Université Lumière Lyon 2

Faculté des Sciences Economiques

Sources of Errors and Biases in Traffic Forecasts for Toll Road Concessions

Thèse pour le Doctorat ès Sciences Economiques

Mention Economie des Transports

Antonio NUNEZ

dirigée par M. le Professeur Alain BONNAFOUS

Présentée et soutenue publiquement le
5 décembre 2007.

Membres du Jury:

M. Alain BONNAFOUS	Pr. à l'IEP de Lyon	*Directeur*
M. Yves CROZET	Pr. à l'Université Lyon 2	
M. Jean DELONS	Chargé de Mission à Cofiroute	
M. Fabien LEURENT	Pr. à l'ENPC	
M. Werner ROTHENGATTER	Pr. à l'Université de Karlsruhe	*Rapporteur*
M. Stéphane SAUSSIER	Pr. à l'Université de Paris 11	*Rapporteur*

Contents

Acknowledgements 11

Abstract 13

Résumé 19

Introduction 27

Plan of the Manuscript 29

1 Errors and Biases in Transport Demand Forecasts 33

1.1 What is Forecasting? 35

1.2 Forecasting in Transport 36

1.2.1 The Classic 4-step Model 38

1.3 Errors in Traffic Forecasts 39

1.4 Sources of Errors 41

1.4.1 Uncertainty About the Future 42

1.4.2 Methodology, Assumptions and Data 43

1.4.3 Behavioural Sources 46

1.4.4 The Particular Case of Road Concessions 49

1.5 Objectives of this Research 51

2 Transport Forecasters' Behaviour and Overconfidence 55

2.1 Introduction 56

2.2 Who Forecasts Transport Demand? 57

2.3 The Latest Forecast 61

2.4 Models 63

2.5 Forecast Errors 66

2.5.1 Sources of Errors 67

2.6 Forecast's Environment 68
2.7 Overconfidence in Transport Forecasts 72
2.8 Econometric Analysis of Biases 76
2.9 Comments Uncommented 77
2.10 Conclusions 78

3 Winner's Curse in Toll Road Concessions 83
3.1 Introduction 85
3.2 Auctions for Toll Road Concessions 90
3.2.1 First-Price, Sealed-Bid Auctions 90
3.2.2 Common Value Auctions 90
3.2.3 Auctions with Differing Levels of Common Uncertainty . 92
3.2.4 Renegotiation in Toll Road Concessions 94
3.3 Bidding for Toll Road Concessions: A Simple Model 96
3.3.1 Model Framework 96
3.3.2 Model Setting 97
3.3.3 Number of Bidders and Traffic Forecast Deviation 99
3.3.4 Number of Bidders and Level of Common Uncertainty . 100
3.3.5 Number of Bidders and Renegotiation 101
3.4 Data on Road Concession Contract Auctions 103
3.4.1 Dependent Variable: Traffic Forecast Deviation 103
3.4.2 Explanatory Variables 104
3.5 Econometric Results 107
3.6 Robustness Analysis 109
3.7 Conclusions 111

4 Decreasing Long-Term Traffic Growth 119
4.1 Introduction 121
4.2 Traffic Growth 122
4.3 Why does Traffic Grow Decreasingly? 124
4.4 Econometric Issues 127
4.4.1 Partial Adjustment 127
4.4.2 Integrated variables, Cointegration and Error-Cor-rection 128
4.5 Data and Estimation 131
4.6 Evidences of Decreasing Growth 131

4.6.1 Cross-section Time Series Analysis 132
4.6.2 Testing for Parameter Stability 134
4.6.3 Moving Regressions 136
4.7 A Functional Form for Decreasing Elasticity 137
4.7.1 Impact on Long-Term Forecasts 140
4.8 Conclusions 142

5 Estimating the Value of Travel Time Savings 147

5.1 Introduction 148
5.2 The Value of Time in Transport 151
5.2.1 VTTS in Freight Transport 154
5.3 Discrete Choice Models 155
5.3.1 The Multinomial Logit 155
5.3.2 The Mixed Logit Model 160
5.4 Bayesian Procedures 162
5.4.1 Overview of Bayesian Concepts 163
5.4.2 Drawing from the Posterior 165
5.4.3 Posterior Mean as a Classical Estimator 169
5.4.4 Posteriors for the Mean and Variance 170
5.4.5 Hierarchical Bayes for Mixed Logit 174
5.5 Challenges in Estimating VTTS 178
5.5.1 Identifying Preference Heterogeneity 178
5.5.2 Selecting Random Parameters 179
5.5.3 Selecting the Distributions of the Random Parameters . 180
5.5.4 Revealed Preference Data 182
5.5.5 Optimization Problems 182
5.5.6 Imposing Constraints 182
5.5.7 Priors 183
5.5.8 Advantages and Problems of Bayesian Procedures 183
5.5.9 The Role of the Alternative Specific Constant 184
5.6 The Survey 185
5.7 Econometric Results 188
5.7.1 Maximum Likelihood estimations 188
5.7.2 Bayesian Estimations 189
5.8 Discussion 192

5.9 Conclusions 195

General Conclusions and Policy Implications 197

A Forecasters' survey questions 201

B Distributions of variables in chapter 3 205

C VTTS survey form 211

List of Figures

1 Ecarts (réel/prévu) 22

2 Distribution de la valeur du temps PL. 25

1.1 Caricature of weather forecasts 36

1.2 Errors on Flyvbjerg et al (2006) sample 40

1.3 Errors variation over time on Flyvbjerg et al. (2003) sample . . 40

1.4 Errors on Standards and Poor's (2005) sample 41

1.5 Forecasting error in 49 road concessions (chapter 3 sample) . . . 41

1.6 From "be" forecast to "do" forecast 45

2.1 In which country do you work?(N=178) 57

2.2 Location of the projects.(N=178) 58

2.3 Degree.(N=178) 58

2.4 Post-grad degree. (N=178) 59

2.5 Sectors forecasters work in.(N=178) 59

2.6 Gerder distribution.(N=178) 60

2.7 Distributions of respondents' age. (N=178) 60

2.8 Number of forecasts.(N=178) 61

2.9 When did you prepare your latest forecast? (N=172). 61

2.10 Has the project been launched?(N=176) 62

2.11 Modes in the last forecast.(N=176) 62

2.12 Financing.(172) 63

2.13 Operation. (N=167) 63

2.14 Constant x Distributed VTTS. (N=153) 64

2.15 initial *versus* growth in demand forecasts. (N=162) 64

2.16 Aggregated or disaggregated modal share.(N=156) 65

2.17 Models forecasters apply. (N=170) 65

2.18 Stated error in the latest forecast.(N=88) 66

2.19 Perception of own's quality of results. (N=147) 67
2.20 Average distribution of under/overestimation.(N=150) 67
2.21 Forecasters under pressure. (N=168) 69
2.22 Would they produce better forecasts without pressure? (N=167) 69
2.23 Role of strategic manipulation.(N=155) 70
2.24 Sense of strategic manipulation.(N=134) 70
2.25 Influence of the technical study on the decision. (N=158) 71
2.26 Knowledge of the minimun demand level. (N=161) 72
2.27 Distributions of forecast errors. 75
2.28 Self-evaluation of competence level.(N=155) 76
2.29 Distributions of self-evaluations. 77

3.1 Length and Forecast Error. 94
3.2 TDF. 104
3.3 Number of Bidders. 105

4.1 From preferences to elasticity. 126
4.2 Traffic on the A10 motorway. 132
4.3 Traffic on the A11 motorway. 132
4.4 LTM long-run elasticities. 133
4.5 PAM long-run elasticities. 133
4.6 ECM long-run elasticities. 134
4.7 PAM short-run elasticities. 134
4.8 ECM short-run elasticities. 135
4.9 Comparing elasticities. 138
4.10 k versus traffic. 139
4.11 γ versus traffic. 139
4.12 A hypothetical example. 141
4.13 Application on the A11 motorway. 141

5.1 Comparison of VTTS distributions. 152
5.2 Survey's Location. 186
5.3 VTTS Distribution for empty and own account by ML 191
5.4 VTTS Distribution for loaded and hire by HB. 192
5.5 VTTS Distribution for empty and own account by HB. 192
5.6 VTTS Distribution for average load and hire dummies by HB. . 193

A.1 Questions in the survey of forecaster's behaviour. 203

B.1 TDF. 207
B.2 Number of Bidders. 207
B.3 Length. 208
B.4 Civil Law. 208
B.5 HIC. 208
B.6 Public Information. 209
B.7 Government Learning. 209

C.1 VTTS survey form 213

List of Tables

1.1 Transport Modelling 53

2.1 Sources of errors. 80
2.2 Comparing ex-post and revealed errors 81
2.3 Comparing drivers and forecasters skilful 81
2.4 Impact of the main characteristics on self-evaluation. 82

3.1 Toll Road Concessions by Country and by Year 114
3.2 Data Definitions and Descriptive Statistics 115
3.3 Econometric results 116
3.4 Econometric results - extended 117

4.1 ADF test - exogenous variables 129
4.2 ADF test - traffic 143
4.3 Summary of descriptive statistics 144
4.4 CUSUM of squares test 145
4.5 Subsamples Elasticities 146

5.1 Sample and traffic count data 187
5.2 Final Sample 187
5.3 Summary of descriptive statistics 188
5.4 Econometric results 194

Chapter 5

Estimating the Value of Travel Time Savings

-Application to the Freight Transport in France-

Abstract

In this study we apply the Logit, the Mixed Logit and the Bayesian Mixed Logit models to estimate the value of time in freight transport in France. We discuss the importance of the value of time and its particular role in the case of private motorways. We present the econometric models currently used to estimate it, giving a special attention to the Bayesian procedures, since it is a relatively new method with only a few results in the literature. We also discuss the main challenges in estimating the value of travel time savings. We then describe the revealed preference survey we realized, including 1027 vehicles in order to study the trade-off between the free road and the tolled motorway. Results show that the Bayesian procedures represent an interesting alternative to the optimization problems the maximum likelihood faces. Also, in line with recent works, we find that using a constant value, representative of an average, can lead to traffic overestimation. Finally, we found average values around €45 per vehicle and per hour, suggesting that the current French standard value should be reviewed upwards.

5.1 Introduction

The value of travel time savings (VTTS) is at the heart of transport projects and transport policies evaluation. It plays a central role in the socio-economic evaluation since time savings usually are the dominant factor in the users' benefit. Despite the importance of the freight transport in the economy, its representativeness in terms of volume of traffic and its contribution for the socio-economic benefits of a new motorway, relatively few studies, in France or abroad, are devoted to the study of the value of time in freight transport.

In order to estimate the welfare produced by the time saving generated by a new infrastructure, econometric models were developed to estimate the value of time. These models are mainly based on discrete choices evaluation of the trade-off between time and money. In models of choice among discrete alternatives, the assumption is made that individual choices are based on perceptions of the relative characteristics of the alternative options; in this way, implicit equivalences are subjectively established. This subjective value of time has concentrated the attention of researchers and policymakers within the industrialized countries. Given this importance, one would like to achieve estimates of subjective value of time that are robust and ideally independent from the functional form of the models used to estimate them (Gaudry et al., 1989).

With the introduction of private finance (and tolling) in transport, willingness to pay is applied to estimate actual out-of-pocket money and then the optimal toll levels and the financial profitability of a project. So, in recent years, an increasingly important application of discrete choice models has been to calculate the potential revenue for tolled roads, and networks with user charges, which offer higher speeds at a higher price. Here the important issue is not the hypothetical willingness to pay, but the actual money that will be handed over. It changes focus from hypothetical to bankable value of time (Hensher and Goodwin, 2004).

In this context, one of the main issues regarding the value of time is its distribution over the population. Heterogeneity in population comes from tastes, revenue, journey characteristics, distance and purpose. In freight transport, it will depend also on the firm's market and financial structures, on the characteristics of the goods, own account or hire transport, among other factors.

While in project evaluation the VTTS is usually taken as constant, for equity reasons (but this practice varies according the current national recommendation, and this social value usually differ from those issued from econometric estimations, representing a more “social” value of time), in revenue forecasts, and so for toll setting, the assumption of a constant VTTS may be very restrictive and lead to significant forecast errors. In fact, if an average value, virtually representative of a symmetric distribution, is taken as representative of a skewed distribution, there will be tendency to overestimate revenue. As a consequence, the value of time represents a main source of uncertainty. Moreover, in the VTTS modelling process, data quality, model structure and statistical or behavioural hypothesis play together; in this way VTTS may be used as a strategic variable, allowing to “adjust” the traffic and revenue levels.

