DOCTORAATSPROEFSCHRIFT

2009 | Interfacultair Instituut voor Verkeerskunde

Multilevel models in traffic safety research: An investigation and illustration of a flexible solution to improve upon classical statistical analysis techniques

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Verkeerskunde, te verdedigen door:

Ward G.M. VANLAAR

Promotor: prof. dr. Geert Wets

universiteit hasselt

INSTITUUT
VOOR MOBILITEIT

DOCTORAATSPROEFSCHRIFT

2009 | Interfacultair Instituut Verkeerskunde

Multilevel models in traffic safety research: An investigation and illustration of a flexible solution to improve upon classical statistical analysis techniques

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Verkeerskunde, te verdedigen door:

Ward G.M. VANLAAR

Promotor: prof. dr. Geert Wets

universiteit hasselt

imob

INSTITUUT
VOOR MOBILITEIT

D/2009/2451/55

"Queequeg was a native of Rokovoko, an island far away to the west and south. It is not down in any map; true places never are."

Herman Melville's 'Moby Dick'

Acknowledgements

Throughout my ten-year career in the field of traffic safety I have worked on my Ph.D., sometimes as part of my regular duties, sometimes above and beyond my regular duties, but not consistently as a full-time Ph.D. student in the strict sense of the word. I have worked on my Ph.D. predominantly from home and the office, either at the Belgian Road Safety Institute (IBSR), or later in my career at the Traffic Injury Research Foundation (TIRF) in Canada. This has certainly made it challenging to strike a balance between assuming the responsibilities of my regular job and completing this academic journey. On the other hand, it has also been rewarding because working on my Ph.D. as part of, or in combination with my regular job as a researcher facilitated the interaction and collaboration with some great people/researchers around the world. Without them, accomplishing this would not have been possible!

First, I am indebted to Prof. Dr. Geert Wets, Director at the Transportation Research Institute of the University of Hasselt, and supervisor of my Ph.D. I am grateful for the opportunity he provided me to graduate as a Ph.D., for his trust in me to successfully achieve this, and for his stimulating input and overall guidance while working on this project.

I also thank Prof. Dr. Tom Brijs, Program Coordinator of Traffic Safety at the Transportation Research Institute of the University of Hasselt, and co-supervisor of my Ph.D. Professor Brijs has been especially helpful in providing guidance with the Bayesian analyses I conducted as part of my Ph.D. and overall critical feedback when working on spin-off manuscripts from this thesis for publication in scientific journals.

I am grateful for the insightful, critical and grounding comments from each of the members of the reviewing committee, including Prof. Dr. Jean Shope, Associate Director of the University of Michigan Transportation Research Institute; Dr. James Hedlund, President of Highway Safety North, New York; Prof. Dr. Dimitris Karlis, Assistant Professor of Statistics at the Athens University of Economics and Business; Prof. Emeritus Jan Pauwels, Catholic University of Leuven; and, Dr. Elke Hermans, Spokeswoman of the Policy Research Centre of Mobility and Public Works, Traffic Safety Track, Transportation Research Institute of the University of Hasselt.

The support of several people at TIRF has been critical in successfully completing this undertaking. Dr. Herb Simpson, Research and Policy Consultant with TIRF and former President and CEO of TIRF has been instrumental in ensuring I could continue working on my Ph.D. after I moved to Canada and began working for TIRF in 2005. Without his endorsement it would not have been possible to continue moving forward in a productive way. Dr. Simpson created a nurturing and stimulating environment for me at TIRF enabling me to successfully

complete my Ph.D. He also provided much needed feedback and comments as papers were being prepared for publication in scientific journals.

I am grateful for the support of Dan Mayhew, Senior Vice President at TIRF and Kyla Marcoux, Research Associate at TIRF for their input in terms of feedback, comments, and technical support, especially when conducting the evaluation research of graduated driver licensing programs in North America.

Other reviewers provided useful comments on my work during the course of the past years and by doing so have helped me improve the quality of my work. They include René Mathijssen, Senior Researcher at the SWOV Institute for Road Safety Research in the Netherlands; Paul Boase, Chief of the Road Safety & Motor Vehicle Regulation Directorate's Road Users' Division with Transport Canada; Dr. Allan Williams, private consultant and former Vice President of Research at the Insurance Institute for Highway Safety in Virginia; Dr. Min Yang, visiting fellow at the Centre for Multilevel Modelling, University of Bristol and Senior Lecturer in Medical Statistics at the Wolfson Institute of Preventive Medicine, Queen Mary, University of London.

I express my gratitude to Laurie Baker, Manager, Road Safety Performance Planning and Analysis with the Insurance Corporation of British Columbia (ICBC), Canada for providing the British Columbia (B.C.) roadside survey data that I used in my thesis, and to Jean Wilson, Manager, Research and Evaluation at ICBC, who formulated feedback that was helpful in improving the quality of the research I did when using the B.C. data.

I also thank Kristel Hertogs who is with the Transportation Research Institute at Hasselt University and who provided logistical support to me. Her excellent ground support was especially useful given that I am based out of Ottawa.

Finally, a special note to my mom and my wife, Robyn. Mom, thanks for never giving up on me and helping to mould me into the man I have become. The values you instilled in me are key to the decisions I make every day. Robyn, you are my best friend, my soul mate and without your continued support at work and at home, I would not have been able to pull this off!

Abstract

This Ph.D. thesis deals with the issue of correct versus incorrect usage of statistics. More precisely, the objective is to demonstrate the applicability, usefulness and added value of multilevel models in the field of traffic safety. Multilevel models, also known as mixed models or random effects models are a family of techniques that can be considered a flexible solution to overcome some limitations of classical analysis techniques. To reach this objective, several sub-goals have been formulated. First, multilevel models are described using an intuitive and a mathematical approach. Second, the use of multilevel models is justified by explaining and illustrating the consequences of not using such models when it is necessary, or at least advisable to do so. It is explained that multilevel models are particularly useful for dealing with hierarchical or nested data, a type of data that is common across scientific disciplines including traffic sciences. Due to the nature of hierarchical data careful consideration is required with respect to the dependence of nested observations and context. Given the dependence of nested observations, ignoring the hierarchical nature of the data will lead to an underestimation of standard errors and an increased level of committing type I errors. Ignoring contextual aspects of hierarchical data typically results in an impoverished conceptualization of the research problem. The third sub-goal of this thesis is to apply multilevel models to a variety of traffic safety research topics and to illustrate their added value. Three case-studies are used to achieve this goal; they involve a case study on drinking driving, one on sleepiness among night-time drivers and one on the effectiveness of graduated driver licensing programs. In the first two case-studies a two-level logistic regression model is used to analyze the data; in the third case study a multilevel meta-regression analysis is carried out using both a 'frequentist' and a 'Bayesian' approach. It is demonstrated that multilevel models are particularly elegant and productive in generating new knowledge about each of the traffic safety issues of interest and that not using such models can lead to faulty conclusions. Finally, the last sub-goal is to define a research agenda. Based on the findings from the case-studies conclusions are formulated regarding the applicability, usefulness and added value of multilevel models in traffic safety research. Special attention is given to practical implications of the findings and their social relevance, both with respect to the applied methods and the actual findings from the analyses. As such, a research agenda for further studying these road safety issues is created.

Contents

List of figures

List of tables

1. Introduction

Today statistical methods are widely used across scientific disciplines. The development of powerful personal computers in the past decades has helped pave the way for statisticians and programmers to create a multitude of statistical software tools, which are widely available to the research community, either as freeware, shareware or at an affordable cost. Such statistical tools are used to summarize data and to help distinguish between random patterns and systematic ones.

As Mlodinow (2008) so eloquently describes in his book "The Drunkard's Walk. How Randomness Rules Our Lives" people tend to misunderstand random patterns. This is often explained from an evolutionary perspective: "if a starving caveman sees an indistinct greenish blur on a distant rock, it is more costly to dismiss it as uninteresting when it is in reality a plump, tasty lizard than it is to race over and pounce on what turns out to be a few stray leaves" (p. 26). In other words, because it was more productive for our ancestors to invest energy in what may turn out to be a false positive than to save energy on what could be a false negative, our brains have been wired during millions of years of evolution "to avoid the former mistake [i.e., dismiss the lizard] at the cost of sometimes making the latter [i.e., pouncing on a few stray leaves]" (p. 26). As a result, so the theory goes, humans like to make connections between different events and interpret them in terms of cause and effect, even in the absence of any meaningful causal or systematic pattern.

If humans are so prone to mistake randomness for a systematic pattern, novel statistical software tools that help us recognize both are invaluable. However, the aforementioned widespread availability of such invaluable statistical software tools that do help us recognize both randomness and systematic patterns comes at a cost, i.e., the possibility of incorrectly using these tools and related consequences. This thesis deals with this issue in the field of traffic safety. The issue of correct versus incorrect usage of statistics has been discussed before, e.g., by Hauer (1983a, 1983b, 2004), Hedlund (1984) and, more recently, by leading research agencies such as the Insurance Institute for Highway Safety (IIHS April 2006). As such, at a general level the topic of this Ph.D. thesis is not new. However, what is new in this thesis is the knowledge that is being generated by applying one particular family of such novel statistical tools, more precisely multilevel models, to a variety of road safety topics. It will be demonstrated in this thesis that multilevel models have been particularly elegant and productive in generating this new knowledge and that not using such models can lead to faulty conclusions.

While it is acknowledged that the world still is – and will continue to be – too complex to be captured completely in, and represented perfectly by statistical models, multilevel models are believed to facilitate modelling the world more truthfully than by using classical statistical analysis techniques. If applied properly, their results are closer to the truth than those coming from classical analysis techniques. In other words, multilevel models can be considered one possible solution to overcome some limitations of classical analysis techniques.

Multilevel models have come of age, especially in educational research. In their introduction to multilevel models Kreft and de Leeuw (2002) give a brief history of this family of techniques, emphasizing that developments similar to those occurring in educational statistics have been, or are occurring elsewhere. More precisely, the authors show that multilevel models are a conglomerate of known models, commonly used in different disciplines including:

- bio-medical sciences where the terms mixed-effects models and random-effects models are used (e.g. growth curve analysis in Lindsey 1993);
- economics (e.g. panel data research in Swamy 1971) and econometrics (e.g. Longford 1993) where the models are referred to as random-coefficient regression models;
- criminology (e.g. drug prevention research in high schools in Kreft 1994); and,
- geography (e.g. spatial analysis to study farms in counties in McMillan and Berliner 1994).

Nevertheless, multilevel modelling is relatively new to the field of traffic safety. This doctoral thesis investigates such multilevel models and their applicability in traffic safety research.

2. Objectives and outline

2.1 Objectives

The objective of this Ph.D. thesis is to demonstrate the applicability, usefulness and added value of multilevel models in the field of traffic safety. This doctoral thesis is predicated on the belief that multilevel models are a family of techniques that can be considered a flexible solution to overcome some limitations of classical analysis techniques.

Of considerable interest, the objective of this thesis is *not* to prove — in the strict sense of the word — that multilevel models are better than other statistical techniques. It is acknowledged that multilevel models have their limitations too, just as any other technique. The objective of this thesis is also not to develop a new statistical method.

As such, the applicability, usefulness and added value of multilevel models will be demonstrated in this thesis by conceptually studying this family of techniques, pinpointing its strengths, and by applying them in several case studies to a variety of road safety research issues. These multilevel models will be applied to those road safety research topics for the very first time and it will be demonstrated that the findings and related conclusions add to the literature. Without the use of multilevel models it would not have been possible to generate this knowledge, or at least not in an elegant and productive way.

To reach this objective, several sub-goals have been formulated.

First sub-goal: introducing multilevel models. Some multilevel models will be introduced using an intuitive and a formal — i.e., mathematical — approach. The models that will be described are the basic multilevel model and the ones that will be used later in the case studies (the binomial model and a meta-analysis model). The purpose of this sub-goal is to provide the reader with some basic building blocks prerequisite to understanding the approach that has been adopted in each of the case studies.

Second sub-goal: justifying the use of multilevel models. Once multilevel models have been introduced and described, their use will be justified by illustrating the consequences of not using a multilevel approach when necessary, or at least advisable.

Third sub-goal: applying multilevel models. A variety of multilevel models will then be applied to road safety issues and their added value will be illustrated. Three case-studies will be used to achieve this goal, more precisely one on drinking driving, one on sleepiness among night-time drivers and one on the effectiveness of graduated driver licensing programs.

Fourth sub-goal: defining a research agenda. Conclusions will be formulated regarding the applicability, usefulness and added value of multilevel models in traffic safety research. Special attention will be given to practical implications of the findings and their social relevance, both with respect to the applied methods and the actual findings from the analyses. As such, a research agenda for further studying these road safety issues will be created.

2.2 Outline

The outline of this doctoral thesis is as follows. To begin, multilevel models are introduced in chapter 3, first using an intuitive approach and then using a more mathematical approach. The basic two-level random intercept model (i.e., the variance components model) and the random slope model, as well as the two-level binomial model are described. A description of a meta-analysis multilevel model is also provided. Both the binomial model and the meta-analysis model will later be used in the case studies. This introduction further discusses how and why multilevel models can overcome some limitations of classical analysis techniques, using three empirical examples, one about seatbelt usage, one about sleepiness among night-time drivers and one about drinking driving. This introductory chapter deals with sub-goals one and two.

Following this introductory chapter, three different case studies are presented (sub-goal three). These are:

- Chapter 4, which contains a case study that investigates the relationship between traffic count (i.e., the number of vehicles driving by at a road site) and drinking driving;
- Chapter 5, which is about sleepiness among night-time drivers; and,
- Chapter 6, which evaluates the effectiveness of graduated driver licensing programs in North-America.

The structure of each of the case studies is the same. First, background information about the road safety topic and the objectives and rationale of the case study are provided in the introduction. The background information contains the information necessary to understand what the issue is about and provides an overview of the literature. The objectives and rationale are each discussed in a separate sub-section in the introduction and briefly state what the goals of the case study are and why multilevel models are being used. In a second section, the applied methods are described, followed by a third section about the results. Finally, these results are discussed in the last section of each case study.

Conclusions are formulated in chapter 7 and translated into a list of recommendations in chapter 8. This last chapter contains a research agenda with suggestions for further research (sub-goal four).

2.2.1 Data acquisition

Different data are used in this thesis. This section about data acquisition briefly describes those data and explains the role that the author of this thesis, Ward Vanlaar, has played in the process of collecting these data (note that analyzing these data was done exclusively by Ward Vanlaar). A detailed description of the data, including references, is provided in the appropriate sections of this thesis.

- Seatbelt use (section 3.3.3.2): These data were collected in 2004 as part of the Belgian Road Safety Institute's (IBSR) efforts to develop a set of indicators to measure levels of road safety in Belgium. The methodology to collect these data (i.e., research design, data collection instruments, procedure) was developed by Ward Vanlaar when he worked at the IBSR as the head of research in the policy and behaviour department. Ward Vanlaar was also in charge of negotiating a contract with the survey firm that collected the data and received guidance from his supervisor at the IBSR during the negotiations. Ward Vanlaar served as a conduit for the survey firm during data collection and was responsible for quality assurance (QA). As such, the data were collected under Ward Vanlaar's supervision.
- Sleepiness among drivers (section 3.3.3.2): These data were collected in 2003 in British Columbia (B.C.), Canada by the Insurance Corporation of British Columbia (ICBC). Ward Vanlaar was not involved in the collection of these data. Permission was obtained from ICBC to analyze the data.
- Drinking driving (section 3.3.4.2): These data were collected in 2003 as part of the IBSR's efforts to develop a set of road safety indicators for Belgium. The methodology to collect these data (i.e., research design, data collection instruments, procedure) was developed by Ward Vanlaar when he worked at the IBSR as the head of research in the policy and behaviour department. Ward Vanlaar was also in charge of negotiating data collection procedures with different police organizations in Belgium — also with guidance from his supervisor at the IBSR. Ward Vanlaar served as a conduit for police during data collection and was responsible for QA — e.g., Ward Vanlaar participated in many road checks across the country when the roadside survey took place. As such, the data were collected under Ward Vanlaar's supervision.
- Case study one — traffic count (chapter 4): The same data about drinking driving from section 3.3.4.2 were used in this case study, as well as the same data from B.C., used in section 3.3.3.2. These data from B.C. also contain information about drinking driving and traffic count in addition to information about sleepiness and have been used in this case study to replicate a model that had previously been built by Ward Vanlaar using the Belgian data.
- Case study two — sleepiness among night-time drivers (chapter 5): Data from 2003 (the same as in section 3.3.3.2) and from 2006 (a second wave of data, comparable to the

2003 data) from ICBC were used in this case study. Data for both datasets were gathered by ICBC; Ward Vanlaar obtained permission to use them.

- Case study three — An evaluation of graduated driver licensing programs in North America using a meta-analytic approach (chapter 6): The methodology to collect the data was developed by Ward Vanlaar. The actual data entry (i.e., data entry of the independent variables needed for the formal description of the different graduated driver licensing programs and of the population counts for the calculation of the relative risk ratios per jurisdiction) was done by Kyla Marcoux, Research Associate at TIRF with guidance from Ward Vanlaar. The data management and programming required to efficiently calculate the relative risk ratios and for the analyses was also done by Ward Vanlaar.

3. An introduction to multilevel models

This chapter introduces multilevel models, first using an intuitive approach and then using a mathematical approach. This introduction further discusses how and why multilevel models can overcome some limitations of classical analysis techniques, using three empirical examples, one about seatbelt usage, one about sleepiness among night-time drivers and one about drinking driving. Conclusions are formulated at the end of this chapter regarding the reasons why to use multilevel models.

3.1. An intuitive approach to multilevel models: Speculations about the relationship between the length and speed of cars[1]

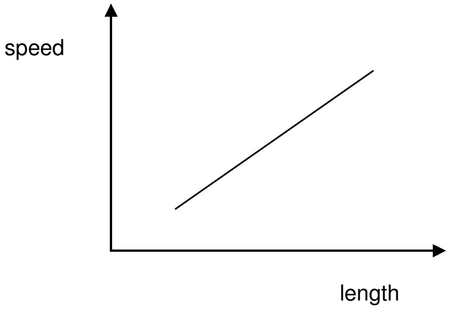
In this section a purely hypothetical example on the relationship between the length and speed of cars is used to introduce the concept of multilevel models. The simple hypothesis in this example is that there is a relationship between vehicle speed and vehicle length – longer vehicles will have more powerful engines and this will correlate with higher speeds. Note that this hypothesis is oversimplified for the sake of clarity in this example.

Assume a roadside survey is conducted to gather information on speeding. To obtain data on the speeding behaviour of individual drivers, a cluster sample is drawn — meaning clusters or groups of drivers are sampled at random rather than individual drivers. Road sites (i.e., the primary sampling units or PSU's) are randomly selected; then, at each road site, each driver's speed is monitored. Along with the speed of each vehicle, its length is also measured. This cluster sample can also be described in terms of a multilevel model, more precisely as a two-level model, with drivers at level one, nested in road sites at level two.

Figure 3.1 contains the results of a traditional regression analysis showing the relationship between length of cars as the independent variable and speed of cars as the dependent variable.

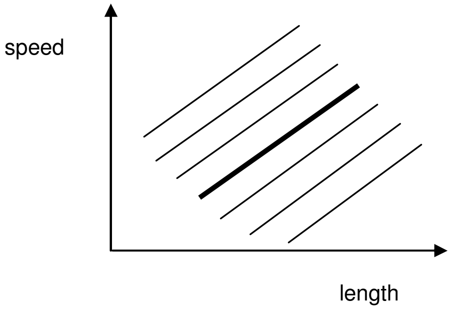
No distinction is made between road sites in this graph, and the information from all cars over all road sites is analyzed together using a classical – one-level – regression analysis. The slope of the regression line measures the increase in speed associated with a unit increase in length. As can be seen, if such findings were obtained from the study, the hypothesis would be confirmed – longer cars drive faster. This relationship does not, however, take account of such things as the road site at which the behaviour is being observed – i.e., the model assumes that road site does not influence the length/speed relationship, which is conceived only in terms of individual characteristics of cars (length and speed).

Figure 3.1: Overall relationship between length of cars and speed of cars



This is remedied in Figure 3.2, with each of the different road sites (six in this figure, with the bold line representing the "average road site") having its own speed/length relationship represented by a separate regression line.

Figure 3.2: Relationship between length of cars and speed of cars per road site, equal slopes



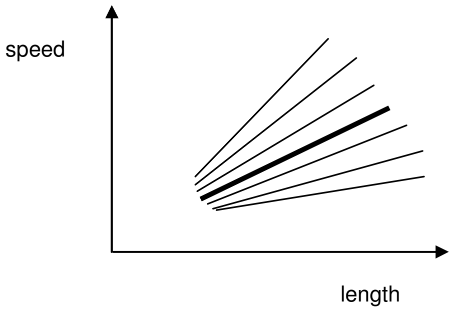
These parallel regression lines imply that while the speed/length relationship at each road site is the same (parallel regression lines having the same slope), some road sites have uniformly higher speeds than others. This might be easily explained, for example by the fact that the different sites were at locations with different speed limits or different roadway geometry. The lowest regression line could for example represent a road site with a speed limit of 30km/h, while the upper regression line could represent a road site with a speed limit of 120km/h.

However, the observed relationships might prove to be more complicated than this and two other possible relationships are illustrated in Figures 3.3 and 3.4. Each regression line, representing a road site, now has a different slope.

[1] This section is based on work done by Vanlaar (2007a, 2007b, 2007c) as part of the European SafetyNet project. For an overview of this project, also see Thomas et al. 2005.

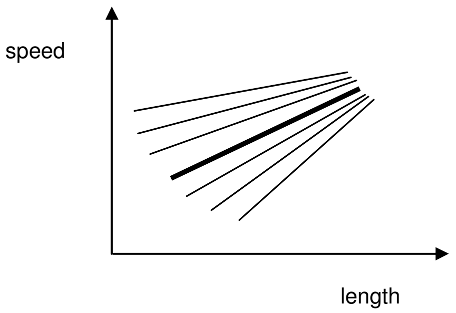
In Figure 3.3 the pattern suggests that the road site makes very little difference for the speed of short cars, but a considerable difference for the speed of long cars. This is called a fanning-out pattern. An explanation could be that the maximum speed of short cars is so low that they can only reach the lowest speed limit of 30km/h (e.g., the fleet of short cars in the town being studied is composed almost exclusively of electric cars). Long, powerful cars, however, can easily reach higher speeds leading to a more diverse speed pattern depending on the different existing speed limits at road sites.

Figure 3.3: Relationship between length of cars and speed of cars per road site, fanning out pattern



Conversely, the data might yield a situation represented in Figure 3.4, which shows relatively large road site-specific differentials for short cars but little variation for long cars. A possible explanation for this fanning-in pattern could be the attitude of drivers of long, i.e., powerful cars: they tend to speed regardless of the speed limit and, therefore, their speed distribution over different locations has a small range; on the other hand, drivers of short cars are more conscientious and tend to respect the speed limits resulting in a broad range of speeds among such vehicles.

Figure 3.4: Relationship between length of cars and speed of cars per road site, fanning in pattern



Using this oversimplified example on the relationship between the length and speed of cars facilitated introducing multilevel models from a conceptual point of view. A straightforward definition of multilevel modelling is given by Heck and Thomas (2000). According to their definition multilevel modelling refers to a variety of statistical methods that may be used to handle such data structures consisting of information available at different levels, i.e., at the level of drivers and at the level of road sites in this example. The next section provides a formal description of multilevel modeling using mathematical equations.

3.2 A formal description of some multilevel models[2]

3.2.1 The basic two-level random intercept and random slope model

3.2.1.1. Objectives

The objectives of this technique include the objectives of ordinary regression analysis, but in addition to that, there is also the objective of taking contextual information into account by letting the intercept and slope vary across road sites. According to Tacq (1997), the four objectives of traditional linear regression analysis are:

- To look for a function, which represents the linear association between the independent variables and the dependent variable better than any other function. This comes down to calculating a regression coefficient for each independent variable.
- To examine the strength of the relationship and to know which share of the variance of the dependent variable is explained by the variances of the independent variables together. This comes down to the calculation of the multiple correlation coefficient R and its square. While the concept of explained variance is well-known in traditional regression analysis, it is problematic in multilevel models according to Snijders and Bosker (1999).
- To investigate whether the associations found in the sample can be generalized to the population using significance tests.
- To examine which independent variable is most important in the explanation of the dependent variable using the standardized beta weights.

3.2.1.2 Model definition

In the case of a single-level bivariate model, i.e., the classical regression model, the general formal equation is:

[2] This section is based on work done by Vanlaar (2007a, 2007b, 2007c).

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i \quad (3.1)$$

Where

- subscript i signifies an individual respondent;
- y and x measure the response and predictor variables, in this example the speed and length of a car;
- β_0 and β_1 are fixed and unchanging parameters, namely the intercept and the slope; the former, when x is centered about its mean, represents the speed of a car of average length; the latter is the change in speed for an increase in length with one unit;
- e signifies the random part, which allows for fluctuations around the fixed part.

This equation is specified only at the micro-level of the individual. To build a multilevel model the *micro-model* has to be re-specified by including road sites using subscript j.

For the random intercept model this yields:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{oij} \quad (3.2)$$

There is one *macro-model* at the road site level:

$$\beta_{0j} = \beta_0 + u_{0j} \quad (3.3)$$

This macro-model allows for the differential road site intercept (u_{0j}) to vary from road site to road site around the overall intercept (β_0). The micro model is seen as a within-road site equation, while the macro model is a between-road site equation in which the parameter of the within model is the response (Jones 1993).

Both equations are combined to form the random two-level model:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \left(u_{0j} + e_{0ij}\right) \quad (3.4)$$

All the elaborations have come in the random part because in addition to allowing individual cars to vary, road sites have also been allowed to vary in having a differential speed for a car of average length. Such models in which the intercept is the only term allowed to vary at level two are commonly referred to as “variance components models” (Rasbash et al. 2004).

The formulas look as follows if the slope is also allowed to vary from road site to road site in addition to a random intercept. The micro model remains unchanged but there are now two macro-models at the road site level:

$$\beta_{0j} = \beta_0 + u_{0j} \quad (3.5)$$

$$\beta_{1j} = \beta_1 + u_{1j} \quad (3.6)$$

These two macro-models allow respectively for the differential road site intercept ($\mathrm{u}_{0\mathrm{j}}$) to vary from road site to road site around the overall intercept (β_0) and for the differential slope ($\mathrm{u}_{1\mathrm{j}}$) to vary around the overall slope (β_1) (Jones 1993). Again, the micro model is seen as a within-road site equation, while the macro models are two between-road site equations in which the parameters of the within model are the responses. Note that this is easy to see when using the notation with $\mathrm{e}_{0\mathrm{ij}}$ as part of the micro model as opposed to the macro model because then only the micro-model contains both subscripts i and j, referring to a within situation, while the macro-models then only contain subscript j, referring to a between situation.

All three equations are combined to form the fully random two-level model:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \left(u_{1j} x_{1ij} + u_{0j} + e_{0ij}\right) \quad (3.7)$$

As with the previously discussed random model all the elaborations have now come in the random part; in addition to allowing individual cars to vary, road sites are allowed to vary in having a differential speed for a car of average length, and a differential speed/length relationship (Jones 1993).

As with any other statistical distribution, and making the usual assumptions of normality, homogeneity and independence, these higher-level distributions can be summarized by measures of the centre, the mean, and spread around the centre, the variance. Relations between the slope and intercept distributions can be summarized by a measure of covariance. "Thus, the higher-level distributions can be summarized in terms of the fixed part (the means β_0 and β_1) and the random part (the variances $\sigma^2_{\mathrm{u}_0}$ and $\sigma^2_{\mathrm{u}_1}$, and the covariance $\sigma_{\mathrm{u}_0\mathrm{u}_1}$)." (Jones 1993: p. 253).

3.2.1.3. Heteroscedasticity

Multilevel models share with many traditional models the assumption that the residuals at each level are homoscedastic, i.e., have constant variance and covariances, and do not depend on the particular values of the explanatory variable(s) included in the model. This assumption is partially relaxed, however, once random slopes are specified in the model: variances at one or both levels are assumed to depend linearly or quadratically on one or more of the explanatory variable(s).

The following reasoning, borrowed from Snijders and Bosker (1999), and applied to the example on speed, illustrates this feature of multilevel models. In case of a fanning-in pattern (see Figure 3.4) a random slope for the effect of car length on speed would indicate that road sites affect the speed of short cars to a larger extent than the speed of long cars. This can be seen in Figure 3.4 as the lines representing the different road sites are farther away from one another at the lower values on the X-axis and closer to one another at the higher values on the X-axis. So at lower values on the X-axis (i.e., for short cars) there is much more variation in speed between road sites, compared to higher values on the X-axis (i.e., for long cars).

In other words, if you drive a short car, which road site you are at will matter a lot and may influence your speed considerably, while road site does not matter as much if you drive a long car. This means that road sites add a large component of variance to the speed of short cars, but little or nothing to the speed of long cars. Therefore, the intra-class correlation for short cars (also known as the Variance Partition Coefficient (VPC)), defined as the proportion of the total residual variation that is due to differences between groups (Goldstein, 2003), will be higher than the intra-class correlation for long cars. This implies that, once random slopes are specified in a model, the intra-class correlation or VPC cannot be uniquely defined any longer because this residual variation (due to differences between groups; road sites in this case) will vary as a function of the explanatory variable's values (short versus long cars in this example).

3.2.1.4 Model assumptions

The assumptions of the basic two-level model are (Snijders & Bosker 1999; Rasbash et al. 2000):

- $\mathrm{e}_{0ij} \sim \mathrm{N}\left(0, \sigma^2_{\mathrm{e}_O}\right)$, i.e., the level-one residuals are assumed to be Normally distributed, with mean zero and constant variance $\sigma^2_{\mathrm{e}_0}$;

- $u_{0j} \sim N\left(0,\sigma_{u_0}^2\right)$ and $u_{1j} \sim N\left(0,\sigma_{u_1}^2\right)$, i.e., the level-two random coefficients are assumed to follow a multivariate Normal distribution with mean zero and constant variance respectively $\sigma_{u_0}^2$ and $\sigma_{u_1}^2$;
- Random coefficients at level 1 (e_{oij}) and at level 2 (u_{0j}, u_{1j}) are assumed to be uncorrelated;
- $y_{ij} = N(XB, \Omega)$, i.e., the response variable is assumed to be Normally distributed, where XB is the fixed part of the model and Ω represents the variances and covariances of the random terms over all the levels of the data;
- The homoscedasticity assumption holds, i.e., the assumption that the variances and covariances estimated at the different levels of the data are constant, just as for many other statistical analysis techniques. However, in multilevel modelling, this assumption can be relaxed.

3.2.1.5 The Variance Partition Coefficient (VPC)

The VPC is the proportion of the total residual variation that is due to differences between groups (Goldstein, 2003), more precisely between road sites in this example. It is also referred to as the intra-class correlation (Snijders & Bosker, 1999), which measures the extent to which the y-values of the dependent variable of individuals in the same group resemble each other as compared to those from individuals in different groups. However, the former interpretation is the more usual one (Rasbash, 2004). The VPC is denoted by:

$$\frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_{e_0}^2} \quad (3.8)$$

3.2.1.6 Extending the model

So far a bivariate two-level model with continuous variables on level one has been considered. Two important extensions of this model will now be discussed. First, a model with a categorical predictor variable will be studied. Second, higher-level predictor variables and contextual effects will be considered.

According to Jones (1993) level-one categorical predictors can be specified in a model in which some or all of the predictors are categorical. A random intercept/random slope model with an independent variable with two categories is achieved by specifying a micro-model with one dummy variable (having a value 0 or 1). In this example the continuous independent variable length could for example be divided in two categories: short cars and long cars. The micro-model looks as follows:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{0ij} \quad (3.9)$$

and additionally two macro-models:

$$\beta_{0j} = \beta_0 + u_{0j} \quad (3.10)$$

$$\beta_{1j} = \beta_1 + u_{1j} \quad (3.11)$$

Another type of extension is to include higher-level variables in the model. Higher-level variables are also referred to as aggregate or ecological variables, or simply context (Snijders and Boskers 1999). They describe the higher-level structures in the dataset. This is achieved by including such variables in the relevant macro-models (Jones 1993). For example, if road site average speed is thought to be affected by traffic count at that road site (C), then the random intercept macro model can be re-specified to include an extra term, as in:

$$\beta_{0j} = \beta_0 + \alpha_1 C_j + u_{0j} \quad (3.12)$$

This main effect could for example mean that the average speed at a road site would decrease with increasing traffic count at that road site.

Similarly, the slope terms can also be related to traffic count at a road site.

$$\beta_{1j} = \beta_1 + \alpha_2 C_j + u_{1j} \quad (3.13)$$

This could be explained as follows. At road sites with a low traffic count the real relationship between length and speed is revealed and consists of a strong association between both variables in that a unit increase in length corresponds to a high increase in speed. At road sites with a high traffic count the real relationship does not emerge because cars are obstructed by one another. Therefore, a unit increase in length only corresponds to a small or no increase in speed. This formulation results in the introduction of an interaction term (the product of x and C) in the combined model. This is defined as a cross-level interaction term: interactions between variables measured at different levels in hierarchically structured data (Kreft and de Leeuw 2002):

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \alpha_1 C_j + \alpha_2 C_j x_{1ij} + \left(u_{1j} x_{1ij} + u_{0j} + e_{0ij} \right) \quad (3.14)$$

3.2.2 The two-level binomial model

3.2.2.1. Objectives

The objective of the two-level binomial model is to look for an appropriate function to model the relationship between a set of explanatory variables (this set can consist of continuous variables, categorical variables or a mixture of both types of variables) and the dependent variable. Specific to the binomial model is that the dependent variable is binary so the responses can only take the values of 0 or 1. The multilevel version of this model allows assigning the observed variance to different hierarchical levels.

3.2.2.2 Model definition

Models for binary data concern the probability π_{ij} that the observed variable y_{ij} from person i in group j takes the value 1 rather than 0. Given an underlying continuous variable, the binary model can be construed as a threshold model (Snijders and Bosker, 1999). For example, the threshold could be the legal blood alcohol limit in road traffic; if a driver's blood alcohol concentration (BAC) — a continuous variable — is equal to or greater than this threshold then the dependent binary variable is one; if BAC, however, is smaller than the threshold, the dependent binary variable is zero.

The model can then be written in terms of the underlying continuous variable y_{ij}^* .
Note that the asterisk is used as a symbol to denote the underlying continuous or latent variable, rather than the observed variable. In case the logit link function is chosen, the model then becomes:

$$y_{ij}^* = \beta_{0j} + \beta_1 x_{1ij} + e_{ij} \quad (3.15)$$

where

e_{ij} ~ logistic (0, $\pi^2/3$), with mean zero and variance $\pi^2/3$ =3.29 (in this case π does not denote a parameter but the number 3.141).

The advantage of constructing the model on the basis of an underlying continuous variable is that the level one errors can be assumed to follow the logistic distribution and therefore the error variance is known. More generally, binary data are assumed to follow the binomial distribution, whether they are derived from an underlying continuous variable (e.g., above/below average, severely injured/slightly injured, passed/failed, etc.) or not (e.g., male/female, yes/no, dead/alive, etc.). The model for logistic regression is based on this distribution.

Besides the logit link function for binary data, other link functions are available, including the probit or loglog functions. In this thesis the logit function will be used, meaning the analyses that are conducted are multilevel logistic regression analyses.

A two level logistic variance components model for binary responses as an equation for the probability π_{ij} is (Rasbash et al., 2004, p. 111):

$$\operatorname{logit}(\pi_{ij}) = \log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_{0j} + \beta_1 x_{1ij} \qquad (3.16)$$

$$\beta_{0j} = \beta_0 + u_{0j} \quad (3.17)$$

To interpret the relationship between the binary response and an explanatory variable, logit coefficients can be transformed into odds ratios using the exponential transformation. In the threshold model for drinking driving, these odds ratios compare the odds for drinking driving of a certain category of a variable (for example the odds for drinking driving for the category “female” of the variable “gender”) to the reference category of that variable (in this example with gender as the independent variable the reference category is “male”).

Taking the exponentials of each side of equation 3.16, we obtain:

$$\frac{\pi_{ij}}{1-\pi_{ij}} = \exp(\beta_{0j}) \times \exp(\beta_1 x_{1ij}) \qquad (3.18)$$

If we increase the independent variable 'x' by one unit, we obtain:

$$\frac{\pi_{ij}}{1-\pi_{ij}} = \exp(\beta_{0j}) \times \exp(\beta_1 (x_{1ij}+1)) = \exp(\beta_{0j}) \times \exp(\beta_1 x_{1ij}) \times \exp(\beta_1) \qquad (3.19)$$

Equation 3.19 is equal to equation 3.18, multiplied by the exponent of β_1. Therefore, increasing the independent variable with one unit corresponds to multiplying the odds ratio with the exponent of the coefficient of this independent variable. If the independent variable is binary (like gender), then the exponent of the coefficient of this binary variable is interpreted as the effect on the odds ratio, comparing the odds for units with x = 1 relative to the odds for units with x = 0, i.e., the reference category. More generally, if x is categorical, then the exponent of its coefficient is interpreted as the effect on the odds ratio, comparing the odds for units with a value for x , different from 0 (1, 2, 3, etc. depending on how many categories the categorical variable consists of) with x = 0, i.e., the reference category.

3.2.2.3. Model assumptions

The model assumptions for the binomial model, including the logistic model are:

$u_{0j} \sim N(0, \sigma_{u0}^2)$, the road-site-specific component of the intercept is assumed to be normally distributed with mean zero and variance σ_{u0}^2 .

$y_{ij} \sim Bin(1, \pi_{ij})$, the observed binary responses are assumed to follow the binomial distribution with expected value π_{ij} and variance $\pi_{ij}(1-\pi_{ij})$.

3.2.3 Meta-analysis using multilevel models

Previous sections focused on some basic multilevel models as well as the two-level binomial model. The basic one-level and the binomial one-level model are commonly used and well understood. Meta-analysis, however, is somewhat less well known and adopts a different approach. The following sections contain a brief introduction to meta-analysis, both as a one-level and multilevel model.

3.2.3.1 Objectives

The objective of a meta-analysis is to summarize all the available scientific information about a topic of interest and to quantify the findings in a meaningful way (see e.g., Egger and Smith 2007). A meta-analysis typically begins with a systematic review of the literature, more precisely of all the available evaluation studies of a particular type of program, for example graduated driver licensing programs (GDL). Once all the evaluation studies have been identified, obtained and scrutinized, a meta-analysis will be applied to summarize and statistically analyze the information from those evaluation studies. In other words, a systematic review typically summarizes all the information from a set of evaluation studies in a 'qualitative' narrative, whereas a meta-analysis takes it one step further and summarizes the information in a 'quantitative' fashion.

In order to facilitate this, a meta-analysis involves two main steps:

- calculating an outcome measure for each evaluation study on the same scale; and,
- translating the methodological features of the evaluation studies and features of the program that is being evaluated into variables.

The outcome measure (or dependent variable in the meta-regression) provides an indication of how effective the program was found to be according to each evaluation study. It is important that the same type of outcome measure be used for each of the evaluation studies in the meta-analysis so as to make the different evaluation studies comparable and to avoid comparing apples to oranges. For example, if one study calculated the effectiveness of GDL using a decrease in the per capita fatality rate and another study used a decrease in the absolute number of injuries, summarizing both types of information into one outcome measure may be problematic. The ability to address this challenge is an advantage of meta-analysis that controls for the different methodologies by comparing the different studies available "on the same scale", i.e., standardizing. Results of a meta-analysis regarding differences in effectiveness will, therefore, be more valid because they truly reflect differences in programs and program features by accounting for differences in evaluation methodologies (see Elvik 2005a and 2005b).

In the end, there will be a set of independent variables that describe the different evaluation studies both from a methodological point of view (e.g., studies with a control group versus studies without a control group; studies with random assignment to groups versus studies without; etc.), as well as in terms of different features of the program that is being evaluated (e.g., GDL programs with at least five components versus GDL programs with less than five components; the number of hours of supervisory driving in each GDL program; etc.).

The effectiveness of the program can then be estimated by summarizing the outcome measure. Also, the relationship between the independent variables and the dependent variable can be investigated using a variety of analysis techniques that resemble a classical regression analysis. The difference with the latter is that in the case of a meta-analysis the unit of analysis is not an individual, but a study. If the unit of analysis is an individual, the relationship between gender of a driver and the chances of drinking driving, for example, could be of interest. If the unit of analysis is a study in a meta-analysis, however, research questions will now focus, for example, on whether studies that evaluated GDL programs with at least five components found GDL to be more effective than studies that evaluated GDL programs with less than five components; or, whether studies of GDL programs that require more hours of supervisory driving found GDL to be more effective than studies of GDL programs with fewer hours of supervisory driving.

Sometimes the methods applied in the available evaluation studies will be too different and it becomes impossible to standardize the outcome measure. In this scenario, summarizing the information on effectiveness from every single evaluation study into one outcome measure can no longer be done in a meaningful way. An alternative approach is then to use raw data from each evaluation study to calculate the outcome measure in the exact same fashion for each of the studies, rather than to standardize the different outcome measures available from the evaluation studies. Actually, using raw data is considered to be the ideal approach because it enables researchers to make the results from the different evaluation studies much more comparable. As a consequence, differences in the outcome measure can now be attributed to differences in the program rather than differences in the applied evaluation methods with a much higher level of certainty. Why then, do researchers typically resort to using published results rather than raw data? The answer simply is that raw data are very hard, if not impossible, to come by. So even if the standard should be to use raw data in a meta-analysis, typically researchers use published results from different evaluation studies because of the logistical limitations related to trying to obtain raw data.

In the last case study of this Ph.D. thesis about the evaluation of GDL programs in North-America, it will be explained that there is too much heterogeneity among the available evaluation studies of GDL, and for this reason raw data are used. This case study is a rare example of a meta-analysis that uses raw data.

