ABSTRACT

Title of Document: A DISCRETE-CONTINUOUS MODELING APPROACH WITH APPLICATIONS TO VEHICLE HOLDING AND USE

Jean-Michel Tremblay, Master of Science, 2012

Thesis directed by: Assistant Professor, Cinzia Cirillo, Department of Civil and Environmental Engineering

Transportation and automobile use is a major concern today in the United-States. The use of automobile has impacts on congestion, urban dynamics, environment and on the economy in general. Good indicators of transportation demand are the number of vehicles owned by a household and the total number of miles traveled.

This thesis aims at building a model that can predict the total vehicle miles traveled and number of cars owned by households, simultaneously. The discrete-continuous model that we present correlates the error terms of a utility-based probit with the error term of an ordinary regression. The objective is to capture the relationship between preferred ownership alternatives and miles traveled.

We successfully show that households with high utility for owning a lot of cars also drive more and that households with high utility for owning few cars drive less. The correlation is between utilities and miles traveled. It also correlates the

two transportation demand indicators without assuming that one precedes the other and, thus, does not suffer from circular variable inclusions.

The thesis ends by incorporating sampling weights into the model before parameters are estimated. We find slight changes in parameters' values calculated with weights. The difference however, is more quantitative than qualitative since the general analysis we make with the weighted coefficients remains the same, only the magnitude of the effects change

A DISCRETE-CONTINUOUS MODELING APPROACH WITH APPLICATIONS TO VEHICLE HOLDING AND USE

by

Jean-Michel Tremblay

Thesis 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
Master of Science
2012

Advisory Committee:
Assistant Professor Cinzia Cirillo, Chair/Advisor
Professor Ali Haghani
Professor Partha Lahiri

Table of Contents

Chapter 1

Introduction

1.1 Motivations

Transportation patterns of western countries have changed drammatically in the last few decades. Most countries have nearly abandoned surface rail systems that were developed in the first half of the twentieth century and have shifted toward the use of private automobiles. Although the tendency to use the automobile is clear, not all regions have the same dependence upon it. Countries like Denmark of the Netherlands and even, to a lesser extend, some regions of the United-States still mantain relatevely high shares of bicycle commuters. In cities like New-York or Montréal, public transportation services are used by a significant number of people and car ownership figures vary accordingly.

Automobile use and dependency have numerous implications. It increases the financial charges of households and individuals that must rely on it. Intensive car-based transportation have a large footprint in cities as more space needs to be allocated to road and parking construction.

Air quality, on a local scale, is greatly affected by the use of fossil fuels that produce acidic rains, dusts, and other pollutants that directly affect individuals' health. Smog episodes indeed pose serious concerns and, more generally, people suffering from asthma or other respiratory diseases are affected by local pollution. On a more global scale, transportation is responsible for a large share of greenhouse gas emissions (about 28% in the US), and there is a general consensus that policies or specific measures should be introduced to reduce emissions from the transportation sector.

In addition, energy prices are expected to increase and car dependent households will have to absorb the raise until they have access to other transportation alternatives. This situation can have negative consequences on the economy as it happened in the 1973 and 1979 oil crisis although the economies of the United-States and most western countries are relatively less dependant on oil than they were then.

Car transportation, as opposed to so-called green commuting, does not favor active lifestyle. With the increasing health care costs faced by developped countries, green commuting is often seen as one of the possible solutions to increase physical activity levels.

From an economic perspective, automobile-based transportation uses resources that are getting scarcer and that are being produced in limited regions of the world. For instance, the United-States are not able to produce all the oil they consume and must rely on imports, which have implications on foreign policies. There are also environmental hazards associated with oil and gas exploration, as reminded by the 2010 oil spill in the Gulf of Mexico. The heavy dependence on fossil fuel has in general a negative impact on the balance of trade for most countries, including the United-States.

Several indicators can be used to quantify automobile use such as the types of vehicles owned, the miles driven by each car in the household, the total number of miles driven, fuel efficiency of cars owned and so on. Some of those variable may be more useful than others and some will be more appropriate for specific problems. For instance, vehicles' body type will certainly be useful for pollution studies but not much for congestion policy analysis.

1.2 Overview

The remaining of this dissertation is organised as follows. Chapter 2 contains a literature review of studies on dicrete-continuous models for the vehicle ownership problem. Chapter 3 proposes the model formulation, and in particular (1) derives the model's properties (2) proposes an innovative kernel density approximation and (3) describes the simulations method adopted to solve this non-closed form maximum log-likelihood estimation problem. Chapter 4 deals with numerical issues concerning the discontinuity of simulated log-likelihood in the estimation of probit models, and proposes Bootstrap resampling techniques for the estimation of standard errors. A Section is also dedicated to the software that has been coded in C++ and that includes all the elements necessary for model estimation and variance analysis. Chapter 5 presents model results from real data extracted from the 2009 National Household Travel Survey and concerning Maryland, DC and Virginia. In Chapter 6, the problem of bias deriving from the non-random distribution of the sample is solved by using pseudo-maximum log-likelihood estimators. Finally, Chapter 7 presents concluding remarks and avenues for future research.

Chapter 2

Literature Review

A large number of publications concerning vehicle ownership can be found in the transportation literature. However, we focus in this Section only on papers that propose advanced econometric methods to predict not only the number of cars owned by a household and their types but also the number of miles travelled. Those models are known as discrete-continuous models as they aims at explaining decisions concerning discrete variables (number of cars owned and/or types-vintages) and continuos variables (total vehicle miles travelled or miles travelled by car type).

Discrete-continuous models have been investigated in marketing studies since the 80's. Marketing researchers developed discrete-continuous models to determine household purchase decisions for frequently purchased packaged goods. Previous studies have predicted one or more of the purchasing decisions by proposing relationships between the observed choices of households and variables such as product price, price cuts, feature advertisements, special displays and observed and unobserved household characteristics [Chi93]. Previous research has focused on three different household purchase decisions (1) the timing of a purchase, (2) the brand choice decision, and (3) the purchase quantity decision.

A comprehensive framework for vehicle ownership and use was developed by Train in 1986 [Tra86]. This model system contains several sub-models:

- a vehicle quantity model,
- a class/vintage model for one-vehicle households,
- a class/vintage model for two-vehicle households,
- an annual vehicle miles traveled (VMT) model for one-vehicle households,
- an annual VMT model for each vehicle for two-vehicle households and
- models for the proportion of VMT in each of two categories (work and shopping) for one- and two- vehicle households, respectively.

Train's model is characterized by the following features: (1) it is a behavioral model that is estimated using choices from a household survey; (2) each household's choices depend on both vehicle class/vintage characteristics (such as vehicle purchase price) and household characteristics (such as household annual income); and (3) the model can be incorporated into a simulation framework to forecast the vehicles' demand and their use. However, the framework proposed is not able to capture interdependecies among the discrete and continuous decisions, which might bias model coefficients and affect model predictions.

In recent years Bhat [Bha05] has proposed a Multiple Discrete Continuous Extreme Value Model (MCDEV) that handles the choice of multiple alternatives simultaneously. His first application concerned a choice situation common in activity-travel analysis where individuls choose the type of discretionary activity (discrete) to participate in and the duration of time investment of the participation (continuous). The model was later applied to analyze the choice of vehicle type/vintage and usage on data from the San Francisco Bay Area Travel Survey [BSE09]. Although this econometric model formulation presents a nice closed form expression for the discrete-continuous probability, it is limited by the assumption that a fixed and limited budget exists for the continuous part. If this assumptions is not restrictive for time assigned to activities (the total time budget available in a day is 24 hours), it might be unrealistic to explain the total household vehicles miles travelled which are expected to vary depending on external circumstancies (introduction of new policies, increasing fuel prices, downturn in the economy).

Later, Spissu [SPPB09] formulated a joint model of vehicle type choice and utilization and estimated the model on a data set extracted from the 2000 San Francisco Bay Area Travel Survey by using a copula-based approach. This method relaxes restrictive distribution assumptions on the dependency structures between the errors in the discrete and continuous choice components. The copula-based methodology was found to provide statistically superior goodness-of-fit when compared to previous estimation approaches for joint discrete-continuous model systems. When applied the models indicates that increasing fuel prices induce individuals to shift vehicle type choices rather than changes in vehicle usage patterns.