Logit is by far the most applied model in discrete choice analysis. The logit model derives from the random utility model, which separates the total utility into deterministic and random components, under the assumptions of independent and identically distributed Gumbel disturbances. Its popularity is due to the fact that the formula of choice probabilities takes a closed form and is readily interpretable with good results related in literature[1]. In this model, heterogeneity, unobserved attributes and measurement errors are captured by the random disturbance and the coefficients of the utility function are fixed, leading to a constant value of time, representative of a virtual average individual.

Advances in simulated estimation techniques have enabled analysts to use increasingly complex models that allow one to define broader behavioural patterns, overshadowing the classic Multinomial Logit (Train, 2003). In the random coefficient random utility model, both coefficients and error term are represented by some PDF (Probability Distribution Function), this model is usually called Mixed Logit (ML) because it can be viewed as a logit with mixtures. ML is a high flexible model than can approximate any random utility model, and it is considered the most promising discrete choice model currently available (Hensher and Goodwin, 2004); it has been known for many years but has only become applicable with the development of simulation techniques. This model do not presents the restrictive properties of logit and allows for a

[1] probably accompanied by less good ones, less released

different PDF for each parameter, but results are also sensitive to the specification of the PDF shape. However, in practice, many difficulties challenge the application of this model, as the choice of the distribution, the starting values and convergence problems in maximum simulated likelihood.

Furthermore, the introduction of prior knowledge is intrinsic even to the classic analysis. First, the analyst usually has some priors about the result (i.e. one should expect that the value of travel time to be positive and to lay within a reasonable set) and second, the set of hypothesis and parameters need to the estimation of mixed logit models like the form of the distributions, eventual constraints and the starting values indirectly represent prior hypothesis.

Bayesian estimations have some strong advantages compared to the classical techniques; they allow for distributed coefficients but the estimation does not require any maximization, rather, draws from the posterior are taken until convergence is achieved, avoiding convergence problems and sample sizes necessary to achieve the convergence are substantially smaller. Moreover, they can properly integrate a priori knowledge on the parameters.

In order to determine the value of time in freight transport in France, an important but misunderstood parameter in project evaluation, and study the impact of model specification a revealed preference survey was conducted, interviewing 1027 truck drivers about their origin, destination and freight characteristics. The survey was conducted in four points; in two tolled motorways and their respective free parallel roads in the north-west of France. This configuration allows to the analysis of the trade off between rapid and more expensive links, and slower free roads.

In this chapter we discuss a number of issues related to the estimation and the interpretation of results in practical estimations of the value of time in transport (i) we analyse the role of model specification in the VTTS estimation, (ii) we identify sources of systematic and random taste variations; (iii) we propose a comparison of the different methods without using relevant prior information; (iv) we measure the benefit of integrating a prior distribution of VTTS and finally (v) we provide a robust estimation of the value of travel time for the freight transport in France. Results show that Bayesian estimations based on a prior knowledge leads to more sound and robust results; furthermore we find that values used currently in France should be reviewed upwards.

The contributions of this study are twofold. First, at the theoretical level, we discuss the importance of estimating distributed value of time in evaluating the willingness to pay for toll roads and show the impact of model structure on the evaluation of the real willingness to pay. Second, at the practical level, we estimate the value of time in freight transport in France and show the sensibility of estimations with respect to the model.

The rest of the chapter is organized as follows. Section 2 briefly discusses the notion and the importance of the value of travel time as well as the scarcity of empirical results in freight transport. Section 3 presents the most used econometric models applied to the VTTS estimation. Section 4 presents the Bayesian procedures and its application to estimate discrete choice models. Section 5 discuss some challenges in estimating the value of travel time savings. Section 6 presents the survey conducted for this study. Section 7 presents the econometric results and compares the different models. Section 8 discusses the results and section 9 concludes the chapter.

5.2 The Value of Time in Transport

The willingness to pay for a unit change in a certain attribute can be computed as the marginal rate of substitution (MRS) between income and the quantity expressed by the attribute, at constant utility levels (Gaudry et al., 1989). The concept is equivalent to computing the compensated variation (Small and Rosen, 1981), as one usually works with linear approximation of the indirect utility function. Thus, the point estimates of the MRS represent the slope of the utility function for the range where this approximation holds. Furthermore, as income does not enter in the truncated indirect utility function, the MRS is calculated with respect to minus the cost variable (Jara-Diaz, 1990). In this way, the WTP in a linear utility function simply equals the ration between the variable of interest and the cost variable. The willingness to pay to save time is usually called the value of time, or, related to the travel time, the value of travel time savings, VTTS.

The value of travel time is certainly the most important number in transport economics. Time savings use to account for the main part of the socio-economic benefit of a new infrastructure. Moreover, it allows the estimation of the market share of a new infrastructure or service and the estimation of the optimal pricing.

The distribution of the VTTS over the population is a fundamental issue. We can classify heterogeneity in the population in two groups, systematic and random. Systematic variations depend on socio-economic and trip specific characteristics. They are estimated either by segmenting the population of by interacting variables. This heterogeneity left is due to factors which can not be observed or are difficult to measure. In these cases, this heterogeneity can take form of a random parameter.

The proportion of a population who will choose to pay a toll t is given by the proportion whose value of the time saved is greater than the toll. The analyst, according to taste, convenience and internal evidence, will select among a number of appropriate analytical distributions in order to find a satisfactory representation of the "true" empirical distribution. The number of people whose value of time savings exceeds the toll charged, who will therefore pay it is then the integral, from toll price to infinite, of that distribution. This is then the measure of revenue to be received by the charging agency. In the case of a symmetric distribution, e.g. normal, in general representing the distribution by its mean will be able to produce the correct revenue. In the case of a substantially skewed distribution (e.g. lognormal) the average will not be in the centre of the distribution, and there will be fewer people in the population actually ready to pay the toll. In this situation revenue will be overestimated for low toll levels.

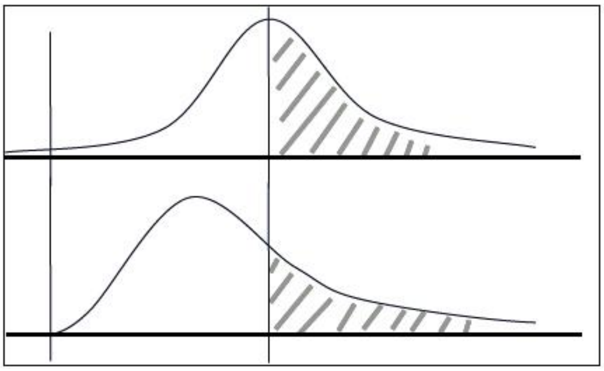

Figure 5.1: Comparison of VTTS distributions.

Hensher and Goodwin (2004) argue that financial institutions have two interests in their negotiations with public agencies on a public-private partnership. First, there is an interest in the best and most reliable possible estimate of the expected revenue. Second, there is interest in figures that strengthen their bargaining position in relation to the case for the scheme to go ahead at all, and on what basis of risk apportionment.

Consider the case where there is a well-established convention, used by the public agency for many years, to represent the distribution of VTTS by the average, partly for reasons of adequacy for purpose in previous applications, and partly because the models and consultants available find it convenient to do so. Then estimates made using the average, other things being equal, will tend to overestimate the revenue. In this case, the financial agency has the choice to go along with the standard procedure, or to "rock the boat" by suggesting using a distribution. The effect of doing so many well put the whole project at risk. So the perceived best interests of the agency are served by accepting the standard procedure, which strengthens the case for the project, but suspecting that it overestimates the revenue, finding a risk-sharing agreement, explicit or implicit, which cushions them against the likely result.

Conversely, the public agency's perceived best interests are served by using the standard practice, since this will increase the probability of raising the funding, anticipating that the public benefits in terms (for example) of congestion and pollution relief will be higher than calculated, and seek to ensure that the risk will be wholly born by the funders.

The paradoxical case is that each will be better served by using the distributions themselves, for internal, confidential reasons, but using the average (or preferably the median) value for public discussion, and hoping that the other party believes. But it is not a long-term solution, since it is almost bound to lead to later disputes, attempts to renegotiate, or collapse of confidence in such deals. There are signs that this can happen. The dilemma is obvious – will the financial advisers prefer to go with an overestimate to secure patronage and the contract (in a bid setting) knowing the likelihood (from previous contractual arrangements) that the risk can be transferred to government, or act as good corporate citizens and promote the more appropriate VTTS across the distribution.

In practice, this question is either ignored, or not expressed in this language (though accepting the underlying significance). The great majority of patronage studies around the world use simple averages for VTTS, so this provides an almost unquestioned benchmark as an always available fallback position, and a handy defensive (but not necessary defensible) instrument.

In this sense, distribution of the value of time in the population represents a number of issues including the choice of behavioural models and estimation procedures as well as the interpretation results will be subject to.

5.2.1 VTTS in Freight Transport

While for passengers transport there is a large literature and an important scientific activity on this topic, for freight transport both scientific and professional studies are very scarce. This little attention given to freight transport is mainly due to the information scarcity in the sector, where the competitiveness is very strong and information on costs play a strategic role. Furthermore, the logistic chain is very complex and has multiple decision takers. In passenger transport the decision maker is the passenger himself; but goods cannot decide, as notes DeJong (1996).

Ortuzar and Willumsen (2001) point out four reasons for the little research in freight transport modelling compared to passenger modelling:

- There are many aspects of freight demand that are more difficult to model than passenger movements.
- For some time urban congestion has been highest in the political agenda of most industrialised countries and in this field passenger play a more important role than freight.
- The movement of freight involves more actors than the movement of passengers; we have the industrial firm or firms sending and receiving the goods, the shippers organising the consignment and modes, the carrier(s) undertaking the movement and several others running transhipment, storage and custom facilities. In some cases two or more of these may coincide, for example in own-account operators, bur there is always scope for conflicting objectives which are difficult to model in detail in practice.

- Recent trends in freight research have emphasised the role it plays in the overall production process, inventory control and management of stocks. These trends are a departure from more traditional passenger modelling techniques and share little in common (Regan and Garrido, 2002).

The value of time of transport is defined as the marginal rate of substitution between travel time and travel cost. While in passenger's transport it comes from the Lagrange multiplier associated to the time constraint in the individual utility maximization, in freight transport time savings enter the financial optimization as they allow to reduce other costs like labour and capital costs and improve productivity.

In France, few studies were devoted to the empirical estimation of the value of time in freight transport; the main studies were realized by Fei Jiang (Jiang, 1998) who utilises revealed preference and Laura Wynter (Wynter, 1994), applying revealed and stated preference of shippers, by phone surveys, both studies in the context of their respective doctoral thesis. Their results range from 27 to 74 €/hour. Massiani (2005, pp.151-155) presents a review of the estimations of the value of time for freight transport in Europe found in literature. Governmental recommendation for the value of time for freight transport in France is 30 €(2000)/hour (Commissariat Général du Plan, 2001).

5.3 Discrete Choice Models

5.3.1 The Multinomial Logit

The most common theoretical base for generating discrete choice models is the random utility theory (Domencich and McFadden, 1975; Williams, 1977)[2]. In random utility models (RUM) the utility that the decision maker n obtains from alternative j is defined by

$$U_{nj} = V_{nj} + \varepsilon_{nj} \tag{5.1}$$

[2] For the hypothesis underlying the model see also Ortuzar and Willumsen (2001) and Ben-Akiva and Lerman (1994)

where U_{nj} is a non-stochastic utility function (called *systematic* or *representative* component of the utility) and ε_{nj} is a random component (or disturbance) which captures the factors that affect utility but the researcher does not or can not observe. The deterministic part is usually assumed to be linear, so that

$$V_{nj} = \beta' x_{nj}$$

.