3.2.3.2 Model definition

Sections 3.2.1.2 and 3.2.2.2 provided a model definition for the basic and the binomial two-level models. This section provides such a definition for a meta-analysis model. The meta-analysis model used to summarize the data, expressed as a multilevel model (or random effects meta-analysis) is (see Hox and de Leeuw 2003):

$$d_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \ldots + \gamma_p Z_{pj} + u_j + e_j \qquad (3.20)$$

In equation 3.20 d_j denotes the observed outcome of study j (j=1,...,J); γ_0 is the intercept, i.e., the estimate for the mean outcome measure across all studies, or the summary effect; $\gamma_1 ... \gamma_p$ are the regression coefficients; $Z_{1j} ... Z_{pj}$ are the observed independent variables, i.e., study and program characteristics; u_j is the residual error term at the level of a study and its variance, σ_u^2, represents the true variation between studies and is assumed to have a normal distribution; and e_j is the residual error term representing the difference between the studies that is the result of sampling variation, which is determined entirely by the within-study variation and sample size, and is assumed to be known from the studies.

3.2.3.3 Model assumptions

The model assumptions are:

$u_j \sim N(0, \sigma_u^2)$, the between study variance is normally distributed with mean zero and variance σ_u^2 .

$e_j \sim N(0, \sigma_e^2)$, the difference between studies that is the result of sampling variation is normally distributed with mean zero and variance σ_e^2, which is determined by the within-study variation and sample size and is assumed to be known from the studies.

3.3 Justifying the use of multilevel models[3]

Three examples are discussed in this section to illustrate what happens when multilevel models have not been used when they should have been used. In order to poignantly illustrate this, first multilevel models are conceptually studied so it can be explained what sets them apart from other models.

3.3.1 Hierarchies

Today several introductory books are available on the market (e.g., Goldstein 2003; Heck and Thomas 2000; Hox 2002; Kreft and de Leeuw 2002; Snijders and Boskers 1999) and each of those defines multilevel models in a specific way. However, these definitions share one concept in particular, namely the concept of hierarchies or nested data structures: "We have variables describing individuals, but the individuals also are grouped into larger units, each unit consisting of a number of individuals. We also have variables describing these higher order units." (Raudenbush and Bryk 2002: p. xix). The individuals are also referred to as micro-units, while the larger units are called macro-units (Tacq 1989).

Hierarchies are very common in the social and the behavioural sciences and often occur naturally: e.g., pupils in classes, classes in schools; employees in departments, departments in firms; suspects in courts; offspring within families. Less obvious examples of hierarchies are observations nested within subjects (repeated measurements) or observations nested in studies (meta-analysis). Leyland and Goldstein (2001) give a rather extensive overview of more advanced applications of multilevel models including repeated measurements, binomial regression, Poisson regression, multivariate models, non-hierarchical structures, spatial analysis, meta-analysis and survival data modelling.

In the field of traffic safety nested data structures can be seen in a variety of different contexts and settings, for example: data on roadside surveys (drivers nested within police checks or locations, police checks or locations nested within regions; e.g., Vanlaar 2005a); on accidents (drivers and passengers in vehicles, vehicles in accidents, accidents in regions; e.g., Jones and Jørgensen 2003); on repeated measurements (e.g., Burns et al. 1999); meta-analysis (e.g., Delhomme et al. 1999; van Driel et al. 2004); etc.

[3] This section is based on Vanlaar (2005a) and (2005b).

3.3.2 Dependence and context

When analyzing nested data structures some conceptual issues calling for a proper approach have to be borne in mind. According to Rasbash et al. (2004: p. 6) "the point of multilevel modelling is that a statistical model explicitly should recognize a hierarchical structure where one is present: if this is not done then we need to be aware of the consequences of failing to do this." In this section, using multilevel modelling techniques as opposed to less sophisticated techniques is justified by means of three empirical traffic safety examples.

Broadly speaking there are two important consequences of ignoring a hierarchical structure: underestimation of standard errors leading to an increased level of committing type I errors, i.e., rejecting the Null Hypothesis when it should not be rejected (Rasbash et al. 2004) and problems related to an impoverished conceptualisation (Raudenbush and Bryk 2002). The first problem is related to the dependence of nested observations while the second problem stems from the existence of variables on different levels of aggregation, describing the micro-units and macro-units and from possible interactions between those different kinds of units. Variables related to macro-units are also referred to as contextual information or context of the micro-units.

The issue of dependence of nested observations has also been recognized in sample survey research and is referred to as clustering effects. Complex sampling designs are developed to model the hierarchical population structure as truthfully as possible in terms of geography or administrative structures. Elaborate procedures are available to analyze data gathered within such sampling designs (Cochran 1963; Kish 1965; Levy and Lemeshow 1999). According to Goldstein (2003: p. 5), however, such procedures usually have been regarded as necessary while they have not generally merited serious substantive interest. "In other words, the population structure, insofar as it is mirrored in the sampling design, is seen as a 'nuisance factor'. By contrast, the multilevel modelling approach views the population structure as of potential interest in itself, so that a sample designed to reflect that structure is not merely a matter of saving costs as in traditional survey design, but can be used to collect and analyse data about the higher level units in the population."

3.3.3 Dependence of nested observations

3.3.3.1 Consequences of ignoring dependence of nested observations

Dependence of observations plays an important role in nested data structures. An assumption made by most statistical analysis techniques that ignore hierarchies is the independence of observations: one observation is supposed to be sampled independently of the other observations. However, observations that are close in time or space are likely to be more similar than observations that are not close in time or space (Kreft and de Leeuw 2002).

Nested data structures are close in time or space by definition, which makes it reasonable to assume that observations within a hierarchical data structure will not be sampled independently from one another. Pupils nested in the same class will be influenced by the same teacher and hence be more alike than pupils from another class. Drivers nested within a certain speed zone are more alike than drivers in another speed zone in that their speed behaviour will be influenced – within certain limits – by the speed limit in that zone. Although speed limits are frequently violated, they do lead to similar behaviour to a certain degree and hence, to dependent observations.

Ignoring the dependence of observations generally causes standard errors of regression coefficients to be underestimated (Rasbash et al. 2004). The mechanism leading to this underestimation is easily explained as follows (Snijders and Boskers 1999). Imagine an extreme case of 10 groups of 100 identical observations each. Applying an ordinary regression analysis to the data leads to the calculation of standard errors based on 1,000 observations. However, since each group contains 100 totally dependent observations, the useful information in the sample really is limited to only 10 observations. Obviously the standard errors will be much greater based on 10 observations, indicating less precision than in the case of 1,000 observations. In reality observations are more likely to be similar to a certain degree instead of being identical. How similar they are exactly is measured by the intra-class correlation (see 3.2.1.5; Snijders & Bosker, 1999).

Multilevel modelling is capable of dealing with the issue of dependence of observations and thus calculates correct standard errors, taking account of the degree of dependence of the observations in the sample under study. It could be argued that a simpler approach to model a two-level sampling design, for example, with cars nested in road sites, would be to include dummy variables for road sites in a one-level model. The model would then contain n-1 dummy variables, with n equal to the number of road sites. This approach is not as efficient as multilevel models since the number of dummy variables can easily increase to a level that may not be manageable. In this regard, some epidemiological studies include several hundred PSUs (road sites in this example). For example, in a road site survey of drinking driving in Belgium (Vanlaar 2005a) there were 413 road sites, so a total of 412 dummy variables would have to be included in the one-level model. This could easily lead to estimation issues: reaching convergence with certain estimation procedures may be problematic with that many parameters to estimate. Also, such an approach could hardly be called parsimonious: trying to summarize and correctly interpret any pattern that emerges from those 412 dummy variables may be very challenging. As a consequence, multilevel models that summarize the variance at different levels provide a more powerful and efficient means of summarizing data.

As a sidebar, comparable issues regarding type-I errors can occur with other types of data such as time series and count data. For example, Commandeur and Koopman (2007) explain that it is not wrong in principle to use classical regression analysis to model time series, but it leads to problems when testing for significance. With respect to count data, Long and Freese (2006: p. 349) argue that "Although the linear regression model has often been applied to count outcomes, this can result in inefficient, inconsistent, and biased estimates". A discussion of type-I error issues with time series and count data is outside the scope of this thesis.

3.3.3.2 Two empirical examples: Seatbelt use and sleepiness among drivers

Table 3.1 contains the results of observational data regarding seatbelt use in Belgium in 2004. Data from this table were gathered during a roadside survey in 2004 according to a stratified two stage cluster sample. The first stage of the roadside survey consisted of randomly selecting road sites (m=149) in each region using a Geographical Information System (Arcview). The second stage of the roadside survey consisted of observing drivers and front seat passengers with regard to seatbelt behaviour (n=21,785). A single level and a two-level binomial model were fit with drivers and front seat passengers at level 1 and road sites (the PSU's) at level 2 in the two-level model.

Table 3.1: Comparison of logit coefficients and standard errors (S.E.) of a single-level and a two-level model regarding seatbelt use

Parameter	Single-level logistic model		Two-level logistic model	
	Logit coefficients	S.E.	Logit coefficients	S.E.
Fixed parameters				
Intercept	0.883	0.169	0.776	0.184
Passenger	-0.260	0.130	-0.205	0.132
Male	-0.663	0.121	-0.670	0.114
Wallonia	-0.454	0.158	-0.510	0.182
Brussels	-0.583	0.137	-0.365	0.140
50km/h	0.648	0.137	0.649	0.171
70km/h	0.921	0.171	0.665	0.155
90km/h	0.461	0.159	0.433	0.191
120km/h	0.795	0.173	0.811	0.188
Weekday night	-0.092	0.214	0.037	0.156
Weekend day	-0.091	0.142	0.151	0.139
Weekend night	0.312	0.156	0.197	0.166
Random parameters				
Level 2 variance: Ω_u	not applicable	not applicable	0.197	0.039
Level 1 variance: Ω_e	1.000	0.000	1.000	0.000

The explanatory variables are Passenger (a dummy variable indicating whether the observed subject was a front seat passenger or a driver with the latter being the reference category), Male (a dummy variable indicating whether the observed subject was male or female with the latter being the reference category), Region (a categorical variable consisting of 2 dummy variables indicating whether the observation took place in Flanders, Wallonia or Brussels with Flanders being the reference category), Speed Limit (a categorical variable consisting of 4 dummy variables indicating the speed limit of the road site where the observation took place; 30km/h is the reference category) and finally the variable Time Span (a categorical variable consisting of 3 dummy variables indicating in what time span the observation took place: Weekday as reference category [week peak hours and week off-peak hours are merged into weekday], Weekday night, Weekend day or Weekend night).

Even though the significance levels of most variables in both the single-level and the two-level model remain unchanged, there are two variables in particular that are interesting when comparing the single-level model, which ignores the hierarchical structure in the data with the two-level model, which acknowledges this structure. Those two variables are Passenger and Weekend night. Both variables are significant at the 5%-level in the single-level model, which can be derived from the logit coefficients since they are more than twice as large as the standard error. A more thorough test, namely a Wald test (Goldstein 2003) confirmed these findings (see Table 3.2). However, these effects are no longer significant according to the two-level model. The p-value of the variable Passenger in Table 3.2 shifts from a significant p-value of 0.046 in the single-level model to a non-significant p-value of 0.121 in the two-level model, while the p-value of the variable Weekend night increases from the significant value of 0.045 to a non-significant value of 0.233.

Table 3.2: Results of the Wald test for the variables Passenger and Weekend night in the single-level and the two-level model

Variable	Single-level logistic model		Two-level logistic model	
	Wald test		Wald test	
Passenger	Joint chi square test	3.989	Joint chi square test	2.402
	Degrees of freedom	1	Degrees of freedom	1
	p-value	0.046	p-value	0.121
Weekend night	Joint chi square test	4.015	Joint chi square test	1.424
	Degrees of freedom	1	Degrees of freedom	1
	p-value	0.045	p-value	0.233

This example illustrates the consequences of ignoring dependence of observations: the significance levels of both variables in the single-level model falsely led us to believe that these two variables are significant while this particular dataset does not contain the evidence to sustain this. The single-level model can therefore lead to erroneous conclusions regarding variables that could have an impact on seatbelt use and ultimately, on increasing the level of traffic safety. Based on the significant negative coefficient of front-seat passengers compared to drivers in the single-level model (meaning that the odds of front-seat passengers' seatbelt use are lower than those of drivers) it could for example have been decided to make front-seat passengers a special target group in a mass media campaign. However, in reality – as demonstrated in the two-level model – there may not be a significant difference in seatbelt use between those two groups.

It warrants mentioning that this illustration could be considered to be not overly convincing because both effects were barely significant in the single-level model to begin with (p-values of 0.046 and 0.045 respectively). This is true, but the 5%-threshold against which significance is evaluated is totally arbitrary, ergo the observation that both p-values are so close to this threshold could also be considered arbitrary. In other words, with a different significance threshold, for example 6%, 8% or 10%, there would be no discussion whatsoever that both effects are significant in the single-level model and they would also become non-significant in the two-level model. The value of this illustration really lies in the fact that the p-values change when going from the single-level model to the multi-level model, more precisely they increase and they do so substantially.

Further to this issue, it could also be argued that there may be a true effect anyway given that the direction of the effects does not change when shifting from one model to the next but that there is not sufficient power to detect this effect. This is true too, but given the lack of power it cannot be confirmed with these data whether there truly is an effect or not. The only conclusion that can be drawn is that there may be such an effect or there may not be such an effect and more data are needed to uncover which hypothesis is true. In other words, the sample is not powerful enough to reject the Null hypothesis that says there is no effect at all.

In this regard, it could be argued that the increased complexity of the two-level model over the single-level model has served to decrease the power of the sample given the fixed sample size. Snijders and Bosker (1999: p. 140) for example state that "A relevant general remark is that the sample size at the highest level is usually the most restrictive element in the design. For example, a two-level design with 10 groups…is at least as uncomfortable as a single-level design with a sample size of 10." Bearing this remark in mind it can indeed be argued that going from a single-level model with 21,785 respondents in the above illustration regarding seatbelts to a multi-level model with only 149 road sites — i.e., units at the highest level — decreases sample power. However, rather than detracting from the rationale to justify the use of multilevel models in this section, this argument most poignantly illustrates the point: given a model that more truthfully captures reality — in this case the multi-level model — it may no longer be possible to reject the Null hypothesis with a certain degree of certainty. The cost of modeling a more complex or more sophisticated model is loss of power; the currency to pay for this cost is a greater sample size.

This same issue of how the sampling design can affect the analysis outcomes is illustrated with a second example. These effects, also called design effects, are described using data from an epidemiological study on subjective sleepiness. Surveys for this study took place in June 2003 at 16 sites in each of three communities in south-western B.C. (see also Wilson et al. 2006 and Beirness et al. 2007 for a description of the methods). In 2003, a total of 2,627 vehicles were selected from the traffic flow. Road site and community were identified as sampling units. The model that accounts for the sampling design contains drivers, nested in road sites (road sites are secondary sampling units); and, road sites, nested in communities (communities are the PSU's).

The dependent variable in this study is self-reported sleepiness; it is a binary variable distinguishing between either being wide awake (value 0) or being somewhat to very sleepy (value 1). The independent variables include: age in six categories (16-18 years, 19-25 years, 26-35 years, 36-45 years, 46-55 years, older than 55 years; the latter category is the reference category); gender, with women being the reference category; BAC in four categories (zero BAC, 0.005g/dL-0.049g/dL, at least 0.05g/dL, test refusal, i.e., those ones who refused to take a breath test, namely 12% in 2003 and 8% in 2006; zero BAC is the reference category); passenger configuration (driver only; family; one passenger, gender different from driver; one passenger, same gender as driver; more passengers, gender different from driver; more passengers, same gender as driver; driver only is the reference category); trip origin (work, friend, restaurant, bar, home, other; home is the reference category); and, interview time (21:00-22:29, 22:30-23:59, midnight-1:29, 1:30-3:00; the latter category is the reference category).

Table 3.3 compares the results of the logistic regression analysis, which ignores the three stage cluster sampling design, with these of a logistic regression analysis that models the data according to such a three stage cluster design. For both models, odds ratios and p-values are reported for each of the independent variables. All the effects in Table 3.3 can be interpreted in a similar fashion and explain how belonging to a particular category of one of the independent variables influences the level of self-reported sleepiness among respondents, compared to the reference category of that independent variable. As such, both models suggest that being younger than 55 increases your chances of feeling sleepy because the odds ratios for any age category in both models are larger than one compared to the reference category of 55+. According to model 2, however, only the youngest age category of 16-18 years old is significantly more likely to report feeling sleepy, compared to the reference category of 55+, while all the other age categories are not.

The point of this comparison is, again, to illustrate how p-values become deflated when no correction for design effects is carried out. As can be seen in Table 3.3, out of the 12 effects that were significant according to the traditional approach (model 1), only three effects remain significant after correction for the design effects (model 2). This means there are really only three effects in this dataset for which there is evidence to reject the Null hypothesis of no effect. As for the other nine effects, this exercise does not imply that there is necessarily no relationship between the involved independent variables and dependent variable. It simply implies that this particular dataset does not provide evidence to reject the Null hypothesis of no relationship. More concretely, there may be a relationship, for example between BAC and subjective sleepiness, but this dataset and the applied model are not powerful enough to reject the Null hypothesis, stating that there is no relationship between BAC and subjective sleepiness. The same holds true for the other independent variables age, gender, passenger configuration, origin and time.

Table 3.3: Odds ratios, standard errors (S.E.) and p-values of the effects between the independent variables and self-reported sleepiness for the model that does not account for the sampling design (model 1) and the model that does (model 2)

	Model 1		**Model 2**	
	Odds Ratio (S.E.)	**p-value**	**Odds Ratio (S.E.)**	**p-value**
Age				
16-18	2.97 (0.77)	0.000	3.40 (1.17)	0.007
19-25	2.91 (0.63)	0.000	2.56 (1.34)	0.111
26-35	2.68 (0.59)	0.000	2.47 (1.31)	0.128
36-45	3.10 (0.69)	0.000	2.65 (1.39)	0.102
46-55	2.31 (0.55)	0.000	1.72 (0.74)	0.238
Gender				
Male	0.69 (0.07)	0.000	0.80 (0.15)	0.247
BAC (g/dL)				
0.005-.049	1.42 (0.24)	0.035	1.62 (0.40)	0.090
0.050+	0.91 (0.22)	0.705	0.99 (0.29)	0.984
Refusal	1.35 (0.28)	0.149	1.37 (0.22)	0.081
Passenger configuration				
Family	0.66 (0.16)	0.080	0.75 (0.21)	0.329
1 pass diff sex	0.83 (0.11)	0.152	1.01 (0.21)	0.955
1 pass same sex	0.56 (0.09)	0.000	0.55 (0.07)	0.002
Group diff sex	0.89 (0.17)	0.551	1.06 (0.17)	0.714
Group same sex	0.49 (0.14)	0.010	0.38 (0.10)	0.005
Trip origin				
Work	1.26 (0.21)	0.157	1.32 (0.29)	0.242
Friend	1.32 (0.20)	0.065	1.19 (0.16)	0.257
Restaurant	0.99 (0.20)	0.955	0.97 (0.28)	0.916
Bar	0.92 (0.26)	0.772	0.99 (0.42)	0.983
Other	1.21 (0.19)	0.207	1.17 (0.31)	0.584
Interview time				
21:00-22:29	0.50 (0.07)	0.000	0.63 (0.15)	0.097
22:30-23:59	0.62 (0.09)	0.001	0.67 (0.14)	0.093
Midnight-1:29	0.73 (0.10)	0.024	0.82 (0.10)	0.135

Despite reasonable arguments such as the possible lack of sleep among younger people as an explanation for subjective sleepiness or experimental findings such as the exacerbating effects of alcohol on feelings of sleepiness (e.g., Horne et al. 2003), data from the 2003 epidemiological study cannot be said to have validated these arguments or findings. More epidemiological research using a more powerful research design is necessary to consolidate such claims.

In conclusion, it is clear that design effects due to a complex sampling design can potentially affect the analysis outcomes, more precisely the standard errors and p-values. This warrants a more conservative analysis approach that is capable of dealing with such design effects. One such solution is multilevel models.

3.3.4 Contextual information of nested observations

3.3.4.1 Consequences of impoverished conceptualisation of contextual information

Contextual analysis is a development in the social sciences, which has focussed on the effects of the social context on individual behaviour (Snijders and Boskers 1999). The second consequence of ignoring a multilevel structure, related to contextual information, is illustrated with the frog pond theory, (Hox 2002: p. 6), "which refers to the idea that a specific individual frog may either be a small frog in a pond otherwise filled with large frogs, or a large frog in a pond otherwise filled with small frogs." These interactions between variables measured at different levels in hierarchically structured data are called cross-level interactions (Kreft and de Leeuw 2002). For example, applied to traffic safety, this metaphor points out that the effect of an explanatory variable like willingness to take risks on the dependent variable speed choice may depend on the average speed of other drivers at a certain location. A moderate risk taker in a speeding environment may thus become a dangerous speeder, while the same driver in a more law-abiding context may respect the speed limit rigorously.

The metaphor clearly illustrates that relationships between variables are not always easily modelled in a simplified way. Failing to acknowledge the complexity of certain problems, for example because of statistical limitations, might induce impoverished conceptualisations of the research problem. A landmark in this regard according to Snijders and Boskers (1999) is the paper by Robinson (1950) about the ecological fallacy, meaning that a correlation between macro-level variables cannot be used to make assertions about micro-level relations. This means for example that one cannot draw conclusions about the relation between individual age and individual odds for having a traffic accident based on a statistical model relating the proportion of young drivers in a geographical region with the proportion of accidents in that region.

Research problems in social and behavioural science often involve relationships between micro-level and macro-level variables and cross level interactions between those different variables. Those complex problems simply cannot be solved with aggregated or disaggregated analyses, which are bound to lead to erroneous conclusions. Multilevel analysis, however, overcomes these obstacles in an elegant and productive way. This technique allows researchers to translate a research problem into a design reproducing a lot of the nuances and without giving in too drastically towards simplifying the nature of the issue under evaluation.

3.3.4.2 An empirical example: Drinking driving

Table 3.4 contains an illustration of the flexibility of multilevel models with respect to conceptualization of the problem using empirical data regarding drinking driving (Vanlaar 2005a).

Data were gathered during a drinking driving roadside survey in 2003 using a stratified two stage cluster sample, including 413 road sites and 11,186 drivers. The methodology strongly resembles the seatbelt roadside survey methodology from the example in the previous section.

The outcome variable is a binary variable based on the BAC of each driver. For the purpose of the multilevel analysis it has been recoded with 0 representing those drivers with a BAC below the legal limit (which is equal to 0.05% in Belgium) and 1 representing those drivers with a BAC at or above the legal limit. Drivers at or above the legal limit are referred to as drinking drivers in this example.

The individual explanatory variables (level 1 explanatory variables) are Gender, Age (a categorical variable consisting of the following age groups: 16-25, 26-39, 40-54, 55+), Previously (a binary variable distinguishing between drivers who previously have been stopped and tested at a road site at least once and drivers who have never been stopped and tested at a road site before) and Probability (a categorical variable representing the driver’s perception of the probability of being tested for drinking driving; drivers could answer: very low, low, medium, high, very high).

The aggregated explanatory variables (level 2 explanatory variables) are Traffic count (a continuous variable indicating the total number of vehicles driving by the road site during the police check) and Intensity (a continuous variable calculated by dividing the number of policemen per road site by traffic count for that road site) — note that this latter variable was included primarily to keep intensity constant to disentangle its effects from the actual effect of traffic count.

Of particular interest is the influence of the variables Gender and Traffic count on the outcome variable. The relationship between Gender and the outcome variable is a nice illustration of the frog pond theory. A cross-level interaction would exist if the influence of Gender on Odds for drinking driving would change according to different values of Traffic count. However, this cross-level interaction effect was found not to be significant according to a Wald test (joint chi square test=1.706, degrees of freedom=1, p-value=0.192).

Table 3.4: Logit and exponential coefficients for the fixed and random effects of the binomial two-level logistic model

Parameter	Binomial model	
	Logit coefficients (S.E.)	Exponential coefficients
Fixed parameters		
Intercept	-4.757 (0.285)	
Traffic count	-0.002 (0.000)	0.998
Intensity	0.896 (0.383)	2.450
Gender (female vs. male)	-1.375 (0.207)	0.253
Previously	0.409 (0.141)	1.505
Probability low	0.537 (0.167)	1.711
Probability medium	0.744 (0.169)	2.104
Probability high	0.312 (0.278)	1.366
Probability very high	1.432 (0.290)	4.187
Age26-39	0.710 (0.242)	2.034
Age40-54	1.314 (0.234)	3.721
Age55+	0.863 (0.272)	2.370
Random parameters		
Level 2 variance: Ω_u	0.991 (0.197)	
Level 1 variance: Ω_e	1.000 (0.000)	

The relationship between Traffic count and the outcome variable is also relevant. According to the binomial two-level model there is a negative relationship between Traffic count and the odds of drinking driving when controlling for intensity of stopping drivers and for the other independent variables. This relationship is significant according to a Wald test (Goldstein 2003; joint chi square test=10.464, degrees of freedom=1, p-value=0.001). For each additional car at a road site the odds of drinking driving are multiplied by a factor of 0.998. This means that the odds of drinking driving decrease by 0.2%, or, per 100 extra cars on a site, the odds are multiplied by a factor of 0.819 (exp(-0.002x100)), meaning that the odds of drinking driving decrease by 18.1%.

In practice this means that police officers should not restrict their enforcement activities to sites where the frequency of vehicle traffic is high. One could argue that this relationship is of a spurious nature caused by the fact that drinking driving takes place primarily on weekend nights with low traffic while there are less drinking drivers during the day when there is much more traffic. Therefore another series of analyses per time span was performed to rule out this explanation. The result confirmed previous findings regarding the negative relationship between traffic count and odds for drinking driving.

Strictly speaking the latter example is not an illustration of a cross-level interaction. Nevertheless, this relationship between an aggregated explanatory variable at level 2 (i.e., road sites) and an individual dependent variable at level 1 (i.e., drivers) does illustrate the relevance of statistical models enabling the examination of the relationships between variables at different levels of aggregation.

Without a technique capable of simultaneously modelling variables at micro- and macro-levels such a relevant research question about the influence of traffic count on drinking driving behaviour may remain unanswered or it may be answered incorrectly, due to a wrong or impoverished conceptualisation of the problem. This example and others of how classical analysis techniques are limited in terms of conceptualization will be discussed in more detail with several case studies.

3.3.5 Summary and conclusion

This chapter illustrated the strengths of multilevel models and what can happen if data patterns that call for a multilevel approach are ignored. In summary, the reasons for using multilevel models are:

- Multilevel models can easily deal with design effects of a sample by taking account of the sampling design when analyzing the data. It could be argued that other types of models exist that also can deal with design effects, such as survey sampling techniques. This is true, but as Goldstein (2003) argues such procedures treat the variance structure in the sample as nuisance rather than something that is of interest, in and of itself, and that can lead to substantive findings if analyzed properly.
- Further regarding such substantial findings, multilevel models overcome issues related to limitations of aggregated or disaggregated analyses (e.g., the ecological fallacy). Often reality cannot be captured in terms of an aggregated or disaggregated analysis, but only a combination of both in one approach will be satisfactory. For example, higher-level variables such as traffic count, speed limit, number of liquor outlet stores per capita, etc. do affect behaviour at the individual level, but without techniques that can combine variables of different levels into one model, such patterns cannot be revealed and understood.
- Multilevel models can be used to model hierarchical data in a very efficient way. A simpler approach to model a two-level sampling design, for example, with cars nested in road sites would be to include dummy variables for road sites in a one-level model. The model would then contain n-1 dummy variables, with n equal to the number of road sites. This approach is not as efficient as multilevel models since the number of dummy variables can easily increase to a level that may not be manageable (e.g., several hundred dummy variables). This could lead to estimation issues and such an approach can hardly be called parsimonious. Furthermore, this approach would also not allow modelling the influence of higher-level independent variables on lower-level dependent variables, or any cross-level

relationship for that matter. Multilevel models, on the other hand, summarize the variance that exists, not only at the level of individuals, but also at the higher levels and such a decomposition of variance allows capturing patterns that may emerge at different levels in a parsimonious and manageable way.

In conclusion, there are several reasons to justify the use of multilevel models. While it has been demonstrated in this section that this family of techniques is especially useful to deal with a particular set of issues (i.e., design effects, the importance of contextual information, efficiency, and the need for a parsimonious model), it is acknowledged that multilevel models are just one possible solution to these problems and that they may not be as suitable for dealing with other sets of problems. Also, even when dealing with problems that are suitable for multilevel models, there are limitations. For example, the increased complexity of multilevel models compared to traditional one-level models serves to decrease the power of the sample so balancing between complexity — i.e., being as truthful as possible — and efficiency — i.e., being as parsimonious as possible — is important. In sum, multilevel models are not a panacea, but it has been demonstrated in this introductory chapter that they do provide an efficient solution to model hierarchical data.

4. Case study one: The influence of traffic count on drinking driving behaviour[4]

4.1 Introduction

4.1.1 The relevance of traffic count

General deterrence theory predicts that the actual likelihood for getting caught and the perceived likelihood of getting caught are important motivators for people to comply with the law (Ross 1992). The perceived likelihood refers to what people think their likelihood is of getting caught and this impression people have can be different from reality or the actual likelihood of getting caught. Homel (1988) tested this theory using data from random breath testing (RBT) and confirmed that not just the actual likelihood of getting caught is important but the perception of the likelihood of getting caught plays an equally important role in preventing people from becoming drinking drivers. Research also investigated the importance of other factors such as the perceived swiftness or celerity of punishment and the perceived severity of punishment (see for example Grosvenor et al. 1999). This case study, however, is about the relevance of the actual likelihood of getting caught and the perception of the likelihood of getting caught.

Findings regarding those two aspects (i.e., actual likelihood versus perceived likelihood of getting caught) led to the practice of using high-visibility roadside checks when enforcing drinking driving laws, primarily to increase the perceived likelihood of getting caught among the public. The objective of such prevention efforts is to make as many people as possible believe that police officers are out on the road, enforcing drinking driving laws and that drinking drivers will most likely be caught. This should prevent drivers from drinking driving. Such practices have generally acknowledged that high-visibility roadside checks have little (or less) impact on the actual likelihood of a drinking driver getting caught but serve to escalate the perceived likelihood of arrest. Further reasoning may lead one to conclude that targeting high traffic count road sites at any time of the day or week with high-visibility roadside checks should be a priority because it serves to increase awareness of the enforcement activity and deter potential drinking drivers from drinking driving.

An alternative to this prevention approach is the “repression” approach that involves targeting times and places where the highest number of drinking drivers are to be expected. Rather than attempting to affect the perceived likelihood of getting caught, this approach seeks to increase the actual likelihood of getting caught — the aim is to apprehend as many drinking drivers as possible.

[4] This section is based on Vanlaar, W. (2008).

Roadside checks are often organized with either a clear prevention or repression objective in mind. However, it has been suggested that the approaches can be combined (e.g., see Goldenbeld and Hway-Liem 1994). According to such a strategy the prevention approach would largely be applied at high traffic count road sites earlier at night when a lot of people are on the road, while the repression approach would be adopted later at night, close to places where a high number of drinking drivers can be expected, such as areas near to drinking establishments.

This chapter attempts to shed some light on the relationship between drinking driving behaviour and traffic count — one variable in particular that can play a role when setting priorities regarding when and where to conduct stop checks.

In 2003, the Belgian Road Safety Institute conducted the third national roadside survey (Vanlaar 2005a). An interesting negative relationship emerged between traffic count and drinking driving — for every 100 extra cars that drove by a survey site, the odds that a driver would be drinking decreased by 18.1%. Possible confounding factors including intensity of stopping drivers (number of police officers per number of cars driving by), time of day and day of the week, and amount of time spent per road site were controlled for. Considering the exploratory nature of these findings, it was suggested that it may be indicative of drinking drivers anticipating higher chances of getting caught at high traffic count road sites and thus avoiding those places where they expect a lot of traffic (Vanlaar 2005a, p. 396).

Although these findings from the Belgian data were considered inconclusive, they are provocative and warrant further investigation, since they bear on the issue of the likelihood of apprehending drinking drivers using a prevention or repression model and, more practically, on decisions of police officers when deciding where to organize roadside checks. This case study uses Canadian data from a comparable roadside survey, carried out in B.C. in 2003, to further examine the influence of traffic count on drinking driving behaviour.

4.1.2 Objectives and rationale

The objectives in this case study are to model the relationship between traffic count at the level of road sites and drinking driving behaviour at the level of the individual driver. To this end, a previously built model with Belgian data will be replicated using Canadian data. Modelling such a relationship between traffic count and drinking driving can only be done properly using a multilevel approach since the independent variable and dependent variable are measured at different levels. The rationale for modelling such a relationship relates to the importance of efficient and effective police enforcement. Given the finite resources of police and the proven effect of police enforcement on traffic safety, the importance of finding novel ways for police to organize their activities goes without saying.

4.2 Methods

Surveys for this epidemiological study took place in June 2003 at 16 sites in each of three communities in south-western B.C., namely Vancouver, Saanich and Abbotsford (see also Wilson et al. 2006 and Beirness et al. 2007 for a description of the methods). To obtain this total sample of 48 sites, road segments were selected randomly and each selected road segment was then searched for possible survey sites. Traffic cones were used to delineate each selected site and to mark off places where interviews were conducted. Drivers were directed into the survey site by an attending police officer and they were subsequently interviewed by a member of the survey team. Roadside surveys were organized on Wednesday, Thursday, Friday or Saturday evening from 21:00 to 03:00. Each site was surveyed for 90 minutes in one of four shifts – from 21:00 to 22:30, 22:30 to midnight, midnight to 01:30, or 01:30 to 03:00. A total of 2,627 vehicles were selected from the traffic flow and asked to – voluntarily – participate; 2,234 (85%) complied and provided a breath sample and 76 drivers (3%) provided a breath sample but refused to answer the questions (a critical appraisal about the possible impact of voluntary participation in this study is available in the discussion section of this case study). Incidentally, data from the Belgian study were collected during a mandatory police check. A total of 12,891 drivers were stopped, ten drivers (0.08%) refused to provide a breath sample and 57 (0.44%) were not able to provide a breath sample.

The dependent variable in this study is drinking driving; a binary variable distinguishing between those drivers who were below a preset alcohol limit and those who were not. This variable is based on breath alcohol concentrations (BrAC) obtained by submitting stopped drivers to a breath test. Two corresponding BAC cut-off values were used when modelling the data, 0.05% and 0.08%. While the legal BAC limit in Canada is 0.08%, the lower cut-off value was used as well to make the model comparable to the Belgian model – the legal BAC limit in Belgium is 0.05%. Moreover, it is the limit for issuing a roadside suspension in B.C. under the Motor Vehicle Act.

Note that the dependent variable's name is "drinking driving", but that its category below the cut-off values contains drivers who had not consumed any alcohol at all, as well as drivers who had, but whose BAC was still below the cut-off value. For practical reasons, we refer to those drivers who had a BAC equal to or above a cut-off value as drinking drivers; drivers with a BAC below the cut-off value are referred to as not having been drinking driving.

The independent variables (all collected by means of a survey) include: traffic count (a continuous variable indicating the total number of vehicles driving by a survey site during the police check), intensity (a continuous variable calculated by dividing the number of staff per survey site by traffic count for that site), gender, previously stopped (a binary variable distinguishing between drivers who previously have been stopped and tested at a road site at least once in the last six months and drivers who have not been stopped in the last six months), probability of being caught (a categorical variable representing the driver's perception of the likelihood of being caught by the police for drinking driving; categories include: unlikely, neutral, likely), and age (16-18 years, 19-25 years, 26-35 years, 36-45 years, 46-55 years, older than 55 years).

The data were modeled using a multilevel approach with drivers (level 1), nested in survey sites (level 2), nested in cities (level 3). Three-level and two-level variance components model were analyzed, although adding the third level did not make any difference with regards to the results for the regression coefficients. The reported results in this chapter come from the two-level model.

The two-level model that was fit to the data included 48 road sites at level 2 (m=48) and 2,627 drivers at level 1 (n=2,627). To model the relationship between the binary response (i.e., smaller than one of the BAC cut-off values versus equal to or greater than the cut-off values) and the set of explanatory variables, the logit link function was used, meaning a two-level logistic regression was performed. To interpret the relationship between the binary response and an explanatory variable, logit coefficients were transformed into odds ratios using the exponential transformation. The odds ratios compare the odds for drinking driving (i.e., a BAC equal to or greater than the cut-off value) of a certain category of an explanatory variable to the reference category of that particular explanatory variable.

Note that there is an important difference between the explanatory variables traffic count and intensity and the remaining explanatory variables. These two variables are aggregated and do not vary for individuals at the same survey site as opposed to the other explanatory individual variables such as gender and age. These aggregated variables only vary at level two (hence they are called level 2 variables) and their influence on a dependent level 1 variable can be modeled properly with multilevel models.

Probabilities for drinking driving were also calculated per level of traffic count. The formula to obtain these probabilities is derived from the logit model and takes the following form (Rasbash et al. 2005):

$$\pi = \frac{1}{1+\exp(-(\beta_0+\beta_1 x))} \qquad (4.1)$$

where π denotes the probability of drinking driving (i.e., BAC equal to or greater than the cut-off value), "exp" denotes the exponential function, β_0 refers to the intercept of the model, β_1

to the regression coefficient of the explanatory variable traffic count and x represents the variable traffic count. The probabilities obtained using this formula were then converted into percentages.

Finally, estimation of the models discussed was performed using the restricted iterative generalized least squares method (RIGLS). A first order penalized quasi likelihood estimation (PQL) was used to compensate for potential bias (see Rasbash et al. 2005 for an explanation of these estimation procedures).

4.3 Results

Table 4.1 contains a comparison of the results of the Belgian model (Vanlaar 2005a) with the model using the 2003 B.C. data. Both models were controlled for the possible confounding influence of time of day and day of week (the results presented in the tables come from models that did not include time of day and day of week because these variables were not significant; strength and direction of the other variables did not change between models with and models without time of day and day of week).

Table 4.1: Comparison of the Belgian two-level logistic model with the B.C. two-level logistic model; Logit coefficients and standard errors (S.E.) are displayed for fixed and random parameters; coefficients for being a drinking driver (i.e., BAC >=0.05%)

Belgian model		**B.C. model**	
	Logit coefficients (S.E.)		*Logit coefficients (S.E.)*
Fixed parameters			
Intercept	-4.757 (0.285)	Intercept	-4.620 (0.864)
Traffic count	-0.002 (0.000)	Traffic count	-0.002 (0.001)
Intensity	0.896 (0.383)	Intensity	0.787 (2.782)
Female	-1.375 (0.207)	Female	-0.367 (0.279)
Previously stopped	0.409 (0.141)	Previously stopped	0.285 (0.250)
Probability low	0.537 (0.167)	Probability neutral	0.275 (0.391)
Probability medium	0.744 (0.169)	Probability likely	0.661 (0.289)
Probability high	0.312 (0.278)	Age 19-25	1.390 (0.768)
Probability very high	1.432 (0.290)	Age 26-35	1.713 (0.772)
Age 26-39	0.710 (0.242)	Age 36-45	1.669 (0.785)
Age 40-54	1.314 (0.234)	Age 46-55	1.242 (0.841)
Age 55+	0.863 (0.272)	Age 56+	1.044 (0.910)
Random parameters			
Level 2 variance	0.991 (0.197)	Level 2 variance	0.192 (0.173)
Level 1 variance	1.000 (0.000)	Level 1 variance	1.000 (0.000)

Coefficients for categorical variables indicate how likely it is for a respondent in a particular category (e.g., aged 26-39 in the Belgian model) to be a drinking driver (i.e., BAC equal to or greater than 0.05%) compared to the reference category of that categorical variable (in this case 18-25 years of age in the Belgian model). Reference categories for the Belgian model are: male for gender; never been caught drinking driving before for previously stopped; very low probability of being caught; and, age 18-25. Reference categories for the B.C. model are: male for gender; not stopped and tested for drinking driving in the last six months; unlikely that you will be caught by police for drinking driving; and, age 16-18).

As can be seen both models are fairly similar, despite some differences in the methodology of the studies. Of significant interest, corresponding coefficients in both models have the same sign, indicative of the same direction for each relationship between the explanatory variables and the dependent variable. The coefficients are also fairly similar with respect to their magnitude.

While the variables probability of being caught and age were categorized differently in the B.C. model, compared to the Belgian model, a comparable pattern emerges in both. The likelihood of drinking driving increases as the perceived probability of being caught increases (this may seem counterintuitive and will be discussed in more detail later in the discussion section). There is a U-shaped relationship between age and drinking driving with the lowest probability for drinking driving assigned to the youngest age category, the highest probability to the middle age categories and, again, a lower probability to the highest age categories. Finally, level 1 variances are constrained to 1, since this is one of the assumptions of the logistic model. Level 2 variance is lower in the B.C. model.

Non-significant parameters were dropped in the next step, resulting in the final model using the B.C. data, displayed in Table 4.2. Level 2 variance has now become significant (as can be derived from its value of 1.327, which is greater than twice its standard error of 0.248). Using a threshold model approach (see sub-section 3.2.2 for a description of a threshold model) it can be calculated that 28.7% of the total variance is level 2 variance due to the variability between survey sites. Such a percent calls for a multilevel approach when analyzing these data.

Exponential coefficients of the final model are presented as well facilitating the interpretation of the explanatory variables' influence on the dependent variable. For example, the odds ratio of 7.941 for the category 19-25 years old means that respondents aged 19-25 are about 7.9 times more likely than the respondents in the reference category (i.e., 16-18) to drink and drive. This effect for respondents aged 19-25 is significant as the value of the corresponding logit coefficient (2.072) is greater than twice its standard error (0.786). Put another way, being 19-25 years of age corresponds to a significant increase in odds for drinking driving of 694.1% ((7.941-1.0)*100).

