ABSTRACT

Title of dissertation: STATISTICAL METHODS FOR ADVANCED ECONOMETRIC MODELS WITH APPLICATIONS TO VEHICLE HOLDING AND SPEED QUANTILES DISTRIBUTION

Jean-Michel Tremblay, 2016

Dissertation directed by: Associate Professor Cinzia Cirillo
Department of Civil and Environmental Engineering

This dissertation proposes statistical methods to formulate, estimate and apply complex transportation models. Two main problems are part of the analyses conducted and presented in this dissertation.

The first method solves an econometric problem and is concerned with the joint estimation of models that contain both discrete and continuous decision variables. The use of ordered models along with a regression is proposed and their effectiveness is evaluated with respect to unordered models. Procedure to calculate and optimize the log-likelihood functions of both discrete-continuous approaches are derived, and difficulties associated with the estimation of unordered models explained. Numerical approximation methods based on the Genz algortithm are implemented in order to solve the multidimensional integral associated with the unordered modeling structure. The problems deriving from the lack of smoothness of the probit model around the maximum of the log-likelihood function, which makes the optimization

and the calculation of standard deviations very difficult, are carefully analyzed. A methodology to perform out-of-sample validation in the context of a joint model is proposed. Comprehensive numerical experiments have been conducted on both simulated and real data. In particular, the discrete-continuous models are estimated and applied to vehicle ownership and use models on data extracted from the 2009 National Household Travel Survey.

The second part of this work offers a comprehensive statistical analysis of free-flow speed distribution; the method is applied to data collected on a sample of roads in Italy. A linear mixed model that includes speed quantiles in its predictors is estimated. Results show that there is no road effect in the analysis of free-flow speeds, which is particularly important for model transferability. A very general framework to predict random effects with few observations and incomplete access to model covariates is formulated and applied to predict the distribution of free-flow speed quantiles. The speed distribution of most road sections is successfully predicted; jack-knife estimates are calculated and used to explain why some sections are poorly predicted.

Eventually, this work contributes to the literature in transportation modeling by proposing econometric model formulations for discrete-continuous variables, more efficient methods for the calculation of multivariate normal probabilities, and random effects models for free-flow speed estimation that takes into account the survey design. All methods are rigorously validated on both real and simulated data.

STATISTICAL METHODS FOR ADVANCED ECONOMETRIC MODELS WITH APPLICATIONS TO VEHICLE HOLDING AND SPEED QUANTILES DISTRIBUTION

by

Jean-Michel Tremblay

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Associate Professor Cinzia Cirillo, Chair/Advisor
Professor Paul M. Schonfeld
Professor Partha Lahiri
Associate Professor Fabian Bastin
Dr. Sevgi Erdoğan

Acknowledgments

This work would not have been possible without the help of many friends. In no particular order, I would like to thank my advisor Cinzia Cirillo for pushing me forward, my wife Lynna Nguyen and my parents for supporting me. Throughout my time as a student, Yangwen Liu found all possible bugs in my software and thus made it better. In the last years, Marco Bassani provided me data that allowed the completion of the second half of this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Transport modeling requires a multidisciplinary approach. In particular, advanced research in this field entails a deep mastery of the theoretical and mathematical bases of the methodologies used in connection with the transportation models, such as optimization methods, mathematical statistics, stochastic processes, dynamic models and so on. All this knowledge is needed to advance the state of the art and to enrich the toolkit available for eventual practical applications [CJM12]. This is especially true for demand modeling based on discrete choice analysis. This branch of the transportation discipline has seen enormous progress in recent years; examples of significant advances include estimation with simulation [Tra09]; dynamics in choice modeling [CXB16]; integrated models for discrete and continuous decision variables [Bha15]; optimization methods for log-likelihood estimation and Bayesian estimation [RAM05]. It is often a real challenge for students in transportation programs to understand all the techniques that have supported such changes and to be prepared to make significant contributions in transportation science.

In this context, I came to a transportation program with a Bachelor degree in Mathematics. Since then I have developed and implemented in ready to use software a number of statistical methods for different transportation applications

with special emphasis on discrete choice models. Although the models presented were motivated by a practical need, the methods that I'm proposing are general and can be applied in different context and on a variety of data. Actually, I'm more motivated by the properties of the models than by the results produced and their implication for policy analysis. Given the diversity of the problems in hands, in this dissertation I have organized my contributions by chapters.

Chapter 2 presents a general framework for continuous and discrete choice variables. Models of this type are required when multiple decisions of different natures are made simultaneously. Joint estimations have the advantage to be statistically efficient and to allow testing the effects of the covariates on all the model outcomes. The model is based on multinomial Probit and regression, where a full variance covariance matrix captures the correlation among the discrete and the continuous parts. This class of models does not have closed mathematical form for the choice probabilities and the underlying optimization problem is usually solved with simulations. Monte-Carlo (MC) based simulations presents a number of major challenges in this case and therefore in this thesis more sophisticated algorithms for the approximation of integrals that involve multivariate normal distributions are proposed. In this Chapter evidences on the advantages offered by the Genz based method for the estimation of discrete continuous problems are explained and numerical results offered.

In Chapter 3, the methods developed on the previous Chapter are applied to the problem of vehicle ownership, type/vintage and use. The analysis is performed on data extracted from the 2009 National Household Travel Survey [UDoT09] and on

an auxiliary dataset derived from the Consumer Report. The discrete-continuous Probit model is estimated both with MC simulations and with the Genz based algorithm. Moreover, a comparison across the unordered model and ordered discrete continuous Probit model is presented. Ordered structure are in general preferred to unordered Probit for the saving in computational costs deriving from the closed mathematical form of the choice probabilities. The results show that discrete-continuous unordered Probit are superior to ordered structures in terms of goodness of fit, but produce comparable results when applied to predict behavioral changes.

Ordered and unordered discrete-continuous models are validated using holdout samples in Chapter 4. Probit based discrete-continuous models have been recently proposed in the transportation literature and have been applied to a number of real case studies. However, it is not clear if the estimation of more complex model formulations produces better forecasts and if the conditioning on continuous variables helps to reproduce the market shares of the discrete alternatives. To this scope experiments on simulated data, for which the true coefficients and the true correlation elements are known to the analyst, have been generated and used to calculate market share on out of samples with both low and high correlation schemes. Results on the car ownership and use problem estimated on both 2001 and 2009 NHTS are presented and discussed.

In Chapter 5, a random effect model is proposed to estimate free-flow speed on two-lane rural highways. The data used for the analysis were collected in the Northwest of Italy using video cameras and a laser speed gun. The model structure adopted separates the estimate of the central tendency of speeds from the typical

deviations of individual speeds. Hence, in the model the same set of variables can be used to determine the mean value and the standard deviation of the speed distribution; the desired speed percentile is then calculated by considering the associated standard normal random variable (Z). Random effects (RE) were included in the model to account for the variability in time and space of the data that contains repeated measurements for the same road/section/direction and to remove the dependency between any estimation errors from individual observations. Fixed effect (FE) models are also calibrated for comparison purposes and the Bayesian information criterion (BIC) is used for variable selection and applied to both the FE and RE models.

The last Chapter of this dissertation is devoted to the transferability of Random Effect models in the context of free-flow speed estimation. This problem is twofold. First, it is necessary to estimate the RE for the road sections where the model is going to be transferred. Second, the speed quantiles used as predictors for the general model are not available in the new road sections under analysis. A method that correlates the RE of the new sections to the RE of the original model is proposed; results show that the method converges when a relatively low number of observations is collected for the out of sample sections. An analysis based on jackknife technique is also proposed to analyze cases where high errors are computed when transferring the model.

Chapter 2: Discrete-Continuous Probit model

2.1 Introduction

In many disciplines it is common to work with mixed data collected from measurements of multiple outcomes. These outcomes are often measured on different scales and their underlying variables can be discrete, ordinal or continuous [TPH13]. In the transportation field, a number of problems are characterized by several decision variables that are not from the same family. For example, in the problem of car ownership both discrete and continuous dependent variables should be taken into account for model formulation. The number of cars in the household can be modeled as a discrete variable or an ordinal variable; the vehicle type is a discrete variable; while the annual mileage or the mileage traveled with each vehicle in the household are continuous variables [LTC14]. In trip generation and activity scheduling, the number of trips is an ordinal variable, the type of activity is a discrete variable and the time allocated to each activity is a continuous variable [PB13]. Models for joint analysis of decision variable of the same type are based on classical multivariate statistics for continuous outcomes and on tree structures (or nested logit models) for sequential decision making over discrete alternatives [BBA01]. Methods for the analysis of variables of different types exist but their development and application

is new especially in transportation demand modeling. One simple approach consists of the analysis of each outcome separately, ignoring the correlation between the different outcomes. The separate analysis produces loss in the efficiency of the covariates and makes it impossible to test the effect of the covariates on all the outcomes to be modeled [TPH13]. Recently, models have been proposed to jointly handle mixed types of dependent variables, including multiple nominal variables, multiple ordinal variables, and multiple count variables, as well as multiple continuous variables [LTC14] [Bha15] [BAS+14].

In this Chapter, we propose multivariate econometric models for decision problems involving both discrete and continuous choices. The approach is based on multinomial probit and regression; it accommodates general covariance structures for the utilities of the discrete alternatives and for the linear regression of the continuous outcome. We also propose a Maximum Likelihood Estimation procedure based on the Genz [Gen92] algorithm to overcome the difficulties deriving from the optimization of probit based functions.

2.2 Literature review

2.2.1 Estimation methods for discrete-continuous Probit

Discrete-continuous probit probabilities do not have a closed-form expression and must be approximated numerically. Non-simulation and simulation procedures have been proposed for multinomial probit and more recently for integrated discrete-

continuous probit based models. Both procedures have proved to be effective in certain circumstances [Tra09]; however, there is no consensus on a general method that can be applied to all types of problems. Early applications use quadrature methods to approximate the integral by a weighted function of specially chosen evaluation points [Gew96]. Quadrature are suggested for low dimensional integrals (max. four or five), which limits the application to discrete problems involving a relatively high number of alternatives. Daganzo introduced another non-simulation procedure based on the Clark algorithm [DBS77]. This algorithm is based on the concept that the maximum of several normally distributed variables is itself approximately normally distributed. Unfortunately, the approximation can be highly inaccurate in some situations, and the degree of accuracy is difficult to assess in any given setting [HSD82].

More recently, simulation methods have been widely applied to approximate non closed form choice probabilities. Hajivassiliou et al. [HMR96] provides a comprehensive summary of the various simulators that have been proposed for probit models. Three simulators are of particular interest for probits of any form: acceptreject [MM81], smoothed acceptreject [McF89], and Geweke-Hajivassiliou-Keane (GHK) simulator [Gek89] [Ker91], [HM98]. Among them, the GHK simulator is the one that produced the best results in our experiments. However we have not been able to adapt it to the models presented in this dissertation. Bhat [Bha11] argued that the computational cost of maximum simulated log-likelihood (MSL) to ensure good asymptotic MSL estimator properties can be prohibitive and that MSL is practically infeasible as the number of dimensions of integration rises. He

introduced a maximum approximate composite marginal likelihood (MACML) estimation approach and applied it to MNP models using simple optimization software for likelihood estimation.

According to Bhat, the MACML estimation of MNP-based models involves only univariate and bivariate cumulative normal distribution function evaluations, regardless of the number of alternatives or the number of choice occasions per individual or the nature of social/spatial dependence structures. From application on both simulated and real data Bhat and his coauthors derive that substantial computational efficiency relative to the MSL inference approach can be obtained [Bha11].

2.3 Methodology

In the integrated model, the discrete problem concerns the forecast of the probability of a choice across a finite number of alternatives using a set of predictors. Suppose there are k alternatives, plus a base alternative U_0, the utility for each alternative consists of one deterministic part and one unobserved part (error term):

$$\begin{aligned} U_0 &= & \epsilon_0 \\ U_1 &= X_1^T\beta_1 + & \epsilon_1 \\ &\dots & \\ U_k &= X_k^T\beta_k + & \epsilon_k \end{aligned}$$

where, U_j is the utility associated to each alternative j; X_j are the explanatory variables associated with the decision maker and the alternative. β_j is the corresponding parameter to be estimated. We say that $X_j^T\beta_j = V_j$ is the deterministic part of the utility.

It is common in discrete-choice modeling to refer to both U_j and V_j as the utility. In this document we avoid this confusion and refer to V_j as the *deterministic part of the utility*.

Another source of confusion is that some of the β_j parameters may share elements together in order to increase modeling flexibility. For example if we want to model two transportation mode, say driving to work or take the bus, and we want to incorporate their price in the utility functions, both predictors could have the same coefficient. This issue is resolved by building a single extended design matrix that generates all utilities for all observations. The details regarding the construction of this design matrix are beyond the scope of this document and can be found at jm.dynddns.us/files/specUtility.pdf.

2.3.1 Choice probability

In an unordered structure, the household is assumed to be rational and to choose the alternative that maximizes its utility. In this case, we adopt multinomial probit model for the the discrete decisions and therefore the error terms follow a multivariate normal distribution with full, unrestricted covariance matrix. The choice probability can be expressed as follow:

$$P\left(Y = y|X, \beta, \Sigma\right) = \int_{R^{k+1}} I\left(V_y + \epsilon_y > V_j + \epsilon_j \qquad \forall j \neq y\right) \phi(\epsilon) d\epsilon \tag{2.1}$$

where:

$$
\begin{aligned}
\mathrm{X} &= (\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{k}) \\
\beta &= (\beta_1, \ldots, \beta_\mathrm{k}) \\
\epsilon &= (\epsilon_0, \ldots, \epsilon_\mathrm{k}) \\
\Sigma &= \text{covariance of } \epsilon, \text{ upon which } \phi \text{ depends}
\end{aligned}
$$

The indicator function $\mathrm{I}(\cdot)$ ensures that the observed choice is indeed the one with the biggest utility. The subscript y indicates the predictors and coefficients of the chosen alternative and the subscript j indicate the other alternatives. The integral correspond to the expectation of the event "U_y is the biggest utility", which is the choice probability.

2.3.2 Difference of error terms

Since only differences in utility matter, the choice probability can be equivalently expressed as a k-dimensional integral over the differences between the errors. Suppose we calculate the differences with respect to alternative y, the alternative for which we are calculating the probability. Then we define for all indexes $\mathrm{j} \neq \mathrm{y}$:

$$
\begin{aligned}
\tilde{\epsilon}_{\mathrm{j–y}} &= \epsilon_\mathrm{j} - \epsilon_\mathrm{y} \\
\tilde{\epsilon}_{\mathrm{–y}} &= (\tilde{\epsilon}_{0\mathrm{y}}, \ldots, \tilde{\epsilon}_{\mathrm{ky}}) \\
\tilde{\mathrm{V}}_{\mathrm{y–j}} &= \mathrm{V}_\mathrm{y} - \mathrm{V}_\mathrm{j} \\
\tilde{\mathrm{V}}_{\mathrm{y–}} &= (\tilde{\mathrm{V}}_{\mathrm{y–}0}, \ldots, \tilde{\mathrm{V}}_{\mathrm{y–k}})
\end{aligned}
$$

The subscript in $\mathrm{q}_{\mathrm{a–b}}$ should read "a minus b" because to compute the result, we calculate q_a minus q_b. The probit is normalized using the procedure proposed by Train [Tra09] to ensure that all parameters are identified. For more details on

the normalization in the context of discrete-continuous models we refer to Liu et al. [LTC14].

A consequence of this normalization argument is that the matrix Σ is over-parametrized for the probit model since, as we shall see, the covariance Σ_{-y} of $\tilde{\epsilon}_{-y}$ is sufficient to express the choice probability. With a few transformations we obtain:

$$\begin{aligned} & V_y + \epsilon_y > V_j + \epsilon_j \forall j \neq y \\ \Leftrightarrow & V_y - V_j > \epsilon_j - \epsilon_y \forall j \neq y \\ \Leftrightarrow & \tilde{V}_{y-} > \tilde{\epsilon}_{-y} \end{aligned} \tag{2.2}$$

This allows to express explicitly the choice probability in terms of $\tilde{V}_{y-}$,$\tilde{\epsilon}_{-y}$ and the multivariate normal cumulative distribution function, as seen in equation 2.3

$$\begin{aligned} P\left(Y = y | X, \beta, \Sigma_{-y}\right) &= \int_{R^{k+1}} I\left(V_y + \epsilon_y > V_j + \epsilon_j \qquad \forall j \neq y\right) \phi(\epsilon) d\epsilon \\ &= \int_{-\infty}^{\tilde{V}_{y-0}} \dots \int_{-\infty}^{\tilde{V}_{y-k}} 1 \cdot \phi(\tilde{\epsilon}_{-y}) d\tilde{\epsilon}_{-y} \\ &= \Phi(\tilde{V}_{y-}) \end{aligned} \tag{2.3}$$

2.3.3 Error terms reparametrization

We have seen how to compute $P(Y = y)$ using $\tilde{V}_{y-}$ and Σ_{-y}, however this is not enough to compute the choice probability of another alternative y'. The key to compute the choice probability of any alternative y' is to set up a matrix M that transforms $\tilde{\epsilon}_{-y}$ into $\tilde{\epsilon}_{-y'}$. For illustration purposes we will work in the case where

k = 0 and we start with Σ_{-0}. In this context we can easily compute P(Y = 0). In order to compute P(Y = 1, 2, 3) we define the following matrices:

$$M_{0,1} = \begin{bmatrix} -1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \tag{2.4}$$

$$M_{0,2} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \tag{2.5}$$

$$M_{0,3} = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \tag{2.6}$$

The purpose of $M_{y,y'}$ is that it can change error terms taken with respect to the y^{th} alternative into error terms taken with respect to the y'^{th} alternative:

$$M_{0,1}\tilde{\epsilon}_{-0} = \begin{bmatrix} -1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \epsilon_1 - \epsilon_0 \\ \epsilon_2 - \epsilon_0 \\ \epsilon_3 - \epsilon_0 \end{bmatrix} = \begin{bmatrix} \epsilon_0 - \epsilon_1 \\ \epsilon_2 - \epsilon_1 \\ \epsilon_3 - \epsilon_1 \end{bmatrix} \tag{2.7}$$

And similarly for the other alternative. The matrices $M_{y,y'}$ define affine transformation of a random vector. We have that, generally, if $a \sim N(\mu, \Sigma)$ then:

$$b = Ma + c \sim N(M\mu + c, M\Sigma M^T) \tag{2.8}$$

We use this result directly to compute any choice probability. Recall that all that was required to compute P(Y = y) was the covariance matrix Σ_{-y}. We apply the affine transformation to compute $\Sigma_{-y'}$ for all y' and we can compute all choice probabilities.

In our case the original base for error term differences was 0, and we assume that $\tilde{\epsilon}_{-0}$ follow a normal with mean zero and covariance Σ_{-0}. It follows that $\tilde{\epsilon}_{-j}$ has mean zero and covariance $\Sigma_{-j} = M_{0,j}\Sigma_{-0}M_{0,j}^T$. This is enough to parametrize the probit model using only Σ_{-0} for the variance of the error terms.

However *any* Σ_{-j} could be used. It may be of modeling interest to use a specific alternative as the basis for error differences, primarily to ease the interpretation, even if the choice of the alternative used for a basis does not affect the model itself. The only challenge will be to construct the transformation matrices. In general we can build a matrix $M_{i,j}$ like this:

- start with a $(k-1)$-dimensional identity matrix
- insert a row vector of 0 after row $i - I(i > j)$
- insert a column vector of –1 after column $j - I(j > i)$

For example for $k = 5$, $M_{2,4}$ is built like this:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix} \tag{2.9}$$

We easily verify that:

$$M_{2,4}\tilde{\epsilon}_{-2} = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \epsilon_0 - \epsilon_2 \\ \epsilon_1 - \epsilon_2 \\ \epsilon_3 - \epsilon_2 \\ \epsilon_4 - \epsilon_2 \\ \epsilon_5 - \epsilon_2 \end{bmatrix} = \begin{bmatrix} \epsilon_0 - \epsilon_4 \\ \epsilon_1 - \epsilon_4 \\ \epsilon_2 - \epsilon_4 \\ \epsilon_3 - \epsilon_4 \\ \epsilon_5 - \epsilon_4 \end{bmatrix} = \tilde{\epsilon}_{-4} \tag{2.10}$$

2.4 Regression

Regression is adopted to model the continuous part of the model. In a regression, the dependent variable Y_{reg} is assumed to be a linear combination of a vector of predictors X_{reg} plus some error term:

$$Y_r = X_r\beta_r + \epsilon_r, \qquad \epsilon_r \sim N(0, \sigma_r^2) \tag{2.11}$$

Given β_{r} , $\mathrm{X_r}$ and σ_{r}^2 , the likelihood of observing $\mathrm{Y_r}$ is given, by design, by the normal density function ϕ:

$$\mathrm{P}\left(\mathrm{Y_r}|\beta_{\mathrm{r}}, \mathrm{X_r}, \sigma_{\mathrm{r}}\right) = \phi\left(\mathrm{Y_r}|\mathrm{X_r^T}\beta_{\mathrm{r}}, \sigma_{\mathrm{r}}^2\right) \tag{2.12}$$

In order to jointly capture the correlation between the discrete and continuous parts, we allow the error term of the regression to be correlated with the error term differences in the probit. Therefore, the specifications of the observable part of the utilities and of the regression remain the same, but the error terms follow an "incremental" normal distribution:

$$\left(\tilde{\epsilon}_{0\text{–}\mathrm{y}}, \ldots, \tilde{\epsilon}_{\mathrm{k}\text{–}\mathrm{y}}, \epsilon_{\mathrm{r}}\right) \sim \mathrm{N}\left(0, \Sigma_{\mathrm{k}+1}\right) \tag{2.13}$$

At this point we assume that we are always working with differences with respect to alternative y = 0. Therefore we will be estimating $\Sigma_{\text{–}0}$, and $\Sigma_{\mathrm{k}+1}$ will include it as a submatrix:

$$\left(\tilde{\epsilon}_{\text{–}0}, \ldots, \tilde{\epsilon}_{\mathrm{k}\text{–}0}, \epsilon_{\mathrm{r}}\right) \sim \mathrm{N}\left(0, \Sigma_{\mathrm{k}+1}\right) \tag{2.14}$$

A more intuitive way to partition the error term is given by:

$$\begin{bmatrix} \tilde{\epsilon}_{\text{–}0} \\ \epsilon_{\mathrm{r}} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{\text{–}0} & \Sigma_{\text{–}0,\mathrm{r}} \\ \Sigma_{\mathrm{r},\text{–}0} & \sigma_{\mathrm{r}}^2 \end{bmatrix}\right) \tag{2.15}$$

This is equivalent to say that $\tilde{\epsilon}_{\text{–}\mathrm{y}}$ and ϵ_{reg} are two normally distributed stan-

dalone entities who are also are jointly normally distributed with covariance $\Sigma_{r,-0}$. The probability of observing Y and Y_r is the product of the probability of observing Y_r ($P(Y_r)$) and the probability of observing Y given Y_r ($P(Y|Y_r)$). It is important to point out that $P(Y_r)$ is not a probability but a density, however these two concepts are not treated differently when performing statistical inference using maximum-likelihood estimation. Densities refer to the probability of observing a random variable in an Infinitesimal interval. If f is the continuous density of the random variable X then:

$$P(x - \frac{\delta}{2} < X < x + \frac{\delta}{2}) \approx \delta f(x) \tag{2.16}$$

We also remark that is is generally true that:

$$P(A, B) = P(A)P(B|A) = P(B)P(A|B) \tag{2.17}$$

The probability $P\left(Y_{reg}\right)$ is simply given by the normal density as previously stated. We are left with the challenge of computing the probability of the conditional probit. We have the following general result for jointly normal vectors:

$$\begin{bmatrix} A \\ B \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right) \tag{2.18}$$

We have that if we observe B = b, then A|B = b is normally distributed with the

following parameters:

$$\begin{aligned}\mu_{\mathrm{a|b}} &= \mu_{\mathrm{a}} + \Sigma_{\mathrm{ab}}\Sigma_{\mathrm{bb}}^{-1}(\mathrm{b} - \mu_{\mathrm{b}}) \\ \Sigma_{\mathrm{a|b}} &= \Sigma_{\mathrm{aa}} - \Sigma_{\mathrm{ab}}\Sigma_{\mathrm{bb}}^{-1}\Sigma_{\mathrm{ba}}\end{aligned} \tag{2.19}$$

We apply this directly by substituting the estimated residual from the regression ϵ_{r} for b:

$$\tilde{\epsilon}_{\mathrm{y}}|\epsilon_{\mathrm{reg}} \sim \mathrm{N}\left(\frac{\Sigma_{\mathrm{y,reg}}}{{\sigma_{\mathrm{reg}}}^2}\epsilon_{\mathrm{reg}}, \Sigma_{\mathrm{y}} - \Sigma_{\mathrm{y,reg}}\sigma_{\mathrm{reg}}^{-2}\Sigma_{\mathrm{reg,y}}\right) \tag{2.20}$$

2.5 Log-likelihood function

We can write the log-likelihood function of the discrete-continuous model as:

$$\mathrm{LL}(\beta, \beta_{\mathrm{r}}, \Sigma_{\mathrm{k+1}}|\mathrm{Y}, \mathrm{Y_r}, \mathrm{X}, \mathrm{X_r}) = \sum_{\mathrm{i}} \log(\phi(\mathrm{Y_{r,i}}|\mathrm{X_{r,i}^T}\beta_{\mathrm{r}}, \sigma_{\mathrm{r}}^2)\Phi(\tilde{\mathrm{V}}_{\mathrm{y_i-,i}}|\Sigma_{\mathrm{-y_i}}) \tag{2.21}$$

where:

$\mathrm{Y_{r,i}}$: $\mathrm{Y_r}$ for the i^{th} observation

$\mathrm{X_{r,i}}$: $\mathrm{X_r}$

$\tilde{\mathrm{V}}_{\mathrm{y_i-,i}}$: $\tilde{\mathrm{V}}_{\mathrm{y-}}$ for the i^{th} observation (depends on the choice $\mathrm{y_i}$)

$\Sigma_{\mathrm{-y_i}}$: $\Sigma_{\mathrm{-y}}$ for the i^{th} observation (depends on the choice $\mathrm{y_i}$)

The estimation of Φ is discussed in section 2.6. At this point we have achieved to reduce the problem to the computation of the multivariate normal CDF function.

2.6 Estimation with numerical computation

In this Section we introduce numerical methods for the computation of the multivariate normal CDF. In particular, we adopt the transformation proposed by Genz [Gen92], which simplifies the problem and transforms [TPH13] into a mathematical form that allows efficient calculation using standard numerical multiple integration algorithms. The use of numerical computation is expected to be faster than Monte Carlo simulation, to produce more precise estimation of the log-likelihood function and a more stable estimation of the Hessian, which is needed to calculate the information matrix.

2.6.1 Genz's algorithm

2.6.1.1 Transformations

Genz suggests an algorithm for the calculation of multivariate normal probabilities:

$$\mathrm{X} \sim \mathrm{N}(0, \Sigma) \tag{2.22}$$

The algorithm estimates $\mathrm{P}(\mathrm{a} < \mathrm{X} < \mathrm{b})$, which can be used to compute the cumulative distribution if we set $\mathrm{a} = 0$. We start with the definition of $\mathrm{P}(\mathrm{a} < \mathrm{X} < \mathrm{b})$:

$$\mathrm{F}(\mathrm{a}, \mathrm{b}) = \mathrm{P}(\mathrm{a} < \mathrm{N} < \mathrm{b}) = \int_{\mathrm{a}_1}^{\mathrm{b}_1} \dots \int_{\mathrm{a}_\mathrm{k}}^{\mathrm{b}_\mathrm{k}} \frac{1}{\sqrt{|\Sigma|(2\pi)^\mathrm{k}}} \mathrm{e}^{\frac{-1}{2}\theta^\mathrm{T}\Sigma^{-1}\theta} \mathrm{d}\theta \tag{2.23}$$

The first transformation that we apply is $\mathrm{Cy} = \theta$, where C is the Cholesky

factor of Σ such that $CC^T = \Sigma$. To perform the substitution we need the Jacobian matrix J of the transformation. The (i, j) element of J is given by derivative the i^{th} element of Cy with respect to y_j:

$$J_{i,j} = \frac{\partial(Cy)_i}{\partial y_j} = \frac{\partial(C_{i,1}y_1 + C_{i,2}y_2 + \cdots + C_{i,k}y_k)}{\partial y_j} = C_{i,j} \tag{2.24}$$

This gives us $J = C$. We also observe that:

$$\sqrt{|\Sigma|} = \sqrt{|CC^T|} = \sqrt{|C||C^T|} = \sqrt{|C|^2} = |C| \tag{2.25}$$

since the determinant of a diagonal matrix is the product of its elements and both C and C^T are diagonal matrices with the same diagonal elements. We also observe that the transformation allows us to factorize the integrand:

$$\begin{aligned} \theta^T\Sigma^{-1}\theta = (Cy)^T(CC^T)^{-1}(Cy) &= (y^TC^T)(C^{-T}C^{-1})(Cy) \\ &= y^T(C^TC^{-T})(C^{-1}C)y = y^Ty \end{aligned} \tag{2.26}$$

such that

$$e^{\frac{-1}{2}\theta^T\Sigma^{-1}\theta} = e^{\frac{-1}{2}y^Ty} = e^{-\frac{1}{2}y_1^2}e^{-\frac{1}{2}y_2^2}\ldots e^{-\frac{1}{2}y_k^2} \tag{2.27}$$

The integration bounds are also transformed but become more complicated. We originally had $a < \theta < b$. For the i^{th} bound we have:

$$a_i < \theta_i = (Cy)_i < b_i$$

$$a_i < \sum_{j=1}^{k} C_{ij} y_j < b_i$$

$$a_i < \sum_{j=1}^{i-1} C_{ij} y_j + C_{ii} y_i < b_i$$

$$(a_i - \sum_{j=1}^{i-1} C_{ij} y_j)/C_{ii} = a_i' < y_i < (b_i - \sum_{j=1}^{i-1} C_{ij} y_j)/C_{ii} = b_i'$$

$$a_i' < y_i < b_i'$$

We can change the indexes of the summation because the elements above the diagonal in C are 0. Also the bound in the i^{th} integral depends on $y_1, \ldots, y_{i-1}$ even if the notation does not show it explicitly. After the first variable substitution the integral becomes:

$$\begin{aligned} F(a, b) &= \int_{a_1'}^{b_1'} \cdots \int_{a_k'}^{b_k'} \frac{1}{|C|\sqrt{(2\pi)^k}} e^{\frac{-1}{2} y^T y} |C| dy \\ &= \frac{1}{\sqrt{(2\pi)^k}} \int_{a_1'}^{b_1'} e^{-\frac{1}{2} y_1^2} \int_{a_2'}^{b_2'} e^{-\frac{1}{2} y_2^2} \cdots \int_{a_k'}^{b_k'} e^{-\frac{1}{2} y_k^2} dy \\ &= \int_{a_1'}^{b_1'} \phi(y_1) \int_{a_2'}^{b_2'} \phi(y_2) \cdots \int_{a_k'}^{b_k'} \phi(y_k) dy \end{aligned} \tag{2.28}$$

where ϕ is the standard normal density function.

The second transformation requires the following result about the derivative of a function inverse. If f an invertible function with inverse f^{-1}, and that the derivative of f in the neighborhood of x is non-zero, then f^{-1} is guaranteed to be differentiable in the neighborhood of x and its derivative is given by:

$$\left(f^{-1}\right)'(x) = \frac{1}{f'(f^{-1}(a))} \tag{2.29}$$

We use this observation to suggest the following *independent* substitutions:

$$y_i = \Phi^{-1}(z_i) \tag{2.30}$$

Then the derivative of the variable change is:

$$(\Phi^{-1})'(z_i) = \frac{1}{\phi(\Phi^{-1}(z_i))} \tag{2.31}$$

We can say that the determinant of the Jacobian of the substitution is given by the product of these derivatives since the substitution is performed independently:

$$|J| = \prod_{i=1}^{k} \frac{1}{\phi(\Phi^{-1}(z_i))} \tag{2.32}$$

We obtain the following integral to solve:

$$\begin{aligned} F(a,b) &= \int_{a_1''}^{b_1''} \phi(\Phi^{-1}(z_1)) \int_{a_2''}^{b_2''} \phi(\Phi^{-1}(z_2)) \dots \int_{a_k''}^{b_k''} \phi(\Phi^{-1}(z_k)) \prod_{i=1}^{k} \frac{1}{\phi(\Phi^{-1}(z_i))} dz \\ &= \int_{a_1''}^{b_1''} \phi(\Phi^{-1}(z_1)) \frac{1}{\phi(\Phi^{-1}(z_1))} \dots \int_{a_k''}^{b_k''} \phi(\Phi^{-1}(z_k)) \frac{1}{\phi(\Phi^{-1}(z_k))} dz \\ &= \int_{a_1''}^{b_1''} \dots \int_{a_k''}^{b_k''} dz \end{aligned} \tag{2.33}$$

with:

$$a_i'' = \Phi((a_i - \sum_{j=1}^{i-1} C_{ij} \Phi^{-1}(z_j))/C_{ii}) \tag{2.34}$$

$$b_i'' = \Phi((b_i - \sum_{j=1}^{i-1} C_{ij}\Phi^{-1}(z_j))/C_{ii}) \tag{2.35}$$

The third transformation turns the integral into a so-called *constant limit form*. The transformation in essentially a linear map between the bounds of the integration space into the the $[0,1]^k$ hypercube:

$$z_i = a_i'' + w_i(b_i'' - a_i'') \tag{2.36}$$

The derivative of the substitution is given by:

$$\frac{\partial(a_i'' + w_i(b_i'' - a_i''))}{\partial w_i} = b_i'' - a_i'' \tag{2.37}$$

Therefore, the final substitution is given by:

$$F(a,b) = (b_1''' - a_1''') \int_0^1 (b_2''' - a_2''') \dots \int_0^1 (b_k''' - a_k''') \int_0^1 dw \tag{2.38}$$

At this point we rename a_i''' into d_i and b_i''' into e_i, and these quantities are given by:

$$d_i = \Phi((a_i - \sum_{j=1}^{i-1} C_{ij}\Phi^{-1}(d_j + w_j(e_j - d_j))/C_{ii}) \tag{2.39}$$

$$e_i = \Phi((b_i - \sum_{j=1}^{i-1} C_{ij}\Phi^{-1}(d_j + w_j(e_j - d_j))/C_{ii}) \tag{2.40}$$

The integral we will apply our algorithm to is:

$$F(a,b) = (e_1 - d_1)\int_0^1 (e_2 - d_2)\ldots\int_0^1 (e_k - d_k)\int_0^1 dw \tag{2.41}$$

2.6.1.2 The algorithm

We estimate $F(a,b)$ using Genz's algorithm [Gen92]:

Data: $\Sigma, a, b, \epsilon, \alpha$ and N_{max}
Result: $\hat{P}(a < N(0,\Sigma) < b)$
set $N \leftarrow 0$;
set Intsum $\leftarrow 0$;
set Varsum $\leftarrow 0$;
compute C the Cholesky factor of Σ;
do
  generate $w_1, \ldots, w_{k-1}$ from a uniform distribution;
  $d_i \leftarrow \Phi((a_i - \sum_{j=1}^{i-1} C_{ij}\Phi^{-1}(d_j + w_j(e_j - d_j))/C_{ii}) \quad i = 1, ..., n;$
  $e_i \leftarrow \Phi((b_i - \sum_{j=1}^{i-1} C_{ij}\Phi^{-1}(d_j + w_j(e_j - d_j))/C_{ii}) \quad i = 1, ..., n;$
  $f_i \leftarrow (e_i - d_i)f_{i-1} \quad i = 1, ..., n;$
  $N \leftarrow N + 1;$
  $\delta \leftarrow (f_k - \text{Intsum})/N;$
  Intsum $\leftarrow$ Intsum $+ \delta$;
  Varsum $\leftarrow \frac{N-2}{N}$Varsum $+ \delta^2$;
  Error $\leftarrow \alpha\sqrt{\text{Varsum}}$;
while *Error* $< \epsilon$ *or* $N = N_{max}$;

Algorithm 1: Genz's algorithm

2.6.2 Acceptance-rejection algorithm

The competing alternative is to use the traditional acceptance-rejection algorithm to compute normal probabilities. We take draws from the normal distribution and estimate the normal probability by the mean of draws that are below the upper

bound of the integral:

$$\hat{P}(N(0,\Sigma) < x) = \sum_{b=1}^{B} I(w_i < x) \tag{2.42}$$

where w_i is a draw from the $N(0,\Sigma)$ distribution.

2.6.3 Comparison of acceptance-rejection and Genz

Figure 2.1: Comparison of Genz and simulation for one observation

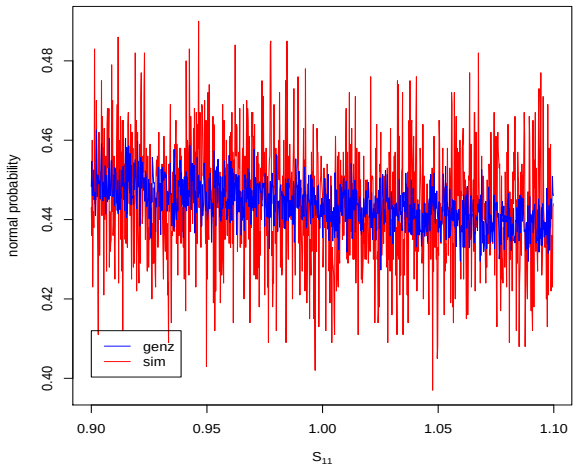
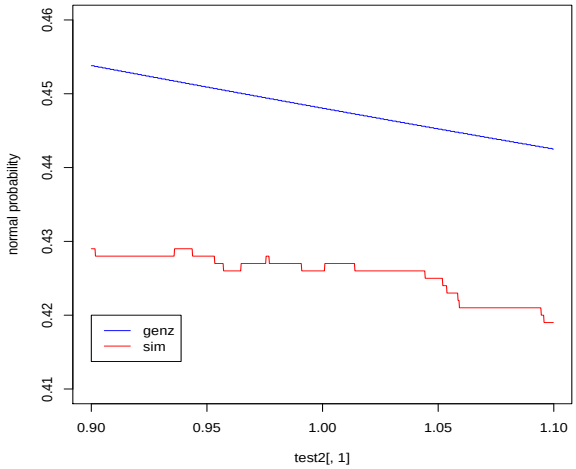


Figure 2.1 shows the difference between Genz and simulation estimation for a single probability. We can see that both methods are centered around the same value and that Genz has much less volatility. These values have been computed with the same number of random draws, but the draws are different for each point on the figure.

Figure 2.2 is the same as figure 2.1, but with the same random stream for all points. Genz's estimation is smooth but simulation estimation is choppy. The

Figure 2.2: Comparison of Genz and simulation with Common Random Stream



distance between the two curves is due to simulation bias. Indeed, using one common stream reduces the noise in the estimation but also sets its vertical location on the figure.

2.6.3.1 Example

We generate the following sample to illustrate the differences between the two methods.

$$\begin{aligned} U_1 &= X_1\beta_1 + X_4\beta_4 + \epsilon_1 \\ U_2 &= X_2\beta_1 + X_5\beta_5 + \epsilon_2 \\ U_3 &= X_3\beta_1 + X_6\beta_6 + \epsilon_3 \end{aligned}$$

where:

- $(\beta_1, \beta_4, \beta_5, \beta_6) = (-1, 1, 1, 1)$
- $\text{chol}(\text{cov}(\epsilon_2 - \epsilon_1, \epsilon_3 - \epsilon_1)) = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.866 \end{pmatrix}$

- the predictors X_i are generated with a standard normal
- we have $n_{obs} = 500$ observations
- for simulations, we use $B = 500$ simulations

To analyze the smoothness of the methods, we take five points where we are inspecting the log-likelihood functions. We take the maximum as found by Genz's algorithm, the real value of the parameters, zero for the coefficients and identity for the covariance matrix, and two "random" points. The value of the points can be seen in table 2.1.

Table 2.1: Five points to inspect log-likelihood functions

parameter	max (p_1)	real (p_2)	zero (p_3)	random 1 (p_4)	random 2 (p_5)
β_1	-0.994	-1.000	0	0.5	-2
β_4	1.065	1.000	0	0.5	-2
β_5	0.924	1.000	0	-2.0	-2
β_6	0.872	1.000	0	-2.0	2
$L_{2,1}$	0.357	0.500	0	1.0	2
$L_{2,2}$	0.691	0.866	1	1.0	2

The computation of gradients is essential to the numerical estimation of parameter estimates. To estimate them, we take these five points and compute the centered differences derivatives for values of δ between 0.1 and 0.001, for both method. The results can be seen in table 2.2,2.3,2.4,2.5 and 2.6. The are several things to notice.

First, Genz's method is relatively robust to a large value of δ for all five points. This looks promising because we would like to use such a large value with simulation and hope to get meaningful derivatives.

Second, simulation produces very unreliable gradients in the vicinity of the maximum likelihood, as seen in tables 2.2 and 2.3. This is troublesome because it

is the area where good precision would be most useful, in particular to estimate standard deviations.

Third, gradients are surprisingly stable at point p_3, p_4 and to some extent p_5 as seen in tables 2.4, 2.5 and 2.6. We believe that this is what explains why the solvers typically find estimates close to the max, but that lack accuracy: they have good enough derivatives to reach the vicinity of the maximum and stall as soon as they reach there. The exact behavior is sample dependent.

