	$\surd$
Intercept	$\widehat{\sigma}^2_I$	0.5352	< 0.0001	$\surd$
F1	$\widehat{\sigma}^2_{F1}$	0.0958	0.0034	$\surd$
F2	$\widehat{\sigma}^2_{F2}$	0.0163	0.2458	$\times$
A1	$\widehat{\sigma}^2_{A1}$	0.1862	< 0.0001	$\surd$
A2	$\widehat{\sigma}^2_{A2}$	0.0584	0.0253	$\surd$
D1	$\widehat{\sigma}^2_{D1}$	0.3055	< 0.0001	$\surd$
B1	$\widehat{\sigma}^2_{B1}$	0.0975	0.0001	$\surd$
P1	$\widehat{\sigma}^2_{P1}$	0.4918	< 0.0001	$\surd$
P2	$\widehat{\sigma}^2_{P2}$	0.1162	0.0093	$\surd$

When comparing these three models, it can be seen that the estimated values for the fixed parameters are the same for all three models. Only two of these estimated parameters are not significantly different from zero, which is an indication of the indifference of the respondents towards these attribute levels. Since positive part-worth utilities, indicate that the respondents prefer a specific attribute level, while negative part-worth utilities is an indication that the respondents do not prefer a certain attribute level, a part-worth utility close to zero show the indifference of respondents toward that attribute level. As the variance-covariance structures changes, so do the estimated variances. For the first model, all the variance components of the random effects are the same and do differ significantly from zero. When the model allows for different variances between the different attribute levels, as in model 2, the variances across the different attribute

levels clearly differ. When the model allows for correlation between the different attribute levels, the variance covariance structure of model 3 is used. Using a model that allows for correlation is more realistic since it can be assumed that there will be correlations in the way in which different respondents' choices are made. A typical example is when a respondent who usually chooses the lowest possible price will always pick the lowest possible price, not really ever allowing to be persuaded to buy a more expensive product that has a higher quality. The estimated variances obtained in model 2 and 3 are very similar. The only variance component that is not significantly different from zero, in model 3, is the variance associated with the concentrated liquid product form. When the variance component of a specific attribute level is close to zero it indicates that the preferences, over all the respondents, for that attribute level was very thinly spread. All respondents had a very similar preference towards that specific attribute. The next unit will compare the models that has been investigated and the results of the estimated parameters will be discussed.

3.4 Model comparison

Researchers are always faced with the question of what model has the best fit. Choosing the best model for the data is not always a straightforward or easy task. One way that this can be accomplished is by using the Akaike Information Criteria (AIC) to choose the best possible model from the different models being considered. The AIC method uses the definition of the Kullback-Leibler information which is a distance measure between two density functions (Ngo and Brand, n.d.). When comparing the densities of the true model and the selected model, one would prefer the distance to be as small as possible as this will indicate that the model is a good fit for the true data. The AIC measure can be represented by

$$AIC = -2\text{Loglikelihood}+2\text{d} \tag{3.46}$$

where d represents the number of parameters that were estimated (Ngo and Brand, n.d.). The fact that this method imposes a penalty for the number of parameters that was estimated is why it is such a good measure to be used to compare models. When using this method to compare the accuracy of the models applied to the industrial cleaner example the model that came out at the top (the model with the smallest AIC value) was the mixed effects model 3 with the unstructured variance components. The traditional approach was also compared to the more comprehensive models, eventhough is has a much simpler covariance structure. Table 3.19 presents the findings:

Table 3.19 Model comparisons

Model	**AIC**
Traditional conjoint model - fixed effect model	5897.4
Mixed effects model 1	5590.4
Mixed effects model 2	5473.2
Mixed effects model 3	5460.9

The addition of the random components to the usually fixed effect model, can effectively account for the correlation between respondents as well as the variation within the respondents. Mixed effects model 1 illustrated that by adding the random effect to the model, the overall model fit did indeed improve. This motivates the initial thought that some of the variation of the model can be explained by the heterogeneity among the respondents. By restricting the variance-components to all be the same, the researcher can't effectively investigate the variance-covariance structure of each attribute level. Thus by modelling the variance-covariance structure as was done in Mixed effects model 2, the researcher can model a different variance component for each attribute level. These variances gives

the researcher a good indication about the variation of the sample of respondents' preferences towards each individual attribute level. The third model, Mixed effects model 3, ignored the assumption of the uncorrelated random variables and allowed a correlation between variables. This resulted in the model with the best fit. Consider the following estimates of all of the fixed effects parameters as well as all the random effects' variances. The values of the parameters that was not present in the model was found by applying a simple reparameterization to the dummy variables and estimating the parameters again.

Table 3.20 Full set of parameters (Model 3)

Effect		**Estimated Value**		**Estimated Variances**
Residual			$\widehat{\sigma}^2$	1
Intercept	$\widehat{\beta}_I$	3.7398	$\widehat{\sigma}^2_{d_I}$	0.5352
F1	$\widehat{\beta}_{F1}$	−0.2171	$\widehat{\sigma}^2_{F1}$	0.0958
F2	$\widehat{\beta}_{F2}$	0.1667	$\widehat{\sigma}^2_{F2}$	0.0163
F3	$\widehat{\beta}_{F3}$	0.0504	$\widehat{\sigma}^2_{F3}$	0.05183
A1	$\widehat{\beta}_{A1}$	−0.3450	$\widehat{\sigma}^2_{A1}$	0.1862
A2	$\widehat{\beta}_{A2}$	0.02326	$\widehat{\sigma}^2_{A2}$	0.0584
A3	$\widehat{\beta}_{A3}$	0.32174	$\widehat{\sigma}^2_{A3}$	0.1464
D1	$\widehat{\beta}_{D1}$	0.5102	$\widehat{\sigma}^2_{D1}$	0.3055
D2	$\widehat{\beta}_{D2}$	−0.5102	$\widehat{\sigma}^2_{D2}$	0.3055
B1	$\widehat{\beta}_{B1}$	−0.1541	$\widehat{\sigma}^2_{B1}$	0.0975
B2	$\widehat{\beta}_{B2}$	0.1541	$\widehat{\sigma}^2_{B2}$	0.0975
P1	$\widehat{\beta}_{P1}$	1.1318	$\widehat{\sigma}^2_{P1}$	0.4918
P2	$\widehat{\beta}_{P2}$	0.0814	$\widehat{\sigma}^2_{P2}$	0.1162
P3	$\widehat{\beta}_{P3}$	−1.2132	$\widehat{\sigma}^2_{P3}$	0.4172

By visually representing these attribute levels and their estimated mean part-worth values and variances, the researcher can obtain important insights into the way in which the overall sample of respondents' preferences where distributed. The estimated fixed effects are similar to those obtained in the fixed effects model and will yield the exact same visual representations as Figure 3.2.1. The researcher needs to consider that this sample of respondents reflects what will hopefully be the preference structure in the market place. When the data of this industrial cleaner study was initially collected, the team of researchers did choose a representative industrial sample of respondents, to ensure that the results obtained from these 86 individuals will be a representation that will closely resemble the preferences in the market place. Although many companies are simply interested in the product that performs the best, it can be seen that there is valuable

information in investigating the difference between these different attribute levels.

Although the level with the highest part-worth is classified as the "best" option, this will not always be the case in all markets. Other possible combinations of attribute levels, that did not necessarily perform the "best" can cater for sectors in the market that would not have chosen the "best" product. If these alternative market segments are large enough, a company can possibly profit from producing more than one product. The following visual representations will represents these estimates obtained for each attribute level. The preference level is an indication of how much, or how little, the respondents liked the selected attribute levels. The higher the preference level, the greater the overall preference. The first graphical representation, Figure 3.4.1, illustrates the overall likelihood of purchase of the respondents.

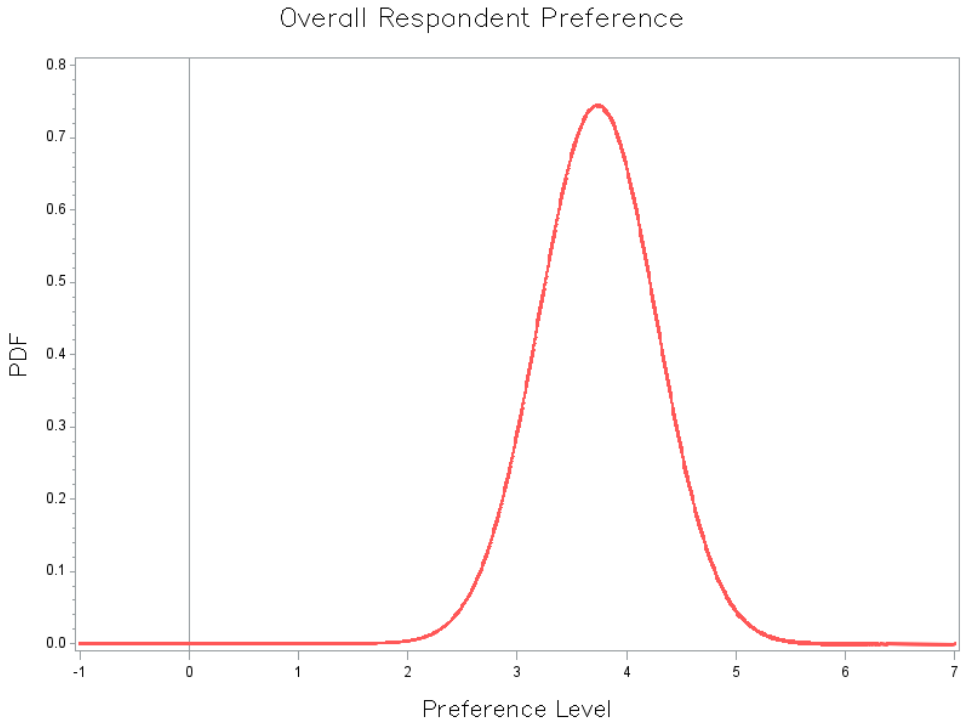


Figure 3.4.1 Overall likelihood of purchase

Overall, the consumers indicated that they did like the 18 profiles that were presented to them, as the distribution is situated to the right of 0. What does greatly influence the variation in the linear mixed model, is the big variance associated with the overall likelihood of purchase. The big spread around the mean indicates that the group of respondents' preferences did vary a great deal when they had to choose between the possible profiles. This variation can encourage a researcher to investigate the possibility of grouping the respondents into specific market segments. If the researcher had additional information about all of these respondents, he/she could use that information to classify

these respondents and model them according to these segments. ACA (Adaptive Conjoint Analysis) originated to solve this problem. If, however the researcher did not have any additional information about these respondents, as was the case with this study, clustering techniques can be applied to each respondents part-worth values ($\beta's$) to group individual customers who have similar preference structures. These segments may then be profiled and assessed for the unique preference structures and market potential (Hair et al., 2006). The first attribute that was considered was the product from.

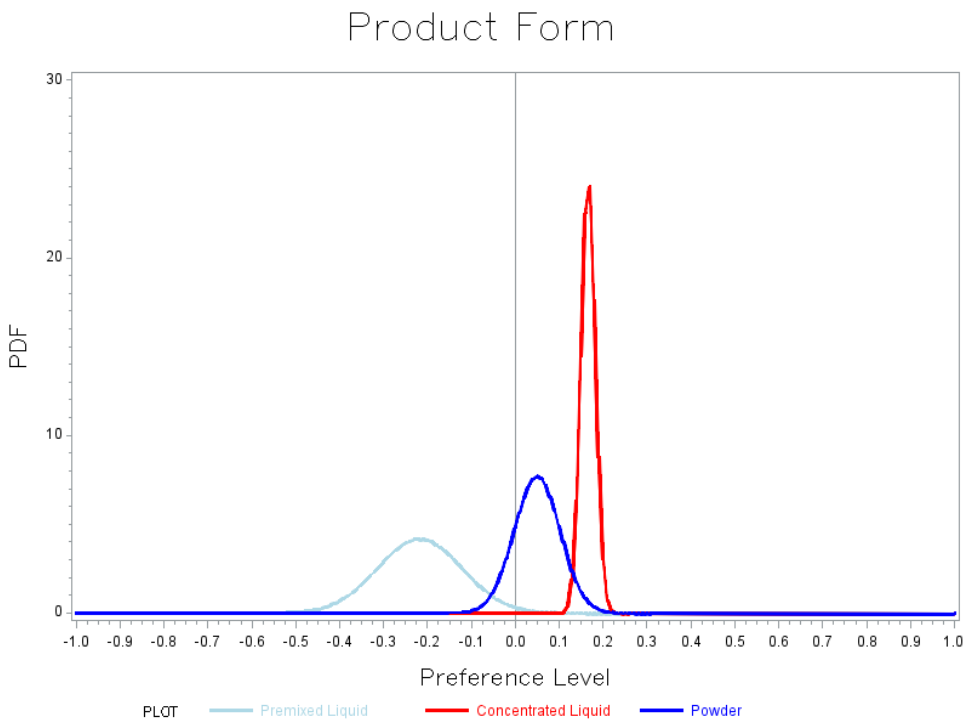


Figure 3.4.2 Product form

When the researcher only consider the part-worth values obtained from the traditional conjoint analysis method then the concentrated liquid product form would have been the preferred choice. The red distribution illustrates the preferences towards this product form. Clearly the concentrated liquid form was preferred by the respondents. Since the variance of this attribute level is very small, it illustrates that the population of respondents was in a general agreement about how much they prefer this attribute level. By further investigating the powder product form, represented by the dark blue distribution, it can be seen that a great portion of the respondents indicated that they do prefer this product form, although not as much as the concentrated liquid form. Although their preference toward the powder product form is not the highest, it is important to realize that they did not dislike this product form. The company can consider to produce

products with these two different forms.

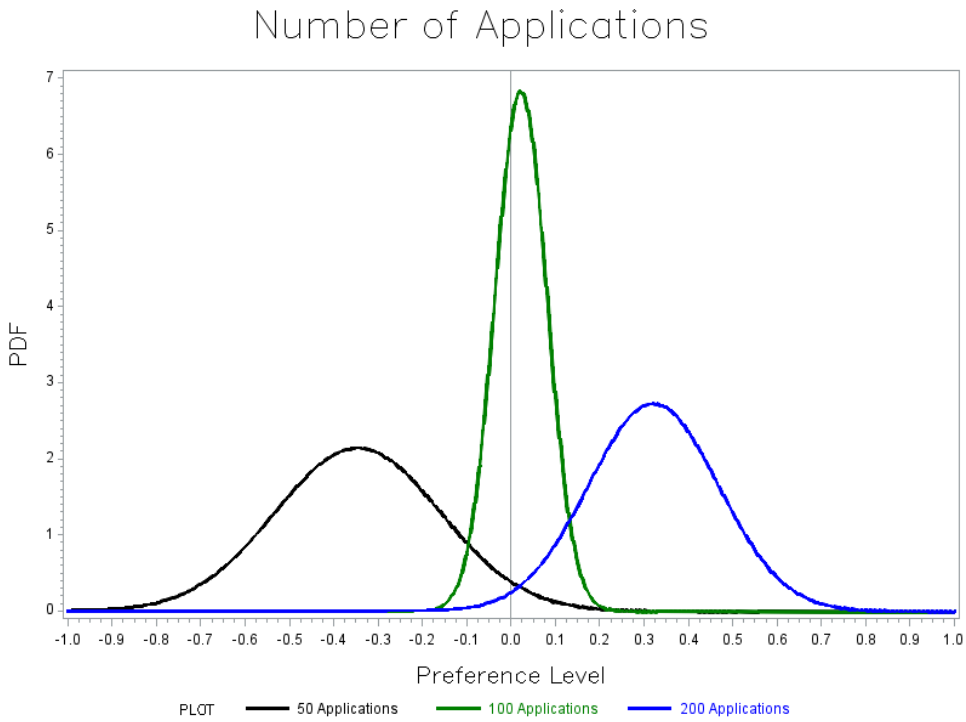


Figure 3.4.3 Number of applications

The second attribute, the number of applications per container, also had three different levels. It was expected that the smallest number of applications would be disliked by almost all of the respondents. When a customer considers a container that has a certain price, it is natural to choose a container that will ensure many applications. The small proportion of the population that preferred the smaller number of applications per container may be respondents who would prefer smaller containers as they are easier to transport. The largest number of applications per container was preferred by almost the whole sample of respondents. The researcher should consider who the users are of this industrial cleaner. Many of the customers may be business-owners themselves, that have to consider their own expenses and the influence it will have on their profit. The greater number of applications they can get for the smallest amount of money, the larger their own profit.

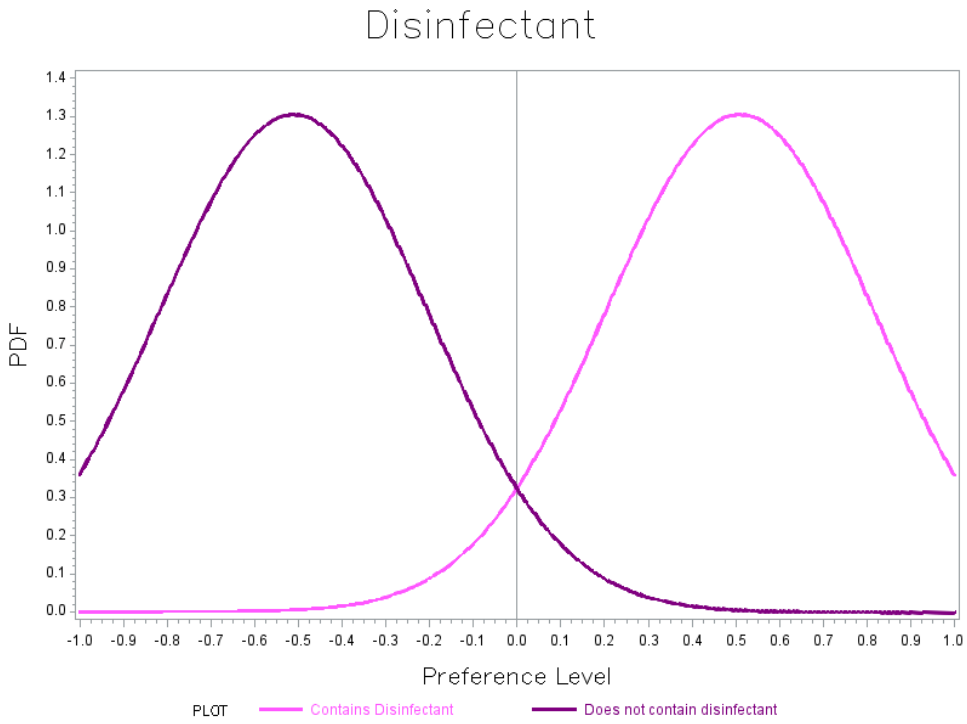


Figure 3.4.4 Disinfectant

When investigating the inclusion of disinfectant in the industrial cleaner, it is clear that this attribute had a big influence on the variation in the linear mixed model. Although a large proportion of the respondents indicated that the inclusion of a disinfectant in the industrial cleaner is important, they were not in agreement about *how* important it actually is. It is because of this uncertainty that this attribute had a large importance value in the traditional conjoint analysis model. The uncertainty surrounding the respondents preference, about this particular attribute, may have a big influence on the prediction of the actual preferred profiles. The company should investigate possible ways to decrease this uncertainty through different possible marketing campaigns or advertising schemes.