Table 4.2: Final B.C. two-level model containing significant effects only; Logit coefficients and standard errors (S.E.), and exponential coefficients are presented for fixed and random parameters; coefficients for being a drinking driver (i.e., BAC >=0.05%)

Parameter	**Logit coefficient (S.E.)**	**Exponential coefficient**
Fixed parameters		
Intercept	-4.026 (0.826)	
Traffic count	-0.004 (0.001)	0.996
Intensity	-5.295 (2.240)	0.005
Age 19-25	2.072 (0.786)	7.941
Age 26-35	2.181 (0.838)	8.855
Age 36-45	2.305 (0.793)	10.024
Age 46-55	2.131 (0.838)	8.423
Age 56+	2.309 (0.869)	10.064
Random parameters		
Level 2 variance	1.327 (0.248)	
Level 1 variance	1.000 (0.000)	

Of particular interest in this case study is the variable traffic count. The odds of drinking driving are multiplied by a factor of 0.996 for each increase in traffic count of one unit. That is, if traffic count at a survey site increases by one car, the odds of finding a drinking driver (i.e., BAC>=0.05%) at that particular site decrease by a factor of 0.996. The range of traffic count in this B.C. study was 1,392 with a minimum of 14 cars and a maximum of 1,406 cars. The probabilities for drivers to be drinking driving at various levels of traffic count are shown in Table 4.3. These probabilities have been calculated using the formula in the Methods section for levels of traffic count ranging from 20 to 1,500 and have been multiplied by 100 to express the results in percentages. Two different calculations have been carried out; one with the parameters of the final B.C. model, excluding the control variables time and day (which yields a β_0-value of -4.026 and a β_1-value of -0.004), and one with the parameters of the final B.C. model, including those two control variables (which yields a β_0-value of -5.456 and a β_1-value of -0.003). The results according to the Belgian model (with a β_0-value of -4.757 and a β_1-value of -0.002) have been inserted for comparison of results between the B.C. data and the Belgian data.

Table 4.3: Comparison of B.C. and Belgian probabilities (in percent) of drinking driving (i.e., BAC>=0.05%) by traffic count (parameters for calculation of Belgian probabilities borrowed from Vanlaar 2005a)

Traffic Count	*BC model, excluding control variables time and day*	*BC model, including control variables time and day*	*Belgian model*
	(β_0=-4.026; β_1=-0.004)	*(β_0=-5.456; β_1=-0.003)*	*(β_0=-4.757; β_1=-0.002)*
20	1.621%	0.401%	0.819%
50	1.440%	0.366%	0.771%
100	1.182%	0.315%	0.698%
500	0.241%	0.095%	0.315%
1,000	0.033%	0.021%	0.116%
1,500	0.004%	0.005%	0.043%

Although the individual probabilities for drinking driving per level of traffic count differ between the three models, more importantly, the same pattern emerges, regardless of what model is used to calculate those probabilities. Overall, there is a clear decreasing trend in the individual probability of drinking driving as the level of traffic count increases; at high traffic count survey sites the chances of finding a driver who has been drinking are significantly lower.

Finally, Table 4.4 displays absolute numbers and row-percentages of BAC by traffic count from the B.C. survey (the grand total in this table does not correspond to the total number of drivers mentioned previously due to missing values in the variables used in Table 4.4). Four levels of traffic count have been included in this table: survey sites with a count of up to 75; survey sites with a count greater than 75 but smaller than or equal to 250; survey sites with a count greater than 250 but smaller than or equal to 500; and, survey sites with a traffic count greater than 500 but smaller than or equal to 1,500. These categories correspond to the range of traffic counts in this study. Three categories of BAC are displayed: BAC lower than 0.05%; BAC greater than or equal to 0.05% but smaller than 0.08%; and, BAC equal to or greater than 0.08%.

Not surprisingly, the significant effect that was found for traffic count using the multilevel approach was reproduced in this 4x3 table according to a Chi-square test (df=6; Chi-square=28.35; p=0.015). There were 15 drinking drivers (BAC>=0.05%) at sites with a count of up to 75, compared to 9 at sites with counts of at least 501. Out of a total of 93 drinking drivers (BAC>=0.05%), 63 (i.e., 68%) were detected at the survey sites with a count of up to 250.

Table 4.4: Traffic count in four categories by BAC in three categories; absolute numbers and row-percentages are presented

Traffic Count	BAC<0.05%	0.05%≤BAC<0.08%	BAC>=0.08%	Total
<=75	147	7	8	162
	90.7%	4.3%	4.9%	100.0%
75<.<=250	1,057	19	29	1,105
	95.7%	1.7%	2.6%	100.0%
250<.<=500	607	9	12	628
	96.7%	1.4%	1.9%	100.0%
500<.<=1,500	410	5	4	419
	97.9%	1.2%	0.9%	100.0%
Total	2,221	40	53	2,314

4.4 Discussion

The objective of this case study was to provide some insights into the relationship between the variable *traffic count* and drinking driving behaviour. The relevance of understanding such a relationship becomes clear when examining two important frameworks — prevention or repression — that are often used when organizing roadside checks to enforce drinking driving legislation. It has been suggested that a prevention approach — designed to increase the perceived likelihood of arrest — would best be applied at high traffic count road sites early at night when a lot of people are on the road. The repression approach — designed to increase the likelihood of detecting and apprehending drinking drivers — would best be adopted at night, close to places where a high number of drinking drivers can be expected.

Regardless of the primary objective, there is a need to understand how traffic count influences drinking driving behaviour as traffic count may play a role in a police officer’s choice of sites for a roadside check.

Despite some differences in the methods applied in both the Belgian and the B.C. study, the latter study replicated the findings of the Belgian model substantiating that the probability for drivers to be drinking driving significantly decreases with an increasing level of traffic count. This supports the suggestion that drinking drivers avoid high traffic count road sites, possibly because they anticipate a higher chance of getting caught at those high traffic count sites. An alternative explanation could be that drinking drivers tend to avoid congestion or more difficult driving situations. It warrants mentioning that this roadside survey provides evidence about the behaviour of drinking drivers, but it provides little or no insight into reasons for this behaviour or motivations. To truly understand why drinking drivers behave this way, it is necessary to adopt a different approach, for example by using a questionnaire designed to survey drinking drivers.

As a side bar, a strange result was identified regarding drivers' perception of being stopped and tested on an average trip, i.e., the perceived likelihood of getting caught. The data clearly support a positive relationship, meaning that drivers who estimate the likelihood of getting tested as very high, are at the highest risk for drinking driving. Based on the theory of deterrence, however, one would expect the opposite, i.e., those drivers who think it is more likely to be tested for drinking driving would be the least likely to actually drink and drive. A possible explanation is that the perception of drivers who are caught on the spot is influenced by this event: because drinking drivers are actually being caught, their perception of how likely it is they will be caught is biased toward the upper extreme, i.e., they estimate the chances of being caught higher than they would if they would not have been caught. This would suggest a roadside survey with police officers involved may not be suitable to investigate this relationship between perceptions of the likelihood of being caught and chances of drinking driving. An alternative explanation could be related to a selective memory bias for alcohol cues (Franken et al., 2003). This memory bias involves an enhanced memory among alcoholics and abstinent alcoholics for alcohol-related cues compared to neutral cues. Typically this memory bias is used to help explain reasons for relapsing. However, in this situation it could perhaps help explain why drinking drivers would overestimate the likelihood of being caught.

Regarding the relationship between traffic count and drinking driving behaviour, in addition to probabilities, absolute numbers are important as well, especially if the objective is more related to the apprehension of as many drinking drivers as possible, rather than just raising the perceived likelihood of arrest. It could be argued that it does not matter if the chances for drinking driving are much higher at low traffic count sites, if the absolute number of drinking drivers at those sites is so low that the costs of organizing a roadside check, compared to the benefits, would be too high. In this regard, results from Table 4.4 were enlightening. The majority of drinking drivers were detected at survey sites with lower traffic counts – 250 cars or less per 90 minutes, corresponding to a maximum of less than three cars per minute. To illustrate, 63 out of 93 drinking drivers with a BAC greater than or equal to 0.05% (i.e., 68% of all drinking drivers) were detected at sites with traffic counts equal to or less than 250. Another 21 drinking drivers (i.e., 22%) were detected at sites with traffic counts greater than 250 but smaller than or equal to 500. Only nine drinking drivers (i.e., 10%) were caught at sites with traffic counts of at least 501.

Despite the convergence of evidence resulting from a comparison of the B.C. data with the Belgian data, it may be premature to regard the findings concerning the relationship between traffic count and drinking driving as conclusive. Possible confounding factors such as time of day and day of week were included both in the Belgian and the B.C. model so they could be kept constant, but the B.C. data are limited in that surveys only took place on Wednesdays through Saturdays from 21:00 to 03:00, while the Belgian data were representative of the entire week. No data were gathered in either study about the differences between urban and rural survey sites or about characteristics of the roads along which these survey sites were located.

Despite the lack of a more detailed classification of sites, there are good reasons to believe that B.C. sites could largely be classified as urban because of the sampling design involving only three major B.C. cities. Aggregated level 2 variables such as urban/rural and variables pertaining to characteristics of roads should be investigated in further research. Also, the B.C. study was based on voluntary participation. As a result only 88% of the sample provided a breath sample. The refusal rate in the Belgian study was much lower because the breath samples were collected as part of an enforcement campaign and not providing a breath sample was punishable and considered equal to blowing over the legal limit. While it is acknowledged that the higher refusal rate in the B.C. study may have biased the results, it is important to note that there is convergence in the results of both studies.

In conclusion, the successful replication of the Belgian model using B.C. data suggests that a more thorough investigation of the influence of traffic count on drinking driving behaviour and possible confounding variables may provide useful insights with respect to the strategic planning of roadside checks to enforce drinking driving legislation.

5. Case study two: Sleepiness among night-time drivers[5]

5.1 Introduction

5.1.1 Background

Despite a relatively extensive body of research on fatigue and drowsiness in relation to driving, their impact on traffic safety is not well understood. This is at least partly due to difficulties and inconsistencies associated with defining fatigue or drowsiness and relating them to the risk of collision.

Although fatigue and drowsiness have different causes and are governed by different processes, they are usually considered together because the results are the same – the person suffering from fatigue or drowsiness becomes less alert or attentive and can, in the extreme, fall asleep.

Drowsiness or sleepiness normally refers to "the urge to fall asleep" (Beirness et al. 2005: p. 6) as the result of a biological need; it is a physiological state of the body that is irreversible in the absence of sleep. It is governed by a circadian sleep-wake cycle that makes most people feel sleepy twice a day – at night and in the afternoon (Dement and Vaughan 1999). Fatigue, on the other hand, has been defined as a "disinclination to continue performing the task at hand" (Brown 1994: p. 239), caused by physical labor or repetitive and monotonous activities, such as monitoring a display screen or driving long distances (Stutts et al. 1999).

Despite the different causes of drowsiness/sleepiness and fatigue, the effects on traffic safety can be the same in that they can compromise the ability to drive safely. They both lead to impaired performance at the wheel and can ultimately result in falling asleep at the wheel (Brown 1994).

According to estimates from a previous Canadian public opinion poll, 20% of drivers in Canada admitted to falling asleep or nodding off at least once while driving in the past 12 months (Beirness et al. 2005). Other estimates, although quite variable, also suggest that the problem is anything but unimportant. The 2002 "Sleep in America" survey found that, in the past year, 51% of drivers admitted to driving while drowsy, 17% admitted to dozing off while driving, and 1% reported having been involved in a crash due to dozing off or fatigue (National Sleep Foundation 2002). Comparable figures have been reported in other studies. For example, in a survey of factors associated with falling asleep at the wheel among long-distance truck

[5] This chapter is based on Vanlaar et al. (2008a) and on Vanlaar et al. (submitted).

drivers, 47.1% of the respondents stated they had fallen asleep at the wheel of a truck; 25.4% had fallen asleep at the wheel in the past year (McCartt et al. 2000).

Fatigued or drowsy drivers would not constitute a major problem if few of these drivers crashed, which is suggested by the "Sleep in America" survey that reported only 1% did so. Other studies suggest differently. In the United States (U.S.), it is believed that up to 20% of serious crashes may be due to fatigued or drowsy driving (Horne and Reyner 1995; Horne 2000). Knipling and Wang (1995) estimated that fatigue likely contributes to 79,000-103,000 collisions and 1,500 fatalities annually in the U.S. These might well be underestimates, given the difficulties in identifying fatigue or drowsiness as a factor in crashes.

In addition to estimating the extent of the problem, some researchers have attempted to identify the characteristics of the problem. It has been suggested that the profile of a sleep-related crash is different from that of other crashes. For example, according to Stutts et al. (1999: p. 9) sleep-related crashes are more likely to:

- occur at night or in mid afternoon, when people have a natural propensity to sleep;
- involve a single vehicle running off the roadway, or rear-end and head-on collisions;
- occur on higher-speed roadways;
- involve only the driver as occupant, who is young and male; and,
- result in serious injuries.

Research has also attempted to identify populations at risk for involvement in crashes due to fatigued or drowsy driving. These include:

- young males (Wang et al. 1996; Pack et al. 1995), because they are more likely to drive late at night and to be sleep deprived;
- persons with sleep disorders (Findley et al. 1989; Cohen et al. 1992; Young et al. 1997), because they are more likely to suffer from acute and/or chronic sleep deprivation;
- drivers under the influence of alcohol (Horne et al. 2003; Lumley et al. 1987), because of the sedating effects of alcohol and because of its exacerbating effects on performance deficits due to fatigue or drowsiness;
- drivers under the influence of certain medications (Ray et al. 1992; Ceutel 1995) with side-effects known to enhance drowsiness;
- night or rotating shift workers (Dalziel and Job 1997; Marcus and Loughlin 1996; McCartt et al. 1996), because they are more likely to get inadequate sleep or experience poor quality sleep; and

- commercial vehicle operators (McCartt el al. 1997; Wylie et al. 1996) who often spend long hours driving, and likely experience both fatigue and drowsiness.

The combined effects of alcohol and sleepiness on crashes are of particular interest in this case study. Even small amounts of alcohol can exacerbate lane-tracking performance on a driving simulator in sleep-deprived subjects (see Wilson et al. 2006). Performance in the combined sleep and alcohol impaired condition is worse than with either sleep impairment or alcohol impairment alone (Horne et al. 2003; Lumley et al. 1987).

5.1.2 Objectives and rationale

Replicating the effects of alcohol on sleepiness that have been established in a laboratory using an epidemiological approach would provide valuable converging evidence about the relationship between this independent variable and the dependent variable. That is the purpose of this study: to apply a multilevel modelling approach to epidemiological data to investigate the effects of the independent variable BAC along with other independent variables such as age, gender, passenger configuration and time of interview on the dependent variable "sleepiness".

The rationale for using a multilevel model in this case study is primarily related to the need to account for a complex sampling design. Also, time of interview, an independent variable of interest, was not measured at the individual level of drivers in this study, but rather at a higher level (road sites where individual drivers were stopped). As a consequence, multilevel models are required to properly model the relationship between these variables that are measured at different levels.

5.2 Methods

5.2.1 Procedure

Data from two epidemiological studies on subjective sleepiness were used. Surveys for these studies took place in June 2003 and in June 2006 at 16 sites in each of three communities in south-western B.C., namely Vancouver, Saanich and Abbotsford (note that the 2003 survey data were also used in the previous case study). Traffic cones were used to delineate each selected site and to mark off places where interviews were conducted. Drivers were directed into the survey site by an attending police officer and they were subsequently interviewed by an interviewer (who was not a police officer).

In 2003, a total of 2,627 vehicles were selected from the traffic flow and asked to – voluntarily – participate; 2,246 (85%) complied and provided a breath sample and 76 drivers (3%)

provided a breath sample but refused to answer the questions. In 2006, 2,374 drivers were stopped and asked to participate; 2,075 (87%) drivers complied and provided a breath sample and another 109 drivers (5%) provided a breath sample but refused to answer the questions. Most of the sites that were used in 2003 were also used in 2006; only a few sites that were no longer available in 2006 had to be replaced.

5.2.2 Sampling design

A two-stage cluster sample was used. First, road segments were selected randomly and each selected road segment was then searched for possible survey sites. As such, 48 sites were selected as PSU's. In a second stage drivers were selected at each PSU. Each site was surveyed for 90 minutes on Wednesday through Saturday evenings in one of four shifts, more precisely from 21:00 to 22:29, or from 22:30 to 23:59, or from midnight to 1:29, or from 1:30 to 3:00.

5.2.3 Data

The dependent variable in this study is self-reported sleepiness; it is a binary variable distinguishing between either being wide awake (value 0) or being somewhat to very sleepy (value 1).

The independent variables include: age in six categories (16-18 years, 19-25 years, 26-35 years, 36-45 years, 46-55 years, older than 55 years; the latter category is the reference category); gender, with women being the reference category; BAC in four categories (zero BAC, 0.005g/dL-0.049g/dL, at least 0.05g/dL, test refusal, i.e., those ones who refused to take a breath test, namely 12% in 2003 and 8% in 2006; zero BAC is the reference category); passenger configuration (driver only; family; one passenger, gender different from driver; one passenger, same gender as driver; more passengers, gender different from driver; more passengers, same gender as driver; driver only is the reference category); trip origin (work, friend, restaurant, bar, home, other; home is the reference category); and, interview time (21:00-22:29, 22:30-23:59, midnight-1:29, 1:30-3:00; the latter category is the reference category).

5.2.4 Data analysis

Two binomial multilevel models were developed, one using the 2003 data and one using the 2006 data. Note that the data from both years have not been collapsed to develop one model because a different scale was used to measure the dependent variable, subjective sleepiness, in each year. In 2003 a three-point scale was used ranging from ‘1’ (wide awake) over ‘2’ (somewhat sleepy) to ‘3’ (very sleepy), while in 2006 a six-point scale was used ranging from ‘1’, corresponding to ‘fully alert’ to ‘6’, corresponding to ‘fighting sleep’.

Both multilevel models have two levels (drivers at level 1 nested in road sites at level 2). The logit link function was used, so a two-level logistic regression analysis was performed. Both models were fit using the restricted iterative generalized least squares (RIGLS) method, more precisely a second order penalized quasi likelihood (second order PQL) estimation (see Rodriguez and Goldman 1995, and Goldstein and Rasbash 1996 for more information about these estimation procedures).

Cases were weighted according to the inverse of their probability of being included in the sample (StataCorp. 2007) — cases with a higher probability of being included will thus be assigned a smaller weight and cases with a lower probability of being included will be assigned a greater weight ensuring over- and under-representation of cases in the sample is accounted for in the analyses. To obtain such weights, information about the number of PSU's (i.e., road sites) included in this study and traffic count (i.e., the number of vehicles driving by at the road site) at each road site was used. More precisely, the weight was calculated as the product of the following probabilities:

- Inverse of the probability of selecting a community in B.C.: 15/3 (three communities out of a possible pool of 15 comparable communities);
- Inverse of the probability of selecting a road site in each of the selected communities: 500/16 (16 road sites per community out of a total population of 500 comparable road sites per community);
- Inverse of the probability of selecting a driver at a road site: COUNT/n, with COUNT equal to traffic count at each road site and n equal to the number of drivers who were actually stopped at each road site.

It warrants mentioning that the numerators of the first two probabilities (15 and 500) are somewhat arbitrary. The number of comparable communities (15) to the communities that were used in this study (3) may be lower or higher. The same is true of the total number of road sites (500), which is particularly difficult to measure (how large is a road site and directly related to this, how many road sites fit in one community?). The number of communities has been discussed with researchers familiar with the demographics of the province of B.C. and it was suggested that the number may actually be lower. However, since in this study the first two probabilities are kept constant and only the third probability varies over drivers, the arbitrary nature of those first two probabilities is irrelevant because the impact of both on the analysis results is cancelled out. The only probability that truly has an impact on the results of the analysis in this study is the third probability, which can be measured objectively based on traffic count and the actual number of tested drivers at each location.

If only the third probability has an impact, why bother with the first two probabilities? Because conceptually it makes more sense to distinguish between all three probabilities, especially when there are reasons to believe that the numerators would vary instead of being constant like in this study. For example, certain communities may be considerably smaller than others, which would suggest that the total number of road sites is smaller in those smaller communities. While it may not be possible to measure the exact number of road sites in each community, it may be possible to estimate the proportion of road sites in one community compared to another based on each community's land mass or total length of roads (e.g., if one community is twice as small as another community the proportion would be 1/2, regardless of how many road sites in total there really are in each community or if one community's network of roads is twice as small as another community the proportion would be 1/2 too). Once the probability at this level of communities would no longer be a constant, then, of course, its impact would not be cancelled out in the analysis results. An example of a situation where communities varied too much is the Belgian drink driving roadside survey, carried out by Vanlaar in 2003 (Vanlaar 2005a). Three regions were included in this study and one of them (Brussels) was considerably smaller than the other two regions (Flanders and the Walloon provinces), so different probabilities for selecting road sites were assigned to those regions.

5.3 Results

The results from both models can be found in Tables 5.1 and 5.2. Table 5.1 contains the logit coefficients and their standard errors (coefficients that are twice as great as their standard erro in this table are significant), while Table 5.2 displays the same results but on the exponential scale, and also contains the p-values, rather than the standard errors.

With both models the results can be interpreted in the same way. The odds ratio for a particular category of an independent variable indicates how likely it is for respondents belonging to that category to report feeling somewhat sleepy to very sleepy, compared to the reference category of that variable. For example, the odds ratio of 3.28 for 16-18 years old in Table 5.2, model 1 (2003 data) means that respondents aged 16-18 are 3.28 times more likely than the respondents in the reference category, i.e., 55+, to report feeling somewhat to very sleepy. This effect for respondents aged 16-18 is significant because its p-value is equal to 0.000.

Table 5.1: Logit coefficients and standard errors (S.E.) of the 2003 and 2006 models about the effects on subjective sleepiness

	Model 1 (2003)		**Model 2 (2006)**	
	Logit coef.	**S.E.**	**Logit coef.**	**S.E.**
Level 1 fixed parameters				
Age				
16-18	1.189	0.298	0.418	0.313
19-25	0.957	0.455	0.684	0.202
26-35	0.942	0.445	0.704	0.188
36-45	1.024	0.432	0.634	0.238
46-55	0.553	0.466	0.451	0.230
Gender				
Male	-0.230	0.159	-0.289	0.091
BAC (g/dL)				
0.005-.049	0.454	0.208	-0.080	0.322
0.050+	0.037	0.227	0.359	0.225
Refusal	0.306	0.217	-0.098	0.343
Passenger configuration				
Family	-0.343	0.318	0.084	0.416
1 pass diff sex	-0.025	0.183	-0.051	0.127
1 pass same sex	-0.626	0.180	-0.469	0.164
Group diff sex	0.063	0.176	-0.190	0.226
Group same sex	-0.901	0.293	-0.385	0.278
Trip origin				
Work	0.234	0.208	0.161	0.208
Friend	0.143	0.104	0.146	0.236
Restaurant	-0.037	0.267	0.031	0.299
Bar	-0.008	0.507	0.167	0.324
Other	0.141	0.208	0.242	0.138
Level 2 fixed parameters				
Interview time				
21:00-22:29	-0.918	0.263	-0.418	0.184
22:30-23:59	-0.485	0.223	-0.342	0.191
Midnight-1:29	-0.554	0.262	-0.085	0.175
Random parameters				
Level 2 variance	0.406	0.085	0.189	0.033
Level 1 variance	1.000	0.000	1.000	0.000

Table 5.2: Odds ratios and p-values of the 2003 and 2006 models about the effects on subjective sleepiness

	Model 1 (2003)		**Model 2 (2006)**	
	Odds ratio	**p-value**	**Odds ratio**	**p-value**
Level 1 fixed parameters				
Age				
16-18	3.28	0.000	1.52	0.182
19-25	2.60	0.035	1.98	0.001
26-35	2.57	0.034	2.02	0.000
36-45	2.78	0.018	1.89	0.008
46-55	1.74	0.235	1.57	0.050
Gender				
Male	0.79	0.148	0.75	0.001
BAC (g/dL)				
0.005-.049	1.57	0.029	0.92	0.803
0.050+	1.04	0.872	1.43	0.110
Refusal	1.36	0.160	0.91	0.775
Passenger configuration				
Family	0.71	0.280	1.09	0.840
1 pass diff sex	0.98	0.890	0.95	0.688
1 pass same sex	0.53	0.001	0.63	0.004
Group diff sex	1.07	0.717	0.83	0.402
Group same sex	0.41	0.002	0.68	0.167
Trip origin				
Work	1.26	0.261	1.17	0.440
Friend	1.15	0.171	1.16	0.538
Restaurant	0.96	0.888	1.03	0.916
Bar	0.99	0.999	1.18	0.607
Other	1.15	0.498	1.27	0.080
Level 2 fixed parameters				
Interview time				
21:00-22:29	0.40	0.000	0.66	0.023
22:30-23:59	0.62	0.029	0.71	0.074
Midnight-1:29	0.57	0.035	0.92	0.629

All the effects in Table 5.2 can be interpreted in a similar fashion and explain how belonging to a particular category of one of the independent variables influences the level of self-reported sleepiness among respondents, compared to the reference category of that independent variable. As such, both models suggest that being younger than 46 significantly increases your chances of feeling sleepy because the odds ratios for any age category below this threshold in both models are larger than one compared to the reference category of 55+. According to model 2, however, the youngest age category of 16-18 years old is not significantly more likely to report feeling sleepy, compared to the reference category of 55+ (p-value of 0.182). Also, while the odds ratios for the category 46-55 are greater than one, suggesting a greater likelihood for this category to report feeling sleepy compared to 55+, these results are not significant (p-values of 0.235 and 0.050 respectively).

The odds ratio for males in Table 5.2 was 0.79 (p-value of 0.148) in 2003 and 0.75 (p-value of 0.001) in 2006. Such odds ratios mean that the chance of men reporting feeling somewhat to very sleepy is 0.79 and 0.75 times smaller than women's chance of feeling somewhat to very sleepy, respectively. Note that this effect is only significant according to the 2006 data.

Results for the variable BAC in Table 5.2, model 1 indicate that having drunk a moderate amount of alcohol (BAC between 0.005% and 0.049%) significantly (p-value of 0.029) increases your chances of feeling somewhat to very sleepy by 57% ((1.57-1)*100), compared to drivers who were sober. The direction of this effect, however, is reversed in the 2006 data: the point estimate of 0.92 suggests that the chances of feeling somewhat to very sleepy are actually smaller for drivers admitting to having consumed moderate amounts of alcohol compared to those drivers who did not consume alcohol at all (or at least did not admit to it). This effect, however, is not significant (p-value of 0.803).

An interesting effect emerges in Table 5.2 in both models with respect to passenger configuration. Apparently drivers who are accompanied by a passenger of the same sex are less likely to report feeling sleepy compared to drivers who are driving by themselves. According to the 2003 data there is a significant (p-value of 0.001) decrease of 47% ((1-0.53)*100) and according to the 2006 data there is a significant (p-value of 0.004) decrease of 37% ((1-0.63)*100). In 2003 the effect for a group of passengers with the same gender was also significant (odds ratio of 0.41, p-value of 0.002), but not in 2006 (odds ratio of 0.68, p-value of 0.167). Note that this variable was later recoded into two categories in both years (same sex passenger and group of same sex passengers versus all the other categories in 2003 and same sex passenger versus all the other categories in 2006). These effects were also significant (2003: odds ratio of 0.49, p-value of 0.001; 2006: odds ratio of 0.63, p-value of 0.009). An interaction effect between gender of the driver and this recoded variable 'passenger configuration' was then included in the model, but this interaction effect was not significant (2003: odds ratio of 1.11, p-value of 0.805; 2006: odds ratio of 1.09, p-value of 0.747).

Finally, the level 2 variable 'interview time' was also significant (this variable is a level 2 variable because it only varies at level 2, i.e., at the level of road sites — all the drivers driving by a particular road site would have been measured in the same time interval of 90 minutes). Reported sleepiness was lower in earlier time intervals in 2003 and 2006. For example, the odds ratios for each category of interview time are smaller than one and significant according to the 2003 data (odds ratios in Table 5.2: 0.40, p-value of 0.000; 0.62, p-value of 0.029; and, 0.57, p-value of 0.035). Only the effect of the interval midnight to 1:29am in 2006 was not significant (odds ratio of 0.92 and p-value of 0.629).

As can be seen in Table 5.1, the level 2 variance in both models is relatively low. According to the intra-class correlation coefficient (see Snijders and Bosker 1999: p. 22) 11% of the total variance is level 2 variance in model 1 ($0.406/(0.406+\pi^2/3)$) and about 5% in model 2 ($0.189/(0.189+\pi^2/3)$).

5.4. Discussion

Results from two roadside surveys were analyzed in this case study using a multilevel modelling approach. It was found that only about 5% to 11% of the variance in the data was variance at level 2, i.e., road sites. This may seem low compared to other studies. For example, Vanlaar (2005a) found that about 23% of the variance in data coming from a roadside survey on drinking driving was level 2 variance. This would suggest that the need for a multilevel approach may not be as pronounced with the data in the present study. However, when comparing the 2003 results from this case study to previously reported 2003 results using the same data but according to a one-level analysis (see Wilson et al. 2006), it becomes clear that the p-values reported in this case study are greater and closer to the 5% threshold than the ones in the previously published paper by Wilson et al. This explains why, for example, gender was significant with the 2003 data according to Wilson et al. (2006) while it was not significant according to the present 2003 results. Despite this low percent of variance at level 2 in this study, the comparison with the previously published results by Wilson et al. in 2006 suggests design effects are at play in this dataset and should be accounted for.

Interestingly, based on the two level analysis using 2006 data, a significant effect of gender on self-reported sleepiness was found, so the effect of gender may be a true one after all. On the other hand, the effect suggests that males report feeling less sleepy than females, which may be somewhat surprising, given that males have been found to be more at risk for driving while sleepy in other research (see e.g., Wang et al. 1996; Pack et al. 1995). This may perhaps be explained by underreporting feelings of sleepiness among males and should be further examined. Special attention should be given to differences between males and females in terms of self-reported sleepiness and how this compares to actual sleepiness.

The effects for age and interview time are in line with the expectations. Younger people are more likely to report feeling sleepy. This can likely be explained by the lack of sleep among younger people, as argued in the background section. Those respondents who were interviewed early at night (those between 21:00 and midnight) are less likely to report feeling sleepy. This finding corresponds to what is known about the circadian sleep-wake cycle according to which people are sleepy twice a day (at night and in the afternoon).

With respect to BAC, only the 2003 data provide evidence to suggest that consuming alcohol may exacerbate feelings of sleepiness. Having consumed moderate amounts of alcohol increased chances for feeling sleepy by about 57%. However, logic dictates that consuming larger amounts of alcohol would then increase feelings of sleepiness even more, but it was not possible to confirm this with 2003 data. Furthermore, the 2006 data do not substantiate the 2003 finding regarding the consumption of moderate amounts of alcohol. A more objective measure of sleepiness may shed light on the relationship between alcohol and sleepiness as it is possible that drivers who are found to have consumed alcohol feel they are already in trouble and do not longer wish to admit to anything else that could make their situation more precarious, even when they are told that the data are collected for the purposes of a study and not enforcement. Also, data of other variables that may have a considerable impact on feelings of sleepiness were not collected in this study. These variables include length of the current trip and number of hours of sleep the past 24 hours (or in extension in the past week or month, since sleep debt can be accumulated over a longer period of time). Inclusion of such important confounding variables would also be useful to get a better handle on the relationship between BAC and sleepiness.

Finally, the only other effect that was found to be significant was passenger configuration, more precisely, drivers who were accompanied by a passenger of the same sex (or passengers of the same sex according to the 2003 data) were less likely to report feeling sleepy compared to drivers who drove alone. After recoding the data, it became clear that drivers who are accompanied by a passenger(s) of the same sex are also less likely to report feeling sleepy compared to any other category, not just drivers alone. It is not clear which mechanism could explain this. As reported by Wilson et al. (2006), it is not clear whether this means drivers are actually less sleepy or simply under-reporting feelings of sleepiness when they are in the presence of a passenger(s) of the same sex due to social dynamics in groups of different composition. In an attempt to further explain this, the variable was recoded and an interaction effect between gender of the driver, passenger configuration and self-reported sleepiness was investigated. Unfortunately, the interaction effects in the 2003 and 2006 data were not significant and did not allow a better explanation of the effect. More research into this finding is needed. It is recommended this effect be studied from a social dynamics point of view.

In conclusion, analyzing these data using a multilevel approach helped shedding some light on the issue of sleepiness at the wheel and the interaction between alcohol and sleepiness. However, many questions remain unanswered and it seems more data are needed to better understand the issue.

6. Case study three: An evaluation of graduated driver licensing programs in North America using a meta-analytic approach[6]

6.1 Introduction

6.1.1 Background

Graduated driver licensing (GDL) programs attempt to provide a more protective environment for novice drivers, typically by lengthening the learning process and imposing a set of restrictions aimed at reducing their risk of collision. To achieve this, most GDL programs are multi-staged and include a learner's stage and an intermediate stage before graduation to a full license.

Despite the "somewhat torturous journey that graduated licensing has experienced in achieving acceptance among the public and policy-makers" (Simpson 2003: p. 25), GDL programs are now commonplace in North America. Indeed, most jurisdictions in Canada and the United States have some version of GDL. Furthermore, a consensus exists in the research community about the effectiveness of GDL programs and today they are widely accepted as an effective safety measure: "The systems that have been evaluated have been found to be very effective in reducing crashes and injuries, and public acceptance is high. This in and of itself provides the compelling case for graduated licensing." (Williams 2003: p. 3).

A sound body of evidence documenting the positive influence of GDL programs on collisions among novice drivers is available. In addition to the more than two dozen evaluation studies of specific GDL programs (see Mayhew et al. 2005 for a descriptive review of these evaluations; see also Hedlund et al. 2006 for an update on research published since 2005 and research in progress; and Shope 2007 for the most recent review of the GDL evaluation literature), two systematic reviews of this literature have been carried out (Foss and Evenson 1999; Hartling et al. 2005). Collectively, these studies and their reviews clearly establish GDL as an effective safety measure.

The authors of a recently published evaluation study that distinguished between different kinds of GDL programs came to the same conclusion (Morrisey et al. 2006; see also Dee et al. 2005). This study, however, was limited geographically in that it only analyzed data from the United States. A more recent study (Baker et al. 2006) was carried out to determine which types of GDL programs are associated with reductions in fatal crashes involving 16-year-old

[6] This section is based on a TIRF report — Vanlaar et al. (2009a) — and a paper published in Accident Analysis and Prevention — Vanlaar et al. (2009b). This case study has also been presented at the Canadian Multidisciplinary Road Safety Conference by Vanlaar et al. (2009c) and the 35th International Forum on Traffic Records and Highway Information Systems.

drivers — evaluations have found a wide range in GDL effectiveness so this particular study sought to determine which types of programs had the greatest impact. This study is also limited geographically to the U.S. and only investigated one age cohort. Furthermore, while the results of this study allow distinguishing between more effective programs (those with at least five GDL components) and less effective programs (those with less than five GDL components), the authors did not estimate the relative importance of individual components of GDL programs — components such as nighttime driving restrictions, supervision or passenger restrictions. Some of the same limitations apply to a more recent study of GDL effectiveness, both in terms of fatal and injury collision involvements, by the same authors (Baker et al., 2007).

Finally, Williams (2007) conducted a literature review to analyze the available evidence regarding GDL components and concluded that extended learner periods, nighttime restrictions and passenger restrictions have contributed to crash reductions, but that there is more to learn about GDL and its components.

Therefore, despite the available literature it is still not known which GDL features contribute most to collision reduction and how exactly this is achieved. Consequently, it is difficult to identify how a GDL program should be best designed or improved. As a result, a priority research need is to identify "effects of specific GDL components and provisions" (Hedlund et al. 2003: p. 109), a need that had already emerged from the Symposium on Graduated Driver Licensing in Chatham, MA on November 5 – 7, 2002.

6.1.2 Objectives and rationale

The objectives of this case study are to calculate a summary statistic of GDL effectiveness, to identify the most effective components of GDL programs, and to help understand how GDL components achieve their effect by applying a meta-analytic approach (see section 3.2.3 for an introduction to meta-analysis and meta-regression).

In their review of the existing GDL literature, Foss and Evenson (1999) were unable to conduct a meta-analysis of GDL evaluation studies and to draw reliable conclusions about the impact of different features of GDL programs on crash numbers, probably because the number of studies available to them at that time was rather small (seven). At best, Foss and Evenson (1999) had to compare outcomes from different studies, using different methodologies. Without controlling for these differences, it is difficult to compare the results of program effectiveness, which, in turn, makes it difficult to determine which features of programs are most effective.

As mentioned previously in section 3.2.3, the ability to address this challenge is an advantage of meta-analysis that controls for the different methodologies by comparing the different studies available “on the same scale”, i.e., standardizing. Results of a meta-analysis regarding differences in effectiveness will, therefore, be more valid because they truly reflect differences in programs and program features by accounting for differences in evaluation methodologies (see Elvik 2005a and 2005b).

A more recent effort by Hartling et al. (2005) to systematically review the GDL literature used the well-acknowledged Cochrane methodology (www.cochrane.org) — note that the Cochrane Collaboration is a global network of volunteers, supported by a small staff; collectively they developed the Cochrane methodology for meta-analysis research. Even though the number of evaluation studies included in their review was higher than in the Foss and Evenson study (1999), the authors decided it was not appropriate to pool the results of the evaluation studies and perform a meta-analysis, “due to statistical heterogeneity and differences among studies with respect to study quality and design, program quality and design, definition of outcomes, baseline rates, and data reported” (Hartling et al. 2005: p. 5).

In an attempt to overcome the issue of heterogeneity, the feasibility of obtaining raw data for each of the evaluated studies was considered in this case study. Using such raw data to conduct a meta-analysis is considered to be "the 'yardstick' against which other forms of systematic review should be measured" (Clarke and Stewart 2007: p. 110) because it becomes possible to standardize the outcome measure to the fullest. In other words, by using raw data to calculate a standardized outcome measure it is no longer necessary to use the different kinds of outcome measures that are being reported in evaluation studies and that are not necessarily suitable to be summarized into one statistic. It was therefore decided to use counts of fatalities per age cohort and by state from the Fatality Analysis Reporting System (FARS) for the U.S. and comparable data for Canadian jurisdictions, contained in Transport Canada’s Traffic Accident Information Database (TRAID), to calculate a standardized outcome measure. Once it was decided to use these raw data, it was no longer desirable to limit the evaluation to those jurisdictions for which an evaluation study was available; hence the present evaluation includes all U.S. and Canadian jurisdictions that have a GDL program in place. As such, a meta-analytic approach was adopted, rather than conducting a meta-analysis according to the more common or colloquial sense of the term.

While the choice to use raw data to calculate a standardized outcome measure for each of the jurisdictions included in this evaluation precluded any difficulties with respect to methodological heterogeneity among different studies, there was still the considerable heterogeneity with respect to programs and their features to deal with; see for example the on-line inventories of GDL programs on the website of the Insurance Institute for Highway Safety (IIHS) and the Traffic Injury Research Foundation (TIRF), respectively at www.iihs.org/laws/graduatedLicenseIntro.aspx and www.trafficinjuryresearch./yndrc/default.asp

However, sophisticated meta-analysis techniques are available today to explicitly model diversity, more precisely random meta-regression analysis using multilevel techniques and Bayesian statistics in multilevel modelling (Goldstein 2003). “The major advantage of using multilevel analysis instead of classical meta-analysis methods is flexibility. In multilevel analysis, it is simple to include study characteristics as explanatory variables in the model. If we have hypotheses about study characteristics that influence the outcomes, we can code these and include them on a priori grounds in the analysis. Alternatively, after we have concluded that the study outcomes are heterogeneous, we can explore the available study variables in an attempt to explain the heterogeneity.” (Hox and de Leeuw 2003: p. 92).

These technical advantages and the flexibility of using 'multilevel' or 'random effects' meta-regression rather than 'fixed effect' meta-analysis can also be expressed in terms of a conceptual advantage. The fixed effect model assumes that variation between programs or heterogeneity is exclusively due to random variation and, therefore, if the data available about each program were infinitely large, the results would be identical (Egger and Smith 2007). Since fixed effect analysis considers a common effect across programs (Smith et al. 1995), this would be equivalent to assuming that each program is equally effective and that differences in effectiveness are really only due to random fluctuations, rather than to, for example, GDL program features. In other words, using a fixed effect model really implies an *a priori* choice that no differences in effectiveness between programs exist, regardless of the different composite features of each program.

A random effects model, on the other hand, "assumes a different underlying effect for each study and takes this into consideration as an additional source of variation" (Egger and Smith 2007: p. 35), which is "mathematically equivalent to assuming these effects are drawn from some population" (Smith et al. 1995: p. 2685). Since this study is predicated on the assumption that GDL programs can be improved through the enhancement of their different features, it seems preferable to at least test for the presence of heterogeneity and to use a random effects model (or multilevel model) to analyze the data accordingly, rather than to simply model the data using a fixed effects model. This approach was adopted in the current study.

6.2 Method

6.2.1 Study population and data

The analyses in this study use data from 46 American States, the District of Columbia and 11 Canadian jurisdictions (see appendix 1 for a list of jurisdictions). The timeframe is 1992 through 2006, inclusive. Jurisdictions were excluded from the analyses only if post data at the level of jurisdictions were not available, e.g., because the implementation of the GDL program occurred too recently (note that at least two years of data post-implementation were needed to calculate the outcome measure as explained below). These jurisdictions include Arizona (implementation in July 2007), Hawaii (implementation in September 2006), Montana (implementation in July 2006), Kansas (pursuing legislation in 2008) and Canada's North West Territories (implementation in 2005) and Nunavut (no GDL program in place).