Table 2.2: Gradient for p_1

parameter	method	δ				
		0.1	0.05	0.01	0.005	0.001
B_1	sim	-3.595	-3.216	-14.097	-19.542	-4.94
B_4	sim	-6.342	-9.648	-17.259	-4.482	19.41
B_5	sim	-3.123	-4.764	-7.439	-11.612	-27.403
B_6	sim	-2.133	2.799	3.235	-8.325	32.072
$L_{2,1}$	sim	0.414	3.824	10.725	11.889	-23.02
$L_{2,2}$	sim	4.846	11.601	29.215	-5.219	-32.202
B_1	Genz	1.275	1.617	1.727	1.73	1.731
B_4	Genz	-0.546	-0.653	-0.687	-0.688	-0.688
B_5	Genz	-0.615	-0.682	-0.703	-0.703	-0.704
B_6	Genz	0.624	0.539	0.512	0.511	0.511
$L_{2,1}$	Genz	-1.903	-1.842	-1.823	-1.822	-1.822
$L_{2,2}$	Genz	1.087	0.339	0.103	0.096	0.093

It is convenient to look at the shape of the log-likelihood with respect to each parameter.

Figures 2.3 and 2.4 illustrates why it is difficult to optimize the simulated log-likelihood functions at the max. The subplot for L_{22} is the most striking as it looks like a step function.

Figure 2.5 shows why it is not too difficult for the solver to iterate around zero since the function is mostly smooth and close to the better approximation produced

Table 2.3: Gradient for p_2

parameter	method	δ				
		0.1	0.05	0.01	0.005	0.001
B_1	sim	-14.801	-12.118	-8.175	-15.88	16.878
B_4	sim	19.051	21.538	29.854	27.422	18.472
B_5	sim	-8.753	-8.941	-12.339	-15.623	-19.969
B_6	sim	-11.592	-12.07	-9.55	-4.245	-8.963
$L_{2,1}$	sim	-13.517	-17.82	-8.545	-6.31	-3.909
$L_{2,2}$	sim	-17.746	-19.455	-22.365	-20.646	-18.965
B_1	Genz	-14.202	-13.883	-13.781	-13.778	-13.777
B_4	Genz	17.86	17.765	17.735	17.734	17.733
B_5	Genz	-7.557	-7.615	-7.633	-7.634	-7.634
B_6	Genz	-13.956	-14.014	-14.033	-14.033	-14.034
$L_{2,1}$	Genz	-11.9	-11.923	-11.93	-11.93	-11.93
$L_{2,2}$	Genz	-12.697	-13.12	-13.254	-13.259	-13.26

by Genz.

Figures 2.6 and 2.7 show mostly discrepancies between the two estimates when we are far from optimal values. These differences are caused by very small default probability values used when the acceptance-rejection algorithm produces zero acceptances.

We generate 375 samples and estimate the coefficients with both methods. It is difficult to assess convergence for each and every sample but it is useful to look at the distribution of the estimates. Table 2.7 shows the mean and standard deviation of all six parameter estimates, for both methods. Genz outperforms simulation in both bias and variance.

Figure 2.8 plots the simulation estimates against the Genz estimates for all six parameters. The horizontal and vertical lines correspond to the true parameters, such that we expect the scatter plots to be distributed equally on both sides of the lines. This is absolutely not the case for the simulation estimates, as they are

Table 2.4: Gradient for p_3

parameter	method	δ				
		0.1	0.05	0.01	0.005	0.001
B_1	sim	-482.603	-478.965	-484.43	-497.348	-508.8
B_4	sim	235.703	237.392	240.335	230.301	223.882
B_5	sim	116.676	122.454	130.476	123.521	152.201
B_6	sim	124.229	124.229	129.231	145.062	153.882
$L_{2,1}$	sim	41.057	42.894	33.322	40.799	26.654
$L_{2,2}$	sim	-0.203	0.331	-0.407	3.69	-2.594
B_1	Genz	-477.343	-476.17	-475.794	-475.782	-475.779
B_4	Genz	232.18	231.907	231.82	231.817	231.816
B_5	Genz	117.246	117.13	117.093	117.092	117.091
B_6	Genz	118.392	118.288	118.254	118.253	118.253
$L_{2,1}$	Genz	37.357	37.351	37.348	37.348	37.348
$L_{2,2}$	Genz	-0.138	-0.194	-0.211	-0.212	-0.212

systematically located on only one side of the horizontal line. Genz estimates, as seen in table 2.7, are centered around the vertical line.

The last thing that we are checking is the computation of standard deviations, using the Hessian matrix estimated at the maximum. Tables 2.8 and 2.9 report estimated standard deviations for ten samples. Not surprisingly, Genz produces valid estimates for the standard deviation for all ten samples. Simulation generates either "NaN" (corresponding to negative variance estimates) or extremely low values that do not compare at all with the actual standard deviations that we computed from the sample of estimates.

Table 2.5: Gradient for p_4

parameter	method	δ				
		0.1	0.05	0.01	0.005	0.001
B_1	sim	-63630.125	-49573.247	-462.302	-421.832	-353.12
B_4	sim	10661.369	14185.278	35248.158	70337.149	16.018
B_5	sim	45804.555	14195.608	35258.397	196.565	323.937
B_6	sim	28211.412	28211.698	70319.588	140507.71	114.084
$L_{2,1}$	sim	-3453.588	7066.678	35182.926	70280.602	27.259
$L_{2,2}$	sim	116260.976	91673.341	70605.627	140924.867	419.104
B_1	Genz	-1355.695	-1355.331	-1355.215	-1355.211	-1355.21
B_4	Genz	119.558	119.461	119.43	119.429	119.429
B_5	Genz	1241.001	1240.998	1240.998	1240.998	1240.998
B_6	Genz	812.548	812.548	812.548	812.548	812.548
$L_{2,1}$	Genz	-701.884	-702.625	-702.869	-702.876	-702.879
$L_{2,2}$	Genz	3283.488	3238.467	3224.271	3223.829	3223.688

Table 2.6: Gradient for p_5

parameter	method	δ				
		0.1	0.05	0.01	0.005	0.001
B_1	sim	-7091.145	-71.023	-103.992	-117.655	-189.794
B_4	sim	14157.719	7140.723	154.233	139.789	123.474
B_5	sim	7134.73	7116.737	118.627	112.184	163.104
B_6	sim	-3502.577	-7014.763	15.261	28.761	15.875
$L_{2,1}$	sim	-3503.259	3.608	3.604	-2.549	-9.236
$L_{2,2}$	sim	7076.781	7081.78	27.708	14.396	-7.037
B_1	Genz	-2867.129	244.117	244.155	244.158	244.105
B_4	Genz	709.739	709.638	709.607	709.626	709.315
B_5	Genz	-2301.384	-5405.096	819.638	819.612	819.549
B_6	Genz	3107.24	-6.911	-6.909	-6.921	-7.024
$L_{2,1}$	Genz	-388.526	-389.334	-386.44	-386.277	-386.196
$L_{2,2}$	Genz	-2752.692	-5874.956	380.991	380.968	380.979

Table 2.7: 375 samples with sim. and Genz estimation

parameter	real	mean-Sim	sd-Sim	mean-Genz	sd-Genz
β_1	-1.000	-1.384	0.173	-1.024	0.130
β_4	1.000	1.391	0.210	1.018	0.139
β_5	1.000	1.408	0.237	1.021	0.152
β_6	1.000	1.415	0.262	1.016	0.146
$L_{2,1}$	0.500	0.611	0.388	0.491	0.145
$L_{2,2}$	0.866	1.308	0.692	0.863	0.164

Figure 2.3: Log-likelihood at p_1

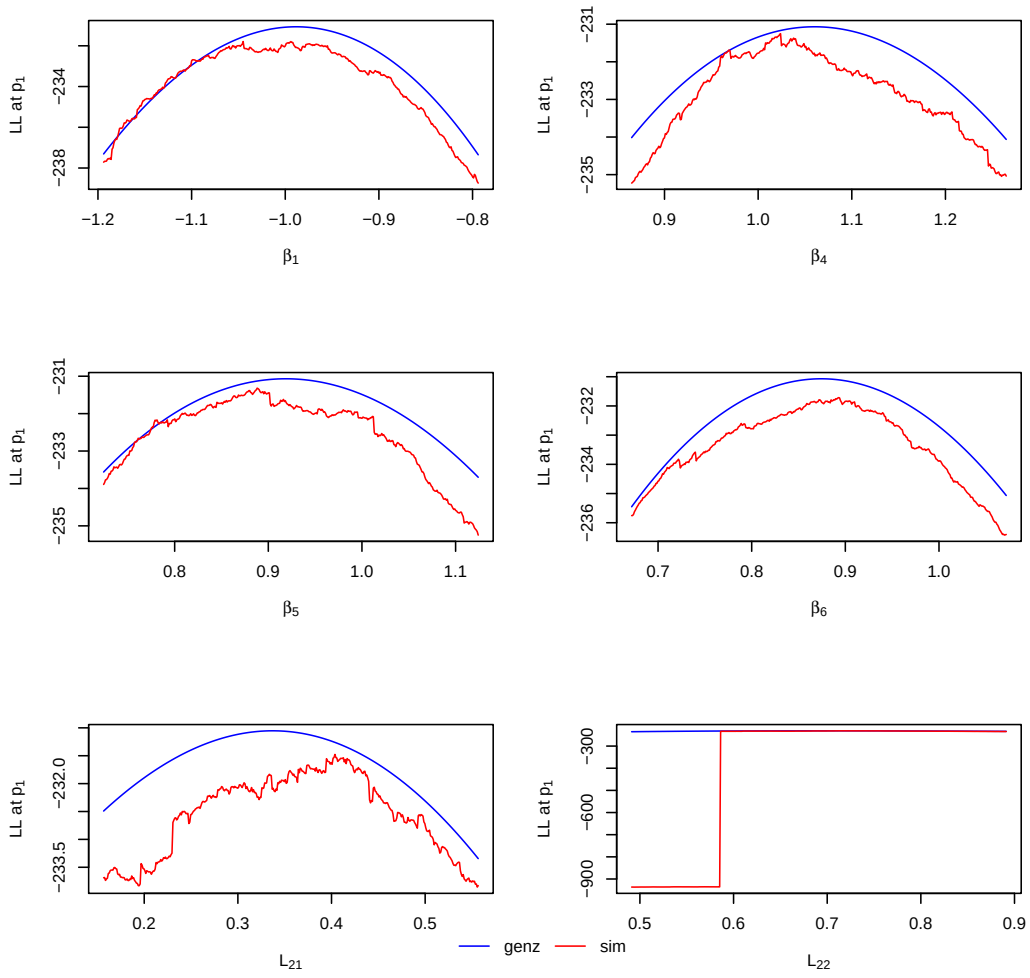


Table 2.8: Standard Errors with Simulation

names \ sample	real	1	2	3	4	5	6	7	8	9	10
β_1	0.182	0.003	0.011	0.006	NaN	0.011	0.000	NaN	0.000	0.000	0.001
β_4	0.216	0.007	0.004	NaN	0.005	0.011	0.004	NaN	0.000	0.000	0.010
β_5	0.245	0.008	0.007	0.005	NaN	0.008	NaN	NaN	0.000	NaN	0.008
β_6	0.190	0.030	NaN	NaN	0.007	0.012	NaN	NaN	0.007	0.007	0.001
$L_{2,1}$	0.348	0.014	0.009	0.008	0.007	0.007	NaN	NaN	0.005	0.011	NaN
$L_{2,2}$	0.256	0.010	NaN	0.016	NaN	0.024	0.004	NaN	0.005	0.019	NaN

Table 2.9: Standard Errors with Genz

names \ sample	real	1	2	3	4	5	6	7	8	9	10
β_1	0.116	0.119	0.107	0.103	0.084	0.144	0.093	0.118	0.102	0.148	0.145
β_4	0.138	0.152	0.114	0.120	0.095	0.140	0.101	0.129	0.108	0.163	0.156
β_5	0.139	0.140	0.148	0.111	0.107	0.153	0.114	0.138	0.115	0.168	0.152
β_6	0.132	0.159	0.132	0.124	0.107	0.165	0.112	0.135	0.120	0.164	0.167
$L_{2,1}$	0.145	0.195	0.132	0.129	0.100	0.143	0.100	0.134	0.110	0.170	0.201
$L_{2,2}$	0.139	0.178	0.149	0.127	0.114	0.168	0.111	0.153	0.120	0.180	0.191

Figure 2.4: Log-likelihood at p_2

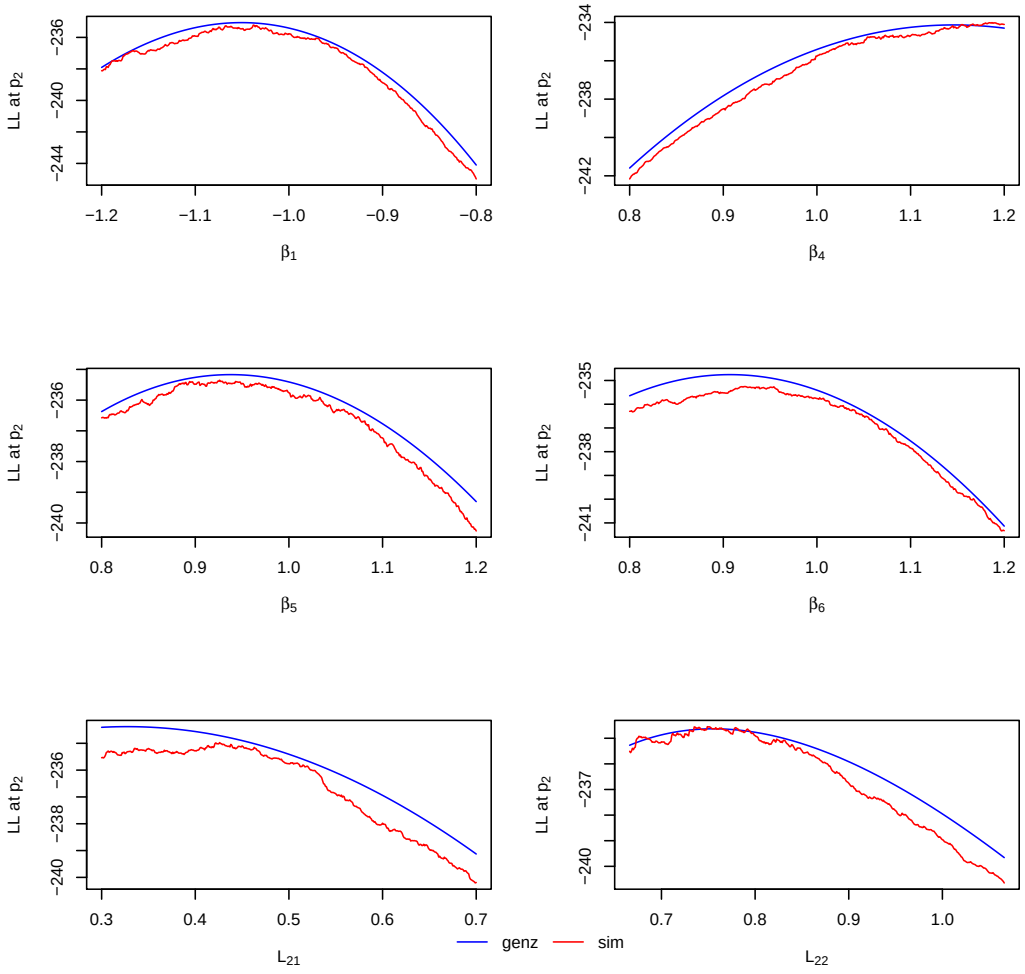


Figure 2.5: Log-likelihood at p_3

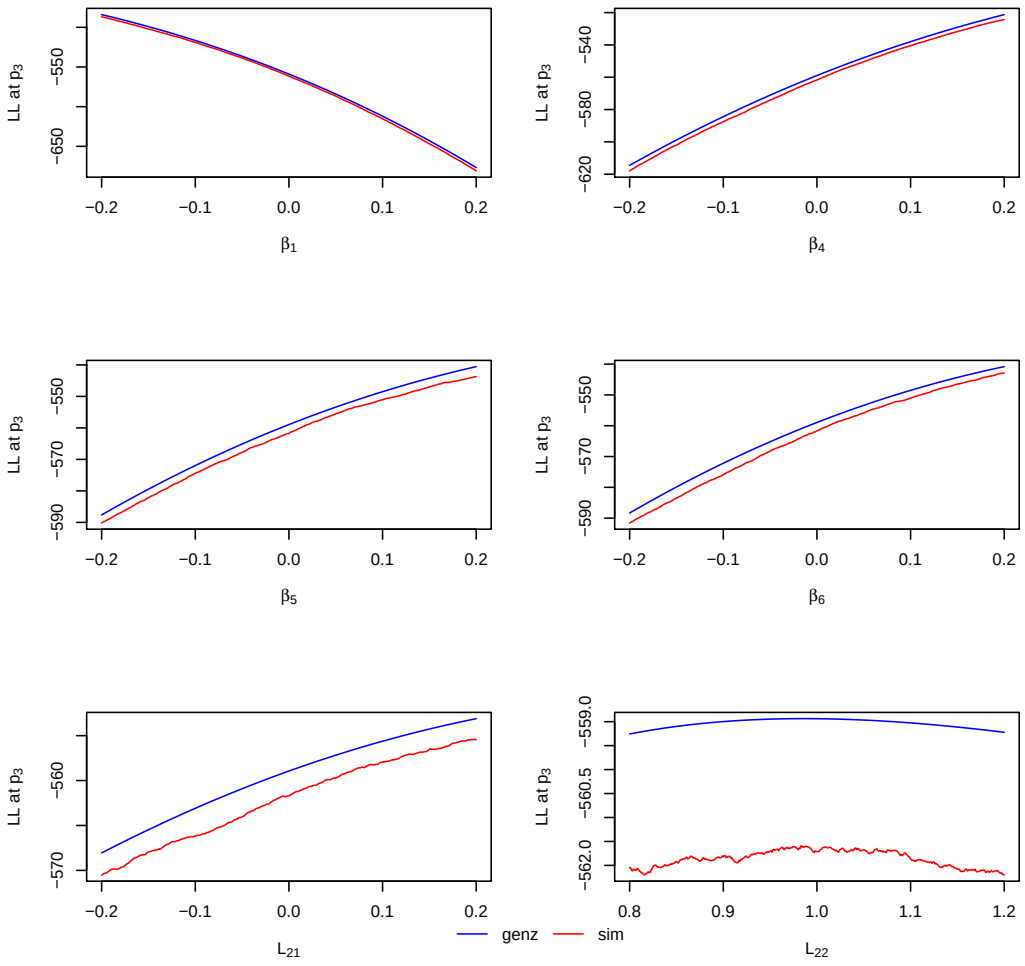


Figure 2.6: Log-likelihood at p_4

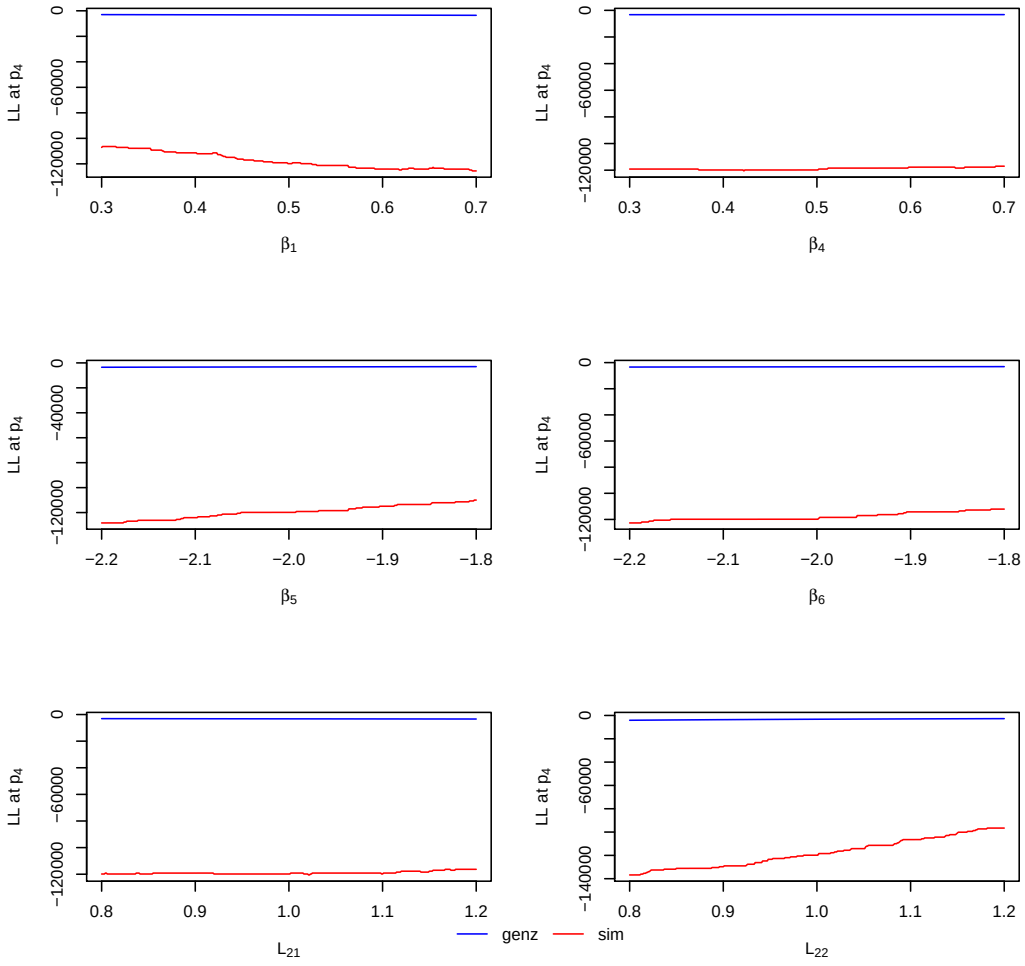


Figure 2.7: Log-likelihood at p_5

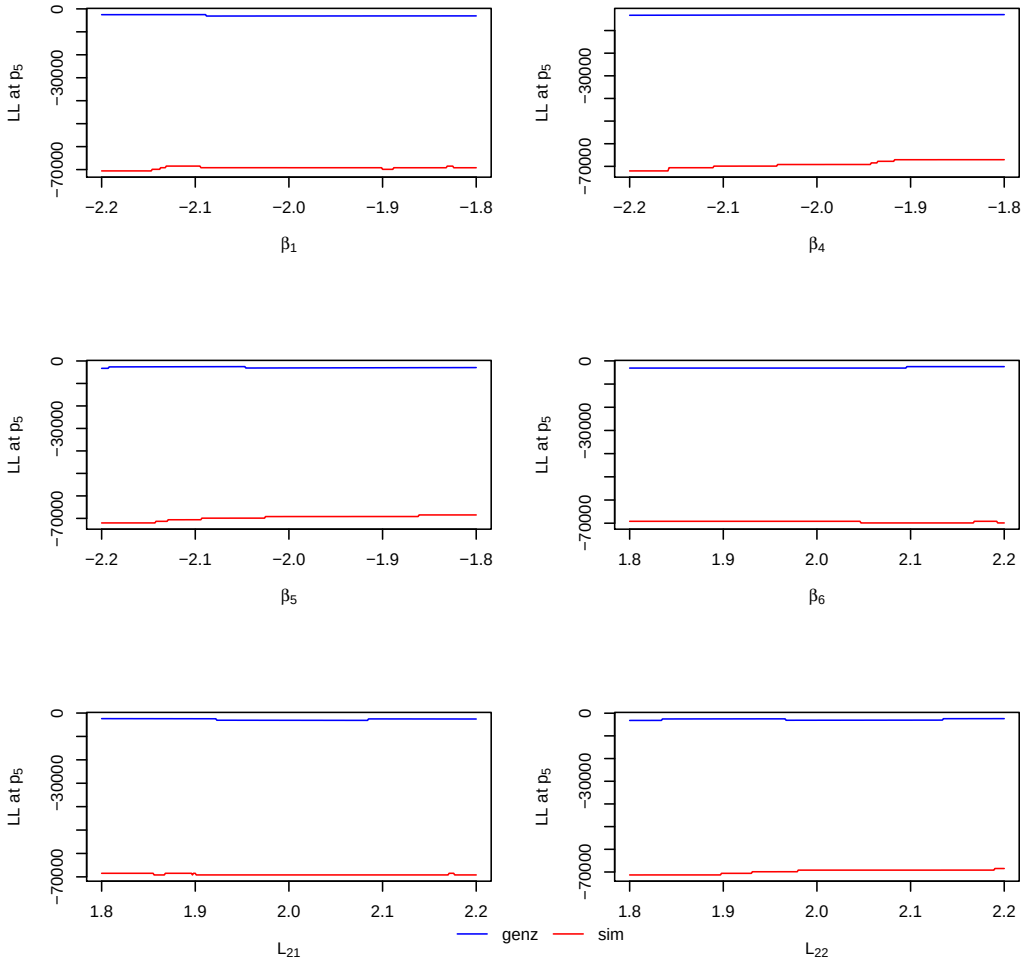
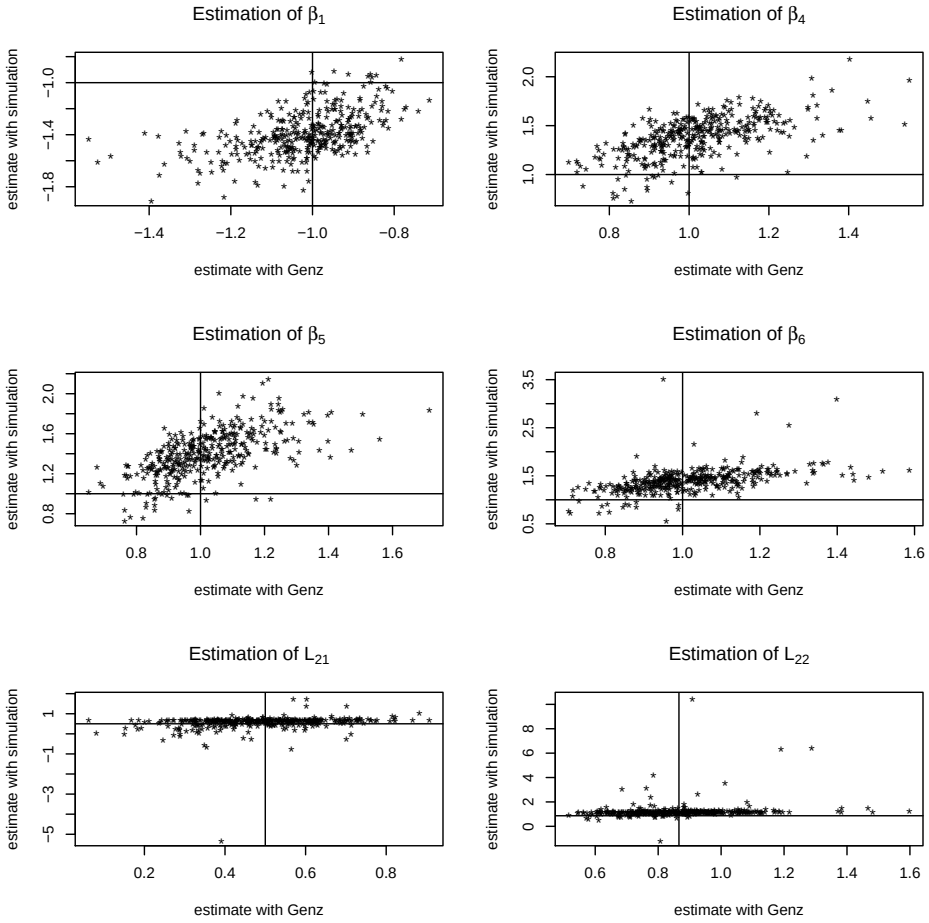


Figure 2.8: Comparison of Genz and simulation estimates



Chapter 3: Ordered and Unordered Discrete-Continuous Probit model with application to vehicle ownership and use

3.1 Introduction

Recently, the advance in the estimation of complex econometric models has allowed analysts to develop comprehensive model systems that jointly analyze multiple choice dimensions. These integrated structures are motivated by the assumption that the different decisions are taken at the same time and are therefore correlated. These modeling frameworks often contain dependent variables that do not belong to the same family (i.e. discrete and continuous variables). This is the case of integrated model for vehicle ownership, where discrete decisions (number of cars, their type and vintage) are naturally linked to continuous decisions (vehicle miles traveled).

In this Chapter we formulate an unordered discrete continuous probit model and we apply it to data on car ownership and use. The unordered discrete-continuous model, is based on multinomial probit and linear regression with unrestricted variance-covariance correlation matrix between the discrete (vehicle holding) and the continuous (vehicle usage) parts. The ordered discrete-continuous structure has a similar

structure with the exception of an ordered probit for the vehicle holding sub-model. The paper also explores the use of numerical approximations methods [Gen92] to overcome problems related to the simulations of probit likelihood functions; notably, the high computational costs, and the instability of the Hessian estimation. The analysis is performed on data extracted from the 2009 National Household Travel Survey [UDoT09]. A summary of this work has recently been published in a peer-reviewed journal [CLT16].

This comparative exercise is motivated by the fact that ordered discrete-continuous models are relatively easier to estimate when compared to unordered model structure due to their closed mathematical form. However, the assumption that vehicle ownership decisions are measured by a single latent variable might affect the goodness of fit of the model and its performance in model application and policy analysis.

3.2 Literature review

Vehicle ownership plays an important role in transportation and land use planning. It is one of the key determinants of people's travel behavior, as it greatly impacts people's mode choice [SW09], frequency of trips [SK12], destination choice, trip timing, activity duration and trip chaining [RCM09]. Vehicle ownership models are also used by policy makers to identify factors that affect vehicle miles traveled (VMT), and therefore address the problems related to traffic congestion, gas consumption and air pollution [DG97], [HKT01]. In this context, ordered probit models

are in general preferred to unordered probit for the saving in computational costs deriving from the closed mathematical form of the choice probabilities.

The earliest generation of discrete-continuous models on vehicle ownership decisions were derived from conditional indirect utility function [Tra86] [HBSM92] [dJ89b] [dJ89a] [dJ91], which is based on micro-economic theory. Originally developed by Dubin and McFadden [DM84], and Hanemann [Han84], the basic concept is that the households choose the combination of vehicle ownership and vehicle usage that gives the highest utility. Roy's identity is applied to estimate vehicle usage and the relationship between the two modeling stages.

Multiple discrete-continuous extreme value (MDCEV) models, developed by Bhat [Bha05] and further applied in Bhat and Sen [BS06] and Bhat et al. [BSE09] are utility-based econometric models that jointly estimate the holding of multiple vehicle types and the miles for each vehicle type. The choice and dependent variable in this model is the mileage for each vehicle type category. Utility for each household is maximized subject to a total mileage budget. Fang [Fan80] developed a BMOPT (Bayesian Multivariate Ordered Probit and Tobit) model, which is composed of a multivariate ordered probit model for the discrete choices and a multivariate Tobit model for the continuous choice. Liu et al. [LTC14] proposed a joint discrete-continuous model to estimate household choices on vehicle holding, type and usage. The discrete components are respectively, multinomial probit for vehicle holding and multinomial logit for the vehicle type sub-models. The joint discrete-continuous model is estimated with unrestricted correlation between the discrete and continuous parts.

Two types of discrete choice modeling structures have been used in the household vehicle ownership studies: ordered-response mechanism and unordered-response mechanism. The ordered-response mechanism assumes that household vehicle ownership is represented as an ordinal variable and the choice is determined by a single latent variable which represents the propensity of the household vehicle ownership decisions. Examples of the application of ordered-response mechanism are Kitamura [Kit87], Golob and Van Wissen [GV89], Golob [Gol90], Kitamura and Bunch [KB92], Bhat and Koppelman [BK93], Kitamura et al. [KGYW99], Hanly and Dargay [HD00], Chu [Chu02], Kim and Kim [KK04] and Cao et al. [CMH07]. The unordered-response mechanism is based on the hypothesis that household vehicle ownership is represented as a nominal variable. It follows the random utility maximization (RUM) principle which assumes that the household makes the vehicle ownership decisions that provides the highest utility among all the possible choices. Examples of the studies with unordered-response mechanism are Mannering and Winston [MW85], Train [Tra86], Bunch and Kitamura [BK90], Purvis [Pur94], Ryan and Han [RH99], Whelan [Whe07]. The reader is referred to Table 3.1 for a more comprehensive description of a sample of these papers.

In the context of the comparison of the ordered and unordered mechanisms, there are several papers that explicitly investigate the empirical performance of the two structures in modeling vehicle ownership decisions. Bhat and Pulugurta [BP98] compared the multinomial logit (MNL) model and the ordered logit (ORL) model on four datasets from Boston, Bay area, Puget Sound area and the Netherlands. The two mechanisms were evaluated by comparing elasticity effects, measure of fit and

predictive performance. The results showed that the MNL model is able to capture elasticity patterns across alternatives, while the ORL is more rigid in elasticity effects. The conclusion from this study is that the appropriate choice mechanism for vehicle ownership modeling is the unordered-response structure. Potoglou and Susilo [PK08] compared the multinomial logit, ordered logit and ordered probit models for car ownership by using data from Baltimore, the Netherlands and Japan. Their results clearly demonstrate the superiority of the MNL to the ordered ORL and ORP. In a study aiming at estimating population heterogeneity in the context of car ownership, Anowar et al. [AYEMM14] propose the application of latent class versions of ordered (ordered logit) and unordered response (multinomial logit) models. The latent class models offer superior data fit compared to their traditional counterparts. In summary, those studies provided strong evidence that the appropriate mechanism is the unordered response mechanism for the vehicle ownership models. However, it is important to stress that the ordered and unordered models have been compared for vehicle holding models only and other applications may lead to different conclusions.

3.3 Ordered discrete-continuous model formulation

The ordered response structure uses latent variables to represent the vehicle ownership propensity of the household, thus it is not consistent with utility maximization theory. Suppose two latent variables Y_d and Y_r represent the preference levels for vehicle holding and vehicle usage (annual miles traveled). The ordered

Table 3.1: Summary of vehicle ownership, type and usage models

Reference	Data Source (Year)	Sample Size	Choices Examined	Model
[Tra86]	US (1978)	1095 household	vehicle quantity, class/vintage, usage	MNL and Regression
[MW85]	US (1978-1980)	3842 households	quantity choice, type choice, utilization model	Nested Logit and OLS regression
[HBSM92]	Sydney (1981-1985)	1444, 1295, 1251, 1197	static vehicle choice and type-mix choice, Static vehicle use, dynamic vehicle choice and use	Nested Logit and 3SLS regression
[KB92]	Dutch National Mobility Panel Data set	Panel, 605 HH, (1984-1987)	vehicle quantity	Ordered Probit
[dJ96]	Dutch National Mobility Panel Data set (Oct, 1992; Oct 1993)	Panel, 3241 respondents	vehicle holding duration, vehicle type choice, annual kilometrage and fuel efficiency	Hazard function, Nested logit, Regression
[BP98]	US (1991, 1990, 1991), Dutch (1987)	3665, 3500, 1822, 1807	vehicle quantity (0, 1, 2, 3, 4)	MNL and Ordered logit
[KGYW99]	California (1993)	Panel (First wave), 4747 households	1) vehicle holding model, and n. of vehicle per HH member and per driver, 2) vehicle type choice, 3) vehicle use	Ordered probit model, Tobit model; MNL; OLS regression
[DG97]	UK, Family Expenditure Survey (1982-1993)	panel, cohort, 7200 households	vehicle quantity	dynamic cohort (panel)
[BS06]	San Francisco (2000)	3500 households	vehicle type holding and usage	MDCEV (multiple discrete-continuous extreme value model)
[Whe07]	UK, (1971-1996) and NTS (1991)	unknown	vehicle quantity (0, 1, 2, 3+)	Hierarchical logit model with saturation level
[Fan80]	NHTS (2001, CA)	2299 households	vehicle choice and usage (BMOPT and MDCEV)	BMOPT (Bayesian Multivariate Ordered Probit and Tobit) and MDCEV
[PK08]	NHTS (2001, Baltimore), Dutch National Mobility Survey (2005), Osaka Metropolitan Trip survey (2000)	3496, 28436, 12632	vehicle quantity	MNL and Ordered logit and Ordered Probit
[BPPL13]	NHTS (2009)	1480	residential and work location, vehicle ownership and tour characteristics	MDCP (multiple discrete-continuous probit)
[LTC14]	NHTS (2009)	1420	Vehicle quantity, vehicle type and vintage, AVMT	Unordered discrete-continuous probit

discrete-continuous model can be written as:

$$\begin{aligned} y_d &= X_d^T \beta_d + \epsilon_d \\ y_r &= X_r^T \beta_r + \epsilon_r \end{aligned} \tag{3.1}$$

where X_d and X_r are explanatory variables for the discrete choice and continuous choice,β_d and β_r are the coefficients to be estimated and ϵ_d and ϵ_r are the

error terms.

The number of vehicles (Y) held by the household is determined by the value of latent variable y_d, which is discretized by a number of cutoff points $\gamma_1, ..., \gamma_{k-1}$. The value of Y is determined by which cutpoints enclose y_d:

$$\begin{array}{lllllll}
 & & y_d & < & 0 & \Rightarrow & Y = 0 \\
0 & < & y_d & < & \gamma_1 & \Rightarrow & Y = 1 \\
\gamma_1 & < & y_d & < & \gamma_2 & \Rightarrow & Y = 2 \\
\dots & & & & & & \\
\gamma_{k-2} & < & y_d & < & \gamma_{k-1} & \Rightarrow & Y = k-1 \\
\gamma_{k-1} & < & y_d & & & \Rightarrow & Y = k
\end{array}$$

Like in the unordered version, in order to jointly to capture the correlation between the discrete and continuous parts the error terms are correlated. Thus, the error terms follow a bivariate normal distribution:

$$(\epsilon_d, \epsilon_r) \sim N(0, \Sigma) \tag{3.2}$$

$$\Sigma = \begin{bmatrix} 1 & \rho\sigma_r \\ \rho\sigma_r & \sigma_r^2 \end{bmatrix} \tag{3.3}$$

Therefore, the model is composed of an ordered probit model and a regression with unrestricted correlation between the error terms.

The estimation scheme is the same than in the unordered model, with a simpler conditioning. The joint probability is separated like this:

$$P(Y, Y_r) = P(Y_r)P(Y|Y_r = y_r) \tag{3.4}$$

The effect on conditioning. is to be able to estimate ϵ_r. For a bivariate normal

distribution, $\epsilon_{\mathrm{d}}|\epsilon_{\mathrm{r}}$ follows a normal distribution with the following parameters:

$$\begin{aligned} \mu_{\epsilon_{\mathrm{d}}|\epsilon_{\mathrm{r}}} &= \rho \ \frac{\epsilon_{\mathrm{r}}}{\sigma} \\ \sigma^2_{\epsilon_{\mathrm{d}}|\epsilon_{\mathrm{r}}} &= 1 - \rho^2 \end{aligned} \tag{3.5}$$

It is interesting to note that the conditional variance $\sigma^2_{\epsilon_{\mathrm{d}}|\epsilon_{\mathrm{r}}}$ does not depend on the observed value ϵ_{r}. The effect of conditioning. on the variance is to uniformly decrease it.

3.4 Data description

Table 3.2: Descriptive Statistics - NHTS 2009

Variables	by Number of cars					Statistics for all vehicle holding cases				
	0	1	2	3	4	min	max	median	mean	s.d.
Vehicle ownership	7.28%	26.72%	43.49%	17.03%	5.48%	0	4	2	1.87	0.96
Hhld. Income level	6.47	10.98	14.55	15.29	15.99	1	18	16	13.21	5.30
Num. of adult	1.31	1.40	1.99	2.21	2.76	1	5	2	1.86	0.66
Num. of workers	0.53	0.69	1.16	1.44	1.63	0	4	1	1.06	0.83
Num. of drivers	0.75	1.26	1.98	2.32	2.83	0	5	2	1.80	0.77
Owned house	0.45	0.75	0.91	0.97	0.98	0	1	1	0.85	0.36
Urban area	0.94	0.84	0.75	0.60	0.56	0	1	1	0.75	0.43
Urban size	4.94	4.25	3.60	2.95	2.55	1	6	5	3.70	2.29
Use of PT	0.26	0.08	0.06	0.05	0.06	0	1	0	0.08	0.27
Age of hhld head	59.13	60.51	53.12	52.16	53.00	18	95	54	55.36	14.91
Female hhld head	0.78	0.64	0.52	0.55	0.44	0	1	1	0.57	0.49
Educ. of hhld head	2.78	3.40	3.67	3.52	3.33	1	5	4	3.49	1.21
Housing unit per sq mile	7233	3637	1341	797	626	50	30000	750	2252	4220
Percent renter-occupied	50	32	22	18	17	0	95	20	26	21
Population per sq mile	12153	6931	3408	2363	1888	50	30000	3000	4725	6333
Workers per sq mile	2870	1899	1042	640	453	25	5000	350	1303	1660
VMT	0	10361	23890	35781	48728	0	91329	19097	21554	16381

The primary data source used in this study is the 2009 National Household Travel Survey [UDoT09]. The analysis is restricted to the Washington D.C. metropolitan area, for which 1,420 observations are available. Household characteristics, land-use variables and information on each household vehicle, are the main variables extracted from the original dataset. Table 3.2 lists the basic statistics relative to the household sample. For the Washington D.C. metropolitan area, the average vehicle ownership per household is 1.87. The percentage of the household

without a car is 7.28%, mainly low-income households; while most households hold 2 cars (43.49%). The number of cars in the household is highly associated with the number of adults and the number of drivers in the family. More than half of the households who do not have a car do not own a house. The land use variables, such as dummy of urban area, urban size, population density and housing density, greatly influence the household car ownership decisions. The households with more cars are generally located in less dense or more rural area. In the Washington D.C. metropolitan area, the average age of the household head is around 55 years old, which is somehow an indication of the aging society happening in western countries. Households with zero or one car have older household head. The average education level in this area is college/bachelors degree; however, households without a car have much lower education level. The average annual mileage traveled by a household is around 20,000 miles per year. The mileage traveled increases accordingly with the household car ownership.