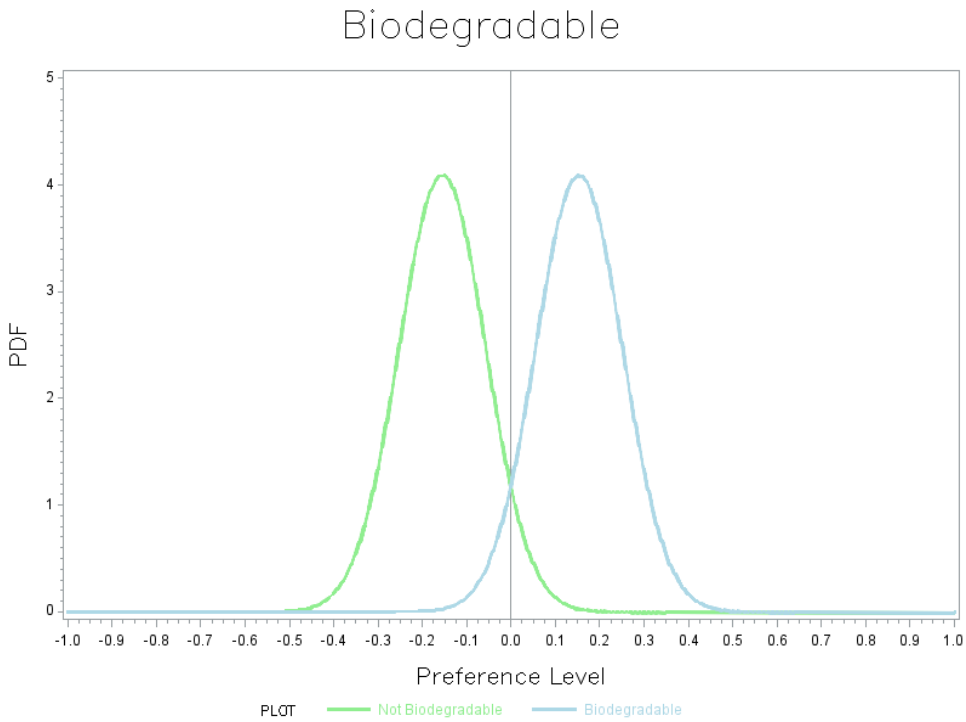


Figure 3.4.5 Biodegradable

Eventhough the part-worth value for the biodegradability attribute is not so high, when compared to other attribute levels, this visual representation indicates that the respondents in general preferred the biodegradability quality of the industrial cleaner. The variation of this attribute level is a lot less than the variation of the disinfectant attribute. This means that the sample of respondents are in general agreement about how important this attribute is. Although the attribute is not the first property that many customer will consider when purchasing industrial cleaners, it will be advised to include this property, as it will most probability be a deciding factor when the customers compare different products. Then of course, there will always be other advantages of producing biodegradable products through governmental grants etc. which promotes the environmental safety factors.

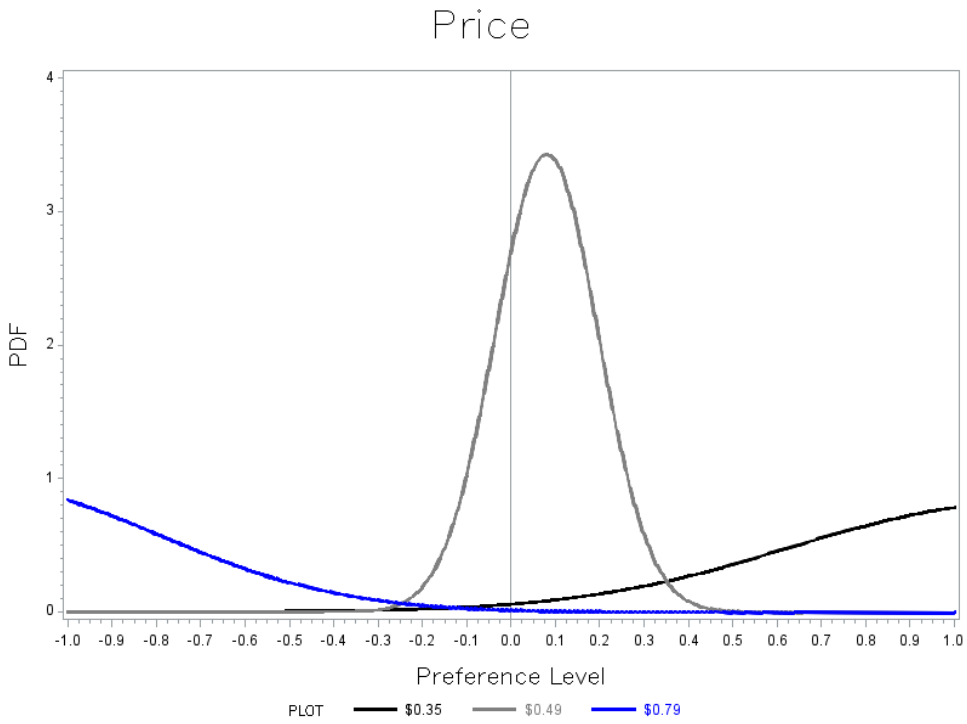


Figure 3.4.6 Price

Finally, price. This attribute is the most influential attribute in this conjoint analysis study. A great proportion of the variation in the dependent variable (responses) can be assigned to this attribute. In the initial stages of this conjoint analysis study it was found that price represented one of the primary determinants of value in the industrial cleaner market, and hence was included in the study. From the visual representation it is quite clear that the whole population of respondents preferred the smallest price per application, while they all had a great dislike towards the largest price value per application. This can be expected in a business environment. A customer will almost always choose the best possible product for the lowest possible price. Defining "the best product" and "the lowest price" will differ for different respondents. The price level in the middle had a much smaller variation and the largest part of the respondents found that price level was reasonable.

Another measure which can be used to determine if a linear mixed model approach is useful is the Intraclass Correlation Coefficient (ICC). The ICC can be calculated by dividing the single attribute levels' variance components by the total variance present in the model. This measure indicates the proportion of the total variance in the response, $\mathbf{y}$ (the dependent variable), that is explained by each attribute level. This can help determine to what extent the preference (which varies among respondents, but not within

a respondent) effects the overall response. This is also a very useful measure as it can give a good indication of whether a linear mixed model is necessary or if a simpler model should be considered. It is also very meaningful to investigate how the ICC changes as variables are added, or removed, from the model. In the industrial cleaner example the following ICC measures where obtained by dividing each variance component in Table 3.20 by the total variation in the model.

Table 3.21 Intraclass correlation coefficients (ICC) for each level

Effect	**ICC**
Residual	0.2550
Intercept	0.1365
F1	0.0244
F2	0.0041
F3	0.0132
A1	0.0475
A2	0.0149
A3	0.0373
D1	0.0780
D2	0.0780
B1	0.0249
B2	0.0249
P1	0.1254
P2	0.0296
P3	0.1064

The largest ICC, except for the residual term, is the intercept term. This term indicates the variation present in the model due to the different preference structures among the different respondents. This ICC clearly indicates that the heterogeneity among the respondents influences the variation within the linear mixed model. Two of the price levels also resulted in relatively large ICC's Usually this variation would be grouped with the residual variance but the fact that such a large portion of the variance can be explained by these differences among the individuals, clearly indicates why the linear mixed model does indeed provide a model that can give the researcher more insight, than the traditional conjoint analysis model. By including additional covariates in the model, which can be used to measure these differences among respondents, the researcher will be able to gain more insights about the respondents preferences than in the traditional model. It is advised that the researchers who employ the conjoint analysis study also

obtain some useful information about each individual that partook in the study. These additional segment of information can be used to classify the different respondents into groups with similar preference structure, and obtain more accurate predictions of the true preferences in the different market segments. Table 3.22 represents the ICC values at an attribute level. These values can typically be compare with the overall importance values of the traditional conjoint analysis approach in Table 3.8. These ICC values indicated how much each attribute contributed to the variation in the model. Another advantage of the mixed effects model approach to conjoint analysis is that this model can determine importance values for each level as well as each attribute. Traditional conjoint methods can just determine the importance values for the attributes.

Table 3.22 Intraclass correlation coefficients (ICC) for each attribute

Effect	**ICC**
Residual	0.2550
Intercept	0.1365
Product form (F)	0.0418
Number of applications (A)	0.0997
Disinfectant (D)	0.1558
Biodegradability (B)	0.0497
Price per application (P)	0.2615

Chapter 4

Hierarchical Bayesian Methods

"*The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise. The canonical example is to imagine that a precocious newborn observes his first sunset, and wonders whether the sun will rise again or not. He assigns equal prior probabilities to both possible outcomes, and represents this by placing one white and one black marble into a bag. The following day, when the sun rises, the child places another white marble in the bag. The probability that a marble plucked randomly from the bag will be white (i.e., the child's degree of belief in future sunrises) has thus gone from a half to two-thirds. After sun rise the next day, the child adds another white marble, and the probability (and thus the degree of belief) goes from two-thirds to three-quarters. And so on. Gradually the initial belief that the sun is just as likely as not to rise each morning is modified to become near-certainty that the sun will always rise.*"

-*The Economist* (2000)

4.1 Introduction

The Bayesian philosophy was introduced in the 18th century by Reverend Thomas Bayes, a Presbyterian minister and mathematician. Although the 19^{th} century was filled with disputes over the validity of the use of Bayesian inference due to the uncertainties around handling the prior distribution, the late 20^{th} century brought retribution. With the evolution of computer capabilities and the development of computational methods like Markov Chain Monte Carlo (MCMC), which made tedious methods feasible, the modern day Bayesian movement began. Thanks to brave thinkers like James O. Berger, Bruno de Finetti, Harold Jeffreys, Jimmy Savagee and Dennis Lindley, the interest in Bayesian statistics grew (Cowels et al., n.d.). This lead to extensive research not only in Bayesian methodology but also in the use of Bayesian methods to address pressing questions in diverse application areas such as astrophysics, weather forecasting, health care policies and criminal justice (Cowels et al., n.d.). This method allows the parameters to be fixed as well as random since all uncertainties are probabilistically modelled (Yan and Su, 2009).

A major advantage of this method is the incorporation of prior knowledge and beliefs from field experts and information into the inference and analysis (Yan and Su, 2009). Other advantages include (Seefeld, 2007):

- The flexibility in how to model distributions of data, and apply it to many statistical situations.
- The method deals well with missing data.
- It effectively models high-dimentional data.
- This method does not require large samples, which is a major advantage for conjoint analysis, since the field constantly deals with small sample sizes.

4.2 Hierarchical Bayes in conjoint analysis

Although Bayesian methods have been introduced in conjoint studies, very few researchers have fully incorporated these methods. Until recently, the individuals advocating hierarchical Bayes (HB) were academics and a few practitioner experts in statistics. HB is demanding both in terms of computational time and complexity (Orme, 2000). Even with these concerns, applications of HB have become more and more apparent. In conjoint analysis studies, HB can be used to estimate the individual part-worth utility levels of each respondent as well as the aggregate part-worth utilities over all the respondents. Although this method can typically be linked to Adaptive Conjoint Analysis (ACA) studies, it has been proved to significantly improve the results of traditional conjoint methods

and choice based conjoint studies. This method increases the quality of individual level utilities by borrowing information from the other individuals which in turn leads to more accurate predictions at an individual and aggregate level (Orme, 2000). Although HB is both computationally demanding and time consuming, this methods can significantly improve on traditional methods like OLS and logit, especially if there is heterogeneity between individuals, as in the case of the mixed model.

The term "hierarchical" refers to the different levels incorporated in the method.

Level 1:
Assumes that the individuals' parameters (part-worth values) are all described by multivariate normal distributions, with a vector of means and a matrix of covariances.
Level 2:
Assumes that, given the individuals' part-worth values, his/her probability of achieving some outcome (like choosing some product) is governed by a particular model, like multinomial logit or linear regression.

The process starts off by choosing initial estimated values for the $\beta's$ (part-worths). These values are then updated using Markov Chain Monte Carlo methods such as the Gibbs Sampler approach and the Metropolis-Hastings algorithm, which should result, after many iterations, in convergence to accurate estimates. The model estimates values for the individual respondents as well as the overall mean and covariance of the estimated $\beta's$. The literature also refers to a term called the "burn in period". This is the set of iterations after which convergence is assumed and can be as much as 10 000 iterations or even more. After the burn in period, a set of the respondents $\beta's$ are saved after every n^{th} iteration. These sets that are saved are called the "draws" of the respondent. The point estimates for each individual can then be calculated by computing the average of each respondents' draws.

Some of the advantages of this approach includes (Orme, 2000):

- HB seems to improve, or at least match, the predictive accuracy of the traditional models.
- HB is known to be a very robust method of estimation.
- This method permits individual level modeling and can thus target specific individuals.
- The aggregate estimation method used in traditional conjoint methods often confuses heterogeneity and noise. By modelling individual level measures rather than an aggregate level, the HB method can separate heterogeneity and noise which results in a more stable and accurate estimation at both individual and aggregate levels.

• The draws of each respondent have also been classified as a rich source of statistical information.

4.3 The Bayesian approach

The basic idea behind the Bayesian method is to assume some prior distribution for the unknown parameters in the model. By combining this distribution with the distribution of the given data, called the likelihood distribution, a posterior distribution can be obtained after which inferences can be made, based on the posterior distribution of the parameters, given the data.

Let $\mathbf{y}$ denote the observed data and let $\boldsymbol{\theta}$ be the vector containing the unknown parameters of interest. Hypotheses are typically expressed through probability distributions of observed data. These distributions depend on unknown parameters. The Bayesian method assumes that the unknown parameter, $\boldsymbol{\theta}$, is a random variable following a distribution with a probability density function of $p(\boldsymbol{\theta})$, also referred to as the prior distribution. By placing a probability distribution on these unknown parameters, the current knowledge about the parameters can be expressed. The observed data, in turn, contains information regarding the model parameters which can be expressed using the likelihood distribution, $p(\mathbf{y}|\boldsymbol{\theta})$, which is the distribution of the observed data given the model parameters. By combining the prior and likelihood distributions, the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$ can be obtained. It is this distribution which is used to perform the Bayesian inference and can be obtained using the Bayesian theorem.

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} \tag{4.1}$$

where $p(\boldsymbol{\theta}, \mathbf{y})$ is the joint density of $\boldsymbol{\theta}$ and $\mathbf{y}$, and $p(\mathbf{y})$ is the marginal density of $\mathbf{y}$. The calculation of the posterior distribution can be illustrated as follows:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) &= \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} \\ &= \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \\ &= \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= c.p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \end{aligned}$$

where c is called the normalizing constant which insures that $p(\boldsymbol{\theta}|\mathbf{y})$ is a valid density function satisfying the property of $\int p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = 1$. Hence in summary the Bayesian idea can be represented as:

$$\text{Posterior distribution} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

which simplifies to

$$\text{Posterior distribution} \propto \text{Likelihood} \times \text{Prior}$$

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

In essence, the prior updates the data likelihood distribution to produce a posterior distribution that combines Bayesian subjective knowledge, with objective observed data. The point estimates and confidence intervals can then be obtained from the joint posterior density, $p(\boldsymbol{\theta}|\mathbf{y})$, and the associated marginal posteriors, $p(\theta_j|\mathbf{y})$, for each θ_j. The marginal posteriors can be obtained by integrating other components of $\boldsymbol{\theta}$ out of the joint posterior. The difficulty level depends on the prior specification and shows why the prior will usually be chosen in such a way as to ensure mathematical convenience and easy solvability. If the integrals involved are more complicated they can be approximated by annalistic techniques or numerical integration (Hoff, 2009).

4.3.1 Choosing priors

The choice of the prior distributions plays a key role in the Bayesian approach. It introduces subjective information into the model. With no specific set of rules to follow when deciding on the "correct" prior distribution the method is often criticized as being too subjective and difficult to implement (Seefeld, 2007). It is however often the case that the prior is simple to choose since it follows a convenient model pattern. A researcher can distinguish between two types of prior specifications, nl. objective priors and subjective priors (Wakefield, 2013). For the first prior specification, objective priors, the researcher will choose a prior distribution that will have a "minimal impact" as to insure that the likelihood information dominates the posterior distribution. Other terms that can be associated with the first prior specifications includes: reference, noninformative and nonsubjective priors (Wakefield, 2013). The second prior specification incorporates subjective information by describing the informed opinion of the expert researcher.

4.4 Bayesian linear regression model

4.4.1 Introduction

As discussed in Chapter 3 the classical linear regression model is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.2}$$

where $\mathbf{y} = (y_1, y_2, .., y_n)'$ is the vector of n responses, $\mathbf{X}$ is an $n \times (p+1)$ design matrix, $\boldsymbol{\beta} = (\beta_0, \beta_2, ..., \beta_p)'$ is the vector of unknown regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ is the vector of errors. It follows that

$$\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2 \backsim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2) \tag{4.3}$$

To obtain the posterior distribution for the parameters the likelihood distribution needs to be combined with the selected prior distributions.