Fang [Fan80] presents a very interesting discrete continuous model formulation that aims at predicting the number of vehicles owned for a certain category and the number of miles traveled with each of them. The model accounts for two different vehicle categories: cars and trucks. The joint model uses an ordered probit basis for the number of vehicles owned and a tobit basis (truncated regression) for the number of miles traveled. Error terms are allowed to be correlated, which is a desirable property of this model. The model coefficients are estimated using a Gibbs sampler, whose complexity grows with the number of vehicle categories. The ordered probit imposes the discrete variable to be ordered, which can be a problem for other kind of discrete-continuous applications. Furthermore, the ordered response mechanism has been found to be inferior to the unordered response model by Bhat and Pulugurta [BP98]

Chapter 3

Models

3.1 Discrete Models

3.1.1 Utilities

Suppose that we want a model that predicts the outcome of a categorical dependent variable Y_{disc} using some set of predictors. In an econometrics setting, the categorical variable is referred to as a choice, and its value is presumed to be the one whose utility for the decision maker is the biggest. From this perspective, it is possible to build a model where all k possible modes of Y_{disc} have a utility that consist of one observable part (systematic utility) and one non-observable part (error term).

$$U_1 = X_1^T \beta_1 + \epsilon_1$$

$$U_2 = X_2^T \beta_2 + \epsilon_2$$

$$\dots$$

$$U_k = X_k^T \beta_k + \epsilon_k$$

Utilities do not have a meaning other than indicating the level of appreciation for one particular choice. For example, suppose a household has to choose how many cars to own and faces the following utilities:

number of cars	utility
0	4
1	7
2	5
3+	-2

the household will choose to own one car. The negative utility for the 3 cars alternative does not mean that the household does not consider owning three cars at all, but simply that the benefit of owning three cars would need to be increased by 9 in order to be equally good for the household. Therefore, we are interested mainly in the differences between utilities, not in their absolute level. [Tra09, p. 19]

3.1.2 Error Terms

The assumptions that the analysts makes on the error terms will determine what model will be used. There are mainly two model families: logit and probit.

In the *logit*, error terms are assumed to be *independently* and *identically* gumbel distributed. These assumptions have numerous pros and cons:

- The model likelihood admits a closed analytical form and can be computed easily
- Because error terms are independent, it is not possible to analyze how deciders trade between alternatives.
- Logit assumes the *independence of irrelevant alternatives*; the addition of one

alternative does not modify the relative share of the previous ones. This is a very restrictive property that will probably not hold if alternatives are similar to each other.

[Tra09, p.34]

In *probit* models, error terms follow a normal distribution, but are not necessarily independent. Typically the error terms will follow a multivariate normal distribution with full, unrestricted, covariance matrix. The main advantages and disadvantages can be summarized as follows.

- The model likelihood is difficult to calculate.
- Analysis of trade offs between alternatives is possible

[Tra09, p. 97]

3.1.3 Likelihood - Probability

In order to compute the likelihood function, let's assume that:

$$X = (X_1, ..., X_k)$$
$$Y = Y_{disc}$$
$$\beta = (\beta_1, ..., \beta_k)$$
$$\epsilon = (\epsilon_1, ..., \epsilon_k)$$
$$\Sigma := \text{Covariance of the error term}$$

No matter what model we consider[Tra09, p. 4], we can express the likelihood of *one* observation by:

$$P(Y = y|X, \beta, \Sigma) = \int_{\mathbb{R}^k} \mathbb{I}(X_y^T \beta_y + \epsilon_y > X_j^T \beta_j + \epsilon_j \quad \forall j \neq y) \phi(\epsilon) d\epsilon$$

The indicator function ensures that the observed choice is indeed the one with the biggest utility. Then, we integrate over the distribution of the random error terms to obtain the probability (average) of the observed choice. The subscript $_y$ indicates the predictors and coefficients of the chosen alternative and the subscript $_j$ indicate the other alternatives. The integration region is the support of the random error term. In our case, we assume that ϵ follows a multivariate normal distribution, hence the support is $\mathbb{R}^k$. The problem is that we do not know how to calculate this integral and, therefore, rely on simulation [Tra09, p. 117]:

$$\hat{P}(Y = y|X, \beta, \Sigma) = \frac{1}{B} \sum_{i=1}^{B} \mathbb{I}(X_y^T \beta_y + \epsilon_y^{(i)} > X_j^T \beta_j + \epsilon_j^{(i)} \quad \forall j \neq y)$$

Where $\epsilon^{(i)}$ is a draw from a multivariate normal with mean 0 and variance Σ. This approximation is just a direct application of the law of big numbers. The estimated probability is the mean (or proportion) of the random variable that indicates that one utility of the chosen alternative is bigger than the others. We may stress the fact that the simulation must be performed for each observation.

3.2 Regression

3.2.1 Fitting a Regression

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:

$$Y_{reg} = X_{reg}^T \beta_{reg} + \epsilon_{reg} \qquad \epsilon_{reg} \sim N(0, \sigma^2)$$

Usually, regression is solved by using the OLS (or WLS) estimator [Wei05], but the same problem can be expressed in the form of a likelihood function to be maximized [MSN08, p. 117]. Indeed, given β_{reg}, X_{reg} and σ^2, the likelihood of observing y_{reg} is given by the normal density function:

$$L(y_{reg}|\beta_{reg}, X_{reg}, \sigma^2) = \phi(y_{reg}|X_{reg}^T \beta_{reg}, \sigma^2)$$

The normal density is centered at $y_{hat} = X_{reg}^T \beta_{reg}$ and has variance σ^2.

3.2.2 Simulation

Although it is not necessary for this problem, in theory we could solve regressions by simulation too, just like for the probit. Given σ^2, we can generate error terms, and from this sample, estimate the likelihood of the observed y_{reg}; by applying this method it is then possible to estimate the density of the regression.

If, for example, we did not have σ^2 available, but we had a sample of error terms, it would be natural to use this approach. We will be using this idea later:

$$\epsilon_1, \epsilon_2, ..., \epsilon_B \sim N(0, \sigma^2)$$

$$\hat{\sigma}^2 = \frac{1}{B-1}\sum_{i=1}^{B}(\epsilon_i - \bar{\epsilon})^2$$

$$\hat{f}(y_{reg}) = \phi(y_{reg}|X_{reg}^T\beta_{reg}, \hat{\sigma}^2)$$

3.3 Discrete-Continuous

3.3.1 Motivations

Now suppose that we would like to modelize Y and Y_{reg} jointly to capture the correlation between them. A very natural approach would be to allow the error term of the regression to be correlated with the error terms of the utilities in the probit.

The specifications of the observable part of the utilities and of the regression's $\hat{y}$ shall remain the same. but the error terms follow an "incremented" normal distribution:

$$(\epsilon_1, ..., \epsilon_k, \epsilon_{reg}) \sim MN(0, \Sigma_{k+1})$$

Given β, β_{reg} and Σ_{k+1} it is easy to simulate data that follow the discrete-continuous (DC) model since we can compute all utilities, select the biggest, and use the remaining residual to compute the regression variable.

3.3.2 Estimating The Likelihood

Estimating the likelihood given the observations is more difficult. We have that:

$$f(Y, Y_{reg}) = f(Y)f(Y_{reg}|Y)$$

This is a general result about conditioning with random variables. [Ric07, p. 88]

We have discussed how we can estimate $f(Y)$ taking a number of draws from the multivariate normal distribution. However it is not clear which functional form can be given to the conditional distribution $f(Y_{reg}|Y)$; however, we have a sample of points from the conditional distribution that can be used in order to estimate it. Consider the following:

$\epsilon^{(i)}$	$i = 1, ..., B$	Draws from the MN distribution
$\epsilon^{(i\|y)}$	$i = 1, ..., B^*$	Subset of draws for which the biggest utility in the probit simulation was the observed Y

In other words, when we simulate the probit probability, we keep the error terms that correspond to the regression whenever (conditional) the biggest utility is the one of the observed choice. We rely on the sample $\{\epsilon^{(i|y)}\}_{i=1}^{B*}$ to estimate $f(Y_{reg}|Y)$.