3.5 Empirical results

In this Section results from model estimation (see table 3.3) are presented. We estimated four models: unordered discrete continuous probit model with Monte Carlo simulation (Model 1); unordered discrete continuous probit model with numerical computation in order to test the performance of the Genz approximation (Model 2); ordered discrete continuous probit model (Model 3) having the same formulation of Model 1 and 2 with the exception of the logsum, and unordered discrete

continuous probit model estimated using Genz method and without logsum (Model 4) to compare the fit of the model with Model 2.

The utility functions for the unordered model are the following:

$$\begin{aligned} &\mathrm{U}_0 = 0 \\ &\mathrm{U}_\mathrm{j} = \mathrm{asc}_\mathrm{j} + \beta_\mathrm{inc} \cdot \text{income} + \beta_\mathrm{gen} \cdot \text{female} + \beta_\mathrm{urb} \cdot \text{urban-size} + \beta_\mathrm{den} \cdot \text{density} \\ &\quad (\mathrm{j} = 1, \ldots, 4) \end{aligned} \tag{3.6}$$

The latent variable for the ordered model is:

$$\mathrm{Y} = \beta_0 + \beta_\mathrm{inc} \cdot \text{income} + \beta_\mathrm{gen} \cdot \text{female} + \beta_\mathrm{urb} \cdot \text{urban-size} + \beta_\mathrm{den} \cdot \text{density} \tag{3.7}$$

Even if the coefficients (for example β_inc) have the same name between the two specifications, it is implicit that they are different, since they do not belong in the same model. "asc" refers to Alternative Specific Constants.

The empirical equation for the regression is:

$$\mathrm{Y}_\mathrm{reg} = \beta_0^\mathrm{r} + \beta_\mathrm{inc}^\mathrm{r} \cdot \text{income} + \beta_\mathrm{home}^\mathrm{r} \cdot \text{own-home} + \beta_\mathrm{den}^\mathrm{r} \cdot \text{density} + \beta_\mathrm{cost}^\mathrm{r} \cdot \text{driving-cost} \tag{3.8}$$

Where the superscript r indicates that even if the coefficients may have the same name as their discrete counterpart, they are distinct from them.

3.5.1 Coefficients estimation

Different types of coefficients enter the final specification of the models in table 3.3. In particular a logsum coefficient is estimated in the unordered structure. The logsum represents a feedback variable from the class/vintage models and reflects the utility derived by the household from its choice of class and vintage for each

car in the household. In this case four different type/vintage models have been estimated respectively for household owning 1, 2, 3 and 4 vehicles and therefore four logsum values have been calculated. For more details on the specification of the type/vintage model the reader is referred to Liu et al [LTC14]. The logsum coefficient is constrained to be equal for all the alternatives as in Train [Tra86]; other specifications containing logsum coefficients specific of the alternatives have been tested but results did not improve significantly the fit of the model. The logsum coefficient is positive, significant and less than one; which is similar to what obtained by Train [Tra86]. It should be noted that it was not possible to estimate the logsum variable in the ordered probit model. This represents a further limitation of the ordered probit model as the logsum variable significantly improves the fit of the model as it can be seen for the comparison of the final log-likelihood value of Models 2 and 4.

Most of the estimated coefficients have the expected sign and are significant. Positive coefficients of household income, in both the discrete and continuous parts of the model, indicate that households with higher income have higher tendency to own more vehicles and drive more. In the unordered models, the magnitude increases as the number of vehicles in the household increases. Similar results can be found in Bhat and Pulugurta [BP98] and Potoglu and Susilo [PK08]. The coefficients of number of drivers in the household are very significant, indicating that this factor has high effects on how many cars a household owns. This coefficient is positive in the ordered structure, and also positive in the unordered structure with an exception for the one-car households. The negative coefficient for one-car household alternative

indicates that, the more drivers in the household, the less likely they own only one car. Similarly, households with female household head are less likely to own more vehicles.

Urban size is an indicator of the urbanization level in the area of the household location, in which the lower value represents urban areas and the higher value represents rural areas. Residential density is an indication of the built environment around the household location. The coefficients of these two variables are significantly negative (with the exception of the one-car household alternative) and have higher magnitude as the households own more cars in the unordered structure. In both modeling structures (ordered and unordered probit models) households located in highly residential areas are more likely to own fewer cars and to drive less; households located in a more rural area have higher probability of having more cars and drive more.

The driving cost (measured in dollars per mile) results to be significant and negative, indicating that higher driving cost induces the households to drive less.

3.5.2 Covariance matrix estimation

The covariance matrices of the four models are reported in tables 3.4, 3.5, 3.6 and 3.7. In the unordered discrete-continuous models, the bottom line of the matrix explains the correlation between the mileage traveled and the utility differences of the vehicle holding alternatives with respect to the alternative of owning zero car. In mathematical terms, the estimation of the correlation factors modifies the

values of differences in utility and reduces the variance of those differences, which ultimately contributes to a better forecasting of both the discrete and the continuous components of the model. Generally, the correlation between annual VMT and vehicle holding levels (with respect to zero-car alternative) increases from the one-car alternative to the three-car alternative, and then declines for the four car alternative, probably because in this case very few observations are available in the sample. These results are consistent across Model 1 and Model 2; while in Model 4 the correlation terms are all positive and follow a consistent increasing trend with the number of alternatives.

In the ordered discrete-continuous models, the correlation between the number of vehicles and mileage traveled is positive and equal to 0.5, which means that the demand of vehicle usage increases the propensity of owning more cars.

3.5.3 Goodness of fit

By comparing the measures of fit of the unordered probit with numerical computation (Model 2) and the one obtained with the Monte-Carlo simulation (Model 1) with 1000 MC draws, it should be noted that the approximation method has a better fit and that the Hessian is more stable, which facilitates the computation of coefficients' t-statistics. However, in this case the values of the standard errors were obtained with Bootstrap techniques for both Model 1 and 2.

The log-likelihood values from the ordered (Model 3) and unordered models (Model 4), both calibrated without the logsum variable) cannot be directly com-

pared because of the different model structure, number of parameters and number of observations. Therefore, the adjusted R^2 is calculated as follows:

$$R^2 = 1 - \frac{LL(\hat{\beta}) - n_{par}}{LL(0)} \tag{3.9}$$

where $LL(\hat{\beta})$ is the value of the log-likelihood function at convergence,$LL(0)$ is the value of the log-likelihood function at 0 ($\beta = 0$), and n_{par} is the number of parameters estimated in the model.

A non-nested test has been also conducted for the ordered and unordered models. This test determines if the adjusted R^2 of two non-nested models are significantly different. The same method is used as in Bhat and Pulugurta [BP98]:

"If the difference in the adjusted R^2 is τ, then the probability that this difference could have occurred by chance, in the asymptotic limit, is bounded by:

$$\Phi(-(-2\tau LL(0) + (n_{par,2} - n_{par,1}))^{0.5}) \tag{3.10}$$

in the asymptotic limit. A small value of the probability of chance occurrence indicates that the difference is statistically significant and that the model with the higher value of adjusted likelihood ratio index is to be preferred. "

The values of the adjusted R^2 and those obtained with the non-nested test for the four models are reported in table 3.3. We observe that the unordered discrete-continuous models have higher goodness of fit. The unordered models have higher log-likelihood and adjusted R^2; the non-nested test attests that the unordered mod-

els significantly improve the model fit when compared to the ordered models.

Table 3.3: Estimation Results

Variable	Model 1		Model 2		Model 3		Model 4	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Dependent variable: Number of cars								
Logsum	0.388	0.012	0.515	0.018				
Constant					0.363	0.111		
1 car	2.863	0.237	-2.948	0.217			1.368	0.096
2 cars	-8.700	0.098	-21.889	0.284			-4.432	1.039
3 cars	-14.404	0.188	-28.931	0.257			-4.918	0.113
4 cars	-21.385	0.201	-35.658	0.245			-11.288	0.114
Income					0.086	0.006		
1 car	-0.051	0.011	-0.101	0.020			0.151	0.010
2 cars	0.056	0.006	0.692	0.072			0.395	0.029
3 cars	0.105	0.010	0.745	0.077			0.429	0.028
4 cars	0.111	0.012	0.693	0.074			0.327	0.026
num. of drivers					0.608	0.057		
1 car	-0.010	0.007	-0.304	0.236			-0.048	0.027
2 cars	3.223	0.079	9.236	0.214			2.111	0.215
3 cars	4.041	0.102	10.167	0.190			1.742	0.086
4 cars	4.432	0.092	10.120	0.165			3.314	0.142
gender (female)					-0.235	0.063		
1 car	-0.129	0.551	-0.063	0.249			-0.054	0.068
2 cars	-0.874	0.054	-3.434	0.213			-0.732	0.245
3 cars	-0.928	0.073	-3.605	0.211			-0.854	0.281
4 cars	-0.885	0.059	-3.667	0.194			-2.208	0.360
urban size					-0.032	0.013		
1 car	0.077	0.035	-0.109	0.058			-0.013	0.028
2 cars	-0.120	0.074	-0.270	0.178			0.103	0.277
3 cars	-0.199	0.093	-0.354	0.186			-0.038	0.018
4 cars	-0.201	0.084	-0.406	0.183			-0.368	0.063
res. Density					-0.103	0.010		
1 car	0.041	0.005	0.101	0.015			-0.168	0.017
2 cars	-0.223	0.034	-1.112	0.159			-0.472	0.036
3 cars	-0.442	0.054	-1.298	0.181			-0.740	0.070
4 cars	-0.484	0.064	-1.262	0.170			-0.599	0.183
α_1					1.580			
α_2					3.149			
α_3					4.201			
Dependent variable: VMT (10k)								
Constant	1.130	0.102	1.385	0.116	1.473	0.105	1.456	0.068
Income	0.129	0.005	0.128	0.007	0.132	0.006	0.127	0.066
own home	0.671	0.277	0.328	0.098	0.258	0.072	0.296	0.060
gender (female)	-0.056	0.034	-0.095	0.061	-0.080	0.059	-0.035	0.013
res. density	-0.113	0.008	-0.118	0.009	-0.120	0.011	-0.117	0.006
driving cost ($ per mile)	-5.103	0.283	-4.670	0.285	-5.133	0.238	-4.967	0.098
Log-likelihood at zero	-9583.87		-9583.87		-9583.87		-9583.87	
Log-likelihood at convergence	-3349.81		-3288.93		-3607.75		-3472.51	
Number of parameters	25		25		10		24	
Number of observations	1420		1420		1420		1420	
Adjusted R^2	0.648		0.654		0.623		0.635	
Likelihood ratio test	367.16 > $\chi_{25,0.01}$				-			
Non-nested test result	-				$\Phi(15.98)$ = 1.34 e-56			

Model 1: unordered discrete continuous probit model with simulation
Model 2: unordered discrete continuous probit model with numerical computation
Model 3: ordered discrete continuous probit model
Model 4: unordered discrete continuous probit model without logsum

3.5.4 Results from model application

The models 2 and 3 have been applied to test policy scenarios; the variables of interest are income, density and driving cost. The scenarios considered are listed in table 3.8.

Results in table 3.3 show the effects of those variables on vehicle holding and

Table 3.4: Covariance Matrix, Model 1, Unordered Monte Carlo

	1 car	2 cars	3 cars	4 cars	VMT
1 car	2.00	1.14	1.31	1.30	0.27
2 cars	1.14	1.63	0.37	0.76	0.10
3 cars	1.31	0.37	2.37	1.68	0.67
4 cars	1.30	0.76	1.68	1.36	0.46
VMT	-0.27	0.10	0.67	0.46	1.23

Table 3.5: Covariance Matrix, Model 2, Genz

	1 car	2 cars	3 cars	4 cars	VMT
1 car	2.00	10.34	10.24	10.57	0.73
2 cars	10.34	58.26	61.44	61.57	4.46
3 cars	10.24	61.44	68.64	67.11	5.21
4 cars	10.57	61.57	67.11	66.34	5.00
VMT	-0.73	4.46	5.21	5.00	1.25

mileage traveled. It appears that results related to vehicle holding are consistent between the ordered and the unordered structures. The effects are very limited in all the cases considered, but in general slightly higher for the ordered model.

Increases in income and density result into slightly more households with 0 and 1 car, and fewer households with 2, 3, and 4+ cars. With reference to income, we calculate that a 10 percent increase in income will result into higher vehicle ownership of about 1.5% and 4.0% according respectively to the unordered and ordered model. These values are lower than those provided by Litman [Lit13], who indicate an average elasticity of vehicle ownership to income of 1.0, but close to the 0.4 calculated by Goodwin et al. [GDH04].

The effect of built environment variables (i.e. density) on vehicle ownership has been studied recently by different authors in transportation and economics and it is interesting to compare our results with those obtained in previous studies. In particular, Fang [Fan80] found from a model calibrated on 2001 data and relative

Table 3.6: Covariance Matrix, Model 3, Ordered DC

	#cars	VMT
#cars	1.00	0.50
VMT	0.50	1.56

Table 3.7: Covariance Matrix, Model 2, Genz (no logsum)

	1 car	2 cars	3 cars	4 cars	VMT
1 car	2.00	3.31	3.95	3.43	1.48
2 cars	3.31	12.89	5.69	4.64	2.38
3 cars	3.95	5.69	11.67	12.19	3.43
4 cars	3.43	4.64	12.19	36.93	4.56
VMT	1.48	2.38	3.43	4.56	1.24

to California that a 50 percent increase in density causes a reduction of 1.2 percent in truck holdings, a larger truck VMT reduction (about 8 percent) and a small car VMT change (- 1.32 percent). Bento et al. [BCMV05] calculate a density elasticity of 0.1. These values compares pretty well with what found in our study which gives a reduction of 1.5 percent in vehicle holdings, a reduction of 5.6 percent in vehicle use and an elasticity of VMT to density equal to 0.11.

Driving cost change has almost no effects on vehicle holding decisions but has a significant effect on car travel demand. Our model predicts that 50 percent increase in fuel cost produces a reduction of 17.7 percent in VMT for the unordered model and about 20 percent for the ordered model. The elasticity of vehicle travel with respect to fuel prices is 0.35 for the unordered model and 0.4 for the ordered model. These values are higher than the average values of 0.3 provided in the literature [Lit13] and found in a number of other studies (TRACE, [TRA99], but close to the value of 0.34 calculated using data from 1968 to 2008 by Li et al. [LLM11].

Table 3.8: Sensitivity Scenarios

variable	income	density	driving cost
% change	-10	-50	-50
	-5	-25	-25
	5	25	25
	10	50	50

3.6 Conclusions

The proposed unordered modeling structure is able to account for full correlation among simultaneous decisions and among the discrete alternatives. This approach overcomes the limitation of existing models by eliminating the hypotheses of fixed budget for the continuous variable and of independence from irrelevant alternatives. The use of numerical methods for the computation of the multivariate normal probabilities has been introduced for the first time in an econometric context and successfully applied to estimate model formulations that include both discrete and continuous dependent variables. This approach is a valid alternative to Monte Carlo simulation, especially in cases when the problem can be modeled just with nominal variables. We have also estimated an ordered response model; although previous literature has shown that ordinal variable vehicle ownership models are inferior to the nominal variable approach in the context of joint discrete-continuous models they might represent an attractive option due to their closed form mathematical formulation.

Empirical results although consistent with previous literature provide a number of interesting insights for policy analysis. The unordered discrete-continuous model always performs better in terms of goodness of fit statistics when compared

to ordered discrete-continuous models. However, the two structures have very similar elasticity values, with the ordered model showing slightly higher elasticities in the discrete part with respect to income and density, and to fuel cost in the continuous part. By using data from 2009 a relatively high elasticity value for fuel cost (about 0.36) is derived, which might be due to the high fuel cost experienced by drivers at that time and to the economic crisis that affected the United States after 2008. We have also experienced difficulty in estimating the logsum variable in the ordered model. Although we do not fully know the reasons for this problem, the impossibility of accounting for the utility deriving from the choice of vehicle type and vintage represents another important limitation of the model based on ordered probit in the context of vehicle ownership modeling.

The proposed models are highly flexible and can be transferred to other integrated decisions that are relevant in transportation and related disciplines (i.e. number of daily activities and time dedicated to each activity, mode choice and departure time; types of car owned and mileage traveled for each type).

Table 3.9: Application Results

Corresponding change	Change in residential density					Change in driving cost			
	0car	1 car	2 cars	3 cars	4 cars	Average veh. ownership		miles	
Actual	7.22%	22.59%	46.82%	17.92%	5.45%	1.92		22,490	
Unordered DC									
Income -10%	7.20%	24.52%	45.93%	16.92%	5.44%	1.89	-1.51%	20,829	-7.39%
Income -5%	7.22%	23.49%	46.39%	17.46%	5.45%	1.90	-0.71%	21,666	-3.67%
Income +5%	7.22%	21.75%	47.18%	18.39%	5.46%	1.93	-0.70%	23,310	3.65%
Income +10%	7.25%	20.70%	47.71%	18.88%	5.47%	1.95	1.48%	24,151	7.38%
Density -50%	7.24%	20.10%	47.99%	19.20%	5.46%	1.96	1.96%	23,730	5.51%
Density -25%	7.25%	21.38%	47.42%	18.49%	5.46%	1.94	0.91%	23,080	2.62%
Density +25%	7.12%	23.79%	46.21%	17.43%	5.45%	1.90	-0.77%	21,850	-2.85%
Density +50%	6.97%	24.91%	45.82%	16.86%	5.45%	1.89	-1.51%	21,231	-5.60%
Fuel cost -50%	7.22%	22.58%	46.86%	17.89%	5.45%	1.92	-0.01%	26,479	17.74%
Fuel cost -25%	7.23%	22.54%	46.81%	17.95%	5.47%	1.92	-0.05%	24,501	8.94%
Fuel cost +25%	7.23%	22.56%	46.74%	18.02%	5.45%	1.92	0.05%	20,489	-8.90%
Fuel cost +50%	7.22%	22.54%	46.84%	17.94%	5.46%	1.92	0.05%	18,501	-17.74%
Ordered DC									
Income -10%	7.66%	27.50%	44.10%	15.98%	4.76%	1.83	-4.23%	20,802	-7.76%
Income -5%	7.36%	26.23%	44.04%	17.01%	5.36%	1.87	-2.08%	21,708	-3.74%
Income +5%	6.77%	24.09%	43.58%	18.83%	6.73%	1.95	2.06%	23,426	3.87%
Income +10%	6.50%	23.21%	43.02%	19.78%	7.50%	1.99	4.09%	24,333	7.90%
Density -50%	5.03%	24.53%	45.05%	18.88%	6.51%	1.97	3.44%	23,857	5.78%
Density -25%	6.00%	24.92%	44.39%	18.40%	6.29%	1.94	1.74%	23,214	2.94%
Density +25%	8.11%	25.37%	43.28%	17.43%	5.80%	1.87	-1.73%	21,904	-2.87%
Density +50%	9.07%	25.49%	42.78%	17.02%	5.64%	1.85	-3.18%	21,278	-5.65%
Fuel cost -50%	7.02%	25.22%	43.85%	17.89%	6.02%	1.91	-0.04%	27,030	19.85%
Fuel cost -25%	7.01%	25.23%	43.85%	17.87%	6.04%	1.91	-0.03%	24,801	9.97%
Fuel cost +25%	7.03%	25.29%	43.78%	17.87%	6.03%	1.91	-0.08%	20,322	-9,89%
Fuel cost +50%	7.05%	25.23%	43.83%	17.94%	5.96%	1.91	-0.11%	18,087	-19.80%

Chapter 4: Validation of Discrete-Continuous Models

4.1 Introduction

In this chapter, ordered discrete-continuous (ODC) and unordered discrete-continuous (UDC) models are validated using both simulated and real data. In both cases, validation is carried out on holdout samples. The real data is relative to the car ownership and use problem and are calibrated on the 2009 NHTS data.

4.2 Simulated data: ordered discrete-continuous (ODC) model

4.2.1 ODC Data generation

For this simulated case, we generate 2,000 observations and we assume that each synthetic individual has a choice across 4 discrete alternatives. By design, the dependent variables are marginally distributed according respectively to an ordered probit and a regression model. We are using 10 predictors X_1 through X_{10}, all assumed to follow a standard normal distribution. Two datasets are generated. The first assumes that low correlation exists between the error terms of the discrete and the continuous models ($\rho = 0.1$); while the second assumes high correlation between the error terms ($\rho = 0.9$). For both cases, the predictors of the ordered

probit are `const`, X_1, X_2, X_3, X_4 and X_5 with coefficients $(1, 1, –1, 0.5, 0.5, 2)$; while the predictors of the regression are `const`, X_6, X_7, X_8, X_9 and X_{10} with coefficients $(–1, 1, 1, 1, 2, –2)$. The model intercepts are denoted by `const` and the variance of the regression is $\sigma^2 = 25$. The models differ only in the correlation value. The cutpoints of the ordered probit are set to $\Gamma = (1, 2)$, such that the differences between them are $\alpha = (1, 1)$. The model is estimated on 1,600 observations and applied to the remaining 400 observations in the sample.

4.2.2 ODC validation results: low correlation

Results relative to the validation of the ordered discrete-continuous model with low correlation ($\rho = 0.1$) are presented in figures 4.1 through 4.6. Table 4.1 summarizes the predicted probability of the observed choice in the validation sample for the following cases: 1) joint ordered discrete-continuous model; 2) ordered probit model and regression; 3) ordered probit only. Not surprisingly all summary statistics are within a narrow range and the joint model does approximately as good as the separate models and the ordered probit model alone. The log-likelihood in the validation sample is approximately the same for all the three models considered.

Figure 4.1 shows how the three models compare with market shares in terms of predictions. In all cases, bigger market shares are associated with bigger predicted probabilities. The circles (joint model), triangles (separate model), and crosses (discrete) are all located at approximately the same height on a vertical scale, meaning that none of them generates better predictions for any alternative.

In Figure 4.2 the probabilities are plotted against their ordered probit counterpart. Since there is almost no correlation between the error terms, both the circles (joint model) and the triangles (separate model) are approximately distributed along the identity line. This is to be expected since the small correlation induces noise in the predictions of the two models.

Predicted values for the regression are not very relevant in the context of a joint model because our model formulation uses the error of the continuous part to improve the prediction of the discrete part. There will be large differences in the predicted values only if the joint model optimizes the regression coefficient in such a way to improve the overall fit, but they cannot benefit directly from the conditioning in our scheme. If we were to report the predicted probability of the discrete model and then use this information to adjust the predicted value of the regression, we might be able to see an improvement in the regression fit. However, as discussed in previous chapters the conditioning is a difficult task.

Figure 4.3 shows the predicted values for the validation sample. It is worth noting that the apparent poor fit is due to the high variance of the residuals and no modeling approach can fix it. The coefficients of the models are given in tables 4.2 through 4.6. Not surprisingly, we are able to recover the true value of the coefficients in all cases, that is typically the case for a sample generated following the true model.

Table 4.1: ODC Validation - $\rho = 0.1$

model	min.	1^{st} q.	median	mean	3^{rd}q.	max.	Log-likelihood
joint ODC	0.003	0.328	0.580	0.603	0.949	1.000	-283.346
separate ODC	0.0026	0.326	0.588	0.601	0.946	1.000	-286.116
just OP	0.003	0.322	0.590	0.602	0.948	1.000	-286.287

Figure 4.1: Simulated ODC, $\rho = 0.1$ - Comp. with Market Shares

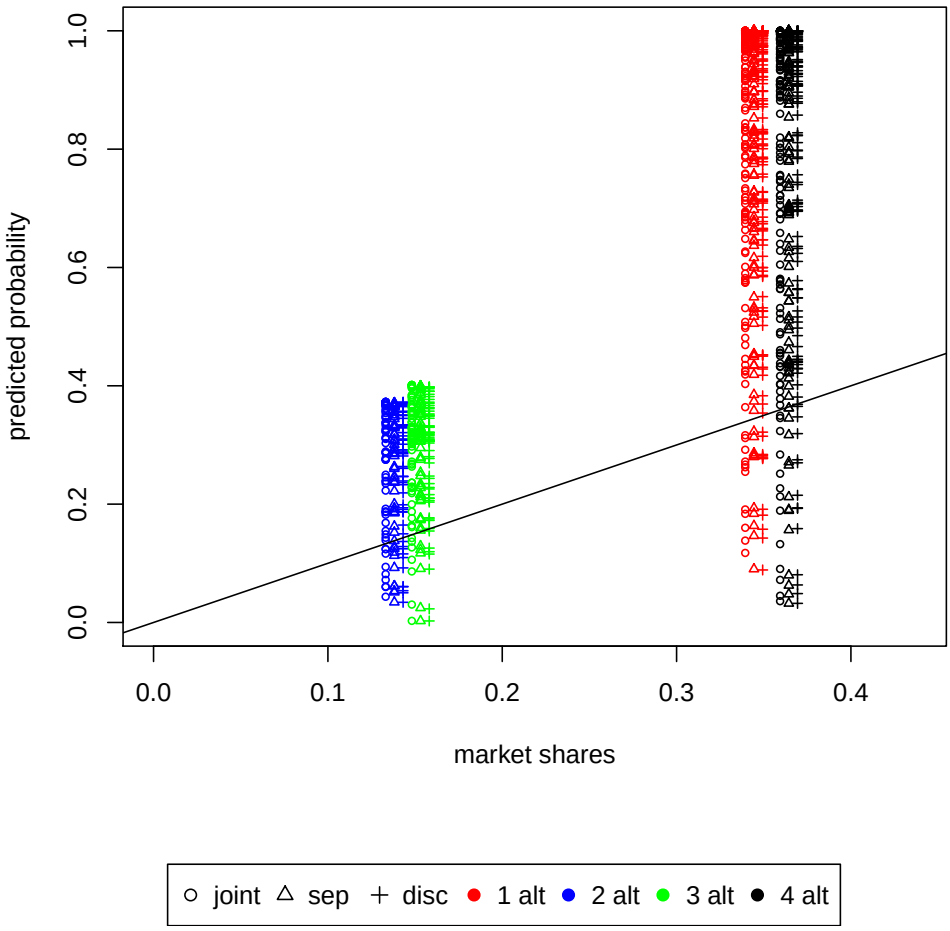


Figure 4.2: Simulated ODC, $\rho = 0.1$ - Comp. with OP

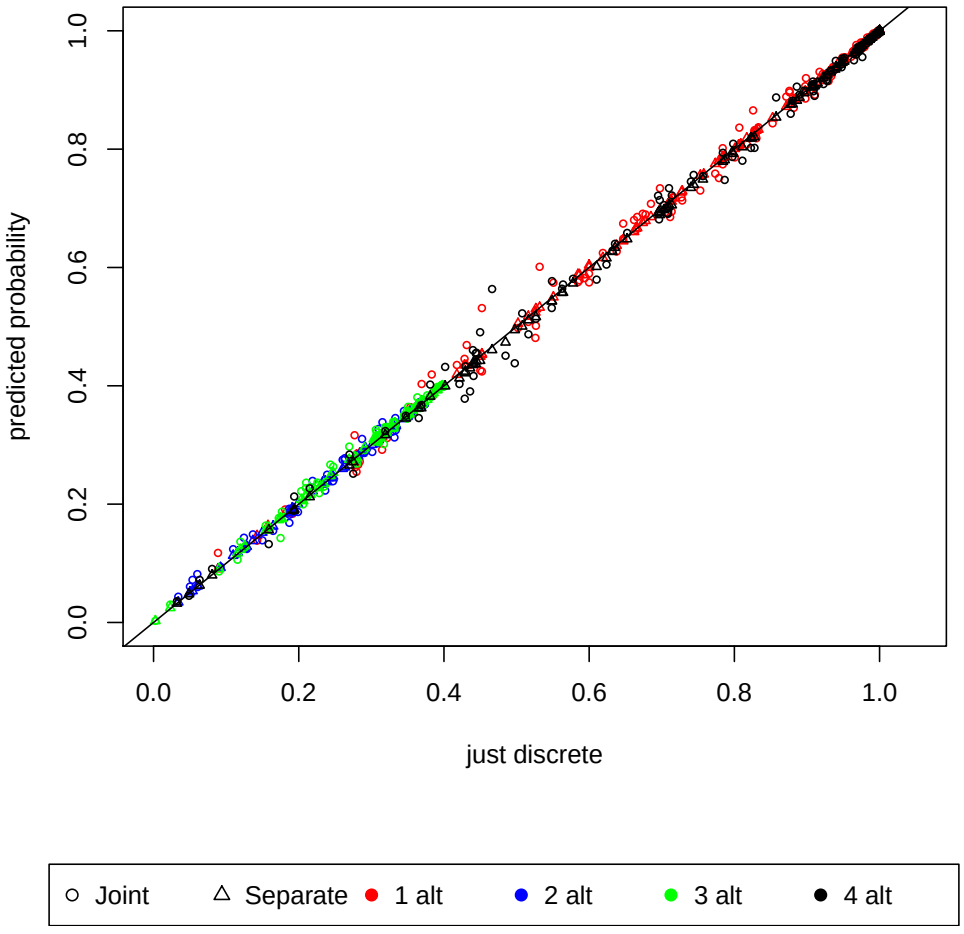


Figure 4.3: Simulated ODC, $\rho = 0.1$ - Regression

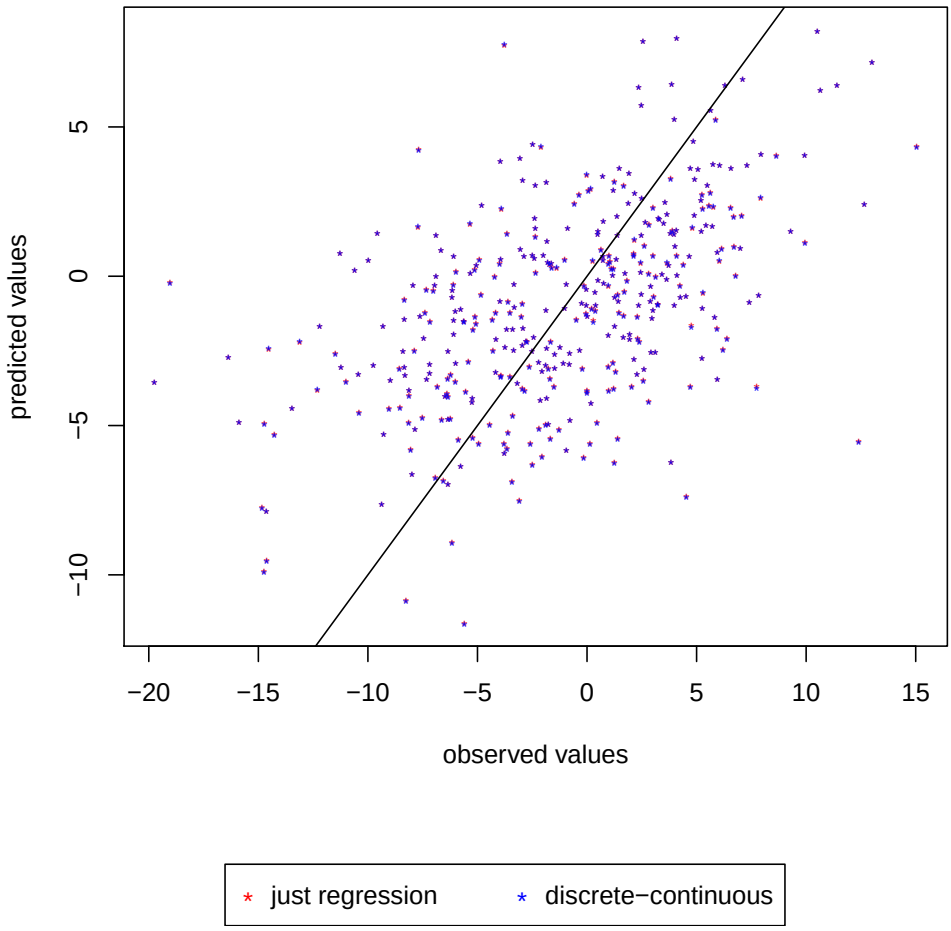


Table 4.2: Simulated ODC, $\rho = 0.1$ - ODC coefficients for validation sample

name	true param.	estimate	SD	t	p
const	1	0.983	0.054	18.196	0.000
X_1	1	0.975	0.046	21.026	0.000
X_2	-1	-0.993	0.045	-21.881	0.000
X_3	0.5	0.405	0.037	10.802	0.000
X_4	0.5	0.512	0.038	13.498	0.000
X_5	2	2.017	0.069	29.287	0.000
α_1	1	0.966	0.058	16.678	0.000
α_2	1	1.048	0.059	17.686	0.000
const	-1	-0.938	0.123	-7.610	0.000
X_6	1	0.875	0.124	7.084	0.000
X_7	1	0.981	0.124	7.916	0.000
X_8	1	1.117	0.123	9.048	0.000
X_9	2	2.128	0.119	17.915	0.000
X_{10}	-2	-2.070	0.120	-17.251	0.000
σ^2	25	24.242	0.856	28.323	0.000
ρ	0.1	0.071	0.035	2.027	0.043
maxLL		-5868.34			
n		1600			

Table 4.4: Simulated ODC, $\rho = 0.1$ - OP coefficients for validation sample

name	true param.	estimate	SD	t	p
const	1	0.991	0.054	18.314	0.000
X_1	1	0.982	0.046	21.139	0.000
X_2	-1	-0.996	0.046	-21.845	0.000
X_3	0.5	0.401	0.038	10.650	0.000
X_4	0.5	0.515	0.038	13.498	0.000
X_5	2	2.024	0.069	29.355	0.000
α_1	1	0.969	0.058	16.707	0.000
α_2	1	1.043	0.059	17.680	0.000
maxLL		-1047.853			
n		1600			

Table 4.6: Simulated ODC, $\rho = 0.1$ - Regression coefficients for validation sample

name	true param.	estimate	SD	t	p
`const`	-1	-0.926	0.124	-7.493	0.000
X_6	1	0.867	0.124	6.994	0.000
X_7	1	0.989	0.124	7.952	0.000
X_8	1	1.115	0.124	9.001	0.000
X_9	2	2.126	0.119	17.835	0.000
X_{10}	-2	-2.060	0.120	-17.106	0.000
sigma	5	4.936			
adj R2		0.346			
n		1600			

4.2.3 ODC validation results: high correlation

All the results for the ordered discrete-continuous model with high correlation ($\rho = 0.9$) are presented in figures 4.8 through 4.13. Table 4.8 reports the summary of predicted probabilities in the validation sample. Except for the minimum and maximum predicted probabilities that are almost zero and one for all three approaches, the quartiles, mean and log-likelihood of the joint model are greatly improved.

Figures 4.4 and 4.2 report predicted probabilities against market shares of ordered probit. As expected all three methods defeat market share predictions. The impact of the high correlation is shown in figure 4.2. The circles denote the conditional probabilities and the triangles the unconditional probabilities. The triangles are distributed along the identity line, meaning that the unconditional probabilities are about the same than their equivalent using just an ordered probit model. The circles are mostly high above the identity line, meaning that the joint model does much better than the ordered probit. Some circles are however, below the identity line meaning that conditioning worsens the predictions for these observations. It appears that while the discrete-continuous model improves the predictions when there is sufficient correlation between error terms, this improvement is never uniform such that there are always units that have a worse conditional prediction.

The model's coefficients are reported in tables 4.9 through 4.13. It is worth noting is that the model estimates are not always close to the true values. It is hard to explain why this happens, but we think that it may have to do with the very high correlation that might affect the identification of the model.