4.4.1.1 Posterior distribution of $\boldsymbol{\beta}'s$

By selecting a normal prior of $\boldsymbol{\beta} \backsim N(\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o)$ where $\boldsymbol{\Sigma}_o = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ the posterior distribution of the $\boldsymbol{\beta}'s$ is (Hoff, 2009):

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2) \propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta})$$

Since the likelihood function is given by:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}[\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}]\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}SSR(\boldsymbol{\beta})\right\} \end{aligned}$$

and the prior distribution will be:

$$\begin{aligned} p(\boldsymbol{\beta}) &= \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_o|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o + \boldsymbol{\beta}_o'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o\right)\right\}. \end{aligned}$$

The posterior distribution, of the $\boldsymbol{\beta}'s$, is given by:

$$\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) &\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)\times p(\boldsymbol{\beta}) \\
&= \exp\left\{-\frac{1}{2}\left[\frac{\mathbf{y}'\mathbf{y}}{\sigma^2}-\frac{\mathbf{2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}}{\sigma^2}+\frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\sigma^2}\right]-\frac{1}{2}\left[\boldsymbol{\beta}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o+\boldsymbol{\beta}_o'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-\frac{\mathbf{2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}}{\sigma^2}+\frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\sigma^2}]\right]-\frac{1}{2}\left[\boldsymbol{\beta}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[-2\boldsymbol{\beta}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o-\frac{\mathbf{2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}}{\sigma^2}\right]-\frac{1}{2}\left[\boldsymbol{\beta}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}+\frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\sigma^2}\right]\right\} \\
&= \exp\left\{\boldsymbol{\beta}'\left[\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o+\frac{\mathbf{X}'\mathbf{y}}{\sigma^2}\right]-\frac{1}{2}\boldsymbol{\beta}'\left[\boldsymbol{\Sigma}_o^{-1}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right]\boldsymbol{\beta}\right\}.
\end{aligned}$$

This is recognized as being proportional to a normal density with an expected value and variance of:

$$\begin{aligned}
E(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) &= \left(\Sigma_o^{-1}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)^{-1}\left(\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o+\frac{\mathbf{X}'\mathbf{y}}{\sigma^2}\right) \qquad (4.4)\\
Var(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) &= \left(\boldsymbol{\Sigma}_o^{-1}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)^{-1}
\end{aligned}$$

If the elements of the prior precision matrix $\boldsymbol{\Sigma}_o^{-1}$ are small in magnitude, then the conditional expectation $E(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2)$ will be approximately equal to the OLS estimate $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. If the measurement precision is very small (σ^2 very large) then the expectation is approximately $\boldsymbol{\beta}_o$ which is the prior expectation (Hoff, 2009).

4.4.1.2 Posterior distribution for σ^2

To obtain the prior distribution off σ^2, let $\gamma=\frac{1}{\sigma^2}$ be the measurement precision. Since the prior distribution chosen for σ^2 is an inverse-gamma distribution, γ will then be distributed as $\gamma \backsim G(\frac{\upsilon_o}{2},\frac{\upsilon_o\sigma^2}{2})$ where $\upsilon_o=n-p-1$. Using this prior and combining the distributions to the likelihood function the posterior distribution of σ^2 can be determined.

$$p(\gamma|\mathbf{y},\mathbf{X},\boldsymbol{\beta}) \propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\gamma)\times p(\gamma)$$

Since the likelihood distribution, in terms of γ, is given by

$$\begin{aligned}
p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\gamma) &= \frac{1}{(2\pi\frac{1}{\gamma})^{n/2}}\exp\left\{-\frac{\gamma}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\right\} \\
&\propto \gamma^{n/2}\exp\left\{-\frac{\gamma}{2}SSR(\boldsymbol{\beta})\right\}
\end{aligned}$$

and the prior distribution is

$$\begin{aligned} p(\boldsymbol{\gamma}) &= \frac{\left(\frac{\upsilon_o\sigma^2}{2}\right)^{\frac{\upsilon_o}{2}}}{\Gamma\left(\frac{\upsilon_o}{2}\right)}\gamma^{\frac{\upsilon_o}{2}-1}\exp\left\{-\gamma\times\frac{\upsilon_o\sigma^2}{2}\right\} \\ &\propto \gamma^{\frac{\upsilon_o}{2}-1}\exp\left\{-\gamma\times\frac{\upsilon_o\sigma^2}{2}\right\} \end{aligned}$$

the posterior distribution is

$$\begin{aligned} p(\gamma|\mathbf{y},\mathbf{X},\boldsymbol{\beta}) &\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\gamma)\times p(\gamma) \\ &\propto \left[\gamma^{\frac{n}{2}}\exp\left\{-\gamma\times\frac{SSR(\beta)}{2}\right\}\right]\times\left[\gamma^{\frac{\upsilon_o}{2}-1}\exp\left\{-\gamma\times\frac{\upsilon_o\sigma^2}{2}\right\}\right] \\ &= \gamma^{\frac{(\upsilon_o+n)}{2}-1}\exp\left\{-\gamma\left[\frac{\upsilon_o\sigma^2}{2}+\frac{SSR(\beta)}{2}\right]\right\}. \end{aligned}$$

This is recognized as the gamma density which leads to the inverse gamma distribution

$$\{\sigma^2|\mathbf{y},\mathbf{X},\boldsymbol{\beta}\}\backsim IG\left(\frac{\upsilon_o+n}{2},\left[\frac{\upsilon_o\sigma^2}{2}+\frac{SSR(\beta)}{2}\right]\right).$$

4.4.1.3 Sampling process

In order to obtain the joint posterior distribution, $p(\boldsymbol{\beta},\gamma|\mathbf{y})$, the Gibbs sampler can be used. Once the form of the hierarchical posterior distributions are obtained the sampling process can begin. The steps to construct the Gibbs sampler to approximate the joint posterior distribution can be represented as follows (Hoff, 2009):

For the first iteration a initial set of values are chosen for the hyper parameters. Given these values ,$\boldsymbol{\beta}^{(0)}$ and $\sigma^{2(0)}$, new values can be generated by using the following steps, where $k=0,1,2...$ represents the iterations.

Step 1:

a) Compute $\mathbf{m}=E(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2)$ and $\mathbf{V}=Var(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2)$ from (4.4)

$$\begin{aligned} \mathbf{m} &= E(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^{2(k)}) = \left(\Sigma_o^{-1(k)}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^{2(k)}}\right)^{-1}\left(\boldsymbol{\Sigma}_o^{-1(k)}\boldsymbol{\beta}_o^{(k)}+\frac{\mathbf{X}'\mathbf{y}}{\sigma^{2(k)}}\right) \\ \mathbf{V} &= Var(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^{2(k)}) = \left(\Sigma_o^{-1(k)}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^{2(k)}}\right)^{-1} \end{aligned}$$

b) Sample $\boldsymbol{\beta}^{(k+1)}\backsim N(\mathbf{m},\mathbf{V})$

Step 2:

a) Compute $SSR(\boldsymbol{\beta}^{(k+1)})$

b) Sample $\sigma^{2(k+1)}\backsim IG\left(\frac{\upsilon_o\sigma^2}{2},\left[\frac{\upsilon_o\sigma^2}{2}+\frac{SSR(\boldsymbol{\beta}^{(k+1)})}{2}\right]\right)$

Step 3: Repeat step 1 and 2, h times to generate the sample $\{(\boldsymbol{\beta}^k, \gamma^k), k = 1, 2, ...h\}$ from $p(\boldsymbol{\beta},\gamma|\mathbf{y})$. This distribution is then used to determine the point estimates and confidence intervals for the estimated parameters. It is good practice to allow for a burn in period, after which convergence is assumed. The researcher can also choose to store only each n^{th} draw, or keep all the information.

4.4.2 Industrial cleaner conjoint analysis example

Using a Bayesian approach to determine the part-worth values for the traditional conjoint analysis model, yielded similar results as those obtained when using an OLS method. For both approaches the parameters could be estimated at a individual level as well as at an aggregate level. Since the generation process stabilized relatively quickly only 10000 iterations was used, with the first 1000 considered as the burn in period. Thereafter, all the parameters were stored and collected in order to obtain the joint posterior distribution. The following tables and figures represents the parameter estimates as well as the posterior distributions obtained in the HB approach (SAS Program 5).

4.4.2.1 Results obtained for respondent 1

Table 4.1 Estimated parameters (Respondent 1)

Level	**Parameters**	**OLS estimation**	**HB estimation**
Variance component	σ^2	0.6237	0.7882
Intercept	β_I	4.1111	4.0992
Premixed Liquid	β_{F1}	-0.0556	-0.0428
Concentrate Liquid	β_{F2}	0.6111	0.5867
50 Applications	β_{A1}	0.4444	0.4571
100 Applications	β_{A2}	0.6111	0.5854
Disinfectant	β_{D1}	-0.2083	-0.2067
Biodegradable	β_{B1}	0.5417	0.5511
Price - \$0.35	β_{P1}	1.4444	1.4553
Price - \$0.49	β_{P2}	0.9444	0.9346

The estimated coefficients are very similar for both approaches, and when comparing the R^2 value for the single respondent with the R^2 obtained in the OLS regression approach in Chapter 3, the HB method yielded very similar result with $R^2 = 0.8625$ compared to $R^2 = 0.8630$ for the OLS regression. On investigation of the convergence rate of the Gibbs sampler approach it was found that the convergence stabilized within approximately 10 iterations. Figure 4.1.1 illustrated the quick convergence stabilization of all 9 parameters

as well as the variance component.

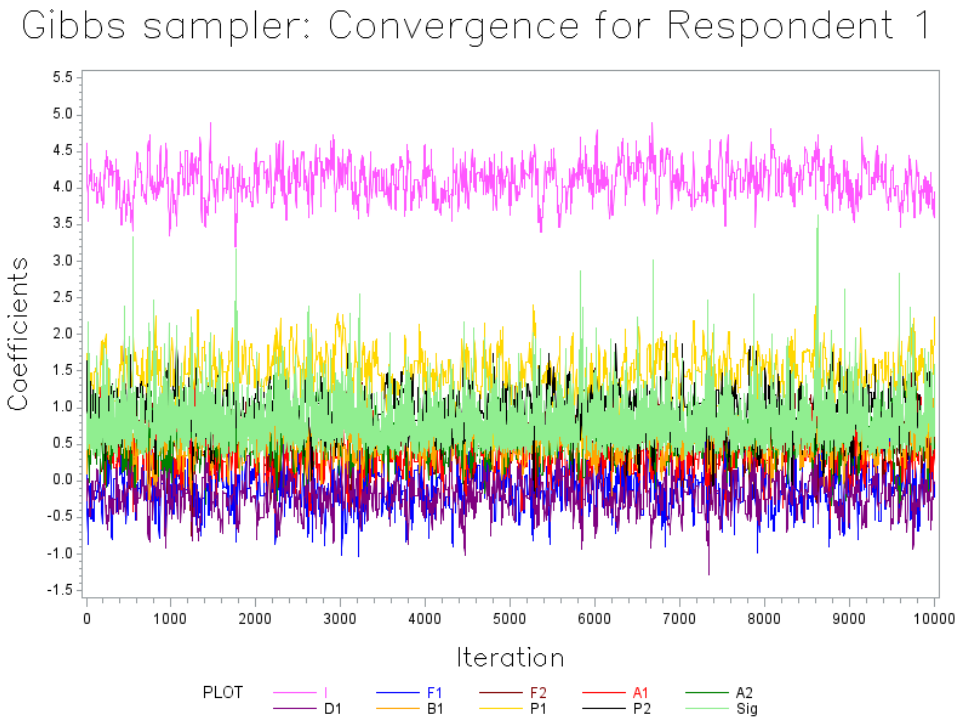
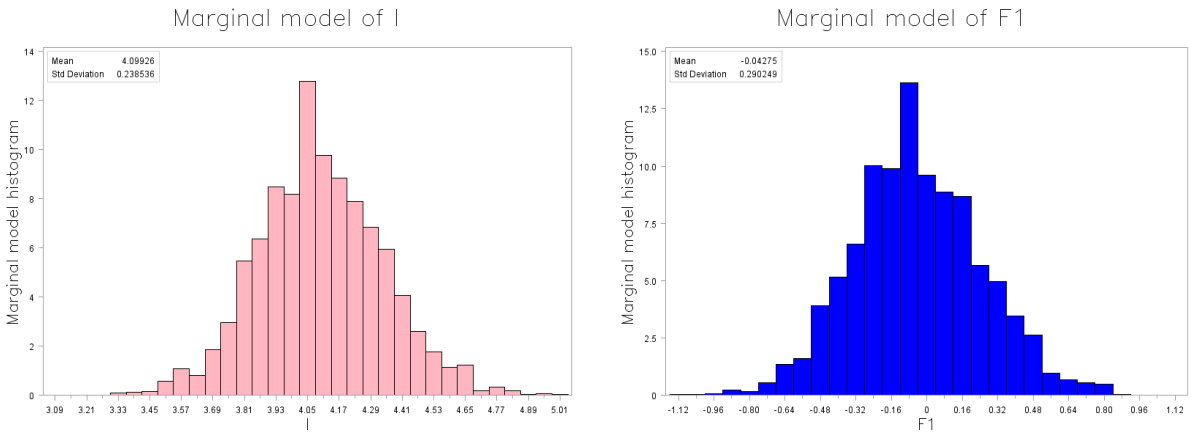


Figure 4.4.1 Convergence of Gibbs sampler (Respondent 1)

The marginal distributions of each of the estimated parameters, illustrated in Figure 4.4.2, all showcase the normal distribution shapes except the residual term which is slightly shewed to the right, as expected



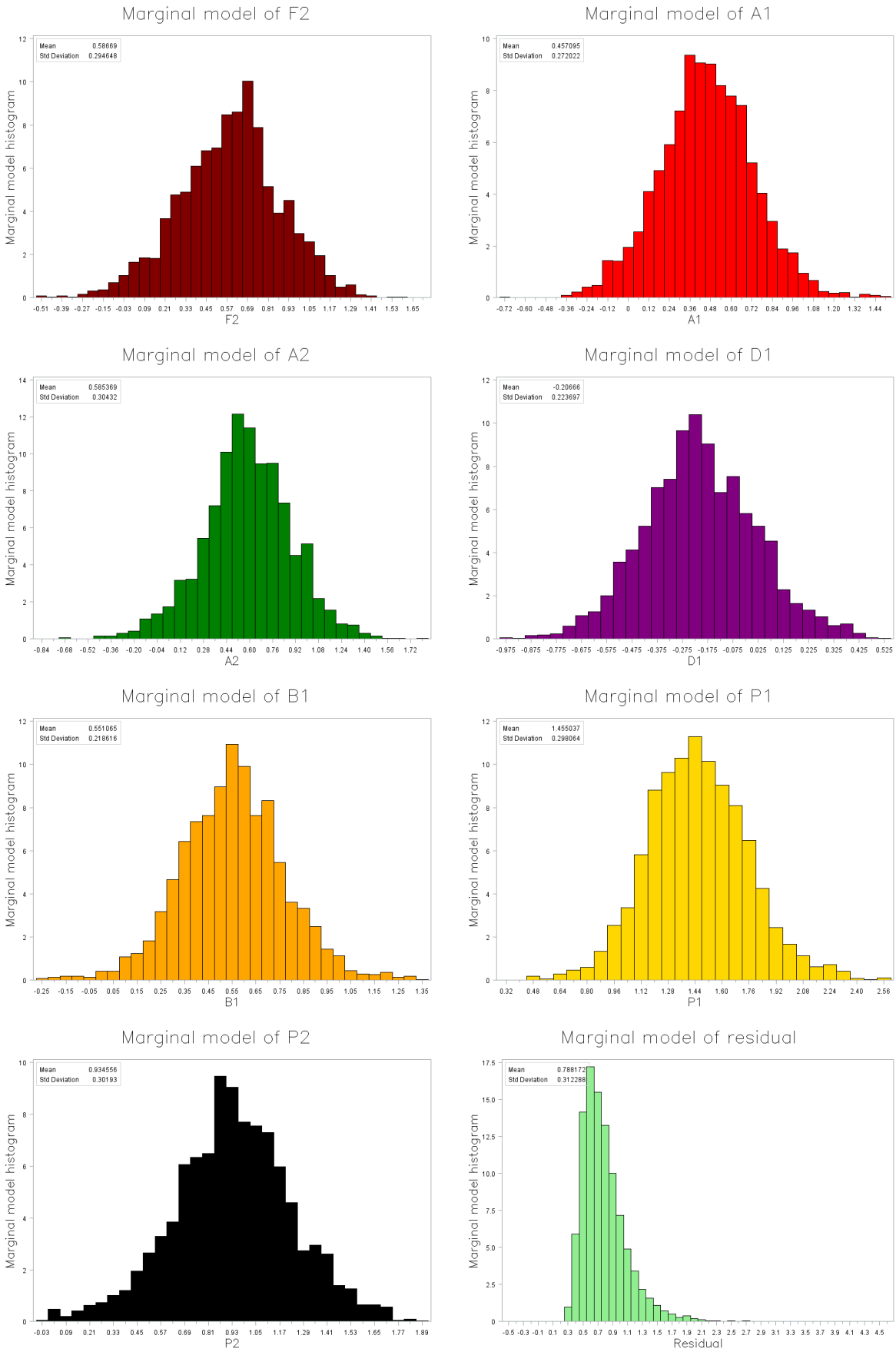


Figure 4.4.2 Marginal posterior distributions (Respondent 1)

4.4.2.2 Results obtained for the aggregate model

The estimated parameter values obtained for the aggregate model also yielded similar results as those obtained in the OLS estimation. These aggregate values where obtained in the traditional conjoint analysis way. For each iteration the average over all the respondents parameters where determined and stored to form the aggregate posterior distribution.

Table 4.2 Estimated parameters (Aggregate model)

Level	**Parameters**	**OLS estimation**	**HB estimation**
Variance component	σ^2	2.6087	2.6139
Intercept	β_I	3.7398	3.7401
Premixed Liquid	β_{F1}	-0.2171	-0.2173
Concentrate Liquid	β_{F2}	0.1667	0.1690
50 Applications	β_{A1}	-0.3450	-0.3449
100 Applications	β_{A2}	0.0233	0.0215
Disinfectant	β_{D1}	0.5102	0.5092
Biodegradable	β_{B1}	-0.1541	-0.1546
Price - \$0.35	β_{P1}	1.1318	1.1302
Price - \$0.49	β_{P2}	0.0814	0.0824

Comparing the AIC of the aggregate model, to the AIC obtained using the OLS method, the results are very similar. The Bayesian estimation method matched the OLS method almost exactly, with a AIC of 5903.8 compare to the AIC of the OLS method, 5897.4. On investigation of the convergence rate of the Gibbs sampler it was found that the estimated parameters stabalized even faster than in the individual respondents case, which can be expected since it is calculated as the average of the part-worth utilities taken over all the respondents. The marginal posterior distributions all showcase the normal distribution form, even the residual which illustrates the effect of the central limit theorem.