If legislative changes to the initial GDL program were passed and took effect in a particular jurisdiction, this jurisdiction was included at least twice, i.e., once to reflect the original implementation of the program and once to reflect the legislative change. For example, the GDL program in B.C. was implemented in 1998 and improved in 2003, so two data points for BC were included in the master data file, and the corresponding outcome measure and independent variables for the jurisdiction were measured accordingly. As such, several jurisdictions have been included more than once (see appendices for a list of jurisdictions and the corresponding dates of implementation or legislative changes) so the master database contains 78 data points rather than 58 (46 States, Washington, DC and 11 Canadian jurisdictions).

Fatality rates were calculated separately for 16, 17, 18 and 19 year old drivers using counts of fatalities per age cohort and by jurisdiction from FARS data for U.S. jurisdictions and TRAID data for Canadian jurisdictions. Population data for each jurisdiction were obtained from the U.S. Census Bureau for U.S. jurisdictions (see www.census.gov/popest/archives/1990s/ for estimates for 1992-1999 and www.census.gov/popest/datasets.html for estimates for 2000-2007) and from Statistics Canada's 2007 Demographic Estimates Compendium for Canadian jurisdictions (see Statistics Canada 2007 for a detailed description of the methodology and the quality of Canadian population estimates). Both IIHS's and TIRF's websites were used to obtain descriptions of the GDL programs for each of the included jurisdictions in the analyses. This information was then coded into a set of 23 independent variables. These variables are described in more detail in Table 6.1 and were included as covariates in the meta-regression analyses.

6.2.2 Outcome measure

The outcome measure in this study, i.e., the dependent variable, was calculated as described in Altman and Bland (2003) and applied by Ulmer et al. (2000), Mayhew et al. (2001, 2002), Foss et al. (2001) and Shope et al. (2001). More precisely, for each jurisdiction eight numbers were obtained:

- the number of fatalities in the post period for the target group (a);
- the population in the post period for the target group (b);
- the number of fatalities in the pre period for the target group (c);
- the population in the pre period for the target group (d);

- the number of fatalities in the post period for the comparison group (e);
- the population in the post period for the comparison group (f);
- the number of fatalities in the pre period for the comparison group (g);
- the population in the pre period for the comparison group (h).

The target groups were 16, 17, 18 and 19 year old drivers. The reason why age groups were used as separate target groups is to help control for confounding by age, for example to control for differences in cognitive development between teenagers of those ages. In all cases, the comparison group was 25-54 year old drivers, who are assumed to be largely unaffected by the GDL program in a jurisdiction.

The post period was defined as a period of 12 months, starting one year after the implementation of the GDL program and ending two years after implementation. The pre period was defined as a period of 12 months, starting two years before the implementation of the GDL program and ending one year before the implementation. If the program was implemented more recently, the post and pre periods were adjusted accordingly – for example GDL was implemented in mid September 2005 in Wyoming so only 3.5 months of post information was available (mid September 2006 to 31 December 2006) and only 3.5 months of pre information was used (June 2004 to mid September 2004).

The reason why such timeframes have been chosen is because it has been shown that the implementation of a GDL program can disrupt the normal licensing patterns. For example, normal licensing patterns can be disrupted before GDL implementation due to novices trying to avoid the change and after GDL implementation due to the time needed for drivers to meet new requirements, as well as to progress through the different GDL stages. This may also affect crash patterns (see e.g., Mayhew et al. 1999, 2001).

The available information was then summarized into fatality rates for the target group ((a/b)/(c/d)) and the comparison group ((e/f)/(g/h)) and then into a fatality ratio (or relative fatality risk) by dividing each fatality rate for each target group (e.g., the fatality rate for drivers aged 16) by the fatality rate for the comparison group (i.e., the fatality rate for drivers aged 25-54). Note that only drivers who died in a fatal crash were counted for the numerators (a, c, e and g), while population numbers pertaining to the entire population and not just drivers were used for the denominators (b, d, f and h). This process of obtaining the fatality rates per age cohort and by jurisdiction and calculating the different relative fatality risks (four in total) for each of the 'jurisdictions' (78 in total) was automated in Stata, release 10 (StataCorp. 2007) with a variety of do-files and automatic do-files.

Using such a relative fatality risk as an outcome measure standardizes the fatalities of a target group to the population of that group as well as to the fatality rate of the comparison group. A ratio of less than one indicates a positive impact of GDL on the fatality risk of young drivers (i.e., a decrease from pre to post period) relative to the comparison group of older drivers in that jurisdiction. The standard error of this measure is used to calculate whether this positive impact is significantly different from the trend in the comparison group or not. Figures 6.1, 6.2 and 6.3 have been inserted to facilitate the interpretation of the outcome measure.

The different analyses used to summarize the data (described below) all assume that the outcome measure and its standard error are measured on the log-scale, so they have been transformed for further analyses and rescaled using the exponential function for the interpretation of the effect of the independent variables on the outcome measure.

Figure 6.1: A visual illustration of the interpretation of the outcome measure used in the meta-analysis; no effect of GDL (target group: 16 year old drivers; comparison group: 25-54 year old drivers)

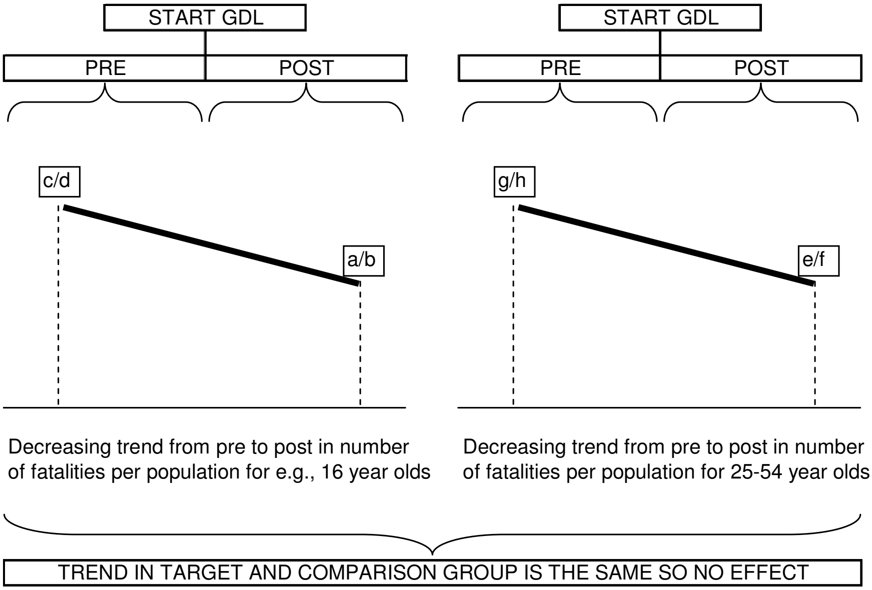

START GDL
PRE
POST
c/d
a/b
Decreasing trend from pre to post in number of fatalities per population for e.g., 16 year olds
START GDL
PRE
POST
g/h
e/f
Decreasing trend from pre to post in number of fatalities per population for 25-54 year olds
TREND IN TARGET AND COMPARISON GROUP IS THE SAME SO NO EFFECT

Figure 6.2: A visual illustration of the interpretation of the outcome measure used in the meta-analysis; a positive effect of GDL (target group: 16 year old drivers; comparison group: 25-54 year old drivers)

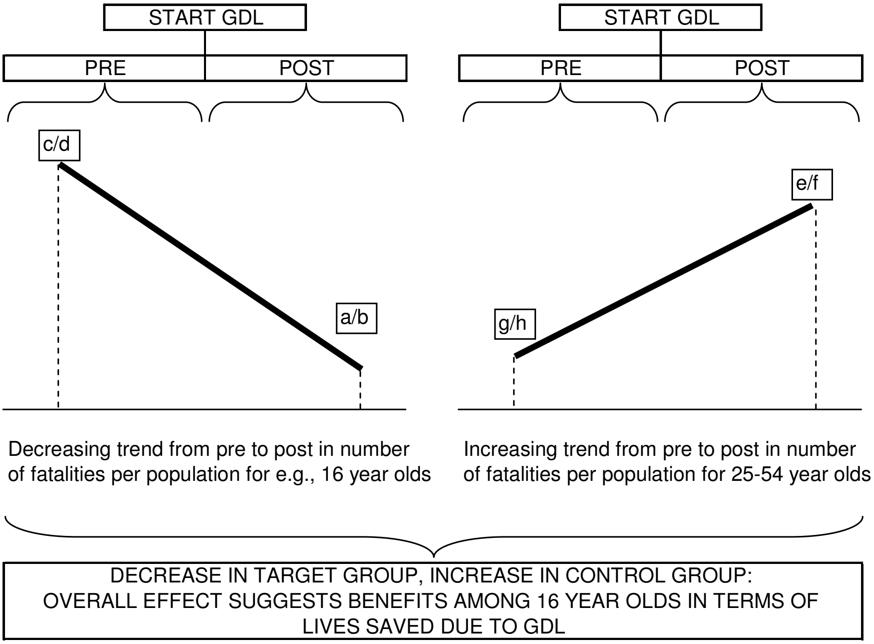
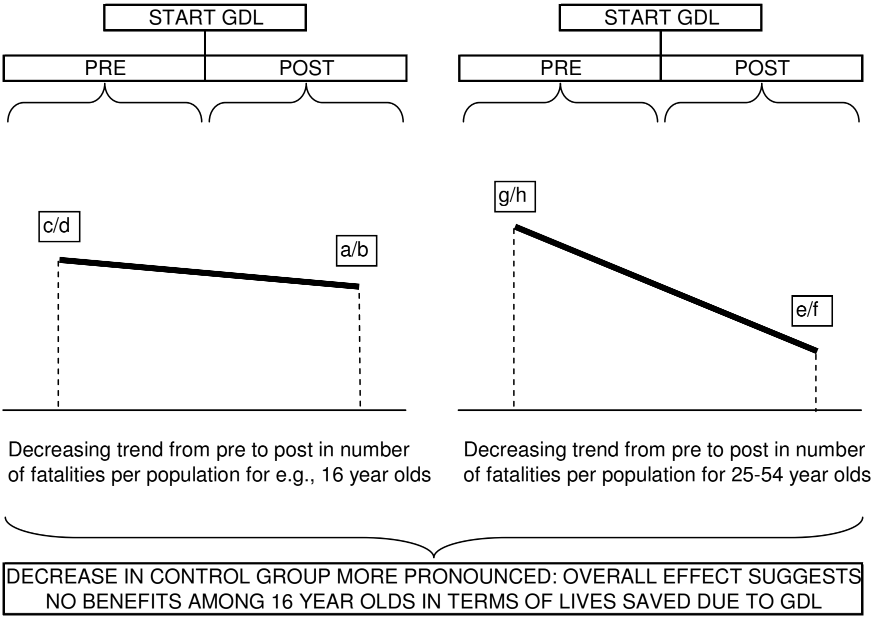

START GDL
PRE
POST
c/d
a/b
START GDL
PRE
POST
e/f
g/h
Decreasing trend from pre to post in number of fatalities per population for e.g., 16 year olds
Increasing trend from pre to post in number of fatalities per population for 25-54 year olds
DECREASE IN TARGET GROUP, INCREASE IN CONTROL GROUP: OVERALL EFFECT SUGGESTS BENEFITS AMONG 16 YEAR OLDS IN TERMS OF LIVES SAVED DUE TO GDL

Figure 6.3: A visual illustration of the interpretation of the outcome measure used in the meta-analysis; no effect of GDL (target group: 16 year old drivers; comparison group: 25-54 year old drivers)


START GDL
PRE
POST
c/d
a/b
START GDL
PRE
POST
g/h
e/f
Decreasing trend from pre to post in number of fatalities per population for e.g., 16 year olds
Decreasing trend from pre to post in number of fatalities per population for 25-54 year olds
DECREASE IN CONTROL GROUP MORE PRONOUNCED: OVERALL EFFECT SUGGESTS NO BENEFITS AMONG 16 YEAR OLDS IN TERMS OF LIVES SAVED DUE TO GDL

6.2.3 Graduated driver licensing programs

A description of the different North-American GDL programs included in this analysis can be found on IIHS's and TIRF's websites. Not surprisingly, there is a lot of variation among the different programs making it particularly challenging to formally describe them in the form of variables that can be included in a meta-regression analysis to investigate potential sources of heterogeneity in the outcome measure. Nevertheless, about two dozen variables were used in this study to capture such differences that can help explain which GDL features are more effective than others, and why. Table 6.1 contains a description of each of the independent variables used in this study.

Table 6.1: Description of the independent variables included in the meta-regression analysis (variable label; categories and frequencies for categorical variables and range and mean for numerical variables)

Variable label	**Categories/range (frequencies/mean)**
effective date of implementation or legislative change	1992-2005
IIHS's rating of the quality of the GDL program	good (28), fair (18),
	marginal (12), poor
minimum length in months of mandatory holding period learner stage	0-12 (6.0)
maximum length in months of mandatory holding period learner stage	0-48 (6.8)
minimum # of hours of supervisory driving required in learner stage	0-60 (22.0)
conditions under which supervisory driving occurs	0=no mandatory hours at night (43)
	1=mandatory hours at night (35)
length of night restriction in hours in learner stage	0-10 (1.3)
night restrictions lifted if supervised	0=no (71)
	1=yes (3)
restriction on passengers in learner stage	0=no restrictions (67)
	1=restrictions (11)
passenger limit lifted in learner stage if passengers are immediate family members	0=no (76)
	1=yes (2)
pasenger restriction lifted for family in learner stage if driver is accompanied by a licensed instructor and driver is in driver education	0=no (77)
	1=yes (1)
minimum entry age for learner stage	14-16 (15.3)
reduction in # of months of mandatory holding period for taking driver education (time discount)	0-8 (0.6)
country	0=US (47)
	1=Canada (12)
driver education requirements in learner stage	0=no requirements (35)
	1=driver education mandatory (17)
	2=time discount if driver ed. (8)
length of night restriction in hours in intermediate stage	0-10 (4.1)
night restrictions lifted for work purpose in intermediate stage	0=no (74)
	1=yes (2)
restriction on passengers in intermediate stage	0=no (31)
	1=yes (47)
passenger limit lifted in intermediate stage if accompanied by a qualified supervisor	0=no (74)
	1=yes (3)
passenger limit lifted in intermediate stage if passengers are immediate family members	0=no (41)
	1=yes (36)
minimum entry age for intermediate stage	14.5-17 (16.1)
driver education requirements in intermediate stage	0=voluntary (54)
	1=mandatory (2)
exit test required to graduate from intermediate stage to full stage	0=no (50)
	1=yes (7)

6.2.4 Model and data analysis

The data in this study were analyzed using three different approaches. First, in an exploratory phase, a summary effect was calculated for each target group and a test for heterogeneity conducted using the random effects DerSimonian and Laird model, as described in Deeks et al.

2007. This analysis was carried out in Stata, release 10 (StataCorp. 2007), using the metan-command. Furthermore, a cumulative meta-analysis was carried out as well in Stata, using the metacum-command and the same random effects model. The summary effect for each age cohort is based on the complete database of 78 'jurisdictions'.

In a second step, a meta-regression was carried out in Stata (using the metareg-command; see Harbord and Higgins 2008) to investigate the relative influence of the different independent variables on the outcome measure, using Restricted Maximum Likelihood (REML) — see Sterne et al. 2007a for details on the REML algorithm. The model used to summarize the data, expressed as a multilevel model has been described in section 3.2.3.

Finally, a full Bayesian analysis (Smith et al. 1995) was conducted in the MLwiN statistical package, using Markov Chain Monte Carlo (MCMC) Gibbs sampling (see Browne 2004, Lambert and Abrams 1995, Turner et al. 1999). The reason why REML and MCMC Gibbs estimates were used was to help increase the robustness of the findings. The meta-regression using the REML estimation procedure in the second step and the MCMC Gibbs sampling in this third step are both based on a sample of 48 'jurisdictions' (rather than 78), due to missing data for the variables included as covariates in these models. The results obtained from the random meta-regression in the second step (obtained with the REML procedure) are used as starting values for each of the parameters that need to be estimated with the MCMC Gibbs sampling procedure.

The length of each MCMC chain was set at 50,000 iterations with a burn-in period of 1,000 iterations. Each of the models was estimated four times with different random number seeds to ensure the results were stable when using different starting values (the results of these different models were found to be similar; see figures with random seeds 1 through 4 in appendices 5, 6 and 7). Diagnostics for each parameter were obtained as well to compare different models (see figures 6.5.1 through 6.7.2 in appendices 5, 6 and 7). These included graphs of Gibbs sampling traces to check for autocorrelation in these traces (see Browne 2004), the Raftery-Lewis diagnostic (Raftery and Lewis 1992) and the Brooks-Draper diagnostic (see Browne 2004) to assess the required length of the MCMC chains and the Deviance Information Criterion (DIC) according to Spiegelhalter et al. (2002), which is a generalization of Akaike's Information Criterion (AIC) to test the complexity of the different models and their goodness of fit. Finally, the different jurisdictions were also ranked according to their effectiveness, although this turned out to be meaningless because the credibility intervals (confidence intervals in Bayesian terminology are called credibility intervals) for each jurisdiction obtained with the MCMC Gibbs sampling all overlapped.

6.2.4.1 A note on a Bayesian approach to statistical inference

With multilevel models both a traditional and a Bayesian approach to statistical inference can be adopted. A Bayesian approach to statistical inference is based on a quite different notion of probability compared to the classical approach and this has far-reaching implications. As Raudenbush and Bryk (2002: p. 400) explain, in the 'classical' or 'frequentist' tradition "…parameters are fixed constants. The data of interest represent a probability sample from a population characterized by these parameters. The probability of a given sample result is its relative frequency of occurrence over many independent probability samples from the population. The laws of repeated random sampling underlie the machinery of inference..." In this case study the REML procedure that was adopted can be considered the classical or frequentist approach. The Bayesian approach, on the other hand, no longer views probability as relative frequency over many repeated samples. "Instead, probability quantifies an investigator's uncertainty about some unknown. The unknowns of most interest are the parameters of the distribution generating the data. Bayesians view these parameters as themselves having probability distributions. These distributions describe the investigator's uncertainty about the parameter values." (Raudenbush and Bryk 2002: p. 400)

As a consequence, testing a hypothesis using the frequentist paradigm is done by investigating whether a null hypothesis can be rejected, i.e., *reductio ad absurdum* (see e.g., Tacq 1997: p. 378). Using the *reductio ad absurdum* approach, it is assumed that the null hypothesis is correct and given this assumption it is calculated how likely it is to find the actual sample result. If the probability of finding the sample result is small (typically smaller than 5%), then it is concluded that the Null Hypothesis must be wrong and it is rejected. So, "We speak about the probability of a given event under the null hypothesis, not the probability that the hypothesis itself is true." (Raudenbush and Bryk 2002: p. 401). In the Bayesian approach, however, the actual probability of the Null Hypothesis will be calculated. This is done by combining prior knowledge (e.g., from previous studies) with new data (i.e., data collected in the current study). Revising this 'prior distribution' in light of the new data produces a 'posterior distribution' and it is this distribution that can be used to calculate probabilities of parameters of interest.

One advantage of the Bayesian approach is that the combination of the prior distribution with new data basically allows for incorporating existing knowledge in the data analysis. Such an iterative process ultimately formalizes a "sequential learning approach" (Browne 2004: p. 3) making it possible to take more information into account with each data run. Such prior knowledge could come from another quantitative study that was previously conducted or from expert opinion. If no prior knowledge exists, the prior distribution can be adjusted accordingly, for example by using low-information priors or simply by using results coming from REML estimates, which was done in this case study.

Carlin and Louis (2000) discuss several other advantages of the Bayesian approach, as does Hauer (2008), including the fact that the null hypothesis can indeed be 'accepted' rather than only 'failing to reject it'. One other advantage that was especially relevant for this case study is that complex models can easily be modeled and model fit statistics can be calculated. As such, it was possible to confirm that the multilevel model or random model fit the data better than the one-level or fixed model. MLwiN uses MCMC methods that can be considered more general than the traditional likelihood based frequentist estimation methods in that they can be used to fit many more statistical models. "They generally consist of several distinct steps making it easy to extend the algorithms to more complex structures." (Browne 2004: p. 4).

Despite several attractive features of adopting a Bayesian approach, researchers only recently started to use Bayesian analysis techniques more often, compared to the frequentist approach. This is due to computational challenges and obstacles that have only been overcome recently with the development of low-cost computers with increased memory capacity and processing speed. As a consequence, such techniques as MCMC can now be applied. Browne (2004) explains that, generally speaking, to calculate the posterior distribution directly, one has to integrate over many parameters and this proves to be particularly challenging. Therefore, MCMC simulates parameter values from the conditional posterior distribution, a distribution that produces equivalent results as those from the more challenging posterior distribution. The MCMC method used in this case study is Gibbs sampling, which works "by simulating a new value for each parameter… from its conditional distribution assuming that the current values for the other parameters are the true values" (Browne 2004: p. 6).

For a more detailed discussion about the advantages and disadvantages of a Bayesian approach used in step three of the analyses versus the more classical 'frequentist' approach, used in steps one and two of the analyses, see Carlin and Louis (2000). For a formal description of the Bayesian model used in this report, see Browne (2004) and Smith et al. (1995). For a technical description about how to use MLwiN to run these analyses, see Lambert and Abrams (1995) and Turner et al. (1999).

6.3 Results

6.3.1 Summary effect

6.3.1.1 Summary effect for 16 year old drivers

Appendix 1 contains a table (entitled "Table 6.1.1: Random Effects Meta-Analysis") showing the outcome measure for 16 year old drivers (ES), i.e., the relative fatality risk (see subsection entitled 'Outcome Measure'), its 95% confidence interval (95%-CI) and its weight for each jurisdiction included in the evaluation, as well as a pooled summary effect for all jurisdictions according to the random effects model. The weight is derived from the variance of each jurisdiction, which means that smaller jurisdictions with more variance will contribute less to the pooled estimate, while larger jurisdictions with less variance will contribute more.

As can be seen, the pooled summary effect for 16 year old drivers is 0.809, with a 95%-CI of (0.714-0.917). The null hypothesis of no effect is rejected (z=3.32, p=0.001). This means the evaluation provides strong evidence in support of GDL because the outcome measure is significantly smaller than one. More precisely, GDL has had a positive and significant impact on fatalities among 16 year old drivers — a decrease in the relative fatality risk of 19.1% ((1-0.809)*100) — when adjusting for a group of older drivers who are assumed not to have been affected by the implementation of GDL.

The results of the test of homogeneity are also displayed in Table 6.1.1 in Appendix 1 and indicate that the null hypothesis of homogeneity cannot be rejected (chi-square=81.17; d.f.=77; p=0.351). In other words, there is no evidence for heterogeneity (i.e., differences in the outcome measure that can be accounted for by differences among GDL programs) according to this test. However, the test of homogeneity has low power (Deeks et al. 2007; Thompson 2007), which means the possibility of a type II error (false negative, or, deciding there is no heterogeneity between jurisdictions while in reality there is) must always be considered. As such, "it is often more useful to quantify heterogeneity than to test for it" (Harbord and Higgins 2008: p. 499). The I-squared measure quantifies heterogeneity (Higgins et al. 2003) and shows that 5.1% of the variance in the outcome measure is due to heterogeneity. Potential sources of heterogeneity will be investigated using meta-regression analysis techniques in the next section.

Appendix 1 also contains a forest plot (see Figure 6.1.1). In a forest plot the contribution of each jurisdiction (its weight) is represented by the area of a square whose centre corresponds to the size of the effect estimated from that jurisdiction. The 95%-CI for the effect from each jurisdiction is also shown and the summary effect is represented by the middle of a diamond whose left and right extremes represent the corresponding confidence interval (Sterne et al. 2007a). This forest plot displays the same results from the table in this appendix, but in a different format.

Finally, a figure from a cumulative meta-analysis is inserted in Appendix 1 as well (see Figure 6.1.2). This figure shows that there has been consistent albeit not significant evidence in support of GDL, since approximately 1996, around the time when Virginia (VA) implemented its program. The pooled summary effect was smaller than one from then on (the horizontal lines are 95%-CIs, the ovals are the point estimates, and the vertical line corresponds to the overall summary effect). Note that findings from this cumulative analysis do not say anything about individual results from Virginia or any other jurisdiction per se. Since 2000, with the implementation of Oregon's program, the overall effect also became significantly smaller than one (95%-CI: 0.710-0.994) and stayed significantly smaller than one. Since then, the effect grew gradually stronger — albeit only slightly stronger — in favor of GDL, as can be seen both from the summary effect, which is more removed from one and from its 95%-CI, which has become narrower.

6.3.1.2 Summary effect for 17 year old drivers

Appendix 2 contains a table (Table 6.2.1) showing the outcome measure for 17 year old drivers. This table uses the same format as the one for 16 year old drivers.

As can be seen, the pooled summary effect for 17 year old drivers is 1.001. However, the 95%-CI for this age cohort (0.906-1.105) contains one, meaning the relative fatality risk is not significantly different from one, and the true effect could be greater or smaller than one. The null hypothesis of no effect is not rejected (z=0.02, p=988). This means the evaluation provides no evidence in support of, or against GDL based on data from 17 year old drivers. In other words, the available data do not support the contention that GDL has had a positive and significant impact on the number of fatalities among 17 year old drivers when adjusting for a group of older drivers who are assumed not to have been affected by the implementation of GDL.

The results of the test of homogeneity are also displayed in Table 6.2.1 in Appendix 2 and indicate that the null hypothesis of homogeneity cannot be rejected (chi-square=53.16; d.f.=77; p=0.983). While there is still the possibility of a type II error occurring, the I-squared measure now shows that none of the variance in the outcome measure is due to heterogeneity. This will have to be kept in mind when further investigating potential sources of heterogeneity — it is unlikely that further investigation of this subpopulation's data will reveal significant relationships between independent variables (i.e., the GDL program features) and the dependent variable (i.e., the effectiveness of GDL in terms of the relative fatality risk) due to this lack of variability.

A forest plot and results from a cumulative meta-analysis are also available in Appendix 2 (Figures 6.2.1 and 6.2.2). As can be seen from the cumulative meta-analysis, the point estimate of the pooled summary effect has been very close — almost equal — to one since 1997 with the implementation of Quebec's program.

6.3.1.3 Summary effect for 18 year old drivers

Appendix 3 contains a table showing the outcome measure for 18 year old drivers (Table 6.3.1). This table again uses the same format as the one for 16 year old drivers.

The pooled summary effect is now equal to 1.083 and the 95%-CI for this age cohort (0.978-1.199) contains one. The null hypothesis of no effect is not rejected (z=1.53, p=0.126). This means the evaluation provides no evidence in support of, or against GDL based on data from 18 year old drivers. In other words, the available data do not support the contention that GDL has had a positive and significant impact on the number of fatalities among 18 year old drivers when adjusting for a group of older drivers who are assumed not to have been affected by the implementation of GDL.

The results of the test of homogeneity indicate that the null hypothesis of homogeneity cannot be rejected (chi-square=86.63; d.f.=77; p=0.212). According to the I-squared measure 11.1% of the variance in the outcome measure is due to heterogeneity.

A forest plot and results from a cumulative meta-analysis are also available in Appendix 3 (see Figures 6.3.1 and 6.3.2). As can be seen in Figure 6.3.2, a pattern comparable to the pattern of 17 year old drivers is apparent in that the pooled summary effect has never been significantly different from one, with the exception of one time, namely after the implementation of New York's program in 1992 when the pooled summary effect was significantly smaller than one.

6.3.1.4 Summary effect for 19 year old drivers

Appendix 4 contains a comparable table (Table 6.4.1) showing the outcome measure for 19 year old drivers.

The pooled summary effect is equal to 1.059 with a 95%-CI of (0.963-1.165). The null hypothesis of no effect is not rejected (z=1.19, p=0.235). This means the evaluation provides no significant evidence in support of, or against GDL based on data from 19 year old drivers. In other words, the available data do not support the contention that GDL has had a positive and significant impact on the number of fatalities among 19 year old drivers when adjusting for a group of older drivers who are assumed not to have been affected by the implementation of GDL.

The results of the test of homogeneity indicate that the null hypothesis of homogeneity cannot be rejected (chi-square=81.48; d.f.=77; p=0.342). According to the I-squared measure 5.5% of the variance in the outcome measure is due to heterogeneity.

A forest plot and results from a cumulative meta-analysis are also available in Appendix 4 (see Figures 6.4.1 and 6.4.2).

6.3.2 The examination of heterogeneity among jurisdictions

In this section heterogeneity in the outcome measure is examined. Potential sources of heterogeneity have been formally described by a variety of variables (see Table 6.1) that have been included in a meta-regression analysis as covariates. As explained in the method section, two approaches were used to obtain the results. First, a random effects meta-regression analysis was run using REML. Then, MCMC Gibbs sampling was used to obtain full Bayesian estimates of the coefficients of the covariates. Results from both approaches are shown in Table 6.2 and compared in this section.

It warrants mentioning that only models for 16, 18 and 19 year old drivers are discussed in this section because variance in the outcome measure due to heterogeneity among jurisdictions was equal to zero for 17 year old drivers. The variance that was found for 16, 18 and 19 year old drivers on the other hand (5.1%, 11.1% and 5.5% respectively) may be low but makes it more likely to identify possible sources of heterogeneity (reasons why there is no variation among 17 year old drivers will be discussed later).

6.3.2.1 Meta-regression for 16 year old drivers

Table 6.2 provides an overview of the significant effects that were found for 16 year old drivers and compares the REML results with the MCMC results. These results come from a model that includes all covariates listed in Table 6.1. The full model can be found in Appendix 5 (see Table 6.5.1), as well as the MCMC diagnostics per significant parameter (see Figure 6.5.1).

Two effects were found to be significant according to the REML estimation and according to the MCMC Gibbs sampling (if a 95%-credibility interval does not contain zero, then the coefficient can be considered significant according to the MCMC Gibbs estimation). These effects are restriction on passengers in the intermediate stage and whether passenger restrictions are lifted in the intermediate stage if passengers are immediate family members. As can be seen, coefficients from both estimation procedures are very similar — the coefficients according to the MCMC Gibbs sampling are somewhat smaller.

To assist with the interpretation of those parameters, the coefficients can be transformed using the exponential function. The exponentiated coefficient for the first variable then becomes 0.115 (REML p-value=0.014). The interpretation is as follows: in a jurisdiction with passenger restrictions in the intermediate stage, the relative fatality risk of 16 year old drivers decreases by a factor of 0.115 or 88.5% ((1-0.115)*100), compared to jurisdictions without such passenger restrictions.

Table 6.2: Comparison of significant effects (on log-scale) according to REML and/or MCMC Gibbs for 16, 18 and 19 year old drivers

Variable	REML		MCMC Gibbs	
	Coefficient	*S.E. (p-value)*	*Coefficient*	*95%-Credibility interval*
16 year old drivers				
Passenger restriction in intermediate stage	-2.160	0.804 (0.014)	-2.102	-3.833;-0.364
No passenger restrictions in intermediate stage if passengers are family	2.114	0.794 (0.014)	2.011	0.237;3.762
18 year old drivers				
Driver education in learner stage mandatory	-0.423	0.189 (0.036)	-0.437	-0.876; -0.004
19 year old drivers				
Length night restriction in learner stage	0.104	0.047 (0.038)	0.102	0.020; 0.181
Country	2.587	1.374 (0.074)	2.543	0.391; 4.682
No night restriction in intermediate stage if work	3.953	2.089 (0.072)	3.966	0.675; 7.237
Driver education in intermediate stage Mandatory	0.746	0.395 (0.073)	0.731	0.057; 1.394
Exit test to graduate from intermediate stage	-3.856	1.612 (0.026)	-3.803	-6.383; -1.195

The transformed result for the second variable is 8.281 (REML p-value=0.014). The interpretation of this coefficient implies that lifting the passenger limit in the intermediate stage if passengers are immediate family members leads to a 728.1% increase ((8.281-1)*100) in the relative fatality risk of 16 year old drivers.

The model was further investigated by checking for outliers and the relative influence of individual jurisdictions to find an explanation for these rather extreme effects. No single jurisdiction had an effect that was particularly greater than that of the other jurisdictions and no outliers were found (see Figure 6.5.2 in Appendix 5).

Also, it was found that 73.7% of the variance between jurisdictions (i.e., heterogeneity) is explained by the covariates in the model and that only 2.5% of the residual variation of the model is due to heterogeneity; the remaining 97.5% is due to within jurisdiction variation (see Table 6.5.1).

Finally, It was argued previously that a random effects model (be it one using REML or MCMC Gibbs) would be more appropriate for a variety of reasons (see section entitled "Objectives and Rationale"). This was formally tested using the Bayesian DIC. The Bayesian DIC was 123.72 for the random model while it was 135.79 for the fixed model. The lower value for the random model confirms such a model is indeed more appropriate than a fixed model because it has a better fit.

6.3.2.2 Meta-regression for 18 year old drivers

Table 6.2 provides an overview of the results of the significant effect that was found for 18 year old drivers and compares both estimation procedures. These results too come from a model that includes all covariates listed in Table 6.1. The full model can be found in Appendix 6 as well as the MCMC diagnostics of the significant parameter (see Table 6.6.1 and Figure 6.6.1).

The transformed coefficient for the significant variable (driver education mandatory in the learner stage) is 0.655 (REML p-value=0.036) and can be interpreted as follows. The relative fatality risk of 18 year old drivers in jurisdictions where driver education is mandatory in the learner phase decreases by a factor of 0.655, or by 34.5% ((1-0.655)*100%), compared to those jurisdictions where driver education is not mandatory in the learner phase. This variable is significant according to both estimation procedures.

The model for 18 year old drivers was further investigated by checking for outliers and the relative influence of individual jurisdictions. No single jurisdiction had an effect that was particularly greater than that of the other jurisdictions and no outliers were found (see Figure 6.6.2 in Appendix 6).

This significant covariate, however, does not explain any of the variance between jurisdictions (i.e., heterogeneity) according to the adjusted R-squared statistic of this model and 11.43% of the residual variation is due to heterogeneity (see Table 6.6.1). This extreme value for R-squared can be explained by the fact that the model may not satisfactorily fit the data. This can be derived from the value of tau-squared, which is equal to zero. As such, this R-squared statistic bears no meaning and it is recommended not to rely on it. Note that the coefficient of the significant effect in this model is not necessarily adversely affected by this lack of model fit. In this regard, it is interesting to see that the Bayesian estimates — that are not affected at all by the potential lack of fit of the REML model — are very comparable to the REML estimates.

Finally, the Bayesian DIC was 110.13 for the random model while it was 115.82 for the fixed model. The lower value for the random model confirms such a model is indeed more appropriate than a fixed model because it has a better fit.

6.3.2.3 Meta-regression for 19 year old drivers

Table 6.2 provides an overview of the significant effects that were found for 19 year old drivers and compares the results from both estimation procedures. These results too come from a model that includes all covariates listed in Table 6.1. The full model can be found in Appendix 7, as well as the MCMC diagnostics per significant parameter.

When checking this model for outliers and the relative influence of individual jurisdictions, one outlier was identified (Maryland) — see Figure 6.7.2 in Appendix 7. Therefore, the results for 19 year old drivers are based on a model that excludes this outlier.

As can be seen, five effects were significant according to the MCMC Gibbs estimation procedure, but only two effects were significant according to the REML results. The first variable is length of night restriction in the learner stage. This variable is significant according to both estimation procedures. Its transformed coefficient is 1.11 (REML p-value=0.038). This means that for an increase in length of the night restriction in the learner stage of one hour, the relative fatality risk of 19 year old drivers increases by 11% ((1.11-1)*100).

The second variable is country and its transformed coefficient is 13.29. As can be seen in Table 6.2, this variable is only significant according to the MCMC Gibbs results (REML p-value=0.074). Its interpretation is as follows. The relative fatality risk of 19 year old drivers in Canadian jurisdictions is 1,229% ((13.29-1)*100) higher than that of 19 year old drivers in U.S. jurisdictions. Apparently, there is something different between Canada and the U.S. that has negative consequences for the relative fatality risk of 19 year old drivers in Canadian jurisdictions.

The transformed coefficient for the third variable is 52.09. This variable is also not significant according to the REML estimates (REML p-value=0.072). It means that the relative fatality risk of 19 year old drivers in jurisdictions where night restrictions are lifted for work purposes in the intermediate stage increases by a factor of 52.09 or 5,109% ((52.09-1)*100%), compared to the relative fatality risk of 19 year old drivers in those jurisdictions where such night restrictions are not lifted.

The fourth variable is only significant according to the MCMC Gibbs estimates. Its transformed coefficient is 2.11 (REML p-value=0.073). According to this variable there is an increase in the relative fatality risk of 19 year old drivers of 111% in jurisdictions with mandatory driver education in the intermediate stage ((2.11-1)*100).

The transformed result for the last variable (exit test) is 0.02 (REML p-value=0.026). This variable is significant according to both estimation procedures. The interpretation of this coefficient implies that the relative fatality risk of 19 year old drivers in jurisdictions that require an exit test to graduate from the intermediate stage is 0.02 times smaller or 98% smaller ((1-0.02)*100) than the relative fatality risk of 19 year old drivers in jurisdictions that do not require such an exit test.

According to the adjusted R-squared statistic of this model all the variance between jurisdictions or heterogeneity is explained (which implies that none of the residual variation is due to heterogeneity) — see Table 6.7.1. This extreme value for R-squared can again be explained by the fact that the model does not fit the data well enough for the R-squared statistic to be meaningful. The tau-squared statistic is also equal to zero. As a consequence, these model fit statistics should not be relied upon. Note again that the coefficients of the significant effects in this model are not necessarily adversely affected by this lack of model fit. Interestingly, the Bayesian estimates are again very comparable to the REML estimates.

Finally, the Bayesian DIC was 71.85 for the random model while it was 76.19 for the fixed model. The lower value for the random model confirms such a model is indeed more appropriate than a fixed model because it has a better fit.

6.4 Discussion

6.4.1 Limitations

When interpreting the results from the analyses in this study, some limitations of the applied methods have to be borne in mind. First, the scope of this study was limited geographically because only data from North-American jurisdictions were used, while other GDL programs exist outside of North America. However, the decision to include raw data from all North-American jurisdictions, rather than to use only published results of the relatively few GDL programs that have been evaluated, considerably broadened the scope of this study. Furthermore, if a more restricted meta-analysis had been carried out based on a systematic review of available evaluation studies, the study's scope would probably also have been limited to North America because the bulk of the available literature comes from evaluations in those jurisdictions.

Second, all the analyses are age-based. As such, it is implicitly assumed that 16, 17, 18 and 19 year old drivers are affected by GDL while 25-54 year old drivers are not. While such an assumption is true to a large extent, some bias may have been introduced in the analyses due to adopting such an age-based approach. For example, in most Canadian and a few U.S. jurisdictions GDL applies to all novice drivers, regardless of their age, which means that for those jurisdictions the comparison group of 25-54 year old drivers may contain some drivers that have been affected by GDL. Also, the denominator used in this study to calculate the relative fatality risks is the entire population of a particular age, rather than the number of licensed drivers of that age. It can be argued that the population truly at risk is the number of licensed drivers rather than the entire population. However, it is extremely difficult, if not impossible to obtain such information about the number of licensed drivers. Therefore, the population was used as a proxy for the population of drivers.

The post period in this study began 12 months after implementation of a GDL program until 24 months after implementation. As such, it can be expected that young drivers aged 16 and 17 would have gone through the new GDL system, as well as many 18 year old drivers and some 19 year old drivers. In Canada and a few U.S. states, GDL applies to new drivers of all ages and not all teenagers apply for a license immediately after they become eligible but may wait until they turn 17, 18 or 19. This means the evaluation period that was chosen likely captured young drivers of any age (i.e., 16, 17, 18 and 19) who have been exposed to the new GDL system. However, it is acknowledged that such an evaluation period may perhaps not have been long enough for a sufficient number of 19 year old drivers (and perhaps 18 year old drivers too) to already have gone through this new GDL system. This may explain why results showed no summary effect of GDL on these older teens. Another interpretation of the lack of a summary effect for older teens could be that too many of them did not go through the GDL system at all and that this cancelled out any effect that may exist among older teens that did go through the GDL system.

Furthermore, this age-based approach does not allow disentangling the effects of the different GDL stages to the same extent that an analysis based on license status would. This also makes it difficult sometimes to explain the reasons why an effect emerges. For example, an increase in crashes after lifting passenger limits when passengers are immediate family members could be due to the negative influence of too many passengers on inexperienced drivers, but it could also be due to exposure, i.e., perhaps novice drivers simply drive more often when such a passenger limit is lifted, which would expose them more to opportunities for crashing. The same is true for the effect that was found with respect to lifting night restrictions for work purposes. While ambitious, challenging, and perhaps not feasible, the results obtained in this study may improve and could be further explained if comparable analyses could be conducted using license status-based data, rather than age-based data, or if these results could be related to exposure data. On the other hand, some would argue that such age-based per capita rates are actually better suited to capture the overall effects of GDL.

The age-based approach makes it difficult to calculate a valid and reliable overall effect for all age groups taken together. This is especially true because in several jurisdictions many 18 and 19 year old drivers do not go through the GDL system and this may offset any effect among these older teens, as discussed previously. A more reliable overall estimate could be obtained by analyzing data from those jurisdictions only where eligibility for participation in a GDL program is not based on age but on experience, like in all Canadian jurisdictions and a few U.S. jurisdictions.

Third, only fatality data were used to investigate the impact of GDL. However, GDL may have a more pronounced effect on injury and property damage only (PDO) crashes involving young drivers because injury and PDO crashes are so much more prevalent than fatal crashes. As such, there could be more to gain in terms of preventing such crashes. The decision to use fatality data, however, was made in light of the search for a common denominator among all North-American jurisdictions and because only standardized fatality data for the U.S. were available.