Table 4.8: ODC Validation - $\rho = 0.9$

model	min.	1st q.	median	mean	3rdq.	max.	Log-likelihood
joint ODC	0.036	0.635	0.911	0.785	1.000	1.000	-134.535
separate ODC	0.020	0.307	0.577	0.591	0.935	1.000	-293.999
just OP	0.009	0.334	0.562	0.600	0.957	1.000	-296.4021

Table 4.9: Simulated ODC, $\rho = 0.9$ - ODC coefficients for validation sample

name	true param.	estimate	SD	t	p
`const`	1	0.846	0.041	20.874	0.000
X_1	1	0.871	0.026	33.201	0.000
X_2	-1	-0.845	0.026	-32.453	0.000
X_3	0.5	0.450	0.021	21.321	0.000
X_4	0.5	0.423	0.020	21.132	0.000
X_5	2	1.785	0.046	38.741	0.000
α_1	1	0.833	0.043	19.284	0.000
α_2	1	0.896	0.043	20.693	0.000
`const`	-1	-1.092	0.125	-8.741	0.000
X_6	1	0.979	0.083	11.795	0.000
X_7	1	1.062	0.084	12.634	0.000
X_8	1	0.962	0.083	11.540	0.000
X_9	2	1.884	0.082	22.983	0.000
X_{10}	-2	-2.023	0.080	-25.316	0.000
σ^2	25	24.950	0.890	28.038	0.000
ρ	0.9	0.786	0.009	86.762	0.000
maxLL		-5354.587			
n		1600.000			

Table 4.11: Simulated ODC, $\rho = 0.9$ - OP coefficients for validation sample

name	true param.	estimate	SD	t	p
`const`	1	1.021	0.055	18.596	0.000
X_1	1	0.954	0.046	20.972	0.000
X_2	-1	-0.970	0.045	-21.463	0.000
X_3	0.5	0.572	0.040	14.418	0.000
X_4	0.5	0.508	0.039	13.146	0.000
X_5	2	2.083	0.071	29.298	0.000
α_1	1	0.992	0.060	16.604	0.000
α_2	1	1.076	0.061	17.773	0.000
maxLL		-1035.574			
n		1600			

Figure 4.4: Simulated ODC, $\rho = 0.9$ - Comp. with Market Shares

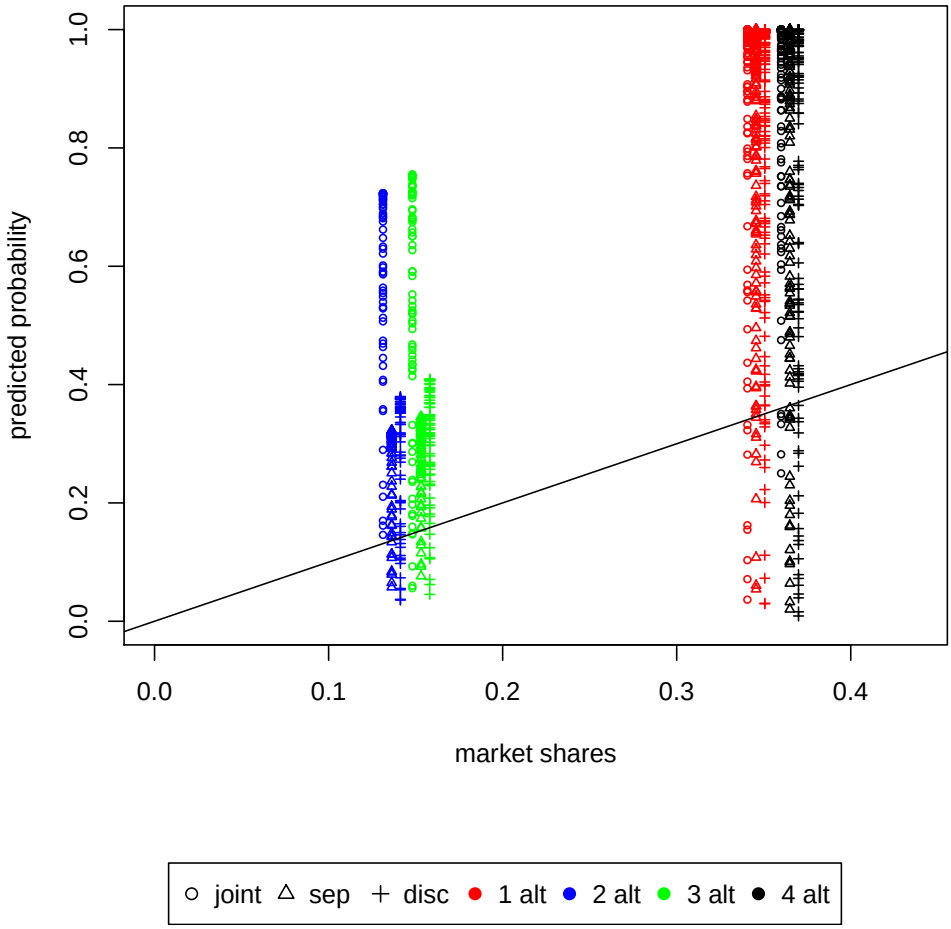


Table 4.13: Simulated ODC, $\rho = 0.9$ - Regression coefficients for validation sample

name	true param.	estimate	SD	t	p
`const`	-1	-1.057	0.125	-8.492	0.000
X_6	1	1.107	0.125	8.866	0.000
X_7	1	1.066	0.125	8.510	0.000
X_8	1	0.960	0.125	7.696	0.000
X_9	2	1.931	0.120	16.088	0.000
X_{10}	-2	-2.089	0.121	-17.225	0.000
sigma	5	4.971			
adj R2		0.340			
n		1600			

Figure 4.5: Simulated ODC, $\rho = 0.9$ - Comp. with OP

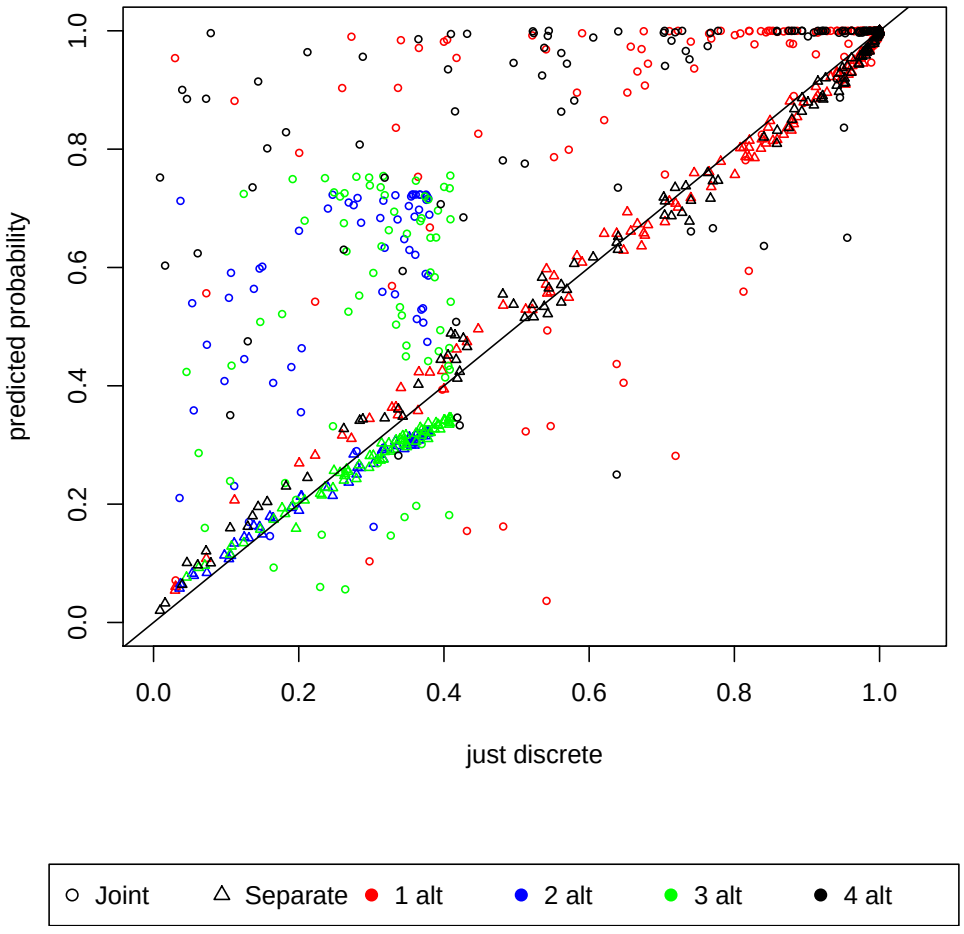
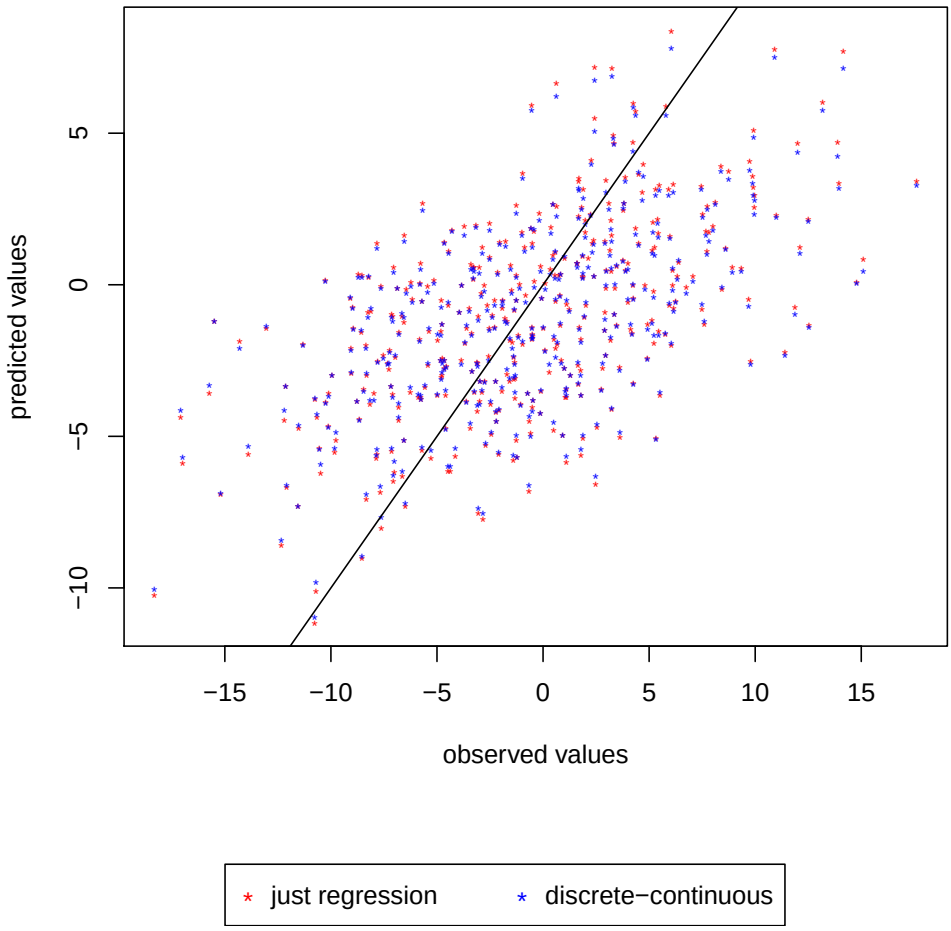


Figure 4.6: Simulated ODC, $\rho = 0.9$ - Regression



4.3 Simulated data: unordered discrete-continuous (UDC) model validation

4.3.1 UDC data generation

For the validation of the unordered discrete-continuous model, we are using a similar procedure. The four utility functions are defined as follows:

$$\begin{array}{lcrcrcrl} U_0 & = & & & X_1 & - & X_2 & +\epsilon_0 \\ U_1 & = & 0.5 & + & 2X_3 & - & 2X_4 & +\epsilon_1 \\ U_2 & = & -0.5 & - & 2X_5 & + & 2X_6 & +\epsilon_2 \\ U_3 & = & 1 & - & X_7 & + & X_8 & +\epsilon_3 \end{array}$$

The continuous part is set to be:

$$Y_r = 1 + X_9 - X_{10} + 2X_1 - 0.5X_2 + \epsilon_r$$

We are considering a case with low correlation, and a case with high correlation. The Cholesky matrix (L) of the covariances (S) is reported because this is what is being optimized in the log-likelihood method. We refer to the joint covariance of differences in utilities and regression residuals. In our case, the first three rows correspond to differences in utilities with respect to the 0^{th} alternative (U_0) and the last row corresponds to the covariance terms between the regression and these differences.

$$S_{\text{low}} = \begin{pmatrix} 1 & 0.25 & 0 & -0.2 \\ 0.25 & 1.0625 & 0.25 & -0.05 \\ 0 & 0.25 & 1.0625 & 0.2 \\ -0.2 & -0.05 & 0.2 & 1.08 \end{pmatrix}$$

$$L_{low} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.25 & 1 & 0 & 0 \\ 0 & 0.25 & 1 & 0 \\ -0.2 & 0 & 0.2 & 1 \end{pmatrix}$$

$$S_{high} = \begin{pmatrix} 1 & 0.8 & -0.3 & -0.6 \\ 0.8 & 1 & 0.24 & -0.48 \\ -0.3 & 0.24 & 0.98 & 0.48 \\ -0.6 & -0.48 & 0.48 & 1.08 \end{pmatrix}$$

$$L_{high} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 \\ -0.3 & 0.8 & 0.5 & 0 \\ -0.6 & 0 & 0.6 & 0.6 \end{pmatrix}$$

4.3.2 UDC validation results: low correlation

The results for the unordered discrete-continuous model are presented in the same way than for the ordered model. Figures 4.15 through 4.20 represent the results obtained for the model with low correlations.

Table 4.15 shows the distribution of predicted probabilities for the joint and unconditional model, as well as for the probit model alone. Although the correlation is "low", it is still higher than for the low correlation of the ordered joint model. We can observe that the mean and quartiles of the joint model are slightly higher

than their unconditional counterpart. Interestingly enough, the median is increased more than the mean and we think that this is due to the observation we made earlier that conditioning actually reduces the predicted probability of some choices. On the other side, the median is a very robust measure of location and is much less affected by this problem.

Figures 4.7 and 4.8 show the predicted probabilities. Again the model outperforms the market share approach. Here the model is more complex because the regression is correlated differently depending on the alternatives. The green circles are closer to the identity line than the circles of other colors. This is likely due to the fact that the regression is less correlated to the third alternative (second column in covariance matrix) than to the other ones with a covariance of -0.05 versus -0.2 and 0.2. Again, the triangles of unconditional probabilities are distributed along the identity line because no information from the correlation is accounted for in their calculation.

Figure 4.9 shows the predicted continuous variable. Just like for the ordered model, we do not use any conditioning information for the regression prediction and we do not expect significant differences for these predictions. The fit appears to be better than for the joint ordered example but this is only due to the smaller variance of the residuals (1.08 versus 25).

Coefficients are reported in tables 4.16 through 4.20. The most interesting information in these tables lie in the difference between the discrete-continuous model coefficients and the probit coefficients. The joint model appears to suffer from some bias that is not present in the probit.

Table 4.15: UDC Validation - low correlations

model	min.	1^{st} q.	median	mean	3^{rd}q.	max.	Log-likelihood
joint UDC	0.012	0.569	0.841	0.738	0.960	1.000	-176.360
separate UDC	0.011	0.548	0.808	0.722	0.939	1.000	-180.794
just Probit	0.007	0.553	0.821	0.729	0.953	1.000	-179.815

Figure 4.7: Simulated UDC, low correlations - Comp. with Market Shares

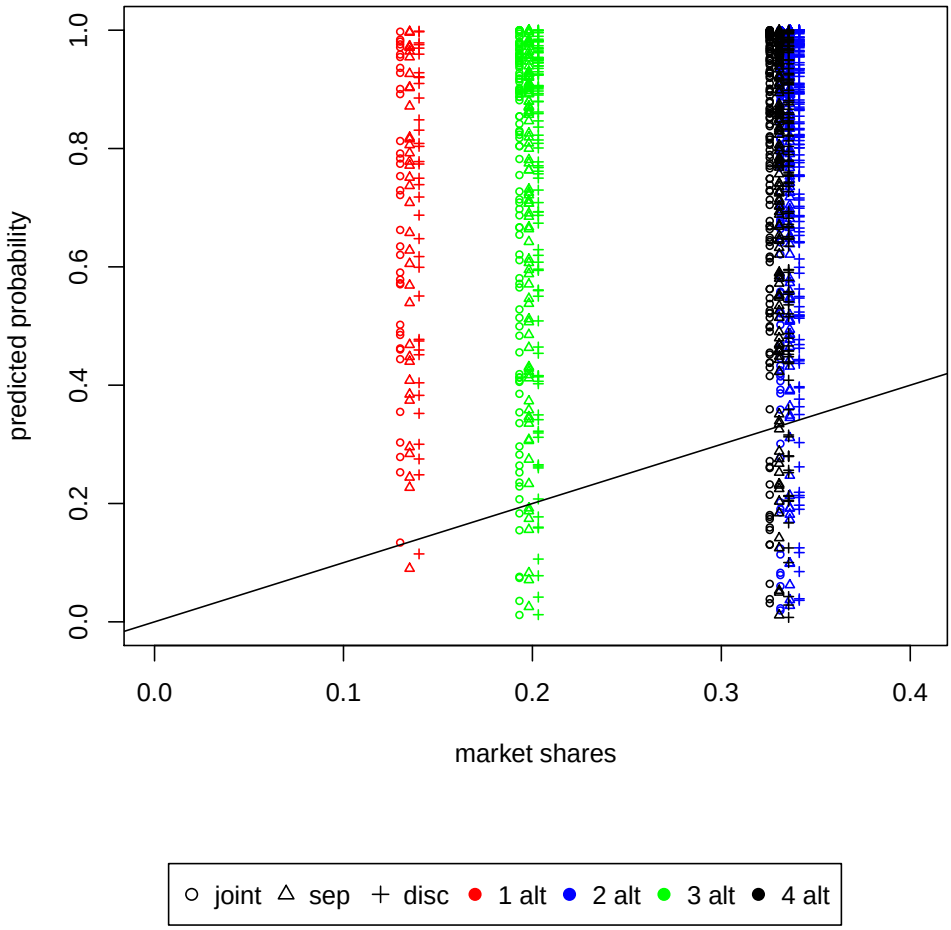


Figure 4.8: Simulated UDC, low correlations - Comp. with OP

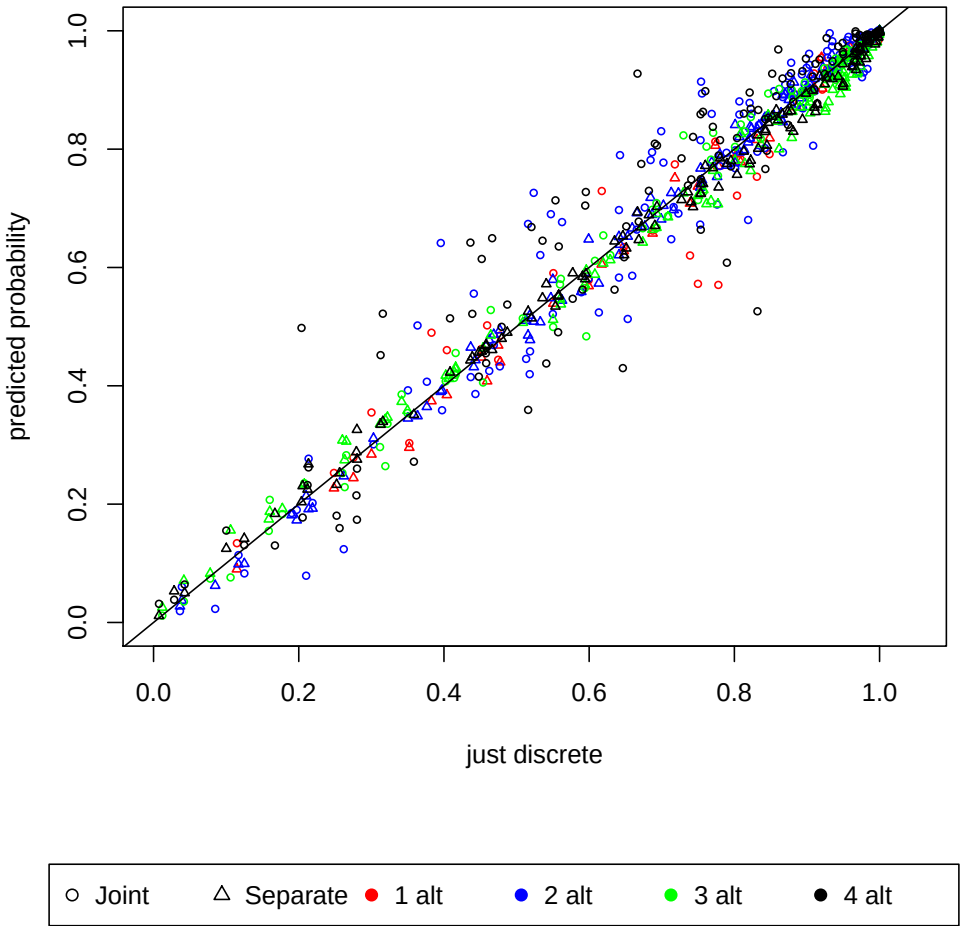


Figure 4.9: Simulated UDC, low correlations - Regression

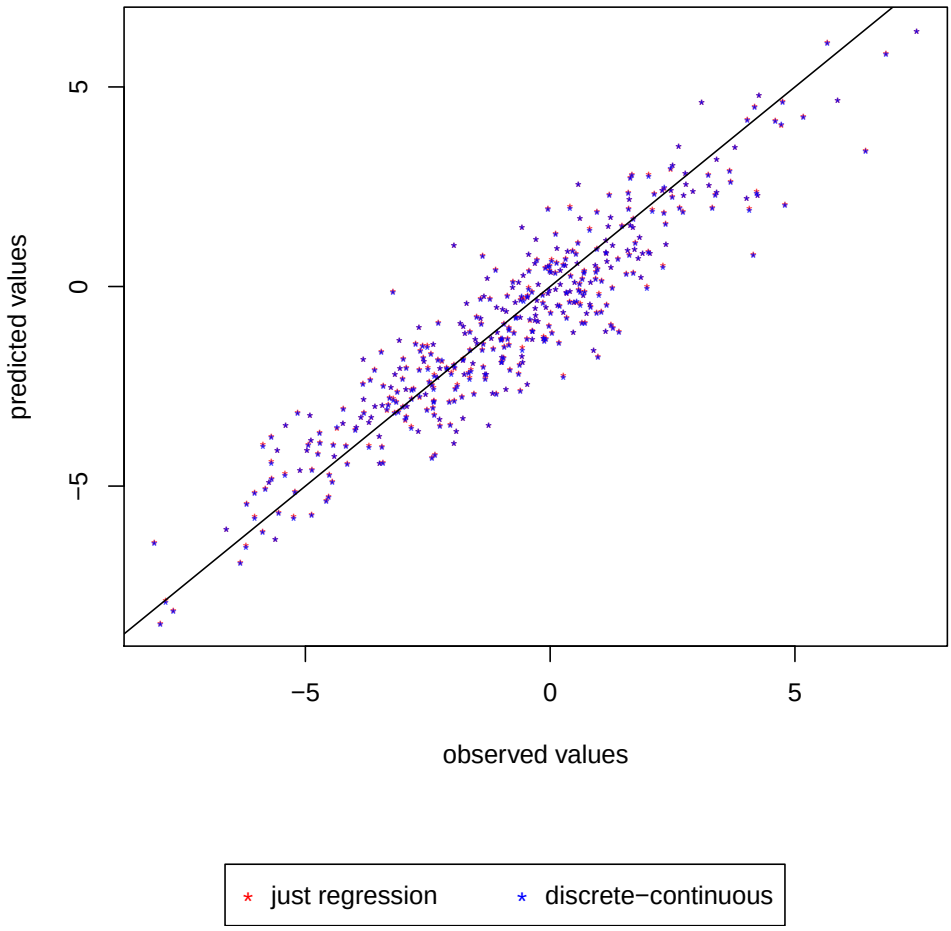


Table 4.16: Simulated UDC, low correlations - UDC coefficients for validation sample

name	true param.	estimate
X_1	1.0000	1.2233
X_2	-1.0000	-1.1565
const	0.5000	0.7500
X_3	2.0000	2.0494
X_4	-2.0000	-2.1669
const	-0.5000	-0.6279
X_5	-2.0000	-2.3091
X_6	2.0000	2.2953
const	1.0000	1.1714
X_7	-1.0000	-1.0469
X_8	1.0000	1.1616
const	-1.0000	-1.0013
X_9	1.0000	1.0057
X_{10}	-1.0000	-0.9749
X_1	2.0000	1.9979
X_2	-0.5000	-0.4525
L_{21}	0.2500	0.0207
L_{22}	1.0625	1.2756
L_{31}	0.0000	0.0755
L_{32}	0.2500	0.4824
L_{33}	1.0625	1.0786
L_{41}	-0.2000	-0.1953
L_{42}	-0.0500	0.0744
L_{43}	0.2000	0.2825
L_{44}	1.0800	0.9938
maxLL		-3099.2520
n		1600.0000

Table 4.18: Simulated UDC, low correlation - Probit coefficients for validation sample

name	true param.	estimate
X_1	1.0000	1.0383
X_2	-1.0000	-0.9861
const	0.5000	0.5816
X_3	2.0000	1.8296
X_4	-2.0000	-1.9304
const	-0.5000	-0.5360
X_5	-2.0000	-1.9665
X_6	2.0000	1.9608
const	1.0000	0.9898
X_7	-1.0000	-0.8733
X_8	1.0000	0.9891
L_{21}	0.2500	0.0492
L_{22}	1.0625	1.0421
L_{31}	0.0000	0.0578
L_{32}	0.2500	0.3195
L_{33}	1.0625	0.9091
maxLL		-768.8513
n		1600.0000

Table 4.20: Simulated UDC, low correlation - Regression coefficients for validation sample

name	true param.	estimate	SD	t	p
const	-1.0000	-0.9862	0.0264	-37.4100	0.000
X_9	1.0000	0.9979	0.0255	39.1600	0.000
X_{10}	-1.0000	-0.9704	0.0257	-37.7300	0.000
X_1	2.0000	1.9969	0.0270	74.1000	0.000
X_2	-0.5000	-0.4651	0.0262	-17.7800	0.000
sigma	1.0400	1.0540			
adj R2		0.8696			
n		1600.0000			

4.3.3 UDC validation results: high correlation

Figures 4.22 through 4.27 report the validation results with high covariance elements. Figure 4.22 summarizes the predicted probabilities. This time, all quartiles and the mean are significantly higher for the joint model. The validation log-likelihood is also much higher (-144 vs -204). However, the log-likelihood for the unconditional discrete-continuous model is lower than the log-likelihood of probit alone. We think that the model estimation may be somehow shifted in the context of a discrete-continuous model. This may be related to the apparent bias is the estimation of the coefficient for the joint model, that we do not observe for the probit alone. Tables 4.23 and 4.25 tend to confirm this hypothesis since the probit coefficients are less biased than the corresponding coefficients of the discrete-continuous model. Also we note that the high correlation appears to increase this apparent bias. Again we suggest that the covariance structure might be weakening the model identification but this requires more research.

Figures 4.10 and 4.11 are basically a more extreme version of their low correlation equivalent. Unconditional probabilities represent by triangle are still distributed along the identity line. Conditional probabilities represented by circles are further from the identity. Most circles are above the line but some are under it.

Table 4.22: UDC Validation - high correlations

model	min.	1^{st} q.	median	mean	3^{rd}q.	max.	Log-likelihood
joint UDC	0.024	0.611	0.910	0.779	0.991	1.000	-144.3271
separate UDC	0.000	0.531	0.825	0.716	0.958	1.000	-204.4712
just Probit	0.001	0.531	0.840	0.725	0.964	1.000	-192.7428

Figure 4.10: Simulated UDC, high correlations - Comp. with Market Shares

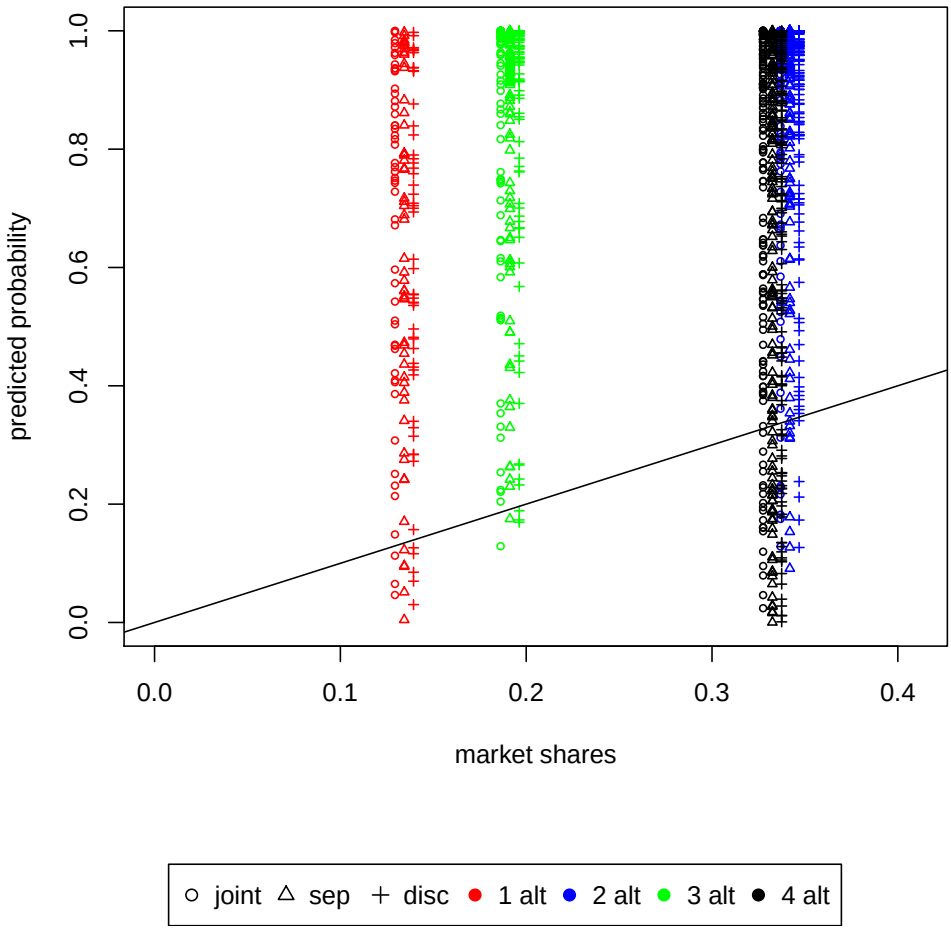


Figure 4.11: Simulated UDC, high correlations - Comp. with OP

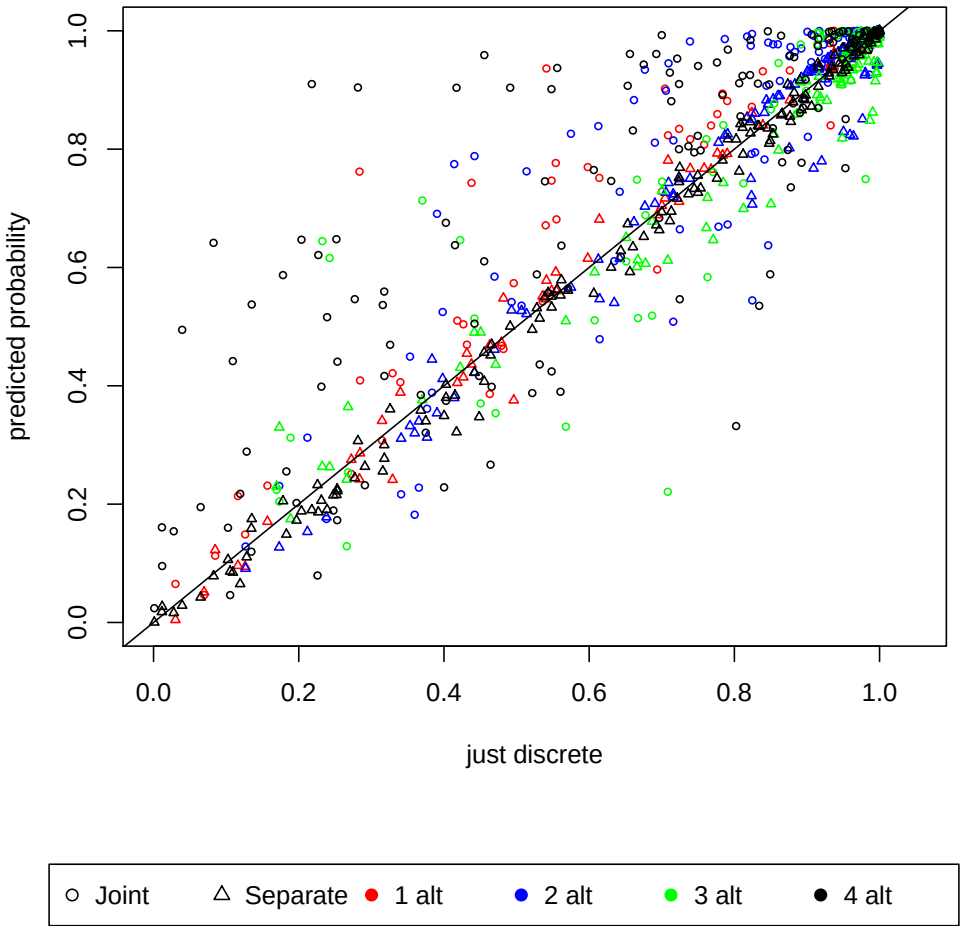


Figure 4.12: Simulated UDC, high correlations - Regression

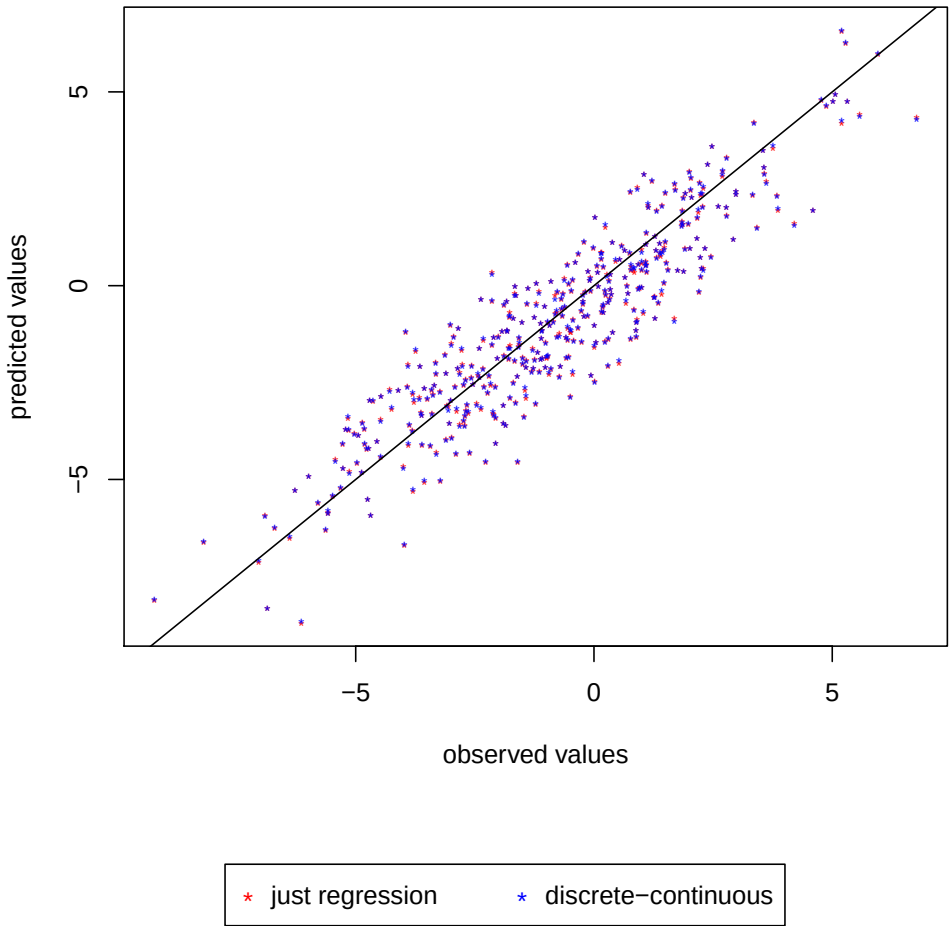


Table 4.23: Simulated UDC, high correlations - UDC coefficients for validation sample

name	true param.	estimate
X_1	1.0000	1.3416
X_2	-1.0000	-1.2953
const	0.5000	0.9029
X_3	2.0000	2.4853
X_4	-2.0000	-2.3999
const	-0.5000	-0.7895
X_5	-2.0000	-2.7010
X_6	2.0000	2.6898
const	1.0000	1.1432
X_7	-1.0000	-1.4073
X_8	1.0000	1.5180
const	-1.0000	-1.0078
X_9	1.0000	1.0315
X_{10}	-1.0000	-1.0001
X_1	2.0000	2.0515
X_2	-0.5000	-0.4667
L_{21}	0.8000	0.9875
L_{22}	1.0000	0.7871
L_{31}	-0.3000	-0.2956
L_{32}	0.2400	0.8750
L_{33}	0.9800	1.1284
L_{41}	-0.6000	-0.6742
L_{42}	-0.4800	-0.0443
L_{43}	0.4800	0.3976
L_{44}	1.0800	0.7264
maxLL		-2949.9730
n		1600.0000

Table 4.25: Simulated UDC, high correlation - Probit coefficients for validation sample

name	true param.	estimate
X_1	1.0000	1.1703
X_2	-1.0000	-1.0785
const	0.5000	0.7951
X_3	2.0000	2.1475
X_4	-2.0000	-2.0838
const	-0.5000	-0.7265
X_5	-2.0000	-2.3638
X_6	2.0000	2.4109
const	1.0000	1.0276
X_7	-1.0000	-1.1679
X_8	1.0000	1.2878
L_{21}	0.2500	1.0586
L_{22}	1.0625	0.6719
L_{31}	0.0000	-0.2334
L_{32}	0.2500	1.0792
L_{33}	1.0625	0.5383
maxLL		-691.1823
n		1600.0000

Table 4.27: Simulated UDC, high correlation - Regression coefficients for validation sample

name	true param.	estimate	SD	t	p
const	-1.0000	-1.0111	0.0268	-37.6900	0.000
X_9	1.0000	1.0112	0.0259	39.0000	0.000
X_{10}	-1.0000	-0.9996	0.0262	-38.1800	0.000
X_1	2.0000	2.0660	0.0274	75.3400	0.000
X_2	-0.5000	-0.4514	0.0266	-16.9600	0.000
sigma	1.0400	1.0720			
adj R2		0.8718			
n		1600.0000			

4.4 Real data: discrete-continuous model validation

Predicting a discrete variable can be done in several ways. It is not possible to assume that the chosen alternative is simply the alternative with the highest probability. To illustrate this, imagine that we want to predict if people will use their car or public transit for their daily commute to work. If the model were to assign a probability of using a car between 70% and 100% and, correspondingly, a probability between 0% and 30% of using public transit, then the most likely alternative would be "car" for 100% of observations. This result does not reflect the actual market share. Also, if we were to predict a rare choice, such as using a bicycle to commute to work, it may be important to identify who are the individuals that are the most likely to make this choice, even if every single decision maker still has a higher probability of not using a bicycle. For the reasons explained above the most used approach is sample enumeration, by which the choice probabilities of each decision maker in a sample are averaged over decision makers.

In a regular regression model, the dependent variable will have fitted values after the model has been calibrated. There are a number of issues that may affect the validity of these fitted values. For example model estimation, goodness of fit or model selection can be challenging. However, once these issues have been addressed, there will be a unique numerical fitted value for the dependent variable. In a regression we would use a F-test to test if the regression model is significantly better than simply using the mean of the dependent variable as a predictor. We argue that we can do a similar analysis by comparing the probabilities predicted by the

model with the sample market shares. If, for example, we have 80% of our sample who use their car and 20% who use public transit, then we want a model that will predict the transit alternative with a probability *more* than 20 % for at least some observations. Otherwise the model is not more useful than just using market shares, and therefore the model is not useful.

In joint discrete-continuous models two two dependent variables, (one discrete and one continuous) are predicted simultaneously. The validation procedure suggested in our work suppose that if the probability of the chosen alternative as predicted by the joint model is higher than the one predicted by the regular model, then we conclude that the more complicated model is more useful, otherwise we should favor the simpler model.

4.4.1 Discrete-continuous car ownership model - NHTS 2009

The NHTS 2009 dataset has been divided into a calibration sample containing 80% of the total observations and a validation sample that has the remaining observations (20% of the total). Coefficients' estimates for the ODC model are reported in 4.30; the coefficients of the OP are in 4.32, and the predictor sif the regression in 4.34. We first perform a BIC variable selection for both the ODC and a separate ordered probit, and then we compute the predicted probability of the actual choice in the validation sample using three methods:

- Joint model using the regression for the conditional probability;
- Joint model ignoring the regression;

- Only the ordered probit.

Table 4.29 highlights some interesting results. First, the joint model slightly outperforms the regular model, as it predicts higher probabilities than the separate DC/regression and OP alone. The minimum and maximum predicted probabilities are practically zero and one for all three models. The three quartiles of predicted probabilities considered are higher for the conditional ODC model, but by just small quantities. For example the median predicted probability is 63%, which is 5% more than the one obtained by using a regular ordered probit. The mean predicted probability is also higher by about 5%.

In figure 4.13 we compare the predicted probabilities obtained with the ODC models to the market share; the line represents the identity line where predicted probabilities equals the market share. It is worth noting that each observation in the validation sample is reported three times in the figure; one for each model estimated (joint, sep, disc). Ideally, all points should be above the identity line, which will indicate that every choice is predicted better than with just the market share.

Figure 4.14 reports these predicted probabilities against the ones obtained by using just an ordered probit alone. The line is still the identity line. The ideal scenario would be to have predicted probabilities consistently above the identity line; in that case the predicted probability of the joint model would always be superior to the base model. However, as it can be seen from 4.14 this is not always the case.