Consider the following illustrative example, where the chosen alternative among "Car", "Bus", "Bike" is "Car". We simulate the utilities B = 10 times:

	Simulation #									
Utilities	1	2	3	4	5	6	7	8	9	10
Car	9.5	8.2	9.5	7.3	10	1.2	4.8	1.4	6.1	7.4
Bus	2.7	1.7	1.7	4.6	4.2	5.8	5.2	4.6	8.1	8
Bike	9.2	8.6	8.9	2.3	8.3	5.5	3.7	9.5	8.5	3.5

In that case we would use the error terms of the regression corresponding to indexes 1,3,4,5 and 10 to estimate the density of the continuous variable.

To conclue, the problem of estimating the DC model likelihood reduces to collecting the regression error terms when we compute the probit. Those error terms are the product of the simulation and our problem reduces to a density estimation problem.

3.4 Density Estimation

3.4.1 Interpretation of a Density Function

We know that the interpretation of a density function is that [Ric07, p. 48]:

$$f(y_{reg})2\delta \approx \mathbb{P}(y_{reg} - \delta < Y_{reg} < y_{reg} + \delta)$$

That is, the density of a random variable Y_{reg} evaluated at y_{reg} times the length of a small interval is approximately equal to the probability that Y_{reg} lies in this interval centered in y_{reg}. We can estimate the left hand side of this expression with random draws, then we estimate $f(y_{reg})$ with:

$$\hat{p} = \hat{\mathbb{P}}(y_{reg} - \delta < Y_{reg} < y_{reg} + \delta) = \frac{1}{B^*}\sum_{i=1}^{B^*} \mathbb{I}\left(y_{reg} - \delta < X_{reg}^T\beta_{reg} + \epsilon_{reg}^i < y_{reg} + \delta\right)$$

$$f(y_{reg}) \approx \frac{\hat{p}}{2\delta}$$

To name only a few problems that can arise, it is possible that $\hat{p} = 0$ and we need to carefully select δ. However, this approximation is computable.

3.4.2 Kernel Density Estimation

Kernel density estimation uses a kernel, that is a density function whose purpose is to "weight" all the points in the sample is order to estimate the density [Par62]:

$$\hat{f}(x) = \frac{1}{B^*}\sum_{i=1}^{B^*} K_h(x_i - x)$$

Where $K(\cdot)$ is a symmetric density function and

$$K_h(x) = \frac{1}{h}K(x/h)$$

Note that $K_h(\cdot)$ is also a symmetric density function that is only a scale transform of $K(\cdot)$. We could use for instance a gaussian kernel (normal) in which case we would not have the problem of estimating the density by 0. However, this method is usually computationally expensive. A sample of 1,000 observations with 1,000 simulations each would require to compute the normal density one million time (!) to estimate the likelihood only once. The problem of finding h remains. The method described in the previous section happens to be the kernel density estimation using a uniform kernel (!) where h was referred to as δ

- Large δ will give a biased estimate of the density
- Small δ will give a volatile estimate of the density

Solvers tend to have a lot of troubles maximizing functions that are not smooth, so choosing small δ will be impossible for a lot of model specifications.

3.4.3 Possible Simple Approximation

There is little hope that we can derive the exact conditional distribution of Y_{reg}, but we may be able to find a known distribution that is a good approximation for it. If a good distribution that can estimate the conditional distribution of Y_{reg} given Y can be found, then it is possible to:

- Find the MLE estimator of the parameter (θ) of this distribution
- Apply it to the conditional residuals from the probit simulation
- Estimate $f(y_{reg}|y)$ with $f(y_{reg}|\hat{\theta}_{mle})$

We tested this idea with the 2001 NHTS Maryland data using the following model [UDoT01]:

Discrete part:

$$U_{0 \text{ car}} = 0$$

$$U_{1 \text{ car}} = \beta_1$$

$$U_{2 \text{ car}} = \beta_2$$

$$U_{3 \text{ car}} = \beta_3$$

$$U_{4+ \text{ car}} = \beta_{4+}$$

Continuous part:

$$\text{MILES\_DRIVEN} = \beta_{\text{const}} + \beta_{\text{inc}} \cdot \text{INCOME}$$

To find reasonable coefficients, we fitted the discrete part with a logit model and the regression part with an ordinary regression, separately. As a covariance matrix we used a diagonal matrix with 1 for the variance of each utilities, and the regression's variance for the variance of the continuous variable. The conditional residuals we obtain with those coefficients are arguably representative of what happens during the optimization process. We started with 25,000 simulations and plotted the density of the conditional residuals for the first 16 observations. The densities are plotted in figure 1. We note that:

- Residuals seem to be normally distributed;
- The mean of the distribution of Y_{reg} is approximately $X_{reg}\beta_{reg} + \mu_{\epsilon(\cdot|y)}$
- The variance of the distribution of Y_{reg} is approximately $\sigma^2_{\epsilon(\cdot|y)}$

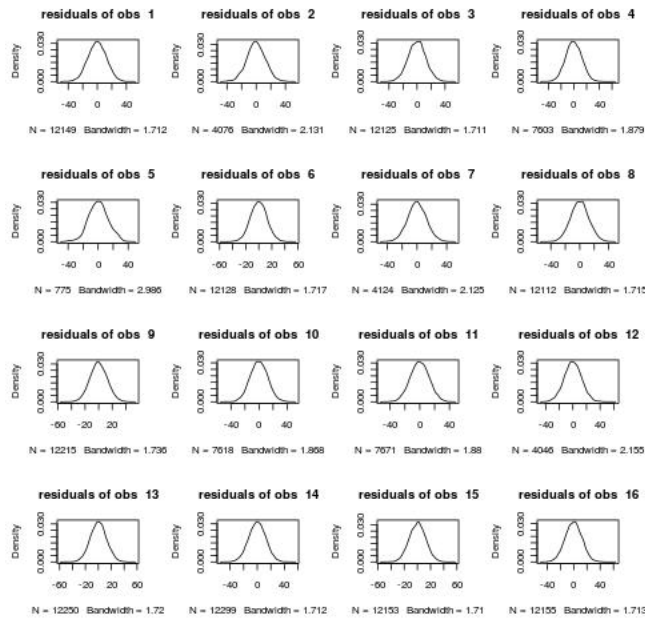

Figure 3.1: Distribution of conditional residuals

Where $\mu_{\epsilon(\cdot|y)}$ and $\sigma^2_{\epsilon(\cdot|y)}$ are the mean and variance of the conditional residuals. Therefore, the computation of the conditional density of y_{reg} is much more stable that the first method and much faster than the second.

In order to be able to estimate the conditional density, at least two conditional residuals are needed. Obviously, the precision of the density estimation depends on how well the discrete variables are predicted earlier in the simulation. However, given that the normal assumption seems to be a good approximation, there is no reason to believe that estimating it with few residuals will cause problems or will deteriorate estimates.

3.5 Simulated Log Likelihood

The final Simulated Log Likelihood of the DC model is given by the following formula:

$$SLL(\beta, \beta_{reg}, \Sigma | Y, Y_{reg}, X, X_{reg}) =$$

$$\sum_{i=1}^{n} \log \left(\sum_{b=1}^{B} \frac{B_i^*}{B} \times \phi(y_{i,reg} | X_{reg}^T \beta_{reg} + \mu_{\epsilon(\cdot|y_i)}, \sigma^2_{\epsilon(\cdot|y_i)}) \right)$$

Where:

$$B_i^* := \text{number of success in } i^{th} \text{ probit simulation}$$

Chapter 4

Numerical Issues

4.1 Covariance Parametrization

In this Chapter we discuss the optimization procedure adopted to solve the maximum likelihood problem defined in the previous Section. Our formulation does not impose any constraints, but requires that Σ, which is the covariance matrix, to be positive-definite. To achieve this, we parametrize the model with respect to the cholesky factor of the covariance matrix. If Σ is positive-definite, then:

$$\Sigma = LL^T$$

where L is a lower triangular matrix, or the cholesky factor of Σ. In drawing from the multivariate normal distribution, we simply multiply the multivariate normal random variable Z, with identity covariance, by L to obtain X, where X has covariance Σ. [Law07, p. 468]:

$$X = LZ \qquad\qquad Z_i \sim N(0,1) \quad \forall i$$

The parametrization of the SLL function proposed, uses L rather than Σ; however, given the relation above Σ can be obtained after L has been estimated.