Fourth, as explained in the method section, certain jurisdictions were included in the analyses more than once to reflect any legislative changes that were passed after the implementation of a GDL program. As such, the final database contained information from 78 'jurisdictions', rather than from 58. The upside of such an approach is that more data are used, which improves precision of the estimates and also enables a more nuanced investigation of differences between programs — when a program is improved it should logically lead to an additional reduction in fatalities, above and beyond the already established reduction due to the implementation of the original program, assuming other relevant circumstances in the jurisdiction have not changed too much. However, the downside is that the observations in the database (or 'data points') for such jurisdictions are dependent inasmuch as improvements to a program are conditional on the features of the already established program. While this may have adversely affected the results, it should be noted that all the data points in the database are dependent, at least to some extent, because it is unlikely that a GDL program would have been implemented in a vacuum, entirely independently from what is being done in other jurisdictions.

In this regard, it had previously been suggested that there may be a learning effect in that jurisdictions that are moving forward with the implementation of a GDL program can benefit from lessons learned in other, pioneering jurisdictions; or, that earlier GDL programs can create a more receptive climate for the implementation of stronger GDL programs. To further bolster this possibility, the different GDL programs were sorted chronologically and for each program the number of components that can make a program more effective were counted. The results can be found in Appendix 8, Table 6.8.1. As can be seen, the number of components increases gradually over time. For example, the average number of components until 1999, inclusive is 5.9, while the average number from 2000 until 2005, inclusive is 7.8. While a significant learning effect was found in earlier versions of the model with 16 year old drivers during an exploratory data analysis phase, unfortunately this could not be confirmed with any of the final models. As such, when looking at how programs evolved over time there seems to be some evidence to suggest there is a need for dissemination of study results as other jurisdictions may benefit from it, but it was not possible to confirm this formally with the meta-analyses.

Finally, while information from 78 jurisdictions was collected, the true sample size of the meta-regression analyses was only 48, due to missing data for several of the covariates. It proved extremely challenging and very labor-intensive to obtain complete GDL data records for each of the jurisdictions included in this study, despite the availability of information about these programs on two on-line inventories (IIHS's and TIRF's) and despite the access to experts in the field.

6.4.2 Findings

Using summary effects coming from a random effects DerSimonian and Laird model, strong evidence in support of GDL reducing fatalities was found. The evidence, however, only showed that GDL has had a positive and significant impact on fatality rates among 16 year old drivers, when adjusting for a group of older drivers who are assumed not to have been affected by the implementation of GDL (reduction in the relative fatality risk of 19.1%). No evidence was found to suggest GDL has had an overall impact on relative fatality risks among 17, 18 and 19 year old drivers when looking at the summary effects only.

The results also showed that there was some heterogeneity among jurisdictions that could provide insight into how certain GDL components have significantly affected fatalities among 16, 18 and 19 year old drivers. None of the variance in the outcome measure for 17 year old drivers, however, was due to heterogeneity, which is hard to explain. Compared to the patterns of variance that emerged among 16, 18 and 19 year old drivers (5.1%, 11.1% and 5.5% respectively), this lack of variance among 17 year old drivers seems like an anomaly, more than anything else. If it would be assumed that there truly is no variance among 17 year old drivers, this would mean that all GDL programs are equally effective for this age group. Given that most GDL programs differ considerably from one another and that 17 year old drivers in this study may have been in the learner, intermediate, or full license stages due to the nature of population data, it seems highly unlikely that this assumption holds true. Perhaps the composition of this group differs from the other age groups in that there is a greater variety of drivers in terms of license status among 17 years old drivers than there is among 16, 18 and 19 year old drivers. For example, the majority of 16 year old drivers may be in the learner stage, and the majority of 18 and 19 year old drivers may be in the full stage, whereas all stages, including the intermediate stage, could be equally represented among 17 year old drivers. However, such an explanation cannot be tested due to the limitations of age-based data and, as a consequence, is only speculative.

As for 18 and 19 year old drivers, the summary effect may not have suggested GDL has had a positive impact, but the proportion of variance due to heterogeneity (11.1% and 5.5% respectively) suggested it may still be possible to distinguish GDL programs which have had a significant impact on relative fatality risks among 18 and 19 year old drivers from those GDL programs that have not had such an impact.

Two variables were significant for 16 year old drivers, namely passenger restrictions in the intermediate stage and whether passenger restrictions are lifted or not in the intermediate stage if passengers are family members. Both variables were significant according to both estimation procedures (REML and MCMC Gibbs).

The first variable suggests that passenger restrictions in the intermediate stage are beneficial in that such restrictions lead to a 88.5% decrease in the relative fatality risk of 16 year old drivers. Conversely, the second variable suggests it is beneficial not to lift passenger limits in the intermediate stage if passengers are family members because lifting such restrictions leads to an increase in the relative fatality risk of 728.1%. Even though it could be argued that some of the 16 year old drivers are in the learner stage, certainly there are 16 year old drivers who have already graduated to the intermediate stage. This finding may suggest that restrictions should not only be license status based but instead be based on a combination of license status and age as lifting such a restriction for 16 year olds, even if they have already graduated to a more advanced stage, proves to have negative consequences.

One variable (mandatory driver education in the learner stage) was significant with 18 year old drivers according to both estimation procedures. Mandatory driver education in the learner stage may be beneficial in that it can lead to a 34.5% reduction in the relative fatality risk of 18 year old drivers. It is noteworthy that this finding stands in marked contrast to what has previously been reported in the literature (see for example Mayhew 2007 for an overview). Keeping the limitations of this study in mind, such a finding should be further investigated, especially in relation to time discounts — i.e., by taking driver education a driver can obtain a time discount in certain jurisdictions and graduate sooner to the next stage. The effect of time discount was not significant in this study. It would be beneficial to have a better understanding of the effect of offering a time discount. For example, if it encourages drivers to take driver education and if driver education truly has a protective effect, then the net effect of a time discount could be positive in that it would help better protect teens from crashing. However, if the effect of graduating sooner to the next, less protective stage would outweigh any beneficial effect of taking driver education, the net result may be negative, simply because teens will be exposed too soon to conditions they are not capable of dealing with. Based on the available evidence to date, it appears as if time discounts primarily serve to expose teens too soon to risky driving conditions and their crash risk increases rather than decreases as a result. For example, Williams and Mayhew (2008) mention several jurisdictions where such an increase has been noticed including Ontario (45% increase in crashes with time discount), Nova Scotia (27% increase) and B.C. (45% increase) — see also Mayhew (2007) for an overview.

Five variables were significant for 19 year old drivers, but not all of them were significant according to both estimation procedures (findings that are only significant according to one procedure could be considered less robust). The first variable (length of night restriction in the learner stage) was significant according to both procedures. According to the results for this variable, an increase in the length of the night restriction in the learner stage of one hour leads to an 11% increase of the relative fatality risk of 19 year old drivers. Assuming the majority of drivers in this age cohort would be in the intermediate or full stage, this may mean that too little night-time driving practice in the learner stage may have negative consequences for drivers entering the less restricted intermediate or full stage, although the limitations of aged based data, as discussed in the previous section, do not allow confirming whether such a residual effect is truly at work or not. For this to be true, a sufficient number of 19 year old drivers would have had to have gone through the new GDL licensing system, which is not certain given the post period that was used in this study. The status of this interpretation is therefore only speculative.

The variable country was only significant according to one procedure. According to this variable, the relative fatality risk of 19 year old drivers in Canadian jurisdictions is 1,229% higher than that of 19 year old drivers in U.S. jurisdictions. Because the results for this variable are controlled for numerous other variables that describe GDL programs, it is likely that other non-GDL related features, specific to Canadian jurisdictions and different from U.S. jurisdictions may explain this extremely high increase in the relative fatality risk of 19 year old drivers in Canadian jurisdictions. Further investigation is needed to better understand this result.

Evidence was found against lifting the night restrictions in the intermediate stage for work purposes, as it is associated with a 5,109% increase in the relative fatality risk of 19 year old drivers. This variable, however, was only significant according to one procedure. Also, note that only two jurisdictions lift night restrictions for work purposes. It seems GDL programs could be enhanced if stricter nighttime driving restrictions were applied. This is consistent with previous findings (see Williams 2007).

The fourth variable (mandatory driver education in the intermediate stage) also was only significant according to one estimation procedure. Its interpretation is as follows. There is an increase in the relative fatality risk of 19 year old drivers of 111% in jurisdictions with mandatory driver education in the intermediate stage. While this finding may seem counterintuitive, it is more in line with other research (see Mayhew 2007). Note that only two jurisdictions have mandatory driver education in the intermediate stage. Further investigation is needed to better understand this result.

Finally, requiring an exit test to graduate from the intermediate stage is beneficial since it leads to a 98% reduction in the relative fatality risk.

It is noteworthy that model fit for each model was further investigated using a variety of model fit statistics, including the proportion of residual variation due to heterogeneity, the proportion of heterogeneity explained by the covariates in the model (adjusted R-squared), the Bayesian DIC and checks for outliers and the relative influence of individual jurisdictions. Model fit statistics for the model with 16 year old drivers indicated a very good fit, while model fit was less good for 18 and 19 year old drivers. Taken together, however, these statistics suggest that the models for 18 and 19 year old drivers fit the data reasonably well. In this regard, it was interesting to see convergence between the coefficients coming from both estimation procedures (REML and MCMC Gibbs).

In conclusion, the meta-analysis used in this study has proven to be useful. Adopting a multilevel modeling approach in this analysis enabled further investigation of the effect of GDL components on relative crash risks — without such a multilevel approach, this would not have been possible. Despite the limitations of the study design, some previously established findings have been confirmed and other interesting and intriguing new findings emerged from these analyses. Several of the results were only significant according to one approach and perhaps they should only serve as exploratory findings that need further investigation and confirmation. Other results were strong and highly significant according to both approaches that were adopted. Such results should also be further investigated and confirmed. In that regard, caution is warranted when interpreting the results from this study. For example, some effects appeared to be very strong, but the accompanying confidence intervals were wide. As such, real effects probably exist, but more research is needed to more reliably estimate the strength of these effects.

While this is true for several effects, it is clearly illustrated particularly with the findings about passenger limitations in the intermediate stage for 16 year old drivers. These findings are remarkable because they corroborate the evidence in favor of passenger limitations found in other studies. However, they are also remarkable because evidence also exists for nighttime restrictions while none of the coefficients in the model for 16 year old drivers pertaining to nighttime restrictions turned out to be significant. The contrast between these extreme effects for passenger restrictions on the one hand and the absence of any significant findings for nighttime restrictions on the other in the model for 16 year old drivers is curious in light of the existing evidence for both types of restrictions. Reasons for this contrast and others (e.g., the existence of evidence in favor of driver education in one model and against driver education in another model) are not entirely understood based on the current findings. Further research along these lines is necessary to help overcome the limitations of this study and the resulting lack of understanding with respect to the effect of certain GDL components. As such, a follow-up study using a comparable approach would be useful and promising.

7. Conclusions

The goal of this Ph.D. thesis was to demonstrate the applicability, usefulness and added value of multilevel models in the field of traffic safety. This doctoral thesis was predicated on the belief that a family of techniques known as multilevel models can be considered a flexible solution to overcome a particular set of limitations of classical analysis techniques. To achieve this goal several sub-goals were formulated:

- First sub-goal: Introducing multilevel models (chapter 3);
- Second sub-goal: Justifying the use of multilevel models (chapter 3);
- Third sub-goal: Applying multilevel models (chapters 4 through 6);

Chapter 3 described some multilevel models in order to achieve the first sub-goal, including the basic two-level random intercept and random slope model, the two-level binomial model and a multilevel model for meta-analysis. This choice of models was guided by the requirements to analyze the data in the case studies later on.

Once these models were described the rationale for using multilevel models was explained in chapter 3 to achieve the second sub-goal. The common concept shared by different definitions of multilevel modelling was identified. Each definition defines multilevel modelling in its own way but they all refer to hierarchies. It was illustrated that such hierarchical patterns are very common in the social and behavioural sciences in general and in traffic safety in particular and that these patterns often occur naturally. Conceptual issues inherently related to hierarchies were then discussed.

It was argued that, broadly speaking, there are two important consequences of ignoring a hierarchical structure in the data. The first consequence, underestimation of standard errors, is related to the dependence of nested observations. Data from an observational study on seatbelt behaviour and a study on sleepiness at the wheel were analyzed according to a single-level model and a multilevel model to illustrate this. Effects that were significant at the 5%-level in the former model were no longer significant in the latter. The single-level model can therefore lead to erroneous conclusions regarding variables that could have an impact on the level of traffic safety.

The second consequence, related to the nature of contextual information and impoverished conceptualisation of the research problem, stems from the existence of variables on different levels of aggregation and from possible interactions between those levels. Data from a roadside survey on drinking driving were analyzed according to a two-level model. Of particular interest was the relationship between traffic count, an aggregated explanatory variable, measured at the level of road sites and odds of drinking driving, a dependent variable, measured at the level of drivers. This relationship, while having relevance to inform the drinking driving enforcement policy, could not have been studied properly without a technique capable of dealing with variables at different levels and cross-level interactions between them.

In summary, the reasons for using multilevel models are the following:

- Multilevel models can easily deal with design effects of a sample by taking account of the sampling design when analyzing the data. It could be argued that other types of models exist that too can deal with design effects, as explained previously. This is true but such procedures treat the variance structure in the sample as 'noise' rather than something that is of interest, in and of itself, and that can lead to substantial findings if analyzed properly.
- Further regarding such substantial findings, multilevel models overcome issues related to limitations of aggregated or disaggregated analyses. Often reality cannot be captured in terms of an aggregated or disaggregated analysis, but only a combination of both in one approach will be satisfactory. For example, higher-level variables such as traffic count, speed limit, number of liquor outlet stores per capita, etc. do affect behahiour at the individual level, but without techniques that can combine variables of different levels into one model, such patterns cannot be revealed and understood.
- Multilevel models can be used to model hierarchical data in a very efficient way. A simpler approach to model a two-level sampling design, for example, with cars nested in road sites would be to include dummy variables for road sites in a one-level model. The model would then contain n-1 dummy variables, with n equal to the number of road sites. This approach is not as efficient as multilevel models since the number of dummy variables can easily increase to a level that may not be manageable. This could lead to estimation issues and it is not parsimonious. Furthermore, unlike multilevel models this approach would also not allow modelling the influence of higher-level independent variables on lower-level dependent variables, or any cross-level relationship for that matter.

Next, chapters 4 through 6 consisted of a series of case studies in which different multilevel models were applied to a specific topic in the field of road safety. The first case study on the influence of traffic count on drinking driving behaviour was discussed. This case study further elaborated on the issue of impoverished conceptualization and replicated the findings that were used to illustrate the consequences of ignoring hierarchies in chapter 3 with comparable information from another jurisdiction. First, the relevance of traffic count was discussed in more detail. The case was made that information on traffic count would be useful for police officers to help them organize roadside checks to either prevent drinking driving or to catch drinking drivers in an efficient fashion. Given limited resources and competing priorities among police today, such information would be valuable. Then, the model from chapter 3 was replicated and the findings were presented and discussed.

The difference between the illustration in chapter 3 and the case study in chapter 4 is the weight of the evidence. The evidence discussed in chapter 3 comes from just one dataset and should only be considered exploratory. However, because the findings from this exploratory study warranted further investigation, more data were used in an attempt to replicate the results. This was done successfully and the case for studying the influence of traffic count, and in extension, of other higher level independent variables (e.g., other variables that characterize road sites such as availability of alcohol), became stronger. It is important to emphasize that this important information about the relationship between traffic count and drinking driving behaviour could not have been obtained without a technique such as multilevel models because variables that are measured at different levels have to be included in one model.

In the second case study, a two-level binomial model was fit to data on sleepiness at the wheel, once using 2003 data and once using 2006 data. While the results from the 2003 model were not always consistent with results coming from the 2006 model, and while the overall conclusions were not always consistent with what can be expected based on the literature, it is important to highlight that building both models without violating assumptions regarding the sampling design used to obtain these data would not have been possible, at least not to the same extent, without a technique such as multilevel models. Some would argue that comparable results could have been obtained without using multilevel models and by simply using a single level analysis that accounts for the sampling design; and, that such an approach has been possible much longer than using multilevel models. While this is true, it would not have been possible to model these data in as efficient a way as multilevel models do (several dozens of dummy variables would have had to be used to account for each of the road sites; as explained previously, this would be very inefficient). Also, it would not have been possible to treat higher level variables such as interview time as a higher level variable and combine it with lower level variables in one model.

The last case study conducted a meta-analysis of the effectiveness of GDL programs. What differentiates this study from the other case studies is the unit of analysis: the data are summarized at the level of jurisdictions (typically using results from evaluation studies but in this case using raw data from jurisdictions), rather than at the level of individuals. It was explained that with multilevel models (or random effects models) it is easy to include characteristics of GDL programs as explanatory variables in the model. This is not possible with a single level meta-analysis or fixed meta-analysis.

However, this technical advantage can also be interpreted in terms of a conceptual advantage. The single level meta-analysis assumes that variation between GDL programs is exclusively due to random variation and, therefore, if the data available about each program were infinitely large, the results would be identical. This would be equivalent to assuming that each

GDL program is equally effective and that differences in effectiveness are really only due to random fluctuations, rather than to, for example, GDL program features. In other words, using a single level or fixed effect model really implies an *a priori* choice that no differences in effectiveness between programs exist, regardless of the different composite features of each program, which would constitute the epitome of impoverished conceptualization. A multilevel meta-analysis (or a random effects meta-analysis model), on the other hand, is based on another assumption, more precisely such an approach assumes there may be differences in effectiveness between different GDL programs and provides the tools to investigate this. Without a multilevel approach, it would have been impossible to study the influence of the different features of GDL programs on their effectiveness.

In conclusion, there are several reasons to justify the use of multilevel models. While it has been demonstrated that this family of techniques is especially useful to deal with a particular set of issues (i.e., design effects, the importance of contextual information, efficiency, and the need for a parsimonious model), it is acknowledged that multilevel models are just one possible solution to these problems and that they may not be as suitable for dealing with other sets of problems. Also, even when dealing with problems that are suitable for multilevel models, there are limitations. In sum, multilevel models are not a panacea, but it has been demonstrated in this thesis that they do provide an efficient solution to model hierarchical data. In a nutshell, applying multilevel models in this thesis made it possible to generate new knowledge, especially about such issues as drinking driving, sleepiness at the wheel and GDL. Without such models this knowledge could not have been produced in an equally productive fashion.

The last sub-goal of this Ph.D. is to define a research agenda. This is the topic of the next and final chapter. Special attention will be given in this last chapter to practical implications of the findings and their social relevance, both with respect to the applied methods and the actual findings from the analyses.

8. Recommendations

Recommendations are formulated at two distinct levels. First, recommendations will be identified with regard to multilevel models; second, recommendations will be formulated with respect to each of the topics that were investigated more thoroughly in the case studies. Those latter recommendations are either based on policy implications of the findings discussed in the case studies or on research implications, more precisely recommendations for further research.

8.1 Recommendations regarding multilevel models

It has been demonstrated in this Ph.D. thesis that multilevel models can be useful in traffic safety research. As explained at the outset of this thesis (see section 3.3) multilevel models are particularly useful for hierarchical data and such data are common in the field of traffic safety. Some straightforward examples of such hierarchical data were briefly described. However, multilevel models can be useful in many more instances including research on education, for example where a particular driver education curriculum is taught in many different classes with different instructors across schools (aspiring drivers are nested in classes and classes are nested in schools) or research on injuries with patients seeing physicians in different hospitals (road crash victims are nested in clusters seeing the same physician and physicians are nested in different hospitals).

Multilevel models are also useful with less straightforward examples of hierarchical data, for example with latent class data. Latent class analysis is a technique used to model the relationship between independent variables that can be directly observed such as gender and age on the one hand and a dependent variable that cannot be directly observed, i.e., a latent variable. Such a latent variable can be considered a rather abstract construct that can only be measured by looking at several observable indicator variables. A good example of such a construct is aggressive driving. “Aggressive driving encompasses a continuum of behaviours that range from extreme acts (e.g., shootings or malicious assaults) to less severe manifestations (e.g., roadside arguments, confrontations, and gestures).” (Beirness et al. 2001: p. 4) Measuring such a concept will generate data that are hierarchical in nature in that observed indicator variables at one level are used to measure an underlying — i.e., unobservable — concept at another level. Vanlaar et al. (2008b) used a generalized linear latent model to investigate the profile of aggressive drivers using six indicator variables that can be directly observed to measure 'aggressive driving'. The observable indicator variables consisted of the self-reported frequency of a variety of driving behaviours that can be considered aggressive including speeding, red-light running, making rude signs and gestures at other drivers, taking risks for fun, using the horn and swearing. By using such a model it is recognized that “aggressive driving” is an abstract construct that can only be observed through several indicators; and that each of these indicators by themselves may not be reliable, but taken together the measurement of aggressiveness while driving becomes more reliable.

It warrants mentioning that latent models have long been studied and understood within the framework of structural equation models, which is another family of techniques altogether. However, it appears that multilevel models are a flexible solution that can deal with many different situations; multilevel models can serve as an umbrella to search for commonalities and to help conceptualize different data problems and their appropriate solutions. This is not to say that every data problem should be confined to a multilevel one but examining the same problem from a different perspective can be enlightening and should be encouraged.

Given that it was demonstrated in this Ph.D. thesis that multilevel models are useful in traffic safety research, knowledge about this family of techniques should be disseminated to help increase their use. More generally speaking, such an attempt should be part of a broader approach to increase the critical mass about sophisticated methods and analysis techniques among researchers in traffic safety.

Research could be more productive if such sophisticated techniques and software tools would be more widely available, understood and used. Since traffic safety, and more generally speaking transportation, has traditionally not been anchored in the organizational structure of universities and schools, and since transportation and traffic safety departments or institutes have only recently begun to emerge, an international forum such as a centre of excellence for the enhancement of traffic safety and transportation data and data analysis could serve as a catalytic impetus to obtain this goal.

8.2 Recommendations regarding traffic safety policy and further research

8.2.1 The influence of traffic count on drinking driving

Given that findings with respect to the influence of traffic count on drinking driving behaviour were replicated with a sample of data that has been collected entirely independently from the original data that were used (i.e., one sample was obtained in Belgium and the other in B.C., Canada by an independent research team), there seems to be convergent evidence to bolster the claim that there is a true relationship between both variables. Often police officers make a conscious choice about the objective of their enforcement campaigns; some prefer to use their campaigns primarily to prevent drinking driving from occurring, while others prefer to catch as many drinking drivers as possible. Regardless of their choice, police officers often simply select road sites with high traffic counts "to make it worth their while". However, the information coming from this case study suggests it may be more productive to select road sites with lower traffic counts. The information from this case study should be disseminated among police officers, if not to convince them to use it right away when designing their enforcement plans, then at least to stimulate a discussion about ways to help render their efforts more cost efficient. Such an approach certainly seems justified in this era of evidence based practices in conjunction with limited resources and competing priorities.

Given the limitations of the evidence — there is convergent evidence but still only from two studies — disseminating this information should perhaps first be done on a small scale among police officers who would be interested in further investigating this. If an alliance could be forged with a police force willing to invest some resources in this, the importance and usefulness of traffic count could be further investigated, along with other important contextual variables, such as the distinction between urban and rural communities, the availability of alcohol (operationalized for example as the number of liquor outlet stores per capita or the number of bars per capita), percent of drinking drivers caught per road site in previous years, enforcement efforts per road site, etc. If such information could then be used to design enforcement campaigns and to evaluate the cost-effectiveness of the approach, this could lead to better informed enforcement policies and allocation of resources to deal with the issue of drinking driving.

Also, while the data coming from the roadside surveys used in this study provide some evidence about the behaviour of drinking drivers, they contain little or no insight into reasons for this behaviour or motivations of drinking drivers. One hypothesis is that drinking drivers tend to avoid high traffic count road sites because they believe their chances for getting caught are greater at such sites. An alternative hypothesis could be that drinking drivers tend to avoid congestion or more difficult driving situations. To bolster the findings regarding the relationship between traffic count and drinking driving behaviour, interviews have to be conducted with caught (or convicted) drinking drivers. During such interviews their motivations and rationale with respect to drinking driving could be further explored.

8.2.2 Sleepiness among night-time drivers

It is becoming more and more apparent that sleepiness among night-time drivers is an important traffic safety issue. If anything, this case study illustrated the need for more information about this. The results in this case study provided some answers but also raised new questions because the findings were not always consistent, either between the two datasets that were used or between the findings from this case study and the literature. Epidemiological research findings that complement laboratory research and what is known from the literature would be invaluable. As such, replicating this epidemiological study and elaborating on the issue of sleepiness among night-time drivers is recommended.

One way to improve the results of this epidemiological study would be to use a more objective measure of sleepiness, rather than self-reported subjective sleepiness. This could be done, for example by using a valid and tested measurement instrument such as the Epworth Sleepiness Scale (ESS), designed by the Sleep Disorders Unit of the Epworth Hospital in Melbourne, Australia (Johns 1991). Such an objective measure could provide insights into the relationship between BAC and sleepiness and gender and sleepiness. The results with respect to these relationships that were obtained in this epidemiological study were rather inconsistent.

Also, more research is needed about the impact of passenger configuration on sleepiness. It is recommended that this relationship be studied from a social dynamics point of view. This could shed light on the question of whether drivers with a particular passenger configuration are reporting different levels of sleepiness because they truly are more or less sleepy or because of over- or underreporting sleepiness.

While there certainly is a need for more research on the effects of sleepiness and fatigue on traffic safety, the available data do suggest action is needed. Education can be used to raise awareness among the public as there is some evidence to suggest that the public's level of concern about sleepiness and fatigue at the wheel is incongruent with the dangers posed by it (see e.g., Vanlaar et al. 2008a, 2008c). Education and support for enforcement may be needed as well especially since information about this issue can help police officers target and enhance enforcement efforts. For example, in a recent survey of some 800 police officers in Ontario, conducted by TIRF, more than half of those surveyed (56.6%) felt that they had not received adequate information about ways to identify drivers who are sleepy or fatigued or to determine the role of fatigue in crashes (Robertson et al. 2009). Other solutions include technology such as rumble strips or in-vehicle warning systems.

8.2.3 Graduated driver licensing

With respect to the case study on GDL, the following recommendations can be formulated. First, a feasibility study would be useful to gauge if it would be possible to conduct a large scale project to replicate the findings of the present study using license status based data rather than age based data and also to estimate at what cost this could be done. This would imply gaining access to driver records systems of each of the jurisdictions included in this study (and preferably jurisdictions elsewhere in the world) and would involve sophisticated and complex data manipulation. This seems very ambitious and such an approach could probably only succeed by relying on several teams of data-analysts who would collaborate under the guidance of a coordinator. If feasible, however, such a project could produce highly relevant and important results and overcome several of the limitations of this case study, as described previously.

In the interim, a follow-up study should be conducted to complete data records of several jurisdictions and to refine the formal description of the GDL programs by means of covariates. Collecting more data for each of the covariates with missing values could increase precision of the estimates and reveal more patterns. For example, this case study found some limited evidence of a learning effect, showing that more recent GDL programs had become more sophisticated and effective, probably because they could benefit from the experience of older GDL programs. Such a finding is important because it means jurisdictions may improve their programs if evidence is available. However, the evidence for this learning effect was weak and the final models in this case study did not confirm that such an effect truly exists. Such results could become statistically significant if more data would be available and the power of the sample would increase. Refining the formal description of GDL programs could also reveal more patterns between independent and dependent variables. This could be accomplished by establishing a panel of experts who could discuss and refine the definitions of the covariates, elaborate on the list of covariates, and use their resources to obtain more data. Comparable analyses could be re-run and could lead to uncovering more patterns, useful for the improvement of GDL programs.

Other avenues for follow-up research include modelling the data by gender. This would allow relating information about the difference in development of the brains between boys and girls during puberty to the results from this case study, which could reveal more findings with important policy implications.

Also, the pool of information about Canadian jurisdictions (and some U.S. jurisdictions) could be further studied separately using information on fatalities and injuries, rather than just fatalities. This may create problems since the pool of jurisdictions would be low (only about a dozen jurisdictions), but the outcome variable may be more stable because it would not be limited to a relatively infrequent event — i.e., a fatality.

It may also be worthwhile to re-analyze the data excluding second and further data points for the same jurisdiction. The analysis would then only focus on features of jurisdictions when they were first implemented and not on features related to improvements to the original GDL programs. This could help further refine findings with respect to the learning effect among jurisdictions.

Changing the tracking periods (e.g., 24 months rather than 12 months) and investigating how this would influence the results may also be worthwhile because it is possible that the tracking periods used may not have been long enough for 18 and 19 year olds to already have gone through the new GDL system.

It may be useful to select those jurisdictions without an age limit for participation in the GDL program for further analyses. This would include all Canadian jurisdictions but also a few U.S. ones. Selecting such jurisdictions would help ensure that older teens who did not go through the GDL system would not be part of the analyses. As such, it would be unlikely that any effect that may exist among older teens would be cancelled out by the lack of an effect among older teens that were not protected by GDL. It would also make it possible to more reliably calculate an overall effect for all age groups taken together.

Finally, it seems worthwhile to disseminate the research results to encourage jurisdictions to begin thinking about their GDL programs and ways they can enhance them. The fact that some evidence was found regarding the existence of a "learning effect" among jurisdictions is encouraging. In that regard, some of the results were only significant according to one approach and perhaps they should only serve as exploratory findings that need further investigation and confirmation. Other results were strong and highly significant according to both approaches that were adopted. Such results should also be confirmed, but could perhaps already be applied with a higher degree of certainty about their merit. It seems the strongest evidence from this case study is that GDL works with 16 year old drivers — a finding that is congruent with findings from many other studies. A significant reduction in the relative fatality rate of 19.1% was found with a model that fit the data very well. Practitioners and policy makers can be certain that the implementation of a GDL program will help save lives, at least with 16 year old drivers. The findings regarding passenger restrictions for 16 year old drivers in the intermediate stage are also in line with other research and suggest that GDL programs can be enhanced by using such a restriction.

References

Altman, D.G., Bland, J.M. (2003). Interaction revisited: the difference between two estimates. *British Medical Journal,* 326, 219.

Baker, S.P., Chen, L.H., Li, G. (2006). National Evaluation of Graduated Driver Licensing Programs. Washington, DC: NHTSA.

Baker, S.P., Chen, L.H., and Li, G. (2007). Nationwide Review of Graduated Driver Licensing. Washington, DC: AAA Foundation for Traffic Safety.

Beirness, D.J., Simpson, H.M., Mayhew, D.R., Pak, A. (2001). The Road Safety Monitor. Aggressive Driving. Ottawa: Traffic Injury Research Foundation.

Beirness, D.J., Simpson, H.M., Desmond, K. (2005). The Road Safety Monitor 2004. Drowsy Driving. Ottawa, Ontario: Traffic Injury Research Foundation.

Beirness, D.J., Foss, R.D., Wilson, R.J. (2007). Trends in Drinking and Driving in British Columbia: A Decade of Roadside Surveys. Presentation presented at the Joint Meeting of The International Association of Forensic Toxicologists (TIAFT) and International Council on Alcohol, Drugs and Traffic Safety (ICADTS), Seattle, Washington, USA, August 26-30, 2007.

Browne, W.J. (2004). MCMC estimation in MLwiN. Version 2.0. Updated version for 2.0 by William J. Browne and Edmond S.W. Ng. Centre for Multilevel Modelling. Institute of Education. Londong: University of London.

Burns, N.R., Nettelbeck, T., White, M., Willson, J. (1999). Effects of car window tinting on visual performance: a comparison of elderly and young drivers. *Ergonomics*, 42, 428-443.

Carlin, B.P., Louis, T.A. (2000). Bayes and Empirical Bayes Methods for Data Analysis. Washington, DC: Chapman & Hall/CRC.

Clarke, M.J., Stewart, L.A. (2007). Obtaining individual patient data from randomised controlled trials. In: M. Egger, G.D. Smith and D.G. Altman (eds.). Systematic Reviews in Health Care: Meta-Analysis in Context (second edition). pp. 109-121. London: BMJ Publishing Group.

Cochran, W.G. (1963). Sampling Techniques (second edition). New York: John Wiley & Sons.

Commandeur, J.J.F., Koopman, S.J. (2007). Practical Econometrics. An introduction to state space time series analysis. Oxford: Oxford University Press.

Dee, T.S., Grabowski, D.C., Morrisey, M.A., (2005). Graduated driver licensing and teen traffic fatalities. *Journal of Health Economics,* 24, 571-589.

Deeks, J.J., Altman, D.G., Bradburn, M.J. (2007). Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: M. Egger, G.D. Smith and D.G. Altman (eds.). Systematic Reviews in Health Care: Meta-Analysis in Context (second edition). pp. 285-312. London: BMJ Publishing Group.

Delhomme, P., Vaa, T., Meyer, T., Harland, G., Goldenbeld, C., Järmark, S., Christie, N., Rehnova, V. (1999). Evaluated Road Safety Media Campaigns: An Overview of 265 Evaluated Campaigns and Some Meta-Analysis on Accidents. Paris: INRETS.

Dement, W. C., & Vaughan, C. (1999). The Promise Sleep. New York: Delacorte Press.

Egger, M., Smith, G.D. (2007). Principles of and procedures for systematic reviews. In: M. Egger, G.D. Smith and D.G. Altman (eds.). Systematic Reviews in Health Care: Meta-Analysis in Context (second edition). pp. 23-42. London: BMJ Publishing Group.

Elvik, R. (2005a). Can We Trust the Results of Meta-Analyses? A Systematic Approach to Sensitivity Analysis in Meta-Analysis. *Journal of the Transportation Research Board*, 1908, 221-229.

Elvik, R. (2005b). Introductory Guide to Systematic Reviews and Meta-Analysis. *Journal of the Transportation Research Board*, 1908, 230-235.

Foss, R.D., Evenson, K.R. (1999). Effectiveness of Graduated Driver Licensing in Reducing Motor Vehicle Crashes. *American Journal of Preventive Medicine*, 16 (1S), 47-55.

Foss, R., Feaganes, J.R., Rodgman, E.A. (2001). Initial effects of Graduated Driver Licensing on 16-Year-Old driver crashes in North Carolina. *Journal of the American Medical Association*, 286 (13), 1588-1592.

Franken, I.H.A., Rosso, M., van Honk, J. (2003). Selective memory for alcohol cues in alcoholics and its relation to craving. *Cognitive Therapy and Research*, 27 (4), 481–488.

Goldenbeld, C., Hway-Liem, O. (1994). Internationale en nationale kennis over politietoezicht in het verkeer (R-94-15). Leidschendam: Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV).

Goldstein, H. (2003). Multilevel statistical models. London: Arnold.

Goldstein H, Rasbash J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society A*, 159, 505-513.

Grosvenor, D., Toomey, T.L., Wagenaar, A.C. (1999). Deterrence and the Adolescent Drinking Driver. *Journal of Safety Research*, 30(3), 187-191.

Harbord, R.M., Higgins, J.P.T. (2008). Meta-regression in Stata. *The Stata Journal*, 8(4), 493-519.

Hartling, L., Wiebe, N., Russell, K., Petruk, J., Spinola, C., Klassen, T.P. (2005). Graduated driver licensing for reducing motor vehicle crashes among young drivers. *The Cochrane Database of Systematic Reviews*, Issue 2. Art. No.: CD003300.pub2.

Hauer, E. (1983a). Reflections on methods of statistical inference in research on the effect of safety countermeasures. *Accident Analysis and Prevention*, 15 (4), 275-286.

Hauer, E. (1983b). An application of the likelihood/Bayes approach to the estimation of safety countermeasures effectiveness. *Accident Analysis and Prevention*, 15 (4), 187-298.

Hauer, E. (2004). The harm done by tests of significance. *Accident Analysis and Prevention*, 36 (3), 495-500.

Hauer, E. (2008). Observational before-after studies in road safety. Estimating the effect of highway and traffic engineering measures on road safety. United Kingdom: Emerald.

Heck, R.H., Thomas, S.L. (2000). An introduction to multilevel modelling techniques. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Hedlund, J. (1984). Comments on Hauer’s approach to statistical inference. *Accident Analysis and Prevention*, 16 (3), 163-166.

Hedlund, J., Shults, R.A., Compton, R. (2003). What we know, what we don't know, and what we need to know about graduated driver licensing. *Journal of Safety Research,* 34, 107-115.

Hedlund, J., Shults, R.A., Compton, R. (2006). Graduated driver licensing and teenage driver research in 2006. *Journal of Safety Research,* 37, 107-121.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., Altman, D.G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal,* 327, 557-560.

Homel, R. (1988). Policing and Punishing the Drinking Driver. A Study of General and Specific Deterrence. New York: Springer Verlag.

Horne, J.A., Reyner, L.A., Barrett, P.R. (2003). Driving impairment due to sleepiness is exarcerbated by low alcohol intake. *Occupational and Environmental Medicine*, 60, 689-692.

Hox, J. (2002). Multilevel Analysis. Techniques and Applications. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Hox, J.J., de Leeuw, E.D. (2003). Multilevel models for Meta-Analysis. In: S.P. Reise & N. Duan (Eds.). Multilevel modelling. Methodological Advances, Issues, and Applications. Mahwah, N.J.: Lawrence Erlbaum Associates.

Insurance Institute for Highway Safety (IIHS) (April 2006). Bad statistics lead to misinformation. Status Report. Vol. 41 (4).

Johns, M.W. (1991). A new method for measuring daytime sleepiness: the epworth sleepiness scale. *Sleep*, 14 (6), 540-545.

Jones, A.P., Jørgensen, S.H. (2003). The use of multilevel models for the prediction of road accident outcomes. *Accident Analysis and Prevention*, 35 (1), 59-70.

Jones, K. (1993). Using multilevel models for survey analysis. *Journal of the Market Research Society*, 35 (3), 249-265.

Kish, L. (1965). Survey Sampling. New York: John Wiley & Sons, Inc.

Kreft, I.G.G. (1994). Multilevel models for hierarchically nested data: Potential applications in substance abuse prevention research. In: Advances in Data Analysis for Prevention Intervention Research, LM Collins, LA Seitz (eds.) Research Monograph 142. pp. 140-183. Washington DC: National Institute on Drug Abuse.

Kreft I, de Leeuw J. (2002). Introducing multilevel modelling (second edition). London: Sage publications.

Lambert, P.C., Abrams, K.R. (1995). Meta-Analysis Using Multilevel Models. *Multilevel Modelling Newsletter*, 7(2), 16-18.

Levy, P.S., Lemeshow, S. (1999). Sampling of Populations. Methods and Applications (third edition). New York: John Wiley & Sons.

Leyland, A.H., Goldstein, H. (2001). Multilevel Modelling of Health Statistics. West Sussex, England: John Wiley & Sons, Ltd.

Lindsey, J.K. (1993). Models for repeated measurements. Oxford: Clarendon Press.

Long, J.S., Freese, J. (2006). Regression Models for Categorical Dependent Variables Using Stata (Second edition). A Stata Press Publication. College Station, Texas: StataCorp LP.

Longford, N. (1993). Random coefficient models. Oxford: Clarendon.

Lumley, M., Roehrs, T., Asker, D., Zorick, F., Roth, T. (1987). Ethanol and caffeine effects on daytime sleepiness/alertness. *Sleep,* 10: 306-312.

Mayhew, D.R. (2007). Driver education and graduated licensing in North America: Past, present and future. *Journal of Safety Research,* 38, 229-235.

Mayhew, D.R., Simpson, H.M., Groseillers M. (1999). Impact of the graduated driver licensing program in Nova Scotia. Ottawa: Traffic Injury Research Foundation.

Mayhew, D.R., Simpson, H.M., Groseillers M., Williams, A.F. (2001). Impact of the graduated driver licensing program in Nova Scotia. *Journal of Crash Prevention and Injury Control*, 2(3), 179-192.

Mayhew, D.R., Simpson, H.M., Williams, A.F., and Desmond, K. (2002). Specific and Long-Term Effects of Nova Scotia's Graduated Licensing Program. Arlington, VA: Insurance Institute for Highway Safety.

Mayhew, D.R., Simpson, H.M, Singhal, D. (2005). Best Practices for Graduated Driver Licensing in Canada. Ottawa: Traffic Injury Research Foundation.

Mayhew, D.R., Simpson, H.M., Singhal, D., Desmond, K. (2006). Reducing the Crash Risk for Young Drivers. Washington, DC: AAA Foundation for Traffic Safety.

McMillan, N.J., Berliner, M.J. (1994). A spatially correlated hierarchical random effect model for Ohio corn yield. Technical report 10. Research Triangle Park, NC: National Institute for Statistical Sciences.

Ministry of Transportation Ontario (1998). Ontario Road Safety Annual Report 1997. Toronto: Ministry of Transportation Ontario.

Mlodinow, L. (2008). The Drunkard's Walk. How Randomness Rules Our Lives. New York: Pantheon Books.

Morrisey, M.A., Grabowski, D.C., Dee, T.S., Campbell, C. (2006). The strength of graduated drivers license programs and fatalities among teen drivers and passengers. *Accident Analysis and Prevention,* 38 (1), 135-141.

Pack, A. I., Pack, A. M., Rodgman, E., Cucchiara, A., Dinges, D. F., Schwab, C. W. (1995). Characteristics of accidents attributed to the driver having fallen asleep. *Accident Analysis and Prevention,* 27, 769−775.

Raftery, A.E., Lewis, S.M. (1992). How many iterations in the Gibbs Sampler? In: J.M. Bernardo, J.O. Berger, A.P. David and A.F.M. Smith (eds.). Bayesian Statistics 4. pp. 763-773. Oxford: Oxford University Press.

Rasbash J, Browne W, Goldstein H, Yang M, Plewis I, Healy M, Woodhouse G, Draper D, Langford I, Lewis T. (2000). A user's guide to MLwiN. Version 2.1c. London: Centre for Multilevel Modelling, Institute of Education, University of London.

Rasbash J, Steele F, Browne W, Prosser B. (2004). A User's Guide to MLwiN. Version 2.1e. London: Centre for Multilevel Modelling, Institute of Education, University of London.

Rasbash, J., Steele, F., Browne, W., Prosser, B. (2005). A User's Guide to MLwiN. Version 2.0. UK: University of Bristol, Centre for Multilevel Modelling.