Table 4.29: 2009 - ODC

model	min.	1st q.	median	mean	3rd q.	max.	Log-likelihood
joint ODC	0.008	0.419	0.632	0.574	0.755	0.999	-209.5722
separate ODC	0.004	0.365	0.557	0.506	0.643	0.998	-239.648
just OP	0.001	0.362	0.585	0.527	0.695	0.999	-238.669

Figure 4.13: ODC 2009 - Validation against market shares

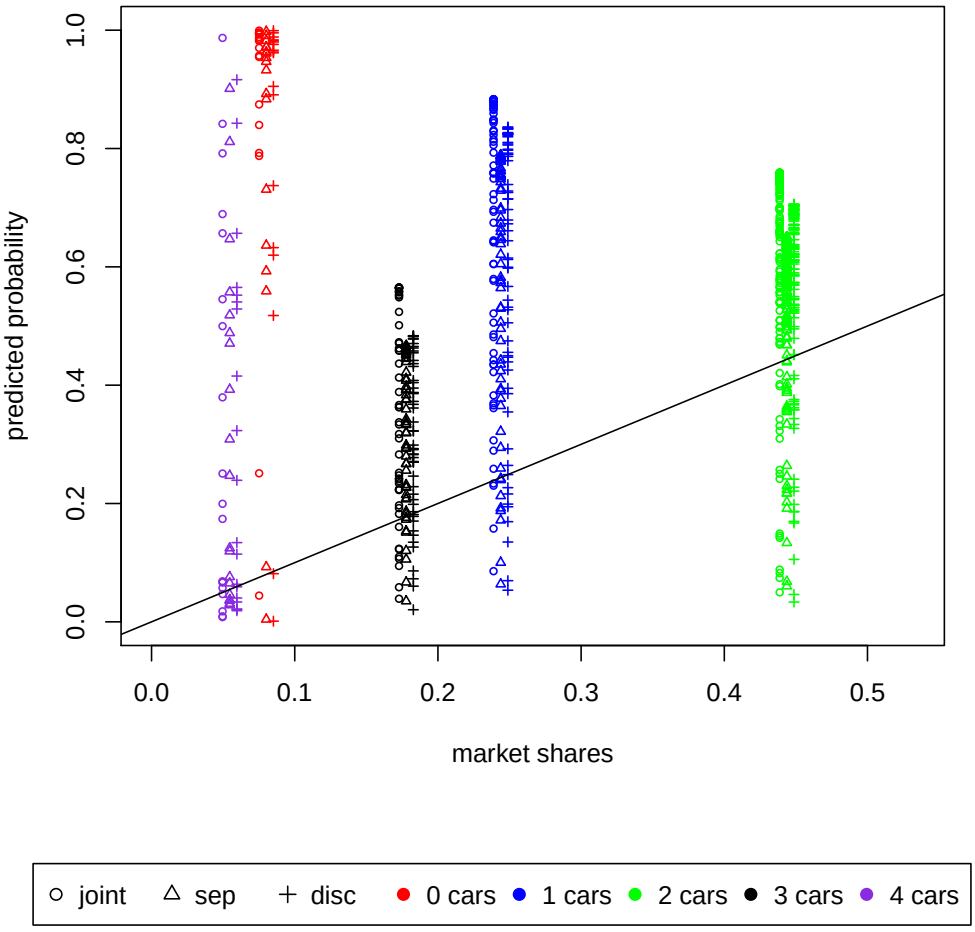


Figure 4.14: ODC 2009 - Validation against ordered Probit

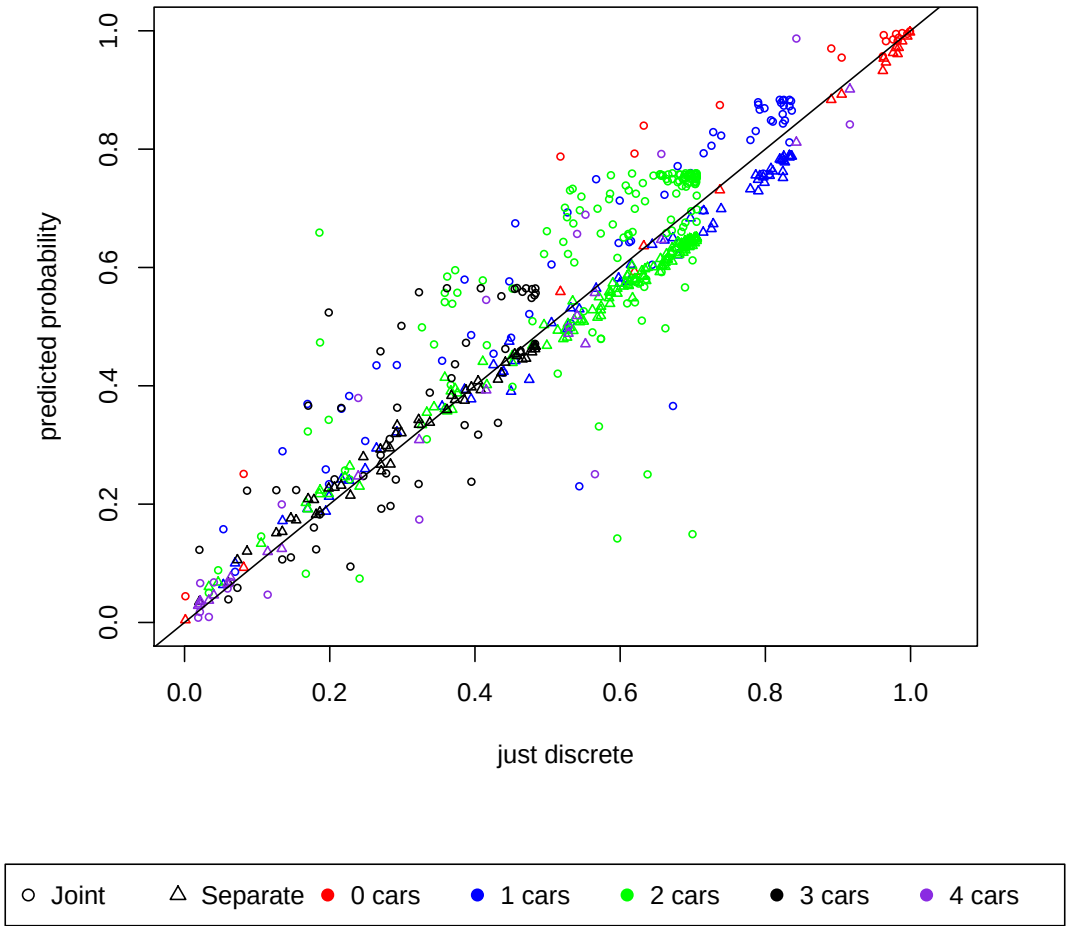


Figure 4.15: ODC 2009 - Validation of regression

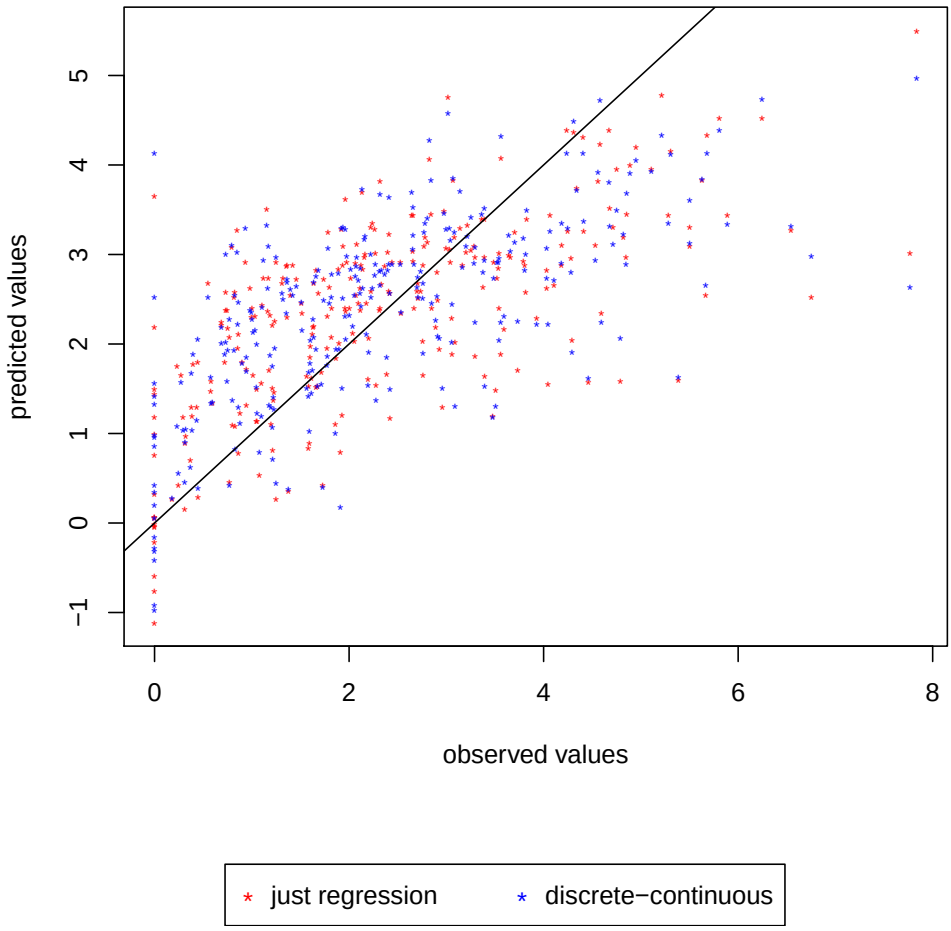


Table 4.30: 2009 NHTS - ODC coefficients

name	estimate	SD	t	p
`const`	-1.734	0.283	-6.125	0.000
driver count	0.885	0.089	9.952	0.000
cost per mile	15.018	1.095	13.709	0.000
density	0.000	0.000	-11.787	0.000
age	-0.002	0.003	-0.829	0.407
own home	0.614	0.113	5.412	0.000
income	0.054	0.008	6.721	0.000
num adults	0.282	0.095	2.963	0.003
transit	-0.368	0.137	-2.679	0.007
urban	-0.140	0.102	-1.363	0.173
gender	-0.141	0.061	-2.313	0.021
urban size	-0.019	0.021	-0.885	0.376
num workers	0.056	0.051	1.101	0.271
α_1	2.503	0.160	15.616	0.000
α_2	1.871	0.075	25.052	0.000
α_3	1.246	0.077	16.098	0.000
`const`	1.522	0.320	4.756	0.000
age	-0.018	0.004	-5.239	0.000
cost per mile	0.196	1.067	0.183	0.855
urban size	-0.076	0.028	-2.746	0.006
num workers	0.221	0.069	3.216	0.001
own home	0.501	0.135	3.706	0.000
density	0.000	0.000	-3.875	0.000
income	0.047	0.011	4.103	0.000
urban	-0.232	0.137	-1.697	0.090
transit	-0.312	0.173	-1.805	0.071
num drivers	0.730	0.108	6.785	0.000
num adults	-0.033	0.120	-0.277	0.782
education	-0.064	0.039	-1.623	0.105
σ^2	2.090	0.087	23.962	0.000
ρ	0.450	0.022	20.266	0.000
max LL	-2836.070			
n	1136.000			

Table 4.32: 2009 NHTS - Ordered Probit coefficients

name	estimate	SD	t	p
`const`	-1.662	0.319	-5.219	0.000
num drivers	1.001	0.097	10.276	0.000
density	0.000	0.000	-3.135	0.002
cost per mile	16.598	1.187	13.981	0.000
income	0.065	0.010	6.499	0.000
own home	0.706	0.125	5.628	0.000
num adults	0.275	0.106	2.606	0.009
transit	-0.367	0.153	-2.402	0.016
urban	-0.159	0.112	-1.415	0.157
gender	-0.213	0.074	-2.886	0.004
age	-0.004	0.003	-1.438	0.150
education	-0.050	0.037	-1.365	0.172
density	0.000	0.000	-1.097	0.273
urban size	-0.021	0.024	-0.841	0.401
num workers	0.056	0.056	0.993	0.321
α_1	2.787	0.173	16.118	0.000
α_2	2.097	0.080	26.181	0.000
α_3	1.298	0.082	15.923	0.000
max LL	-906.556			
n	1136.000			

Table 4.34: 2009 NHTS - Regression coefficients

name	estimate	SD	t	p
driver count	0.747	0.064	11.663	0.000
density	0.000	0.000	-4.446	0.000
income	0.054	0.010	5.530	0.000
num workers	0.334	0.064	5.200	0.000
urban size	-0.094	0.023	-4.154	0.000
own home	0.422	0.118	3.564	0.000
σ	1.477			
adj R2	0.744			

The UDC version of this analysis does not favor joint models. First, table 4.36 illustrates that the probit model outperforms the joint model. This appears to be partly due to failed convergence in the estimation of the UDC. It is still the case however that conditional probabilities of the UDC are higher for all quartiles and for mean versus their unconditional counterpart.

Figures 4.16 and 4.18 provide more details about the predicted probabilities. Unfortunately, UDC predicted probabilities appear to be distributed at random, which is far from ideal. We can see that the green point that represent the 2 cars alternative are consistently under predicted by the joint model, which may indicate that the the component if this alternative's utility are either poorly estimated, or poorly specified. On the other side, the 1 and 4 cars alternative are still predicted better with the joint model, although it is not clear whether the circles, representing the conditional probabilities, offer a substantial improvement.

The coefficients used for this validation are reported in tables 4.37 through 4.40.

Table 4.36: 2009 - UDC

model	min.	1^{st} q.	median	mean	3^{rd}q.	max.	Log-likelihood
joint UDC	0.000	0.250	0.363	0.390	0.517	0.896	-348.752
separate UDC	0.000	0.268	0.321	0.342	0.388	0.765	-367.763
just probit	0.000	0.335	0.569	0.481	0.637	0.973	-279.552

Figure 4.16: UDC 2009 - Validation against market share

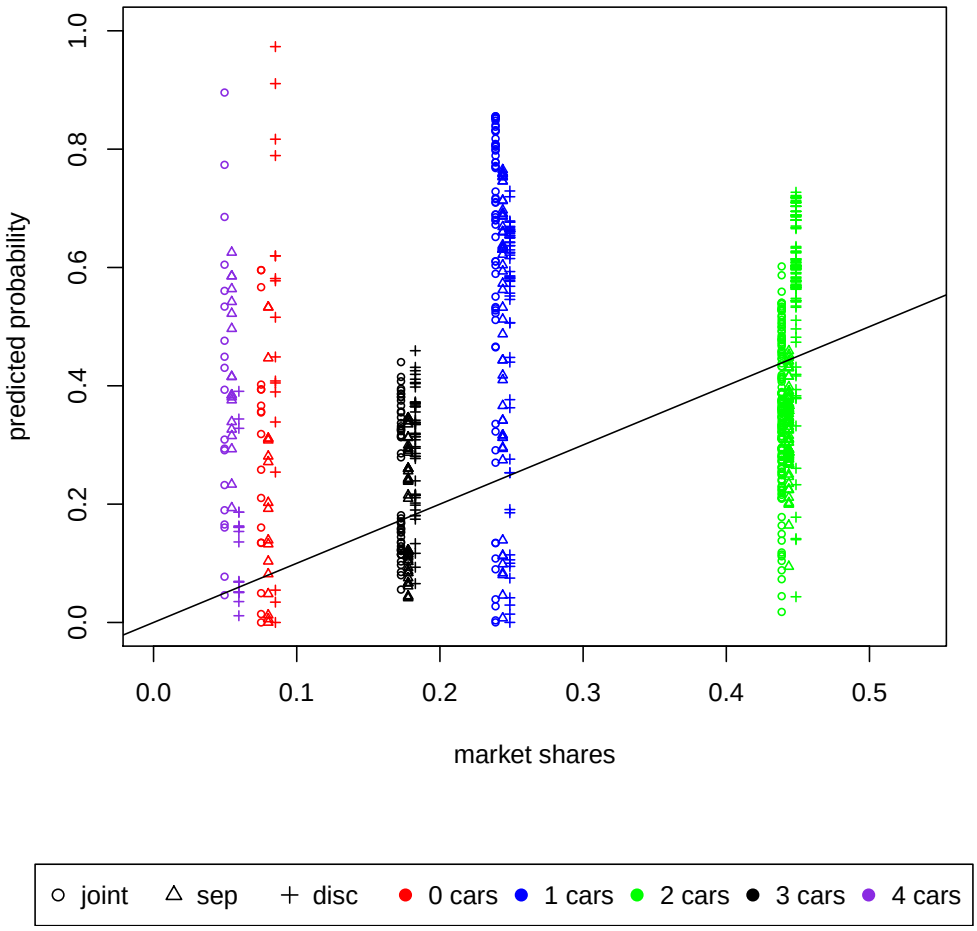


Figure 4.17: UDC 2009 - Validation against Probit

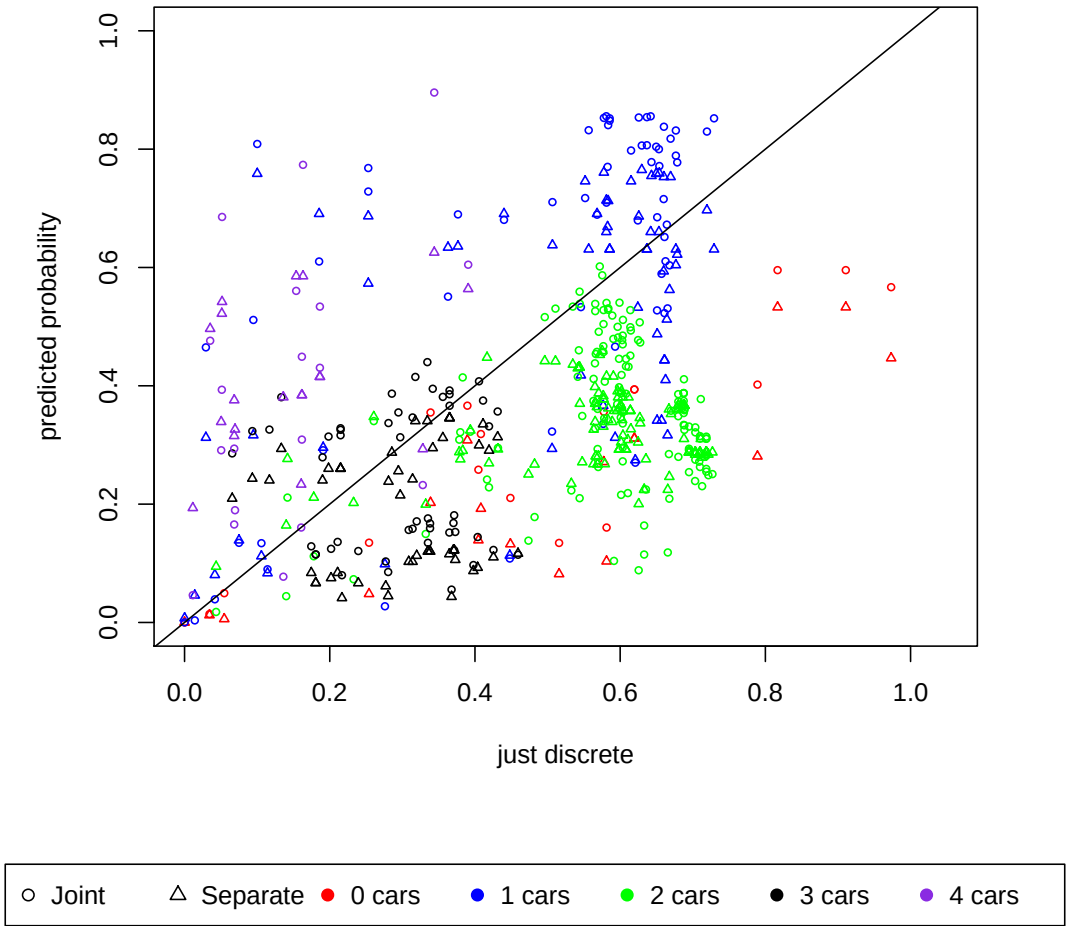


Figure 4.18: UDC 2009 - Validation of regression

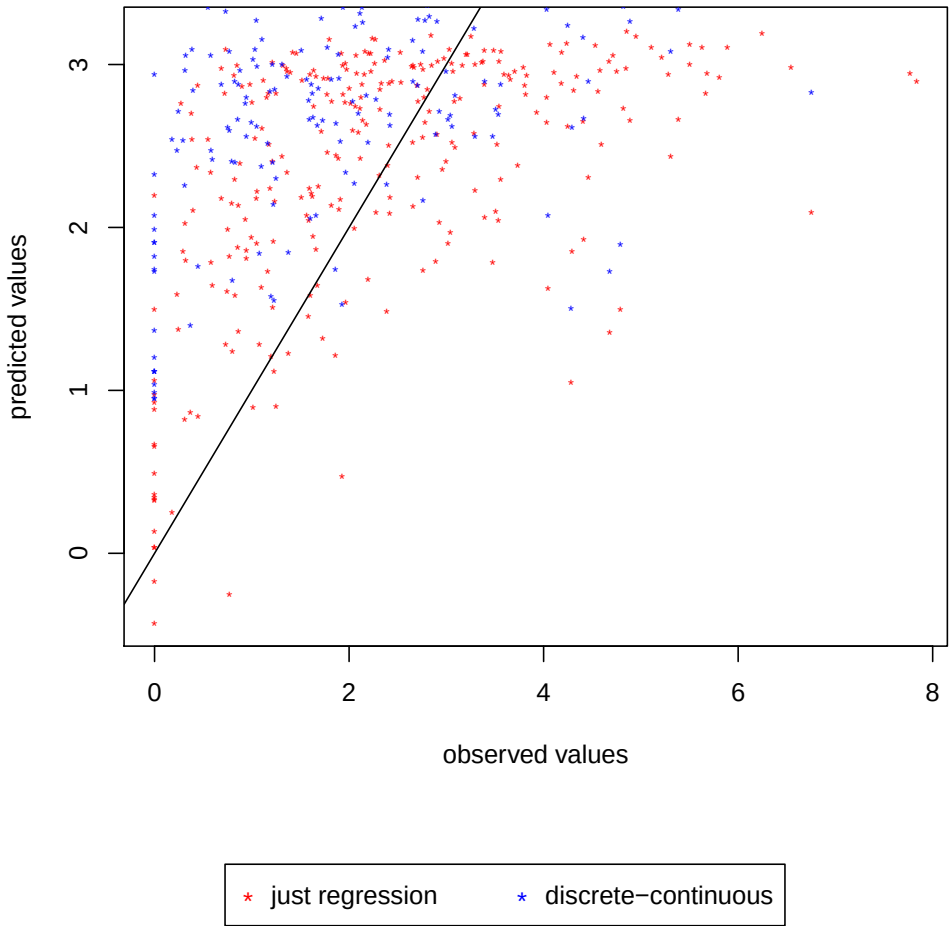


Table 4.37: 2009 NHTS - UDC coefficients

name	estimate
const	1.315
income	0.106
num drivers	0.302
gender	-0.242
urban size	-0.212
density	0.000
const	-2.543
income	0.285
num drivers	2.449
gender	0.255
urban size	-0.944
density	0.000
const	-7.547
income	0.564
num drivers	2.536
gender	-4.045
urban size	-1.244
density	0.000
const	-4.887
income	0.438
num drivers	2.915
gender	0.775
urban size	-1.440
density	0.000
const	0.909
income	0.083
own home	0.791
gender	-0.035
density	0.000
cost per mile	2.554
L_{21}	3.148
L_{22}	0.173
L_{31}	-3.759
L_{32}	-3.091
L_{33}	2.322
L_{41}	1.943
L_{42}	-1.306
L_{43}	2.117
L_{44}	-0.751
L_{51}	0.448
L_{52}	-0.894
L_{53}	0.196
L_{54}	-0.851
L_{55}	1.130
max LL	-3654.819
n	1136.000

Table 4.38: 2009 NHTS - Probit coefficients

name	estimate
const	0.224
income	0.109
num drivers	0.823
gender	-0.481
urban size	-0.106
density	0.000
const	-18.882
income	0.796
num drivers	11.923
gender	-3.564
urban size	-0.594
density	-0.001
const	-30.443
income	0.911
num drivers	15.916
gender	-3.533
urban size	-1.418
density	-0.001
const	-27.793
income	0.852
num drivers	15.454
gender	-3.709
urban size	-1.114
density	-0.002
L_{21}	8.464
L_{22}	4.471
L_{31}	0.218
L_{32}	3.950
L_{33}	6.264
L_{41}	2.567
L_{42}	4.395
L_{43}	3.408
L_{44}	-0.027
max LL	-1059.813
n	1136.000

Table 4.40: 2009 NHTS - Regression coefficients

name	estimate	SD	t	p
`const`	0.505	0.217	2.324	0.020
income	0.102	0.010	10.416	0.000
own home	0.429	0.148	2.903	0.004
gender	-0.146	0.098	-1.483	0.138
density	0.000	0.000	-7.985	0.000
cost per mile	2.619	1.173	2.234	0.026
σ	1.628			
adj R2	0.689			

4.5 Conclusions

In this chapter, we have argued that using actual choice probabilities is a good way to assess the predictive power of a discrete-continuous model. We have seen, using simulated samples that the conditional predictions offer a significant improvement over unconditional predictions when there is a high correlation between the error terms. Using the 2009 NHTS data, we have shown that ordered discrete-continuous models offer a modest improvement in predicting vehicle holding along with vehicle miles traveled. Both conditional and unconditional predicted probabilities using ordered models greatly outperform their unordered part. For instance, the median prediction using ODC is 63%, approximately twice as much as the with the UDC. The results obtained from just the probit model, whose prediction compare with the ordered probit and ODC, illustrate that optimal values of UDC have probably not been reached in the maximization process. Altogether, we have illustrated that unordered models pose great challenges and suffer from competition with ordered ones for this application.

Chapter 5: Random Effect Models for Free-Flow speed estimation

5.1 Introduction

Speed is a fundamental variable in the geometric design of highways and streets. Transportation engineers normally refer to the design speed to calculate the characteristics of geometric elements of the road, and to the operating speed to assess the consistency of the adopted design values along the designed road alignment. Hence, operating speed models are fundamental in road design since they can anticipate the speeds that will be adopted by drivers.

In this Chapter, we present results obtained from the estimation of free-flow speed on two-lane rural highways. The model structure adopted separates the estimate of the central tendency of speeds from the typical deviations of individual speeds. In the model the same set of variables can be used to evaluate the mean value and the standard deviation of the speed distribution; the desired speed percentile is then calculated considering the associated standard normal random variable (Z). Fixed effect (FE) models are also calibrated for comparison purposes and the (Bayesian Information Criterion) BIC criterion is used for variable selection and applied to both the FE and RE models. A summary of this work has recently been published in a peer-reviewed journal [BCMnt].

5.2 Literature review

Operating speeds reflect the speed behavior of drivers who are affected by the horizontal and vertical alignments as well as the cross section. As a consequence, operating speed data are collected from isolated vehicles moving in free-flow conditions. Free-flow speeds are generally normally distributed as indicated in many contributions [Tra11] [BDMC14]. Usually, to assess if speed design consistency has been achieved, the 85th percentile of the distribution (V85) is considered [Has04] [MFT05], since it conventionally separates the population of prudent drivers from the small group of more aggressive drivers. Despite this widespread approach, some commentators contend that a knowledge of the parameters describing the entire distribution is more powerful and useful for applications and inferences [Tra11] [MFT05] [Bon01].

Figueroa and Tarko [MFT05] emphasized this concept with the following given example: of two different distributions, the first with a low mean speed (50 km/h) and high standard deviation (15 km/h), the second with high mean speed (60 km/h) and a low standard deviation (5 km/h), that have the same V85 (65 km/h). Hence, the V85 alone is not able to provide a comprehensive interpretation of the effect of road geometrics on operating speeds.

Several variables influence the operating speed. If on urban streets cross sectional and environmental variables seem to be more significant in modeling [WDLH06] [BM13], on rural highways the characteristics describing the horizontal and vertical alignments are found to be significant in a number of scientific and technical con-

tributions [Tra11] [MFT05] [Tra03]. Despite the fact that a considerable amount of research on operating speeds has been produced in the last twenty years, more recent works have aimed at improving model predictions and at extending the spectrum of road typologies to those types that have not been fully investigated.

In July 2011 the Operational Effects of Geometrics Committee of the Transportation Research Board sponsored the publication of the E-C 151 Circular [Tra11]. This document included some important remarks regarding the current level of research on the topic, and some criticisms and suggestions were made to improve model applicability and speed predictability.

According to the E-C 151 Circular [Tra11], the majority of the models available in literature can predict the V85th percentile speeds for cars on horizontal curves, assuming they (the cars) maintain a constant speed throughout. Only a few models predict truck speeds, and speeds on tangents; some can estimate speed variations approaching and exiting the curves, and a few consider the possibility of variation in speeds within a curve; only one model [MFT05] allows for the evaluation of the entire speed distribution.

In the E-C 151 Circular [Tra11], some remarks were also made on variables: many models contain the parameters describing the horizontal curvature, while few include variables related to the vertical alignment (i.e., grade), tangent, horizontal–vertical combinations, cross-section elements (i.e., width of lanes), available sight distance, and posted speeds.

Starting from this state of the art, this research aims at the calibration of linear regression models able to predict any speed percentile. The proposed methodology

takes into account possible random effects caused by the structure of the data where sections are randomly selected from roads forming part of the two-lane highway network in the Northwest of Italy.

5.3 Speed database

The speed database was assembled at various stages from 2005 to 2011 with data from several road sections in the provinces of Turin, Vercelli and Alessandria (Italy). A series of observational surveys was carried out on several typologies of rural roads including freeways, multi-lane and two-lane highways. In this investigation, only the sub-database of two-lane rural roads was considered for model calibration. Speed data were collected from a total of 13 roads and 37 sections with the final database containing 6,567 speed observations, which were eventually used to calibrate the speed models. During the surveys, no data were measured in one lane, thus leading to 73 individual lanes investigated. Table 5.1 lists the identification code for roads and sections, the name, the length of the road, and the main geometric characteristics of sections. In particular, the table reports the lane width, the radius of the centerline (the symbol ∞ denotes tangent sections) and the average longitudinal grade across the sections. All the geometric characteristics were derived from regional GIS databases. For the some roads, several cross sections were selected when differences in the geometric characteristics and/or margin treatments were observed. The minimum distance between the closest sections of the same road was set equal to 2 km. Surveys were carried out at different time periods; thus it

is reasonable to assume that the same vehicles were not surveyed multiple times at different road sections.

Table 5.1: Geometric and operative characteristics of the selected road sections

Road #	**Section** #	**Road name**	**Road length** km	**Lane width** m	**Radius** m	**Grade** ±%	V_{min} km/h	V_{max} km/h	V_{85} km/h	**PSL** km/h	n_{obs}
1	1	SP70-VC	3.830	3.60	∞	1.50	39.0	97.0	76.0	70	429
2	2	SP8-VC	7.790	3.70	178.47	3.00	45.0	128.0	87.0	90	618
	3			3.75	∞	0.00	32.0	157.0	85.0	50	972
3	4	SP299-VC	57.250	3.60	334.57	0.50	45.0	128.0	87.0	90	799
	5			3.70	∞	1.00	38.0	130.0	89.0	90	669
4	6	SS460-TO	61.757	3.75	1000.00	1.44	57.0	122.6	93.2	70	192
	7			3.75	304.00	5.14	24	114.0	94	70	312
5	8	SS565-TO	18.180	3.75	∞	2.09	57.0	124.0	98.0	70	101
	9			3.75	3226.00	4.69	52.0	114.0	91.0	70	101
	10			3.25	150.00	8.50	46.0	77.0	70.1	50	87
	11			3.00	∞	1.50	42.6	129.4	90.3	70	107
	12			3.00	∞	0.00	46.7	128.6	94.3	70	120
	13			3.00	∞	0.00	50.0	114.0	90.4	70	108
6	14	SP55-AL	18.180	3.00	∞	2.00	49.6	129.4	101.7	70	128
	15			3.00	∞	0.00	49.0	127.8	95.9	70	127
	16			3.00	∞	0.00	46.3	128.3	98.9	70	138
	17			3.00	∞	0.00	49.4	149.6	95.3	70	128
	18			3.50	∞	0.50	42.0	115.2	88.5	90	41
	19			3.50	∞	0.50	73.0	130.0	106.6	90	29
	20			3.50	2250.00	0.50	70.0	118.0	107.0	90	26
	21			3.50	∞	0.50	58.0	127.0	112.9	90	28
	22			3.50	∞	0.00	55.0	130.0	100.3	90	30
	23			3.50	1502.00	1.00	32.0	72.0	57.4	50	32
	24			3.50	∞	0.00	59.0	110.0	92.6	70	27
7	25	SP230-VC	39.466	3.50	909	1.00	44.0	96.0	78.0	70	37
	26			3.50	8351.25	0.00	47.0	98.0	85.5	70	38
	27			3.50	452.00	0.00	44.0	70.0	57.5	50	36
	28			3.50	1503.00	1.00	56.0	130.0	97.0	90	43
	29			3.50	∞	1.00	65.0	119.5	107.9	90	38
	30			3.50	∞	0.00	54.7	108.1	85.1	70	42
	31			3.50	∞	0.00	44.3	120.2	83.6	70	140
8	32	SP177-TO	10.927	3.20	300.00	2.50	27.6	93.3	75.8	70	116
9	33	SP176-TO	4.961	3.80	∞	0.00	42.6	106.6	75.3	90	67
10	34	SP267-TO	8.839	3.50	∞	0.00	34.7	105.4	68.5	50	154
11	35	SP220-TO	4.045	3.80	∞	0.00	44.2	89.1	79.8	70	100
12	36	SP183-TO	2.156	3.50	∞	0.00	36.9	134.8	82.5	50	161
13	37	SP23-TO	92.223	3.50	550.00	0.50	36.4	192.1	102.1	90	249

Speed data were collected using two different techniques: longitudinal measurement by means of a laser speed gun, and cross-sectional measurement by means of a digital video camera positioned perpendicularly to the road axis. The type of technique was decided on the basis of the characteristics of the survey site, and the observation points were carefully selected so as to minimize any disturbance to traffic and to avoid any change in driver behavior.

To merge the data coming from the longitudinal and transversal surveys into the same dataset, a preliminary comparison of speeds detected with the use of the two devices was carried out. In particular, differences in speed distribution ascrib-

able to the precision and accuracy of the two measurement systems were carefully checked. A comparison of the speeds detected for the same vehicle with the two methods is reported in figure 5.1. All the data are included between the 10% lines across the equality line. The high coefficient of determination and the small standard error confirm the possibility of using the two datasets jointly to calibrate the same speed model. Speed data were collected under free-flow conditions assuming a minimum headway of 6 seconds, and at points where drivers assumed stationary speeds, hence along tangents in sections far from curves, and in sections located at the center of curves. Each observation included in the database was associated with first the corresponding percentile p and then the standardized normal variable Zp, derived from the average and standard deviation calculated on the sub-dataset of the lane (Zp = 0 when p = 50%, and Zp = 1.036 when p = 85%).

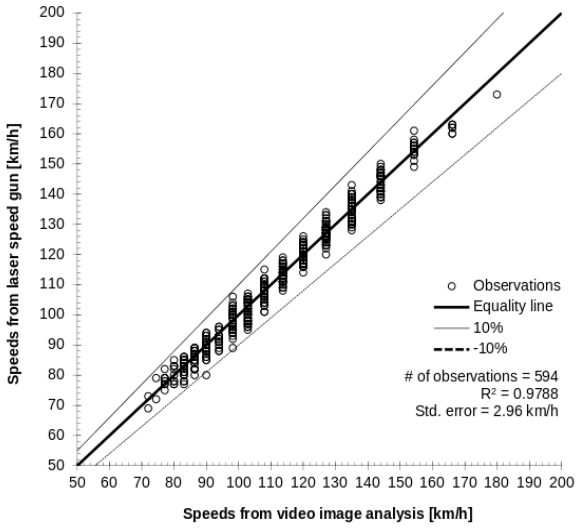
Field inspections were carried out to collect information on the geometric characteristics of the transversal section (i.e., lane and shoulder width, posted speed limit, as well as the presence of driveways, retaining walls, ramps and barriers). Such investigations were supported by the collection of the same information from aerial views on Google Earth . Regarding the curvature (corresponding to the inverse of the radius) and the longitudinal grade, data were obtained from GIS databases thanks to the cooperation of the public agencies that manage the road networks.

Each set of measurements was subjected to the chi-squared and the Kolmogorov Smirnov tests [CVC88] to check if data were normally distributed. In all cases the tests were successful. Table 5.1 also includes the minimum and the maximum speed values, the 85th percentile of observed speeds, and the posted speed

limit. Fifteen sections belonging to roads SP230-VC and SP177-TO had less than 100 speed observations, which is typically the minimum number of observations recommended in speed surveys. This was due to the very low traffic volumes that occurred during the surveys. In these cases, the goodness of fit tests demonstrated that in each direction the number of speed data were sufficient to form a normal distribution. Moreover, minimum and maximum speed ranges were wide enough; hence, reasonably, the recorded speed data were considered representative of the speed conditions in those sections.

The posted speed limit was between 50 and 90 km/h, the maximum grade of 8.5%. Fourteen of thirty-seven sections were on curves with radii ranging between 150 and 8351.25 m. The lane width was between 3 and 3.8 m. The road characteristics in Table 5.1 are representative of typical Italian two-lane rural highways, where the V85 is frequently observed to be above the posted speed limit (in the dataset this happens 32 times out of a total of 37 observations).

Figure 5.1: Comparison of operating speeds detected for the same vehicle from video image analysis and laser speed gun



Data were differentiated according to the direction of driving, since the same road feature may produce different effects depending on the side occupied with respect to the driving trajectory (right and closer, left and farther). Hence, all the elements located on the roadside have been considered two times for the two driving directions.

Table 5.2 contains the list of the thirty-four variables considered in the investigation with the raw statistics in order to understand the variability of each parameter, and consequently the field of validity for the models. Furthermore, the frequency, which counts the percentage of non-zero values, with which the variable is present in the database has also been reported. The variables included in Table 5.2 are a mix of numerical continuous, numerical discrete and Boolean, and are denoted in the table by the symbols NC, ND and B respectively. In the case of Boolean variables, 0 indicates that the element is absent, while the value 1 indicates that the element is present.

Some variables have been considered twice in an effort to understand if either their presence or density affects driver behavior. All the variables characterized by units expressed in No./km were estimated by summing the number of elements (i.e., ramps, driveways, intersections, and pedestrian crossings) in a section of 1 km across each investigated cross section. Throughout the surveys, particular attention was given to the roadside characteristics, since they are normally not taken into account in operating speed investigations on two-lane rural roads [Tra11].

For both the right (R) and left (L) sides, the presence and width of shoulders (SR, SL, SRW, and SLW respectively), the presence and density of ramps (RLS,

Table 5.2: Summarized raw statistics of considered variables

Variable	symbol	type	unit	min.	max.	μ	σ	Frequency	
Posted speed limit	PSL	ND	km/h	50	90	73.0	13.610	73	100%
Posted speed limit variation	ΔPSL	ND	km/h	0	20	0.5	3.287	2	3%
Lane width	LW	NC	m	3.0	3.8	3.5	0.260	73	100%
Curvature	1/R	NC	m-1	0	6.67e-3	8.28e-4	1.61e-3	28	38%
Longitudinal grade	LG	NC	%	−8.85	8.50	-0.012	2.110	41	56%
Shoulder right	SR	B	-	0	1	-	-	69	95%
Shoulder left	SL	B	-	0	1	-	-	69	95%
Shoulder right width	SRW	NC	m	0.0	1.5	0.9	0.437	69	95%
Shoulder left width	SLW	NC	m	0.0	1.5	0.9	0.437	69	95%
Ramp left side	RLS	B	-	0	1	-	-	5	7%
Ramp right side	RRS	B	-	0	1	-	-	5	7%
Ramp density left side	TRDLS	NC	km^{-1}	0.0	2.0	0.1	0.323	5	7%
Ramp density right side	TRDRS	NC	km^{-1}	0.0	2.0	0.1	0.323	5	7%
Driveways left side	DLS	B	-	0	1	-	-	55	75%
Driveways right side	DRS	B	-	0	1	-	-	55	75%
Driveway density left side	DDLS	NC	km^{-1}	0.0	8.0	2.3	2.321	55	75%
Driveway density right side	DDRS	NC	km^{-1}	0.0	8.0	2.3	2.325	55	75%
Intersections left side	ILS	B	-	0	1	-	-	39	53%
Intersections right side	IRS	B	-	0	1	-	-	39	53%
Intersection density left side	IDLS	NC	km^{-1}	0.0	5.0	1.1	1.362	39	53%
Intersection density right side	IDRS	NC	km^{-1}	0.0	5.0	1.1	1.362	39	53%
Stopping places Lay-by left side	SPLS	B	-	0	1	-	-	25	34%
Stopping places Lay-byright side	SPRS	B	-	0	1	-	-	27	37%
Sidewalk left side	SLS	B	-	0	1	-	-	13	18%
Sidewalk right side	SRS	B	-	0	1	-	-	13	18%
Pedestrian crossing	Ped	B	-	0	1	-	-	12	16%
Pedestrian crossing density	PedD	NC	km^{-1}	0.0	4.0	0.3	0.845	12	16%
Parking lanes left side	PKLLS	B	-	0	1	-	-	1	1%
Parking lanes right side	PKLRS	B	-	0	1	-	-	1	1%
Safety barrier left side	SBLS	B	-	0	1	-	-	19	26%
Safety barrier right side	SBRS	B	-	0	1	-	-	20	27%
(Retaining) Wall left side	WLS	B	-	0	1	-	-	1	1%
(Retaining) Wall right side	WRS	B	-	0	1	-	-	1	1%

RRS, TRDLS, and TRDRS), the presence and the density of driveways (DLS, DRS, DDLS, and DDRS), the presence of lay-bys (LBR-SR and LBL-SL), sidewalks (SLS and SRS), parking lanes (PKLLS and PKLRS), safety barriers (SBLS and SBRS), and retaining walls (WLS and WRS) were carefully noted, as well as, the presence and density of pedestrian crossings (Ped and PedD).

5.4 Modeling approach

In order to evaluate the operating speed, we adopted the model structure proposed by Figueroa and Tarko [MFT05]. This model structure separates the estimate of the central tendency of speeds from the typical deviations of individual speeds, which is a function of the driving skills and decisions of individual drivers. In the speed dataset each speed data is a speed percentile (p) for a single direction

(d), on a specific section (s) and on a specific road (r); thus the sample dataset consists of data randomly extracted from sections and roads in the road network.

5.4.1 Fixed and random effect models

The data collected according to the methodology described in Section 5.3, contains repeated measurements; multiple observations are in fact available for the same road, the same section and for both directions. Random effects were included in the model to account for the dependency between any estimation errors from individual observations. They evaluate the existence of any differences between the speed predictions for all directions/sections/roads and the corresponding predictions for a specific direction/section/road. They are considered normally distributed according to the following distributions:

$$\begin{aligned} \alpha_{\mathrm{r}} &\sim \mathrm{N}(0, \sigma_{\mathrm{r}}^2) \\ \alpha_{\mathrm{s}} &\sim \mathrm{N}(0, \sigma_{\mathrm{s}}^2) \\ \alpha_{\mathrm{d}} &\sim \mathrm{N}(0, \sigma_{\mathrm{d}}^2) \end{aligned} \tag{5.1}$$

The dependent variable (Vrsd,i), which represents the generic observed speed (i) at a certain percentile (p) in a direction (d), section (s) and road (r), is then derived from a random effect (RE) model as follows:

$$\mathrm{V}_{\mathrm{rsd,i}} = \beta_0 + \sum_{\mathrm{k}} \beta_{\mathrm{k}}^{\mathrm{C}} \mathrm{X}_{\mathrm{rsd,k}}^{\mathrm{C}} + \sum_{\mathrm{j}} \beta_{\mathrm{j}}^{\mathrm{D}} (\mathrm{Z}_{\mathrm{p}} \mathrm{X}_{\mathrm{rsd,j}}^{\mathrm{D}}) + \sum_{\mathrm{m} \in \{\mathrm{r,s,d}\}} \alpha_{\mathrm{m}} + \epsilon_{\mathrm{rsd,i}} \tag{5.2}$$

in which β_0 is the general model intercept, β^{C} and β^{D} are calibration parameters for the variables affecting the estimated mean X^{C}, and the estimated standard deviation X^{D} respectively. Z_p is the standardized normal variable. The sum $\sum_{m\in\{r,s,d\}} \alpha_m$ is the cumulation of the random effects. $\epsilon_{rsd,i}$ is the error associated with each measurement; this is the same "regular" error terms than in regular regressions. In equation 5.2, the second term represents the central tendency term, while the third represents the dispersion term. If random effects from equation 5.2 are assumed to be constants to be estimated (instead of being normally distributed), the model becomes a fixed effect (FE) model. It is customary to use Greek letters for random effects and Latin letters for fixed effects.

5.4.2 Variable selection

The Information Criterion (BIC) postulated by Schwarz [Sch78] was used to select the variables that significantly affect driver speed behavior from all the possible covariates. The model with the lowest BIC function value f_{BIC}, calculated according to formula 5.3, would be preferred:

$$f_{BIC} = -2\hat{LL} + k\ln(n) \tag{5.3}$$

In this equation, $\hat{LL}$ is the maximized value of the log-Likelihood function, n the number of observations, and k is the number of parameters included in the model. According to the structure of equation 5.2, the number of parameters k is equal to the sum of the size of coefficients β_0, β^{C} and β^{D}. Only the variables that

contributed to the minimization of the BIC function were selected and included in the model.

5.5 Model calibration

A total of three models have been calibrated. Model #1 is a fixed effect (FE) model, in which a simple multiple regression analysis was performed by including all the variables selected according to the BIC criterion. Model #2 is a FE model that was calibrated with the variables selected by the random effects (RE) model and, finally, model #3 is an RE model, which was also calibrated using variables selected according to the BIC criterion.

Analyses have been carried out through the use of the R-software version 3.0.2, in particular using the REML algorithm running the lme4 package [R C15] [BMBW15]. The synthesis of results from model calibration can be found in tables 5.3 and 5.4. From an analysis of the table, it can be observed that the BIC criterion selects different variables when applied to FE or RE regression types. In the FE model, both geometric characteristics and Z values are selected, while the same method when applied to the RE model only keeps the section curvature and the pedestrian density in addition to the Z variables.