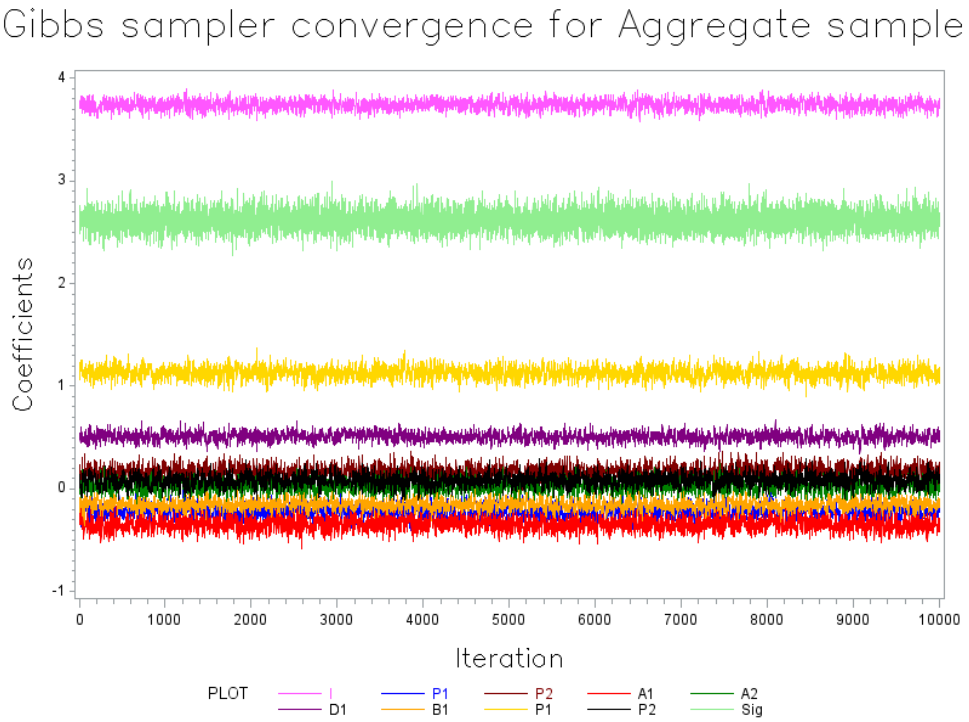
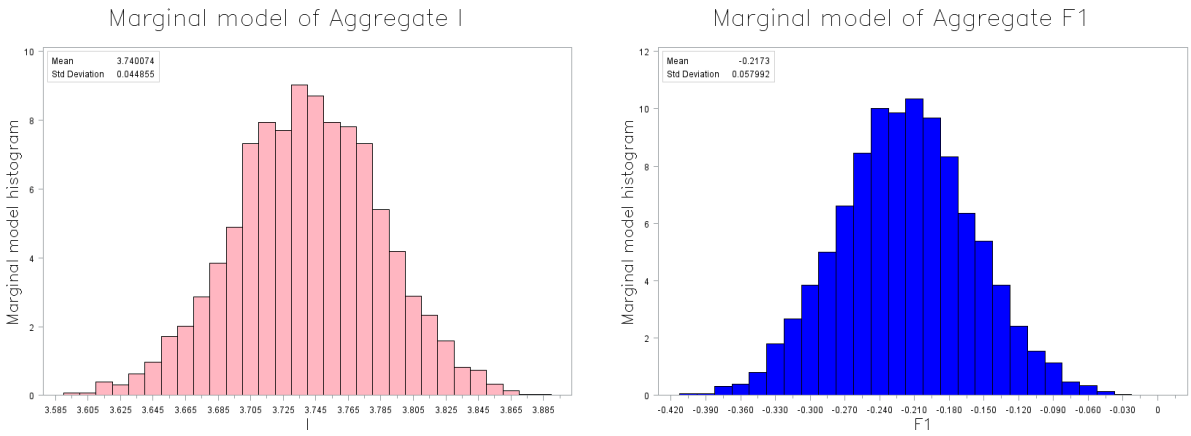


Figure 4.4.3 Convergence of Gibbs sampler (Aggregate model)



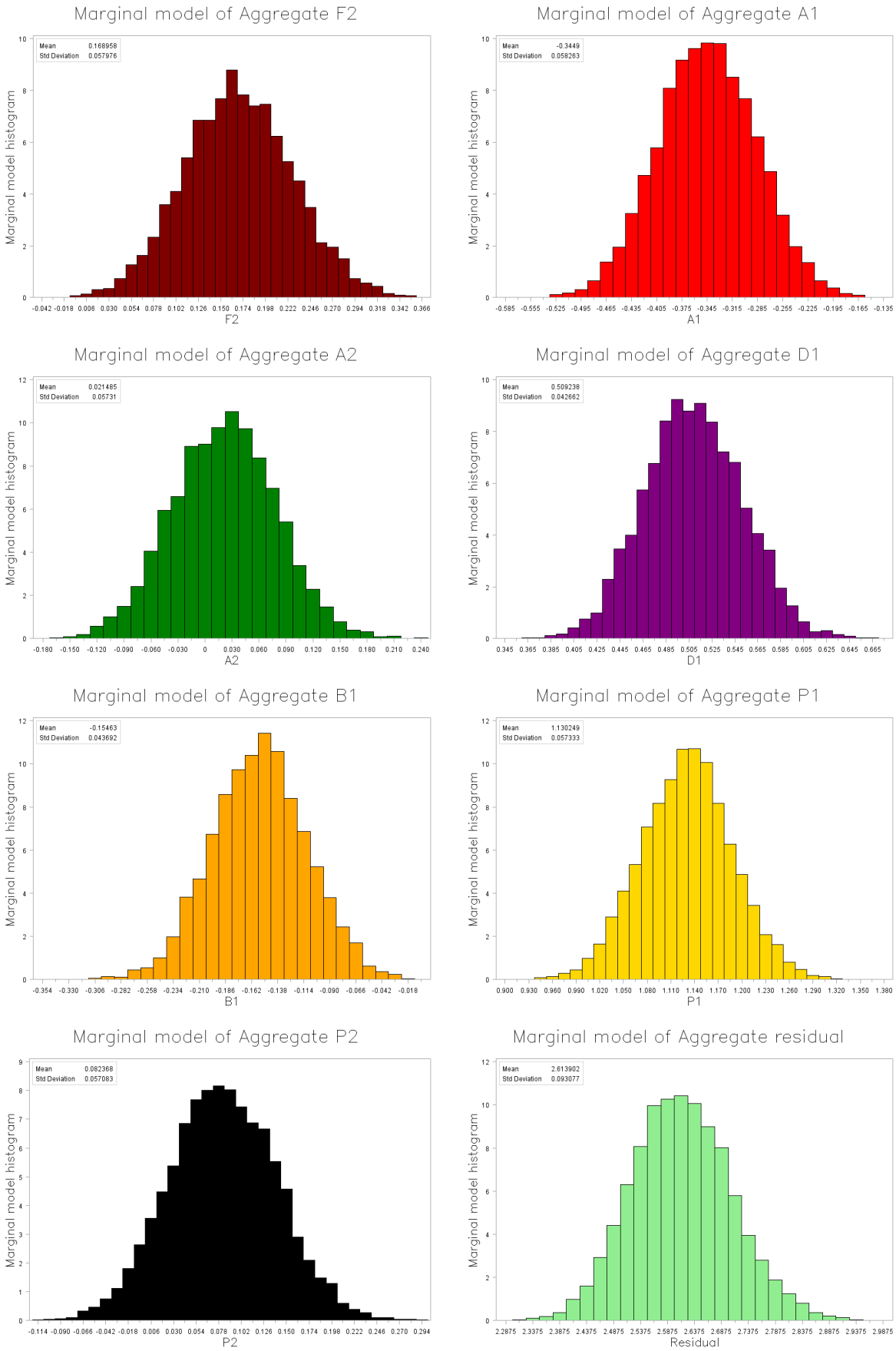


Figure 4.4.4 Marginal posterior distributions (Aggregate model)

4.5 Bayesian mixed effects model

The Bayesian approach can also be applied to linear mixed effect models. In the classical approach, used in Chapter 3, a distinction of fixed vs random effects are required. For a Bayesian approach it is only important whether the distributional parameters are known or unknown. Consider the stacked linear mixed effects model as described in Chapter 3.3.3. This model literally stack all the respondents matrices on top of the others to form one big matrix that contains the information of all m respondents

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}. \tag{4.5}$$

$\mathbf{y}$: $N \times 1$ vector of stacked responses, $\mathbf{X}$: $N \times (p+1)$ and $\mathbf{Z}$: $N \times (q+1)$ are considered fixed and known matrices of stacked $\mathbf{X}_i's$ and $\mathbf{Z}_i's$, $\boldsymbol{\beta} : (p+1) \times 1$ is the vector containing the unknown fixed parameters, $\mathbf{b} : m(q+1) \times 1$ is a stacked vector of unknown random variables and $\boldsymbol{\varepsilon}$:$N \times 1$ is the stacked vector of unknown error terms. Visually this model can be represented as:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{Z}_m \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{pmatrix}$$

Note that $\mathbf{R} = diag(\mathbf{R}_1, \mathbf{R}_2 ... \mathbf{R}_m)$ and $\mathbf{D} = diag(\mathbf{D}, \mathbf{D}, ... \mathbf{D})$ are both symmetric block designs.

For the Bayesian method the estimation of $\boldsymbol{\beta}$ and $\mathbf{b}$ will be handled in the exact same way and no distinction will be made between fixed and random components (Searle et al., 2006). The Bayesian approach takes into consideration the two possible cases that can occur in the linear mixed effect model. The variance components will either be known ($\mathbf{V}$ known) or, as is the case in almost all the real world situations, the variance components can be unknown ($\mathbf{V}$ unknown).

4.5.1 Estimation process when the variance components are known

When $\mathbf{V}$ is known, $\mathbf{D}$ and $\mathbf{R}$ are also assumed to be known. The only parameter that will then need to be estimated are the overall means, $\boldsymbol{\beta}$, and the random effects, $\mathbf{b}$. This means, prior distributions will be assumed for the mean parameters but not for the variance components.

4.5.1.1 Posterior distribution of $\boldsymbol{\beta}$

Since the distributions of $\mathbf{y}$, $\boldsymbol{\beta}$ and $\mathbf{b}$ are all normal, the joint distribution will be also be normal. The distribution of interest will be the posterior distribution of $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}|\mathbf{y})$ and can be determined by using Bayes' Theorem (Searle et al., 2006):

$$p(\boldsymbol{\beta}|\mathbf{y}) = \frac{\int p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b})p(\boldsymbol{\beta})p(\mathbf{b})d\mathbf{b}}{\int\int p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b})p(\boldsymbol{\beta})p(\mathbf{b})d\mathbf{b}d\boldsymbol{\beta}}.$$

Given that

$$\mathbf{y}|\boldsymbol{\beta},\mathbf{b},\mathbf{R} \backsim N(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b},\mathbf{R})$$

$$\boldsymbol{\beta} \backsim N(\boldsymbol{\beta}_o,\boldsymbol{\Sigma}_o) \text{ and } \mathbf{b} \backsim N(\mathbf{0},\mathbf{D})$$

it can be calculated that

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b})p(\boldsymbol{\beta})p(\mathbf{b}) &= \frac{1}{(2\pi)^{N/2}|\mathbf{R}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b})\right\} \\ &\times\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_o|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_o)'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_o)\right\} \\ &\times\frac{1}{(2\pi)^{q/2}|\mathbf{D}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}\right\} \end{aligned}$$

where N, p and q are the dimensions of $\mathbf{y}, \boldsymbol{\beta}$ and $\mathbf{d}$. This joint density can now be factored into the following order:

$$p(\mathbf{y},\boldsymbol{\beta},\mathbf{b}) = p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b})p(\boldsymbol{\beta})p(\mathbf{b}) = p(\mathbf{b}|\boldsymbol{\beta},\mathbf{y})p(\boldsymbol{\beta}|\mathbf{y})p(\mathbf{y})$$

where $p(\boldsymbol{\beta}|\mathbf{y})$ is the density of interest. To obtain this decomposition requires some algebraic work and repeated use of *Result 11.* These manipulations result in (Searle et al., 2006):

$$\begin{aligned} p(\mathbf{y},\boldsymbol{\beta},\mathbf{b}) &= p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b})p(\boldsymbol{\beta})p(\mathbf{b}) = p(\mathbf{b}|\boldsymbol{\beta},\mathbf{y})p(\boldsymbol{\beta}|\mathbf{y})p(\mathbf{y}) \\ &= \frac{1}{(2\pi)^{q/2}|\mathbf{A}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{b}-E\left[\mathbf{u}|\boldsymbol{\beta},\mathbf{y}\right]\right)'\mathbf{A}^{-1}\left(\mathbf{b}-E\left[\mathbf{u}|\boldsymbol{\beta},\mathbf{y}\right]\right)\right\} \\ &\times\frac{1}{(2\pi)^{p/2}|\mathbf{C}^{-1}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-E\left[\boldsymbol{\beta}|\mathbf{y}\right]\right)'C\left(\boldsymbol{\beta}-E\left[\boldsymbol{\beta}|\mathbf{y}\right]\right)\right\} \\ &\times\frac{1}{(2\pi)^{N/2}|\mathbf{L}-\mathbf{L}\mathbf{X}\mathbf{C}^{-1}\mathbf{X}'\mathbf{L}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{y}-E\left[\mathbf{y}\right]\right)'\left(\mathbf{L}-\mathbf{L}\mathbf{X}\mathbf{C}^{-1}\mathbf{X}'\mathbf{L}\right)\left(\mathbf{y}-E\left[\mathbf{y}\right]\right)\right\} \end{aligned}$$

where

$$\mathbf{A} = \mathbf{D}^{-1}+\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}$$

$$\mathbf{L} = \mathbf{R}^{-1}+\mathbf{R}^{-1}\mathbf{Z}\mathbf{A}^{-1}\mathbf{Z}'\mathbf{R}^{-1} = (\mathbf{Z}\mathbf{D}\mathbf{Z}'+\mathbf{R})^{-1} = \mathbf{V}^{-1}$$

$$\mathbf{C} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \boldsymbol{\Sigma}_o^{-1}$$
$$E\left[\mathbf{u}|\boldsymbol{\beta},\mathbf{y}\right] = \mathbf{A}^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$
$$E\left[\boldsymbol{\beta}|\mathbf{y}\right] = \mathbf{C}^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + \boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o)$$
$$E\left[\mathbf{y}\right] = (\mathbf{L} - \mathbf{L}\mathbf{X}\mathbf{C}^{-1}\mathbf{X}'\mathbf{L})^{-1}\mathbf{L}'\mathbf{X}\mathbf{C}^{-1}\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o.$$

From this manipulations the density of interest, $p(\boldsymbol{\beta}|\mathbf{y})$, was obtained and since it is known that all the densities are normal the conditional distribution is given by

$$\boldsymbol{\beta}|\mathbf{y} \backsim \mathbf{N}\left(E\left[\boldsymbol{\beta}|\mathbf{y}\right], \mathbf{C}^{-1}\right)$$

with the Bayesian estimates determined by

$$\widehat{\boldsymbol{\beta}}_B = E\left[\boldsymbol{\beta}|\mathbf{y}\right] = \mathbf{C}^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + \boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o)$$

$$Var(\widehat{\boldsymbol{\beta}}_B) = Var(E\left[\boldsymbol{\beta}|\mathbf{y}\right]) = \mathbf{C}^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{C}^{-1}.$$

4.5.1.2 Estimation of b

The estimation of the random coefficients follows in a similar way. This joint density can now be factored into the following order:

$$p(\mathbf{y},\boldsymbol{\beta},\mathbf{b}) = p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b})p(\boldsymbol{\beta})p(\mathbf{b}) = p(\boldsymbol{\beta}|\mathbf{y},\mathbf{b})p(\mathbf{b}|\mathbf{y})p(\mathbf{y})$$

where $p(\mathbf{b}|\mathbf{y})$ is the density of interest. This decomposition can once again be obtained using algebraic calculations and *Result 11* to obtain (Searle et al., 2006):

$$\begin{aligned} p(\mathbf{y},\boldsymbol{\beta},\mathbf{b}) &= p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b})p(\boldsymbol{\beta})p(\mathbf{b}) = p(\beta|\mathbf{b},\mathbf{y})p(\mathbf{b}|\mathbf{y})p(\mathbf{y}) \\ &= \frac{1}{(2\pi)^{p/2}|\mathcal{A}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta} - E\left[\boldsymbol{\beta}|\mathbf{b},\mathbf{y}\right]\right)'\mathcal{A}^{-1}\left(\boldsymbol{\beta} - E\left[\boldsymbol{\beta}|\mathbf{b},\mathbf{y}\right]\right)\right\} \\ &\quad\times\frac{1}{(2\pi)^{q/2}|\mathcal{C}|^{-\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{b} - E\left[\mathbf{u}|\mathbf{y}\right]\right)'\mathcal{C}(\mathbf{b} - E\left[\mathbf{u}|\mathbf{y}\right])\right\} \\ &\quad\times\frac{1}{(2\pi)^{N/2}|\mathbf{L} - \mathbf{L}\mathbf{X}\mathbf{C}^{-1}\mathbf{X}'\mathbf{L}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{y} - E\left[\mathbf{y}\right]\right)'\left(\mathbf{L} - \mathbf{L}\mathbf{X}\mathbf{C}^{-1}\mathbf{X}'\mathbf{L}\right)\left(\mathbf{y} - E\left[\mathbf{y}\right]\right)\right\} \end{aligned}$$

where

$$\mathcal{A} = \boldsymbol{\Sigma}_o^{-1} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}$$
$$\mathcal{L} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}\mathbf{A}^{-1}\mathbf{X}'\mathbf{R}^{-1} = (\mathbf{X}\boldsymbol{\Sigma}_o^{-1}\mathbf{X}' + \mathbf{R})^{-1}$$
$$\mathcal{C} = \mathbf{Z}'\mathcal{L}\mathbf{Z} + \mathbf{D}^{-1}$$
$$E\left[\boldsymbol{\beta}|\mathbf{b},\mathbf{y}\right] = \mathcal{A}^{-1}\left[\mathbf{X}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{Z}\mathbf{b}) + \boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o\right]$$
$$E\left[\mathbf{u}|\mathbf{y}\right] = \mathcal{C}^{-1}\mathbf{Z}'\mathcal{L}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_o).$$

From this manipulation, it can be seen that the distribution of interest will be normally distributed as:

$$\mathbf{b}|\mathbf{y} \backsim \mathbf{N}(\mathbf{E}\left[\mathbf{u}|\mathbf{y}\right], \mathcal{C}^{-1})$$

with the Bayesian estimates determined by

$$\widehat{b}_B = \mathbf{E}\left[\mathbf{u}|\mathbf{y}\right] = \mathcal{C}^{-1}\mathbf{Z}'\mathcal{L}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_o)$$
$$Var(\widehat{b}_B) = \mathcal{C}^{-1} = \left(\mathbf{Z}'\mathcal{L}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}.$$

4.5.2 Estimation process when variance components are unknown

When the variance components, $\mathbf{D}$ and $\mathbf{R} = \sigma^2\mathbf{I}_N$, are not known, prior distributions need to be chosen for the variance components as well. The following prior distributions were considered:

$$\mathbf{D} \backsim \mathbf{W}_q^{-1}(\mathbf{D}_o^{-1}, \boldsymbol{\rho}_o) \quad \text{and} \quad \sigma^{-2} \backsim G\left(\frac{\upsilon_o}{2}, \frac{\upsilon_o\tau_o}{2}\right) \tag{4.6}$$

where $\mathbf{W}_q^{-1}(.)$ denotes an inverse Wishart distribution and $G(.)$ represents a Gamma distribution. The joint posterior distribution for all unknown parameters and random effects will then be given by:

$$p(\boldsymbol{\beta},\sigma^2, \mathbf{D}, \mathbf{b}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\beta}, \sigma, \mathbf{b})p(\mathbf{b}|\mathbf{D})p(\boldsymbol{\beta}, \sigma, \mathbf{D})}{p(\mathbf{y})}$$

where

$$p(\mathbf{y}) = \iiiint p(\mathbf{y}|\boldsymbol{\beta}, \sigma, \mathbf{b})p(\mathbf{b}|\mathbf{D})p(\boldsymbol{\beta}, \sigma, \mathbf{D})d\boldsymbol{\beta}d\sigma d\mathbf{D}d\mathbf{b}$$

hence

$$p(\boldsymbol{\beta},\sigma^2, \mathbf{D}, \mathbf{b}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma, \mathbf{b})p(\mathbf{b}|\mathbf{D})p(\boldsymbol{\beta}, \sigma, \mathbf{D}).$$

It is common practice to assume that the prior distributions are independent which greatly simplifies the computations. This can be illustrated by:

$$p(\boldsymbol{\beta}, \sigma, \mathbf{D}) = p(\boldsymbol{\beta})p(\sigma)p(\mathbf{D})$$

Another assumption that is common when modelling linear mixed models, is to assume that the covariance matrices $\boldsymbol{\Sigma}_o$ and $\mathbf{D}_o$ are diagonal, which greatly reduces the number of parameters that need estimation (Wu, 2010).