The individual selects the maximum-utility alternative so that user n chooses alternative i if and only if

$$U_{ni} \geq U_{nj} \ \forall j \neq i$$

From this perspective, the choice probability of alternative i is equal to the probability that the utility of alternative i is greater than or equal to the utilities of all other alternatives in the choice set. This can be written as

$$P_{ni} = Prob(U_{ni} \geq U_{nj} \ \forall j \neq i)$$

Using the random utility model in expression (5.1), this can be rewritten as

$$P_{ni} = Prob(V_{ni} + \varepsilon_{ni} \geq V_{nj} + \varepsilon_{nj} \ \forall j \neq i)$$

To derive a specific random utility model, we require an assumption about the joint probability distribution of the full set of disturbances $\varepsilon_{nj}, \forall j$. The issues therefore are what distribution is assumed for each model, and what is the motivation for these different assumptions.

The logit model is derived under the assumptions of independent and identically distributed Gumbel (IID) disturbances, which means that the unobserved factors are uncorrelated over alternatives and have the same variance for all alternatives. The density for each unobserved component of utility is

$$f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}} \tag{5.2}$$

and the cumulative distribution is

$$F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}} \tag{5.3}$$

The variance of this distribution is $\pi^2/6$. By assuming the variance is $\pi^2/6$ we are implicitly normalizing the scale of the utility. If ε_{ni} and ε_{nj} are independent and identically Gumbel (or type I extreme value) distributed, then $\varepsilon_n = \varepsilon_{nj} - \varepsilon_{ni}$ is logistically distributed

$$F(\varepsilon_n) = \frac{e^{\varepsilon_n}}{1 + e^{\varepsilon_n}}$$

If ε_{ni} is considered given, the choice probability is the cumulative distribution for each ε_{nj} evaluated at $\varepsilon_{ni} + V_{ni} - V_{nj}$, which, according to (5.3) is $exp(-exp(-(\varepsilon_{ni} + V_{ni} - V_{nj})))$. Since the ε's are independent, this cumulative distribution over all $j \neq i$ is the product of the individual cumulative distributions:

$$P_{ni}|\varepsilon_{ni} = \prod_{j \neq i} e^{-e^{-(\varepsilon_{ni}+V_{ni}-V_{nj})}}$$

Of course, ε_{ni} is not given, and so the choice probability is the integral of $P_{ni}|\varepsilon_{ni}$ over all values of ε_{ni} weighted by its density (5.2):

$$P_{ni} = \int (\prod_{j \neq i} e^{-e^{-(\varepsilon_{ni}+V_{ni}-V_{nj})}}) e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{nj}}} \tag{5.4}$$

Some algebraic manipulation of this integral (Domencich and McFadden, 1975) results in a succinct, closed-form expression:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \tag{5.5}$$

which is the logit choice probability.

Limitations of Logit

In addition to the well know property of independence of irrelevant alternatives (IIA) and it's inability to deal with correlated choices over time (panel data), the constant parameters represents a restrictive assumption. If one or more characteristics (parameters) vary randomly across the population, the assumptions of the standard logit calibration are not satisfied, and the error term is no longer distributed independently of the explanatory variables. Thus the coefficients estimates from the calibration will be biased.

Another problem due to heterogeneity arises because the estimation produces estimates of time and cost parameters that are averages over the sample, and they are then used in ratio form to give the value of time (Fowkes and Wardman, 1988). The true value of time would be the average over the sample of individuals' value of time, these values being the ratio of their individual time and cost coefficients. It is easy to demonstrate that the ratio of the means and the means of the ratio are not necessary equal (unless the denominator is constant or the ratios are constants).

Moreover, as the parameters for time and cost are estimates from the model, they are not really constants but random variables with a certain probability density function (PDF). For this reason the value of time (calculated as the ratio between the time and cost parameters in a linear in parameters model) is also a random variable with an unknown PDF. We know the maximum likelihood parameters are asymptotically distributed multivariate Normal. Consequently the VTTS point estimate is a random variable governed by an unknown PDF, the probability function for the ratio between two Normally distributed variables is unknown a priori); only some things are known is special cases. For example, the ratio between two independently distributed standard Normal variables follows a Cauchy PDF (Arnold and Brockett, 1992), but this is unstable since it has an infinite variance and its mean does not have an analytical expression.

However, some econometric methods were developed in order to estimate confidential intervals for the value of time calculated as the ratio between the time and cost parameters, say β_t and β_c. The most applied is the asymptotic t-test.

The asymptotic t-test is generally used to prove if a normally distributed parameter is significant different from zero. Ben-Akiva and Lerman (1994) present an extension of this test for a linear combination of the parameters. As β_t and β_c are asymptotically distributed normal, the following null hypothesis can be postulated:

$$H_0 : \beta_t - VT\beta_c = 0,$$

where VT represents the value of time point estimate. The confidence interval is given by the set of VT values for which it is not possible to reject H_0 at a given level of significance. The corresponding test statistic is (Armstrong et al., 2001):

$$t = \frac{\beta_t - VT\beta_c}{\sqrt{Var(\beta_t - VT\beta_c)}}$$

This expression distributes normal for linear models and asymptotically normal for non-linear models like the MNL (see Ben-Akiva and Lerman (1994)). Armstrong et al. (2001) also derive the upper and lower bounds for the interval as follows:

$$V_{S,I} = (\frac{\beta_t}{\beta_c}\frac{t_c}{t_t})\frac{(t_t t_c - \rho t^2)}{(t_c^2 - t^2)} \pm (\frac{\beta_t}{\beta_c}\frac{t_c}{t_t})\frac{\sqrt{(\rho t^2 - t_t t_c)^2 - (t_t^2 - t^2)(t_c^2 - t^2)}}{(t_c^2 - t^2)} \quad (5.6)$$

where t_t and t_c correspond to the t-statistic for β_t and β_c, respectively; t is the critical value of t given the degree of confidence required and sample size and ρ is the coefficient of correlation between both parameter estimates. Expression (5.6) is a real number only if the radical argument is non-negative; it can be shown that this condition is met when the parameters β_t and β_c are statistically significant (so that t_t and t_c are greater than t). This condition assures positive upper and lower bounds.

It can be observed that the confidence interval derived from this formulation is not symmetrical with respect to the VT point estimate (β_t/β_c), and that the interval's mid-point is greater than β_t/β_c as well. Another feature is that the value of ρ has a strong influence on the size of the interval. In fact, the bigger the value of ρ the narrower the interval and vice versa, all other things being equal. In addition, the more significant the t-statistics are, the narrower the intervals (for details, Armstrong et al. (2001)).

5.3.2 The Mixed Logit Model

The specification of the random coefficients logit model (or mixed logit)[3] is the same as for the standard logit except that varies over decision makers rather than being fixed. As in the MNL the utility of person n from alternative j is specified as

$$U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj} \tag{5.7}$$

where x_{nj} are observed variables that relate to the alternative and decision maker, β_n is a vector of coefficients of these variables for person n representing that person's tastes, and ε_{nj} is a random term that is iid extreme value. The coefficients may vary over decision makers in the population with density $f(\beta)$. This density is a function of parameters θ that represent, for example, the mean and variance of the β's in the population.

The decision maker knows the value of his own β_n and ε_{nj}'s for all j and chooses the alternative i if and only if $U_{ni} \geq U_{nj} \forall j \neq i$. The researcher observes the x's but not β_n or the ε_{nj}'s. If the researcher observed β_n, then the choice probability would be standard logit, since the ε_{nj}'s are iid extreme value. That is, the probability *conditional* on β_n is

$$L_{ni}(\beta_n) = \frac{e^{\beta_n' x_{ni}}}{\sum_j e^{\beta_n' x_{ni}}}$$

However, the researcher does not know β_n and therefore can not condition on β. The unconditional choice probability is therefore the integral of $L_{ni}(\beta_n)$ over all possible variables of β_n.

$$P_{ni} = \int \frac{e^{\beta_n' x_{ni}}}{\sum_j e^{\beta_n' x_{ni}}} f(\beta) d\beta$$

[3] Random coefficients is the most widely used derivation of mixed logit models, but not the only one; each derivation provides a particular interpretation (Train, 2003).

which is the random coefficients probability.

The researcher specifies a distribution for the coefficients and estimates the parameters of that distribution.

McFadden (2000) show that any random utility model can be approximated to any degree of accuracy by a mixed logit with appropriated choice of variables and mixing distribution.

The researcher specifies the functional form $f(\cdot)$ and wants to estimate the parameters θ. The choice probabilities are

$$P_{ni} = \int L_{ni}(\beta) f(\beta|\theta) d\beta f(\beta) d\beta.$$

where

$$L_{ni}(\beta) = \frac{e^{\beta_n' x_{ni}}}{\sum_j e^{\beta_n' x_{ni}}}$$

The probabilities are approximated through simulation for any given value of θ:

(1) Draw a value of β from $(\beta|\theta)$, and label it β^1 with the superscript r=1 referring to the first draw.

(2) Calculate the logit formula $L_{ni}(\beta^r)$ with this draw.

(3) Repeat steps 1 and 2 many times, and average the results.

This average is the simulated probability:

$$\check{P}_{ni} = \frac{1}{R} \sum_{r=1}^{R} L_{ni}(\beta^r),$$

where R is the number of draws. $\check{P}_{ni}$ is an unbiased estimator of P_{ni} by construction. Its variance decreases as R increases. It is strictly positive, so that $ln\check{P}_{ni}$ is defined, which is useful for approximating the log-likelihood function below. $\check{P}_{ni}$ is smooth (twice differentiable) in the parameters θ and variables x, which facilitates the numerical search for the maximum likelihood function and the calculation of elasticities. And $\check{P}_{ni}$ sums to one over alternatives, which is useful in forecasting.

The simulated probabilities are inserted into the log-likelihood function to give a simulated log-likelihood:

$$SLL = \sum_{n=1}^{N} \sum_{j=1}^{J} y_{nj} ln \check{P}_{nj},$$

where $y_{nj} = 1$ if n chose j and zero otherwise. The maximum simulated likelihood estimator is the value of θ that maximizes SLL. Usually, different draws are taken for each observation. This procedure maintains independence over decision makers of the simulated probabilities that enter SLL.

5.4 Bayesian Procedures

This section aims at introducing the bayesian procedures used to estimate mixed logit models. As they represent relatively new procedures they are described in more details than the precedent procedures, drawn on material in Train (2003).

A powerful set of procedures for estimating discrete choice models has been developed within the Bayesian tradition. The breakthrough concepts were introduced by Albert and Chib (1993) and McCulloch and Rossi (1994) in the context of probit, and by Allenby and Lenk (1994) for mixed logits with normally distributed coefficients. These authors showed how the parameters of the model can be estimated without needing to calculate the choice probabilities. Their procedures provide an alternative to the classical estimation methods. Rossi et al. (1996) and Allenby and Rossi (1999) showed how the procedures can also be used to obtain information on individual-level parameters within a model with random taste variation. Train (2001) extended the Bayesian procedure for mixed logit to nonnormal distributions of coefficients, including lognormal, uniform, and triangular distributions.

Two important notes are required regarding the Bayesian perspective. First, the Bayesian procedures, and the term "Hierarchical Bayes" that is often used in the context of discrete choice models, refer to an estimation method, not a behavioural model. Probit, mixed logit, or any other model that the researcher specifies can, in principle, be estimated by either classical or Bayesian procedures. Second, the Bayesian perspective from which these procedures arise provides a rich and intellectually satisfying paradigm for inference and decision making. Nevertheless, a researcher who is uninterested in the Bayesian perspective can still benefit from Bayesian procedures: the use of Bayesian procedures does not necessitate that the researcher adopt a Bayesian perspective on statistics. The Von-Misses theorem shows that the Bayesian procedures provide an estimator whose properties can be examined and interpreted in purely classical ways.

5.4.1 Overview of Bayesian Concepts

Consider a model with parameters θ. The researcher has some initial ideas about the value of these parameters and collects data to improve this understanding. Under Bayesian analysis, the researcher's ideas about the parameters are represented by a probability distribution over all possible values that the parameters can take, where the probability represents how likely the researcher thinks it is for the parameters to take a particular value.

Prior to collecting data, the researcher's ideas are based on logic, intuition, or past analyses. These ideas are represented by a density on θ, called the prior distribution and denoted $K(\theta)$ [4].