In model #1, twenty of the thirty-three variables resulted significant with respect to direction of the central tendency of data, while eighteen variables are responsible for the dispersion of speed data as a function of the percentile p. It was noted that the central tendency did not appear to be affected by the presence

on the left side of shoulders, ramps, barriers, retaining walls and driveways, nor by the presence of stopping places and by the intersection density along the road. The presence on the right side of barriers and retaining wall did not prove relevant either. The dispersion around the average speed is affected by a number of variables. It is worth noting that the presence of stopping places and retaining walls on both sides are excluded from the model.

In the case of models #2 and #3, the group of significant variables is the same since it was selected using the BIC criterion for model #3, as explained before. Table 3 shows that a drastic change of significant variables occurred, since the central value of speed distribution is affected by the curvature and the presence of a sidewalk on the right side (i.e., the closest sidewalk to the driving trajectory), while the dispersion is not affected by the presence of a shoulder and stopping places located on the left side.

Tables 5.3 and 5.4 contain the synthesis of the statistical analyses carried out with the three models. The variables in the two tables are described in table 5.2.

The quality of the results obtained can be appreciated in the following figures, where a comparison of observed vs. predicted speed values has been presented, with the results for model #1 reported in figure 5.2 and those for model #2 in figure 5.6. The results of these two models are compared with those of model #3 in figures 5.3 and 5.7 respectively. The graph in figure 5.7 compares the RE model #3 and the FE model #1, for which the variable selection is not the same.

In figure 5.3 sets of points appear to be distributed randomly along the identity line (y=x). However, the predicted values of the RE model are much closer to the

Table 5.3: Model 1 and 2 - coefficients and significant variables

Variable	Model 1			Model 2		
	estimate	s.d.	p-value	estimate	s.d.	p-value
(Intercept)	98.48	1.07	0.000	77.27	0.10	0.000
DDLS	0.71	0.05	0.000			
DDRS	0.81	0.05	0.000			
DRS	-8.03	0.19	0.000			
ILS	-1.87	0.17	0.000			
IRS	-2.78	0.17	0.000			
LG	-0.25	0.03	0.000			
LW	-10.59	0.27	0.000			
Ped	-5.66	0.45	0.000			
PedD	-0.47	0.21	0.026	-7.12	0.14	0.000
PKLLS	6.15	0.74	0.000			
PKLRS	11.71	0.68	0.000			
PSL	0.22	0.01	0.000			
RRS	4.45	0.67	0.000			
SLS	-4.32	0.26	0.000			
SLW	16.74	0.52	0.000			
SR	3.22	0.28	0.000			
SRS	-6.29	0.25	0.000			
SRW	-13.47	0.52	0.000			
TRDRS	0.86	0.50	0.086			
1/R	-1151.00	38.90	0.000	-1175.00	42.80	0.000
Z · 1/R	-202.20	27.21	0.000	-33.67	55.06	0.541
Z · DDRS	0.98	0.05	0.000	1.02	0.10	0.000
Z · DLS	-2.52	0.16	0.000	-2.29	0.62	0.000
Z · DRS	-3.20	0.20	0.000	-2.78	0.56	0.000
Z · IDRS	0.18	0.09	0.046	-0.14	0.23	0.545
Z · ILS	-1.27	0.16	0.000	-1.40	0.52	0.007
Z · IRS	-3.25	0.21	0.000	-2.40	0.50	0.000
Z · Ped	2.08	0.30	0.000	1.21	0.95	0.206
Z · PKLLS	-3.46	0.68	0.000	-3.28	1.35	0.015
Z · PSL	0.07	0.01	0.000	0.16	0.03	0.000
Z · SBLS	-2.14	0.16	0.000	-1.28	0.54	0.018
Z · SBRS	-2.62	0.17	0.000	-1.41	0.53	0.008
Z · SLS	-2.59	0.22	0.000	-2.91	0.57	0.000
Z · SLW	12.07	0.52	0.000	11.78	0.89	0.000
Z · SR	2.65	0.29	0.000	4.47	0.79	0.000
Z · SRS	-3.32	0.22	0.000	-3.22	0.55	0.000
Z · SRW	-14.66	0.52	0.000	-14.45	0.89	0.000
Z · TRDLS				-4.01	1.01	0.000
Z · TRDRS				-2.40	1.01	0.018
Z · WLS				-2.77	1.09	0.011
Z · WRS				-4.18	1.16	0.000
Z · DDLS				0.13	0.11	0.208
Z · ΔPSL				-0.13	0.06	0.046
Z · IDLS				0.22	0.23	0.341
Z · LG				0.09	0.05	0.075
Z · LW				-1.60	1.16	0.167
Z · PKLRS				1.64	1.36	0.229
Z · RLS				5.32	1.62	0.001
Z · RRS				3.58	1.58	0.024
Z · SPRS				0.39	0.45	0.383
Zlane	13.31	0.65	0.000	9.33	3.24	0.004
	Residual standard error: 4.048 Degrees of freedom (df): 6528 Corrected multiple R^2: 0.927 F-statistic: 2182 with 38 and 6828 df $P(F_{38,6828} > 2182) < 10e-16$			Residual standard error: 6.673 df: 6532 Corrected multiple R^2: 0.8006 F-statistic: 776 with 34 and 6532 df $P(F_{34,6832} > 776) < 10e-16$		

line, which indicates a much better fit. A similar result can be seen in figure 5.7, with model #2 points that are further from the equality line. This is to be expected since the model selection for the regression is not optimal according to the BIC criterion.

Finally, figures 5.4, 5.5, 5.8, 5.9, 5.11 and 5.12, report the residual distributions for the three models, organized by road and section. The expected results for these box plots would be for all 13 boxes of road residuals and all 37 boxes of

Table 5.4: Model 3 - coefficients and significant variables

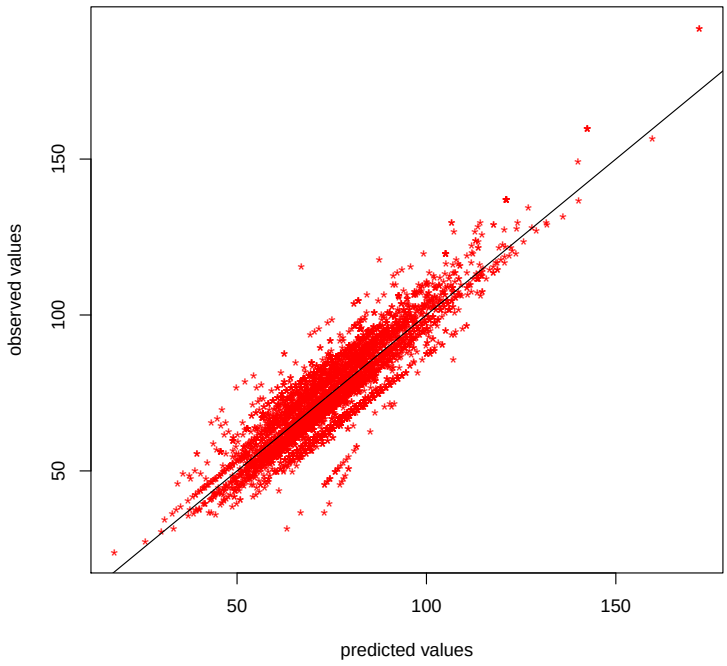
Variable	Model 3		
	estimate	s.d.	p-value
Intercept	79.34	1.82	0.000
PedD	-7.61	1.94	0.000
1/R	-1949.00	1063.00	0.067
Z	10.99	0.69	0.000
Z · 1/R	-134.10	11.69	0.000
Z · SLS	-2.81	0.12	0.000
Z · SBLS	-1.27	0.11	0.000
Z · IDRS	-0.13	0.05	0.009
Z · DLS	-2.34	0.13	0.000
Z · Ped	1.38	0.20	0.000
Z · SRS	-3.14	0.12	0.000
Z · SBRS	-1.44	0.11	0.000
Z · SRW	-14.40	0.19	0.000
Z · SLW	11.94	0.19	0.000
Z · ILS	-1.37	0.11	0.000
Z · PSL	0.15	0.01	0.000
Z · DDRS	1.00	0.02	0.000
Z · DRS	-2.80	0.12	0.000
Z · IRS	-2.40	0.11	0.000
Z · SR	3.99	0.17	0.000
Z · PKLLS	-3.29	0.28	0.000
Z · TRDLS	-4.21	0.21	0.000
Z · RLS	5.58	-0.34	0.000
Z · LW	-1.83	0.24	0.000
Z · WRS	-3.38	0.24	0.000
Z · IDLS	0.22	0.05	0.000
Z · LG	0.09	0.01	0.000
Z · DDLS	0.11	0.02	0.000
Z · ΔPSL	-0.11	0.01	0.000
Z · WLS	-1.99	0.23	0.000
Z · PedD	0.73	0.08	0.000
Z · PKLRS	1.63	0.29	0.000
Z · TRDRS	-2.53	0.21	0.000
Z · RRS	3.74	0.33	0.000
Z · LBRS	0.40	0.09	0.000
	BIC at convergence: 36,775		
	R^2: 0.9332		
	Random effect	Variance	
	Direction:Section	19.45	
	Section:Road	85.96	
	Road	0.00	
	Residuals	3.86	

section residuals to have approximately the same height and shape, and to be centered around zero. This would indicate that errors are independent and identically distributed for all roads and sections.

The plots 5.4 and 5.5 indicate that model 1 cannot be considered satisfactory due to the large dispersion of the residuals. For example, road #7 has very high residuals in absolute value terms, suggesting a high variance, section #18 has very low negative residuals for almost all observations suggesting a strong negative effect, and section 19 has very high residuals indicating a positive effect.

Model #2 is not satisfactory either, since the residuals for roads (figure 5.8) are never centered around zero, while the RE model #3 fixes most of the regression

Figure 5.2: Fitted and Observed Values for Model 1

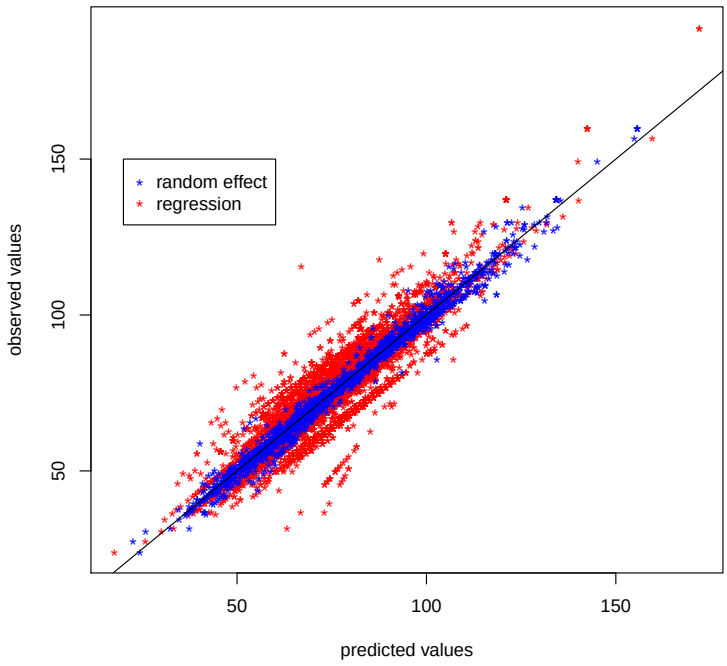


problems. In fact, in this last case the errors for all roads and sections are centered around zero, hence the fitted values are not biased in any road or any section. The variance of the residuals is much lower than in the case of model #1 and #2. This illustrates a very interesting property of RE models: they do not simply shift the error terms for each value of the effects, but they also allow to reduce the spread of the residuals.

5.6 Conclusions

In this chapter fixed (FE) and random effect (RE) linear models have been applied to predict a full range of percentile values of the operating speed along tangents or curves of two lane rural highways. The model has been calibrated with

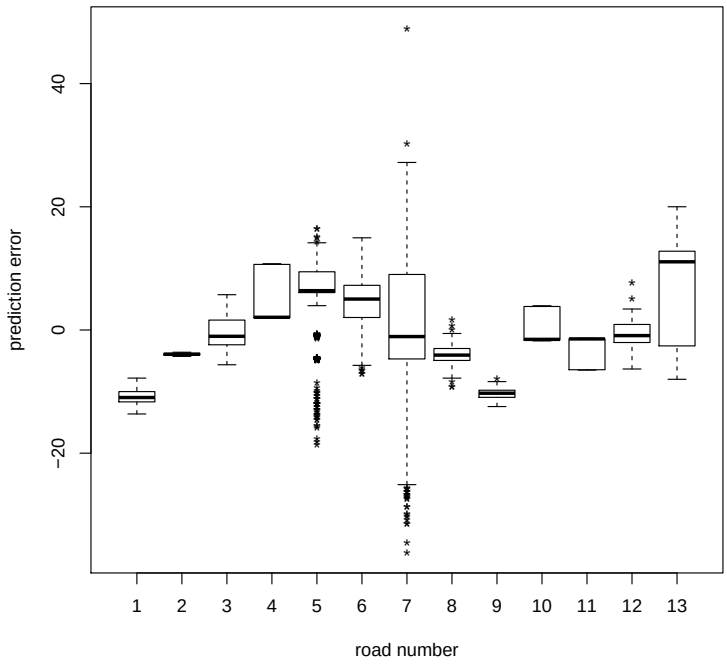
Figure 5.3: Comparison of Fitted and Observed Values - Model 1 and 3



data collected from surveys of isolated vehicles on a number of road sections in the Northwest of Italy. A total of 6,567 observations have been collected from 37 sections randomly selected from 13 roads. In total, 33 geometric and environmental variables have been taken into account for model estimation. The structure of the model separates the central tendency from the dispersion of speed data allowing the estimation of not only the 85th percentile (which is usually regarded as a reference measure for operating speeds) but also the evaluation of any percentile through the standardized normal variable Zp.

The comparison across fixed effect and random effect models containing variables selected according to the BIC criterion demonstrates that the latter perform better from a statistical point of view. The free-flow speed data collected in this

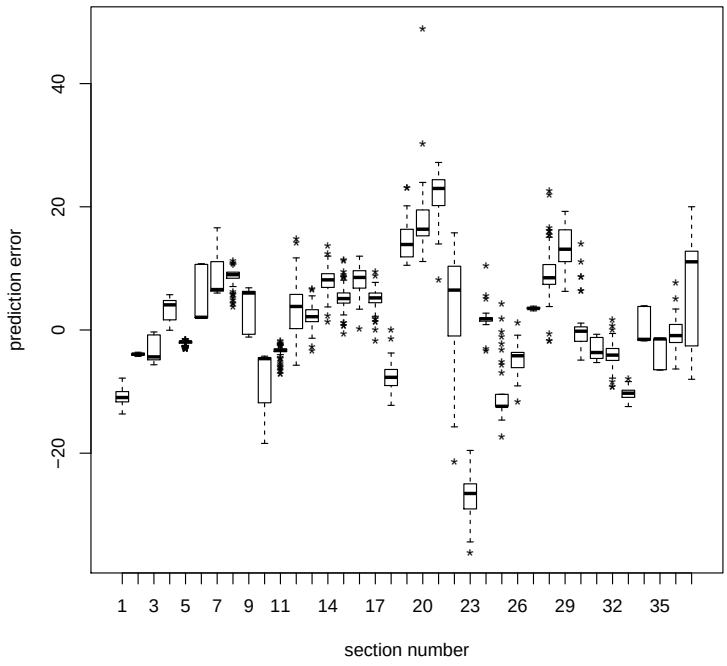
Figure 5.4: Residuals for Model 1, by Road



study violate the assumptions of ordinary least squares regression which produces biased estimates and large dispersion of the residuals. The RE model, which correctly accounts for the sampling design, produces a very low level of errors and it does not suffer from the presence of outliers.

The RE model results (Model #3) highlights once again [Tra11] the effects of the curvature on the central tendency of speed distribution. The density of pedestrian crossing is the only other variable that significantly affects the mean value of free-flow speeds. According to the same model several cross-sectional elements located in the roads margins contribute to the dispersion of speed data around the central value. Each element affects driver behavior in a small measure, but when all these effects are taken into account the differences between drivers behavior become

Figure 5.5: Residuals for Model 1, by Section



significant. When analyzing the single REs it is possible to conclude that road effects are negligible, that most of the errors are associated with a road section and for a lesser extent to the direction effect, and that the residuals have low standard deviation.

The results obtained also clearly demonstrate that speed predictions are less variable and more transferable if they are obtained from random effect models than from fixed effect models.

Future investigation might be aimed at the validation of the model presented here with speed data relative to different driving environments and collected with different measurement technologies.

Figure 5.6: Fitted and Observed Values for Model 2

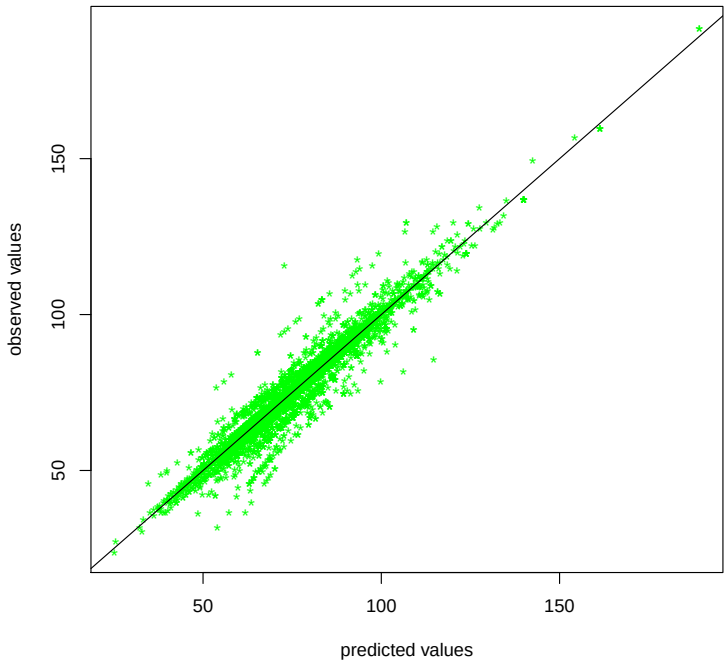


Figure 5.7: Comparison of Fitted and Observed Values - Model 2 and 3

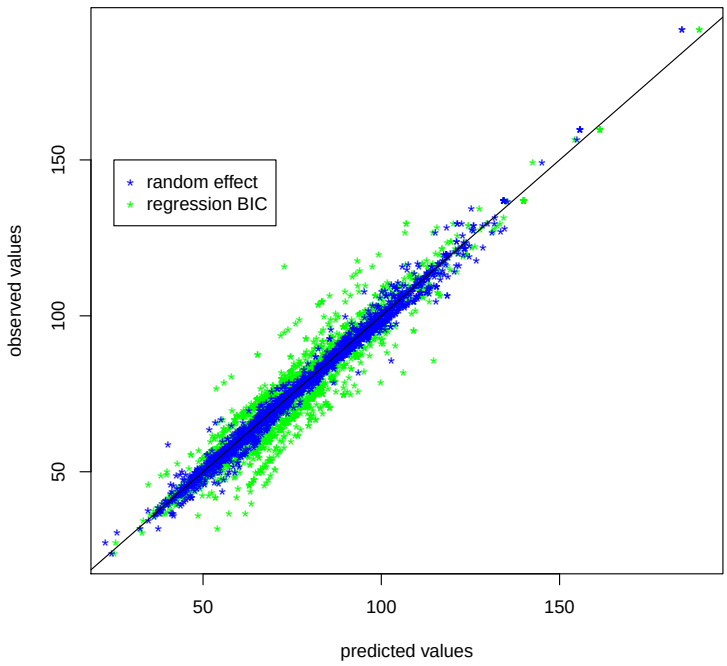


Figure 5.8: Residuals for Model 2, by Road

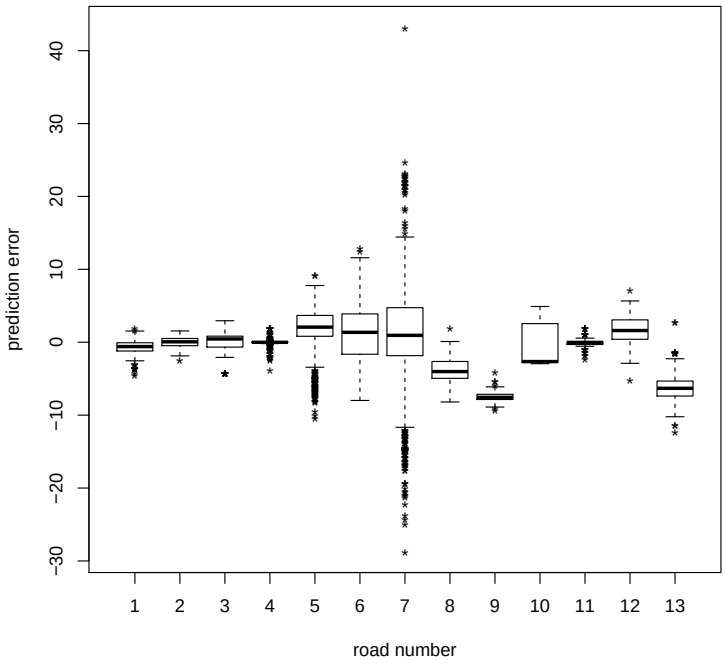


Figure 5.9: Residuals for Model 2, by Section

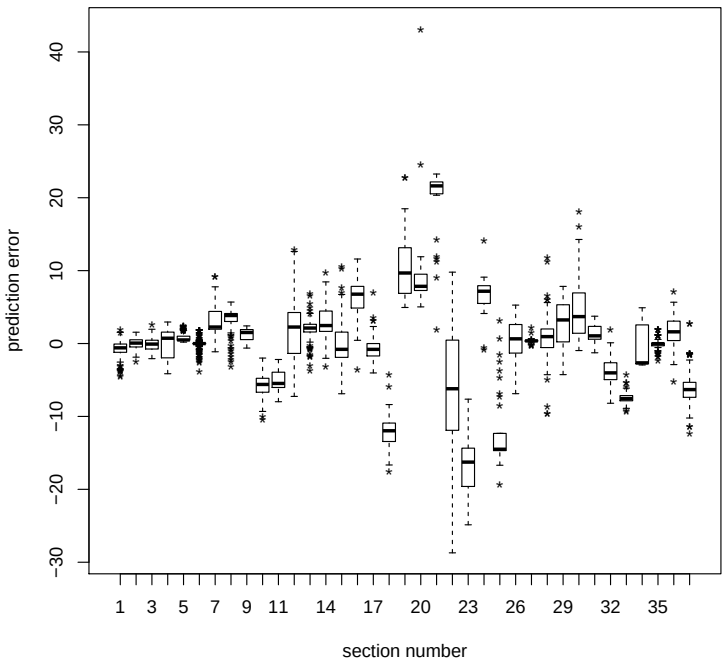


Figure 5.10: Fitted and Observed Values for Model 3

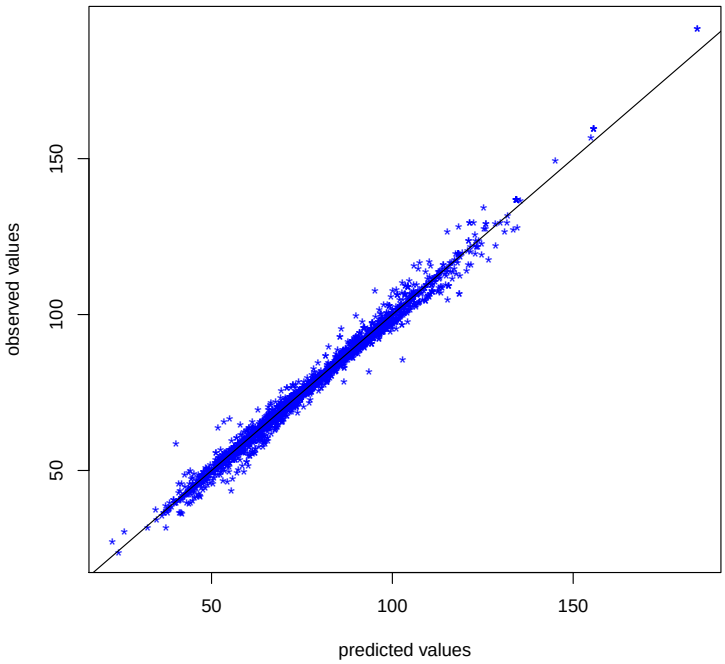


Figure 5.11: Residuals for Model 3, by Road

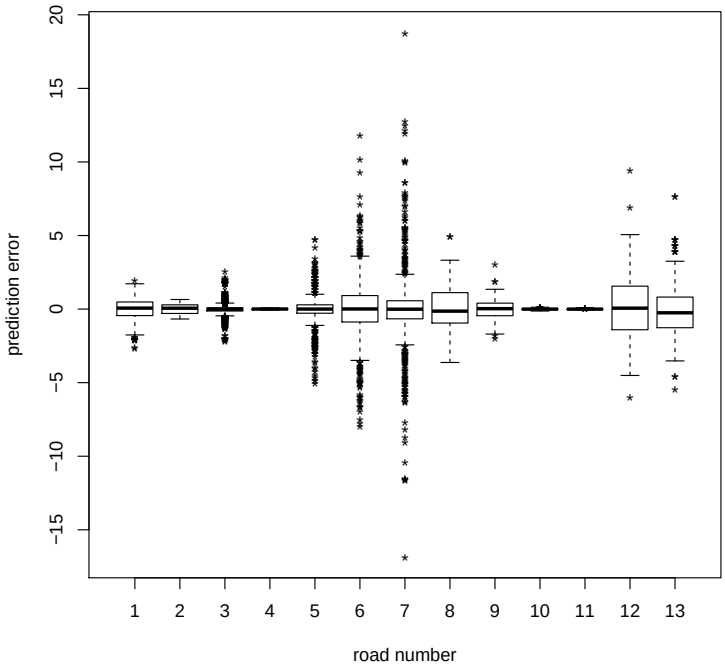
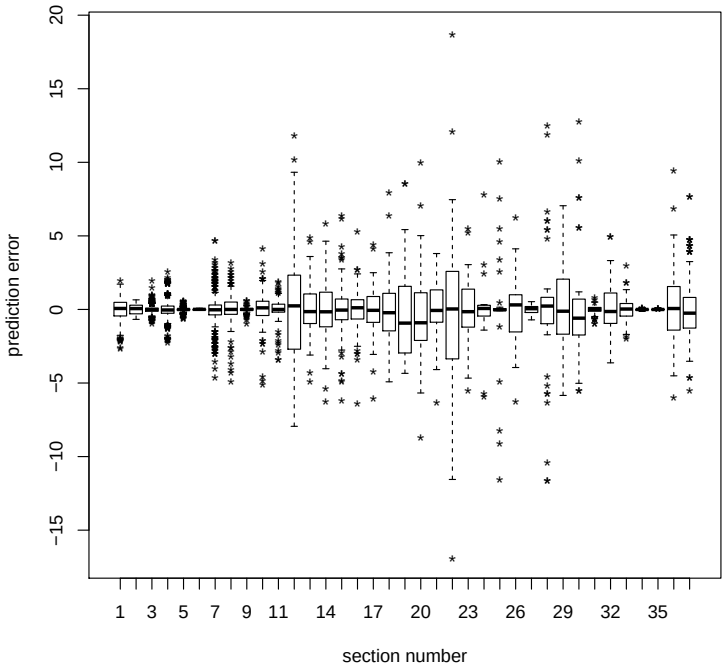


Figure 5.12: Residuals for Model 3, by Section



Chapter 6: Validation of Random Effects

6.1 Introduction

In this chapter, we propose methods to predict quantiles of speed data distributions for road sections that were not in the original sample and for which we suppose to have very few observations available. We have shown that the theory of linear models can be used to estimate random effect models and to predict their realized values in the sample. Our problem, however, poses two challenges. First, we are interested in making out of sample predictions, and in order to do that we need to calculate random effects in the new road section. Second, our modeling approach makes use of speed quantiles as predictors of the linear model; those are not readily available for the new sections.

Random effects in most situations are assumed to have zero mean and therefore the best *a priori* predictor for random effects in a new road section is zero. We claim that better predictors could be produced by considering an auxiliary, simpler random effect model whose purpose is to overcome the unavailability of quantiles in the validation sample. We show that this auxiliary model, although inferior to the actual model, can provide good random effect predictors by sampling as little as five observations for each new section.

Searle [MSN08, p. 260] suggests to model random effect pairs jointly; this way to proceed takes advantage of the potential correlation existing between them. We followed this idea and we use it to build our auxiliary model. We correlate the random effects of the original model with the random effects to be defined for the new section and we use the first one to predict the latter.

Our validation approach is based on jackknife technique [ET93, p. 141,237]. The validation is performed on the quantiles of 31 sections; predictions for each section are made based on the information contained in the remaining sections. This is similar to regular validation schemes that calibrate a model on 80% of the sample and use it to make predictions for the remaining 20% of the sample. With the jackknife scheme, however, each set of predictions is made using a different training sample.

6.2 One random effect models

We describe our methodology for a simple case where the response variable is affected by a set of predictors X, one random effect α_{s} and an error term ϵ:

$$\mathrm{Y_{si}} = \mathrm{X_{si}}\beta + \alpha_{\mathrm{s}} + \epsilon_{\mathrm{si}}$$

Where α_{s} and ϵ_{si} are independent and follow a normal distribution:

$$\alpha_{\mathrm{s}} \sim \mathrm{N}(0, \sigma_{\mathrm{s}}^2)$$

$$\epsilon_{\mathrm{si}} \sim \mathrm{N}(0, \sigma^2)$$

Once the model is calibrated, estimates for β, σ_{s}^2 and σ^2 are available. Predictors for the realized values of α_{s} in the sample can be computed using either the true or the estimated values of the parameters. If our interest is focused on the value of $\alpha_{\mathrm{s}'}$ in a new section that is not in the sample, the best we can do without further modeling assumptions is to predict the overall mean of the random effects: $\hat{\alpha}_{\mathrm{s}'} = 0$.

It is worth noting that we *estimate* β, σ_{s}^2 and σ^2 because they are parameters of the model. On the other side we *predict* α_{s} and Y_{si} because they are regular random variables.

Let's assume that we are interested in predicting random effects for a new section s' with unknown random effect $\alpha_{\mathrm{s}'}$. The best *a priori* estimate of $\alpha_{\mathrm{s}'}$ is zero since $\alpha_{\mathrm{s}'} \sim \mathrm{N}(0, \sigma_{\mathrm{s}}^2)$. In some contexts this may be satisfying, but we are exploring methods to sample a few observations in the new section s' in order to improve our knowledge about $\alpha_{\mathrm{s}'}$ and build a better *a posteriori* estimator.

The terms *a priori* and *a posteriori* are used in their classic meaning here. *a priori* refers to what happens before we observe data to predict $\alpha_{\mathrm{s}'}$ and *a posteriori* refers to what happens after we observe data for the prediction. No Bayesian inference is performed in this chapter.

6.2.1 Conditional mean

Suppose that we observe k observations $\mathrm{Y}_{\mathrm{s}'1}, \mathrm{Y}_{\mathrm{s}'2}, ..., \mathrm{Y}_{\mathrm{s}'\mathrm{k}}$ in the new section. For illustration purposes we will work with $\mathrm{k} = 3$. Our objective is to find a good

predictor of $\alpha_{s'}$.

First, we note that we cannot observe $\alpha_{s'}$ directly: we always observe $\alpha_{s'} + \epsilon_{s'i}$. We start by defining this sum of residuals:

$$r_{s'i} = Y_{s'i} - X_{s'i}\hat{\beta}$$

We cannot observe $Y_{s'i} - X_{s'i}\beta$ because we do not know the value β, so we will use the estimated value $\hat{\beta}$ from model calibration to estimate the total residuals.

We know by hypothesis that $u = (\alpha_{s'})$ follows a normal distribution with mean zero and variance $D = (\sigma_s^2)$.

We know that $r = (r_{s'1}, r_{s'2}, r_{s'3})$ has approximately a zero mean and variance:

$$V = \begin{bmatrix} \sigma_s^2 + \sigma^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 + \sigma^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 + \sigma^2 \end{bmatrix}$$

We can also write V like this:

$$V = \sigma^2 I + \sigma_s^2$$

V is approximate because r was not calculated with the real parameter of the model. The approximate covariance between the random effect $\alpha_{s'}$ and one given total residual is:

$$\mathrm{Cov}(\alpha_{s'}, r_{s'i}) \approx \sigma_s^2$$

So we can write that the approximate covariance C between u and r is:

$$C = \begin{bmatrix} \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \end{bmatrix}$$

or, with $\mathbf{1}_{m \times n}$ being a $m \times n$ matrix of ones:

$$C = \sigma_s^2 \mathbf{1}_{1 \times k}$$

and then:

$$\begin{bmatrix} \mathrm{u} \\ \mathrm{r} \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathrm{D} & \mathrm{C} \\ \mathrm{C}^{\mathrm{T}} & \mathrm{V} \end{bmatrix} \right)$$

Here we do not make claims on the joint distribution of (u, r). It is sufficient to know the first and second moments of (u, r) to derive the *best linear unbiased predictor* (BLUP) of u. If the joint distribution were normal, the BLUP would be the overall best predictor of u, however it will not be the case for our final application.

6.2.2 Best predictor

Once we observe r, the BLUP of u is the conditional mean. The expectation of $\mathrm{u}|\mathrm{r}$ is given by:

$$\mathrm{E}(\mathrm{u}|\mathrm{r}) = \mathrm{C}\mathrm{V}^{-1}\mathrm{r}$$

If (u, r) were jointly normal, the BLUP would happen to be the best predictor.

In the case of only one random effect u will have dimension one and is equal to $\alpha_{\mathrm{s}'}$. The predicted value for an observation i in section s' will be:

$$\hat{\mathrm{Y}}_{\mathrm{s}'\mathrm{i}} = \mathrm{X}_{\mathrm{s}'\mathrm{i}}\hat{\beta} + \hat{\alpha}_{\mathrm{s}'}$$

We can estimate $\hat{\beta}$ relatively easily so the main challenge for this problem is to predict $\hat{\alpha}_{\mathrm{s}'}$. Our objective is to investigate what is the smallest sample in section s' we can use to predict $\alpha_{\mathrm{s}'}$ satisfactorily. Generally, the prediction converges faster

when σ^2 is low and σ_s^2 is high because then one single observation of r is less noisy and more correlated with the unknown realized value $\alpha_{s'}$.

6.3 Numerical examples

To illustrate our methodology, we use the speed database with 37 sections described in 5.3. In our first case study we ignore the direction effect. For all 31 sections considered, we calibrate the model on 36 sections, and we compute the estimated random effect with $k = 1, 2, 3, ...$ observations taken from the section that is left out for validation. This will provide a series of predicted random effects. We want to assess the convergence of the prediction.

Figures 6.1, 6.2, 6.3 and 6.4 contain the results obtained for each section. The subplots represent the predicted random effects in the validation section. The x-axis corresponds to the number of observations that were used from the validation section in order to predict the random effects. For example, an x-value of 10 for section 3 means that we estimated a model with all sections but the third one, and then looked at 10 observations in the third section in order to predict the (realized) effect of section 3. Each subplot contains a solid dark red line, a solid pale red line and a dotted dark red line. The solid dark red line shows the predicted effects for the full model, assuming we knew the quantile information in the validation sample. This is not a realistic case for this problem but it is nevertheless the effect that we ultimately want to predict. The pale red line shows the predicted effects with the auxiliary model. These effects can be computed but we are only interested in how

they can help us predict the ones from the full model. The dashed red line shows the predicted effects for the full model but using only information from the auxiliary model. This is the prediction that we ultimately use.

From the figures 6.1, 6.2, 6.3 and 6.4 we can conclude the following. First, a relative convergence in the predicted effects as the number of observations grows is observed. Second, there is a substantial difference between the solid pale and dark red lines. This is not a problem in itself although the closer the two lines are, the more likely it will be to predict one from the other. Lastly, the dashed dark red line does not approximate the dark red line and it is mostly superposed to the pale red line, meaning that the predicted effect of the full model using the auxiliary model are no better approximation than just the predicted effects of the auxiliary model. This is obviously disappointing but we will see later that accounting for the design of the sample mostly fixes this problem.