Raudenbush, S.W., Bryk, A.S. (2002). Hierarchical Linear Models. Applications and Data Analysis Methods (second edition). Thousand Oaks, California: Sage Publications.

Rice, N. (2001). Binomial Regression, in Multilevel Modelling of Health Statistics, Ed. AH Leyland, H Goldstein, John Wiley & Sons, Ltd., West Sussex, England, pp. 27-43.

Robertson, R., Holmes, E., Vanlaar, W. (2009). The facts about fatigued driving in Ontario. A guidebook for police. Ottawa: Traffic Injury Research Foundation.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.

Rodriguez, G., Goldman, N. (1995). An assessment of estimation procedure for multilevel models with binary responses. *Journal of the Royal Statistical Society A*, 158, 73-89.

Ross, H.L. (1992). Confronting Drunk Driving: Social policy for saving lives. New Haven, Connecticut: Yale University.

Shope, J.T. (2007). Graduated driver licensing: Review of Evaluation results since 2002. *Journal of Safety Research,* 38(2), 165-175.

Shope, J.T., Molnar, L.J., Elliott, M.R., Waller, P.F. (2001). Graduated Driver Licensing in Michigan: Early impact on motor vehicle crashes among 16-year-old drivers. *Journal of the American Medical Association*, 286 (13), 1599-1598.

Simpson, H.M., (2003). The evolution and effectiveness of graduated licensing. *Journal of Safety Research,* 34, 25-34.

Smith, T.C., Spiegelhalter, D.J., Thomas, A. (1995). Bayesian Approaches to Random-Effects Meta-Analysis: A Comparative Study. *Statistics in Medicine,* 14, 2685-2699.

Snijders, T., Bosker, R. (1999). Multilevel analysis. An introduction to basic and advanced multilevel modelling. London: Sage Publications.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64, 583-640.

StataCorp. (2007). Stata Statistical Software. Release 10. College Station, Texas: StataCorp LP.

Statistics Canada 2007. Demographics Estimates Compendium 2007 (CD-ROM). 91-213-SCB.

Sterne, J.A.C., Bradburn, M.J., Egger, M. (2007a). Meta-analysis in Stata. In: M. Egger, G.D. Smith and D.G. Altman (eds.). Systematic Reviews in Health Care: Meta-Analysis in Context (second edition). pp. 347-372. London: BMJ Publishing Group.

Sterne, J.A.C, Egger, M., Smith, G.D. (2007b). Investigating and dealing with publication and other biases. In: M. Egger, G.D. Smith and D.G. Altman (eds.). Systematic Reviews in Health Care: Meta-Analysis in Context (second edition). pp. 347-372. London: BMJ Publishing Group.

Swamy, PAVB. (1971). Statistical Inference in Random Coefficient Regression Models. New York: Springer.

Tacq J. (1989). Van multiniveau-probleem naar multiniveau-analyse(RISBO-papers S 04). Rotterdam: Rotterdams Instituut voor Sociaal-wetenschappelijk Beleidsonderzoek, Erasmus universiteit.

Tacq, J. (1997). Multivariate Analysis Techniques in Social Science Research. From Problem to Analysis. London: Sage Publications Ltd.

Thomas, P., Morris A., Yannis, G., Lejeune, P., Wesemann, P., Vallet, G., Vanlaar, W. (2005). Designing the European Road Safety Observatory. *International Journal of Injury Control and Safety Promotion*, 12 (4), 251-253.

Thompson, S.G. (2007). Why and how sources of heterogeneity should be investigated. In: M. Egger, G.D. Smith and D.G. Altman (eds.). Systematic Reviews in Health Care: Meta-Analysis in Context (second edition). pp. 157-175. London: BMJ Publishing Group.

Turner, R.M., Omar, R.Z., Yang, M., Goldstein, H., Thompson, S.G. (1999). Random effects meta-analysis of trials with binary outcomes using multilevel models in MLwiN. *Multilevel Modelling Newsletter,* 11(1), 6-9.

Ulmer, R.G., Preusser, D.F., Williams, A.F., Ferguson, S.A., Farmer, C.M. (2000). Effect of Florida's graduated licensing program on the crash rate of teenage drivers. *Accident Analysis and Prevention,* 32(4), 527-532.

van Driel, C.J.G., Davidse, R.J., van Maarseveen, M.F.A.M. (2004). The effects of an edgeline on speed and lateral position: a meta-analysis. *Accident Analysis and Prevention*, 36, 671-682.

Vanlaar, W. (2005a). Drink driving in Belgium: results from the third and improved roadside survey. *Accident Analysis and Prevention*, 37, 391-397.

Vanlaar, W. (2005b). Multilevel Modelling in Traffic Safety Research: Two Empirical Examples Illustrating the Consequences of Ignoring Hierarchies. *Traffic Injury Prevention*, 6 (4), 311-316.

Vanlaar, W. (2007a). An intuitive introduction to multilevel modelling. In: E. Dupont and H. Martensen (eds.). Multilevel modelling and time series analysis in traffic safety research – methodology. Deliverable D7.4 of the EU FP6 project SafetyNet. pp. 33-36.

Vanlaar, W. (2007b). Basic two level random intercept and random slope models. In: E. Dupont and H. Martensen (Eds.). Multilevel modelling and time series analysis in traffic safety research – methodology. Deliverable D7.4 of the EU FP6 project SafetyNet. pp. 37-51.

Vanlaar, W. (2007c). Binary and general binomial responses. In: E. Dupont and H. Martensen (Eds.). Multilevel modelling and time series analysis in traffic safety research – methodology. Deliverable D7.4 of the EU FP6 project SafetyNet. pp. 71-78.

Vanlaar, W. (2008). Less is more: The influence of traffic count on drinking and driving behavior. *Accident Analysis and Prevention,* 40, 1018-1022.

Vanlaar, W., Simpson, H., Mayhew, D., Robertson, R. (2008a). Fatigued and drowsy driving: a survey of attitudes, opinions and behavior. *Journal of Safety Research*, 39, 303-309.

Vanlaar, W., Simpson, H., Mayhew, D., Robertson, R. (2008b). Aggressive driving: A survey of attitudes, opinions and behaviors. *Journal of Safety Research*, 39, 375-381.

Vanlaar, W., Simpson, H., Robertson, R. (2008). A perceptual map for understanding concern about unsafe driving behaviours. *Accident Analysis and Prevention*, 40, pp. 1667-1673.

Vanlaar, W., Wets, G., Brijs, T. (submitted). Sleepiness among night-time drivers: A multilevel model. *Accident Analysis and Prevention*.

Vanlaar, W., Mayhew, D., Marcoux, K., Wets, G., Brijs, T., Shope, J. (2009a). An Evaluation of Graduated Driver Licensing Programs in North-America. An Analysis of Relative Fatality Risks of 16, 17, 18 and 19 Year Old Drivers Using a Meta-Analytic Approach. Ottawa: Traffic Injury Research Foundation.

Vanlaar, W., Mayhew, D., Marcoux, K., Wets, G., Brijs, T., Shope, J. (2009b). An Evaluation of Graduated Driver Licensing Programs in North-America Using a Meta-Analytic Approach. *Accident Analysis and Prevention*, 41, 1104-1111.

Vanlaar, W., Mayhew, D., Marcoux, K., Wets, G., Brijs, T., Shope, J. (2009c). An Evaluation of Graduated Driver Licensing Programs in North-America Using a Meta-Analytic Approach. Proceedings of the 19th Canadian Multidisciplinary Road Safety Conference, Saskatoon, Saskatchewan, June 7-10, 2009.

Wåhlberg, A.E.AF. (2001). The theoretical features of some current approaches to risk perception. *Journal of Risk Research,* 4(3), 237-250.

Wang, J., Knipling, R. R., & Goodman, M. J. (1996). The role of driver inattention in crashes; new statistics from the 1995 Crashworthiness Data System. 40th Annual Proceedings, Association for the advancement of Automotive Medicine, Vancouver.

Williams, A.F. (2003). The compelling case for graduated licensing. *Journal of Safety Research*, 34, 3-4.

Williams, A.F. (2007). Contribution of the components of graduated licensing to crash reductions. *Journal of Safety Research,* 38, 177-184.

Williams, A.F., Mayhew, D.R. (2008). Graduated Licensing and Beyond. *American Journal of Preventive Medicine*, 35(3S), S324-S333.

Wilson, R.J., Fang, M., Cooper, P.J. (2006). Sleepiness Among Night-Time Drivers: Relationship to Blood Alcohol Concentration and Other Factors. *Traffic Injury Prevention*, 7, 15-22.

Appendices

Appendix 1: Summary effect for 16 year old drivers

Table 6.1.1: Random Effects Meta-Analysis

```
   jurisdiction     |     ES    [95% Conf. Interval]     % Weight
--------------------+-------------------------------------------------
AK                  | 1.914       0.061    60.256          0.13
AL                  | 1.272       0.627     2.581          2.79
AR1                 | 0.852       0.311     2.338          1.45
AR2                 | 0.815       0.290     2.293          1.39
CA                  | 1.351       0.738     2.473          3.67
CO                  | 0.224       0.048     1.052          0.64
CT1                 | 0.603       0.098     3.699          0.47
CT2                 | 1.185       0.073    19.342          0.20
DC                  | 0.275       0.005    14.926          0.10
DE                  | 0.914       0.142     5.867          0.45
FL                  | 3.389       1.459     7.873          2.03
GA1                 | 0.629       0.344     1.148          3.70
GA2                 | 1.283       0.722     2.278          4.01
IA                  | 0.574       0.190     1.739          1.22
ID                  | 0.999       0.061    16.296          0.20
IL                  | 1.000       0.521     1.922          3.22
IN                  | 0.668       0.344     1.298          3.12
KY1                 | 0.357       0.154     0.832          2.02
LA                  | 0.889       0.362     2.183          1.80
MA1                 | 1.341       0.412     4.359          1.08
MA2                 | 2.103       0.511     8.651          0.76
MD1                 | 1.086       0.216     5.464          0.59
MD2                 | 0.544       0.160     1.849          1.01
ME1                 | 1.186       0.159     8.852          0.38
ME2                 | 6.405       0.329   124.496          0.18
ME3                 | 0.273       0.031     2.435          0.32
MI                  | 0.479       0.233     0.982          2.71
MN1                 | 1.901       0.580     6.230          1.07
MN2                 | 1.142       0.461     2.831          1.77
MO                  | 0.682       0.349     1.331          3.08
MS                  | 1.239       0.556     2.765          2.22
NC                  | 0.258       0.110     0.604          2.00
ND                  | 1.757       0.056    54.994          0.13
NE                  | 1.458       0.405     5.253          0.92
NH1                 | 1.049       0.162     6.780          0.44
NH2                 | 2.504       0.244    25.693          0.29
NH3                 | 1.075       0.224     5.145          0.62
NJ                  | 0.460       0.015    13.841          0.13
NM                  | 0.243       0.011     5.452          0.16
NV                  | 1.013       0.220     4.660          0.66
NY1                 | 1.169       0.435     3.147          1.51
NY2                 | 0.541       0.183     1.603          1.26
OH1                 | 0.486       0.226     1.046          2.42
OH2                 | 0.626       0.276     1.422          2.14
OK                  | 1.899       0.733     4.924          1.62
OR                  | 0.259       0.030     2.252          0.33
PA                  | 0.515       0.237     1.115          2.38
RI1                 | 1.257       0.073    21.539          0.19
RI2                 | 2.002       0.064    62.588          0.13
SC1                 | 0.722       0.349     1.493          2.66
SC2                 | 0.472       0.159     1.397          1.26
SD                  | 0.545       0.018    16.781          0.13
TN                  | 0.546       0.284     1.051          3.21
TX                  | 0.640       0.393     1.045          5.22
UT1                 | 2.297       0.430    12.262          0.55
UT2                 | 2.061       0.333    12.750          0.46
UT3                 | 0.509       0.125     2.063          0.78
VA1                 | 0.703       0.320     1.543          2.31
VA2                 | 1.594       0.631     4.027          1.70
VA3                 | 0.827       0.345     1.985          1.90
VA4                 | 0.836       0.379     1.847          2.28
VT                  | 0.266       0.028     2.543          0.30
WA                  | 3.318       0.679    16.221          0.61
WI                  | 0.684       0.295     1.589          2.03
WV                  | 0.630       0.164     2.416          0.84
WY                  | 0.333       0.010    10.615          0.13
Alberta             | 0.203       0.057     0.716          0.95
BC1                 | 1.504       0.594     3.811          1.69
BC2                 | 0.354       0.071     1.772          0.59
Manitoba            | 0.200       0.010     3.896          0.18
NewBrunswick        | 0.475       0.016    14.553          0.13
Newfoundland        | 1.174       0.036    37.906          0.13
NovaScotia          | 0.458       0.046     4.555          0.29
Ontario             | 0.598       0.299     1.197          2.89
```

```
PrinceEdwardI        |  0.430       0.033     5.691          0.23
Quebec               |  1.329       0.443     3.984          1.24
Saskatchewan         |  1.811       0.057    57.181          0.13
Yukon                |  2.900       0.042   199.675          0.09
---------------------+--------------------------------------------------
D+L pooled ES        |  0.809       0.714     0.917        100.00
---------------------+--------------------------------------------------

  Heterogeneity chi-squared =  81.17 (d.f. = 77) p = 0.351
  I-squared (variation in ES attributable to heterogeneity) =   5.1%
  Estimate of between-study variance Tau-squared =  0.0158

  Test of ES=1 : z=   3.32 p = 0.001
```

Figure 6.1.1: Forest Plot

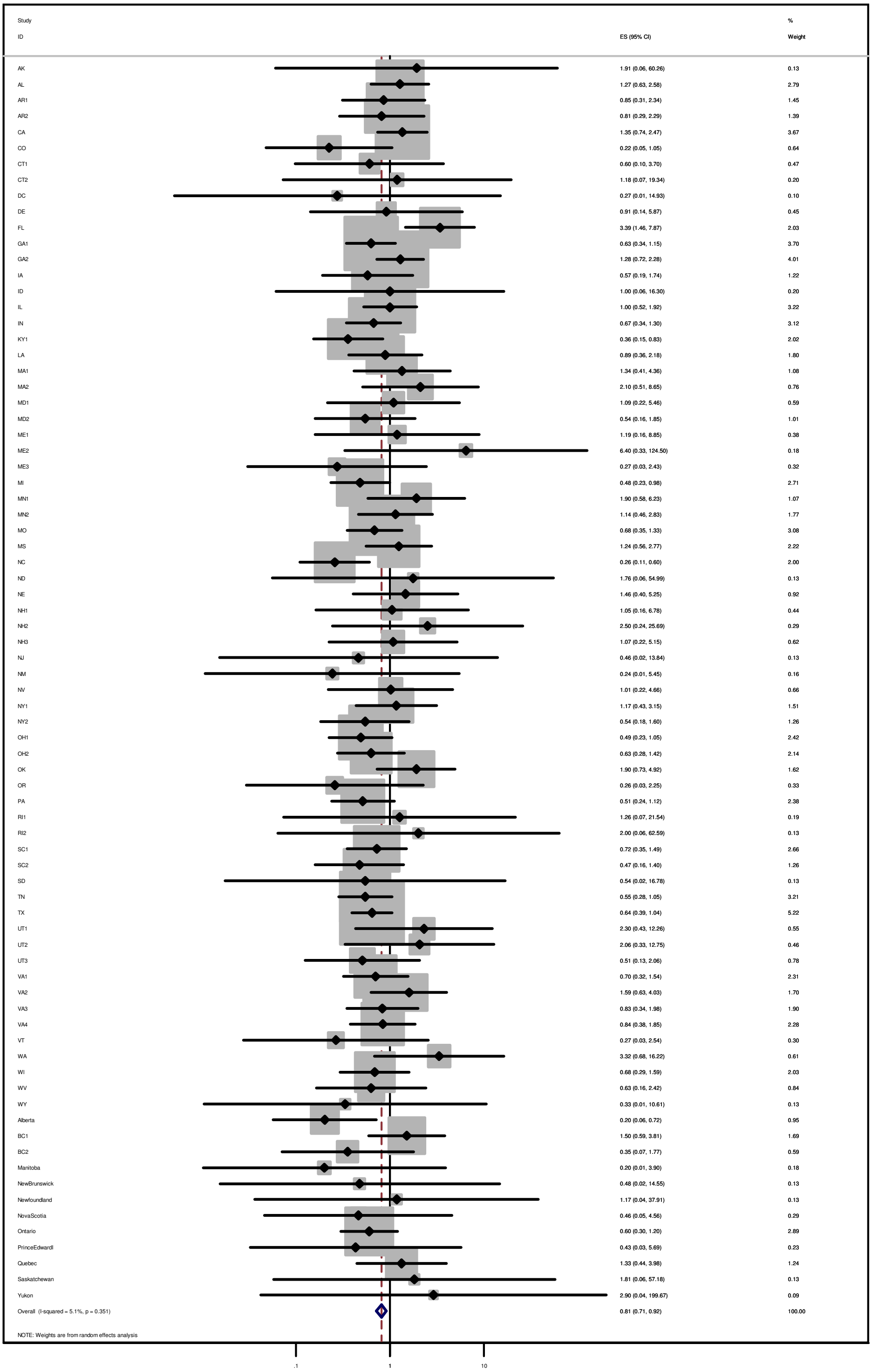
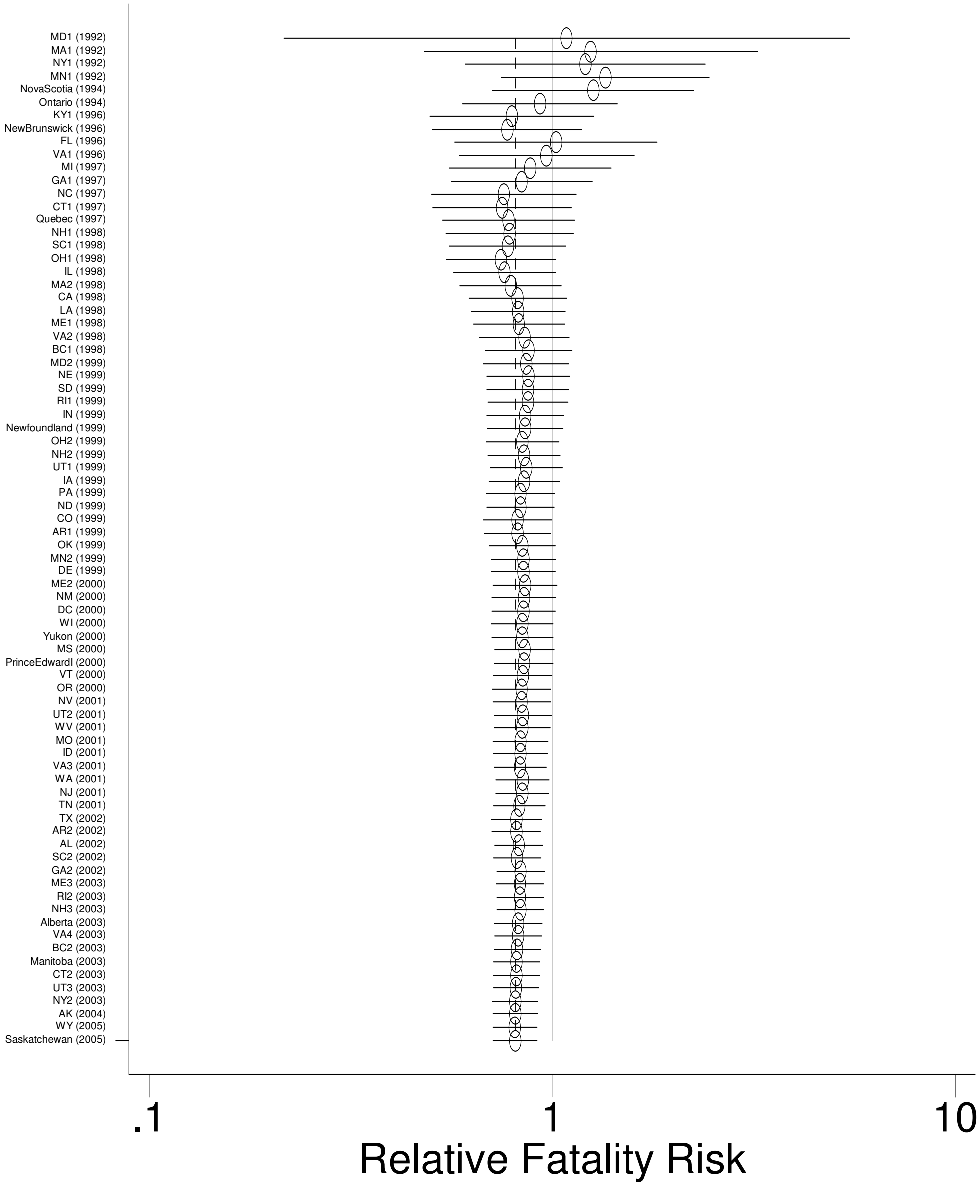


Figure 6.1.2: Cumulative Random Effects Meta-Analysis



Appendix 2: Summary effect for 17 year old drivers

Table 6.2.1: Random Effects Meta-Analysis

```
   Jurisdiction     |     ES    [95% Conf. Interval]     % Weight
---------------------+---------------------------------------------------
AK                   |  0.245       0.010     5.764          0.10
AL                   |  1.223       0.569     2.626          1.69
AR1                  |  0.852       0.376     1.927          1.48
AR2                  |  1.172       0.519     2.648          1.49
CA                   |  1.132       0.677     1.892          3.75
CO                   |  0.874       0.310     2.464          0.92
CT1                  |  0.708       0.154     3.257          0.42
CT2                  |  1.595       0.343     7.408          0.42
DC                   |  0.383       0.007    20.828          0.06
DE                   |  0.712       0.121     4.185          0.32
FL                   |  1.308       0.751     2.280          3.20
GA1                  |  0.795       0.454     1.393          3.14
GA2                  |  1.465       0.806     2.663          2.77
IA                   |  0.787       0.288     2.148          0.98
ID                   |  0.984       0.236     4.099          0.49
IL                   |  1.089       0.584     2.034          2.54
IN                   |  1.031       0.519     2.048          2.10
KY1                  |  1.148       0.589     2.238          2.22
LA                   |  0.979       0.391     2.452          1.17
MA1                  |  0.498       0.146     1.706          0.65
MA2                  |  1.464       0.564     3.800          1.09
MD1                  |  0.250       0.053     1.177          0.41
MD2                  |  1.652       0.693     3.942          1.31
ME1                  |  1.828       0.289    11.552          0.29
ME2                  |  0.589       0.175     1.988          0.67
ME3                  |  0.694       0.163     2.957          0.47
MI                   |  1.649       0.876     3.104          2.47
MN1                  |  0.365       0.152     0.877          1.28
MN2                  |  1.021       0.444     2.347          1.43
MO                   |  0.672       0.369     1.224          2.76
MS                   |  1.217       0.597     2.484          1.94
NC                   |  0.891       0.488     1.627          2.73
ND                   |  0.423       0.112     1.605          0.56
NE                   |  0.618       0.154     2.482          0.51
NH1                  |  0.181       0.008     4.194          0.10
NH2                  |  0.253       0.025     2.591          0.18
NH3                  |  1.072       0.224     5.135          0.40
NJ                   |  2.843       0.753    10.742          0.56
NM                   |  2.520       0.478    13.279          0.36
NV                   |  0.972       0.252     3.741          0.54
NY1                  |  1.017       0.525     1.971          2.26
NY2                  |  0.855       0.473     1.547          2.81
OH1                  |  1.183       0.569     2.459          1.85
OH2                  |  0.841       0.458     1.546          2.67
OK                   |  1.059       0.435     2.575          1.25
OR                   |  1.037       0.271     3.969          0.55
PA                   |  0.718       0.390     1.321          2.66
RI1                  |  0.582       0.018    18.344          0.08
RI2                  |  0.687       0.105     4.495          0.28
SC1                  |  0.867       0.416     1.807          1.83
SC2                  |  1.471       0.702     3.086          1.80
SD                   |  0.222       0.025     2.001          0.20
TN                   |  0.933       0.484     1.797          2.30
TX                   |  1.025       0.682     1.541          5.94
UT1                  |  0.301       0.059     1.544          0.37
UT2                  |  0.981       0.296     3.252          0.69
UT3                  |  0.202       0.024     1.733          0.21
VA1                  |  0.677       0.308     1.487          1.60
VA2                  |  1.457       0.738     2.875          2.14
VA3                  |  0.733       0.364     1.477          2.02
VA4                  |  1.345       0.605     2.990          1.55
VT                   |  2.231       0.072    69.485          0.08
WA                   |  1.298       0.508     3.312          1.13
WI                   |  0.765       0.355     1.648          1.68
WV                   |  0.691       0.204     2.338          0.66
WY                   |  0.707       0.013    37.759          0.06
Alberta              |  0.928       0.361     2.388          1.11
BC1                  |  1.184       0.544     2.577          1.64
BC2                  |  1.314       0.534     3.236          1.22
Manitoba             |  1.250       0.316     4.950          0.52
NewBrunswick         |  1.897       0.166    21.714          0.17
Newfoundland         |  0.098       0.004     2.138          0.10
NovaScotia           |  1.940       0.413     9.122          0.41
Ontario              |  1.113       0.658     1.881          3.58
PrinceEdwardI        |  0.909       0.016    51.292          0.06
Quebec               |  1.592       0.835     3.036          2.37
Saskatchewan         |  0.235       0.010     5.544          0.10
```

```
Yukon                |  5.574       0.131   237.211          0.07
---------------------+---------------------------------------------------
D+L pooled ES        |  1.001       0.906     1.105        100.00
---------------------+---------------------------------------------------

  Heterogeneity chi-squared =  53.16 (d.f. = 77) p = 0.983
  I-squared (variation in ES attributable to heterogeneity) =   0.0%
  Estimate of between-study variance Tau-squared =  0.0000

  Test of ES=1 : z=   0.02 p = 0.988
```

Figure 6.2.1: Forest Plot

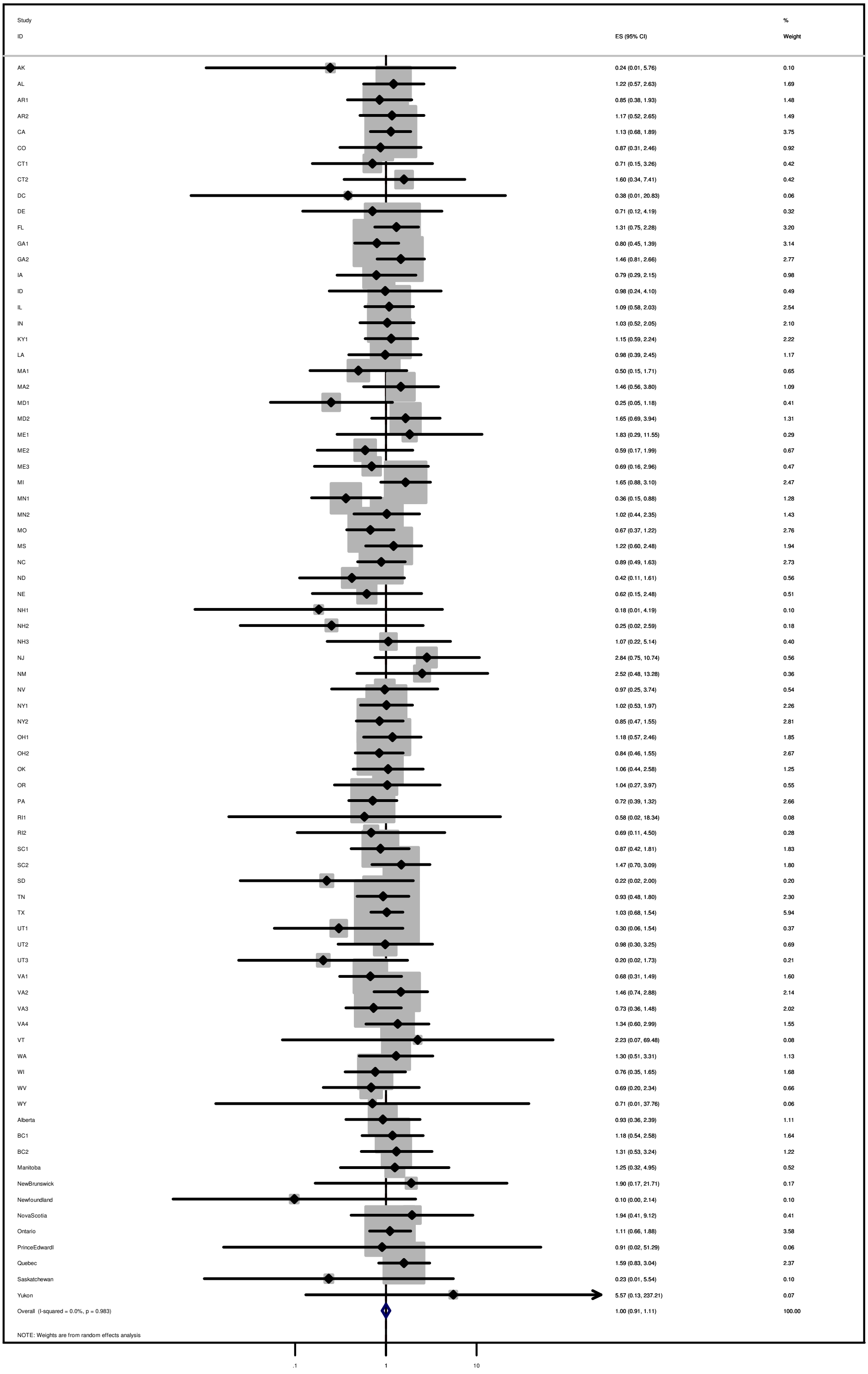
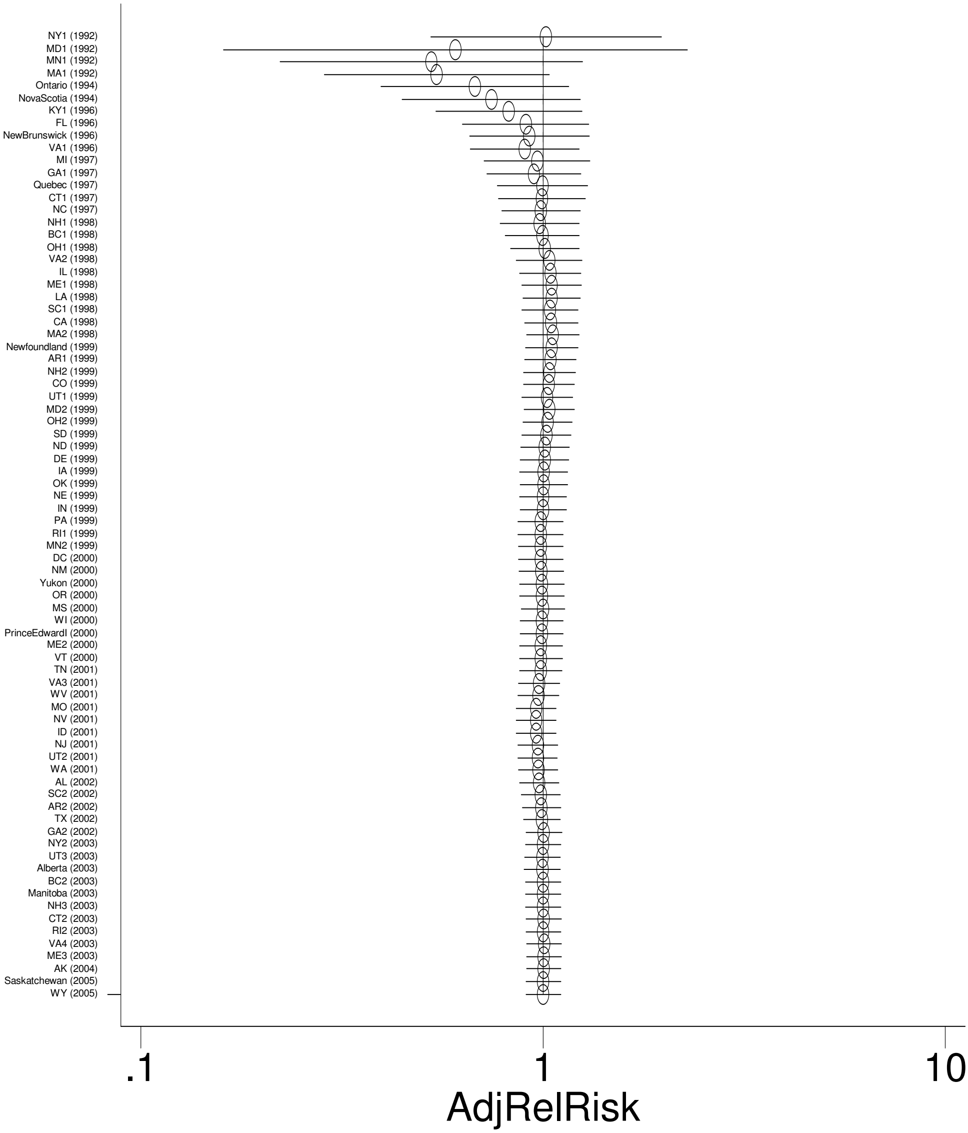


Figure 6.2.2: Cumulative Random Effects Meta-Analysis



Appendix 3: Summary effect for 18 year old drivers

Table 6.3.1: Random Effects Meta-Analysis

```
   Jurisdiction     |     ES    [95% Conf. Interval]     % Weight
--------------------+---------------------------------------------------
AK                  |  0.990       0.019     52.330          0.07
AL                  |  1.023       0.522      2.004          1.94
AR1                 |  0.822       0.388      1.743          1.60
AR2                 |  1.118       0.523      2.388          1.57
CA                  |  1.547       1.005      2.381          3.85
CO                  |  0.530       0.192      1.465          0.93
CT1                 |  3.182       0.643     15.758          0.39
CT2                 |  1.465       0.376      5.699          0.54
DC                  |  0.256       0.005     13.891          0.06
DE                  |  7.945       0.381    165.505          0.11
FL                  |  1.308       0.807      2.121          3.27
GA1                 |  0.981       0.582      1.653          2.92
GA2                 |  0.933       0.538      1.617          2.69
IA                  |  3.408       1.204      9.651          0.89
ID                  |  0.471       0.155      1.428          0.79
IL                  |  0.907       0.484      1.702          2.17
IN                  |  1.177       0.587      2.362          1.83
KY1                 |  1.444       0.797      2.617          2.38
LA                  |  1.385       0.642      2.988          1.54
MA1                 |  0.386       0.164      0.911          1.27
MA2                 |  0.534       0.155      1.831          0.65
MD1                 |  0.583       0.195      1.745          0.81
MD2                 |  3.778       1.213     11.761          0.76
ME1                 |  7.408       0.359    152.841          0.11
ME2                 |  0.987       0.282      3.453          0.63
ME3                 |  0.769       0.210      2.818          0.59
MI                  |  1.020       0.536      1.941          2.09
MN1                 |  1.557       0.517      4.693          0.80
MN2                 |  0.280       0.098      0.795          0.88
MO                  |  1.175       0.644      2.143          2.34
MS                  |  1.648       0.720      3.775          1.35
NC                  |  1.556       0.792      3.054          1.93
ND                  |  0.812       0.105      6.271          0.24
NE                  |  2.358       0.815      6.825          0.86
NH1                 |  0.713       0.042     11.983          0.13
NH2                 |  1.420       0.121     16.658          0.17
NH3                 |  0.966       0.240      3.890          0.51
NJ                  |  1.039       0.366      2.948          0.89
NM                  |  0.572       0.133      2.453          0.47
NV                  |  2.541       0.499     12.942          0.38
NY1                 |  0.394       0.208      0.748          2.10
NY2                 |  0.877       0.488      1.576          2.43
OH1                 |  1.598       0.940      2.718          2.84
OH2                 |  1.282       0.769      2.138          3.01
OK                  |  1.499       0.685      3.280          1.49
OR                  |  1.018       0.366      2.834          0.92
PA                  |  0.900       0.524      1.545          2.77
RI1                 |  0.969       0.057     16.600          0.13
RI2                 |  0.517       0.017     16.150          0.09
SC1                 |  1.097       0.560      2.145          1.95
SC2                 |  1.477       0.704      3.098          1.64
SD                  |  0.814       0.169      3.910          0.41
TN                  |  1.650       0.953      2.857          2.70
TX                  |  1.391       0.969      1.998          4.84
UT1                 |  1.709       0.303      9.652          0.34
UT2                 |  2.655       0.926      7.611          0.87
UT3                 |  2.075       0.577      7.465          0.60
VA1                 |  0.892       0.416      1.911          1.56
VA2                 |  0.987       0.529      1.842          2.20
VA3                 |  1.017       0.541      1.912          2.15
VA4                 |  1.026       0.506      2.080          1.78
VT                  |  0.168       0.019      1.500          0.21
WA                  |  0.940       0.428      2.064          1.48
WI                  |  0.555       0.271      1.140          1.73
WV                  |  0.466       0.166      1.311          0.90
WY                  |  4.248       0.197     91.426          0.11
Alberta             |  1.148       0.509      2.591          1.39
BC1                 |  1.478       0.683      3.196          1.53
BC2                 |  0.832       0.412      1.680          1.80
Manitoba            |  0.307       0.034      2.727          0.21
NewBrunswick        |  1.004       0.019     51.797          0.07
Newfoundland        |  0.201       0.019      2.177          0.18
NovaScotia          |  0.475       0.048      4.725          0.19
Ontario             |  1.085       0.639      1.842          2.85
PrinceEdwardI       |  1.745       0.132     23.075          0.15
Quebec              |  0.755       0.443      1.285          2.83
Saskatchewan        |  4.032       0.171     95.208          0.10
```

```
Yukon                |  2.643       0.038   181.963          0.06
---------------------+---------------------------------------------------
D+L pooled ES        |  1.083       0.978     1.199        100.00
---------------------+---------------------------------------------------

  Heterogeneity chi-squared =  86.63 (d.f. = 77) p = 0.212
  I-squared (variation in ES attributable to heterogeneity) =  11.1%
  Estimate of between-study variance Tau-squared =  0.0218

  Test of ES=1 : z=   1.53 p = 0.126
```

Figure 6.3.1: Forest Plot

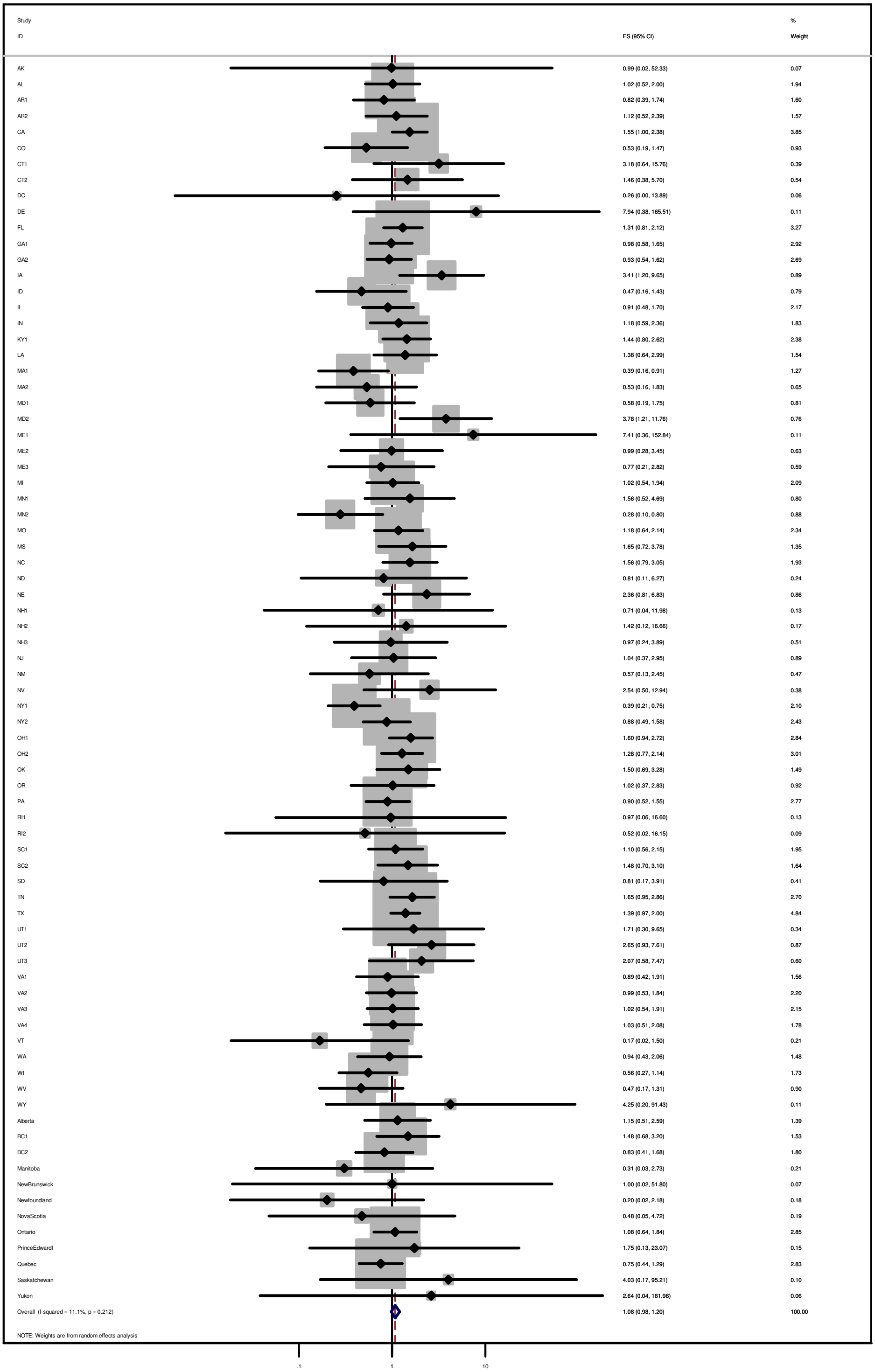
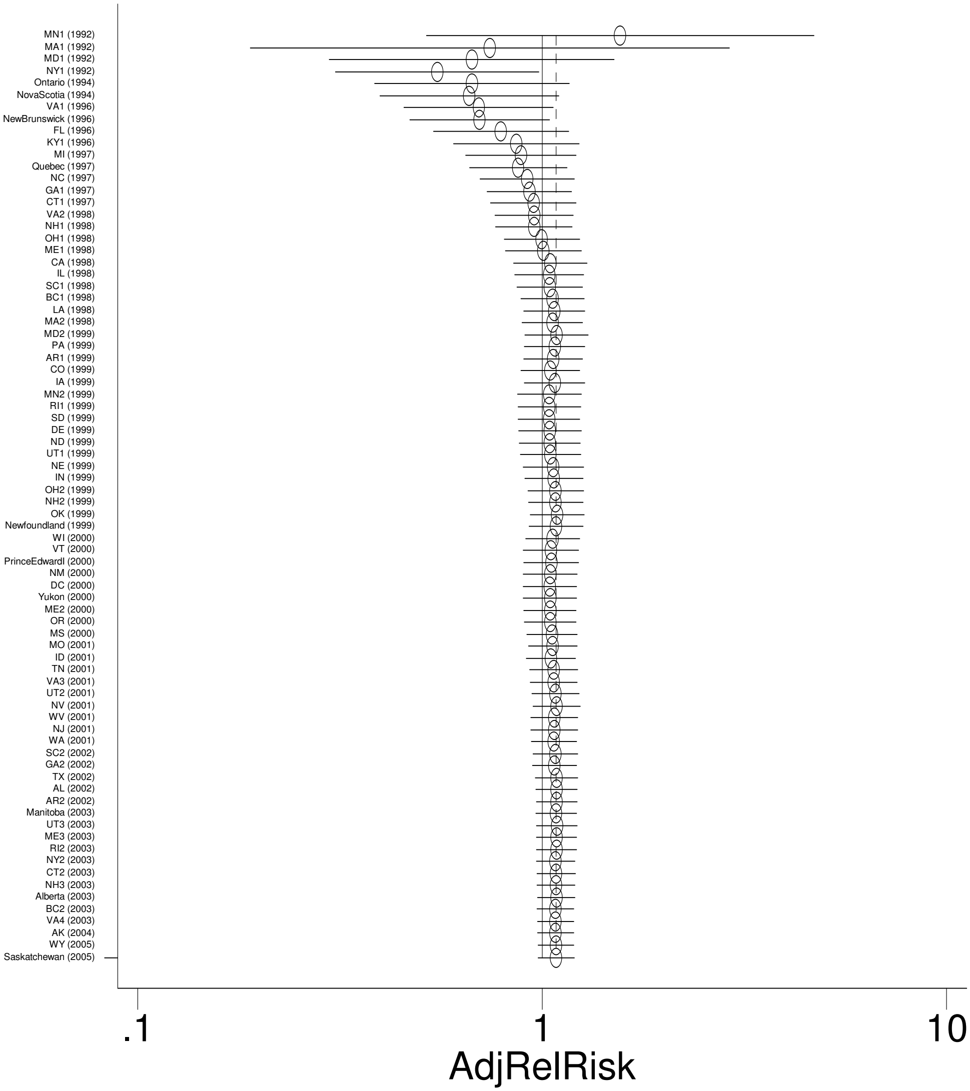