The researcher collects data in order to improve her ideas about the value of θ. Suppose the researcher observes a sample of N independent decision makers. Let y_n denote the observed choice (or choices) of decision maker n, and let the set of observed choices for the entire sample be labeled collectively as $Y = y_1, \ldots, y_N$. Based on this sample information, the researcher changes, or updates, her ideas about θ. The updated ideas are represented by a new density on θ, labeled $K(\theta|Y)$ and called the posterior distribution. This posterior distribution depends on Y, since it incorporates the information that is

[4] In the traditional literature we often find phrases such as "x is random" or "we shall treat w as random" or even "we shall treat x as fixed, i.e. as not random" where "random" means that the object in question will be assigned a probability distribution. In the Bayesian approach all objects appearing in a model are assigned probability distributions and are random in this sense. The only distinction between objects is whether they will become known for sure when the data are in, in which case they are data (!); or whether they will not become known for sure, in which case they are parameters. Generally, the words "random" and "fixed" do not figure in a Bayesian analysis and should be avoided (Lancaster, 2006).

contained in the observed sample.

There is a precise relationship between the prior and posterior distribution, established by Bayes' rule. Let $P(y_n|\theta)$ be the probability of outcome y_n for decision maker n. This probability is the behavioural model that relates the explanatory variables and parameters to the outcome, though the notation for the explanatory variables is omitted for simplicity. The probability of observing the sample outcomes Y is

$$L(Y|\theta) = \prod_{n=1}^{N} P(y_n|\theta)$$

This is the likelihood function (not logged) of the observed choices. Note that it is a function of the parameters θ.

Bayes' rule provides the mechanism by which the researcher improves her ideas about θ. By the rules of conditioning,

$$K(\theta|Y)L(Y) = L(Y|\theta)k(\theta) \tag{5.8}$$

where $L(Y)$ is the marginal probability of Y, marginal over θ:

$$L(Y) = \int L(Y|\theta)k(\theta)d\theta.$$

Both sides of equation (5.8) represent the joint probability of Y and θ, with the conditioning in opposite directions. The left-hand side is the probability of Y times the probability of θ given Y, while the right-hand side is the probability of θ times the probability of Y given θ. Rearranging, we have

$$K(\theta|Y) = \frac{L(Y|\theta)k(\theta)}{L(Y)} \tag{5.9}$$

This equation is Bayes' rule applied to prior and posterior distributions. In general, Bayes rule links conditional and unconditional probabilities in any setting and does not imply a Bayesian perspective on statistics. Bayesian statistics arises when the unconditional probability is the prior distribution (which reflects the researcher's ideas about θ *not* conditioned on the sample information) and the conditional probability is the posterior distribution (which gives the researcher's ideas about θ conditioned on the sample information).

We can express equation (5.9) in a more compact and convenient form. The marginal probability of Y, $L(Y)$, is constant with respect to θ and, more specifically, is the integral of the numerator of (5.9). As such, $L(Y)$ is simply the normalizing constant that assures that the posterior distribution integrates to 1, as required for any proper density. Using this fact, equation (5.9) can be stated more succinctly by saying simply that the posterior distribution is proportional to the prior distribution times the likelihood function:

$$K(\theta|Y)\alpha L(Y|\theta)k(\theta).$$

Intuitively, the probability that the researcher ascribes to a given value for the parameters after seeing the sample is the probability that she ascribes before seeing the sample times the probability (i.e., likelihood) that those parameter values would result in the observed choices. The mean of the posterior distribution is

$$\overline{\theta} = \int \theta K(\theta|Y)d\theta \tag{5.10}$$

This mean has importance from both a Bayesian and a classical perspective. From a Bayesian perspective, $\overline{\theta}$ is the value of θ that minimizes the expected cost of the researcher being wrong about θ, if the cost of error is quadratic in the size of the error (Lancaster, 2006; Train, 2003). From a classical perspective, $\overline{\theta}$ is an estimator that has the same asymptotic sampling distribution as the maximum likelihood estimator.

5.4.2 Drawing from the Posterior

Usually, the posterior distribution does not have a convenient form from which to take draws. For example, we know how to take draws easily from a joint untruncated normal distribution; however, it is rare that the posterior takes this form for the entire parameter vector. Importance sampling can be useful for simulating statistics over the posterior. Geweke (1992, 1997) describes the approach with respect to posteriors and provides practical guidance on appropriate selection of a proposal density. Two other methods are particularly useful for taking draws from a posterior distribution: Gibbs sampling and the Metropolis-Hasting algorithm. These methods are often called Monte Carlo Markov chain, or MCMC, methods. Formally, Gibbs sampling is a special type of Metropolis-Hasting algorithm (Gelman, 1992).However, the case is so special, and so conceptually straightforward, that the term Metropolis-Hasting (MH) is usually reserved for versions that are more complex than Gibbs sampling. That is, when the MH algorithm is Gibbs sampling, it is referred to as Gibbs sampling, and when it is more complex than Gibbs sampling, it is referred to as the MH algorithm.

As stated, the mean of the posterior is simulated by taking draws from the posterior and averaging the draws. Instead of taking draws from the multidimensional posterior for all the parameters, Gibbs sampling allows the researcher to take draws of one parameter at a time (or a subset of parameters), conditional on values of the other parameters (Casella and George, 1992). Drawing from the posterior for one parameter conditional on the others is usually much easier than drawing from the posterior for all parameters simultaneously. In some cases, the MH algorithm is needed in conjunction with Gibbs sampling. The MH algorithm is particularly useful in the context of posterior distributions because the normalizing constant for the posterior need not be calculated. Recall that the posterior is the prior times the likelihood function, divided by a normalizing constant that assures that the posterior integrates to one. The MH algorithm can be applied without knowing or calculating the normalizing constant of the posterior. In summary, Gibbs sampling, combined if necessary with the MH algorithm, allows draws to be taken from the posterior of a parameter vector for essentially any model.

Gibbs Sampling

For multinomial distributions, it is sometimes difficult to draw directly from the joint density and yet easy to draw from the conditional density of each element given the values of the other elements. Gibbs sampling can be used in these situations. A general explanation is provided by Casella and George (1992).

Consider two random variables ε_1 and ε_2. Generalization to higher dimension is obvious. The joint density is $f(\varepsilon_1, \varepsilon_2)$, and the conditional densities are $f(\varepsilon_1|\varepsilon_2)$ and $f(\varepsilon_2|\varepsilon_1)$. Gibbs sampling proceeds by drawing iteratively from the conditional densities: drawing ε_1 conditional on a value of ε_2, drawing ε_2 conditional on this draw of ε_1, drawing a new ε_1 conditional on the new value of ε_2, and so on. This process converges to draws from the joint density. To be more precise:

1. Choose an initial value for ε_1, called ε_1^0 Any value with nonzero density can be chosen.

2. Draw a value of ε_2 called ε_2^0, from $f(\varepsilon_2|\varepsilon_1^0)$.

3. Draw a value of ε_1, called ε_1^1 from $f(\varepsilon_1|\varepsilon_2^0)$

4. Draw ε_2^1 from $f(\varepsilon_2|\varepsilon_1^1)$, and so on.

The Metropolis-Hastings Algorithm

If all else fails, the Metropolis-Hastings (MH) algorithm can be used to obtain draws from a density. Initially developed by Metropolis et al. (1953) and generalized by Hastings (1970), the MH algorithm operates as follows. The goal is to obtain draws from $f(\varepsilon)$.

1. Start with a value of the vector ε, labeled ε^0.

2. Choose a trial value of ε^1 as $\tilde{\varepsilon}^1 = \varepsilon^0 + \eta$, where η is drawn from a distribution $g(\eta)$ that has zero mean. Usually a normal distribution is specified for $g(\eta)$.

3. Calculate the density at the trial value $\tilde{\varepsilon}^1$, and compare it with the density at the original value ε^0. That is, compare $f(\tilde{\varepsilon}^1)$ with $f(\varepsilon^0)$. If $f(\tilde{\varepsilon}^1) \geq f(\varepsilon^0)$, then accept $\tilde{\varepsilon}^1$, label it ε^1, and move to step 4. If $f(\tilde{\varepsilon}^1) = f(\varepsilon^0)$, then accept $\tilde{\varepsilon}^1$ with probability $f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$, and reject it with probability $1 - f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$. To determine whether to accept or reject $\tilde{\varepsilon}^1$ in this case, draw a standard uniform μ. If $\mu \leq f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$, then keep $\tilde{\varepsilon}^1$. Otherwise, reject $\tilde{\varepsilon}^1$. If $\tilde{\varepsilon}^1$ is accepted, then label it ε^1. If $\tilde{\varepsilon}^1$ is rejected, then use ε^0 as ε^1.

4. Choose a trial value of ε^2 as $\tilde{\varepsilon}^2 = \varepsilon^1 + \eta$, where η is a new draw from $g(\eta)$.

5. Apply the rule in step 3 to either accept $\tilde{\varepsilon}^2$ as ε^2 or reject $\tilde{\varepsilon}^2$ and use ε^1 as ε^2.

6. Continue this process for many iterations. The sequence et becomes equivalent to draws from $f(\varepsilon)$ for sufficiently large t.

The draws are serially correlated, since each draw depends on the previous draw. In fact, when a trial value is rejected, the current draw is the same as the previous draw. This serial correlation needs to be considered when using these draws. The MH algorithm can be applied with any density that can be calculated. The algorithm is particularly useful when the normalizing constant for a density is not known or cannot be easily calculated. Suppose that we know that ε is distributed proportional to $f^*(\varepsilon)$. This means that the density of ε is $f(\varepsilon) = \frac{1}{k} f^*(\varepsilon)$, where the normalizing constant $k = \int f^*(\varepsilon) d\varepsilon$ assures that f integrates to 1. Usually k cannot be calculated analytically, for the same reason that we need to simulate integrals in other settings. Luckily, the MH algorithm does not utilize k. A trial value of et is tested by first determining whether $f(\tilde{\varepsilon}^t) > f(\tilde{\varepsilon}^{t-1})$. This comparison is unaffected by the normalizing constant, since the constant enters the denominator on both sides. Then, if $f(\tilde{\varepsilon}^t) \leq f(\tilde{\varepsilon}^{t-1})$, we accept the trial value with probability $f(\tilde{\varepsilon}^t)/f(\tilde{\varepsilon}^{t-1})$. The normalizing constant drops out of this ratio. The MH algorithm is actually more general than described here, though in practice it is usually applied as described. Chib and Greenberg (1995) provide an excellent description of the more general algorithm as well as an explanation of why it works. Under the more general definition, Gibbs sampling is a special case of the MH algorithm, as Gelman (1992) pointed out. The MH algorithm and Gibbs sampling are often called Markov chain Monte Carlo (MCMC, or MC-squared) methods; a description of their use in econometrics is provided by Chib and Greenberg (1996). The draws are Markov chains because each value depends only on the immediately preceding one, and the methods are Monte Carlo because random draws are taken.

5.4.3 Posterior Mean as a Classical Estimator

The Bayesian procedure provides draws from the joint posterior of the parameters. In a Bayesian analysis, these draws are used in a variety of ways depending on the purpose of the analysis. The mean and standard deviation of the draws are simulated approximations to the mean and standard deviation of the posterior. These statistics have particular importance from a classical perspective, due to the Bernstein-von Mises theorem. Consider a model with parameters θ whose true value is θ^*. The maximum of the likelihood function is $\widehat{\theta}$, and the mean of the posterior is $\overline{\theta}$ for a prior that is proper and strictly positive in a neighbourhood of θ^*. Three interrelated statements are established in different versions of the theorem (e.g., Rao (1987); Cam and Yang (1990); Lehmann and Casella (1998); Bickel and Doksum (2000)

1. The posterior distribution of θ converges to a normal distribution with covariance B^{-1}/N around its mean, where B is the information matrix. Stated more precisely: $\sqrt{N}(\theta - \overline{\theta}) \xrightarrow{d} N(0, B^{-1})$, where the distribution that is converging is the posterior rather than the sampling distribution.

2. The posterior mean converges to the maximum of the likelihood function: $\sqrt{N}(\overline{\theta}) - \widehat{\theta} \xrightarrow{p} 0$. This result is a natural implication of statement (1). Asymptotically, the shape of the posterior becomes arbitrarily close to the shape of the likelihood function, since the posterior is proportional to the likelihood function times the prior and the prior becomes irrelevant for large enough N. The mean and mode of a normal distribution are the same.