4.2 Drawing Error Terms

4.2.1 Crude Monte-Carlo

We have discussed that, in order to estimate the SLL function, we draw error terms from the specified distribution, simulate the utilities of all alternatives and then estimate the probability of the observed choice. The standard deviation of the estimate we obtain with such a raw procedure is in the order of $1/\sqrt{B}$ (where B is the number of simulations). [Owe03]

4.2.2 Quasi Monte-Carlo

There is no need, however, that each of the draws be independently distributed over the whole support of the error terms. We are interested in getting a collection of error terms that replicate as closely as possible the properties of the exact distribution they come from. For example if we wanted to sample 100 points on the interval $[0, 1]$, we would get a better coverage if we sampled 10 points on the interval $[0, 0.1]$, 10 points on the interval $[0.1, 0.2]$ and so on. This is the idea of *partitioning*.

The principle of partitioning can be extended to more than one dimension. For this problem we are working with an error term that is distributed on $\mathbb{R}^k$ so it is rather impractical to divide the support into subspaces and drawing terms in each of them since there would be too many subspaces. It is possible to select a subset of those many subspaces to draw from using a Latin hypercube scheme that we do not discuss here. The standard deviation of the estimation using such a scheme will be at most in the order of $1/\sqrt{B-1}$ and potentially better if we take advantage of the structure of the distribution and model at hand. [Owe03]

Another famous method is the *Halton Sequence*. The error terms are created by iteratively adding new error terms in order to density the coverage we have of the support. It is not easy however to determine the variance of the estimator we obtain using such a procedure. [Tra09, p. 221]

For this model, we use crude Monte-Carlo draws, but we need to bear in mind that the resulting imprecision may affect the quality of the results we obtain.

4.3 Dealing With Probabilities Estimated at 0

4.3.1 Probit Alone

When estimating the likelihood of the probit alone, it is possible that no successes are observed in the indicator function. In this case, the literature suggests to fix the number of successes to some arbitrary small values (like 0.1 success). [Tra09, p. 118]

If this number is too low, the optimization algorithm may converge to a point where there is no extreme values and the algorithm will never reach the maximum of the LL function.

If this number is too high, then we underestimate how bad are the particular coefficients to predict the observation that yields to zero successes in the simulation. For example if the actual probability of one observation is $10e-10$, and we set it to $10e-3$, the resulting log-likelihood will not reflect that one observation, given the current parameters, was very unlikely to observe.

4.3.2 Joint Model

When dealing with joint discrete continuous model, the probit probability is multiplied by some density and the problem becomes more complex. We may obtain 0 success in the discrete part, in which case it is not possible to estimate the conditional continuous density. Having no success in the computation of the probit likelihood gives no indication about the value of the conditional density of the regression. Unlike in the probit, it is very hazardous to set an arbitrary value here because the conditional density may be low *or* high.

A possible solution to this problem would be to ignore the correlation. When computing the simulated likelihood, we do know the covariance of the error terms; from the covariance matrix, we can use the diagonal term that correspond to the variance of the regression and estimate the conditional density by the unconditional density:

$$f(y_{reg}|Y) \approx \phi(y_{reg}|X_{reg}^T\beta_{reg}, \Sigma_{k+1,k+1})$$

From our empirical experience it has been observed that once an optimal point has been found in the SLL function, at least a few successes for each observation are registered. Therefore, such approximation is never applied at the maximum.

4.4 Continuity Problems

4.4.1 Gradient of the Simulated Likelihood

Our problem does not have a closed mathematical form for the likelihood function. As a consequence, it is not possible to compute the gradient. We use a five-point stencil gradient approximation given by the following formula [BF05, p. 172]:

$$f'(x) = \frac{-f(x+2\delta) + 8f(x+\delta) - 8f(x-\delta) + f(x-2\delta)}{12\delta} + \frac{\delta^4}{30} f^{(5)}(c)$$

where:

$$c \in [x - 2\delta, x + 2\delta]$$

and $f^{(5)}(\cdot)$ is the fifth derivative of $f(\cdot)$.

4.4.2 Discontinuity of the SLL Function

Let's consider the smallest difference between simulated utilities and call this quantity h. Then if the β parameter used to compute the utilities changes by only a small amount compared to h, the value of the indicator function will not change at all. Hence, the gradient with respect to the parameter is zero everywhere it is defined in the SLL function. When the indicator function changes, the SLL function, calculated at corresponding value of β, is not continuous. Thus, the SLL function is piecewise constant. [Tra09, p. 119] Consider the next figure, where the function plotted is piecewise constant.

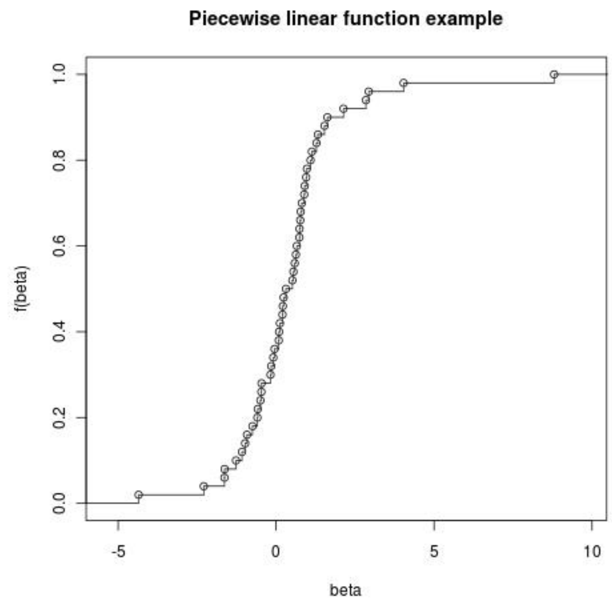

Figure 4.1: ill-conditioned piecewise function

The SLL is piecewise constant, therefore it is only possible to calculate the "arc-gradient". [Tra09, p. 120]. When computing this arc gradient using finite differences, it is important to take a difference that is not too small. This may cause high values of the approximate derivative if we are close to a change in the values of the indicator, or 0 if the indicator is not close to change.

We observe that the numerical approximation error goes to zero much faster than δ. However, the SLL function must be computed many time in order to achieve this precision. In addition, computation of the SLL takes more time as the number of predictors and the number of alternatives increases. The time required to compute the error terms depends on the number of simulations and the number of alternatives.

Adding more predictors increases the computational effort in terms of SLL

δ	δ^4
0.1	1e-4
0.05	6.25e-6
0.01	1e-8

Figure 4.2: error term magnitude

		number of alternatives (probit)				
		2	3	4	5	6
	10	60	78	96	120	148
	15	80	96	116	140	168
#predictors	20	100	116	136	160	188
	25	120	136	156	190	208
	30	149	156	186	200	228

Figure 4.3: function evaluations to evaluate gradient

evaluations, but it does not modify much the time required to compute the SLL only once.

We must keep in mind that δ^4 is not the error, but an upper bound on the magnitude of the error times an unknown constant. This constant may be relatively big so we shall remain conservative and take $\delta = 0.1$. The "best" value for δ may depend on the problem at hand.

4.4.3 Hessian Matrix of the Simulated Likelihood

It is well known from the asymptotic theory of maximum likelihood estimators that, for a random variable that has a (multivariate) parameter θ [Ric07, p. 279]:

$$\hat{\theta} \to MN(\theta, I)$$

Where I is Fisher's information matrix, the elements of I are given by:

$$I_{ij} = -\mathbb{E}\left(\frac{\partial}{\partial\theta_i\partial\theta_j}SLL(\theta)\right)$$

The elements of the expectation are the mixed derivatives of the SLL function. We cannot compute this expectation so we approximate it with the elements of the Hessian matrix evaluated at the SLL's maximum. This procedure is computationally very expensive. In our problem we have 35 predictors and 5 alternatives for a total of 55 parameters to estimate. For the mixed derivative between the first parameters and the others, we must consider finite differences 55 times. Then for the mixed derivatives between the second parameter and the others, we must consider 61 finite differences, and so on. The total number of finite differences to compute is $55+54+\cdots+2+1$. Each of them require the calculation of the simulated likelihood four times for a grand total of 6,160 SLL function evaluations.