Figure 6.3.2: Cumulative Random Effects Meta-Analysis



Appendix 4: Summary effect for 19 year old drivers

Table 6.4.1: Random Effects Meta-Analysis

```
    Study     |     ES    [95% Conf. Interval]     % Weight
---------------------+--------------------------------------------------
AK                   |  1.834       0.058    57.727          0.08
AL                   |  0.944       0.547     1.630          2.71
AR1                  |  1.383       0.569     3.365          1.10
AR2                  |  1.363       0.654     2.839          1.58
CA                   |  1.062       0.698     1.616          4.25
CO                   |  1.278       0.434     3.760          0.76
CT1                  |  0.587       0.184     1.865          0.66
CT2                  |  0.075       0.004     1.326          0.11
DC                   |  0.233       0.004    12.664          0.06
DE                   |  1.254       0.075    20.988          0.11
FL                   |  1.155       0.688     1.941          2.96
GA1                  |  1.238       0.669     2.294          2.18
GA2                  |  1.026       0.548     1.920          2.11
IA                   |  1.904       0.730     4.966          0.95
ID                   |  1.075       0.330     3.499          0.64
IL                   |  0.762       0.415     1.401          2.23
IN                   |  0.827       0.407     1.679          1.69
KY1                  |  0.978       0.500     1.912          1.87
LA                   |  2.428       1.211     4.870          1.74
MA1                  |  0.717       0.251     2.042          0.80
MA2                  |  1.543       0.529     4.504          0.77
MD1                  |  0.622       0.244     1.583          1.00
MD2                  |  4.420       1.441    13.552          0.70
ME1                  |  0.399       0.040     4.007          0.17
ME2                  |  1.040       0.297     3.640          0.57
ME3                  |  0.450       0.086     2.356          0.33
MI                   |  1.365       0.709     2.628          1.95
MN1                  |  0.322       0.122     0.850          0.93
MN2                  |  1.029       0.454     2.336          1.28
MO                   |  1.301       0.765     2.212          2.85
MS                   |  0.747       0.330     1.692          1.29
NC                   |  1.132       0.640     2.001          2.51
ND                   |  0.389       0.033     4.602          0.15
NE                   |  1.933       0.587     6.371          0.62
NH1                  |  2.093       0.205    21.389          0.17
NH2                  |  1.394       0.234     8.295          0.28
NH3                  |  0.267       0.027     2.691          0.17
NJ                   |  2.107       0.890     4.986          1.17
NM                   |  1.013       0.316     3.242          0.65
NV                   |  0.938       0.184     4.779          0.34
NY1                  |  0.709       0.387     1.297          2.26
NY2                  |  0.728       0.392     1.351          2.17
OH1                  |  1.432       0.845     2.427          2.88
OH2                  |  1.217       0.693     2.134          2.57
OK                   |  1.939       0.839     4.482          1.23
OR                   |  1.395       0.533     3.651          0.94
PA                   |  0.589       0.334     1.038          2.53
RI1                  |  0.887       0.052    15.197          0.11
RI2                  |  1.021       0.060    17.309          0.11
SC1                  |  0.540       0.281     1.038          1.96
SC2                  |  1.004       0.525     1.921          1.99
SD                   |  2.762       0.501    15.213          0.31
TN                   |  1.869       1.007     3.468          2.17
TX                   |  0.901       0.649     1.251          6.25
UT1                  |  1.003       0.383     2.627          0.94
UT2                  |  1.281       0.439     3.742          0.77
UT3                  |  0.392       0.101     1.520          0.49
VA1                  |  1.215       0.599     2.462          1.70
VA2                  |  1.663       0.828     3.343          1.73
VA3                  |  0.850       0.446     1.622          2.00
VA4                  |  0.704       0.369     1.346          1.99
VT                   |  1.018       0.060    17.180          0.11
WA                   |  1.013       0.468     2.194          1.43
WI                   |  1.136       0.549     2.350          1.61
WV                   |  0.940       0.382     2.315          1.07
WY                   |  0.176       0.007     4.202          0.09
Alberta              |  0.642       0.299     1.379          1.46
BC1                  |  1.298       0.651     2.589          1.77
BC2                  |  1.098       0.517     2.331          1.51
Manitoba             | 10.441       1.235    88.251          0.20
NewBrunswick         |  1.512       0.240     9.509          0.27
Newfoundland         |  0.295       0.024     3.641          0.14
NovaScotia           |  9.676       1.148    81.583          0.20
Ontario              |  0.971       0.612     1.541          3.63
PrinceEdwardI        |  0.212       0.008     5.433          0.09
Quebec               |  1.558       0.904     2.687          2.72
```

```
Saskatchewan         |  1.990       0.063    62.843          0.08
Yukon                |  2.520       0.037   173.486          0.05
---------------------+--------------------------------------------------
D+L pooled ES        |  1.059       0.963     1.165        100.00
---------------------+--------------------------------------------------

  Heterogeneity chi-squared =  81.48 (d.f. = 77) p = 0.342
  I-squared (variation in ES attributable to heterogeneity) =   5.5%
  Estimate of between-study variance Tau-squared =  0.0098

  Test of ES=1 : z=   1.19 p = 0.235
```

Figure 6.4.1: Forest Plot

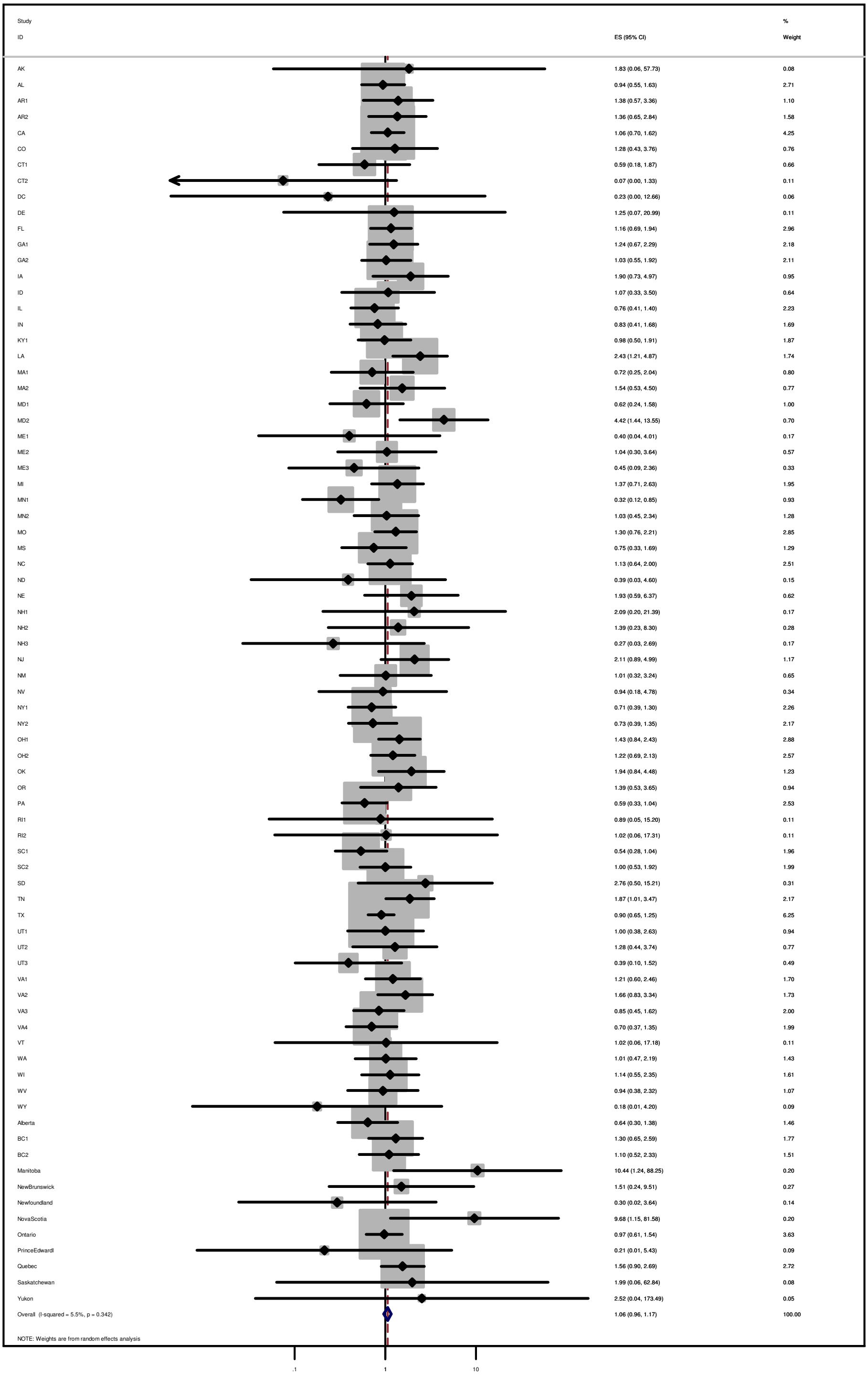
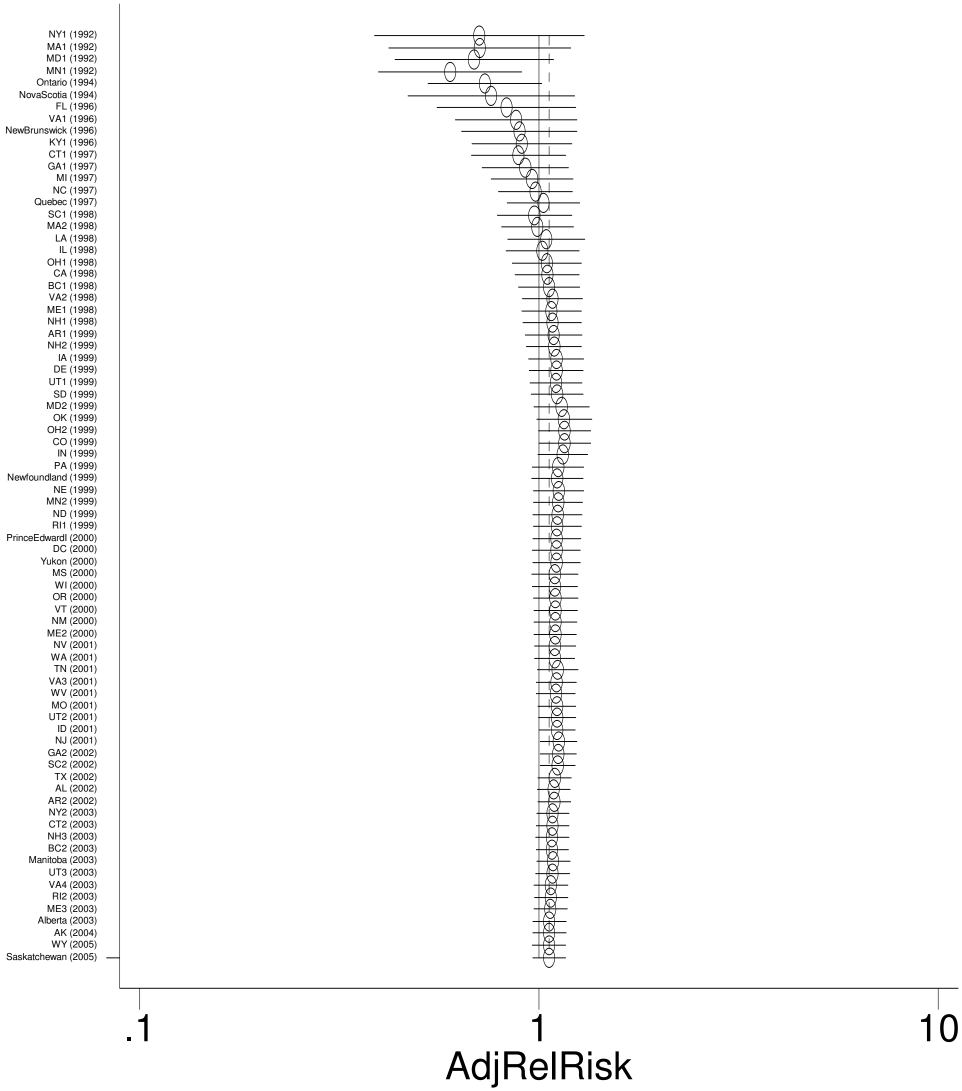


Figure 6.4.2: Cumulative Random Effects Meta-Analysis



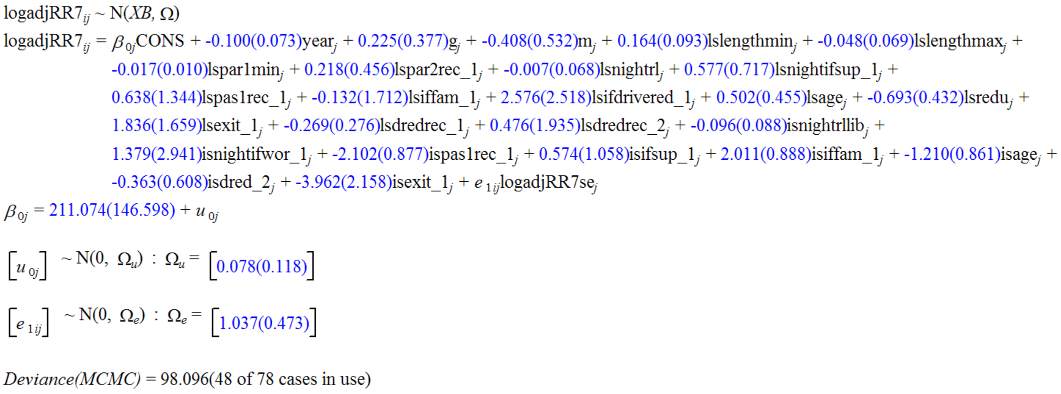
Appendix 5: Meta-regression for 16 year old drivers

Table 6.5.1: Meta-regression

```
Meta-regression                                       Number of obs  =      48
REML estimate of between-study variance               tau2           =   .0301
% residual variation due to heterogeneity             I-squared_res  =   2.46%
Proportion of between-study variance explained        Adj R-squared  =  73.69%
Joint test for all covariates                         Model F(25,22) =    1.18
With Knapp-Hartung modification                       Prob > F       =  0.3466
------------------------------------------------------------------------------
   logadjRR7 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.1074513   .0644252    -1.67   0.110    -.2410609    .0261584
    _Iiihs_2 |   .2196012   .3505599     0.63   0.537    -.5074155     .946618
    _Iiihs_3 |  -.3659963   .4823285    -0.76   0.456    -1.366284    .6342918
 lslengthmin |    .170582   .0852881     2.00   0.058    -.0062947    .3474587
 lslengthmax |  -.0454361    .065892    -0.69   0.498    -.1820876    .0912155
   lspar1min |  -.0172798   .0097601    -1.77   0.091    -.0375211    .0029614
_Ilspar2re~1 |   .2156036   .4304987     0.50   0.621    -.6771961    1.108403
    lsnightrl |  -.0020125   .0600872    -0.03   0.974    -.1266257    .1226007
lsnightifsup |   .5086947   .6584168     0.77   0.448    -.8567781    1.874168
   lspas1rec |   .6685581   1.296098     0.52   0.611    -2.019386    3.356502
     lsiffam |  -.1769388   1.641356    -0.11   0.915    -3.580903    3.227025
lsifdrivered |    2.59863    2.42896     1.07   0.296    -2.438725    7.635985
       lsage |   .5202102    .416306     1.25   0.225    -.3431557    1.383576
      lsredu |  -.6997212   .4169904    -1.68   0.107    -1.564506     .165064
      lsexit |   1.778373   1.589924     1.12   0.275    -1.518927    5.075673
_Ilsdredre~1 |  -.2630512   .2516453    -1.05   0.307    -.7849315    .2588292
_Ilsdredre~2 |   .4599831   1.870905     0.25   0.808    -3.420036    4.340002
isnightrllib |  -.0955647   .0806419    -1.19   0.249    -.2628058    .0716763
isnightifwor |   1.290175    2.80328     0.46   0.650    -4.523471    7.103821
   ispas1rec |  -2.160185   .8043738    -2.69   0.014    -3.828354   -.4920157
     isifsup |   .6340943   .9821616     0.65   0.525    -1.402784    2.670973
     isiffam |   2.114694   .7944557     2.66   0.014     .4670933    3.762294
       isage |  -1.211882   .8049229    -1.51   0.146     -2.88119    .4574256
   _Iisdred_2 |   -.322202   .5473806    -0.59   0.562      -1.4574    .8129958
      isexit |    -3.9084   2.041057    -1.91   0.069    -8.141293    .3244924
       _cons |   226.3739   129.2742     1.75   0.094    -41.72433    494.4721
------------------------------------------------------------------------------
```

Figure 6.5.1: Full Bayesian Meta-regression

BAYES, GIBBS SAMPLING, RANDOM MODEL, 16 YEAR OLDS (50,000 iterations, burn-in period of 1000 iterations), RANDOM SEED 1



$$\text{logadjRR7}_{ij} \sim \text{N}(XB,\ \Omega)$$

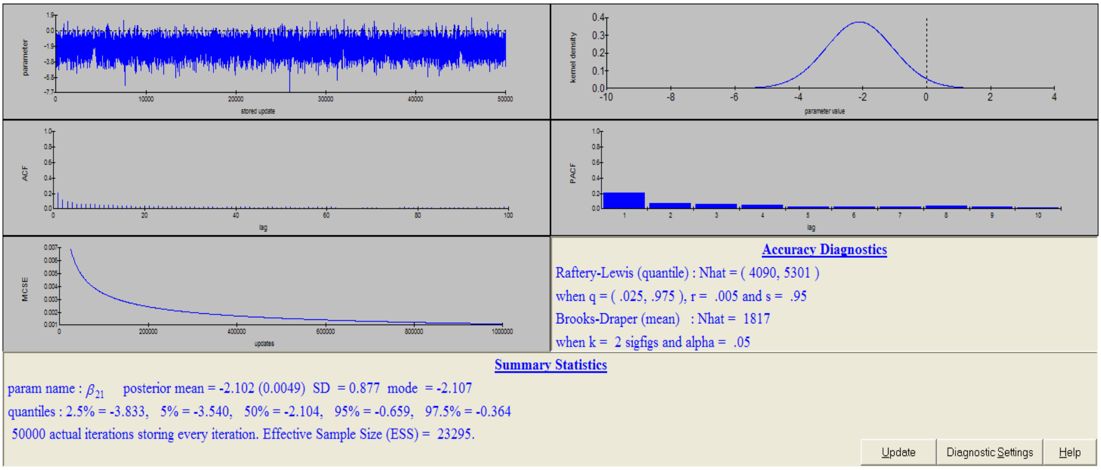
$$\text{logadjRR7}_{ij} = \beta_{0j}\text{CONS} + \text{-}0.100(0.073)\text{year}_j + 0.225(0.377)\text{g}_j + \text{-}0.408(0.532)\text{m}_j + 0.164(0.093)\text{lslengthmin}_j + \text{-}0.048(0.069)\text{lslengthmax}_j + \text{-}0.017(0.010)\text{lspar1min}_j + 0.218(0.456)\text{lspar2rec\_1}_j + \text{-}0.007(0.068)\text{lsnightrl}_j + 0.577(0.717)\text{lsnightifsup\_1}_j + 0.638(1.344)\text{lspas1rec\_1}_j + \text{-}0.132(1.712)\text{lsiffam\_1}_j + 2.576(2.518)\text{lsifdrivered\_1}_j + 0.502(0.455)\text{lsage}_j + \text{-}0.693(0.432)\text{lsredu}_j + 1.836(1.659)\text{lsexit\_1}_j + \text{-}0.269(0.276)\text{lsdredrec\_1}_j + 0.476(1.935)\text{lsdredrec\_2}_j + \text{-}0.096(0.088)\text{isnightrllib}_j + 1.379(2.941)\text{isnightifwor\_1}_j + \text{-}2.102(0.877)\text{ispas1rec\_1}_j + 0.574(1.058)\text{isifsup\_1}_j + 2.011(0.888)\text{isiffam\_1}_j + \text{-}1.210(0.861)\text{isage}_j + \text{-}0.363(0.608)\text{isdred\_2}_j + \text{-}3.962(2.158)\text{isexit\_1}_j + e_{1ij}\text{logadjRR7se}_j$$

$$\beta_{0j} = 211.074(146.598) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.078(0.118) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 1.037(0.473) \end{bmatrix}$$

Deviance(MCMC) = 98.096(48 of 78 cases in use)



Accuracy Diagnostics

Raftery-Lewis (quantile) : Nhat = (4090, 5301)

when q = (.025, .975), r = .005 and s = .95

Brooks-Draper (mean) : Nhat = 1817

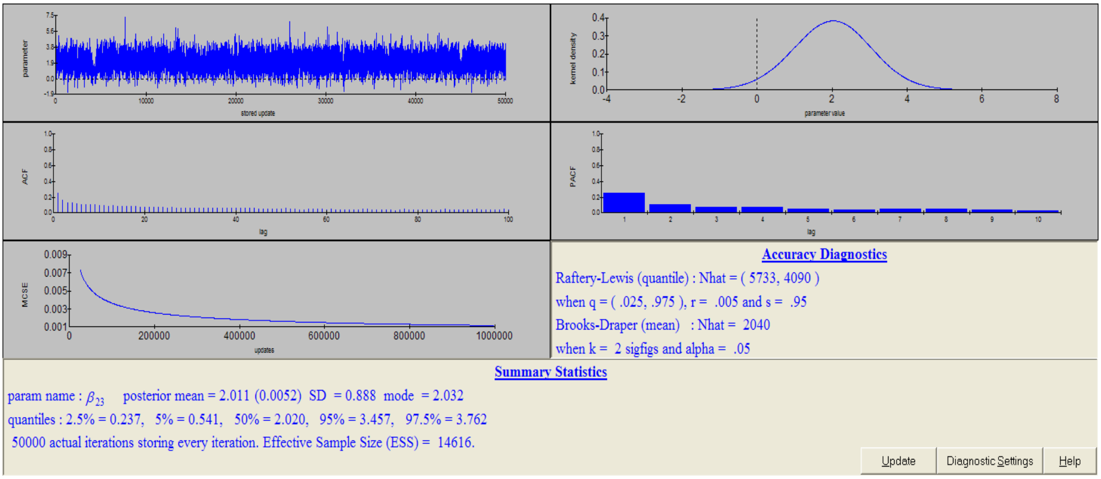
when k = 2 sigfigs and alpha = .05

Summary Statistics

param name : β_{21} posterior mean = -2.102 (0.0049) SD = 0.877 mode = -2.107

quantiles : 2.5% = -3.833, 5% = -3.540, 50% = -2.104, 95% = -0.659, 97.5% = -0.364

50000 actual iterations storing every iteration. Effective Sample Size (ESS) = 23295.



Accuracy Diagnostics

Raftery-Lewis (quantile) : Nhat = (5733, 4090)

when q = (.025, .975), r = .005 and s = .95

Brooks-Draper (mean) : Nhat = 2040

when k = 2 sigfigs and alpha = .05

Summary Statistics

param name : β_{23} posterior mean = 2.011 (0.0052) SD = 0.888 mode = 2.032

quantiles : 2.5% = 0.237, 5% = 0.541, 50% = 2.020, 95% = 3.457, 97.5% = 3.762

50000 actual iterations storing every iteration. Effective Sample Size (ESS) = 14616.

BAYES, GIBSS SAMPLING, RANDOM MODEL, 16 YEAR OLDS (50,000 iterations, burn-in period of 1,000 iterations), RANDOM SEED 2

$$\text{logadjRR7}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\begin{aligned}\text{logadjRR7}_{ij} = {} & \beta_{0j}\text{CONS} + -0.098(0.073)\text{year}_j + 0.225(0.376)\text{g}_j + -0.412(0.525)\text{m}_j + 0.163(0.092)\text{lslengthmin}_j + -0.048(0.069)\text{lslengthmax}_j + \\ & -0.017(0.010)\text{lspar1min}_j + 0.221(0.454)\text{lspar2rec\_1}_j + -0.007(0.068)\text{lsnightrl}_j + 0.573(0.717)\text{lsnightifsup\_1}_j + \\ & 0.640(1.335)\text{lspas1rec\_1}_j + -0.134(1.709)\text{lsiffam\_1}_j + 2.577(2.477)\text{lsifdrivered\_1}_j + 0.498(0.450)\text{lsage}_j + -0.690(0.429)\text{lsredu}_j + \\ & 1.831(1.647)\text{lsexit\_1}_j + -0.266(0.273)\text{lsdredrec\_1}_j + 0.472(1.926)\text{lsdredrec\_2}_j + -0.096(0.088)\text{isnightrllib}_j + \\ & 1.400(2.927)\text{isnightifwor\_1}_j + -2.095(0.869)\text{ispas1rec\_1}_j + 0.566(1.050)\text{isifsup\_1}_j + 2.000(0.880)\text{isiffam\_1}_j + -1.207(0.858)\text{isage}_j + \\ & -0.369(0.606)\text{isdred\_2}_j + -3.976(2.157)\text{isexit\_1}_j + e_{1ij}\text{logadjRR7se}_j\end{aligned}$$

$$\beta_{0j} = 207.961(145.489) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ : \ \Omega_u = \begin{bmatrix} 0.080(0.120) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ : \ \Omega_e = \begin{bmatrix} 1.019(0.477) \end{bmatrix}$$

Deviance(MCMC) = 97.318(48 of 78 cases in use)

BAYES, GIBSS SAMPLING, RANDOM MODEL, 16 YEAR OLDS (50,000 iterations, burn-in period of 1,000 iterations), RANDOM SEED 3

$$\text{logadjRR7}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\begin{aligned}\text{logadjRR7}_{ij} = {} & \beta_{0j}\text{CONS} + -0.099(0.074)\text{year}_j + 0.223(0.381)\text{g}_j + -0.403(0.537)\text{m}_j + 0.163(0.093)\text{lslengthmin}_j + -0.047(0.069)\text{lslengthmax}_j + \\ & -0.017(0.011)\text{lspar1min}_j + 0.222(0.456)\text{lspar2rec\_1}_j + -0.007(0.068)\text{lsnightrl}_j + 0.570(0.720)\text{lsnightifsup\_1}_j + \\ & 0.631(1.345)\text{lspas1rec\_1}_j + -0.130(1.710)\text{lsiffam\_1}_j + 2.605(2.534)\text{lsifdrivered\_1}_j + 0.500(0.454)\text{lsage}_j + -0.692(0.434)\text{lsredu}_j + \\ & 1.816(1.663)\text{lsexit\_1}_j + -0.270(0.274)\text{lsdredrec\_1}_j + 0.487(1.942)\text{lsdredrec\_2}_j + -0.096(0.088)\text{isnightrllib}_j + \\ & 1.362(2.961)\text{isnightifwor\_1}_j + -2.095(0.882)\text{ispas1rec\_1}_j + 0.569(1.066)\text{isifsup\_1}_j + 2.003(0.898)\text{isiffam\_1}_j + -1.201(0.863)\text{isage}_j + \\ & -0.363(0.610)\text{isdred\_2}_j + -3.955(2.182)\text{isexit\_1}_j + e_{1ij}\text{logadjRR7se}_j\end{aligned}$$

$$\beta_{0j} = 210.169(148.128) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ : \ \Omega_u = \begin{bmatrix} 0.081(0.128) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ : \ \Omega_e = \begin{bmatrix} 1.038(0.481) \end{bmatrix}$$

Deviance(MCMC) = 97.334(48 of 78 cases in use)

BAYES, GIBSS SAMPLING, RANDOM MODEL, 16 YEAR OLDS (50,000 iterations, burn-in period of 1,000 iterations), RANDOM SEED 4

$$\text{logadjRR7}_{ij} \sim \text{N}(XB, \Omega)$$

$$\begin{aligned}\text{logadjRR7}_{ij} = {} & \beta_{0j}\text{CONS} + -0.095(0.076)\text{year}_j + 0.225(0.381)\text{g}_j + -0.414(0.538)\text{m}_j + 0.163(0.094)\text{lslengthmin}_j + -0.048(0.068)\text{lslengthmax}_j + \\ & -0.017(0.010)\text{lspar1min}_j + 0.217(0.458)\text{lspar2rec\_1}_j + -0.009(0.069)\text{lsnightrl}_j + 0.592(0.718)\text{lsnightifsup\_1}_j + \\ & 0.625(1.330)\text{lspas1rec\_1}_j + -0.115(1.699)\text{lsiffam\_1}_j + 2.558(2.501)\text{lsifdrivered\_1}_j + 0.493(0.453)\text{lsage}_j + -0.689(0.427)\text{lsredu}_j + \\ & 1.838(1.647)\text{lsexit\_1}_j + -0.273(0.279)\text{lsdredrec\_1}_j + 0.476(1.918)\text{lsdredrec\_2}_j + -0.096(0.089)\text{isnightrllib}_j + \\ & 1.442(2.929)\text{isnightifwor\_1}_j + -2.071(0.884)\text{ispas1rec\_1}_j + 0.554(1.062)\text{isifsup\_1}_j + 1.968(0.910)\text{isiffam\_1}_j + -1.199(0.869)\text{isage}_j + \\ & -0.366(0.608)\text{isdred\_2}_j + -3.994(2.167)\text{isexit\_1}_j + e_{1ij}\text{logadjRR7se}_j\end{aligned}$$

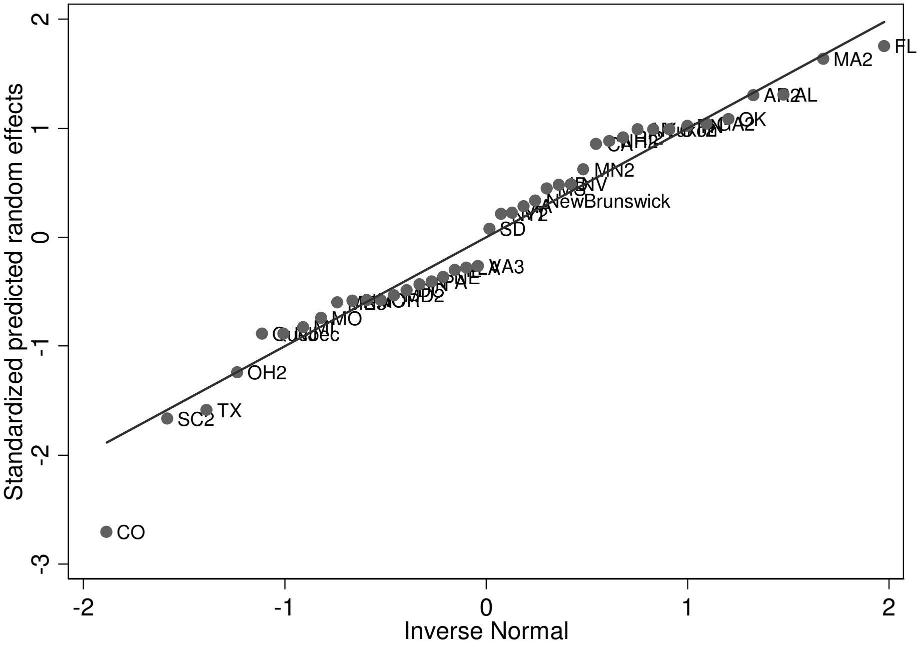
$$\beta_{0j} = 202.522(151.786) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.093(0.140) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 1.007(0.505) \end{bmatrix}$$

Deviance(MCMC) = 93.962(48 of 78 cases in use)

Figure 6.5.2: Normal probability plot of standardized shrunken residuals



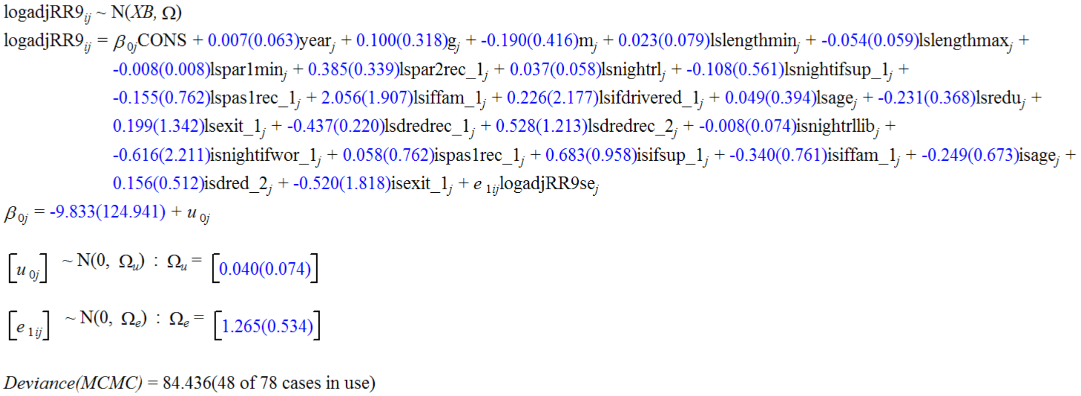
Appendix 6: Meta-regression for 18 year old drivers

Table 6.6.1: Meta-regression

```
Meta-regression                                     Number of obs  =      48
REML estimate of between-study variance             tau2           =       0
% residual variation due to heterogeneity           I-squared_res  =  11.43%
Proportion of between-study variance explained      Adj R-squared  =      .%
Joint test for all covariates                       Model F(25,22) =    0.87
With Knapp-Hartung modification                     Prob > F       =  0.6351
------------------------------------------------------------------------------
   logadjRR9 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0108451   .0527708     0.21   0.839    -.0985949    .1202851
    _Iiihs_2 |   .0730795   .2698139     0.27   0.789    -.4864803    .6326394
    _Iiihs_3 |  -.2167965   .3504076    -0.62   0.542    -.9434974    .5099044
 lslengthmin |   .0286035   .0680428     0.42   0.678    -.1125086    .1697156
 lslengthmax |   -.055624   .0538544    -1.03   0.313    -.1673112    .0560632
   lspar1min |  -.0082214   .0070555    -1.17   0.256    -.0228536    .0064109
_Ilspar2re~1 |   .4045334   .2982076     1.36   0.189    -.2139113    1.022978
    lsnightrl |   .0426433   .0478848     0.89   0.383    -.0566638    .1419503
lsnightifsup |  -.1207104   .4950782    -0.24   0.810     -1.14744     .906019
   lspas1rec |  -.1823846   .6793941    -0.27   0.791    -1.591362    1.226593
     lsiffam |   2.119684   1.775762     1.19   0.245    -1.563021    5.802389
lsifdrivered |   .2768912   2.012674     0.14   0.892    -3.897139    4.450921
       lsage |   .1051972   .3343059     0.31   0.756    -.5881109    .7985052
      lsredu |  -.2408143   .3354282    -0.72   0.480    -.9364498    .4548212
      lsexit |   .2461586   1.209511     0.20   0.841    -2.262214    2.754531
_Ilsdredre~1 |  -.4232671   .1889674    -2.24   0.036    -.8151615   -.0313728
_Ilsdredre~2 |   .4547635   1.101381     0.41   0.684    -1.829361    2.738888
isnightrllib |   -.018547   .0615327    -0.30   0.766    -.1461581    .1090641
isnightifwor |  -.4938242   1.961843    -0.25   0.804    -4.562438    3.574789
   ispas1rec |   .0643973   .6731866     0.10   0.925    -1.331706    1.460501
     isifsup |   .6135038   .8539684     0.72   0.480    -1.157518    2.384526
     isiffam |  -.3337816   .6759174    -0.49   0.626    -1.735548    1.067985
       isage |  -.3207081   .5763126    -0.56   0.583    -1.515907    .8744911
  _Iisdred_2 |   .1629145   .4435683     0.37   0.717    -.7569899    1.082819
      isexit |  -.6391904   1.569569    -0.41   0.688    -3.894278    2.615897
       _cons |  -17.42221   105.2834    -0.17   0.870    -235.7666    200.9222
------------------------------------------------------------------------------
```

Figure 6.6.1: Full Bayesian Meta-regression

BAYES, GIBBS SAMPLING, RANDOM MODEL, 18 YEAR OLDS (50,000 iterations, burn-in period of 1000 iterations), RANDOM SEED 1



$$\text{logadjRR9}_{ij} \sim \mathrm{N}(XB, \Omega)$$

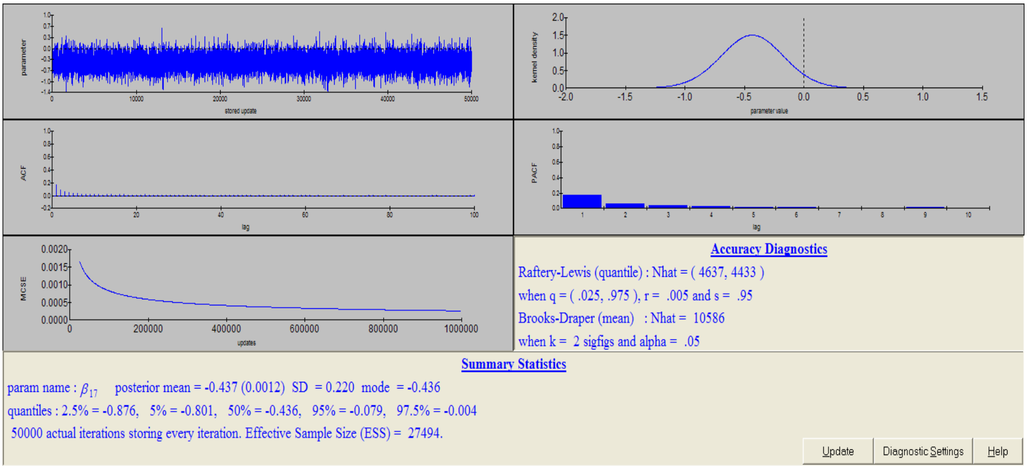
$$\begin{aligned}\text{logadjRR9}_{ij} = {} & \beta_{0j}\text{CONS} + 0.007(0.063)\text{year}_j + 0.100(0.318)\text{g}_j + -0.190(0.416)\text{m}_j + 0.023(0.079)\text{lslengthmin}_j + -0.054(0.059)\text{lslengthmax}_j + \\ & -0.008(0.008)\text{lspar1min}_j + 0.385(0.339)\text{lspar2rec\_1}_j + 0.037(0.058)\text{lsnightrl}_j + -0.108(0.561)\text{lsnightifsup\_1}_j + \\ & -0.155(0.762)\text{lspas1rec\_1}_j + 2.056(1.907)\text{lsiffam\_1}_j + 0.226(2.177)\text{lsifdrivered\_1}_j + 0.049(0.394)\text{lsage}_j + -0.231(0.368)\text{lsredu}_j + \\ & 0.199(1.342)\text{lsexit\_1}_j + -0.437(0.220)\text{lsdredrec\_1}_j + 0.528(1.213)\text{lsdredrec\_2}_j + -0.008(0.074)\text{isnightrllib}_j + \\ & -0.616(2.211)\text{isnightifwor\_1}_j + 0.058(0.762)\text{ispas1rec\_1}_j + 0.683(0.958)\text{isifsup\_1}_j + -0.340(0.761)\text{isiffam\_1}_j + -0.249(0.673)\text{isage}_j + \\ & 0.156(0.512)\text{isdred\_2}_j + -0.520(1.818)\text{isexit\_1}_j + e_{1ij}\text{logadjRR9se}_j\end{aligned}$$

$$\beta_{0j} = -9.833(124.941) + u_{0j}$$

$$\left[u_{0j}\right] \sim \mathrm{N}(0,\ \Omega_u)\ :\ \Omega_u = \left[0.040(0.074)\right]$$

$$\left[e_{1ij}\right] \sim \mathrm{N}(0,\ \Omega_e)\ :\ \Omega_e = \left[1.265(0.534)\right]$$

Deviance(MCMC) = 84.436(48 of 78 cases in use)



Accuracy Diagnostics

Raftery-Lewis (quantile) : Nhat = (4637, 4433)

when q = (.025, .975), r = .005 and s = .95

Brooks-Draper (mean) : Nhat = 10586

when k = 2 sigfigs and alpha = .05

Summary Statistics

param name : β_{17} posterior mean = -0.437 (0.0012) SD = 0.220 mode = -0.436

quantiles : 2.5% = -0.876, 5% = -0.801, 50% = -0.436, 95% = -0.079, 97.5% = -0.004

50000 actual iterations storing every iteration. Effective Sample Size (ESS) = 27494.