3. The asymptotic sampling distribution of the posterior mean is the same as for the maximum of the likelihood function: $\sqrt{N}(\overline{\theta}) - \theta^* \xrightarrow{d} N(0, B^{-1})$. This result is obvious from statement (2).

The third statement says that the mean of the posterior is an estimator that, in classical terms, is equivalent to MLE. The first statement establishes that the standard deviations of the posterior provide classical standard errors for the estimator. The true mean and standard deviation of the posterior cannot be calculated exactly except in very simple cases. These moments are approximated through simulation, by taking draws from the posterior and calculating the mean and standard deviation of the draws. For fixed number of draws, the simulated mean, denoted $\breve{\theta}$, is consistent and asymptotically normal, with variance equal to $1 + (1/R)$ times the variance of the non-simulated mean, where R is the number of (independent) draws. If the number of draws (whether independent or not) is considered to rise with N at any rate, the simulation noise disappears asymptotically such that $\breve{\theta}$ is efficient and asymptotically equivalent to MLE. In contrast, MSLE is inconsistent for a fixed number of draws. For consistency, the number of draws must be considered to rise with N, but even this condition is not sufficient for asymptotic normality. The number of draws must be considered to rise faster than $\sqrt{N}$ for MSLE to be asymptotically normal, in which case it is also equivalent to MLE. Since it is difficult to know in practice how to satisfy the condition that the number of draws rises faster than $\sqrt{N}$, $\breve{\theta}$ is attractive relative to MSLE, even though their non-simulated counterparts are equivalent.

The researcher can therefore use Bayesian procedures to obtain parameter estimates and then interpret them the same as if they were maximum likelihood estimates. A highlight of the Bayesian procedures is that the results can be interpreted from both perspectives simultaneously, drawing on the insights afforded by each tradition. This dual interpretation parallels that of the classical procedures, whose results can be transformed for Bayesian interpretation as described by Geweke (1989). In short, the researcher's statistical perspective need not dictate her choice of procedure.

5.4.4 Posteriors for the Mean and Variance

The posterior distribution takes a very convenient form for some simple inference processes. We describe two of these situations, which, as we will see, often arise within more complex models for a subset of the parameters. Both results relate to the normal distribution. We first consider the situation where the variance of a normal distribution is known, but the mean is not. We then turn the tables and consider the mean to be known but not the variance. Finally, combining these two situations with Gibbs sampling, we consider the situation where both the mean and variance are unknown.

Result A: Unknown Mean, Known Variance

We discuss the one-dimensional case first, and then generalize to multiple dimensions. Consider a random variable β that is distributed normal with unknown mean b and known variance σ. The researcher observes a sample of N realizations of the random variable, labeled β_n, $n = 1, \ldots, N$. The sample mean is $\overline{\beta} = (1/N) \sum_n \beta_n$. Suppose the researcher's prior on b is $N(b_0, s_0)$; that is, the researcher's prior beliefs are represented by a normal distribution with mean b_0 and variance s_0. Note that we now have two normal distributions: the distribution of β, which has mean b, and the prior distribution on this unknown mean, which has mean b_0. The prior indicates that the researcher thinks it is most likely that $b = b_0$ and also thinks there is a 95 percent chance that b is somewhere between $b_0 - 1.96\sqrt{s_0}$ and $b_0 + 1.96\sqrt{s_0}$. Under this prior, the posterior on b is $N(b_1, s_1)$ where

$$b_1 = \frac{\frac{1}{s_0}b_0 + \frac{N}{\sigma}\overline{\beta}}{\frac{1}{s_0} + \frac{N}{\sigma}}$$

and

$$s_1 = \frac{1}{\frac{1}{s_0} + \frac{N}{\sigma}}$$

The posterior mean b_1 is the weighted average of the sample mean and the prior mean[5].The weight on the sample mean rises as sample size rises, so that for large enough N, the prior mean becomes irrelevant. Often a researcher will want to specify a prior that represents very little knowledge about the parameters before taking the sample. In general, the researcher's uncertainty is reflected in the variance of the prior. A large variance means that the researcher has little idea about the value of the parameter. Stated equivalently, a prior that is nearly flat means that the researcher considers all possible values of the parameters to be equally likely. A prior that represents little information is called *diffuse*.

The multivariate versions of this result are similar. Consider a K-dimensional random vector $\beta \tilde{N}(b, W)$ with known W and unknown b. The researcher observes a sample β_n, $n = 1, \ldots, N$, whose sample mean is $\overline{\beta}$. If the researcher's

[5] For the proof, see Train (2003)

prior on b is diffuse (normal with an unboundedly large variance), then the posterior is $N(\overline{\beta}, W/N)$. To take draws from this posterior let L be the Choleski factor of W/N. Draw K iid standard normal deviates, $\eta_i, i = 1, \ldots, K$, and stack them into a vector $\eta = \langle \eta_1, \ldots, \eta_K \rangle'$. Calculate $\tilde{b} = \overline{\beta} + L\eta$. The resulting vector $\tilde{b}$ is a draw from $N(\overline{\beta}, W/N)$.

Result B: Unknown Variance, Known Mean

Consider a (one-dimensional) random variable that is distributed normal with known mean b and unknown variance s. The researcher observes a sample of N realizations, labeled β_n, $n = 1, \ldots, N$. The sample variance around the known mean is $\overline{s} = (1/N)\sum_n(\beta_n - b)^2$. Suppose the researcher's prior on s is inverted gamma with degrees of freedom v_0 and scale s_0. This prior is denoted $IG(v_0, s_0)$. The density is zero for any negative value for s, reflecting the fact that a variance must be positive. The mode of the inverted gamma prior is $s_0 v_0/(1 + v_0)$. Under the inverted gamma prior, the posterior on σ is also inverted gamma $IG(v_1, s_1)$, where

$$v_1 = v_0 + N,$$

$$s_1 = \frac{s_0 v_0 + N\overline{s}}{v_0 + N}.$$

The inverted gamma prior becomes more diffuse with lower v_0. For the density to integrate to one and have a mean, v_0 must exceed 1. It is customary to set $s_0 = 1$ when specifying $v_0 \to 1$. Under this diffuse prior, the posterior becomes $IG(1 + N, (1 + N\overline{s})/(1 + N))$. The mode of this posterior is $(1 + N\overline{s})/(2 + N)$, which is approximately the sample variance $\overline{s}$ for large N. The multivariate case is similar. The multivariate generalization of an inverted gamma distribution is the inverted Wishart distribution. The result in the multivariate case is the same as with one random variable except that the inverted gamma is replaced by the inverted Wishart. A K-dimensional random vector $\beta\tilde{N}(b, W)$ has known b but unknown W. A sample of size N from this distribution has variance around the known mean of $\overline{S} = (1/N)\sum_n(\beta_n - b)(\beta_n - b)'$. If the researcher's prior on W is inverted Wishart with v_0 degrees of freedom and scale matrix S_0, labeled $IW(v_0, S_0)$, then the posterior on W is $IW(v_1, S_1)$ where

$$v_1 = v_0 + N,$$

$$S_1 = \frac{S_0 v_0 + N\overline{S}}{v_0 + N}.$$

The prior becomes more diffuse with lower v_0, though v_0 must exceed K in order for the prior to integrate to one and have means. With $S_0 = I$, where I is the K-dimensional identity matrix, the posterior under a diffuse prior becomes $IW(K+N, (KI+N\overline{S})/(K+N))$. Conceptually, the prior is equivalent to the researcher having a previous sample of K observations whose sample variance was I. As N rises without bound, the influence of the prior on the posterior eventually disappears. Consider first an inverted gamma $IG(v_1, s_1)$. Draws are taken as follows:

1. Take v_1 draws from a standard normal, and label the draws $\eta_i, i = 1, \ldots, v_1$.
2. Divide each draw by $\sqrt{s_1}$, square the result, and take the average. That is, calculate $r = (1/v_1)\sum_i(\sqrt{1/s_1\eta_i})^2$, which is the sample variance of v_1 draws from a normal distribution whose variance is $1/s_1$.
3. Take the inverse of $r : \tilde{s} = 1/r$ is a draw from the inverted gamma.

Draws from a K-dimensional inverted Wishart $IW(v_1, S_1)$ are obtained as follows:

1. Take v_1 draws of K-dimensional vectors whose elements are independent standard normal deviates. Label these draws $\eta_i, i = 1, \ldots, v_1$.
2. Calculate the Choleski factor of the inverse of S_1, labeled L, where $LL' = S_1^{-1}$.
3. Create $R = (1/v_1)\sum_i(L\eta_i)(L\eta_i)'$. Note that R is the variance of draws from a distribution with variance S_1^{-1}.
4. Take the inverse of R. The matrix $\tilde{S} = R^{-1}$ is a draw from $IW(v_1, S_1)$.

Unknown Mean and Variance

Suppose that both the mean b and variance W are unknown. For neither of these parameters does the posterior take a convenient form. However, draws can easily be obtained using Gibbs sampling and results A and B. A draw of b is taken conditional on W, and then a draw of W is taken conditional on b. Result A says that the posterior for b conditional on W is normal, which is easy to draw from. Result B says that the posterior for W conditional on b is inverted Wishart, which is also easy to draw from. Iterating through numerous cycles of draws from the conditional posteriors provides, eventually, draws from the joint posterior.

5.4.5 Hierarchical Bayes for Mixed Logit

In this section we show how the Bayesian procedures can be used to estimate the parameters of a mixed logit model. We utilize the approach developed by Allenby (1997), implemented by Software (2000), and generalized by Train (2001). Let the utility that person n obtains from alternative j in time period t be

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt},$$

where ε_{njt} is iid extreme value and $\beta_n \tilde{N}(b, W)$.

Giving β_n' a normal distribution allows us to use results A and B, which speeds estimation considerably. The researcher has priors on b and W. Suppose the prior on b is normal with an unboundedly large variance. Suppose that the prior on W is inverted Wishart with K degrees of freedom and scale matrix I, the K-dimensional identity matrix.

Note that these are the priors used for results A and B. More flexible priors can be specified for W, using the procedures of, for example, McCulloch and Rossi (2000), though doing so makes the Gibbs sampling more complex.

A sample of N people is observed. The chosen alternatives in all time periods for person n are denoted $y_n' = \langle y_{n1}, \ldots, y_{nT} \rangle$, and the choices of the entire sample are labelled $Y = \langle y_1, \ldots, y_T \rangle$. The probability of person n's observed choices, conditional on β, is

$$L(y_n|\beta) = \prod_t \frac{e^{\beta' x_{ny_{nt}t}}}{\sum_j e^{\beta' x_{njt}}}.$$

The probability not conditional on β is the integral of $L(y_n|\beta)$ over all β:

$$L(y_n|b, W) = \int L(y_n|\beta) f(\beta|b, W) d\beta,$$

where $f(\beta|b, W)$ is the normal density with mean b and variance W. This $L(y_n|b, W)$ is the mixed logit probability. The posterior distribution of b and W is, by definition,

$$K(b, W|Y) \alpha \prod_n L(y_n|b, W) k(b, W), \qquad (5.11)$$

where $k(b, W)$ is the prior on b and W described earlier (i.e., normal for b times inverted Wishart for W).

It would be possible to draw directly from $K(b, W|Y)$ with the MH algorithm. However, doing so would be computationally very slow. For each iteration of the MH algorithm, it would be necessary to calculate the right-hand side of (5.11). However, the choice probability $L(y_n|b, W)$ is an integral without a closed form and must be approximated through simulation. Each iteration of the MH algorithm would therefore require simulation of $L(y_n|b, W)$ for each n. That would be very time-consuming, and the properties of the resulting estimator would be affected by it. Recall that the properties of the simulated mean of the posterior were derived under the assumption that draws can be taken from the posterior without needing to simulate the choice probabilities. MH applied to (5.10) violates this assumption.

Drawing from $K(b, W|Y)$ becomes fast and simple if each β_n is considered to be a parameter along with b and W, and Gibbs sampling is used for the three sets of parameters b, W, and $\beta_n \forall n$. The posterior for b,W, and $\beta_n \forall n$ is

$$K(b, W, \beta_n \forall n|Y) \alpha \prod_n L(y_n|\beta_n) f(\beta_n|b, W) k(b, W).$$

Draws from this posterior are obtained through Gibbs sampling. A draw of each parameter is taken, conditional on the other parameters:

1. Take a draw of b conditional on values of W and $\beta_n \forall n$.

2. Take a draw of W conditional on values of b and $\beta_n \forall n$.

3. Take a draw of $\beta_n \forall n$ conditional on values of b and W.

Each of these steps is easy, as we will see. Step 1 uses result A, which gives the posterior of the mean given the variance. Step 2 uses result B, which gives the posterior of the variance given the mean. Step 3 uses an MH algorithm, but in a way that does not involve simulation within the algorithm. Each step is described in the following.