Figure 6.1: Sections 1-12

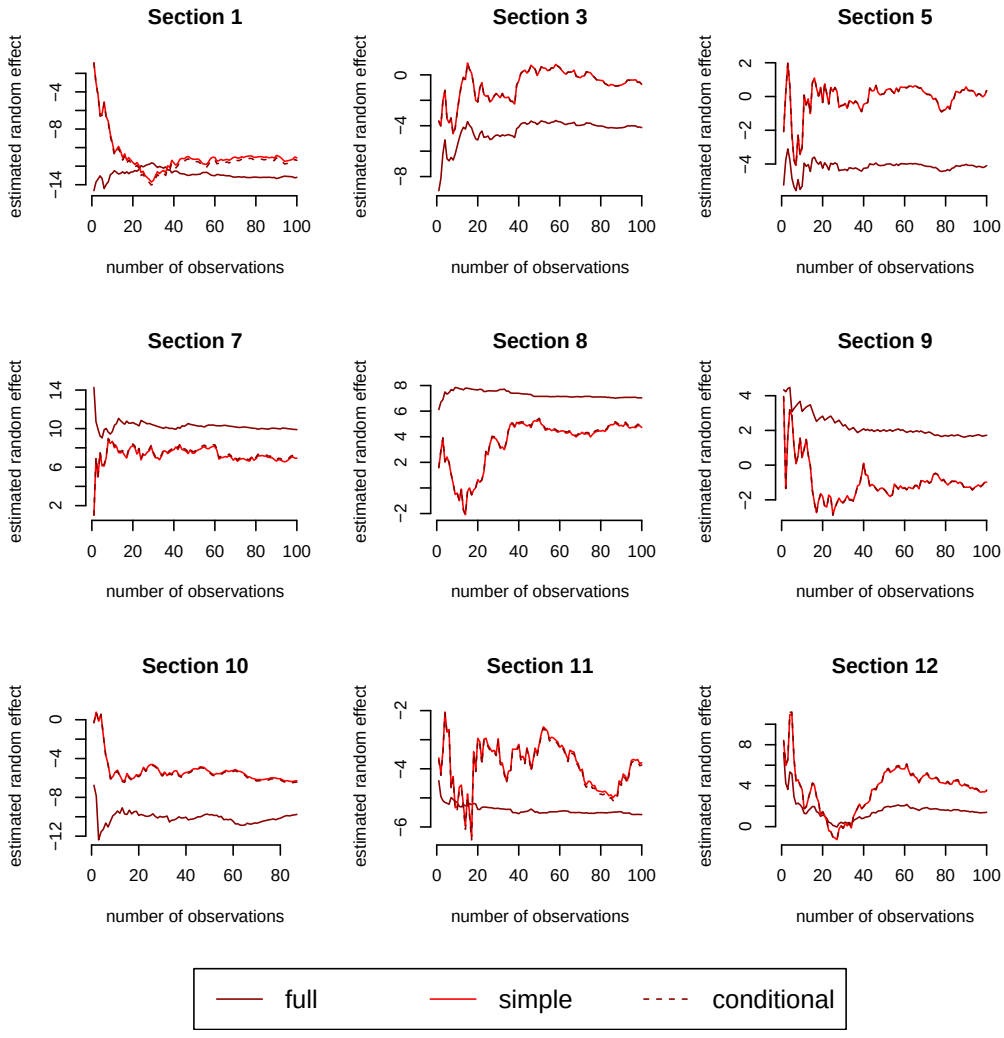


Figure 6.2: Sections 13-22

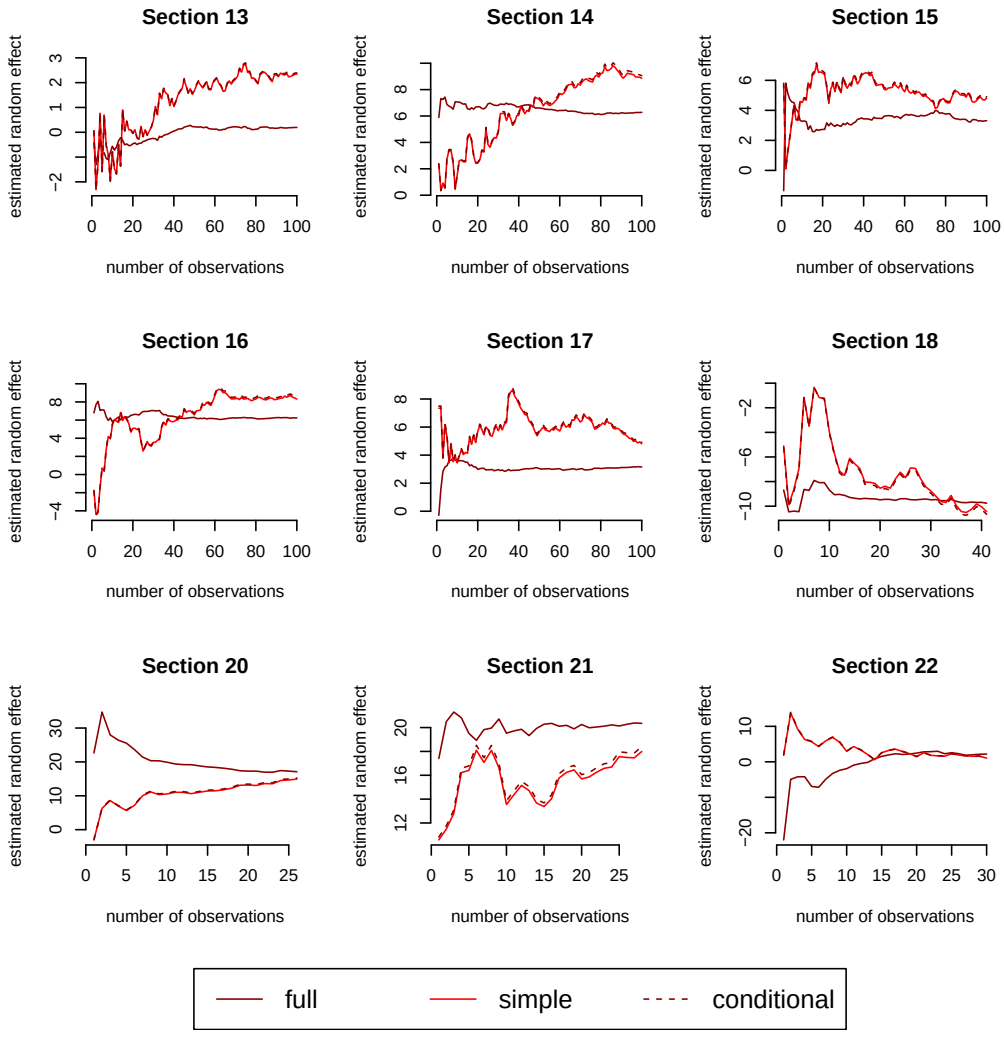


Figure 6.3: Sections 23-31

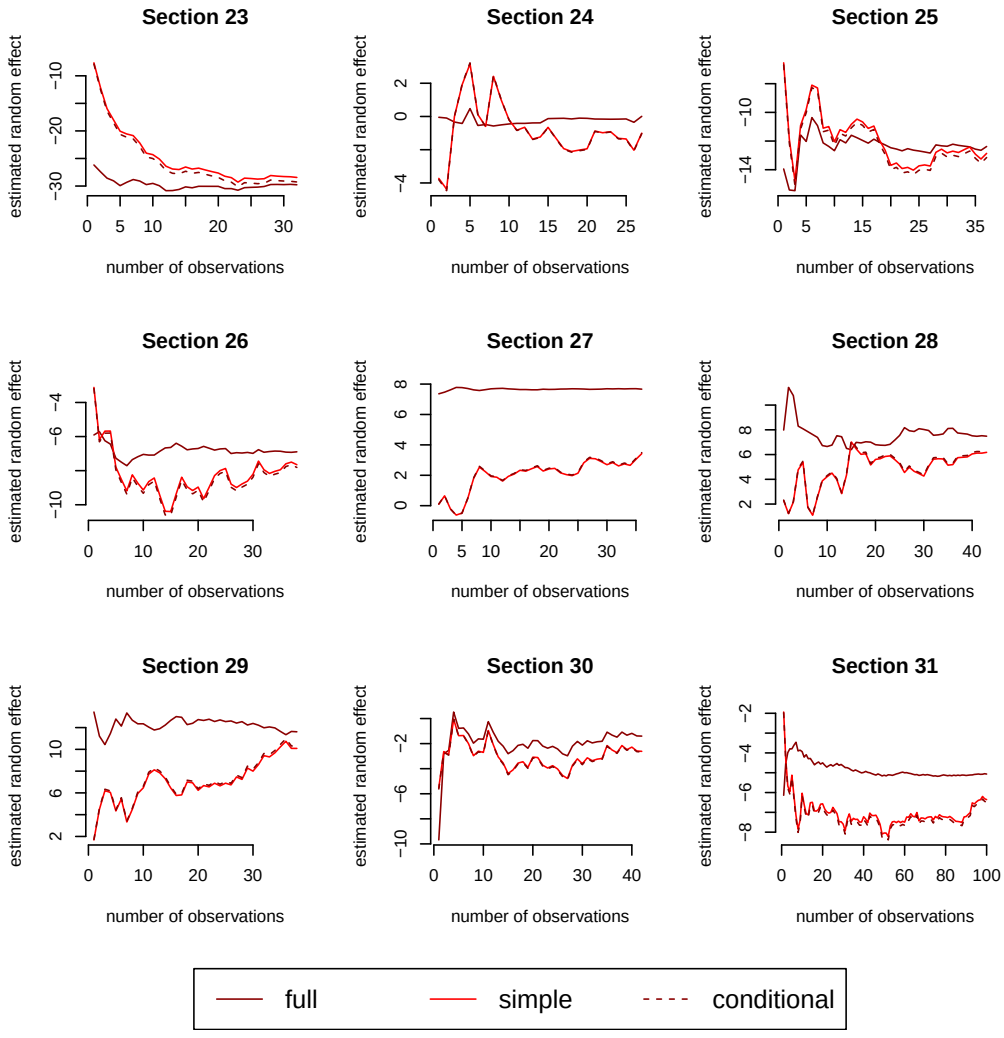
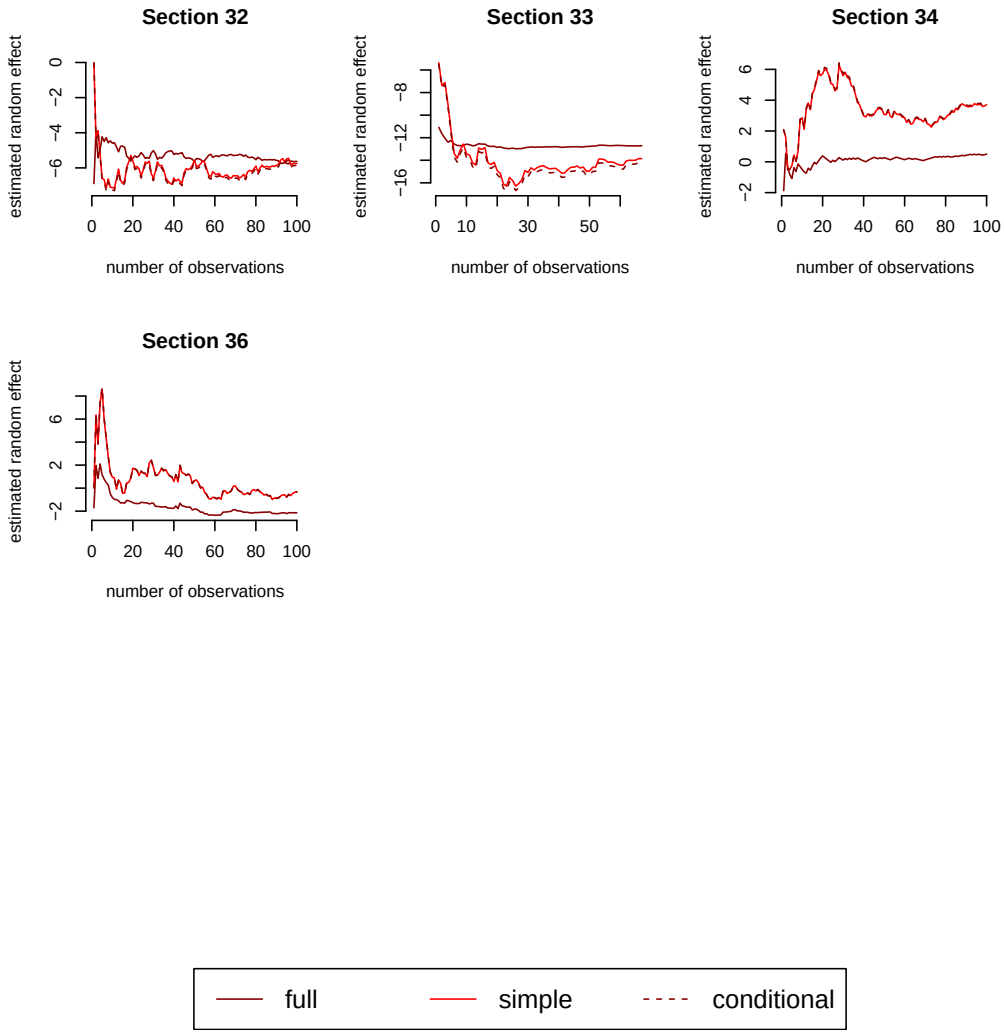


Figure 6.4: Sections 32-36



6.4 Two random effects models

The case with two nested random effects is described by the following equation:

$$\mathrm{Y_{sdi}} = \mathrm{X_{sdi}}\beta + \alpha_{\mathrm{s}} + \beta_{\mathrm{d}} + \epsilon_{\mathrm{sdi}}$$

where the random effect β_{d} is nested within the levels of α_{s}. To simplify the notation we write β_{d} and not $\beta_{\mathrm{d|s}}$, but the latter is more precise.

In this model, a sample from the new section s' also includes some levels of the direction effect. In our case there are two directions for all sections. For illustration purpose we will assume that we sample one section, two directions and $\mathrm{k} = 3$ observations per direction. The random effect to be predicted is:

$$\mathrm{u} = (\alpha_{\mathrm{s}'}, \beta_1, \beta_2)$$

and the total residuals are given by:

$$\mathrm{r_{s'di}} = \mathrm{Y_{s'di}} - \mathrm{X_{s'di}}\beta$$

Similar to the one random effect model, we do not observe a single random effect by itself, instead we always observe a sum of effects. We have the following covariance structure for the total residuals:

$$\mathrm{var}(\mathrm{r_{s'di}}) \approx \sigma_{\mathrm{s}}^2 + \sigma_{\mathrm{d}}^2 + \sigma^2$$

$$\mathrm{cov}(\mathrm{r_{s'di}}, \mathrm{r_{s'dj}}) \approx \sigma_\mathrm{s}^2 + \sigma_\mathrm{d}^2$$

and:

$$\mathrm{cov}(\mathrm{r_{s'di}}, \mathrm{r_{s'd'i}}) \approx \sigma_\mathrm{s}^2$$

These covariance components are approximated because the $\mathrm{r_{sdi}}$ terms are computed using estimated values of $\hat{\beta}$. Moreover, we will also plug in estimated values for the variance components. Even if we do not address here the effect of such an approximation, it is always useful to keep it in mind.

The covariance matrices D, C and V of u and r are:

$$\mathrm{D} = \begin{bmatrix} \sigma_\mathrm{s}^2 & 0 & 0 \\ 0 & \sigma_\mathrm{d}^2 & 0 \\ 0 & 0 & \sigma_\mathrm{d}^2 \end{bmatrix}$$

$$\mathrm{V} = \sigma_\mathrm{s}^2 + \begin{bmatrix} \sigma_\mathrm{d}^2 + \sigma^2 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 & 0 & 0 & 0 \\ \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 + \sigma^2 & \sigma_\mathrm{d}^2 & 0 & 0 & 0 \\ \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 + \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\mathrm{d}^2 + \sigma^2 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 \\ 0 & 0 & 0 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 + \sigma^2 & \sigma_\mathrm{d}^2 \\ 0 & 0 & 0 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 + \sigma^2 \end{bmatrix}$$

or just, with $\mathbf{0}_{\mathrm{m\times n}}$ being a matrix of zeros:

$$\mathrm{V} = \sigma_\mathrm{s}^2 + \begin{bmatrix} \sigma_\mathrm{d}^2 + \sigma^2\mathrm{I} & \mathbf{0}_{\mathrm{k\times k}} \\ \mathbf{0}_{\mathrm{k\times k}} & \sigma_\mathrm{d}^2 + \sigma^2\mathrm{I} \end{bmatrix}$$

$$\mathrm{C} = \begin{bmatrix} \sigma_\mathrm{s}^2 & \sigma_\mathrm{s}^2 & \sigma_\mathrm{s}^2 & \sigma_\mathrm{s}^2 & \sigma_\mathrm{s}^2 & \sigma_\mathrm{s}^2 \\ \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 & \sigma_\mathrm{d}^2 \end{bmatrix}$$

or:

$$\mathrm{C} = \begin{bmatrix} \sigma_\mathrm{s}^2\mathbf{1}_{\mathrm{1\times k}} & \sigma_\mathrm{s}^2\mathbf{1}_{\mathrm{1\times k}} \\ \sigma_\mathrm{d}^2\mathbf{1}_{\mathrm{1\times k}} & \mathbf{0}_{\mathrm{1\times k}} \\ \mathbf{0}_{\mathrm{1\times k}} & \sigma_\mathrm{d}^2\mathbf{1}_{\mathrm{1\times k}} \end{bmatrix}$$

The predicted random effects are still given by the expected mean:

$$\mathrm{E(u|r)} = \mathrm{CV^{-1}r}$$

The predicted value for an observation i in section s$'$ and direction d will be:

$$\hat{\mathrm{Y}}_{\mathrm{s'di}} = \mathrm{X}_{\mathrm{s'di}}\hat{\beta}_{\mathrm{s'di}} + \hat{\alpha}_{\mathrm{s'}} + \hat{\beta}_{\mathrm{d}}$$

The sum $\hat{\alpha}_{\mathrm{s'}} + \hat{\beta}_{\mathrm{d}}$ is always used in predicted values so we are always interested in assessing that it converges fast enough. For example, if $\hat{\alpha}_{\mathrm{s'}}$ is underestimated but $\hat{\beta}_{\mathrm{d}}$ is overestimated, their sum might still be a good predictor of the real sum of random effects.

6.4.1 Numerical examples

For this second case, we use the same speed database and we keep the direction effect. For all the 31 sections considered, we calibrate the model on 36 sections, and we compute the predicted random effects with $\mathrm{k} = 1, 2, 3, ...$ observations taken from *each direction*. We want to assess the convergence of $\hat{\alpha}_{\mathrm{s'}}$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

Figures 6.5 through 6.8 should be read in the same way as figures 6.1 through 6.4; please note that we do not include the predicted random effect of the auxiliary model. The solid red line corresponds to the predicted section effect, and the two solid blue lines to the predicted direction effects. The dashed lines correspond to the

predicted effects using the auxiliary model and can be computed in the validation section.

For both the sections and directions, the predicted effects using the auxiliary model are very close to the ones using the full model. However there are still noticeable differences when very few observations are used for the prediction. For example, predictions in sections 8 and 27 are not so precise for only two or three observations.

We also note striking differences between predictions with one and two random effects. Effects with only a section component are not close to the "true" effects of the full model but incorporating the directions, thus accounting for the design of the sample, drastically improves the predictions.

Figure 6.5: Sections 1-12

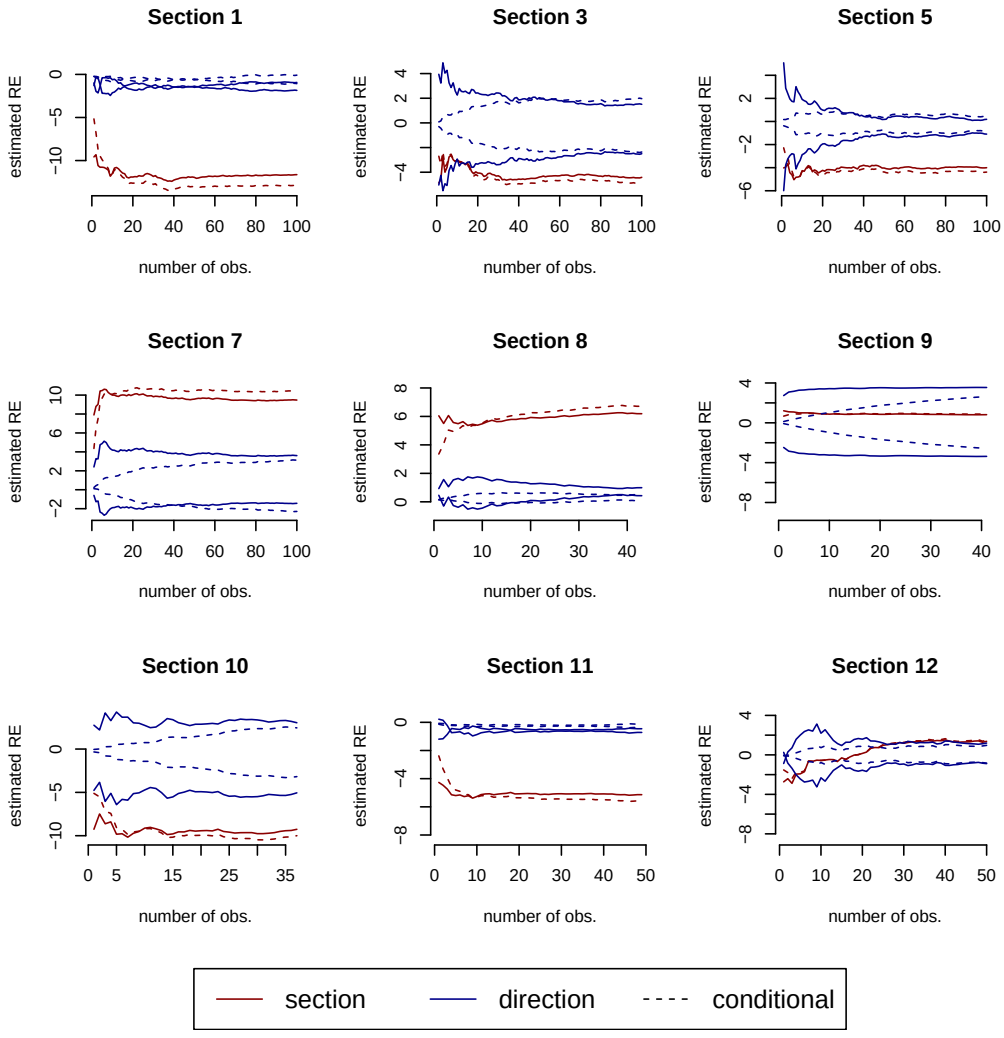


Figure 6.6: Sections 13-22

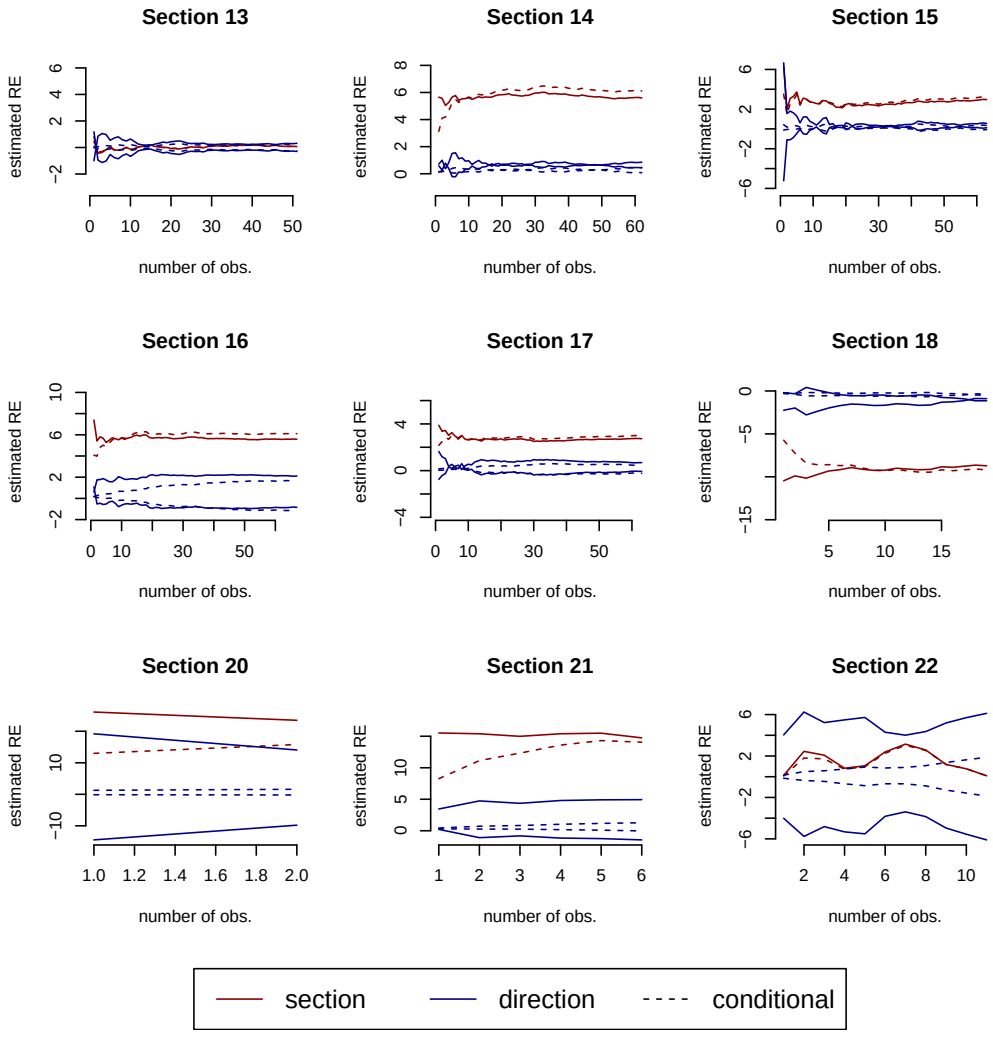


Figure 6.7: Sections 23-31

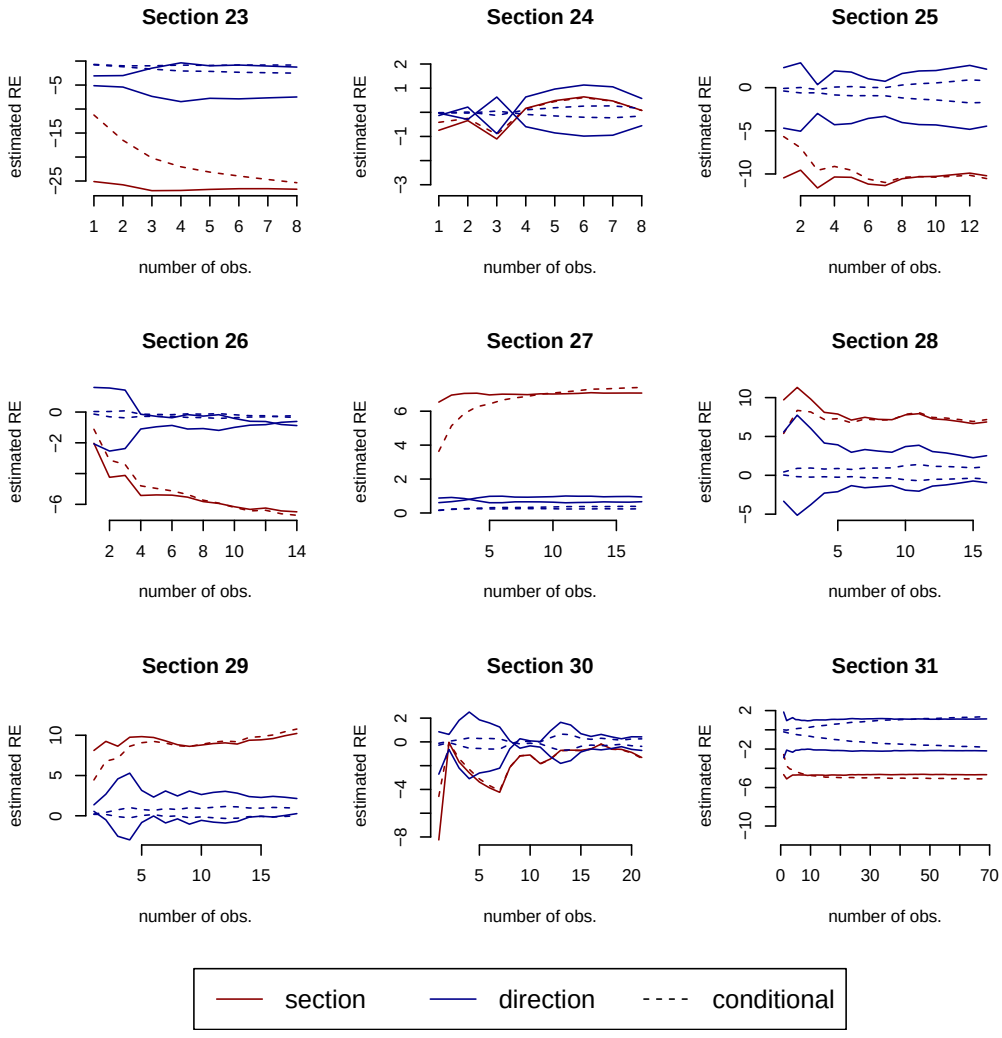
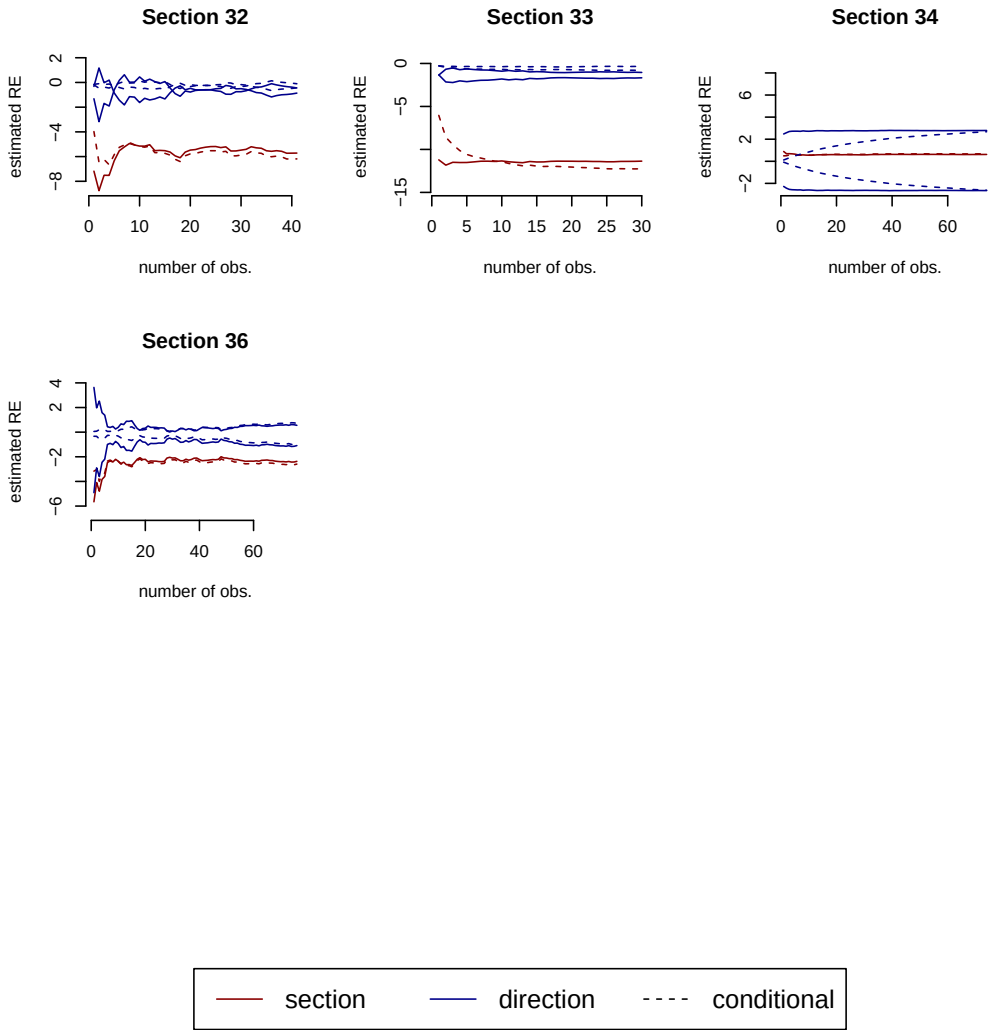


Figure 6.8: Sections 32-36



6.5 Computation of the residuals

In the previous examples we were calculating:

$$r_{sdi} = Y_{sdi} - X_{sdi}$$

However it is not always possible to compute r. Suppose that we have the following model specification:

$$Y_{sdi} = X_{sdi}\beta_x + Z_{sdi}\beta_z + \alpha_s + \beta_d + \epsilon_{sdi}$$

where Z_{sdi}, which represents the regular variables multiplied by the normal quantiles Z, is not available in the validation sample. In this case, it is not possible to isolate the sum of random effects and we must rely on alternative methods.

The strategy here is to use the estimated total residuals from the first model, with only observable predictors X, to predict the random effects of the complicated model.

The simple (with X) and expanded (with X and Z) models are both calibrated and parameters are estimated from both models. In addition we also calculate the empirical correlation between the effects from both models. The ideal scenario is that corresponding effects from both models are perfectly correlated:

$$Y_{sdi} = X_{sdi}\beta_* + \alpha_{s*} + \beta_{d*} + \epsilon_{sdi*}$$

$$\alpha_{s*} \sim N(0, \sigma^2_{s*})$$

$$\beta_{\mathrm{d}*} \sim \mathrm{N}(0, \sigma^2_{\mathrm{d}*})$$

$$\epsilon_{\mathrm{sdi}*} \sim \mathrm{N}(0, \sigma^2_*)$$

$$\mathrm{Y_{sdi}} = \mathrm{X_{sdi}}\beta_{\mathrm{x}} + \mathrm{W_{sdi}}\beta_{\mathrm{w}} + \alpha_{\mathrm{s}} + \beta_{\mathrm{d}} + \epsilon_{\mathrm{sdi}}$$

$$\alpha_{\mathrm{s}} \sim \mathrm{N}(0, \sigma^2_{\mathrm{s}})$$

$$\beta_{\mathrm{d}} \sim \mathrm{N}(0, \sigma^2_{\mathrm{d}})$$

$$\epsilon_{\mathrm{sdi}} \sim \mathrm{N}(0, \sigma^2)$$

By design, the marginal distributions of the random effects are normal. However, the empirical analysis does not support the assumption that they are jointly normal so we can only hope to estimate first and second moments of all the effects and residuals to compute the BLUP.

We will assume that $(\alpha_{\mathrm{s}}, \alpha_{\mathrm{s}*})$ and $(\beta_{\mathrm{d}}, \alpha_{\mathrm{s}*})$ have correlation ρ_α and ρ_β, respectively, and derive the moments of r$*$ and u using this correlation.

Figure 6.9 and 6.10 plot the predicted effects of one model against the other. The more correlated they are, the easier will be to predict one using the other.

Figure 6.9: Comparison of Section Effects

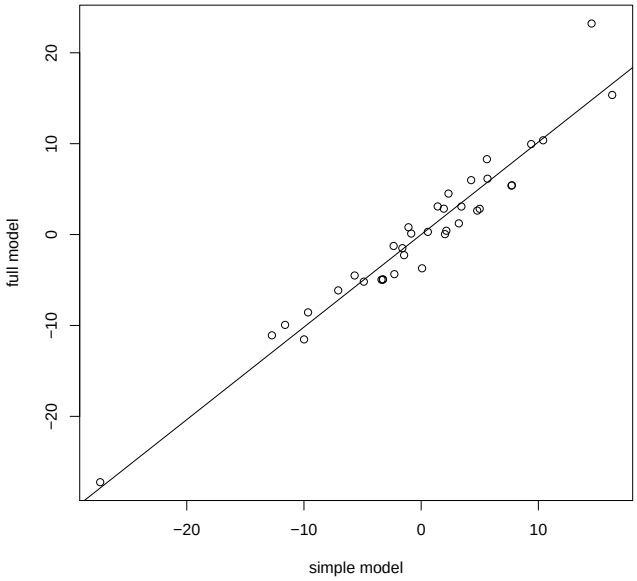
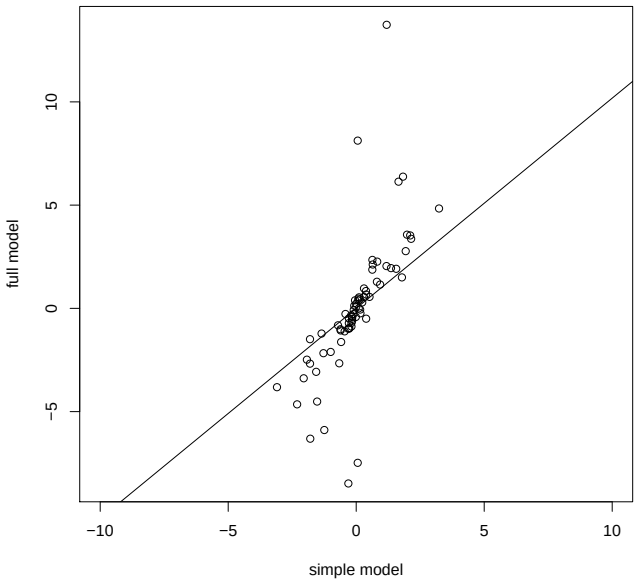


Figure 6.10: Comparison of Direction Effects



We can define the total residuals is the simple model like this:

$$\mathrm{r_{sdi*}} = \mathrm{Y_{sdi}} - \mathrm{X_{sdi}}\beta_*$$

We note that $\mathrm{r_{sdi*}}$ can be calculated easily in the validation sample. The variance of one of those residual is given by:

$$\mathrm{Var}(\mathrm{r_{sdi*}}) = \sigma^2_{\mathrm{s*}} + \sigma^2_{\mathrm{d*}} + \sigma^2_*$$

the covariance of two residuals in the same direction is:

$$\mathrm{Cov}(\mathrm{r_{sdi*}}, \mathrm{r_{sdj*}}) = \sigma^2_{\mathrm{s*}} + \sigma^2_{\mathrm{d*}}$$

The covariance of two residuals not in the same direction is:

$$\mathrm{Cov}(\mathrm{r_{sdi*}}, \mathrm{r_{sd'i*}}) = \sigma^2_{\mathrm{s*}} + \sigma^2_{\mathrm{d*}}$$

The covariance between residuals and random effects are given by:

$$\mathrm{Cov}(\mathrm{r_{sdi*}}, \alpha_{\mathrm{s}}) = \mathrm{Cov}(\alpha_{\mathrm{s*}} + \beta_{\mathrm{d*}} + \epsilon_{\mathrm{sdi*}}, \alpha_{\mathrm{s}}) = \mathrm{Cov}(\alpha_{\mathrm{s*}}, \alpha_{\mathrm{s}}) = \rho_\alpha \sigma_{\mathrm{s}} \sigma_{\mathrm{s*}} = \sigma_{\mathrm{ss*}}$$

$$\mathrm{Cov}(\mathrm{r_{sdi*}}, \beta_{\mathrm{d}}) = \mathrm{Cov}(\alpha_{\mathrm{s*}} + \beta_{\mathrm{d*}} + \epsilon_{\mathrm{sdi*}}, \beta_{\mathrm{d}}) = \mathrm{Cov}(\beta_{\mathrm{s*}}, \beta_{\mathrm{s}}) = \rho_\beta \sigma_{\mathrm{d}} \sigma_{\mathrm{d*}} = \sigma_{\mathrm{dd*}}$$

Therefore, the joint covariance of r_* and $\mathrm{u} = (\alpha_{\mathrm{s}}, \beta_1, \beta_2)$ is described by the following variance components:

D is the same as before:

$$\mathrm{D} = \begin{bmatrix} \sigma^2_{\mathrm{s}} & 0 & 0 \\ 0 & \sigma^2_{\mathrm{d}} & 0 \\ 0 & 0 & \sigma^2_{\mathrm{d}} \end{bmatrix}$$

V is different because the random effects in r are from the simpler model:

$$\mathrm{V} = \sigma^2_{\mathrm{s}*} + \begin{bmatrix} \sigma^2_{\mathrm{d}*} + \sigma^2_* & \sigma^2_{\mathrm{d}*} & \sigma^2_{\mathrm{d}*} & 0 & 0 & 0 \\ \sigma^2_{\mathrm{d}*} & \sigma^2_{\mathrm{d}*} + \sigma^2_* & \sigma^2_{\mathrm{d}*} & 0 & 0 & 0 \\ \sigma^2_{\mathrm{d}*} & \sigma^2_{\mathrm{d}*} & \sigma^2_{\mathrm{d}*} + \sigma^2_* & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{\mathrm{d}*} + \sigma^2_* & \sigma^2_{\mathrm{d}*} & \sigma^2_{\mathrm{d}*} \\ 0 & 0 & 0 & \sigma^2_{\mathrm{d}*} & \sigma^2_{\mathrm{d}*} + \sigma^2_* & \sigma^2_{\mathrm{d}*} \\ 0 & 0 & 0 & \sigma^2_{\mathrm{d}*} & \sigma^2_{\mathrm{d}*} & \sigma^2_{\mathrm{d}*} + \sigma^2_* \end{bmatrix}$$

C is also affected:

$$\mathrm{C} = \begin{bmatrix} \sigma_{\mathrm{ss}*} & \sigma_{\mathrm{ss}*} & \sigma_{\mathrm{ss}*} & \sigma_{\mathrm{ss}*} & \sigma_{\mathrm{ss}*} & \sigma_{\mathrm{ss}*} \\ \sigma_{\mathrm{dd}*} & \sigma_{\mathrm{dd}*} & \sigma_{\mathrm{dd}*} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\mathrm{dd}*} & \sigma_{\mathrm{dd}*} & \sigma_{\mathrm{dd}*} \end{bmatrix}$$

6.6 Results: quantiles calculation

Tables 6.1 through 6.8 report predicted speed deciles. A number of patterns can be observed. First, some sections are predicted accurately. For example, sections 7, 13, 18 and 27 have good predictions within 2-3 miles per hour for almost all the quantiles. Most central deciles (40^{th} to 60^{th} quantiles) have good predictions. However, this is not enough to the scope of this research, which aims at estimating quantiles of speed distributions. Furthermore, the additional observations collected to predict random effects could be used to predict the mean speed of the section with some accuracy. The worst value obtained for the predicted median (50^{th} quantile) is for section 29, direction 1, with an observed median of 93.1 and a prediction of

88. This compare with the 10^{th} quantile that predicts 63.2 for an observed value of 72.3.

Second, we can observe that some sections are likely to have an error in part due to the prediction error of the random effects. One way to assess this is to observe the extreme quantiles and check what sections appear to underestimate both the 10^{th} and the 90^{th} quantiles. Two good examples of this are section 22, direction 1 and section 23, direction 1. In the first case most deciles are underestimated and in the second case they are overestimated. Going back to the plots of predicted random effects, we see in figure 6.6 that the direction effects are not so precise with only five observations and this is likely a case where the prediction has created a small error. For section 23 we see in figure 6.7 that the section effect and one of the direction effect are overestimated with 5 observations, which would explain this component of the prediction error.

Third, the most obvious prediction error is the overestimation of low deciles and the underestimation of high deciles, or vice-versa. This can be observed for example in section 10, direction 1 that underestimates the 10^{th} quantile by 5.3 but overestimates the 90^{th} quantile by 8.6. There might still be an error in the predicted random effect for this section but it cannot be fixed because the random effect is a constant added to all predictions in the same section and direction. The most likely cause for this kind of error lies in the estimated jackknife coefficients. All sections with severe quantile prediction errors also have a lot of jackknife coefficients that greatly differ from the full sample coefficients. We think that this is the most important source of error.