► **Bayesian estimation for random coefficient model:**
Consider the linear mixed effect model discussed in Chapter 3.3, also sometimes referred

to as the random coefficient model. For this model $\mathbf{X}_i = \mathbf{Z}_i$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{b}_i$ Let $\beta_* = \{\boldsymbol{\beta}_{1*}, \boldsymbol{\beta}_{2*} ... \boldsymbol{\beta}_{m*}\}$ be the individual-spesific parameters. The conditional distributions for the parameters will then be given by (Wu, 2010):

$$\{\boldsymbol{\beta}|\mathbf{y}, \sigma, \mathbf{D}, \boldsymbol{\beta}_*\} \backsim N(\mathbf{U}(m\mathbf{D}^{-1}\overline{\beta} + \boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o), \mathbf{U})$$

$$\{\mathbf{D}|\mathbf{y}, \sigma, \boldsymbol{\beta}, \boldsymbol{\beta}_*\} \backsim W^{-1}\left(\left[\sum_{i=1}^{m}(\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})' + \mathbf{D}_o\right]^{-1}, m + \boldsymbol{\rho}_o\right)$$

$$\left\{\sigma^{-2}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\beta}_*\right\} \backsim G\left(\frac{\nu_o + N}{2}, \frac{1}{2}\left[\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i) + \nu_o\tau_o\right]\right)$$

$$\left\{\beta_i|\mathbf{y}, \boldsymbol{\beta}, \sigma, \mathbf{D}, \boldsymbol{\beta}_j, j \neq i\right\} \backsim N(\mathbf{H}_i(\sigma^{-2}\mathbf{X}_i'\mathbf{y}_i + \mathbf{D}^{-1}\boldsymbol{\beta}), \mathbf{H}_i)$$

where

$$\overline{\beta} = \sum_{i=1}^{m}\frac{\boldsymbol{\beta}_i}{m}, \quad H_i = \sigma^{-2}\mathbf{X}_i'\mathbf{X}_i + \mathbf{D}^{-1}, \quad \mathbf{U}^{-1} = m\mathbf{D}^{-1} + \sigma_o^{-1}.$$

Two very important assumptions about this estimation method is (Wu, 2010):

1. The responses $\mathbf{y}_i$, the parameters $\boldsymbol{\beta}$, the random effects $\mathbf{b}_i$, and the error terms $\boldsymbol{\varepsilon}_i$, are all linked in a linear form.

2. The parameters $\boldsymbol{\beta}$, the random effects $\mathbf{b}_i$, and the error terms $\boldsymbol{\varepsilon}_i$, are all assumed to be normally distributed.

The Bayesian mixed effects model approach will now be applied to the industrial cleaner example, and compared to the mixed effects model discussed in Chapter 3.

4.5.3 Industrial cleaner conjoint analysis example

For the hierarchical Bayes approach applied to the linear mixed effects model for the industrial cleaner data the burn in period was specified as 5000 iteration, afterwhich each 10^{th} iterations' parameter estimates was stored for 25000 iterations. The point estimates of these sample of stored values was then determined and represented in Table 4.3 and Table 4.4. The part-worth utilities are very similar to those obtained in the classical linear mixed effect model. The variance components are similar but there are a few variance components that deviate from those obtained in the classical approach.

Table 4.3 Estimated part-worth utilities ($\beta's$) for the mixed effects model

Level	**Classical estimation**	**HB estimation**
Intercept	3.7398	3.7363
Premixed Liquid	−0.2171	−0.2178
Concentrate Liquid	0.1667	0.1692
50 Applications	−0.3450	−0.3435
100 Applications	0.0233	0.0236
Disinfectant	0.5102	0.5176
Biodegradable	−0.1541	−0.1516
Price - \$0.35	1.1318	1.1323
Price - \$0.49	0.0814	0.0857

Table 4.4 Estimated variance components for the mixed effects model

Level	**Classical estimation**	**HB estimation**
Residual	1	1.2190
Intercept	0.5352	0.6046
Premixed Liquid	0.0958	0.1135
Concentrate Liquid	0.0163	0.1559
50 Applications	0.1862	0.2409
100 Applications	0.0584	0.1052
Disinfectant	0.3055	0.2614
Biodegradable	0.0975	0.0623
Price - \$0.35	0.4918	0.3902
Price - \$0.49	0.1162	0.0626

When comparing the AIC of the classical approach to linear mixed effects model, 5460.9, the hierarchical Bayesian approached showed an improvement with an AIC of 4784.6.

This improvement is due to the fact that the Bayesian approach can separate the noise form the heterogeneity which is present in the model. The differences in the variance components is the biggest difference between the two models. On investigation of the convergence of the part-worth utility parameters and the variance components of the random effect, it can be seen that these estimates did indeed stabilize well, as illustrated in Figure 4.5.1 and 4.5.2.

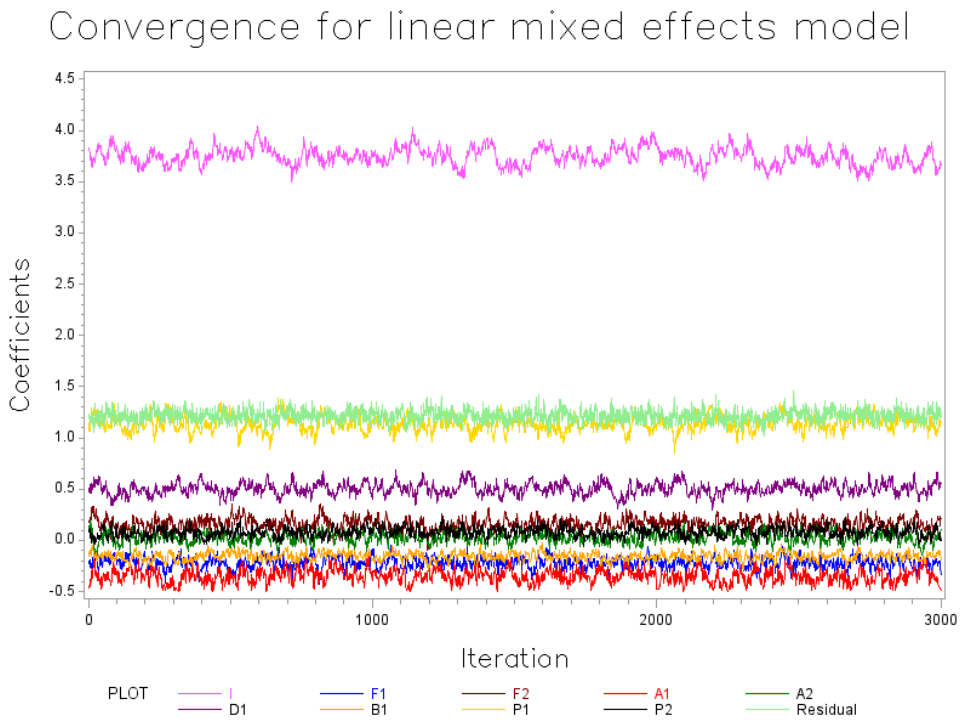


Figure 4.5.1 Convergence of part-worth utilities for HB mixed effects model

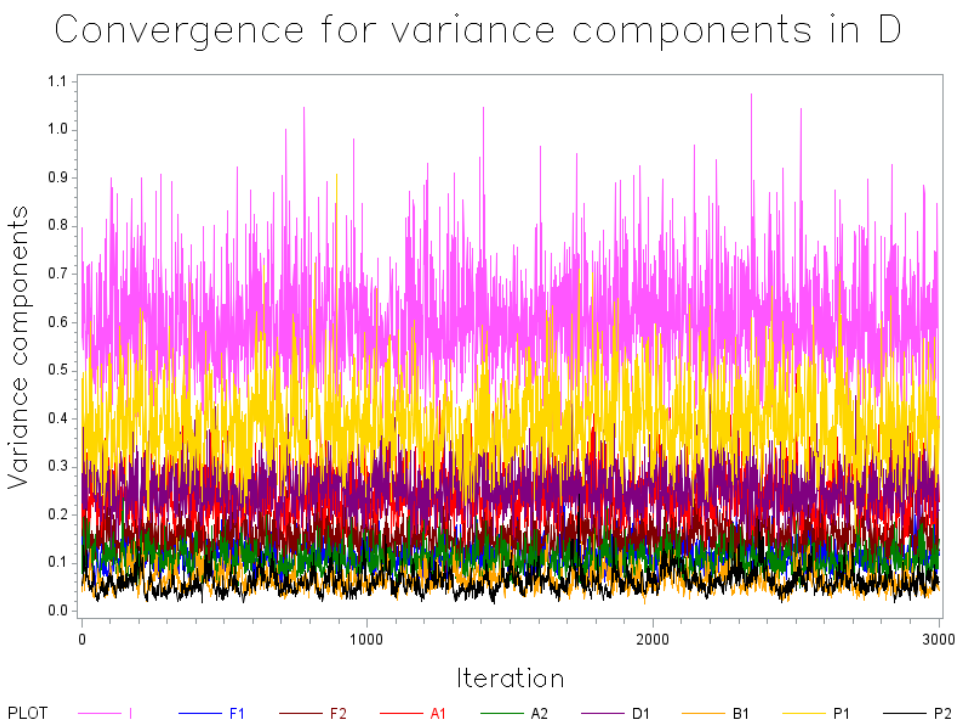
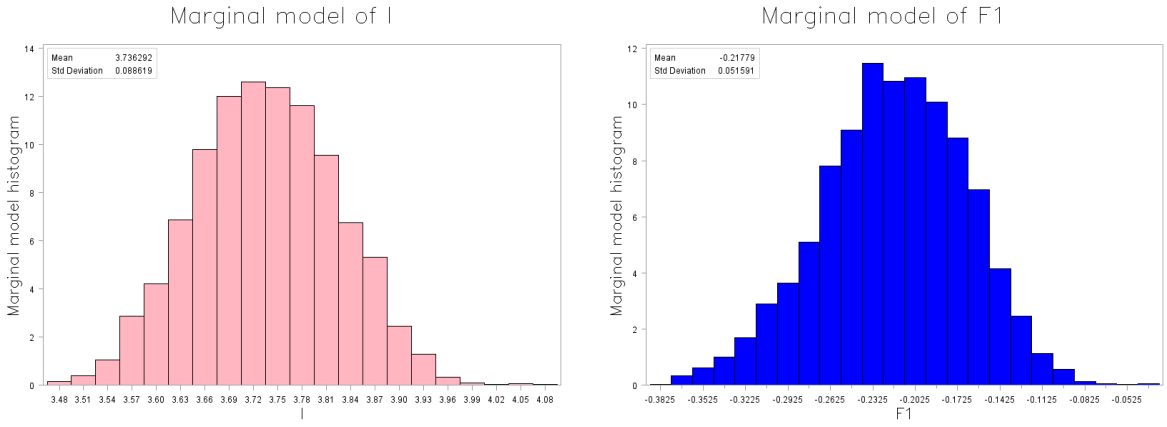


Figure 4.5.2 Convergence of variance components for HB mixed effects model

The marginal posterior distributions of the part-worth utilities (the fixed effects) as well as the marginal posterior distributions of the variance components of each attribute level modelled in the HB model is illustrated in Figure 4.5.3 and 4.5.4.



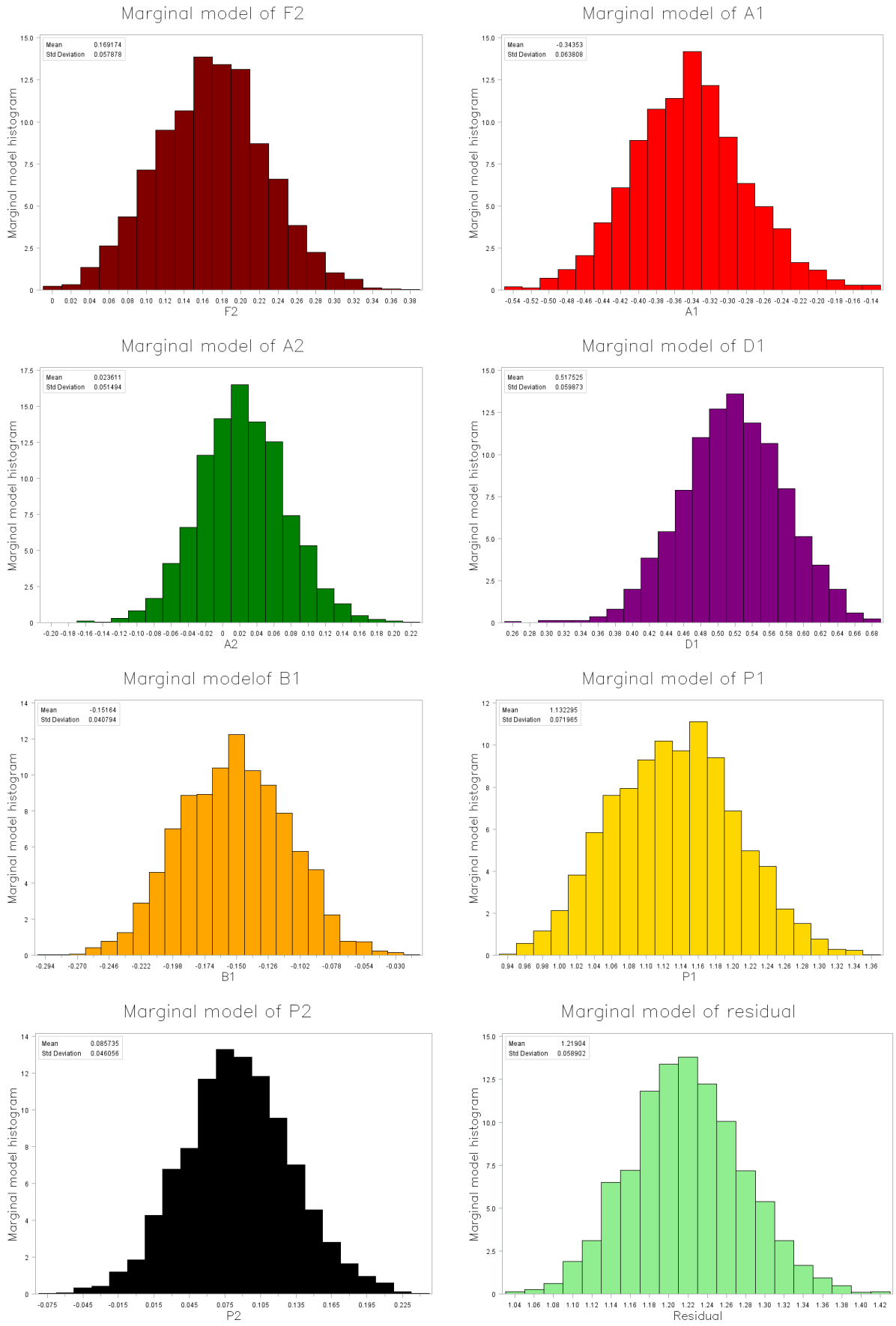
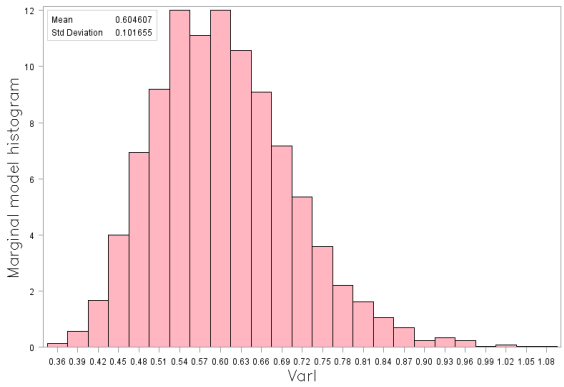
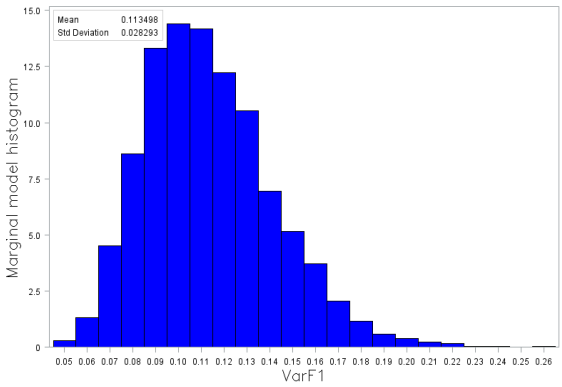