1. $b|W, \beta_n \forall n$. We condition on W and each person's β_n in this step, which means that we treat these parameters as if they were known. Result A gives us the posterior distribution of b under these conditions. The β_n's constitute a sample of N realizations from a normal distribution with unknown mean b and known variance W. Given our diffuse prior on b, the posterior on b is $N(\overline{\beta}, W/N)$, where $\overline{\beta}$ is the sample mean of the β_n's. To take draws from this posterior proceed as Result A described in section 5.4.4.

2. $W|b, \beta_n \forall n$. Result B gives us the posterior for W conditional on b and the β_n's. The β_n's constitute a sample from a normal distribution with known mean b and unknown variance W. Under our prior on W, the posterior on W is inverted Wishart with $K + N$ degrees of freedom and scale matrix $(KI{+}NS_1)/(K{+}N)$, where $S_1 = (1/N)\sum_n(\beta_n{-}b)(\beta_n{-}b)'$is the sample variance of the β_n's around the known mean b. It is easy to take draws from inverted gamma and inverted Wishart distributions, as shown before.

3. $\beta_n|b, W$. The posterior for each person's β_n, conditional on their choices and the population mean and variance of β_n, is

$$K(\beta_n|b, W, y_n)\alpha L(y_n|\beta_n)f(\beta_n|b, W), \tag{5.12}$$

There is no simple way to draw from this posterior, and so the MH algorithm is used. Note that the right-hand side of (5.12) is easy to calculate: $L(y_n|\beta_n)$ is a product of logits, and $f(\beta_n|b, W)$ is the normal density. The MH algorithm operates as follows:

(a) Start with a value β_n^0.

(b) Draw K independent values from a standard normal density, and stack the draws into a vector labeled η^1.

(c) Create a trial value of β_n^1 as $\tilde{\beta}_n^1 = \beta_n^0 + \rho L \eta^1$, where ρ is a scalar specified by the researcher and L is the Choleski factor of W. Note that the proposal distribution is specified to be normal with zero mean and variance $\rho^2 W$.

(d) Draw a standard uniform variable μ^1.

(e) Calculate the ratio

$$F = \frac{L(y_n|\tilde{\beta}_n^1)\rho(\tilde{\beta}_n^1|b, W)}{L(y_n|\tilde{\beta}_n^0)\rho(\tilde{\beta}_n^0|b, W)}.$$

(f) If $\mu^1 \leq F$, accept $\tilde{\beta}_n^1$ and let $\beta_n^1 = \tilde{\beta}_n^1$. If $\mu^1 > F$, reject $\tilde{\beta}_n^1$ and let $\beta_n^1 = \beta_n^0$.

(g) Repeat the process many times. For high enough t, β_n^t is a draw from the posterior.

We can know draw from the posterior for each parameter conditional on the other parameters. We combine the procedures into a Gibbs sampler for the three sets of parameters. Start with any initial values b^0, W^0, and β_n^0. The tth iteration of the Gibbs sampler consists of these steps:

1. Draw b^t from $N(\tilde{\beta}^{t-1}, W^{t-1}/N)$, where $\tilde{\beta}^{t-1}$ is the mean of the β_n^{t-1}'s.

2. Draw W_t from $IW(K + N, (KI + NS^{t-1})/(K + N))$, where $S^{t-1} = \sum_n (\beta_n^{t-1} - b^t)(\beta_n^{t-1} - b^t)'/N$.

3. For each n, draw β_n^t using one iteration of the MH algorithm previously described, starting from β_n^{t-1} and using the normal density $f(\beta_n|b^t, W^t)$.

These three steps are repeated for many iterations. The resulting values converge to draws from the joint posterior of b, W, and $\beta_n \forall n$. Once the converged draws from the posterior are obtained, the mean and standard deviation of the draws can be calculated to obtain estimates and standard errors of the parameters. Note that this procedure provides information about β_n for each n, similar to the procedure using classical estimation.

5.5 Challenges in Estimating VTTS

The value of time in transport has usually been estimated though classical multinomial logit which, assuming homogeneous tastes, can derive a single value of time for a fictitious average individual. Recently, the mixed logit model has been applied with different specifications and various degrees of sophistication. Although the theory is in general relatively clear, practical specification and estimation represent real challenges. Some important topics are discussed here focusing in the objective of estimating the VTTS.

5.5.1 Identifying Preference Heterogeneity

The most popular way of acknowledging systematic variations on preferences (or systematic taste variations) has been (within a specific trip purpose) to segment a sample based on exogenous criteria such as income, trip length and time of day for passengers and in length, type of commodity and ownership (own account or hire) for freight. This segmentation is achieved through estimating separate models for each segment or by interacting the travel time with an individual socio-economic or specific trip characteristics (Gaudry et al., 1989; Revelt and Train.K., 1998; Ortuzar and Willumsen, 2001; Amador et al., 2004). Hensher and Goodwin (2004) note that in practice, the selection of the number and dimensions of discrimination is not usually driven by questions of statistical diagnostics, research hypothesis and evidence. It is constrained by the specific properties of the forecasting and appraisal models within which the empirical values will be used.

However, even after controlling for observable characteristics, there is a lot of heterogeneity left. This heterogeneity is due to factors which can not be observed or are difficult to measure. In these cases, this heterogeneity can take form of a random parameter. One disadvantage of specifying random parameters is that information is not provided about factors determining these variations. To maximise the explanatory power of the model, one should explain as much systematic variation as possible, and allow for a random variation where it is significant.

5.5.2 Selecting Random Parameters

McFadden (2000) propose a Lagrange Multiplier test as a basis for accepting/rejection the preservation of fixed parameters in the mode. Brownstone (2001) provides a succinct summary of the test. These tests work by constructing artificial variables as in equation (5.13):

$$z_n = (x_{in} - \overline{x}_i)^2, \ with \ \overline{x}_i = \sum_j x_{jn} P_{jn}, \tag{5.13}$$

and P_{jn} is the conditional choice probability. The conditional logit is then re-estimated including these artificial variables, and the null hypothesis of no random coefficients on attributes x is rejected if the coefficients of the artificial variables are significantly different from zero. The actual test for the joint significance of the variables can be carried out using either a Wald or Likelihood Ratio test statistics. Brownstone (2001) suggests that these tests are easy to calculate and appear to be a quite powerful omnibus test; however they are not as good for identifying which error components to include in a more general mixed logit specification. Another test (Hensher and Greene, 2003) is to assume all parameters are random and then examine their estimated standard deviations, using a zero-based t-test for individual parameters and the likelihood ratio test to establish the overall contribution of the additional information. While appealing, this is very demanding for a large number of explanatory variables and might be problematic in establishing the model with a full set of random parameters.

5.5.3 Selecting the Distributions of the Random Parameters

If there is one single issue that can cause much concern it is the influence of the distributional assumptions of random parameters (Hensher and Greene, 2003). Except for the sign of VTTS, we appear to have no theoretical arguments to support one distribution or another. However, there is evidence of a left skewed distribution of VTTS. Abraham and Blanchet (1973) proposed a lognormal distribution in analogy with the income distribution. In effect, it is quite intuitive that there is substantially more individuals with relatively low value of time and not prepared to pay much to save time; in contrast a smaller number of individuals are willing to pay high tolls. This evidence has been being validated by non-parametric studies (Fosgerau, 2007) and by good fits provided by left skewed distributions (lognormal, but also Sb, Raylagh and others).

The lognormal distribution is very popular for the following reasoning (Hensher and Greene, 2003). The central limit theorems explain the genesis of a normal curve. If a large number of random shocks, some positive, some negative, change the size of a particular attribute, x, in an additive fashion, the distribution of that attribute will tend to become normal as the number of shocks increases. But if these shocks act multiplicatively, changing the value of x by randomly distributed proportions instead of absolute amounts, the central limit theorems applied to $y = ln(x)$ tend to produce a normal distribution. Hence x has a lognormal distribution.

The substitution of multiplicative for additive random shocks generates a positively skewed, leptokurtic, lognormal distribution instead of a symmetric, mesokurtic normal distribution. The degree of skewness and kurtosis of the two-parameter lognormal distribution depends only on the variance, and so if this is low enough, the lognormal approximates the normal distribution. Lognormals are appealing in that they are limited to the non-negative domain; however they typically have a very long right-hand tail which is a disadvantage (especially for willingness-to-pay calculations). It is this large proportion of "unseasonable" values that often casts doubt on the appropriateness of the lognormal. Moreover, in parameter estimation, experience has demonstrated that entering an attribute in a utility expression specified with a random parameter that is lognormally distributed, and which is expected a priori to produce a negative mean estimate, typically causes the model either not converge or converge with unacceptable large mean estimates. The trick to overcome this is to reverse the sign of the attribute prior to model estimation.

The simplest way to derive VTTS is to take the ratio of the means of the parameter distributions involved. This is not the mean of the VTTS, but the VTTS derived from coefficients of the "average individual" for each parameter. If the denominator is a constant, as in our case, both values are identical. If it is distributed, the distribution of the ratio can be computed by simulation, as in Sillano and Ortuzar (2005). Revelt and Train.K. (2001) cites three reasons for fixing the cost coefficient: (1) As Ruud (1996) points out, mixed logit models have a tendency to be unstable when all coefficients are allowed to vary. Fixing the price coefficient resolves this instability. (2) If the price coefficient is allowed to vary, the distribution of willingness to pay is the ratio of two distributions, which is often inconvenient to evaluate. With a fixed price coefficient, willingness to pay for an attribute is distributed the same as the coefficient of the attribute. (3) The choice of distribution to use for a price coefficient is problematic. The price coefficient is necessarily negative, such that a normal distribution is inappropriate. With a lognormal distribution (which assures that the price coefficient is always negative), values very close to zero are possible, giving very high (implausibly high) values for willingness to pay.

However, as noted by Train and Weeks (2004), this restriction is counter-intuitive as the marginal utility of money can vary across respondents according to factors that can be independent of observed socio-economic covariates. A fixed price coefficient implies that the standard deviation of unobserved utility, which is called the scale parameter, is the same for all observations; if the price coefficient is constrained to be fixed when in fact scale varies over observations, then the variation in scale will be erroneously attributed to variation in willingness to pay.

In this context the choice of the distribution is dictated not only by the researcher's preferences but also by the model characteristics and uses. Number of recent works (for example Hensher and Greene (2003); Hess et al. (2005)

demonstrate that the choice of distributional assumptions have a significant impact on estimation results, particularly and predictably, in the inferences that can potentially be drawn regarding extreme values. Although selecting distributions for individual parameters is challenge enough, it is compounded when interest focuses on ratios of random parameters, as in the derivation of estimates of willingness to pay (WTP).

5.5.4 Revealed Preference Data

The main advantage of revealed preference data is that it represents the actual choices. Flyvbjerg et al. (2003) for example, point the stated preference approach as a main source of errors in forecasting due to divergences between the stated and the actual behaviour. However, one of the main problems with revealed data is that it usually does not provide a high variation in the choice set (usually no more than two or three options) and in the attributes of these options, making the identification of random variations very difficult.

5.5.5 Optimization Problems

With mixed logit models (especially those with lognormal distributions), maximization of the simulated likelihood function can be difficult numerically. Often the algorithm fails to converge for various reasons. The choice of starting values is often critical, with the algorithm converging from starting values that are close to the maximum but not from other starting values. The issue of local versus global maxima complicates the maximization further, since convergence does not guarantee that the global maximum has been attained. This fact emphasizes the importance of appropriate starting values. In effect in the mixed logit model, the use of inadequate starting points may cause the model not converge or stop in a local maximum.