Table 6.1: Predicted quantiles sections 1,3,5 and 7

quantile	Section 1				Section 3				Section 5				Section 7			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.
10	52.4	56	51.3	54	66.2	58	67	63	63.8	60	60.8	57	64.3	66	64.6	72.6
20	58.2	59	57.6	58	69.3	63	70.4	68	67.5	64	65.8	63	70.1	69	70.7	76
30	62.4	62	62.1	60	71.5	66	72.8	71	70.2	68	69.5	67	74.3	73	75.1	79.8
40	66	65.8	66	62	73.4	69	74.9	74	72.4	71	72.6	70	77.8	77	78.9	82
50	69.3	67	69.6	65	75.2	72	76.9	77	74.6	73	75.5	75	81.1	80	82.4	85
60	72.6	69	73.2	67	77	74	78.8	79	76.7	77	78.4	77	84.5	82	85.9	87
70	76.2	71	77.1	70	78.9	77	80.9	82	79	81	81.5	81.9	88	86.8	89.7	89
80	80.4	73	81.6	74	81.2	81	83.4	85	81.6	85	85.1	87	92.2	90	94.1	93
90	86.2	77.3	87.9	79	84.3	88	86.8	89	85.3	90	90.1	94	98	95	100.2	97.4

Table 6.2: Predicted quantiles sections 8,9,10,11

quantile	Section 8				Section 9				Section 10				Section 11			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.
10	68.1	69.4	68.1	74.2	61.9	60	66.5	70	44.7	50	50.7	56	54.5	59.3	55.2	59.3
20	73.7	75.4	73.6	77	67.8	67	71.2	72.8	50.2	51.2	54.8	60	61.4	63.2	61.8	61.7
30	77.8	81	77.6	77.6	72.1	68	74.6	76	54.2	54.6	57.7	63	66.3	65.1	66.6	67.1
40	81.3	84	81	81.6	75.7	73	77.5	79.6	57.6	55	60.2	64	70.6	68.1	70.7	71
50	84.6	86	84.2	85	79.1	75	80.2	82	60.8	56	62.6	65	74.5	70.6	74.5	72.5
60	87.9	88.2	87.4	87.2	82.5	79	82.9	86.4	64	57	64.9	67	78.5	71.9	78.4	74
70	91.3	90	90.8	90	86.2	86	85.8	88	67.5	59	67.4	68	82.7	75.9	82.5	76.4
80	95.4	94.2	94.8	94.8	90.4	87	89.2	91.2	71.5	60.8	70.4	71.2	87.6	81.6	87.3	81.6
90	101.1	103.5	100.3	100.6	96.3	90	94	94.1	77	64.4	74.5	73	94.5	95.6	93.9	93

Table 6.3: Predicted quantiles sections 12,13,14,15

quantile	Section 12				Section 13				Section 14				Section 15			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.
10	63.1	59.7	61.7	59	65.6	62.7	64	65.8	64	67.1	61.1	67.8	67.2	64.3	58.5	65.7
20	67.9	64.2	67.3	67.5	70.3	67.4	69.3	68.5	71	70.6	69.2	72.1	72.5	69.5	66.9	73.8
30	71.3	71.3	71.3	72	73.7	70.7	73.2	72	76.1	76.3	75.1	75.3	76.3	74.3	73	75.6
40	74.3	73.9	74.7	78.3	76.6	75.4	76.4	74.3	80.4	79.6	80.1	79.1	79.6	78	78.1	80.2
50	77	76.9	77.9	82.9	79.3	79.8	79.5	81.9	84.4	83.1	84.8	83	82.6	81.9	83	83.8
60	79.8	85.7	81	86	82	84.2	82.5	84.4	88.5	85.9	89.4	90.6	85.7	84.3	87.8	85.2
70	82.7	88.8	84.5	90	84.9	87.8	85.8	84.8	92.8	91	94.4	94.1	89	88.6	93	88.7
80	86.1	92.1	88.5	91	88.3	89.1	89.7	86.9	97.8	96.3	100.3	97.9	92.8	92.6	99.1	93.7
90	90.9	94.8	94	106.8	93	93.2	95	93.4	104.8	107.1	108.4	105.9	98.1	103.1	107.5	97.2

Table 6.4: Predicted quantiles sections 16,17,18,21

quantile	Section 16				Section 17				Section 18				Section 21			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.
10	64.8	71.7	64.4	71.3	60.4	66.2	63.4	67.8	51.5	51.9	50.5	52.6	73.2	73	73.6	81.7
20	71.5	75.2	71.1	74.6	67.9	69.2	69.9	70.9	58	54.2	57.5	56.6	80	88	80.7	89.2
30	76.4	80.2	76	76.3	73.4	72.9	74.6	74.4	62.8	58.4	62.6	60.9	85	92.5	85.8	95.3
40	80.5	82.5	80.1	79.6	78	74.5	78.6	77.4	66.8	65.3	66.9	62.2	89.2	97	90.1	98.4
50	84.4	85	84	82.2	82.4	76.9	82.3	80	70.5	67.2	70.9	68.5	93.1	98	94.2	104
60	88.3	88.3	87.9	84.4	86.8	85.4	86.1	83.3	74.3	69.7	74.9	71.8	97.1	99	98.3	105
70	92.4	90.5	92	91.5	91.4	92.3	90.1	88.7	78.3	74.9	79.2	75.4	101.3	99.5	102.6	105
80	97.3	93.5	96.9	94.3	96.9	96.3	94.8	92.4	83	80.3	84.2	84.9	106.2	100	107.7	111.8
90	104	108.6	103.6	100.2	104.4	99.3	101.3	94.1	89.6	91.2	91.2	91.7	113.1	105	114.8	114.9

Table 6.5: Predicted quantiles sections 22,23,24,25

quantile	Section 22				Section 23				Section 24				Section 25			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.
10	59.8	64.8	55.8	64	53.3	39.1	50	46	62.5	67	65.6	59.7	58.3	54	59.1	50.4
20	67.1	69	63.9	70	53.9	42.4	52.2	47	68.4	68.6	70.5	66	62	56.2	63	55.2
30	72.5	71.4	69.8	71	54.4	46.1	53.7	48	72.6	69	74.1	75.3	64.8	59.8	65.7	65.4
40	77	78	74.8	72	54.8	46.8	55.1	50	76.2	72.4	77.1	77.4	67.1	62.6	68.1	71.4
50	81.2	82	79.4	72	55.1	47.5	56.4	50.5	79.6	74	80	79	69.2	66	70.3	73
60	85.5	94	84.1	75	55.5	48.2	57.6	51.8	83	79.4	82.8	81.8	71.4	68.6	72.5	78.4
70	90	97.8	89.1	76	55.9	48.9	59	54.1	86.6	81	85.9	88.1	73.7	71	74.9	81.2
80	95.3	101	94.9	79	56.3	50.2	60.6	57.4	90.9	86	89.4	90.2	76.4	75	77.6	84.2
90	102.7	108.4	103	82	56.9	51	62.8	61.5	96.8	99	94.4	96.1	80.2	75	81.5	89

Table 6.6: Predicted quantiles sections 26,27,28,29

quantile	Section 26				Section 27				Section 28				Section 29			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.
10	59.4	59.4	61.9	57.2	41.3	44.6	43	45.8	63.2	67	63.6	76.6	63.2	72.3	64.1	78.5
20	64.6	65.6	66.2	60.6	44.7	46.2	45.8	47.6	71.1	75	71.7	78	71.7	77.1	72.6	80.1
30	68.2	67	69.3	65.5	47.1	47.8	47.9	48.4	76.7	77	77.6	83.8	77.9	84	78.7	82.6
40	71.4	67.6	72	68.8	49.2	49.4	49.6	49.2	81.6	78	82.6	86.4	83.1	89.2	83.9	85.1
50	74.3	73	74.5	76	51.2	51	51.3	51	86.1	87	87.3	87	88	93.1	88.8	85.1
60	77.3	75.8	77	76.8	53.1	53.6	52.9	53.6	90.7	87	91.9	87.6	93	93.1	93.7	93.7
70	80.4	77.1	79.6	77.3	55.2	54.2	54.6	55.6	95.5	88.5	97	91	98.2	93.1	98.9	101.2
80	84.1	80.8	82.8	80.4	57.7	55	56.7	56.8	101.2	89	102.8	96.8	104.4	94.9	105	103.9
90	89.2	86.7	87.1	86.6	61.1	60	59.5	60.2	109.1	103	111	100.8	112.9	109.7	113.5	108.8

Table 6.7: Predicted quantiles sections 30,31,32,33

quantile	Section 30				Section 31				Section 32				Section 33			
	dir 1		dir 2		dir 1		dir 2		dir 1		dir 2		dir 1		dir 2	
	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.
10	60.8	56.3	65.3	57.4	60.4	61.2	63.2	67.2	40.4	52.1	46.5	55.2	53.4	52.7	57	58.2
20	63.3	56.3	66.6	62.1	65.3	66.5	67.3	68.9	49.2	57.5	53.1	59.2	58.6	58.1	61.1	60.9
30	65.1	57.5	67.5	62.1	68.9	68	70.3	70	55.5	59.6	57.9	62.2	62.4	60	64	64
40	66.7	61.4	68.3	63.8	71.9	70.3	72.9	73	60.9	59.6	62.1	63.9	65.6	62.7	66.5	67.4
50	68.1	65.9	69	71.8	74.7	71.9	75.3	74.4	65.9	64.2	65.9	67.4	68.6	66.2	68.9	67.4
60	69.6	69.3	69.7	74.1	77.5	73.2	77.7	77.1	71	66.7	69.7	69.3	71.6	67.6	71.2	67.4
70	71.1	69.3	70.5	74.1	80.6	74.6	80.3	79.6	76.4	69.5	73.8	71.4	74.8	71.1	73.7	71.1
80	73	87.2	71.4	79.2	84.1	76.9	83.3	83.4	82.7	79.4	78.7	73.5	78.5	73.8	76.6	71.9
90	75.5	93.2	72.6	79.2	89	84.1	87.4	87.5	91.5	83.4	85.3	77.3	83.7	81.3	80.7	75.3

Table 6.8: Predicted quantiles sections 34,36

quantile	Section 34				Section 36			
	dir 1		dir 2		dir 1		dir 2	
	pred.	emp.	pred.	emp.	pred.	emp.	pred.	emp.
10	44.1	42.5	40.9	43.1	52	44.6	53.6	56.7
20	48.4	46.5	46.5	50.6	57.5	50.9	58.8	59.1
30	51.4	49.1	50.6	52.7	61.5	59.2	62.5	62.2
40	54	51.2	54.1	56.2	64.9	64.2	65.7	65.2
50	56.5	53.2	57.4	58.8	68	68.2	68.7	69.5
60	58.9	54.7	60.7	60.2	71.2	70.9	71.7	72.4
70	61.5	58	64.2	64.9	74.6	74.9	74.9	76.8
80	64.6	62.2	68.3	70.3	78.6	79.3	78.7	78.8
90	68.8	68.4	73.9	76	84.1	94.6	83.9	82.5

6.7 Jackknifed coefficients

The poor predictions, especially for the high and low quantiles, in some combinations of section and direction appear to be caused by the instability of the jackknife coefficients used for the validation.

Tables 6.9, 6.10 and 6.11 present the jackknife coefficients used for speed data validation and should be read together (the tables are only split in order to fit in the document). Each line corresponds to estimates with one section removed, except the first one that contains the estimations on the whole sample. Each column corresponds to the estimated coefficients for a specific predictor. The gray cells indicate the coefficients that are very different from the full sample equivalent.

The validation scheme used in this chapter is a variant of the jackknife method in statistics. We have discussed in details in the previous section how to compute out of sample predictors for random effects. However, the prediction of speed quantiles in a new section also requires that the model's coefficients be estimated for the validation sample. Ideally, the estimated coefficients in a given column would be stable and comparable to the estimated coefficient for the whole sample. Large differences between a coefficient estimated with one section out and with the whole sample raises concerns about that section or predictor. For example the jackknife coefficients for `Z_times_SLW` are very similar for all 31 sections removed and with the overall sample, whereas the coefficients for `Z_times_X1_over_R` appear to be off for sections 1,3,5,7,10,12 and 30.

The sections that generate extreme jackknife coefficients are consistently the

same for all predictors. Sections 1,3,5,7,10,15,29,30,32 and 36 generate most of the extreme jackknife coefficients. This might be an indication that these sections behave according to a different model, or that the measurement of predictors and speed data was less reliable on these sections.

Table 6.9: Jackknife coefficients 1

section out	intercept	Zlane	Z_times_X1_over_R	Z_times_SLS	Z_times_SBLS	Z_times_IDRS	Z_times_DLS	Z_times_Ped	Z_times_SRS	Z_times_SBRS	Z_times_SRW	Z_times_SLW
none	79.34	10.99	-134.12	-2.82	-1.27	-0.13	-2.34	1.38	-3.14	-1.44	-14.4	11.94
1	79.8	6.88	-64.43	-1.63	-1.89	-0.05	-1.92	1.51	-2.22	-1.85	-15.11	11.13
3	79.52	19.35	-88.29	-2.19	-1.68	-0.22	-2.3	1.62	-2.5	-1.91	-13.44	12.7
5	79.49	16.02	-88.36	-1.97	-1.68	-0.27	-2.49	1.92	-2.47	-1.26	-13.96	12.36
7	79.22	10.67	-94.03	-2.64	-0.44	-0.03	-2.23	1.7	-3.01	-0.85	-14.5	12.06
8	79.1	11.27	-143.02	-2.77	-1.16	-0.12	-2.33	1.46	-3.1	-1.35	-14.4	11.97
9	79.31	10.92	-135.47	-2.78	-1.21	-0.11	-2.31	1.36	-3.2	-1.54	-14.39	11.94
10	79.21	20.94	-92.76	-2.04	-1.32	-0.33	-2.68	1.99	-2.28	-1.52	-13.87	12.32
11	79.54	11.26	-133.6	-2.89	-1.27	0.02	-2.28	1.12	-3.28	-1.53	-14.4	11.94
12	79.29	6.79	-108.67	-2.98	-1.62	0.18	-2.53	0.72	-3.54	-1.88	-14.43	11.86
13	79.34	11.01	-129.85	-2.82	-1.33	-0.17	-2.34	1.46	-3.12	-1.51	-14.36	11.96
14	79.12	12.74	-133.41	-2.65	-1.15	-0.09	-2.07	1.39	-3.3	-1.5	-14.31	11.97
15	79.23	10.55	-146.97	-2.43	-1.03	-0.21	-5.05	1.74	-3.1	-1.55	-14.62	11.69
16	79.12	13.2	-137.08	-2.85	-1.41	-0.12	-2.14	1.49	-3.16	-1.55	-14.31	12.03
17	79.23	11.44	-132.53	-2.77	-1.05	-0.18	-2.3	1.19	-3.23	-1.31	-14.45	11.89
18	79.68	11.87	-139.33	-2.71	-1.21	-0.12	-2.33	1.47	-3.09	-1.38	-14.43	11.92
20	78.41	9.39	-135.25	-2.76	-1.1	-0.2	-2.15	1.43	-3.13	-1.39	-14.48	11.87
21	78.72	9.76	-134.22	-3	-1.45	-0.02	-2.45	1.02	-3.29	-1.64	-14.4	11.93
22	79.33	9.46	-127.27	-2.74	-1.03	-0.14	-2.37	1.3	-3.26	-1.47	-14.47	11.89
23	80.42	12.08	-127.73	-2.42	-1.13	-0.27	-2.52	2.29	-2.72	-1.24	-14.47	11.8
24	79.34	10.84	-134.9	-2.83	-1.14	-0.14	-2.35	1.55	-3.05	-1.4	-14.46	11.91
25	79.74	11.12	-129.57	-2.94	-1.63	-0.06	-2.53	1.1	-3.25	-1.74	-14.43	11.85
26	79.59	10.6	-133.18	-2.97	-1.19	-0.12	-2.44	1.44	-2.98	-1.46	-14.46	11.9
27	79.48	11.05	-135.3	-2.84	-1.3	-0.12	-2.31	1.3	-3.16	-1.46	-14.38	11.97
28	79.09	8.72	-123.17	-2.94	-1.45	-0.08	-1.68	1.2	-3.36	-1.79	-14.4	11.86
29	78.94	12.35	-137.62	-2.58	-1.15	-0.36	-1.98	2.02	-2.9	-1.15	-13.95	12.32
30	79.37	7.44	-73.32	-4.76	-2.23	-0.32	-2.1	-2.06	-4.79	-2.25	-14.3	11.95
31	79.52	11.39	-135.74	-2.64	-1.36	-0.11	-2.25	1.33	-3.29	-1.41	-14.35	11.99
32	79.47	11.94	-97.53	-3.01	-1.06	0.17	-1.42	4.08	-3.45	-1.14	-14.34	11.94
33	79.78	11.29	-134.52	-2.81	-1.33	-0.12	-2.39	1.31	-3.23	-1.44	-14.41	11.93
34	79.34	11.06	-134.27	-2.84	-1.3	-0.12	-2.31	1.4	-3.17	-1.45	-14.38	11.97
36	79.4	13.11	-129.52	-3	-1.58	-0.16	-2.65	-0.03	-2.05	-0.28	-14.39	11.85

Table 6.10: Jackknife coefficients 2

section out	Z_times_ILS	Z_times_PSL	Z_times_DDRS	Z_times_DRS	Z_times_IRS	Z_times_SR	Z_times_PKLLS	Z_times_TRDLS	Z_times_RLS	Z_times_LW	Z_times_WRS	Z_times_IDLS
none	-1.37	0.15	1	-2.8	-2.4	3.99	-3.3	-4.21	5.58	-1.84	-3.38	0.22
1	-1.08	0.13	0.9	-2.78	-2.18	4.97	-6.59	-2.31	2.19	0	-4.93	0.29
3	-1.06	0.14	0.94	-2.88	-1.93	0.88	-2.98	-2.18	2.07	-3.71	-2.27	0.19
5	-0.8	0.13	0.99	-3.17	-2.11	5.87	-3.27	-5.08	7.46	-3.59	-1.48	0.03
7	-1.02	0.17	1.08	-3.07	-2.65	4.68	-3.66	-4.71	5.09	-2.3	-3.39	0.03
8	-1.32	0.15	1.01	-2.84	-2.41	4.01	-3.35	-4.31	5.58	-1.97	-3.29	0.2
9	-1.3	0.15	1.01	-2.85	-2.46	3.96	-3.34	-4.19	5.53	-1.77	-3.47	0.21
10	-1.08	0.13	0.95	-2.95	-1.78	1.46	-3.2	-5.1	7.57	-3.83	-1.94	0.06
11	-1.43	0.15	1.04	-2.83	-2.55	4	-3.26	-4.13	5.43	-1.84	-3.3	0.21
12	-1.69	0.14	1.14	-3.38	-3.15	3.95	-3.81	-3.85	4.87	-0.2	-3.53	0.47
13	-1.58	0.15	1.02	-2.84	-2.44	3.92	-3.45	-4.21	5.6	-1.78	-3.35	0.3
14	-1.34	0.15	1.1	-2.95	-2.67	3.95	-3.57	-4.19	5.56	-2.42	-3.12	0.18
15	-0.71	0.16	0.96	-0.79	-2.36	4.4	-3.3	-6.53	10.16	-1.69	-3.82	0.04
16	-1.65	0.14	0.96	-2.64	-2.67	3.84	-2.89	-4.03	5.24	-2.17	-2.87	0.24
17	-1.24	0.16	1.09	-2.95	-2.27	4.24	-3.65	-4.29	5.74	-2.33	-3.79	0.24
18	-1.28	0.15	1.03	-2.92	-2.37	3.92	-3.52	-4.2	5.55	-2.11	-3.24	0.2
20	-1.01	0.16	0.98	-3.15	-2.14	4.29	-3.32	-4.11	5.35	-1.66	-3.9	0.12
21	-1.51	0.14	0.98	-2.77	-2.48	3.95	-3.13	-4.16	5.55	-1.29	-3.77	0.35
22	-1.13	0.16	1.06	-2.88	-2.36	4.35	-3.47	-4.34	5.84	-1.68	-3.84	0.14
23	-1.13	0.17	1.02	-2.97	-2.26	4.24	-3.79	-4.61	6.5	-2.71	-2.7	-0.02
24	-1.27	0.15	1	-2.8	-2.3	4.11	-3.37	-4.29	5.74	-1.93	-3.41	0.15
25	-1.48	0.14	1.02	-3.03	-2.48	3.73	-3.32	-4.02	5.37	-1.44	-3.19	0.31
26	-1.43	0.15	1	-2.76	-2.28	4.1	-3.31	-4.29	5.72	-1.77	-3.49	0.19
27	-1.38	0.15	1	-2.78	-2.4	3.93	-3.23	-4.21	5.58	-1.82	-3.34	0.22
28	-1.51	0.15	0.97	-3.7	-2.71	4.26	-2.92	-3.31	3.79	-0.92	-3.74	0.27
29	-1.13	0.18	1.03	-2.32	-2.13	3.73	-3.47	-5.64	8.62	-3.35	-2.69	0.01
30	-1.27	0.11	0.91	-2.29	-1.96	3.72	0.68	-3.31	4.14	0.31	-3.74	0.17
31	-1.26	0.15	1	-2.84	-2.54	3.87	-3.22	-4.15	5.45	-1.9	-3.15	0.21
32	-1.49	0.17	0.98	-2.49	-2.73	4.19	-5.47	-4.23	5.63	-2.92	-4.5	0.45
33	-1.48	0.15	0.99	-2.79	-2.37	3.94	-3.2	-4.16	5.49	-1.83	-3.33	0.25
34	-1.37	0.15	1	-2.78	-2.41	3.93	-3.22	-4.21	5.58	-1.82	-3.33	0.22
36	-2.26	0.18	0.74	-2.19	-1.97	4.17	-1.77	-4.46	6.07	-3.25	-3.48	0.58

Table 6.11: Jackknife coefficients 3

section out	Z_times_LG	Z_times_DDLS	Z_times_delta_PSL	Z_times_WLS	Z_times_PedD	Z_times_PKLRS	X1_over_R	PedD	Z_times_TRDRS	Z_times_RRS	Z_times_SPRS
none	0.09	0.11	-0.11	-1.99	0.73	1.63	-1949.12	-7.6	-2.53	3.74	0.4
1	0.09	-0.11	-0.12	-4.32	-0.31	-0.34	-2044.73	-7.79	-0.76	0.61	1.11
3	0.12	0.07	-0.06	-0.48	0.51	1.99	-1985.25	-7.67	1.09	-3.04	0.45
5	0.09	0.05	-0.1	-0.73	0.33	1.81	-1979.94	-7.66	-3.49	5.8	0.37
7	-0.07	0.06	-0.2	-3.5	0.73	1.65	-2280.23	-7.56	-2.91	3.48	0.43
8	0.08	0.11	-0.12	-2.04	0.72	1.64	-1899.47	-7.51	-2.63	3.77	0.42
9	0.1	0.1	-0.11	-1.9	0.73	1.67	-1946.11	-7.59	-2.55	3.75	0.37
10	0.15	0.09	-0.09	-0.12	0.31	1.71	-1216.14	-7.56	-3.61	5.92	0.47
11	0.09	0.09	-0.1	-2.29	0.75	1.9	-1990.04	-7.68	-2.39	3.47	0.48
12	0.08	0.21	-0.07	-2.21	0.57	1.26	-1939.59	-7.59	-2.44	3.6	0.4
13	0.09	0.12	-0.1	-1.6	0.69	1.68	-1948.84	-7.6	-2.56	3.8	0.44
14	0.11	0.05	-0.09	-1.72	0.73	2.29	-1904.54	-7.52	-2.65	3.93	0.62
15	0.08	0.1	-0.19	-1.8	0.69	0.74	-1925.73	-7.56	-0.37	-0.63	-0.22
16	0.09	0.09	-0.07	-1.47	0.75	1.97	-1904.2	-7.52	-2.41	3.52	0.37
17	0.09	0.11	-0.13	-2.14	0.77	2.02	-1927.38	-7.56	-2.61	3.9	0.57
18	0.09	0.09	-0.12	-1.98	0.68	1.73	-2019.71	-7.73	-2.57	3.81	0.52
20	0.08	0.11	-0.16	-2.3	0.8	1.28	-1758.79	-7.25	-2.97	4.6	0.18
21	0.09	0.13	-0.11	-2.22	0.76	1.53	-1821.74	-7.37	-2.51	3.78	0.29
22	0.08	0.11	-0.15	-2.2	0.78	1.52	-1947.73	-7.6	-2.57	3.83	0.23
23	0.08	0.12	-0.14	-1.77	0.64	1.13	-2171.51	-8.02	-2.91	4.65	0.46
24	0.08	0.09	-0.13	-2.11	0.75	1.6	-1948.27	-7.6	-2.52	3.73	0.37
25	0.09	0.13	-0.09	-1.88	0.69	1.7	-2031.33	-7.76	-2.36	3.58	0.46
26	0.08	0.11	-0.12	-2.05	0.74	1.52	-1999.88	-7.7	-2.43	3.56	0.31
27	0.09	0.11	-0.11	-1.97	0.79	1.72	-1978.23	-8.74	-2.54	3.75	0.41
28	0.09	0.16	-0.11	-1.76	0.79	1.16	-1898.56	-7.51	-3.29	5.25	-0.06
29	0.09	0.07	-0.1	-1.72	0.92	2.26	-1866.59	-7.45	-3.64	6.13	0.81
30	0.1	0.09	-0.05	-2.15	2.39	5.53	-1956.46	-7.57	-1.52	2.04	0.52
31	0.09	0.12	-0.1	-1.99	0.72	1.68	-1986.33	-7.67	-2.64	3.95	0.46
32	0.08	-0.13	-0.1	-4.06	0.22	1.37	-1976.37	-7.45	-2.66	4.04	1.26
33	0.09	0.13	-0.11	-1.94	0.74	1.69	-2040.42	-7.77	-2.46	3.61	0.43
34	0.09	0.11	-0.11	-2	0.68	1.72	-1949.95	-7.68	-2.54	3.75	0.42
36	0.13	0.33	-0.12	-3.14	1.11	1.71	-1961.17	-7.54	-2.86	4.3	0.93

6.8 Conclusions

In this chapter we have explored methods to predict speed distributions using mixed linear models.

The difficulties of making prediction were twofold. First we are using a model with random effects and predictions for new road sections require some way to predict these random effects. Second, we are using normal quantiles of the dependent variables to build more predictors, effectively making out-of-sample predictions of random effects more difficult.

We have discussed that it was necessary to observe at least some speed data in the section for which decile predictions were computed. To overcome the problem created by the use of normal quantiles in the calculation of residuals, we proposed the use of a simple auxiliary model. We have shown how to set the relationship between this auxiliary model and the full actual model used in prediction in order to derive the *best linear unbiased predictor* (BLUP) in this context. We observed that this method was not performing well when used to make random effect predictions for a model that ignored the sampling design of the data, but turned out to be very precise when the full sampling sampling design was accounted for in the model.

The other part that was required to validate the model are the coefficients associated with the variables in the model. We have seen that their estimation was roughly stable except for some variables that generated more extreme coefficients for some sections. The sections with extreme coefficients were consistently the same for all the affected variables.

By applying a jackknife technique we obtained some interesting results. Some road sections were predicted satisfactorily. The sections that presented large errors in the prediction of random effect were mostly affected by errors in the coefficients of the model. Some sections appeared to have a disproportionate effect on the model, suggesting that they should be modeled in a different way.

Chapter 7: Conclusions

7.1 Summary

This dissertation has proposed econometric techniques to model decisions that involve both discrete and continuous dependent variables and a mixed linear model that accounts for the survey design. Procedures to apply these models in a predictive context have been formulated and applied to simulated data and real case studies.

Probit and ordered probit are estimated jointly with linear regression by correlating the error terms, that are assumed to follow a multivariate normal distribution. It has been shown that the existence of a closed form for the choice probability of ordered models makes them much more stable numerically. On the other hand, the calculation of choices probabilities in the probit model, that involves the integration of multivariate normal probabilities, has resulted to be challenging. A method, first suggested by Genz, has proved to greatly reduce the simulation error compared to a naive Monte-Carlo approach, to be more stable and generally more suited for the problem. There are still, however, several limitations to the use of such models, notably the long computation time and the numerical instability.

Discrete models are peculiar to use for predictions because they only produce choice probabilities. In a cross-validation scheme actual choices have been compared

to predicted probabilities. Market shares of alternatives are taken into account when looking at predicted probabilities, mostly because it is very hard to predict a choice that has a very low occurrence. It was found that both discrete-continuous models only improves marginally the predictive power of vehicle ownership and use models estimated on 2009 NHTS data, and that ordered models appear to have a slightly better predicting power. However, results obtained on simulated data attest that improvements that these models offer can be significant when high correlation exists between the continuous and the discrete variables.

The second part of this dissertation reviewed the use of linear mixed models in the context of free-flow speed distribution estimation. Random effect models offer the possibility to account for the sampling design of our data, that contain multiple observations for the same direction, road and section. A variable selection methods based on BIC has been used to select the final model specification. Among the results obtained, it was found that there is no contribution from the road to free-flow speed and that instead only road sections have an impact, that is sometimes very high. The two directions on a road section do have an impact on speeds, although this is more limited.

Finally, a method to predict the distribution of free-flow speeds on new road sections is offered. An auxiliary model that deals with the unavailability of certain predictors in the validation sample is used to predict the value of the random effects. The methods was successfully applied to speed distribution predictions on most of the road sections in our sample, although significant errors were calculated in a limited number of other sections. A jackknife based analysis has shown that

predictor coefficients are likely the cause of this problem, as some road sections have a disproportionate impact on them.

7.2 Contributions

The major contributions from this dissertation can be summarized as follows:

- Models for dependent variables that do not belong to the same family are emerging in the transportation literature and related fields (i.e. marketing, economics). This thesis has offered a theoretical framework, computational tools and numerical results for discrete-continuous models.

- The discrete-continuous models have been not only estimated but also validated to measure their ability to reproduce choice probabilities in validation samples.

- The discrete-continuous models have been applied in the context of car ownership and use, which is a very relevant subject in our societies that are highly dependent on cars and fossil fuel. The models proposed here can be used to calculate energy consumption from private transportation and Greenhouse Gas Emissions. The models are general and can be applied in different contexts and in different disciplines.

- The Random Effect model for free-flow speed distribution represents a significant step in the ability to understand, model and transfer operational speed measurements. This method was suggested by several analysts, but never fully

explored in this context. Moreover, the transferability study offers the possibility to reduce cost associated with data collection and to extend on a large scale the results obtained.

- The techniques developed in this thesis are highly multi-disciplinary. Elements from statistics, econometrics, optimization, survey design and transportation have been assembled to produce these results.

7.3 Future work

There any many questions left unanswered in this document.

The unordered discrete-continuous model suffers from numerical problems in estimation. The application of more advanced optimization procedures should be investigated to solve numerical instability and to compute standard errors for model parameters. The discrete-continuous models can be further expanded to take into account heterogeneity in the form of latent classes or random parameters. Although these extensions are possible, the numerical problems and the long computation time might prevent practical applications. The proposed models are general and can be applied to several transportation problems that include discrete and continuous decision variables: activity type and duration; number of trips and distance traveled, departure time and trip duration.

Model selection is just partially addressed in this dissertation; other methods than BIC could be implemented and their effectiveness explored with respect to model fit and model prediction. One problem that has not been studied is the

incorporation of sampling weights in the analysis, which would make inferences to the whole population more robust. The modeling approach for free-flow speed has proved to be useful, but more validation is needed in order to make the coefficients of the model robust. This may involve the use of mixture of models. Finally, more realistic validation runs can be performed by collecting a handful of observations on new roads.

Bibliography

[AYEMM14] S. Anowar, S. Yasmin, N. Eluru, and L. Miranda-Moreno. Analyzing car ownership in Quebec City: a comparison of traditional and latent class ordered and unordered models. *Transportation*, 41(5):1013–1039, 2014.

[BAS+14] C.R. Bhat, S. Astroza, R. Sidharthan, M. Jobair Bin Alam, and W.H. Khushefati. A joint count-continuous model of travel behavior with selection based on a multinomial probit residential density choice model. *Transportation Research Part B*, 68:31–51, 2014.

[BBA01] J.L. Bowman and M.E. Ben-Akiva. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A*, 31:1–28, 2001.

[BCMnt] M. Bassani, C. Cirillo, and J.M. Molinari, S.and Tremblay. Random effect models to predict operating speed distribution on rural two-lane highways. *Journal of Transportation Engineering*, 0:0, in print.

[BCMV05] A.M. Bento, M.L. Cropper, A.M. Mobarak, and K. Vinha. The impact of urban spatial structure on travel demand in the United States. *Review of Economics and Statistics*, 87:466–478, 2005.

[BDMC14] M. Bassani, D. Dalmazzo, G. Marinelli, and C. Cirillo. The effects of road geometrics and traffic regulations on driver-preferred speeds in northern Italy. An exploratory analysis. *Transportation Research Part F*, 25:10–26, 2014.

[Bha05] C.R. Bhat. A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B*, 39(8):679–707, 2005.

[Bha11] C.R. Bhat. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B*, 45(7):923–939, 2011.

[Bha15] C.R. Bhat. A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B*, forthcoming:0, 2015.

[BK90] D.S. Bunch and R. Kitamura. Multinomial probit model estimation revisited: testing estimable model specifications, maximum likelihood algorithms and probit integral approximations for trinomial models of car ownership. Technical report, Institute of Transportation Studies Technical Report, University of California, Davis, CA, 1990.

[BK93] C.R. Bhat and F.S. Koppelman. An endogenous switching simultaneous equation system of employment, income, and car ownership. *Transportation Research Part a-Policy and Practice*, 27:447–459, 1993.

[BM13] M. Bassani and G. Mutani. Effects of environmental lighting conditions on operating speeds on urban arterials. *Transportation Research Record: Journal of the Transportation Research Board*, 2298:78–87, 2013.

[BMBW15] D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2015. R package version 1.1-9.

[Bon01] J.A. Bonneson. Controls for horizontal curve design. *Transportation Research Record: Journal of the Transportation Research Board*, 1751:82–89, 2001.

[BP98] C.R. Bhat and V. Pulugurta. A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B*, 32(1):61–75, 1998.

[BPPL13] C.R. Bhat, R. Paleti, R.M. Pendyala, and K.C. Lorenzini, K and. Konduri. Accommodating immigration status and self selection effects in a joint model of household auto ownership and residential location choice. *Transportation Research Record*, 2382:142–150, 2013.

[BS06] C.R. Bhat and S. Sen. Household vehicle type holdings and usage: an application of the multiple discrete-continuous extreme value (MDCEV) model. *Transportation Research Part B*, 40:35–53, 2006.

[BSE09] C.R. Bhat, S. Sen, and N. Eluru. The impact of demographics, built environment attributes, vehicle characteristics, and gasoline prices on household vehicle holdings and use. *Transportation Research Part B*, 43(1):1–18, 2009.

[Chu02] Y.B. Chu. Automobile ownership analysis using ordered probit models. travel demand and land use. *Planning and Administration*, 2002:60–67, 2002.

[CJM12] J.Y.J. Chow, R. Jayakrishnan, and H. Mahamassani. Is transport modeling education too multi-disciplinary? a manifesto on the search for its evolving identity, 2012.

[CLT16] C. Cirillo, Y. Liu, and J.M. Tremblay. Simulation, numerical approximation and closed forms for joint discrete continuous models with an application to household vehicle ownership and use. *Transportation*, 2016.

[CMH07] X.Y. Cao, P.L. Mokhtarian, and S.L. Handy. Cross-sectional and quasi-panel explorations of the connection between the built environment and auto ownership. *Environment and Planning A*, 39:830–847, 2007.

[CVC88] P. Capéraà and B. Van Cutsem. *Méthodes et modèles en statistique non paramétrique exposé fondamental*. Presses de l'Université Laval, Paris, France, first edition, 1988.

[CXB16] C. Cirillo, R. Xu, and F. Bastin. A dynamic formulation for car ownership modeling. *Transportation Science*, 50(1):322–335, 2016.

[DBS77] C. Daganzo, F. Bouthelier, and Y. Sheffi. Multinomial probit and qualitative choice: A computationally efficient algorithm. *Transportation Science*, 11:338–358, 1977.

[DG97] J. Dargay and D. Gately. Vehicle ownership to 2015: Implications for energy use and emissions. *Energy Policy*, 25:1121–1127, 1997.

[dJ89a] G.C. de Jong. *Simulating car cost changes using an indirect utility model of car ownership and car use;*. PTRC SAM, Brighton, 1989.

[dJ89b] G.C. de Jong. *Some joint models of car ownership and car use; Ph.D. thesis*. Faculty of Economic Science and Econometrics, University of Amsterdam, 1989.

[dJ91] G.C. de Jong. An indirect utility model of car ownership and car use. *European Economic Review*, 34(5):971–985, 1991.

[dJ96] G.C de Jong. A disaggregate model system of vehicle holding duration, type choice and use. *Transportation Research Part B*, 30:263–276, 1996.

[DM84] J.A. Dubin and D.L. McFadden. An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52(2):345–362, 1984.

[ET93] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman et Hall, 1993.

[Fan80] H.A. Fang. A discretecontinuous model of households vehicle choice and usage, with an application to the effects of residential density. *Transportation Research Part B*, 42:736–758, 1980.

[GDH04] P. Goodwin, J. Dargay, and M. Hanly. Elasticities of road traffic and fuel consumption with respect to price and income: A review. *Transport Reviews*, 24(3):275–292, 2004.

[Gek89] J. Gekewe. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57:1317–1339, 1989.

[Gen92] A. Genz. Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149, 1992.

[Gew96] J Geweke. *Monte-Carlo simulation and numerical integration*, pages 731–800. Handbook of Computational Economics. Elsevier Science, Amsterdam, 1996.

[Gol90] T.F. Golob. The dynamics of household travel time expenditures and car ownership decisions. *Transportation Research Part a-Policy and Practice*, 24:443–463, 1990.

[GV89] T.F. Golob and L. Vanwissen. A joint household travel distance generation and car ownership model. *Transportation Research Part B-Methodological*, 23:471–491, 1989.

[Han84] W.M. Hanemann. Discrete continuous models of consumer demand. *Econometrica*, 52:541–561, 1984.

[Has04] Y. Hassan. Highway design consistency: Refining the state of knowledge and practice. *Transportation Research Record: Journal of the Transportation Research Board*, 1881:63–71, 2004.

[HBSM92] D.A. Hensher, P.O. Barnard, N.C. Smith, and F.W. Milthorpe. *Dimensions of automobile demand; a longitudinal study of automobile ownership and use*. North-Holland, Amsterdam, 1992.

[HD00] M. Hanly and J.M. Dargay. Car ownership in Great Britain - panel data analysis. activity pattern analysis and exploration: Travel behavior analysis and modeling. *Planning and Administration*, -99:83–89, 2000.

[HKT01] Y. Hayashi, H. Kato, and R.V.R. Teodoro. A model system for the assessment of the effects of car and fuel green taxes on CO_2 emission. *Transportation Research Part D: Transport and Environment*, 6:123–139, 2001.

[HM98] V. Hajivassiliou and D. McFadden. The method of simulated scores for the estimation of LDV models. *Econometrica*, 66:863–896, 1998.

[HMR96] V. Hajivassiliou, D. McFadden, and P. Rudd. Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics*, 72:85–134, 1996.

[HSD82] J. Horowitz, J. Sparmann, and C. Daganzo. An investigation of the accuracy of the Clark approximation for the multinomial probit model. *Transportation Science*, 16:382–401, 1982.

[KB92] R. Kitamura and D.S. Bunch. *Heterogeneity and state dependence in household car ownership: A panel analysis using ordered-response probit models with error components*. Elsevier, Amsterdam, 1992.

[Ker91] E.M. Keramidas, editor. *Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints*. Fairfax: Interface Foundation of North America, Inc., 1991.

[KGYW99] R. Kitamura, T.F. Golob, T. Yamamoto, and G. Wu. Accessibility and auto use in a motorized metropolis. In *UC Irvine: Center for Activity Systems Analysis*, 1999.

[Kit87] R. Kitamura. A panel analysis of household car ownership and mobility, infrastructure planning and management. In *In Proceedings of the Japan Society of Civil Engineers*, pages 13–27, 1987.

[KK04] H.S. Kim and E. Kim. Effects of public transit on automobile ownership and use in households of the usa. In *RURDS-The Applied Regional Science Conference*, page 245262, 2004.

[Lit13] T. Litman. Understanding transport demands and elasticities: How prices and other factors affect travel behavior. Technical report, Victoria Transport Policy Institute, 2013. pp. 1-76.

[LLM11] S. Li, J. Linn, and E. Muehlegger. Gasoline taxes and consumer behavior. Stanford University, 2011.

[LTC14] Y. Liu, J.M. Tremblay, and C. Cirillo. An integrated model for discrete and continuous decisions with application to vehicle ownership, type and usage choices. *Transportation Research Part A*, 69:319–328, 2014.

[McF89] D. McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.

[MFT05] A. Medina Figueroa and A.P. Tarko. Speed factors on two-lane rural highways in free-flow conditions. *Transportation Research Record: Journal of the Transportation Research Board*, 1912:39–46, 2005.

[MM81] C. Manski and D. McFadden, editors. *Manski, C. and Lerman, S.*, pages 305–319. Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA, 1981.

[MSN08] C.E. McCulloch, S.R. Searle, and J.M. Neuhaus. *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Hoboken, New-Jersey, second edition, 2008.

[MW85] F. Mannering and C. Winston. A dynamic empirical-analysis of household vehicle ownership and utilization. *Rand Journal of Economics*, 16:215–236, 1985.

[PB13] R. Paleti and C.R. Bhat. Integrated model of residential location, work location, vehicle ownership, and commute tour characteristics. *Transportation Research Record*, 2382:162–172, 2013.

[PK08] D. Potoglou and P.S. Kanaroglo. Modelling car ownership in urban areas: a case study of Hamilton, Canada. *Journal of Transport Geography*, 16:42–54, 2008.

[Pur94] L.C. Purvis. Using census public use micro data sample to estimate demographic and automobile ownership models. *Transportation Research Record*, 1443:21–30, 1994.

[R C15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[RAM05] P.E. Rossi, G.M. Allenby, and R.E. McCulloch. *Bayesian statistics and marketing*. Number vol. 13 in Wiley series in probability and statistics. Wiley, 2005.

[RCM09] M.J. Roorda, J.A. Carrasco, and E.J. Miller. An integrated model of vehicle transactions, activity scheduling and mode choice. *Transportation Research Part B-Methodological*, 43:217–229, 2009.

[RH99] J. Ryan and G. Han. Vehicle-ownership model using family structure and accessibility application to honolulu, hawaii. *Transportation Research Record: Journal of the Transportation Research Board*, 1676:1–10, 1999.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[SK12] E. Shay and A.J. Khattak. Household travel decision chains: Residential environment, automobile ownership, trips and mode choice. *International Journal of Sustainable Transportation*, 6:88–110, 2012.

[SW09] S. Srinivasan and J. Walker. Vehicle ownership and mode use: the challenge of sustainability. *Transportation*, 36:367–370, 2009.

[TPH13] A. Texeira-Pinto and J. Harezlak. *Factorization and latent variable models for joint analysis of binary and continuous outcomes*, pages 81–91. Analysis of Mixed Data: Methods & Applications. CRC Press, Taylor & Francis Group, Boca Raton, FL, 2013.

[Tra86] K.E. Train. *Qualitative choice analysis: Theory, econometrics and an application to automobile demand*. The MIT press, Cambridge, Massachusetts, first edition, 1986.

[TRA99] *TRACE: Elasticity Handbook: Elasticities for Prototypical Contexts, European Commission, Directorate-General for Transport*, 1999. www.transport-research.info/Upload/Documents/200310/trace.pdf.

[Tra03] Transportation Research Board of the National Academies. *Design Speed, Operating Speed, and Posted Speed Practices*, number 504 in National Cooperative Highway Research Program, Washington, D.C., 2003.

[Tra09] K.E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, England, second edition, 2009.

[Tra11] Transportation Research Board of the National Academies. *Modeling Operating Speed*, volume E-C151 Synthesis Report of *Transportation Research Circular*, Washington, D.C., 2011.

[UDoT09] Federal Highway Administration U.S. Department of Transportation. National household travel survey, 2009. `http://nhts.ornl.gov`.

[WDLH06] J. Wang, K.K. Dixon, H. Li, and M. Hunter. Operating speed model for low-speed urban tangent streets based on in-vehicle global positioning system data. *Transportation Research Record: Journal of the Transportation Research Board*, 1961:24–33, 2006.

[Whe07] G. Whelan. Modelling car ownership in Great Britain. *Transportation Research Part A-Policy and Practice*, 41:205–219, 2007.