BAYES, GIBSS SAMPLING, RANDOM MODEL, 18 YEAR OLDS (50,000 iterations, burn-in period of 1,000 iterations), RANDOM SEED 2

$$\text{logadjRR9}_{ij} \sim \mathrm{N}(XB,\ \Omega)$$

$$\begin{aligned}\text{logadjRR9}_{ij} = {} & \beta_{0j}\text{CONS} + 0.009(0.061)\text{year}_j + 0.095(0.311)\text{g}_j + -0.199(0.406)\text{m}_j + 0.023(0.078)\text{lslengthmin}_j + -0.054(0.059)\text{lslengthmax}_j + \\ & -0.008(0.008)\text{lspar1min}_j + 0.387(0.339)\text{lspar2rec\_1}_j + 0.038(0.057)\text{lsnightrl}_j + -0.114(0.560)\text{lsnightifsup\_1}_j + \\ & -0.163(0.759)\text{lspas1rec\_1}_j + 2.072(1.913)\text{lsiffam\_1}_j + 0.190(2.170)\text{lsifdrivered\_1}_j + 0.055(0.389)\text{lsage}_j + -0.226(0.368)\text{lsredu}_j + \\ & 0.198(1.345)\text{lsexit\_1}_j + -0.435(0.219)\text{lsdredrec\_1}_j + 0.509(1.221)\text{lsdredrec\_2}_j + -0.009(0.074)\text{isnightrllib}_j + \\ & -0.584(2.177)\text{isnightifwor\_1}_j + 0.060(0.760)\text{ispas1rec\_1}_j + 0.665(0.952)\text{isifsup\_1}_j + -0.340(0.760)\text{isiffam\_1}_j + -0.250(0.670)\text{isage}_j + \\ & 0.163(0.504)\text{isdred\_2}_j + -0.529(1.795)\text{isexit\_1}_j + e_{1ij}\text{logadjRR9se}_j\end{aligned}$$

$$\beta_{0j} = -13.980(121.271) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.035(0.064) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 1.270(0.503) \end{bmatrix}$$

Deviance(MCMC) = 85.632(48 of 78 cases in use)

BAYES, GIBSS SAMPLING, RANDOM MODEL, 18 YEAR OLDS (50,000 iterations, burn-in period of 1,000 iterations), RANDOM SEED 3

$$\text{logadjRR9}_{ij} \sim \mathrm{N}(XB, \Omega)$$

$$\begin{aligned}\text{logadjRR9}_{ij} = {} & \beta_{0j}\text{CONS} + 0.007(0.062)\text{year}_j + 0.102(0.316)\text{g}_j + -0.190(0.413)\text{m}_j + 0.022(0.078)\text{lslengthmin}_j + -0.054(0.059)\text{lslengthmax}_j + \\ & -0.007(0.008)\text{lspar1min}_j + 0.379(0.342)\text{lspar2rec\_1}_j + 0.037(0.058)\text{lsnightrl}_j + -0.105(0.564)\text{lsnightifsup\_1}_j + \\ & -0.152(0.761)\text{lspas1rec\_1}_j + 2.063(1.892)\text{lsiffam\_1}_j + 0.185(2.173)\text{lsifdrivered\_1}_j + 0.035(0.403)\text{lsage}_j + -0.227(0.368)\text{lsredu}_j + \\ & 0.193(1.338)\text{lsexit\_1}_j + -0.441(0.222)\text{lsdredrec\_1}_j + 0.551(1.228)\text{lsdredrec\_2}_j + -0.005(0.076)\text{isnightrllib}_j + \\ & -0.641(2.189)\text{isnightifwor\_1}_j + 0.059(0.762)\text{ispas1rec\_1}_j + 0.685(0.954)\text{isifsup\_1}_j + -0.340(0.760)\text{isiffam\_1}_j + -0.229(0.681)\text{isage}_j + \\ & 0.163(0.512)\text{isdred\_2}_j + -0.498(1.795)\text{isexit\_1}_j + e_{1ij}\text{logadjRR9se}_j\end{aligned}$$

$$\beta_{0j} = -10.417(123.904) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ : \ \Omega_u = \begin{bmatrix} 0.047(0.083) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ : \ \Omega_e = \begin{bmatrix} 1.238(0.535) \end{bmatrix}$$

Deviance(MCMC) = 83.332(48 of 78 cases in use)

BAYES, GIBSS SAMPLING, RANDOM MODEL, 18 YEAR OLDS (50,000 iterations, burn-in period of 1,000 iterations), RANDOM SEED 4

$$\text{logadjRR9}_{ij} \sim \mathrm{N}(XB, \Omega)$$

$$\begin{aligned}\text{logadjRR9}_{ij} = {} & \beta_{0j}\text{CONS} + 0.008(0.062)\text{year}_j + 0.102(0.314)\text{g}_j + -0.196(0.412)\text{m}_j + 0.024(0.079)\text{lslengthmin}_j + -0.054(0.059)\text{lslengthmax}_j + \\ & -0.008(0.008)\text{lspar1min}_j + 0.382(0.342)\text{lspar2rec\_1}_j + 0.038(0.058)\text{lsnightrl}_j + -0.108(0.567)\text{lsnightifsup\_1}_j + \\ & -0.165(0.761)\text{lspas1rec\_1}_j + 2.059(1.906)\text{lsiffam\_1}_j + 0.206(2.202)\text{lsifdrivered\_1}_j + 0.044(0.399)\text{lsage}_j + -0.228(0.370)\text{lsredu}_j + \\ & 0.206(1.355)\text{lsexit\_1}_j + -0.437(0.222)\text{lsdredrec\_1}_j + 0.519(1.231)\text{lsdredrec\_2}_j + -0.007(0.075)\text{isnightrllib}_j + \\ & -0.574(2.184)\text{isnightifwor\_1}_j + 0.051(0.765)\text{ispas1rec\_1}_j + 0.684(0.964)\text{isifsup\_1}_j + -0.336(0.765)\text{isiffam\_1}_j + -0.238(0.681)\text{isage}_j + \\ & 0.156(0.512)\text{isdred\_2}_j + -0.535(1.798)\text{isexit\_1}_j + e_{1ij}\text{logadjRR9se}_j\end{aligned}$$

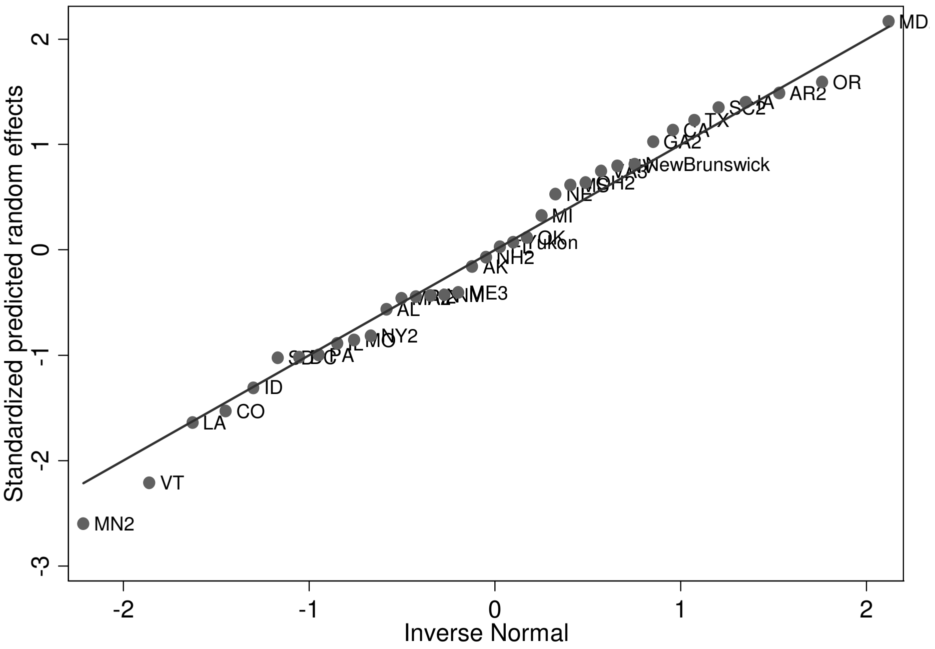
$$\beta_{0j} = -12.145(123.415) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ : \ \Omega_u = \begin{bmatrix} 0.041(0.072) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ : \ \Omega_e = \begin{bmatrix} 1.271(0.548) \end{bmatrix}$$

Deviance(MCMC) = 84.931(48 of 78 cases in use)

Figure 6.6.2: Normal probability plot of standardized shrunken residuals



Appendix 7: Meta-regression for 19 year old drivers

Table 6.7.1: Meta-regression

```
Meta-regression                                        Number of obs  =      47
REML estimate of between-study variance                tau2           =       0
% residual variation due to heterogeneity              I-squared_res  =   0.00%
Proportion of between-study variance explained         Adj R-squared  = 100.00%
Joint test for all covariates                          Model F(25,21) =    1.53
With Knapp-Hartung modification                        Prob > F       =  0.1630
------------------------------------------------------------------------------
  logadjRR10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0477804   .0491847    -0.97   0.342    -.1500655    .0545047
     _Iiihs_2 |   .1787797   .2516661     0.71   0.485    -.3445887    .7021481
     _Iiihs_3 |   -.12238   .3198451    -0.38   0.706    -.7875344    .5427744
 lslengthmin |  -.0325761   .0668439    -0.49   0.631    -.1715856    .1064335
 lslengthmax |   .0362948   .0530123     0.68   0.501    -.0739503    .1465399
   lspar1min |   -.005056   .0067624    -0.75   0.463    -.0191193    .0090073
_Ilspar2re~1 |   .0971577   .2767202     0.35   0.729    -.4783135     .672629
   lsnightrl |    .103704   .0468253     2.21   0.038     .0063254    .2010826
lsnightifsup |  -.1092041    .443524    -0.25   0.808    -1.031563    .8131545
   lspas1rec |   .2855166    .599157     0.48   0.639    -.9604986    1.531532
     lsiffam |   .3092276   1.539185     0.20   0.843    -2.891682    3.510137
lsifdrivered |  -2.581543   1.863311    -1.39   0.180    -6.456511    1.293425
       lsage |  -.0221124   .3083675    -0.07   0.944    -.6633977    .6191728
      lsredu |  -.0666995    .260502    -0.26   0.800    -.6084432    .4750441
      lsexit |   2.586603    1.37388     1.88   0.074    -.2705375    5.443743
_Ilsdredre~1 |     .04917   .1744854     0.28   0.781    -.3136923    .4120323
_Ilsdredre~2 |  -2.029129   1.570542    -1.29   0.210     -5.29525    1.236992
isnightrllib |  -.0791428   .0577926    -1.37   0.185     -.199329    .0410435
isnightifwor |   3.952962   2.088877     1.89   0.072     -.391096    8.297021
   ispas1rec |  -.8806373   .6443105    -1.37   0.186    -2.220554    .4592797
     isifsup |  -.7353308    .835918    -0.88   0.389    -2.473717    1.003056
     isiffam |   .9064637   .6406229     1.41   0.172    -.4257846    2.238712
       isage |  -.4324577   .5493029    -0.79   0.440    -1.574796    .7098802
   _Iisdred_2 |   .7456382   .3945228     1.89   0.073    -.0748169    1.566093
      isexit |  -3.856032   1.611587    -2.39   0.026    -7.207511   -.5045527
       _cons |   103.1957   97.99118     1.05   0.304    -100.5881    306.9795
------------------------------------------------------------------------------
```

Figure 6.7.1: Full Bayesian Meta-regression

BAYES, GIBBS SAMPLING, RANDOM MODEL, 19 YEAR OLDS (50,000 iterations, burn-in period of 1000 iterations), RANDOM SEED 1

NOTE THAT MD2 HAS BEEN DELETED FROM THE FILE BECAUSE IT WAS IDENTIFIED AS AN OUTLIER.

$$\text{logadjRR10}_{ij} \sim \text{N}(XB,\ \Omega)$$

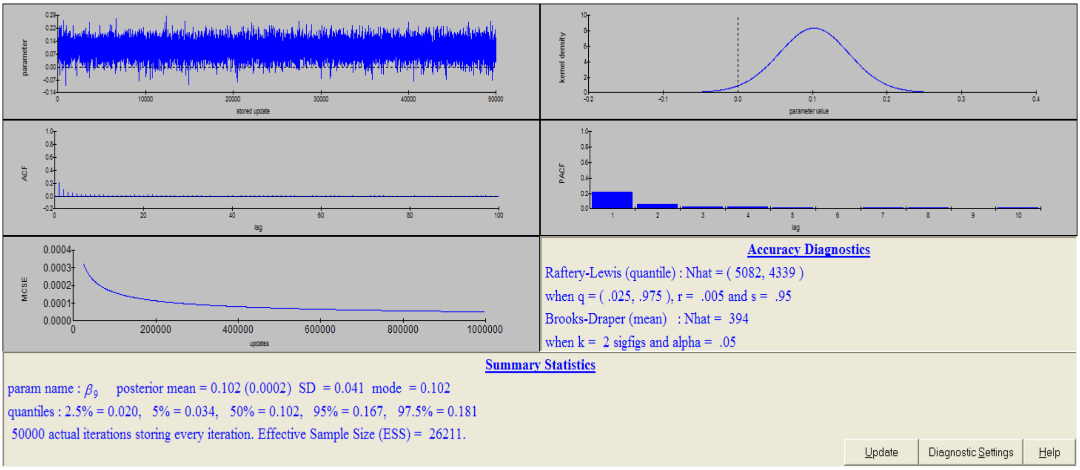
$$\begin{aligned}\text{logadjRR10}_{ij} = {} & \beta_{0j}\text{CONS} + -0.051(0.042)\text{year}_j + 0.176(0.213)\text{g}_j + -0.116(0.274)\text{m}_j + -0.032(0.055)\text{lslengthmin}_j + 0.037(0.042)\text{lslengthmax}_j + \\ & -0.005(0.006)\text{lspar1min}_j + 0.084(0.229)\text{lspar2rec\_1}_j + 0.102(0.041)\text{lsnightrl}_j + -0.102(0.371)\text{lsnightifsup\_1}_j + \\ & 0.265(0.492)\text{lspas1rec\_1}_j + 0.315(1.201)\text{lsiffam\_1}_j + -2.603(1.452)\text{lsifdrivered\_1}_j + -0.021(0.262)\text{lsage}_j + -0.060(0.210)\text{lsredu}_j + \\ & 2.543(1.085)\text{lsexit\_1}_j + 0.048(0.149)\text{lsdredrec\_1}_j + -2.072(1.241)\text{lsdredrec\_2}_j + -0.078(0.049)\text{isnightrllib}_j + \\ & 3.966(1.660)\text{isnightifwor\_1}_j + -0.890(0.540)\text{ispas1rec\_1}_j + -0.721(0.674)\text{isifsup\_1}_j + 0.924(0.535)\text{isiffam\_1}_j + \\ & -0.422(0.466)\text{isage}_j + 0.731(0.337)\text{isdred\_2}_j + -3.803(1.310)\text{isexit\_1}_j + e_{1ij}\text{logadjRR10se}_j\end{aligned}$$

$$\beta_{0j} = 109.521(83.920) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ : \ \Omega_u = \begin{bmatrix} 0.019(0.027) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ : \ \Omega_e = \begin{bmatrix} 0.586(0.241) \end{bmatrix}$$

Deviance(MCMC) = 44.924(47 of 77 cases in use)

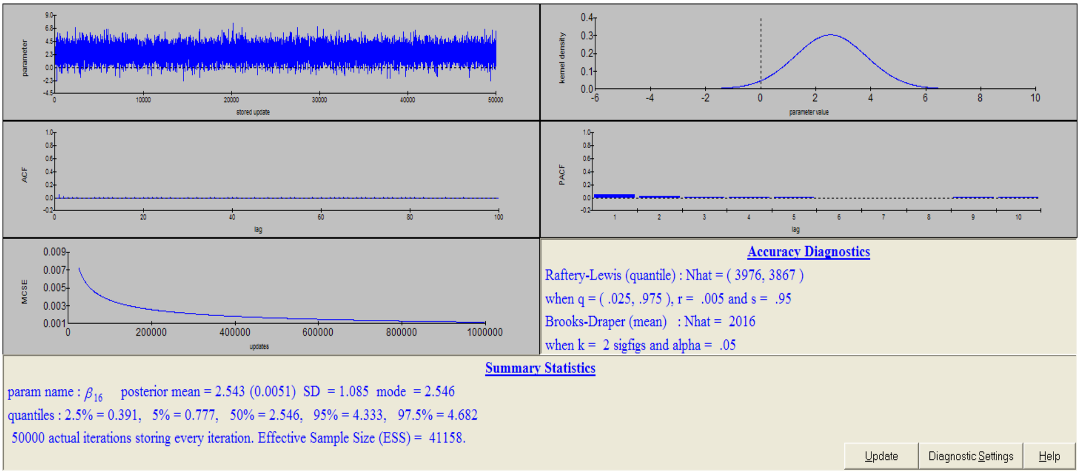


Accuracy Diagnostics

Raftery-Lewis (quantile) : Nhat = (5082, 4339)
when q = (.025, .975), r = .005 and s = .95
Brooks-Draper (mean) : Nhat = 394
when k = 2 sigfigs and alpha = .05

Summary Statistics

param name : β_9 posterior mean = 0.102 (0.0002) SD = 0.041 mode = 0.102
quantiles : 2.5% = 0.020, 5% = 0.034, 50% = 0.102, 95% = 0.167, 97.5% = 0.181
50000 actual iterations storing every iteration. Effective Sample Size (ESS) = 26211.

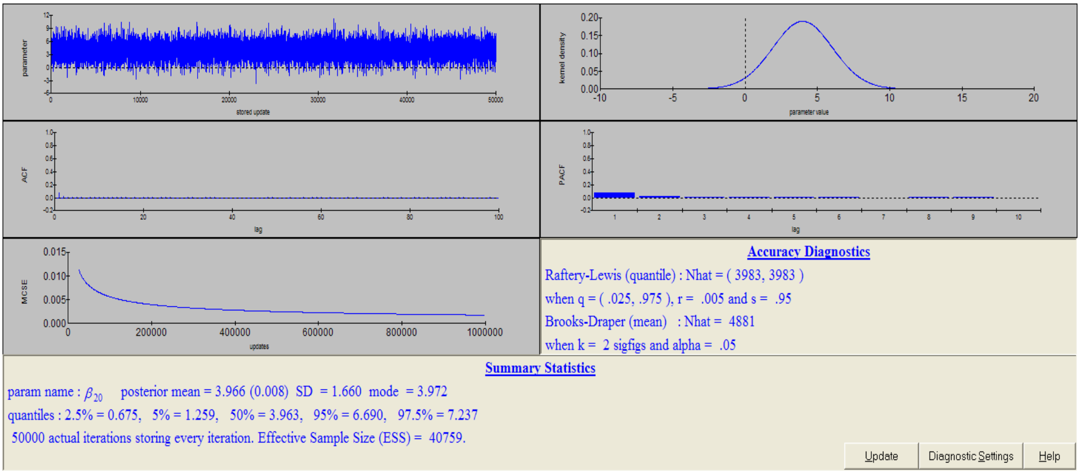


Accuracy Diagnostics

Raftery-Lewis (quantile) : Nhat = (3976, 3867)
when q = (.025, .975), r = .005 and s = .95
Brooks-Draper (mean) : Nhat = 2016
when k = 2 sigfigs and alpha = .05

Summary Statistics

param name : β_{16} posterior mean = 2.543 (0.0051) SD = 1.085 mode = 2.546
quantiles : 2.5% = 0.391, 5% = 0.777, 50% = 2.546, 95% = 4.333, 97.5% = 4.682
50000 actual iterations storing every iteration. Effective Sample Size (ESS) = 41158.

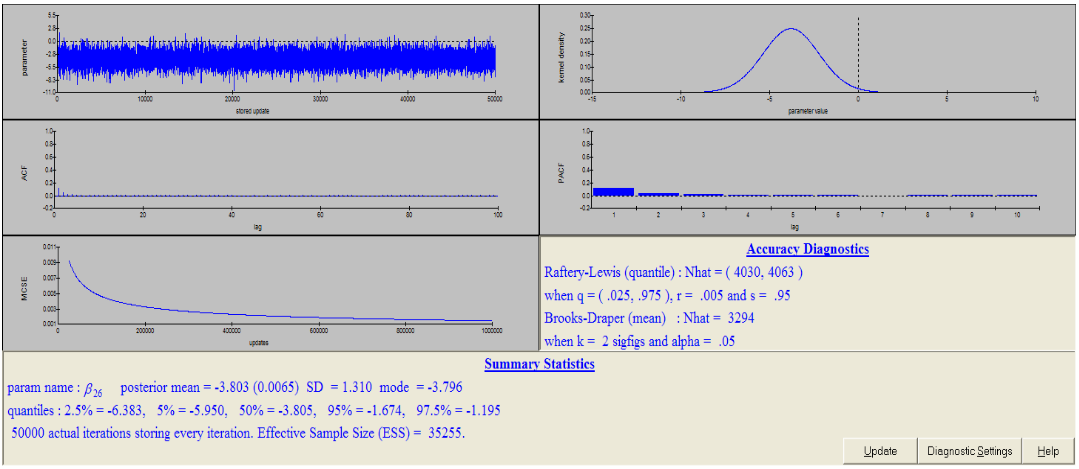


Accuracy Diagnostics

Raftery-Lewis (quantile) : Nhat = (3983, 3983)
when q = (.025, .975), r = .005 and s = .95
Brooks-Draper (mean) : Nhat = 4881
when k = 2 sigfigs and alpha = .05

Summary Statistics

param name : β_{20} posterior mean = 3.966 (0.008) SD = 1.660 mode = 3.972
quantiles : 2.5% = 0.675, 5% = 1.259, 50% = 3.963, 95% = 6.690, 97.5% = 7.237
50000 actual iterations storing every iteration. Effective Sample Size (ESS) = 40759.

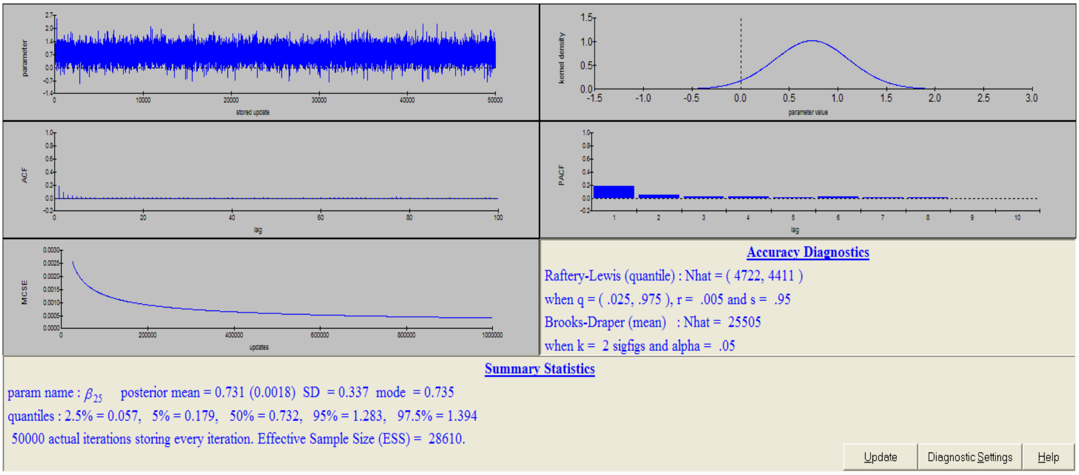


Accuracy Diagnostics

Raftery-Lewis (quantile) : Nhat = (4030, 4063)
when q = (.025, .975), r = .005 and s = .95
Brooks-Draper (mean) : Nhat = 3294
when k = 2 sigfigs and alpha = .05

Summary Statistics

param name : β_{26} posterior mean = -3.803 (0.0065) SD = 1.310 mode = -3.796
quantiles : 2.5% = -6.383, 5% = -5.950, 50% = -3.805, 95% = -1.674, 97.5% = -1.195
50000 actual iterations storing every iteration. Effective Sample Size (ESS) = 35255.



Accuracy Diagnostics

Raftery-Lewis (quantile) : Nhat = (4722, 4411)
when q = (.025, .975), r = .005 and s = .95
Brooks-Draper (mean) : Nhat = 25505
when k = 2 sigfigs and alpha = .05

Summary Statistics

param name : β_{25} posterior mean = 0.731 (0.0018) SD = 0.337 mode = 0.735
quantiles : 2.5% = 0.057, 5% = 0.179, 50% = 0.732, 95% = 1.283, 97.5% = 1.394
50000 actual iterations storing every iteration. Effective Sample Size (ESS) = 28610.

BAYES, GIBBS SAMPLING, RANDOM MODEL, 19 YEAR OLDS (50,000 iterations, burn-in period of 1000 iterations), RANDOM SEED 2

$$\text{logadjRR10}_{ij} \sim \mathrm{N}(XB, \Omega)$$

$$\begin{aligned}\text{logadjRR10}_{ij} = {} & \beta_{0j}\text{CONS} + -0.051(0.042)\text{year}_j + 0.178(0.211)\text{g}_j + -0.113(0.274)\text{m}_j + -0.031(0.055)\text{lslengthmin}_j + 0.037(0.041)\text{lslengthmax}_j + \\ & -0.005(0.006)\text{lspar1min}_j + 0.088(0.226)\text{lspar2rec\_1}_j + 0.102(0.040)\text{lsnightrl}_j + -0.100(0.368)\text{lsnightifsup\_1}_j + \\ & 0.262(0.487)\text{lspas1rec\_1}_j + 0.321(1.197)\text{lsiffam\_1}_j + -2.607(1.461)\text{lsifdrivered\_1}_j + -0.022(0.258)\text{lsage}_j + -0.061(0.209)\text{lsredu}_j + \\ & 2.544(1.076)\text{lsexit\_1}_j + 0.047(0.149)\text{lsdredrec\_1}_j + -2.072(1.233)\text{lsdredrec\_2}_j + -0.079(0.049)\text{isnightrllib}_j + \\ & 3.986(1.642)\text{isnightifwor\_1}_j + -0.894(0.532)\text{ispas1rec\_1}_j + -0.713(0.670)\text{isifsup\_1}_j + 0.925(0.527)\text{isiffam\_1}_j + \\ & -0.419(0.462)\text{isage}_j + 0.730(0.333)\text{isdred\_2}_j + -3.811(1.299)\text{isexit\_1}_j + e_{1ij}\text{logadjRR10se}_j\end{aligned}$$

$$\beta_{0j} = 109.415(82.965) + u_{0j}$$

$$\left[u_{0j}\right] \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \left[0.019(0.028)\right]$$

$$\left[e_{1ij}\right] \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \left[0.579(0.244)\right]$$

Deviance(MCMC) = 44.406(47 of 77 cases in use)

BAYES, GIBBS SAMPLING, RANDOM MODEL, 19 YEAR OLDS (50,000 iterations, burn-in period of 1000 iterations), RANDOM SEED 3

$$\text{logadjRR10}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\begin{aligned}
\text{logadjRR10}_{ij} = {} & \beta_{0j}\text{CONS} + \text{-}0.051(0.043)\text{year}_j + 0.177(0.213)\text{g}_j + \text{-}0.113(0.274)\text{m}_j + \text{-}0.031(0.055)\text{lslengthmin}_j + 0.037(0.042)\text{lslengthmax}_j + \\
& \text{-}0.005(0.006)\text{lspar1min}_j + 0.085(0.228)\text{lspar2rec\_1}_j + 0.102(0.040)\text{lsnightrl}_j + \text{-}0.100(0.369)\text{lsnightifsup\_1}_j + \\
& 0.263(0.490)\text{lspas1rec\_1}_j + 0.325(1.185)\text{lsiffam\_1}_j + \text{-}2.591(1.459)\text{lsifdrivered\_1}_j + \text{-}0.024(0.262)\text{lsage}_j + \text{-}0.060(0.210)\text{lsredu}_j + \\
& 2.547(1.083)\text{lsexit\_1}_j + 0.046(0.150)\text{lsdredrec\_1}_j + \text{-}2.078(1.237)\text{lsdredrec\_2}_j + \text{-}0.078(0.050)\text{isnightrllib}_j + \\
& 3.977(1.655)\text{isnightifwor\_1}_j + \text{-}0.892(0.537)\text{ispas1rec\_1}_j + \text{-}0.708(0.669)\text{isifsup\_1}_j + 0.925(0.533)\text{isiffam\_1}_j + \\
& \text{-}0.420(0.466)\text{isage}_j + 0.727(0.334)\text{isdred\_2}_j + \text{-}3.812(1.312)\text{isexit\_1}_j + e_{1ij}\text{logadjRR10se}_j
\end{aligned}$$

$$\beta_{0j} = 110.343(84.802) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.020(0.029) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 0.576(0.241) \end{bmatrix}$$

Deviance(MCMC) = 44.104(47 of 77 cases in use)

BAYES, GIBBS SAMPLING, RANDOM MODEL, 19 YEAR OLDS (50,000 iterations, burn-in period of 1000 iterations), RANDOM SEED 4

$$\text{logadjRR10}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\begin{aligned}
\text{logadjRR10}_{ij} = {} & \beta_{0j}\text{CONS} + \text{-}0.051(0.042)\text{year}_j + 0.175(0.214)\text{g}_j + \text{-}0.113(0.274)\text{m}_j + \text{-}0.032(0.055)\text{lslengthmin}_j + 0.037(0.042)\text{lslengthmax}_j + \\
& \text{-}0.005(0.006)\text{lspar1min}_j + 0.084(0.230)\text{lspar2rec\_1}_j + 0.101(0.040)\text{lsnightrl}_j + \text{-}0.101(0.372)\text{lsnightifsup\_1}_j + \\
& 0.260(0.494)\text{lspas1rec\_1}_j + 0.312(1.189)\text{lsiffam\_1}_j + \text{-}2.608(1.466)\text{lsifdrivered\_1}_j + \text{-}0.023(0.262)\text{lsage}_j + \text{-}0.059(0.210)\text{lsredu}_j + \\
& 2.547(1.076)\text{lsexit\_1}_j + 0.048(0.149)\text{lsdredrec\_1}_j + \text{-}2.084(1.240)\text{lsdredrec\_2}_j + \text{-}0.078(0.050)\text{isnightrllib}_j + \\
& 3.986(1.659)\text{isnightifwor\_1}_j + \text{-}0.892(0.534)\text{ispas1rec\_1}_j + \text{-}0.713(0.677)\text{isifsup\_1}_j + 0.925(0.531)\text{isiffam\_1}_j + \\
& \text{-}0.416(0.467)\text{isage}_j + 0.728(0.335)\text{isdred\_2}_j + \text{-}3.807(1.301)\text{isexit\_1}_j + e_{1ij}\text{logadjRR10se}_j
\end{aligned}$$

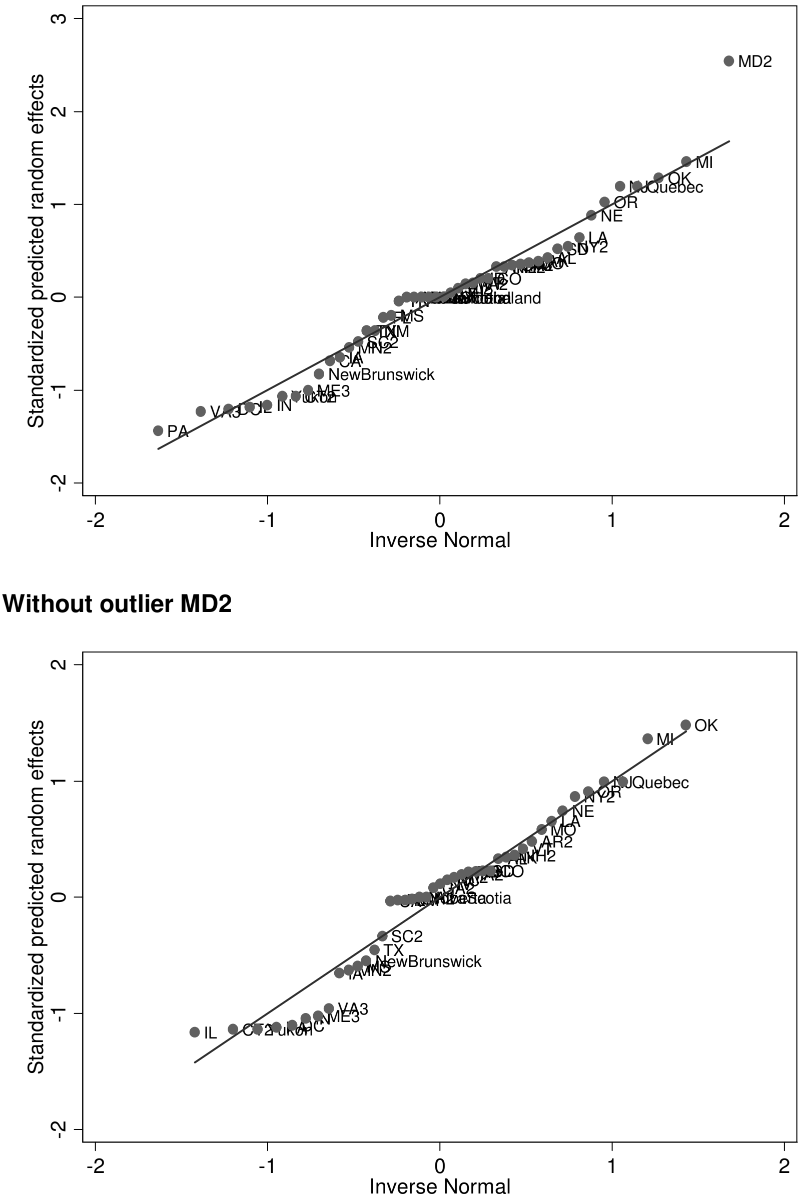
$$\beta_{0j} = 110.044(84.443) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.020(0.027) \end{bmatrix}$$

$$\begin{bmatrix} e_{1ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 0.582(0.251) \end{bmatrix}$$

Deviance(MCMC) = 44.452(47 of 77 cases in use)

Figure 6.7.2: Normal probability plot of standardized shrunken residuals



With outlier MD2

Without outlier MD2

Appendix 8: Number of GDL components per program

Table 6.8.1: Count of number of GDL components per GDL program

Legend

- juris – jurisdiction
- date – GDL implementation date (year)
- iihs – iihs rating: good(≥6 points) fair(4-5 points) marginal(2-3 points) poor(0-1 points)
- gdl – gdl stages: both learner and intermediate stage

Learner Stage

- lsc1 - Learner stage entry age (≥16)
- lsc2 - Learner stage length - minimum mandatory holding period (≥6 months)
- lsc3 - Learner stage length - minimum mandatory holding period (≤3 months)
- lsc4 - Minimum amount of supervised driving (≥30 hours)
- lsc5 - Minimum amount of supervised driving (≥50 hours)
- lsc6 - Mandatory hours of driving at night and/or inclement weather/before age 16 (yes/no)
- lsc7 - Learner stage night restriction (begins at 21:00 or 22:00)
- lsc8 - Learner stage night restriction (begins after 22:00)
- lsc9 - Learner stage passenger restriction (≤1 passengers allowed)
- lsc10 - Learner stage passenger restriction (≤2 passengers allowed)
- lsc11 - Learner stage exit test (yes/no)

Intermediate Stage

- isc1 - Intermediate stage entry age – difference between entry age for learner stage and intermediate stage is ≥12 months
- isc2 - Intermediate stage – most conservative night restriction (begins at 21:00 or 22:00)
- isc3 - Intermediate stage – most conservative night restriction (begins after 22:00)
- isc4 - Intermediate stage passenger restriction (≤1 passengers allowed)
- isc5 - Intermediate stage passenger restriction (≤2 passengers allowed)
- isc6 - Intermediate stage exit test (yes/no)

- total – total number of 'x's checked off

juris	date	iihs	gdl	lsc1	lsc2	lsc3	lsc4	lsc5	lsc6	lsc7	lsc8	lsc9	lsc10	lsc11	isc1	isc2	isc3	isc4	isc5	isc6	total
MD1	1992		x														x				2
MA1	1992		x	x													x				3
MN1	1992		x												x						2
NY1	1992		x	x												x	x				4
ON	1994	m		x	x	x					x			x						x	6
NS	1994	f	x	x	x	x						x	x	x			x			x	9
NB	1996	m	x	x	x	x						x	x	x							7
FL	1996	f	x		x	x	x	x	x						x		x				8
VA1	1996		x		x	x									x						4
KY1	1996	m		x	x	x					x										4
CT1	1997			x		x															2
MI	1997	f	x		x	x	x	x	x						x		x				8
QC	1997	m	x	x	x	x					x										5
GA1	1997		x		x	x									x		x				5
NC	1997	g	x		x	x					x				x	x	x	x	x		9
IL	1998	g	x		x	x	x	x	x						x	x	x	x	x		11
LA	1998	f	x		x	x									x		x				5
NH1	1998		x			x											x				3
CA	1998	g			x	x	x	x	x								x	x	x		8
OH1	1998		x																		1
SC1	1998		x			x										x	x				4
VA2	1998		x		x	x									x						4
ME1	1998		x			x	x		x						x						5
BC1	1998	m		x		x					x	x	x	x							6
MA2	1998	g	x	x	x	x	x				x						x	x	x		9
IN	1999	f	x			x									x		x	x	x		6
IA	1999	f	x		x	x			x						x		x				6
MN2	1999	g	x		x	x	x		x						x		x	x	x		9
NE	1999	g	x		x	x			x						x		x	x	x		8
NF	1999	f	x	x	x	x					x	x	x	x			x			x	10
OH2	1999	g	x		x	x	x	x	x								x				7
RI1	1999		x	x	x	x											x				5
SD	1999	m	x		x	x				x	x					x	x				7
OK	1999	g	x		x	x	x		x								x	x	x		8
CO	1999	g			x	x	x	x	x						x		x	x	x		9
DE	1999	g	x	x	x	x	x	x	x			x	x			x	x	x	x		13
MD2	1999	g	x		x	x	x	x	x								x	x	x		9
UT1	1999		x				x		x								x				4
AR1	1999				x	x									x						3
ND	1999	m			x	x					x										3
NH2	1999	f	x			x											x				3
PA	1999	g	x	x	x	x	x	x									x				7
PE	2000	f	x		x	x						x	x								5
NM	2000	m	x		x	x	x	x	x								x	x	x		9
OR	2000	g	x		x	x	x	x							x		x	x	x		9
DC	2000	g	x	x	x	x	x		x	x	x						x	x	x		11
MS	2000	m	x		x	x															3
VT	2000	f	x		x	x	x		x						x	x	x	x	x		9
WI	2000	g	x		x	x	x		x				x				x	x	x		9
ME2	2000		x			x	x		x						x			x	x		7
YU	2000	g	x		x	x	x	x	x		x	x	x	x	x		x				12
ID	2001	m	x		x	x	x	x	x							x	x	x	x		10
MO	2001	g	x		x	x	x		x						x		x	x	x		9
NJ	2001	g	x	x	x	x					x	x	x		x		x	x	x		11
WV	2001	f	x		x	x									x		x				5
NV	2001	g	x		x	x	x	x	x							x	x	x	x		10
TN	2001	g	x		x	x	x	x	x	x	x				x	x	x	x	x		13
UT2	2001		x				x		x								x	x	x		6
VA3	2001	g	x		x	x	x		x								x	x	x		8
WA	2001	g	x		x	x	x	x	x						x		x	x	x		10
GA2	2002	g	x		x	x	x		x						x			x	x		8
TX	2002	f	x		x	x									x		x	x	x		7
SC2	2002	m	x		x	x	x		x		x					x	x	x	x		10
AR2	2002	m			x	x									x			x	x		5
AL	2002	f			x	x									x		x				4
MB	2003	f	x	x	x	x								x					x		6
NH3	2003		x														x	x	x		4
AB	2003	f	x		x	x					x			x	x					x	7
UT3	2003	f	x				x		x								x	x	x		6
VA4	2003		x		x	x	x		x								x	x	x		8
RI2	2003	g	x	x	x	x	x	x	x								x				8
NY2	2003	g	x	x	x	x										x	x		x		7
ME3	2003	g	x		x	x	x		x						x		x	x	x		9
CT2	2003	g	x	x	x	x												x	x		6
BC2	2003		x	x	x	x					x	x	x	x					x	x	10
AK	2004	g			x	x	x		x						x		x	x	x		8
SA	2005		x		x	x								x						x	5
WY	2005	f	x				x	x	x						x		x	x	x		8