5.5.6 Imposing Constraints

This point is directly related to the choice of the distributions. In practice we often find that any one distribution has strengths and weaknesses. The weakness is usually associated with the spread or standard deviation of the distribution at its extremes including behaviourally unacceptable sign changes for the symmetrical distributions. One appealing 'solution' is to constrain the distribution in terms of domain (for instance, a truncated normal) or dispersion (constraining the coefficient of variation). Hensher and Greene (2003) simulated the resulting VTTS with lognormal distributions and derived and unusually high mean. They managed to lower it to more plausible values by truncating the simulated distribution, but found it very sensitive to this kind of constraint.

5.5.7 Priors

The introduction of prior knowledge is intrinsic to even the classic analysis. First, the analyst usually has some priors about the result (i.e. one should expect that the value of travel time to be positive and to lay within a reasonable set) and second, the set of hypothesis and parameters need to the estimation of mixed logit models like the form of the distributions and the starting values indirectly represent a prior hypothesis.

5.5.8 Advantages and Problems of Bayesian Procedures

The Bayesian procedures avoid two of the most prominent difficulties associated with classical procedures. First, the Bayesian procedures do not require maximization of any function. Second, desirable estimation properties, such as consistency and efficiency, can be attained under more relaxed conditions with Bayesian procedures than classical ones. Maximum simulated likelihood is consistent only if the number of draws used in simulation is considered to rise with sample size; and efficiency is attained only if the number of draws rises faster than the square root of sample size. In contrast, the Bayesian estimators that we describe are consistent for a fixed number of draws used in simulation and are efficient if the number of draws rises at any rate with sample size.

Nevertheless, to simulate relevant statistics that are defined over a distribution, the Bayesian procedures use an iterative process that converges, with a sufficient number of iterations, to draws from that distribution. This convergence is different from the convergence to a maximum that is needed for classical procedures and involves its own set of difficulties. The researcher cannot easily determine whether convergence has actually been achieved. Thus, the Bayesian procedures trade the difficulties of convergence to a maximum for the difficulties associated with this different kind of convergence. The researcher will need to decide, in a particular setting, which type of convergence is less burdensome.

As we have shown, the Bayesian procedures provide an estimator whose properties can be examined and interpreted in purely classical ways. The researcher can therefore use Bayesian procedures to obtain parameter estimates and then interpret them the same as if they were maximum likelihood estimates. From an estimation perspective, for some behavioural models and distributional specifications, Bayesian procedures are far faster and, after the initial learning that a classicist needs, are more straightforward from a programming perspective than classical procedures. For other models, the classical procedures are easier. The differences can be readily categorized, through an understanding of how the two sets of procedures operate. The researcher can use this understanding in deciding which procedure to use in a particular setting.

However, the use of Bayesian procedures within a Bayesian perspective provides the fascinating opportunity of properly integrating prior beliefs in the analysis. The use of bayesian estimation with a bayesian perspective, which means that the researcher wants to update his prior information based on the new data (and do not use a diffuse prior), also rise some questions.

5.5.9 The Role of the Alternative Specific Constant

The alternative specific constant in a logit-like model assures that the market share estimated by the model corresponds to the actual (sample) market share, for each alternative. It captures the captive market share (which is not affected by the concurrent modes) and also the deterministic part of the utility function which is not explained by the explanatory variables. While this property is very suitable in many market analysis, in traffic forecasting it can poses a major problem. Suppose we could include all the decision variables (usually cost, time, alternative specific variables and decision maker specific variables), there are few reasons to believe that users have a preference for a road or another (behaviour effects like habit can affect the choice in the short term but have few implications in the long term). Affecting a bonus for one option reduces the part of the population willing to change of mode. This characteristic is few realistic and is not compatible with traditional assignment procedures which computes the generalized cost for each route and allocate traffic based on it.

5.6 The Survey

Our empirical analysis relies on a Revealed Preference study based on an Origin-Destination survey. The approach given is the concurrence between a tolled motorway and a free national road (autoroute and route nationale, in French, respectively), in order to compare the trade-off between a faster and tolled and a slower free option. This survey was realized in two pairs Motorway/National Route:

- A28 (Toll bridge of Alençon Nord) and N138, direction Le Mans-Alençon ;
- A11 (Toll bridge of Ancenis) and N23, direction Angers-Nantes ;

These points are illustrated in figure 5.2. We interviewed 1173 truck drivers about:

- The origin and the destination of the trip (last and next points of loading/ unloading);
- OD's frequency;
- Own account or hired;
- Number of employees of the transport company;
- Kind of product transported;
- Type of vehicle (visual observation).

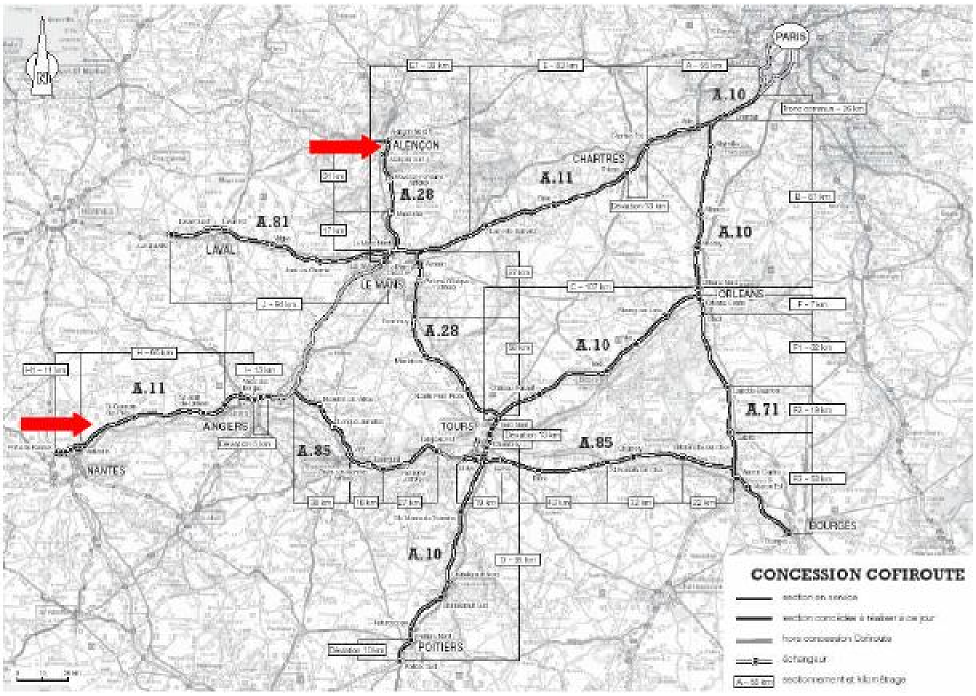

Figure 5.2: Survey's Location.

The traditional econometric approach to estimate the parameters of the discrete choice model is the maximum likelihood, producing the value of the parameters for which the observed sample is most likely to have occurred. Assuming that the observations in the sample are drawn independently at a random from the population, the likelihood of the sample is the product of individual likelihoods.

In an Origin-Destination survey, observations are collected based on their *ex-ante* choice, which characterizes a choice-based sample. The problem of finding tractable estimation procedure possessing desirable statistical properties is not an easy one, and the state of the art is provided by the papers of Coslett (1981) and Manski and McFadden (1981).

It has been found in general that maximum likelihood estimators specific to choice-base sampling are impractical, except in very restrict circumstances, due to computational intractability. However, if the analyst knows the fraction of the decision making population selecting each alternative then a tractable method can be introduced. The approach modifies the familiar maximum likelihood estimator of random sampling by weighting the contribution of each observation to the log-likelihood by the ratio Q_i/S_i, where the numerator is the fraction of the population and the denominator the fraction, of the sample selecting option i. This approach is applied in this study. Sample sizes and count data for the motorway (M) and free road (R), are reported in table 5.1.

Table 5.1: Sample and traffic count data
Sources: Cofiroute; Service des Routes de la DDE de la Sarthe;
Service des Routes de la DDE de Loire Atlantique.

		sample	Count	Weight	Integer Weight
Ancenis	M	400 (50%)	2412 (76%)	6.03	6
	R	395 (50%)	767 (24%)	1.94	2
Alençon	M	183 (48.5%)	962 (50%)	5.25	5
	R	195 (51.5%)	954 (50%)	4.89	5
Total	M	583 (50%)	3374 (66%)	-	-
	R	590 (50%)	1721 (34%)	-	-

Once the sample has been weighed, we remove from the analysis those observations presenting one of the following characteristics:

- No real choice, i.e. the other alternative is too expensive or inexistent;
- Local traffic, distance shorter than 25 km;
- Recorded OD pair disconnected to the site of survey.

After removing these observations, the sample was reduced to 1027 observations, shared as shown in table 5.2.

Table 5.2: Final Sample

		sample	Weighted Sample
Ancenis	M	385	2310
	R	343	686
Alençon	M	170	850
	R	129	645
Total	M	555	3160
	R	472	1331

Table 5.3 presents the summary statistics for the main variables in the sample.

Table 5.3: Summary of descriptive statistics

Variable	Mean	Median	Std dev	Min	Max	Definition
Travel Cost (TC)	34.49	23.01	31.27	0.87	290.68	Toll in €
Travel Time (TT)	-1.24	-0.93	1.19	-8.02	7.92	Δ time in hours
distance	343.96	249.50	317.45	25.80	2227.40	distance in km
loaded	0.91	1	0.28	0	1	1 if loaded
hire	0.75	1	0.43	0	1	1 if for hire

5.7 Econometric Results

5.7.1 Maximum Likelihood estimations

We introduce the variables “hire” and “loaded” as sources of systematic variation as we could imagine that transport for hire (against own account) and loaded vehicles (against empty) have higher values of time. The variable distance was also tested as a source of systematic variation but not kept in the model due to a high correlation (0.81) with the travel cost. This fact represents a weakness of the revealed preference approach as discussed before. Using the Lagrange Multiplier test presented before, we have found that the travel time parameter was the only one presenting a significant random variation over the population.

As pointed by many authors, the simplicity of the MNL represents an strong advantage due to its properties and well-known estimation procedure; in this sense, the classic MNL shall be the starting point of any discrete choice estimation. We first estimate the model without the sources of systematic heterogeneity. The results of this model are shown in model MNL(1) in table 5.4. The value of time estimated by this model is €52 /h.

We then add the interaction between the travel time and the variables “loaded” and “for hire” in the model MNL(2). We can see that these factors strongly affect the value of time, which can be whiten as:

$$VTTS_{MNL} = 46 + 10loaded + 16hire \tag{5.14}$$

The average in the sample using MNL (2) is €67.1, ranging from €46 for empty and €72 for hire and loaded.

Results estimated by MNL are extremely close to those find by Alvarez et al. (2007) in Spain (€64.1) using the same model, but far higher than the national standard values used in both countries.

We then estimate models with distributed coefficients. We tested the Matlab code developed by Kenneth Train [6] and the *R* code developed by Ryuich Tamura. The Matlab code of Kenneth Train was kept for the final estimations.

We first estimated, as usual, considering the travel time parameter as lognormally distributed. Model 5.15 uses only time and cost as explanatory variables. Model 5.16 includes interactions. Note that in the models using lognormal distributions, the travel time was multiplied by -1 to get positive coefficients.

$$VTTS_{ML} = \frac{1}{0.0017} e^{2.87+1.99N(0,1)} \tag{5.15}$$

$$VTTS_{ML} = \frac{1}{0.024} e^{1.98+0.10 loaded+0.21 hire+1.80N(0,1)} \tag{5.16}$$

Results show unacceptable mean and variance. This result confirms the difficulty in estimating mixed logit models with lognormal distributions discussed before.

We tried also to estimate the model using the cost variable following a lognormal PDF and the cost normally distributed and the time lognormally distributed. In both cases we failed to achieve convergence. There is a folk concept floating among researchers in the field that the variance of random coefficients are identified empirically only if with repeated choices for each person. This concept is probably too severe, but it indicates the difficulty we face.[7]

5.7.2 Bayesian Estimations

Within the Bayesian approach, instead of proceeding adding constraints or changing the PDF in order to find more reasonable values, we include our

[6] Available at http://elsa.berkeley.edu/t̃rain/software.html

[7] based in a discussion with Kenneth Train.

beliefs as “priors”.