As those SLL evaluations do not depend on each other, it is possible to parallely compute them with, *a priori*, no limit in the number of parallel threads. As of 2012, processors with 8 physical cores were available on the market. Multi-thread computations, however, requires specific programming skills. Re-sampling technique, such as bootstrapping, appear to be more appropriate in this context.

Another problem is that we are interested in finding standard errors of the covariance matrix, not the cholesky matrix elements. To achieve that it is necessary, at time of computing standard errors, to re-parametrize the model with respect to the covariance matrix.

Suppose we have been able to maximize the SLL function and we have $\Sigma_{mle} = L_{mle}L_{mle}^T$. In order to compute the Hessian matrix of the whole parameter, it is necessary to add and subtract small quantities to the elements of Σ_{mle}. The matrix we obtain will not necessarily be positive-definite if we take δ too big, hence finite differences cannot be computed if this happens.

Simulations involved in the computations of the SLL will cause the maximum likelihood estimates to be biased [Tra09, p. 239]. So we will consider the Hessian matrix not at the maximum of the SLL, but at some other point that is close to the real maximum. This will also cause errors in the computations of the standard errors.

One may wonder if the Hessian matrix we obtain with such a process is reliable enough to proceed to standard errors estimation using asymptotic theory. Furthermore, After obtaining the Hessian matrix of the SLL at its maximum, we still need to invert it. It is possible that the estimated Hessian matrix is not even invertible or that the variances we estimate are negative. [AGM03, p. 150]

4.5 Value of Hessian Based Standard Errors

4.5.1 Simulation Using Logit Model

To convince ourself of the value of such an approximated Hessian matrix, we consider fitting a logit model using the same utility specifications than in the DC model, but using a logit model. We will first compute the real maximum of the log-likelihood (LL) function. The logit's LL function is easy to compute and maximize. Consequently, we can estimate very well the Hessian matrix at the maximum, H. Then we will perturb H and see how well this perturbed matrix will do at estimating the standard errors.

For some values of α, we modify H by adding a perturbation term that is randomly distributed on the interval $[-\alpha, \alpha]$, compute the standard errors using this inexact Hessian matrix, and check whether this procedures gives an estimation on the standard deviation that is good. We want the perturbed Hessian matrix to give estimated standard errors that are centered at the real values and that have low variance.

We have 28 coefficients to estimate and we want reliable estimation of their standard errors for each of them. We perform 1,000 perturbations of the Hessian matrix for $\alpha = 0.1, 1, 10$. For each value of α we have 1,000 estimates of the standard errors. We consider the expectation and variance of those estimates and then we compute the mean and maximum over the 28 coefficients. We discard variances estimated at negative values (NAs). We consider the absolute value of deviations between "real"

	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$
mean abs. bias	0.49%	10.6 %	29.4%
max abs. bias	1.79%	21.6 %	72.8 %
mean rel. var.	0.06 %	32.14 %	37.81 %
max rel. var.	0.62 %	218.09 %	199.70 %
mean # NA	0	129	311
max # NA	0	543	571

Figure 4.4: standard error estimation for 28 coefficients

standard errors (with original H) and estimated standard errors (with perturbed H). The following table summarizes the results

The relative bias in the computations of the standard errors is quite high. In addition, the variance of the standard errors, when computed from different hessian matrices, is high too.

One problem that arises quickly is that we get negative variance estimates for the coefficients. On the average, for perturbations that are as low as in the $[-1, 1]$ range, it was not possible to compute the standard errors 12.9 % of the time. For one coefficient, failure to compute standard errors happened 54.3 % of the time.

We tried to compute Hessian-based standard errors for simpler model specifications and always had a lot of negative variance estimates, indicating that we have not a reliable hessian matrix estimate at hand.

4.6 Bootstrap Variance Estimation

4.6.1 Parametric Bootstrap

Our objective is to calculate the variance of the maximum likelihood estimator conditional on the values of our predictors, the values of the dependent variables and the value of the parameters.

$$V(\theta_{mle}|X, Y, \theta)$$

To simplify our notation here, Y and X are tuples containing respectively the dependent variables and the predictors of the probit and the regression. θ is a triple that contains $\beta_{probit}, \beta_{reg}$ and Σ.

The source of variability of the MLE estimator is that, we are assuming that the Ys we observe are from an infinite population. However. the sample only gives a finite number of them. If the sample did output different values of the Ys, we would obtain different MLE estimates.

We do not have at hand the real θ parameter of the model, but we have $\hat{\theta}$, a good approximation of it. We have X, on which the dependant variable depends. The parametric bootstrap procedure generates a new sequence of the dependant variable Y, and estimate the parameter θ for this new Y. If we repeat this procedure B times, we obtain a sequence of $\hat{\theta}_i$. This sequence of $\hat{\theta}_i$ follow approximately the same distribution than the real $\hat{\theta}$ in the actual sample. [Efr87, p. 50]:

$$Y_i^* \sim DC(\hat{\theta}_{mle})$$

$$\theta_i^* = \arg\max_{\theta} SLL(\theta|X, Y)$$

The subscripts here denote replications of the vectors or matrices, not indexes. Suppose from simplicity that θ is a one-dimensional parameter. We can order the re-sampled θs to build a confidence interval for it. If we denote by $\theta^*_{(i)}$ the i^{th} smallest θ we obtained in the bootstrap process, then a good confidence interval would be to discard the lowest and highest $\frac{\alpha}{2}$% observations and consider the range of the remaining re-sampled θs [CVC88, p. 83]:

$$IC = [\theta_{(B\frac{\alpha}{2})}, \theta_{(B(1-\frac{\alpha}{2}))}]$$

B is the number of bootstrap samples that we used. This number will depend on the computing time we are willing to spend. Bootstrap sample estimates may be considerably faster to compute because we have a good starting value for the optimization process : θ_{mle}. This can reduce dramatically the number of iterations that the solver will need to perform in order to maximize the SLL function.

4.6.2 Non-Parametric Bootstrap

Another alternative to select resamples is to select a simple random sample from the original sample, with replacement. For a sample of size n, select n times any observation and call that selection the resample. Of course some observation will be selected more than once. The remaining of the procedure remains the same. [ET93, p. 45]

The difference between the parametric and non-parametric versions of the bootstrap lies in the fact that the actual distribution from which we draw the resamples are different. In the first case, we draw from the *true* model, in the second case we draw actual observations.

Both methods work in theory, but some investigation will be required to assess if one is better than the other for this problem. We are using the parametric bootstrap without further justification at this moment.

4.7 Model Identification

It is important to make sure that the model is identifiable (or estimable).

For the regression, it is important that the predictors are not co-linear, that is, no predictor can be expressed as a linear combination of the others. This property can be "observed" in the OLS formula used to compute the coefficients of the regression: $\beta_{ols} = (X^T X)^{-1} X^T y$. The matrix inversion will not be possible if all the columns of X (the predictors) are not independent from each others. [Wei05, p. 214]

For the probit, we are considering the differences between the utilities, not their absolute value. Suppose, for demonstration purposes, that we want to predict the mode of transportation used with utilities being a constant plus an error term:

$$U_{car} = \beta_{car} + \epsilon_{car}$$

$$U_{bike} = \beta_{bike} + \epsilon_{bike}$$

$$U_{bus} = \beta_{bus} + \epsilon_{bus}$$

The identification problem lies in the fact that adding a constant to all the βs will not change the differences between utilities, hence $(\beta_{car}, \beta_{bike}, \beta_{bus})$ and $(\beta_{car} + k, \beta_{bike} + k, \beta_{bus} + k)$ will give the same results in terms of utility selection. The solution is simply to set one of those constants to zero. The situation is similar if we use a common predictor across alternatives (for example income, that is not specific to one alternative). [Tra09, p. 20]

If we add a variable whose value is different for each alternative (for example price), we do not have identification problem since the (different) predictors will multiply the coefficients. If we add a constant to the coefficients, the constant will have a different effect on each utility, so that it will not be possible to keep the same differences in utilities.