Figure 4.5.3 Marginal distributions for the part-worth utilities of HB mixed effects model

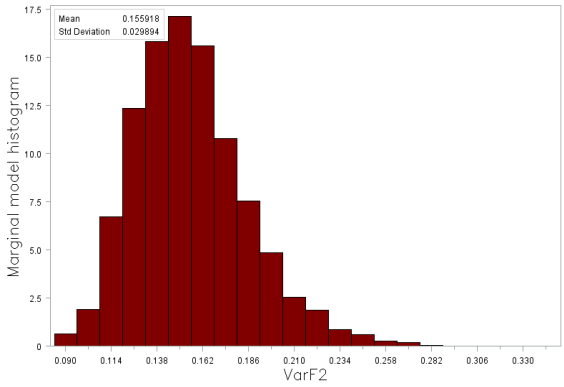
Marginal model of VarI



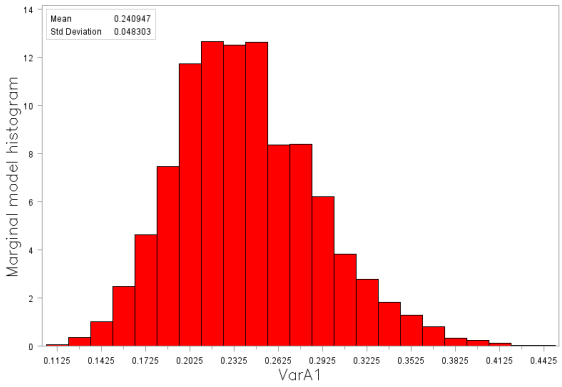
Marginal model of VarF1



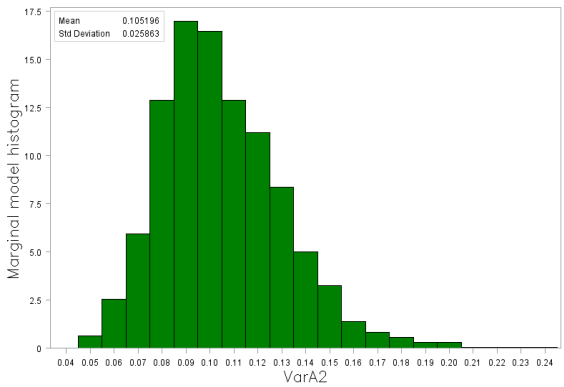
Marginal model of VarF2



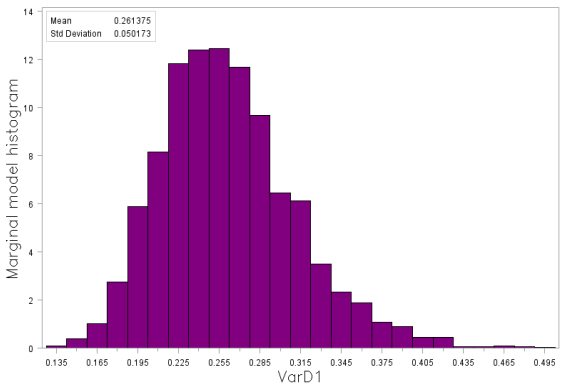
Marginal model of VarA1



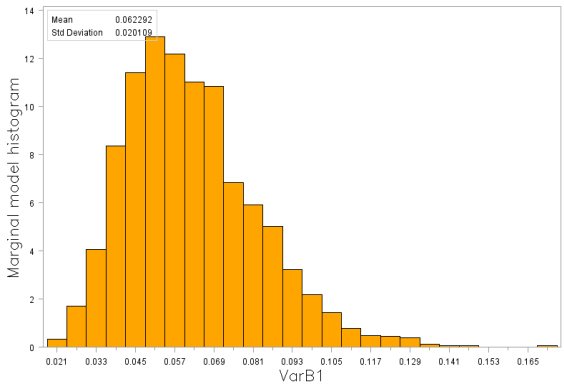
Marginal model of VarA2



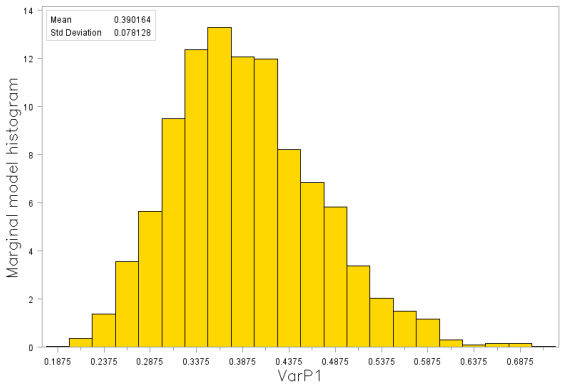
Marginal model of VarD1



Marginal model of VarB1



Marginal model of VarP1



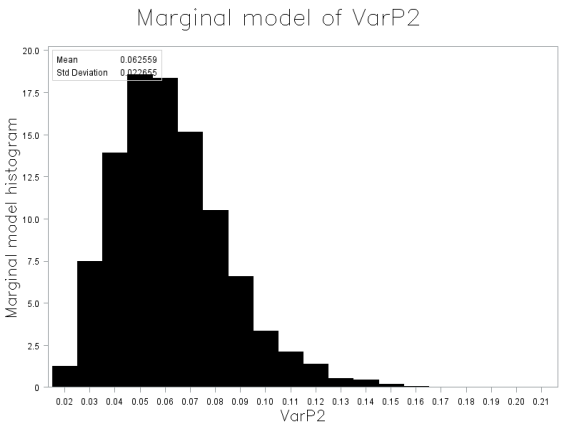


Figure 4.5.4 Marginal posterior distributions for variance components of the random effects of HB mixed effects model

All of the marginal posterior densities of the part-worth utilities are close to that of normal densities while the variance components all have a slight shewed to the right distributional form. The direct accessibility of the parameter densities presents a major advantage of the MCMC estimation. Irregular densities immediately highlight serious violation of distributional assumptions that might go undetected otherwise. Relying on asymptotic distributions and standard errors from the information matrix, as the classical approached do, might give misleading results in such cases (Frühwirth-Schatter and Otter, 1999). This example illustrates the power of the hierarchical Bayesian approach and motivates the inclusion of this method in conjoint studies as it can yield valuable results. Since conjoint analysis studies do often have heterogeneity present in the model, methods like these could be very beneficial.

Chapter 5

Conclusion and Future Directions

"Study the past if you would define the future."
-Confucius

5.1 Conclusion

In the business driven world conjoint analysis is a powerful tool that companies can employ to predict the preferences of consumers in certain market segments. Being able to accurately model consumers preferences has immense value and can increase a companies profits. In Chapter 2 some of the main conjoint analysis designs and techniques was introduced and Chapter 3 started off by introducing the traditional conjoint analysis method. Although the traditional conjoint analysis method yields valuable insight and can relatively accurately model consumers preferences, there is room for improvement. By modeling the conjoint analysis data as a linear mixed effects model, insights regarding the misspesification of traditional models, can be gained. A large part of the variation in the model can be explained by the possible heterogeneity between the individuals' preference structures. The researcher can build a more accurate predictive model by considering these differences. By obtaining additional covariates that can possibly explain these differences, a model that explains more variation can be constructed. Finally, Chapter 4 showcased the hierarchical Bayesian approach which revealed that the heterogeneity present in the model could be slightly better managed. The inclusion of Bayesian approaches can hold valuable information regarding the distributional assumptions.

5.2 Future direction

Possible extension and future directions that can be considered include, the possibility of the violation of the normality assumption. When this is the case in a conjoint analysis

study the researcher can consider heavy tailed distributions such as the t-distribution or the elliptical distribution. Another extension which can yield possible interesting results, is considering a Structural Equation Model (SEM) approach. This technique is also based on multivariate regression models, but unlike the traditional linear model, the response variable in one regression equation in the SEM may appear as a predictor in another equation (Lee, 2007). This approach can hold valuable results, as one of the main problems in conjoint analysis is finding latent structures between consumers.

APPENDICES

A. Abbreviations and Notation

pdf	Probability density function
$N(\mu, \sigma^2)$	Normal distribution with mean μ, and variance σ^2
$G(\alpha, \beta)$	Gamma distribution with shape α and scale β parameters
$IG(\alpha, \beta)$	Inverse-gamma distribution with shape α and scale β parameters
$\chi^2(\nu)$	Chi-square distribution with degrees ν of freedom
$N(\boldsymbol{\mu}, \Sigma)$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance covariance matrix $\Sigma = \sigma^2\mathbf{I}_n$
$W_p(V, n)$	Wishart distribution with n degrees of freedom and a positive definite $p \times p$ matrix V

Chapter 2

$p = 1, 2...t$	Attributes
$k = 1, 2...K$	Attribute levels
$j = 1, 2...J$	Stimuli
x_{jp}	Value of the level of the p^{th} attribute in the j^{th} stimuli
y_j	Individual respondents preference rating for the j^{th} stimuli
ω_p	Individual respondents importance weight for attribute p
x_p	Individual respondents ideal point for attribute p
d_j^2	Weighted squared distance of x_{jp} from x_p
β_{jp}	Part-worth utility for the level of attribute p in stimuli j
I_{ji}	Importance of attribute j for person i
D_{jki}	Desirability for level k of attribute j for person i
E_i	Evaluation of product for person i
$\mathbf{X}_{n\times p}$	Full rank design matrix
$\boldsymbol{\beta}_j$	Vector of part-worth utilities for respondent j
$\Phi(x)$	Gaussian link function
$\Phi^{-1}(x)$	Inverse Gaussian link function

Chapter 3 and 4

n	Number of observations
n_i	Number of observations in group i
k	Number of groups
m	Number of respondents
p	Number of fixed effects
q	Number of random effects
$N = nm$	Total number of observations
μ	Overall mean
τ_i	Random effect for group i
σ^2	Residual variance component
σ^2_τ	Random effect variance component
$\mathbf{y}_i : n \times 1$	Vector of responses
$\mathbf{X}_i : n \times p$	A full rank, unstochastic design matrix for fixed components
$\boldsymbol{\beta}_i : p \times 1$	Vector of unknown regression coefficients
$\mathbf{Z}_i : n \times q$	A full rank, unstochastic design matrix for random components
$\mathbf{b}_i : q \times 1$	Vector of unknown random components
$\boldsymbol{\varepsilon}_i : n \times 1$	Vector of error terms
σ^2	Variance of residual
$\mathbf{D}_i : q \times q$	Positive definite diagonal variance-covariance matrix
$\mathbf{R}_i : n \times n$	Positive definite diagonal variance-covariance matrix of residual terms
$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}_i\mathbf{Z}_i' + \mathbf{R}_i$	A $n \times n$ positive definite variance-covariance matrix
$\boldsymbol{\theta}$	Vector containing parameters of interest.
$\mathbf{y} : N \times 1$	Stacked vector of responses
$\mathbf{X} : N \times p$	Stacked design matrix for fixed components
$\boldsymbol{\beta} : p \times 1$	Vector of unknown regression coefficients
$\mathbf{Z} : N \times mq$	Block design such that $\mathbf{Z} = diag(\mathbf{Z}_i)$
$\mathbf{b} : mq \times 1$	Stacked vector of unknown random components
$\boldsymbol{\varepsilon} : N \times 1$	Stacked vector of error terms
$\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$	$N \times N$ positive definite matrix
$\mathbf{D} : qm \times qm$	Block design such that $\mathbf{D} = diag(\mathbf{D}_i)$
$R : nm \times nm$	Block design such that $\mathbf{R} = diag(\mathbf{R}_i)$
$\boldsymbol{\Sigma}_o = \sigma^2(X'X)^{-1}$	Chosen variance-covariance component
$\gamma = \frac{1}{\sigma^2}$	Measurement precision
υ_o	Degrees of freedom

B. Results

Result 1 *(Johnson and Wichern, 2007)*
Given that

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}.\end{aligned}$$

The expected value of $\widehat{\boldsymbol{\beta}}$ *will be:*

$$\begin{aligned}E(\widehat{\boldsymbol{\beta}}) &= E\left[\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right] \\ &= E(\boldsymbol{\beta})+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')E(\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}.\end{aligned}$$

The variance of $\widehat{\boldsymbol{\beta}}$ *will be:*

$$\begin{aligned}Var(\widehat{\boldsymbol{\beta}}) &= Var(\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}) \\ &= Var(\boldsymbol{\beta})+Var((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})+2Cov(\boldsymbol{\beta},(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}) \\ &= 0+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}+0 \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Then the estimated parameter will be normally distributed: $\widehat{\boldsymbol{\beta}} \backsim N_{p+1}(\boldsymbol{\beta},\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

Result 2 *(Johnson and Wichern, 2007)*
Given that

$$\begin{aligned}\widehat{\boldsymbol{\varepsilon}} &= \mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{Y}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \left[I-X(X'X)^{-1}X'\right]\mathbf{Y} \\ &= \left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\left[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right] \\ &= \left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}.\end{aligned}$$

The expected value of $\widehat{\varepsilon}$ *will be:*

$$\begin{aligned}E(\widehat{\boldsymbol{\varepsilon}}) &= E(\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}) \\ &= \left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]E(\boldsymbol{\varepsilon}) \\ &= \left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right](\mathbf{0}) \\ &= \mathbf{0}.\end{aligned}$$

The variance of $\widehat{\varepsilon}$ *will be:*

$$\begin{aligned} Cov(\widehat{\boldsymbol{\varepsilon}}) &= \left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] Cov(\boldsymbol{\varepsilon}) \left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]' \\ &= \left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}+\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\right]-\sigma^2 \\ &= \sigma^2\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]. \end{aligned}$$

The error terms are then normally distributed as $\widehat{\boldsymbol{\varepsilon}} \backsim N_n(\mathbf{0}, \sigma^2[\mathbf{I}-\mathbf{H}])$ *where* $\mathbf{H}$ *is an indempotent matix. A matrix will be considered indempotent if and only of* $\mathbf{HH}=\mathbf{H}$. *Given that*

$$\begin{aligned} \widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}' &= \boldsymbol{\varepsilon}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon} \\ &= tr(\boldsymbol{\varepsilon}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}) \\ &= tr(\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'). \end{aligned}$$

The expected value of $\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}'$ *will be:*

$$\begin{aligned} E(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}') &= E(tr(\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')) \\ &= tr(\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')) \\ &= \sigma^2 tr(\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]) \\ &= \sigma^2 tr(\mathbf{I})-\sigma^2 tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= n\sigma^2-\sigma^2 tr\begin{pmatrix}\mathbf{I}\\ (r+1)\times(r+1)\end{pmatrix} \\ &= n\sigma^2-\sigma^2(r-1) \\ &= \sigma^2(n-r-1). \end{aligned}$$

Result 3 *(Bain and Engelhardt, 1992, p271)*

Theorem 8.3.5

$$\textit{If } \mathbf{Z} \backsim N(0,1) \textit{ then } \mathbf{Z}^2 \backsim \chi^2(1)$$

Corollary 8.3.2

$$\textit{If } \mathbf{X}_1,\ldots\mathbf{X}_n \textit{ denotes a random sample from } N(\mu,\sigma^2) \textit{ then}$$

$$\sum_{i=1}^{n}\frac{(\mathbf{X}_i-u)^2}{\sigma^2} \backsim \chi^2(n)$$

Result 4 *Given that*

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

The expected value of $\mathbf{Y}$ *will be:*

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\ &= E(\mathbf{X}\boldsymbol{\beta}) + E(\boldsymbol{\varepsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{0} \\ &= \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

The variance of Y *will be:*

$$\begin{aligned} Var(\mathbf{Y}) &= Var(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\ &= Var(\mathbf{X}\boldsymbol{\beta}) + Var(\boldsymbol{\varepsilon}) + 2Cov(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\varepsilon}) \\ &= 0 + Var(\boldsymbol{\varepsilon}) + 0 \\ &= \sigma^2\mathbf{I}_n. \end{aligned}$$

Result 5 *Maximum Likelihood estimation derivations*
In order to pursue the theory of the ML method, the derivative with respect to $\underline{\boldsymbol{\theta}}$ *should be determined, to find the value for the parameter that will maximize the function. The solution of the following should be found:*

$$\frac{\partial LogL}{\partial \underline{\boldsymbol{\theta}}} = 0.$$

Since the dependent variable $\mathbf{y}$ *is normally distributed, the likelihood function will be given by:*

$$L = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\left(\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'}{\sigma^2}\right)}.$$

Determining the log of the likelihood function leads to the following expression:

$$\begin{aligned} LogL &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\boldsymbol{\beta}'\mathbf{y} + \mathbf{X}'\boldsymbol{\beta}'\mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

Firstly the estimates of the $\boldsymbol{\beta}'s$ *can be determined as follows:*

$$\begin{aligned} \frac{\partial LogL}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2}(\mathbf{0} - \mathbf{y}'\mathbf{X} - \mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) \\ &= \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\mathbf{X}' \\ &= \mathbf{0}. \end{aligned}$$

The results of the estimated values are exactly the same as those obtained using the OLS method in (3.3):

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

When calculating the estimated variance the same process can be followed, only in terms

of σ^2 :

$$\begin{aligned}\frac{\partial LogL}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\boldsymbol{\beta}'y + \mathbf{X}'\boldsymbol{\beta}'\mathbf{X}\boldsymbol{\beta}) \\ &= \frac{2\sigma^4}{n}\left[-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\boldsymbol{\beta}'y + \mathbf{X}'\boldsymbol{\beta}'\mathbf{X}\boldsymbol{\beta})\right] \\ &= -\sigma^2 + \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= 0.\end{aligned}$$

The estimate for the variance, determined using the ML method, is slightly different than the estimate obtained using the OLS method given by (3.5):

$$\widehat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n}.$$

However, this difference becomes negligibly small as $n \to \infty$.