As a prior distributions for the Bayesian estimations we adopt as mean the current value used in France. Jiang (1998) finds an average VTTS of 195 FF, or approximately €30, which is also the value adopted as the governmental recommendation in the “Rapport Boiteux” (Commissariat Général du Plan, 2001). Since the VTTS from a linear in parameters utility function is the ration between the time and cost estimates, we decided to keep the cost parameter from the model (ML); the mean of the prior distribution for time becomes the mean of the value of time used today (€30) inflated by the economic growth between 2000 and 2005 (€32.3) multiplied by 0.01. We specify a large standard deviation (3.0) in order to diffuse the prior. The result of this estimation is shown in model HB. The estimation was performed using the Matlab code developed by Kenneth Train. It should be noted that the HB reproduces the maximum likelihood estimations when the coefficients are considered fixed and when the prior information is very diffuse and the simulation is long enough. We have used a very large number of draws in order to be able to identify the variance.

Note that the approach adopted to represent the real market share, weighting observations (and then the likelihood function) was derived and is usually applied for maximum likelihood estimations. Although we believe the same approach can be applied to Bayesian estimations without further concerns, we did not find any application or theoretical discussions on this point.

We first estimate the model considering the cost coefficient as fixed and the time as lognormally distributed. Results show that the VTTS distibution (in €/h) can be written as:

$$VTTS_{HB} = \frac{1}{0.03} e^{0.294+0.083loaded+0.175hire+0.0059N(0,1)} \tag{5.17}$$

Even after a very high number of draws, the bayesian algorithm was unable to get apart from the initial solution and to identify the heterogeneity (small variance). The average value of time in the population is 54.6€. Figure 5.3 shows the VTTS distribution when both loaded and hire dummy variables are equal to zero.

We then estimate the model with the cost coefficient following a normal

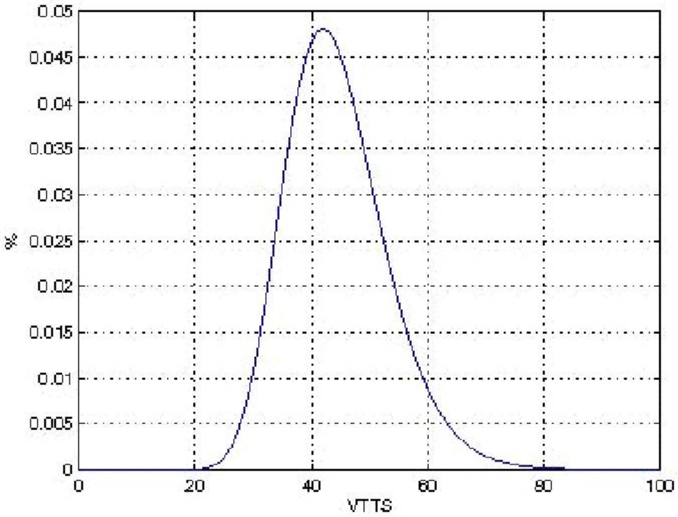

Figure 5.3: VTTS Distribution for empty and own account by ML

distribution and the time coefficient following a lognormal PDF.

$$VTTS_{HB} = \frac{1}{N(0.303, 0.027)} e^{2.207+0.256 loaded+0.196 hire+0.297 N(0,1)} \tag{5.18}$$

Although the ratio of a lognormal by a normal distribution is not a trivial analytical issue, we can use simulation to calculate the ratio of points the both distributions and then derivate the resulting distribution, taking in account the correlation among the coefficients (-0.0136). We used the trial version of @Risk to perform this simulation (Latin Hypercube sampling with 10000 iterations). The resulting distribution when both load and own account dummy variables are one is given in figure 5.4 and the resulting distribution when both load and own account dummy variables are null is given in figure 5.5. Figure 5.6 shows the distribution for the average values of load and own account dummy variables in the sample.

This result seems to be much more reliable than the previous since the solution obtained is quite far from the priors and it accommodates the variations of the utility of money in the sample.

Estimation results are given in table 5.4. TT is the travel time and TC is the travel cost; standard errors are given in parentheses. Note that the

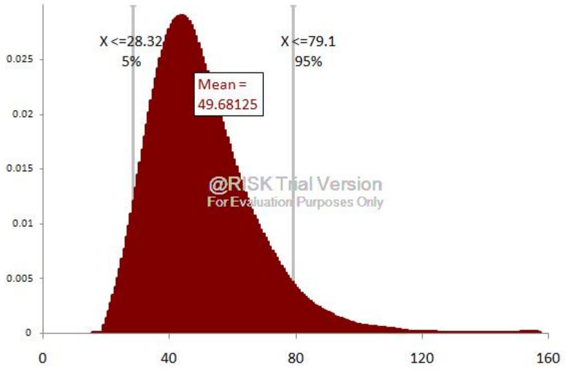

Figure 5.4: VTTS Distribution for loaded and hire by HB.

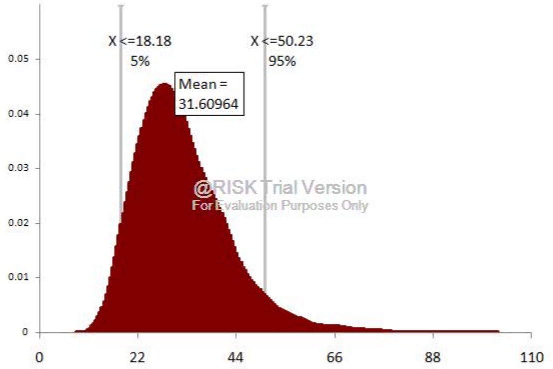

Figure 5.5: VTTS Distribution for empty and own account by HB.

log-likelihood for the Bayesian estimations is simulated, in order to be able to compare models in a single base.

5.8 Discussion

In line with many recent studies in this field, we faced here many difficulties in estimating the VTTS, especially when the mixed logit model is applied; we faced many convergence problems and even when convergence was achieved, the values provided were unrealistic. The Bayesian estimation provides a very attractive way of avoiding these optimization problems, accommodating both cost and time variables following PDFs, most in line with the theory.

Two points are of particular interest in our study. The differences between

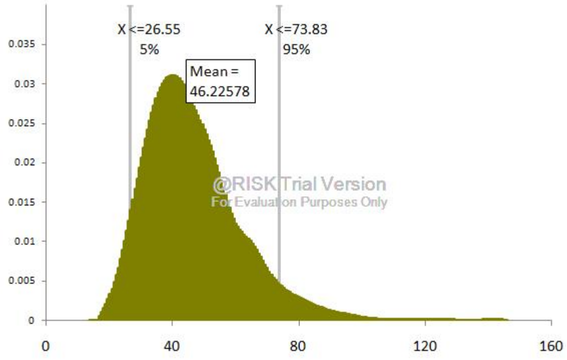

Figure 5.6: VTTS Distribution for average load and hire dummies by HB.

the constant and the distributed values of time in forecasting demand and the magnitude of the value itself.

It is easy to see that if the researcher believes that the average value of time is €52 (from model MNL(1)), but in fact it follows the distribution represented in figure 5.6 than instead of 50%, only 29,7% of the population will be willing to pay more than €52, leading to a rude demand overestimation. However if the atual values are given by the distribution in figure 5.6 but the researcher applies the current value used in France (€32) then most of users will actually be willing to pay more than this value, and the demand will be underestiamted.

Many recent results in the literature converge to the conclusion that using constant instead of distributed values of time tends to overestimate the demand. Two effects have to be isolated. First the skewness of the distribution. If the means are roughly the same, the constant value (or symmetrical distribution) will tend to overestimate the market share. Another point is whether the classic logic model and the distributed parameters model tend to produce different means. International experience suggests that this is not a general conclusion but depends on the nature of the data and specifications used in each study. For example Algers et al. (1998) and Gaudry et al. (1989) also found that more restrictive models lead to higher average values. However Amador et al. (2004) and Hensher (2001a,b) conclude that more restrictive models tend to underestimate the value of time; finally, other authors have found no significant differences between the values produced by different models (Train, 1998; Carlsson, 1999).

Table 5.4: Econometric results

		MNL	ML(1)	ML(2)	HB(1)	HB(2)
PDF TT		fixed	Lognormal	Lognormal	Lognormal	Lognormal
PDF TC		fixed	fixed	fixed	fixed	Normal
TT	Mean	0.4613	2.8797	1.9833	0.2941	-2.2069
		(0.0619)	(0.7202)	(0.2439)	(0.0041)	(0.9283)
	Sdt Dev		1.9933	1.8072	0.0059	0.2968
			(0.1566)	(0.1253)	(0.0011)	(0.0603)
TC	Mean	-0.01	-0.0017	-0.0238	-0.0302	0.3029
		(0.0031)	(0.0068)	(0.0065)	(0.003)	(0.0669)
	Sdt Dev					0.0267
						(0.0113)
Loaded		0.1004		0.1079	0.0833	0.2557
		(0.0216)		(0.0456)	(0.0243)	(0.0819)
Hire		0.163		0.2103	0.1746	0.1966
		(0.0203)		(0.0436)	(0.0215)	(0.071)
Intercept		-0.1347	-4.3057	-2.4119	-0.2703	-3.2377
		(0.0576)	(2.4688)	(0.3957)	(0.063)	(0.3002)
LL		-2467	-2359	-2338	-2529	-2242

(standard errors in parentheses)

Using wrong national standard values, of course, can lead to either over or underestimation. This point lead us to discuss the magnitude of the value of time in freight transport in France. Our results suggest that they should be reviewed upwards. Recent studies in other European countries have found similar results. Alvarez et al. (2007) found €64.1 in Spain, Fowkes et al. (2004) obtain values ranging from €55 to €200 in UK. We can conclude that the current standard French value can be on a downward bias.

5.9 Conclusions

The value of travel time savings is a fundamental concept in transport economics and its size strongly affects the socio-economic evaluation of transport schemes. Financial assessment of tolled roads rely upon the value of time as the main (or even the unique) willingness to pay measure. Values of time estimates, which primarily represent behavioural values, as then increasingly been used as measures of out-of-pocket money. In this setting, one of the main issues regarding the value of time is its distribution over the population.

Logit is by far the most applied discrete choice model used in estimations of values of time. Its popularity is due to its easy closed form. However, using a single value (representative of a mean or median) may lead to significant errors in evaluating the optimal toll and the revenue from a tolled road. In this perspective, the generalised used of logit models in the context of tolled infrastructure may lead to consequent traffic and revenue forecast errors.

The ambition of using distributed values for the parameters of discrete choice models associated with the recent progresses in hardware and software performances lead researchers to focus in more flexible structures. In this way a partial simulation partial closed form discrete choice model called mixed logit has been developed, allowing for distributed coefficients, estimated by simulated likelihood. In practice, however, the use of such models has been limited to cases where the kind of data associated with the choice of the distribution lead to model convergence and coherent results. Researchers and practitioners usually want to estimate lognormal distributed values of time, which in practice present convergence problems and tend to produce unacceptable high values for some share of the population. In this context, the use of constraints under the form of censure or caps for the standard deviation has been the solution find to overcome such problems. These constraints are then set according to the researcher's beliefs and prior works. The introduction of à priori knowledge is intrinsic to the econometric analysis. First, the analyst usually has some à priori about the result (i.e. one should expect that the value of travel time to be positive and to lay within a reasonable set) and second, the set of hypothesis, constraints and starting values of mixed logit models represent a priori hypothesis.

Bayesian estimations have some strong advantages compared to the classical techniques; they allow for distributed coefficients but the estimation does not require any maximization, rather, draws from the posterior are taken until convergence is achieved, avoiding convergence problems and sample sizes necessary to achieve the convergence are substantially smaller. Moreover, they can properly integrate a priori knowledge on the parameters.

In this chapter we present the main econometric models currently used for VTTS estimation. We apply these methods to the estimation of the value of travel time savings in freight transport in France. For this analysis a revealed preference study on two couple of tolled motorways and free roads was conducted. For the Bayesian estimation, we conjugate the data from this survey with the precedent studies guiding the current value used in France.

Estimations with mixed logit faced many difficulties, as expected. These difficulties could be avoided using the Bayesian procedures, providing also the opportunity of properly integrating a priori beliefs.

Results show that 1) using a single constant value of time, representative of an average, can lead to demand overestimation, 2) the estimated average value of time of freight transport in France is about €45, depending on the load/empty and hire/own account variables, which implies that 3) the standard value recommended in France should be reviewed upwards.