Now, suppose that the variance of the error terms are unrestricted and that the errors are independent. Then it will not be possible to identify the scale of the model. Indeed, $(\beta_{car}, \beta_{bike}, \beta_{bus})$ and $(k\beta_{car}, k\beta_{bike}, k\beta_{bus})$ will not give the same utilities, but the ratio between the differences will remain the same. If the variance of the error terms is modified accordingly, the utilities will not be the same, but the previous biggest utility will remain the biggest. Thus we force one error term to have a fixed variance. Usually we will just say that $\Sigma_{1,1} = 1$. [Tra09, p. 22]

4.8 Software and Libraries

4.8.1 Programming

Our estimation procedure intensively relies on simulation, therefore computing efficiency is a major concern. As a consequence, all computations were performed using C++ and the GNU C++ compiler, version 4.5.2. [Str97]

4.8.2 Matrices

Armadillo version 2.4.3, a wrapper library for BLAS, is used for matrix computation. It provides a good balance between speed and facility of use. [San10]

4.8.3 Solver

The solver adopted is NLopt, which has been developed for non-linear optimization. The library provides local and global optimization algorithms. We are using version 2.2.4 of the library. [Joh07]

4.9 Algorithms

For this specific problem we use the LBFGS algorithm, that stands for low-storage BFGS. BFGS is a gradient based, local, solving algorithm and L-BFGS is a version that uses less memory. [Noc80, DN89]

4.9.1 Rentix : A Discrete Choice Modeling Framework

When we specify statistical models, we usually refer the estimation process as maximizing the likelihood function. This requires the following actions to be taken:

- Read and organize the raw data set
- Organize utility and regression based datasets
- Initialize and run solvers
- Display results
- Track runtime errors
- Compute derivatives
- etc.

We propose Rentix, that is a programming interface that solves statistical models and that only requires the user to specify the likelihood function. The implementation is done in C++ such that time-intensive simulations are done reasonably fast.

Chapter 5

The Real Model

5.1 Motivations

The discrete continuous formulation proposed in Section 3 is applied here in the context of vehicle ownership and usage; models are estimated at household level. The problem is characterized by two dependent variables: the total number of vehicles owned by a household and the total number of miles traveled in a year. Those variables are somehow related and therefore it is desirable to model them jointly to account for possible correlation across error terms.

By adopting this approach, it is not necessary to make any assumption about the hierarchy of the two decisions, as both miles traveled and number of cars owned could be caused by a third (latent) phenomenon.

5.2 Data

For the empirical analysis, we use data extracted from the 2009 National Household Travel Survey (NHTS) and relative to Maryland, Virginia and the District of Columbia. Only complete observations are included in our final sample, which contains 1,497 observations. [UDoT09]

5.2.1 Descriptive Statistics

Eight variables have been retained in our final specification:

- Income, categorized into 18 ordered categories, treated as continuous. It should be noted that the last (18^{th}) income category includes households that earn over 100,000$ per year, so a quite large number of households fall into this category.

- Education (of the head of the household), is categorized into 5 ordered categories and treated as a continuous variable. Categories correspond to the following education levels : some high school, high school, some college, college and graduate degree.

- Number of drivers in the household

- Number of workers in the household

- House ownership, binary (owns or rents)

- Urban density, in thousands of units per square mile. Density also suffers from lack of variation as most households live in low-density areas, but few live in very dense areas.

- Number of vehicles owned. Households that own 4 cars or more are resumed into one "4+ cars" category.

- Total number of miles traveled (in thousands). Most households travel between 0 and 50 thousand miles per year, with some households traveling much more

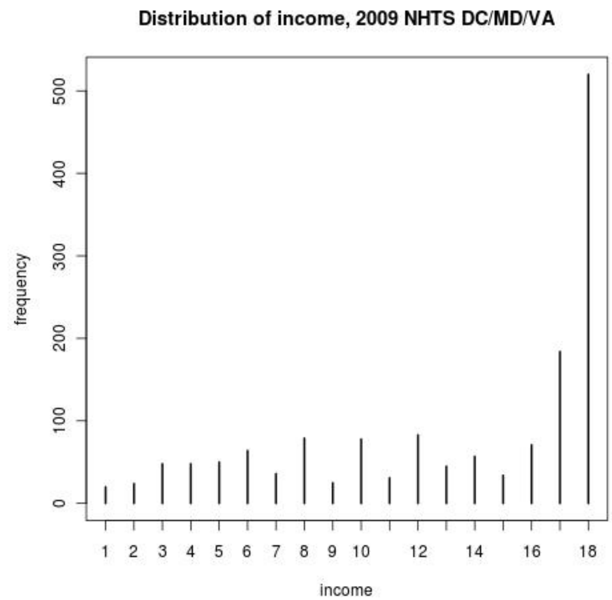

Figure 5.1: Distribution of income

than that.

5.3 Estimated Coefficients

In car ownership model specification the dependent variable is the number of vehicles owned by the household. For identification purposes, the utility of owning 0 car is set to zero. Consequently, each utility, except the one that corresponds to the 0 cars alternative, uses all the selected predictors.

The confidence intervals are fixed at 90 % and calculated from 100 bootstrap samples. The bootstrap process took approximately 24 hours. The lower bound was calculated by taking the mean between the 5^{th} and the 6^{th} smallest coefficient and the upper bound by the mean across the 95^{th} and 96^{th} smallest coefficients.

predictor	alternative	estimate	CI lower bound	CI upper bound
constant	1 car	1.33	1.26	1.40
	2 cars	-0.91	-1.00	-0.85
	3 cars	-1.49	-1.51	-1.42
	4+ cars	-0.74	-0.80	-0.70

income	1 car	0.3	0.22	0.41
	2 cars	1.06	0.98	1.19
	3 cars	1.05	0.94	1.16
	4+ cars	0.65	0.54	0.79
education	1 car	1.75	1.58	1.91
	2 cars	-0.7	-0.83	-0.55
	3 cars	-2.07	-2.20	-1.95
	4+ cars	-2.87	-2.96	-2.80
num. of drivers	1 car	-2.95	-3.04	-2.82
	2 cars	1.2	1.10	1.29
	3 cars	1.58	1.50	1.75
	4+ cars	0.98	0.87	1.04
num. of workers	1 car	-1.85	-1.92	-1.71
	2 cars	-0.15	-0.22	-0.08
	3 cars	0.65	0.59	0.75
	4+ cars	0.68	0.63	0.73
urban density	1 car	-0.23	-0.41	-0.11
	2 cars	-1.68	-1.91	-1.51
	3 cars	-3.03	-3.20	-2.94
	4+ cars	-3.4	-3.51	-3.31
house ownership	1 car	0.63	0.59	0.71
	2 cars	0.11	0.01	0.15
	3 cars	-0.34	-0.38	-0.27
	4+ cars	0.06	-0.00	0.09

log-likelihood at zero : -8106.00
log-likelihood at max. : -7680.13
difference : 425.87
degrees of freedom : 30
$\mathbb{P}(\chi^2_{30} > 425.87) \approx 0$

For the regression the dependent variable is the (scaled) number of miles traveled

The predictors are:

predictor	estimate	CI lower bound	CI upper bound
constant	1.14	1.09	1.17
income	0.69	0.63	0.75
education	-0.44	-0.59	-0.38
num. of drivers	5.70	5.59	5.81
num. of workers	3.52	3.48	3.64
urban density	-0.84	-0.96	-0.70
house ownership	1.36	1.33	1.42

The estimated covariance is :