Result 6 *(Bain and Engelhardt, 1992, p 275)*
If $V_1 \backsim \chi^2(\nu_1)$ *and* $V_2 \backsim \chi^2(\nu_2)$ *are independent, then the random variable*

$$\mathbf{X} = \frac{\mathbf{V}_1/\nu_1}{\mathbf{V}_2/\nu_2} \backsim F(\nu_1, \nu_2)$$

will have a F-distribution with ν_1 *and* ν_2 *degrees of freedom.*

Result 7 *(Johnson and Wichern, 2007)*

$$If \mathbf{Z} \backsim N_{\upsilon \times p}(\mathbf{0}, \mathbf{I}_\upsilon \otimes \boldsymbol{\Sigma})$$

$$then \mathbf{Z}'\mathbf{Z} \backsim W_p(\upsilon, \boldsymbol{\Sigma})$$

Result 8 *Conditional distributions*

Given the following variables that are both normally distributed:

$$\begin{pmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{pmatrix} = \left\{\begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}; \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right\}.$$

The distribution of x_1 *conditional on* x_2 *will then be a multivariate normal distribution with a mean of*

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_2 - \boldsymbol{\mu}_2)$$

and a variance-covariance structure of

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Result 9 *Since*

$$Var(\mathbf{X}) = E(\mathbf{X}^2) - (E(\mathbf{X}))^2$$

then

$$E(\mathbf{X}^2) = (E(\mathbf{X}))^2 + Var(\mathbf{X})$$

Result 10 *(Hoff, 2009) Gamma and inverse -gamma distributions.*
A random variable $\mathbf{X}$ *has a G(*α*,*β*) distribution for* $a > 0$, $b > 0$ *and*

$$p(x) = \frac{1}{\Gamma(a)\beta^{\alpha}} x^{a-1} e^{-\frac{x}{\beta}}$$

The random variable X *will have an IG(*α*,*β*) distribution if* $\frac{1}{X}$ *has a gamma(a,b) distribution. In other words, if* $Y \backsim gamma(\alpha, \beta)$ *and* $X = \frac{1}{Y}$, *then* $X \backsim IG(\alpha, \beta)$

$$p(x) = \frac{\beta^{\alpha}}{\Gamma(a)} x^{-a-1} e^{-\frac{\beta}{x}}$$

Result 11 *(Searle et al., 2006)*

$$\mathbf{x}'\mathbf{G}\mathbf{x} - 2\mathbf{x}'\mathbf{H}\mathbf{y} + \mathbf{y}'\mathbf{H}'\mathbf{G}^{-1}\mathbf{H}\mathbf{y} = (\mathbf{x} - \mathbf{G}^{-1}\mathbf{H}\mathbf{y})'\mathbf{G}(\mathbf{x} - \mathbf{G}^{-1}\mathbf{H}\mathbf{y}).$$

C. Illustrations, Figures and Tables

List of Illustrations

List of Figures

List of Tables

D. SAS 9.3 Program Codes

Program 1 - Traditional conjoint analysis model

```
proc iml;
       use rr.x;
       read all into X;
       use rr.y;
       read all into y;

   *Estimate part-worth values for all individuals;
       Beta = inv(X‘*X)*X‘*Y;
       A = mean(beta‘);*global means - Aggregate model;
       Estimated_fixed_effects = round(A‘,0.01);
       B = var(beta‘);
       Estimated_variance = round(B‘,0.01);
       print "Traditional Conjoint Approach to Industrial Cleaner Example";
       print Estimated_fixed_effects Estimated_variance; *Aggregate model;
       print Beta;

   *Determine R2 for all respondents with own beta’s;
       yhat = (X*beta);

   /*Stack for 86 respondents*/
   Do i = 1 to ncol(Y);
       YStack = YStack//Y[,i];
       yhatStack = YhatStack//yhat[,i];
   end;
   /* End of Stack Process*/

   SSR = sum((Ystack-yhatstack)##2);
   SST = sum((Ystack-mean(Ystack))##2);
   R2_Model_ownbetas = 1-SSR/SST;
   print R2_Model_ownbetas;
   YhatAgg = X*A‘;

   /*Stack for 86 respondents*/
   Do i = 1 to ncol(Y);
       yhatAggStack = YhatAggStack//yhatAgg;
   end;

   SSRAgg = sum((Ystack-yhataggstack)##2);
   SSTAgg = sum((Ystack-mean(Ystack))##2);
   R2_Model_avebeta = 1-SSRAgg/SSTAgg;
   print R2_Model_avebeta;

   /* Conjoint analysis for respondent 1 */
       yhat1 = (X*beta[,1]);
       Y1 = Y[,1];
       print Y1 yhat1;
       SSR = sum((Y1-yhat1)##2);
```

```
    SST = sum((Y1-mean(Y1))##2);
    R2_resp1_Model = 1-SSR/SST;
    print R2_resp1_Model;

*Calculate Overall Importance of each Attribute/Factor;
Range_Atr_1 = 0.22+0.17;
Range_Atr_2 = 0.35+0.33;
Range_Atr_3 = 0.51+0.51;
Range_Atr_4 = 0.15+0.15;
Range_Atr_5 = 1.13+1.21;

Total_Range = Range_Atr_1+Range_Atr_2+Range_Atr_3
       +Range_Atr_4+Range_Atr_5;
Range = (Range_Atr_1// Range_Atr_2// Range_Atr_3//
       Range_Atr_4// Range_Atr_5)||(Total_Range*J(5,1,1));
Name2 = {'Product Form' 'Number of Applications'
       'Disinfectant' 'Biodegradable' 'Price'};
Overall_Importance = round((Range[,1]/Range[,2])*100,0.01);

Title "The Overall Importance of each Attribute";
print Overall_Importance[rowname = name2];
Over_data = {1,2,3,4,5}||Overall_Importance;
name3 = {'Attribute' 'Overall_Importance'};

Create Over from Over_data[colname = name3];
append from Over_data;

/* Determine the predicted values of the 18 profiles */
Yhat = X*Estimated_fixed_effects;
rounded_yhat = round(yhat);
Print 'Estimated Profile Utilities ' Yhat rounded_yhat;

quit;
```

Program 2 - Officer ratings example

```
Data Random_example;
   input rating officer;
   cards;
    76 1
    65 1
    85 1
    74 1
    59 2
    75 2
    81 2
    67 2
    49 3
    63 3
    61 3
```

```
 46 3
 74 4
 71 4
 85 4
 89 4
 66 5
 84 5
 80 5
 79 5
;
goptions reset = all;
symbol1 color = lightpurple value = dot w = 2 h = 2;
title h = 3 f = simplex 'Graphical Representation of the Data';
axis1 lable = (a = 90 f = simplex h = 1.75);
axis2 lable = ( f = simplex h = 1.75) order = (0 to 6 by 1);
proc gplot data = random_example;
plot rating*officer/ vaxis = axis1 haxis = axis2;
run;
title;

proc means data = random_example;
output out = Averages mean = ave;
var rating;
by officer;
run;

goptions reset = all i = join;
title h = 4 f = simplex 'Graphical Representation of the Means';
symbol1 color = lightgreen value = dot line = 3 h = 2 w = 2;
axis1 label = (a=90 f = simplex h = 1.75 'Average rating');
axis2 label = (f = simplex h = 1.75 'Officer') order = (0 to 6 by 1);
proc gplot data = Averages;
plot ave*officer/vaxis = axis1 haxis = axis2;
run;
title;

proc iml;
use random_example;
read all into data;
Y1 = 0;
Y2 = 0;
Y3 = 0;
Y4 = 0;
Y5 = 0;
Do i = 1 to nrow(data);
    if data[i,2] = 1 then Y1 = Y1//data[i,1];
    if data[i,2] = 2 then Y2 = Y2//data[i,1];
```

```
    if data[i,2] = 3 then Y3 = Y3//data[i,1];
    if data[i,2] = 4 then Y4 = Y4//data[i,1];
    if data[i,2] = 5 then Y5 = Y5//data[i,1];
end;

Y1 = Y1[2:5,];
Y2 = Y2[2:5,];
Y3 = Y3[2:5,];
Y4 = Y4[2:5,];
Y5 = Y5[2:5,];
YY = Y1||Y2||Y3||Y4||Y5;
k = ncol(YY); *number of groups;
n = nrow(Y1);
NN = nrow(data);

GroupMean = mean(YY);
OverallMean = groupMean[,+]/k;

SE = (YY - GroupMean#J(nrow(YY), ncol(YY),1))##2;
SEE = SE[+,];
SSE = SEE[,+];
print 'The SSE is:  '  SSE;

MSE = SSE/(NN-k);
print 'The MSE is:  '  MSE;

SB = n#((GroupMean -
    OverallMean*J(nrow(groupmean),ncol(groupmean),1))##2);
SSB = sum(SB);
print 'The SSB is:  '  SSB;

MSB = SSB/(k-1);
print 'The MSB is:  '  MSB;

*Estimated parameter;
mu_hat = overallmean;
print 'The estimated mean is:  '  mu_hat;

SigE2_hat = MSE;
print 'The estimated error variance is:  '  SigE2_hat;

SumSum = k*((n**2)/NN);
n0 = (NN - SumSum)/(k-1);
SigR2_hat = (MSB - MSE)/n0;
print 'The estimated random variables variance is:  '  SigR2_hat;

*Calculate the test statistic;
F = MSB/MSE;
print 'The F test statistic is:  '  F;

F_crit = Finv(0.05,NN-k,k-1);
print 'The F critical value is:  '  F_crit;
```

```
   if F > F_crit then print 'The Null Hypothesis will be rejected';
   else print 'The Null Hypothesis will not be rejcted';
   quit;
```

Program 3 - Cholesterol intake example

EM apporach

```
*Diet mixed example - Using Em with ML;
   proc iml;
   use rr.diet;
   read all into Dat;

   Age = Dat[,1];
   chol = Dat[,2];
   Gender = Dat[,3];
   sub = Dat[,4];
   n = nrow(Chol);
   p = 2;
   X = J(n,1,1)||Age||Gender;
   Z = sub;
   Y = Chol;
   q = ncol(Z);

   *Startin values;
   Beta = J(ncol(X),1,1);
   SigE = 10;
   SigD = 150;
   count = 0;
   k = 0; *number of itterations;

   Do until (count = 1);

       V = Z*Z'*(SigD) +SigE*I(n);
       r = Y-X*beta;
       Beta_k = beta + SigE#(inv(X'*X)*X'*inv(V)*r);
       SigE_k = SigE +((SigE**2)/n)*
                  (r'*inv(v)*I(n)*I(n)'*inv(V)*r - trace(I(n)'*inv(v)*I(n)));
       SigD_k = SigD +((SigD**2)/n)*
                  (r'*inv(v)*Z*Z'*inv(V)*r - trace(Z'*inv(v)*Z));

       *Test for convergence;
       TestOld = Beta//SigE//SigD;
       TestNew = Beta_k//SigE_K//SigD_K;
       Test = abs(TestOld - TestNew);
       Sum = Test[+];
       if sum < 0.1 then count = 1;
       else count = 0;

       *reset values;
```

```
        beta = beta_k;
        SigE = SigE_k;
        SE = sqrt(SigE);
        SigD = SigD_k;
        SD = Sqrt(SigD);
        k = k+1;
        if k = 300 then count = 1;
    end;
    print k beta SigE SE SigD SD;
    random_effect = sigD*Z'*inv(V)*(Y-X*Beta);
    print random_effect;
    quit;
```

NR approach

```
*Diet mixed example - Using NR with ML;
    title 'Mixed Model Diet Example - ML method';
    proc mixed data = rr.diet method = ml;
    class subsidiary;
    model chol = age gender/ s;
    random subsidiary/s;
    run;
    title;
```

Program 4 - Mixed effects model

Model 1

```
proc mixed data = cleanerDataMix covtest method = ml;
    model Y = X1 X2 X3 X4 X5 X6 X7 X8/ solution outp=ind_predict
                outpm=mean_predict;
    random intercept X1 X2 X3 X4 X5 X6 X7 X8/ subject=resp
                type=cs solution G;
    ods output SolutionF=solModel SolutionR = REffects;
    run;
```

Model 2

```
proc mixed data = cleanerDataMix covtest method = ml;
    model Y = X1 X2 X3 X4 X5 X6 X7 X8/ solution outp=ind_predict
                outpm=mean_predict;
    random intercept X1 X2 X3 X4 X5 X6 X7 X8/ subject=resp
                type=vc solution G;
    ods output SolutionF=solModel SolutionR = REffects;
    run;
```

Model 3

```
proc mixed data = cleanerDataMix method = ml covtest ratio;
   model Y = X1 X2 X3 X4 X5 X6 X7 X8/ solution outp=ind_predict
                 outpm=mean_predict;
   random intercept X1 X2 X3 X4 X5 X6 X7 X8/ subject=resp
                 type=un solution G V ;
   ods output SolutionF=solModel SolutionR = REffects;
   run;
```

Program 5 - Bayesian linear regression estimation

PROC IML

```
resetline;
   proc iml;
   use rr.x;
   read all into X;
   use rr.y;
   read all into Y;

   beta = inv(X‘*X)*X‘*Y;
   print ’Estimate using OLS’ Beta;
   YhatOLS = X*Beta;

   *Stack the two Y’s;
   Do i = 1 to ncol(Y);
       betastack = betastack||beta[,i];
    Ystack = Ystack//Y[,i];
    Yhats = Yhats//YhatOLS[,i];
       Xstack = Xstack//X;
   end;

   nnn = nrow(Xstack);*number of obs;
   ppp = ncol(Xstack)-1;*number of variables;
   n = nrow(X); *Number of Profiles;
   p = ncol(X)-1; *Number of beta’s without intercept;
   m = ncol(Y); *Number of Responses;

   *Bayesian approach;

   *Initail values;
   Beta0 = J(nrow(beta), ncol(Beta), 1);
   var00 = J(1,ncol(Y),25);
       it = 0;*itteration counter;
       kk = 0;*draw counter;
       Allbeta = J(9,1,0); *J(nrow(Beta), ncol(beta),0);*reset matrix;
       allVar = 0; *J(nrow(Var00), ncol(var00),0);*reset matrix;
       YhatHBStack = J(nrow(y),1,0);
       count = 0; *R2 calculations;
```

```
    burn_in = 1000; *10000;
    ndraws = 1; *100;
Iteration = 10000;

Do j = 1 to Iteration; * itterations;
 Do resp = 1 to m;*for each respondent;
        Var0 =var00[1,resp];
        Sig0 = var0*inv(X'*X);

 *Update Beta;
 VarBeta = inv(inv(sig0)+(X'*X)/var0);
 MeanBeta = VarBeta*(inv(sig0)*(beta0)+(X'*y[,resp])/var0);
 MeanBeta = MeanBeta[,1];
 SampleBeta = (randnormal(1,MeanBeta,VarBeta))';

 *update Sig2;
 SSRBeta_New =(Y[,resp] - X*SampleBeta)'*
                          (Y[,resp] - X*SampleBeta);
 vo = n-p-1;
 k = (vo+n)/2;
 thetha = ((vo*var0+SSRBeta_new)/2);
 seed = 0;
 SampleSig = 1/(inv(thetha)*rangam(seed,k));

 *update values;
 var0 = samplesig;
 sig0 = var0*inv(X'*X);
 beta0 = sampleBeta;
AllVar = AllVar||Var0;
 AllBeta = AllBeta||Beta0;
 end;
*end of respondents loop;

*Reset values ;
Beta0 = AllBeta[,2:87];
Var00 = AllVar[,2:87];

* ------------------ Aggregate sample distribution-------------------;
AAA = mean(Beta0');
AAV = mean(Var00');
Agg_Sample = (randnormal(1,AAA,AAV))';
n_a = nrow(ystack);

SRR_agg =(Ystack - Xstack*Agg_sample)'*(Ystack - Xstack*Agg_sample);
 voo = n_a-p-1;
 kkk = (voo+n_a)/2;
 thetha_2 = ((voo*AAV+SRR_agg)/2);
 seed = 0;

 Agg_sample_var = 1/(inv(thetha_2)*rangam(seed,kkk));

it = it + 1;*counter for number of itterations;
```

```
   kk = kk + 1;

   do hh = burn_in to it by ndraws;
   if (it > 2)&(kk = hh) then StoreBeta =
                                  StoreBeta//Allbeta[,2:87];
   if (it > 2)&(kk = hh) then StoreVar =
                                  StoreVar//Allvar[,2:87];
   if (it > 2)&(kk = hh) then StoreAgg =
                                  StoreAgg//Agg_Sample;
   if (it > 2)&(kk = hh) then StoreAgg_var =
                                 StoreAgg_var//Agg_Sample_var;
   end;

   Allbeta = J(9,1,0);*J(nrow(Beta), ncol(beta),0);*reset matrix;
   allVar = 0;*J(nrow(Var00), ncol(var00),0);*reset matrix;
   end;
   *end of itteration loop;

   nameB = "BetaY1":"BetaY86";
   create BetaDist from StoreBeta[colname = nameB];
   append from StoreBeta;

   nameV = "Var1":"Var86";
   create Vardist from StoreVar[colname=nameV];
   append from storeVar;

   namea = {'Agg_Betas'};
   create AggBetaSample from StoreAgg[colname = namea];
   append from StoreAgg;

   nameb = {'Agg_Var'};
   create AggVarSample from StoreAgg_var[colname = nameb];
   append from StoreAgg_var;

   quit;
```

PROC MCMC

```
resetline;
   *MCMC for respondent1;
   proc iml;
   use rr.x;
   read all into X;
   X1 = X[,2];
   X2 = X[,3];
   X3 = X[,4];
   X4 = X[,5];
   X5 = X[,6];
   X6 = X[,7];
   X7 = X[,8];
```

```
X8 = X[,9];

use rr.y;
read all into YY;
Y = YY[,1];

resp1 = X1||X2||X3||X4||X5||X6||X7||X8||Y;
name1 = {'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7' 'X8' 'Y'};
create MCMC_Resp1 from resp1[colname = name1];
append from resp1;
quit;

title 'Simple Linear Regresion for respondent1';
ods graphics on;
proc mcmc data = MCMC_resp1 outpost=resp1_out ntu=1000
           nmc=10000 thin=1 seed=0;

    parms beta0 0 beta1 0 beta2 0 beta3 0 beta4 0 beta5
           0 beta6 0 beta7 0 beta8 0;
    parms sigma2 1;

    prior beta0 beta1 beta2 beta3 beta4 beta5 beta6
           beta7 beta8 ~normal(mean = 0, var = 1e6);
    prior sigma2 ~igamma(shape = 8/2, scale = 2/8);

    mu = beta0 + beta1*x1 + beta2*x2 + beta3*x3 +
           beta4*x4 + beta5*x5 + beta6*x6 + beta7*x7 + beta8*x8;

    model Y ~n(mu, var = sigma2);

    run;
    ods graphics off;

*MCMC for aggregate model;

proc iml;
use rr.x;
read all into X;
use rr.y;
read all into Y;
m = ncol(Y);

Do i = 1 to m; *Stack;
    Ystack = Ystack//Y[,i];
    Xstack = Xstack//X;
end;

X1 = Xstack[,2];
X2 = Xstack[,3];
X3 = Xstack[,4];
X4 = Xstack[,5];
X5 = Xstack[,6];
X6 = Xstack[,7];
```

```
X7 = Xstack[,8];
X8 = Xstack[,9];

Agg = X1||X2||X3||X4||X5||X6||X7||X8||Ystack;
name1 = {'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7' 'X8' 'Y'};
create MCMC_Agg from Agg[colname = name1];
append from Agg;
quit;

title 'Simple Linear Regresion for Aggregate model';
ods graphics on;
proc mcmc data = MCMC_agg outpost=agg_out nmc=50000
           thin=5 seed=246810;
    parms beta0 0 beta1 0 beta2 0 beta3 0 beta4 0 beta5 0
           beta6 0 beta7 0 beta8 0;
    parms sigma2 1;

    prior beta0 beta1 beta2 beta3 beta4 beta5 beta6 beta7
           beta8 ~normal(mean = 0, var = 1e6);
    prior sigma2 ~igamma(shape = 8/2, scale = 2/8);

    mu = beta0 + beta1*x1 + beta2*x2 + beta3*x3 + beta4*x4
           + beta5*x5 + beta6*x6 + beta7*x7 + beta8*x8;

    model Y ~n(mu, var = sigma2);
    run;
    ods graphics off;
```