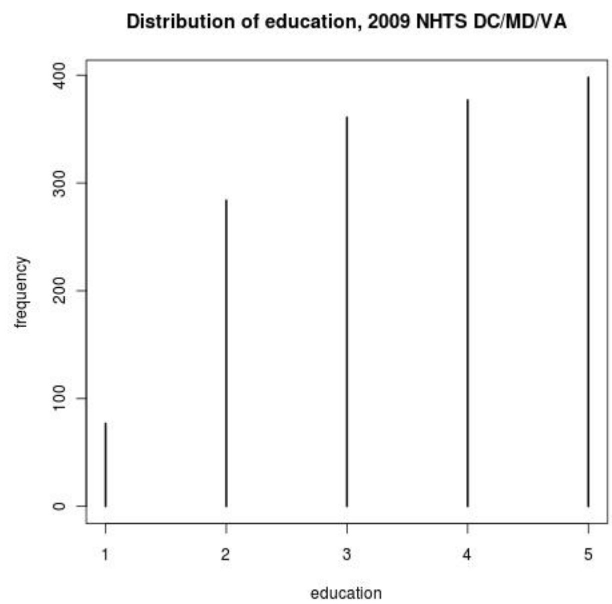

Figure 5.2: Distribution of education

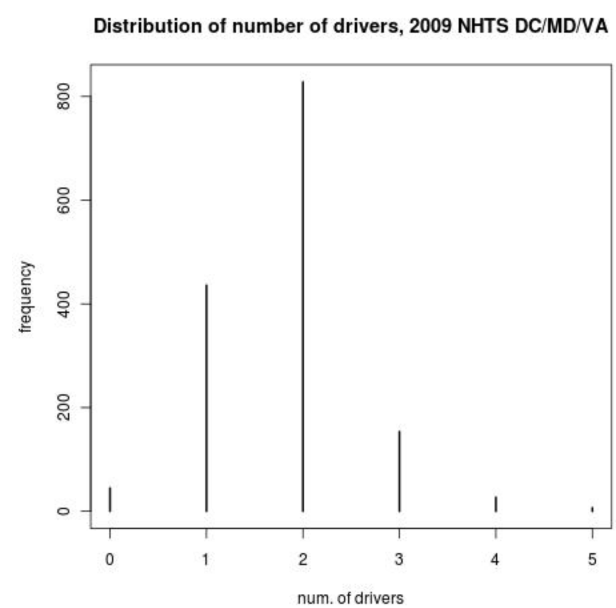

Figure 5.3: Distribution of number of drivers

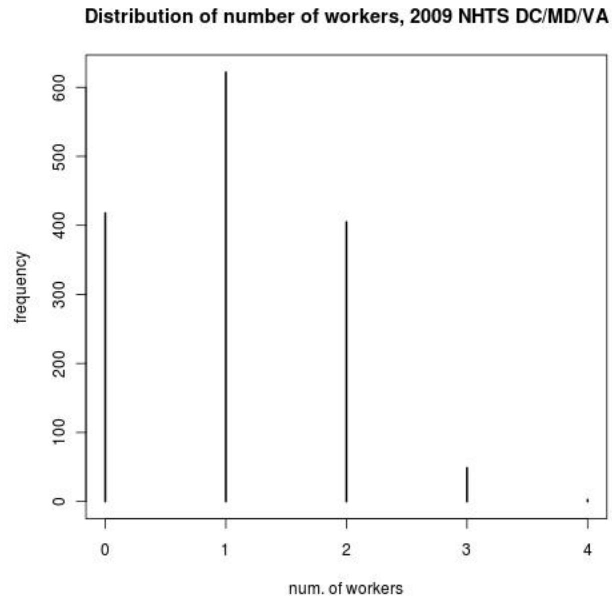

Figure 5.4: Distribution of number of workers

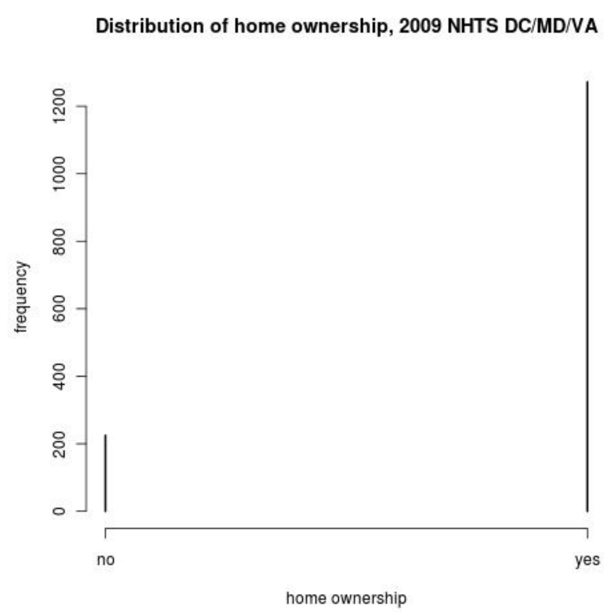

Figure 5.5: Distribution of home ownership

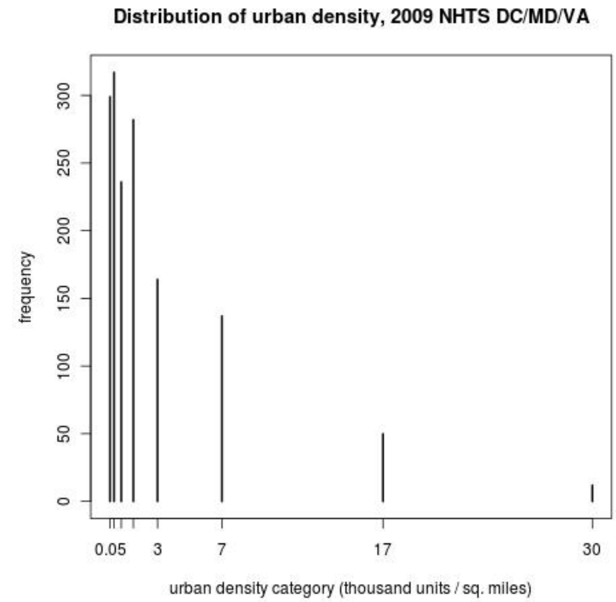

Figure 5.6: Distribution of urban density

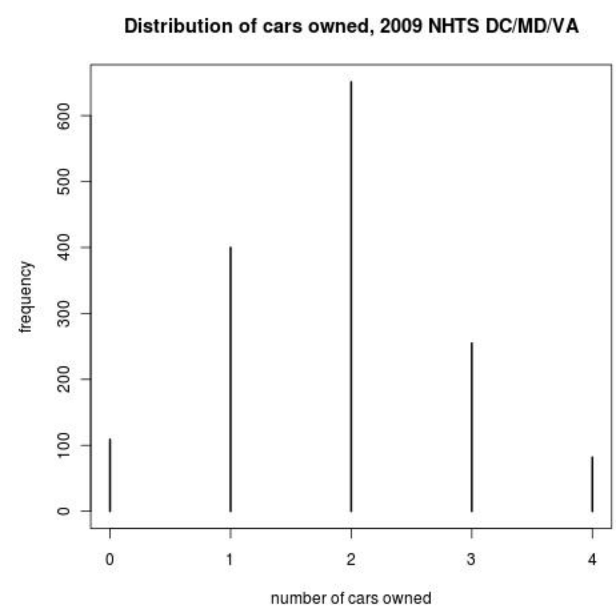

Figure 5.7: Distribution of cars owned

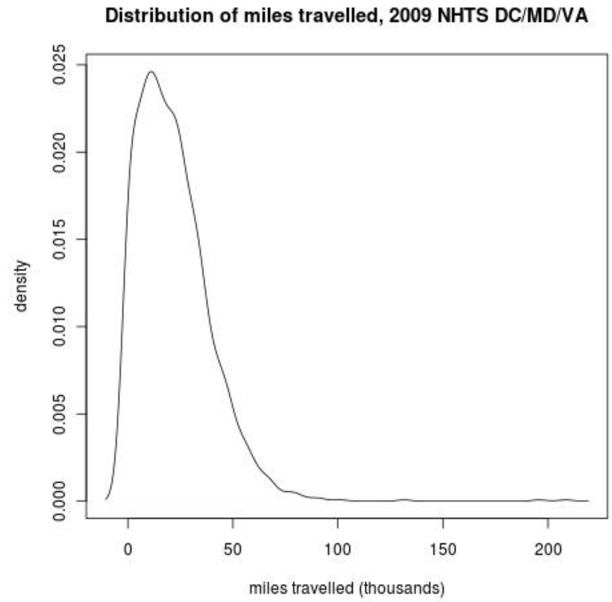

Figure 5.8: Distribution of miles traveled

$$\hat{\Sigma} = \begin{bmatrix} 1.00 & 0.28 & 0.30 & -0.58 & -0.20 & -0.47 \\ 0.28 & 109.67 & 3.57 & 9.82 & 0.80 & -25.29 \\ 0.30 & 3.57 & 109.82 & 7.94 & 2.70 & -1.55 \\ -0.58 & 9.82 & 7.94 & 93.53 & 1.63 & 23.36 \\ -0.20 & 0.80 & 2.70 & 1.63 & 94.12 & 28.05 \\ -0.47 & -25.29 & -1.55 & 23.36 & 28.05 & 209.14 \end{bmatrix}$$