Program 6 - Bayesian mixed effects model estimation

```
resetline;
   *MCMC Mixed model;
   proc iml;
   use rr.x;
   read all into X;
   use rr.y;
   read all into y;

   m = ncol(Y);

   Do i = 1 to m; *Stack;
       Ystack = Ystack//(J(nrow(Y),1,i)||Y[,i]);
       Xstack = Xstack//X;
   end;

   X1 = Xstack[,2];
   X2 = Xstack[,3];
   X3 = Xstack[,4];
   X4 = Xstack[,5];
   X5 = Xstack[,6];
   X6 = Xstack[,7];
```

```
X7 = Xstack[,8];
X8 = Xstack[,9];

mixed = X1||X2||X3||X4||X5||X6||X7||X8||Ystack;
name1 = {'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7' 'X8' 'Resp' 'Y'};
create MCMC_mixed from mixed[colname = name1];
append from mixed;
quit;

proc mcmc data = mcmc_mixed ntu = 5000 thin = 10
           nmc=30000 outpost=MixedPost seed=0
propcov = quanew Diag = (mcse ess);
 ods select Parameters REParameters PostSummaries;

*placeholders;
 array theta[9] w0 w1 w2 w3 w4 w5 w6 w7 w8;
 array Sig_c[9,9];

*initial values;
 array mu0[9] (0 0 0 0 0 0 0 0 0);
 array S[9,9] (10 0 0 0 0 0 0 0 0 0 1.3 0 0 0 0 0 0 0 0 0 5 0 0 0
0 0 0 0 0 0 4.3 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0.58 0 0 0 0
 0 0 0 0 0 0.36 0 0 0 0 0 0 0 0 0 0.36 0 0 0 0 0 0 0 0 0 0.4 );

  parms beta0 0 beta1 0 beta2 0 beta3 0 beta4 0 beta5 0 beta6 0
        beta7 0 beta8 0;*fixed parameters;
    parms sigma2 1;
    parms Sig_c {0.23 0 0 0 0 0 0 0 0 0 0.14 0 0 0 0 0 0 0 0 0 0.25
   0 0 0 0 0 0 0 0 0 0.79 0 0 0 0 0 0 0 0 0 0.26 0 0 0 0 0 0 0 0 0 0.14
   0 0 0 0 0 0 0 0 0 0.25 0 0 0 0 0 0 0 0 0 0.47 0 0 0 0 0 0 0 0 0 0.51};

    prior beta:  ~normal(mean = 0, var = 20);
    prior sigma2 ~igamma(shape = 10/2, scale = 2/10);

    hyperprior Sig_c ~iwish(18, S);

    random theta ~mvn(mu0, Sig_c) subject=resp;

    mu = beta0 + beta1*x1 + beta2*x2 + beta3*x3 + beta4*x4 +
  beta5*x5 + beta6*x6 + beta7*x7 + beta8*x8+ w0+ w1*x1
  + w2*x2 + w3*x3 + w4*x4 + w5*x5 + w6*x6 + w7*x7 + w8*x8;

model Y ~n(mu, var = sigma2);
run;
```

Bibliography

[Ajjan Associates (n.d.)] Ajjan Associates (no date) *Conjoint analysis specialists,* viewed 4 March 2014, from http://consulting.ajjan.com/conjoint.htm.

[Allison and Christakis, 1994] Allison, P.D. and Christakis, N.A. (1994).'Logit models for sets of ranked items', *Journal of Sociological Methodology*, **24**, 199-228.

[Bain and Engelhardt, 1992] Bain, L.J. and Engelhardt, M. (1992). 'Introduction to probability and mathematical statistics*',* 2nd edn., Duxbury Classic Series, California.

[Böckenholt, 2001] Böckenholt, U. (2001). 'Mixed-effects analysis of rank-ordered data', *Journal of Psychometrika*, **66**(1), 45-62.

[Chipperfield and Steel, 2010] Chipperfield, J.O. and Steel, D.G. (2010). '*Multivariate Random Effect Models with complete and incomplete data*', Centre for Statistical and Survey Methodology, University of Wolllongong, Working paper 16-10, 29.

[Christensen, 2011] Christensen, R. (2011). '*Bayesian Ideas and Data Analysis: An Intoduction for Scientists and Statisticians*', CRC Press.

[Corbeil and Searle, 1979] Corbeil, R.R. and Searle, S.R. (1979), 'Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model', *Journal of Technometrics*, **18**(1), 31-38.

[Cowels et al., n.d.] Cowles, K., Kass, R., and O'Hagan, T. (no date), *What is Bayesian Analysis?*, viewed 10 March 2014, form https://bayesian.org/Bayes-Explained.

[Cunningham et al., 2010] Cunningham, C.E., Deal, K., and Chen, Y. (2010). 'Adaptive Choice-Based Conjoint Analysis: A new patient-centered approach

to the assessment of health service preference', Review article, Patient 2010: **3**(4), 257-273

[Damaraju et al., 2011] Damaraju, R., Wiley, J.B., and Chitturi, P. (2011). 'Choice-Based Conjoint Analysis: Models and Designs', CRC Press.

[Dempster et al., 1977] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). 'Maximum Likelihood from Incomplete Data via the EM Algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1-38.

[DeSarbo, 2006] DeSarbo, W.S. (2006). 'The Handbook of Market Research: Uses, Misuses and Future Advances. Chapter 19 Latent Structure Regression'. SAGE Publications, Inc.

[Douglas, 1964] Douglas, C.J. (1969). '*Categorical Conjoint Measurement*', paper presented at the annual meeting of the association of Mathematical Psychology, Ann Arbor, Michigan.

[Draper and Smith, 1998] Draper, N.R and Smith, H. (1998). 'Applied Regression Analysis', 3rd edn., Wiley Series in Probability and Statistics.

[The Economist, 2000] The Economist (no date), *In Praise of Bayes*, viewed 12 November 2014, from https://www.economist.com/node/382968.

[Elmore et al., 1999] Elmore, J.R., Heymann, H., Johnson, J., and Hewett, J.E. (1999). 'Preference mapping: relating acceptance of "creaminess" to a descriptive sensory map of semi-solid', *Journal of Food Quality and Preference*. **10**, 465-475.

[Fichet et al., 2011] Fichet, B., Verde, R., Piccolo, D, Vichi, M. (2011). 'Classification and Multivariate Analysis of Complex Data Structures', Studies in Classification, Data Analysis, and Knowledge Organizations, Springer Verlag, Berlin Heidelberg.

[Fox, 2002] Fox, J. (2002). 'Linear Mixed Models', Appendix to R and S-Plus Companion to Applied Regression. http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-mixed-models.pdf

[Frühwirth-Schatter and Otter, 1999] Frühwirth-Schnatter, S. and Otter, T. (1999). 'Conjoint Analysis Using Mixed Effect

Models', Department of Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna. http://epub.wu.ac.at/806/1/document.pdf

[Goffin et al., 2010] Goffin, K., and Lemke, F. (2010). 'Identifying Hidden Needs: Creating Breakthrough Products', Palgrave MacMillan.

[Green et al., 2001] Green, P.E., Krieger, A.M., and Wind, Y. (2001). 'Thirty Years of Conjoint Analysis: Reflections and Prospects', *Journal of Marketing Engineering* **31**(3), S56-S73.

[Green and Schaffer, 1991] Green, P.E. and Schaffer, C.M. (1991). 'Importance Weight Effects On Self-explicated Preference Models: Some Empirical Findings' *Advances in Consumer Research*, Provo, UT, Association for Consumer Research, 476-482

[Green and Shrinivasan, 1978] Green, P.E. and Shrinivasan, V. (1978). 'Conjoint analysis in Consumer Research: Issues and Outlook', *Journal of Consumer Research*, **5**(2), 103-123.

[Grover and Vries, 2006] Grover, R. and Vries, M. (2006). 'The Handbook of Market Research: Uses, Misuses and Future Advances. Chapter 15 Conjoint Analysis: Understanding Consumer Desicion-Making', SAGE Publications, Inc.

[Guidici and Figini, 2009] Giudici, P. and Figini, S., (2009). 'Applied Data Mining for Business and Industry',2nd edn. , John Wiley and Sons, Ltd.

[Gustafsonn et al., 2007] Gustafsonn, A., Herrmann, A. and Huber, F. (2007). 'Conjoint Measurement: Methods and Applications', 4th edn., Springer Verlag.

[Hair et al., 2006] Hair, J.F., Black, B., Babin, B., Anderson, R.E., and Tatham, R.L. (2006). 'Multivariate Data Analysis', 6th edn., Pearson Prentice Hall.

[Halloand and Cravens, 1973] Holland, C.W. and Cravens, D.W. (1973). 'Fractional Factorial Experimental Design in Marketing Research', *Journal of Marketing Research*, **10**(3), 270-276.

[Hastie et al., 2008] Hastie, T., Tibshirani, R., and Friedman, J. (2008). 'The Elements of Statistical Learn-

ing: Data Mining, Inference, and Prediction', 2nd edn., Springer Verlag.

[Hedeker et al., 1991] Hedeker, D., Gibbons, R.D., and Davis, J.M. (1991). 'Random Regression Models for Multicenter Clinical Trails Data', *Psychopharmacology Bulletin.* **17**(1), 73-77.

[Hoff, 2009] Hoff, P.D. (2009). 'A First Course in Bayesian Statistical Methods', Springer Verlag.

[Isaacson and Lesnick, 2012] Isaacson, B. and Lesnick, D. (2012). 'MaxDiff vs Conjoint: Which is better to measure consumer preference*?'*, MMR Strategy Group, viewed 24 June 2013, from http://www.mmrstrategy.com/wp-content/uploads/2012/08/MaxDiff-vs-Conjoint-Which-is-Better-to-Measure-Consumer-Preferences.pdf

[Jensen et al., 2007] Jensen, W.A., Jeffrey, B.B., and Woodall, W.H. (2007). 'Profile Monitoring via Linear Mixed Models*'*, Virginia Polytechnic Institute and State University, Blacksburg. http://www.stat.vt.edu/research/Technical_Reports/TechReport06-2.pdf

[Jiang, 2010] Jiang, J. (2010). 'Large Sample Techniques for Statistics*'*, Springer Verlag.

[Johnson and Wichern, 2007] Johnson, R.A. and Wichern, D.W., (2007). 'Applied Multivariate Statistical Analysis', 6th edn., Pearson International Edition.

[Khattree and Naik, 2003] Khattree, R., and Naik, D.N. (2003). Applied Multivariate Statistics with SAS Software, 2nd. edn., SAS Institute Inc., Cary, NC, USA.

[Kleinbaum et al., 2008] Kleinbaum, D.G., Lawrence, L.K., Nizam, A., and Muller, K.E. (2008). 'Applied Regression Analysis and other Multivariate Methods', Thomson Learning, Inc.

[Krantz, 1964] Krantz, D. H. (1964). 'Conjoint Measurement: The Luce-Tukey Axiomatization and Some Extensions', *Journal of Mathematical Psychology*, **1**, 248-77.

[Kruskal, 1965] Kruskal, J. B. (1965) 'Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data', *Journal of the Royal Statistical Society*, Series B, **27**, 251-63.

[Kufheld, 2005] Kufheld, W.F. (2010). 'Conjoint Analysis'. SAS Institute Inc., SAS Technical Report R-109, Conjoint Analysis Examples, Cary, NC: SAS Institute Inc., 1993.85.

[Kutner et al., 2005] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005) 'Applied Linear Statistical Models', 5th edn.,McGraw Hill International Edition.

[Lasaffre and Lawson, 2012] Lasaffre, E. and Lawson, A.B. (2012). '*Bayesian Biostatistics*', John Wiley and Sons, Ltd.

[Lee, 2007] Lee, S. (2007). 'Structural equation modeling: A Bayesian apporach', Wiley series of prabability and statistics.

[Lindstrom and Douglas, 1988] Lindstrom, M.J. and Douglas, M.B. (1988). 'Newton-Rhapson and EM Algorithms for Linear Mixed-effect Models for Repeated-Measures Data', *Journal of the American Statistical Association*, **83**(404), 1014-1022.

[Louviere et al., 2000] Louviere, J.J., Hensher, D.A., Swait, J.D., and Adamonice, W. (2000). 'Stated Choice Methods: Analysis and Application', Cambridge University Press.

[Madsen and Thyregod, 2011] Madsen, H. and Thyregod, P. (2011). 'Introduction to General and Generalized Linear Models', CRC Press, Taylor and Francis Group.

[Mayden-Olivares and Böckenholt, 2009] Maydeu-Olivares, A. and Böckenholt, U. (2009). 'Modeling Preference Data, Chapter 12' The SAGE Handbook of Quantitative Methods in Psychology. SAGE Publications Ltd.

[McCulloch et al., 2008] McCulloch, C.E., Searle, S.R. and Neuhaus, J.M. (2008). 'Generalized, Linear, and Mixed Models', 2nd edn., Wiley.

[Meng and Van Dyk, 1997] Meng, X. and Van Dyk, D. (1997) 'The EM Algorithm – An Old Folk-Song Sung to a Fast New Tune', *Journal of the Royal Statistical Society. Series B (Methodological)*, **59**(3), 511-567.

[Neter et al., 1996] Neter, J., Kutner, M., Wasserman, W. and Nachtshelm, C. (1996).'Applied Linear Statistical Models', 4th edn., McGraw Hill/Irwin Series: Operations and Decision Sciences.

[Ng et al., 2012] Ng, S., Krishnan, T., and McLachlan, G.H. (2012).'Chapter 6: The EM algorithm. Handbook of Computational Statistics. Concepts and Methods', 2nd end., Springer Verlag.

[Ngo and Brand, n.d.] Ngo, L. and Brand, R., (no date), 'Model Selection in Linear Mixed Models Using SAS PROC MIXED', viewed 17 August 2013, from http://www2.sas.com/proceedings/sugi22/STATS/PAPER284.PDF

[Oded et al., 2008] Netzer, O., Toubia, O., Bradlow, E.T., Dahan, E., Evgeniou, T., Feinberg, F.M., Feit, E.M., Hui, S.K., Johnson, J., Liechty, J.C., Orlin, J.B., Rao, V.R (2008). Beyond conjoint analysis: Advances in preference measurements. Springer Science + Business Media, LLC 2008.

[Orme, 2010] Orme, B. (2010) 'Chapter 9: Interpreting the Results of Conjoint Analysis', Getting started with Conjoint Analysis: Strategies for Product Design and Pricing Research, 2nd edn., Madison, Wisconsin, Research Publishers LLC.

[Orme, 2000] Orme, B. (2000) 'Hierarchical Bayes: Why all the attentions', Research Paper Series. Sawtooth Software Inc.

[Owusu, 2013] Owusu, A. (2013). 'Influence of Price and Quality On Consumer Purchase of Mobile Phone in the Kumasi Metropolis in Ghana: A Comparative Study', *European Journal of Business and Management,* **5**(1).

[Pinheiro, 1994] Pinheiro, J.C. (1994). 'Topics in Mixed Effects Models', A PhD Thesis. University of Wisconsin, Madison..

[Pinheiro and Bates, 2004] Pinheiro, J.C. and Bates, D.M. (2004). 'Mixed-Effects Models in S and S-PLUS', Springer Verlag.

[Qualtrics, 2011] Qualtrics (2011). 'Conjoint Analysis: Explaining Full Profile and Self Explicated Approaches', viewed May 2013, from http://www.qualtrics.com/docs/ConjointOverview.pdf

[Roodman, 1974] Roodman, G.M. (1974). 'A Procedure for Optimal Stepwise MSAE Regression Analysis', *Journal of Operational Research*, **22**(0), 393-399.

[Searle et al., 2006] Searle, S.R., Casella, G., and McCulloch, C.E. (2006) 'Variance Components', Wiley Series in Probability and Statistics.

[Seefeld, 2007] Seefeld, K. (2007). 'Statistics using R with Biological Examples', Department of Mathematics and Statistics, University of New Hampshire, Durham, NH.

[Snijders, 2005] Snijders, T.A.B. (2005).'Fixed and Random Effects', B.S. Everitt and D.C. Howell (eds.). Encyclopedia of Statistics in Behavioral Science. **2**, 66-665. Chicester (etc.). Wiley.

[Tversky, 1967] Tversky, A. (1967).'A General Theory of Polynomial Conjoint Measurement', *Journal of Mathematical Pshychology*, **4**, 1-20.

[Wakefield, 2013] Wakefield, J. (2013).'*Bayesian and Frequentist Regression Methods*' Springer Series in Statistics. Springer Verlag.

[Wittink and Cattin, 1989] Wittink, D and Cattin, P. (1989). 'Commercial use of conjoint analysis: An update', *Journal of Marketing*, **53**(3), 91-96.

[Wittink et al., 1994] Wittink, D., Vriens, M., and Burhenne, W. (1994). 'Commercial use of conjoint in Europe: Results and critical reflections', *International Journal of Research in Marketing*, **11**(1), 41-52.

[Wu, 2010] Wu, L. (2010). 'Mixed Effect Models for Complex Data', CRC Press. Taylor and Francis Group.

[Yan and Su, 2009] Yan, X. and Su, X.G. (2009). 'Linear Regression Analysis: Theory and Computing', World Science Publishing Co. Pte. Ltd

[Young, 1969] Young, F. W. (1969). 'Polynomial Conjoint Analysis of Similarities: Definitions for a Special Algorithm', Research Paper No. 76, Psychometric Laboratory, University of North Carolina. http://files.eric.ed.gov/fulltext/ED072085.pdf