The lower and upper bounds on the covariance elements are given by:

$$\hat{\Sigma}_{low} = \begin{bmatrix} 1 & 0.25 & 0.24 & -0.62 & -0.25 & -0.49 \\ 0.25 & 107.93 & 2.89 & 9.34 & 0.32 & -26.12 \\ 0.24 & 2.89 & 108.07 & 7.23 & 2.03 & -2.59 \\ -0.62 & 9.34 & 7.23 & 91.79 & 0.88 & 21.29 \\ -0.25 & 0.32 & 2.03 & 0.88 & 93.15 & 26.94 \\ -0.49 & -26.12 & -2.59 & 21.29 & 26.94 & 207.25 \end{bmatrix}$$

$$\hat{\Sigma}_{up} = \begin{bmatrix} 1 & 0.34 & 0.35 & -0.53 & -0.16 & -0.45 \\ 0.34 & 111.07 & 4.08 & 10.34 & 1.44 & -24.03 \\ 0.35 & 4.08 & 110.95 & 8.60 & 3.02 & 0.43 \\ -0.53 & 10.34 & 8.60 & 94.40 & 1.88 & 24.19 \\ -0.16 & 1.44 & 3.02 & 1.88 & 94.84 & 28.26 \\ -0.45 & -24.03 & 0.43 & 24.19 & 28.26 & 213.82 \end{bmatrix}$$

5.4 Analysis of coefficients

5.4.1 Car Ownership

Income As expected, income has a positive (when compared to the 0 car alternative) effect on the utility of all ownership alternatives. Income highly affects the utility of owning 2 cars; which is also the alternative with the highest market share.

Education Education increases the utility of owning only one car (0.57) and reduces the utility of owning more than one car. The more cars, the more education reduces the utility (down to -2.67 for four or more cars)

Drivers Households with more drivers have bigger utilities for owning more than one car. More drivers reduces the utility of owning only one car, although it reduces it *less* than the utility of no car at all.

Workers Households with a higher nemuber of workers have higher utility of owning 3 cars and lower utility for owning only one car; the effects on other ownership options are milder.

Density High urban density *significantly* reduces the utility of owning 3 cars or more, and reduces the utility of owning two cars. However it does not have a significant effect on the utility of owning one car. Households in dense areas seem to still rely on at least one car.

Tenure House ownership produces a rather unexpected effect on car ownership. Households that own a house have higher utilities for owning 1 car or 4 cars.

5.4.2 Miles Traveled

All predictors except education and urban density increase the number of miles traveled. We note that the drivers explain nearly twice as much miles traveled than workers.

5.4.3 Covariance of Error Terms

First alternative It should be noted that the covariance matrix does not provide a lot of insights for the analysis of the error terms of "0 car" alternative. We have mainly two different results to be explained. There are few households that do not own at least a car, so it is not clear if it is really possible to make a strong distinction between owning 0 cars and other ownerships. To overcome this problem, one may think to combine households with zero and one car as "low car ownership" households; however, this way to proceed is not common in car ownership modeling. Before doing that, we should investigate the homogeneity of those households. Also, we set the utility of "0 car" alternative to zero, so this may have an impact on the covariance terms between this alternative and the others.

Other alternatives Other alternatives error terms are correlated between each other as expected. Owning one car alternative is negatively correlated with owning more than 4 cars or more. Owning 2 cars is somewhat correlated with owning 3 cars, but negatively correlated with owning 4 cars or more. Roughly, alternatives are more correlated with similar alternatives and less correlated with much different alternatives. Indeed, we expect that when the error terms (what we cannot predict in the utilities) of owning 1 car is big, then the error term of owning 4 cars or more may be small (or negative).

Limitations Covariances between error terms of alternatives may seem relatively small, but we must keep in mind that the error terms are the part of the utilities that we cannot explain. Utilities already include a relatively large number of predictors, so error terms don't need to be strongly correlated for the model to offer good prediction power. The various socio-economic variables that we include in the utilities *may* be enough to explain the differences between alternatives.

On the other hand, if we use such a model to extrapolate car ownership over a full population, correlations of error terms will add a component that is not available when the assumption of independence across error terms is made. For example, if two alternatives have a negative covariance between their error terms, even if the utilities alone fail to predict the generally least predicted one, error terms may be able to compensate for that. This may produce a more accurate portrait of the heterogeneity in preferences. From our analysis, it has apperaed that there are households who chose an alternative that seems completely dominated by others, and we rely on error terms (or un-observable utility component) to reflect this reality.

Miles traveled and cars owned   The most interesting part of the covariance matrix is the relationship between the error parts of the regression and the alternatives utilities.

We note that the error term of the number of miles traveled is negatively correlated with the utility of owning only 1 car, and it is positively correlated with the utility of owning 3 car 4 or more cars. The correlation with the utility of owning 2 cars is not appreciable, but owning 2 cars is the *de facto* most chosen alternative of our choice set. High covariance terms between utilities of 1, 3 and 4+ cars alternatives and number of miles traveled helps to understand the high and low values of number of miles traveled.

5.5 Change in Regression Coefficients

Finally, we compare the results obtained with the joint discrete-continuous model with the results obtained by applying simple regression to the number of miles traveled. It can be seen that when adding covariance elements to the regression model some coefficients change in size. Some variables are observed to have a stronger effect on the predicted number of miles traveled while some others loose their importance:

predictor	disc.-continuous	regression only
constant	1.14	-0.79
income	0.69	0.65
education	-0.44	-1.07
num. of drivers	5.70	7.27
num. of workers	3.51	3.37
urban density	-0.84	-0.76
house ownership	1.36	3.43
σ^2	231.60	209.19

Cells in gray indicate those variables that will contribute more to the number of miles traveled in the discrete continuous model than in the regression alone. We also note that the variance of the error term in the DC model is bigger than in a simple regression. This is an interesting results, indicating that there is a "compensation effect" for the incorporation of car ownership utilities into the model.

5.6 Applications

5.6.1 Coefficients

The coefficients we obtained already give us some information that may be use to elaborate policies or to take decisions. For instance, we already know without further work how many additional miles a new driver in a household will produce. We are also able to determine, a relative way, how changes in predictors affect the utilities of different vehicle holding alternatives.

5.6.2 Policy Implications

Another possibility is to modify one variable in the sample (multiply or add a constant), and apply the model to the modified sample with the new variable.

For example if we are interested in the effect of population density we create a new density that is 125 % of the previous one, apply the model, and estimate vehicle holding under different population settings.

It is then possible to evaluate the benefit of public policies. A decision maker will be able to determine precisely the effect of policies on urban planning, gasoline taxes and other variables. For example, is it worth trying to increase the population density in a given area.

Chapter 6

Weights in the Sample

6.1 Selection Probability

National surveys are assumed to be random sample of the population of interest; or in other words individuals/households have equal probability of being selected. However, some individuals are much more likely to be included in the sample than others. Suppose for example that low income people were targeted by the survey because it was decided that further action must be taken to improve their mobility. Then the coefficients we estimate will tend to fit better low income observations than other observations.

Let's assume that in the sample process each individual in the population has a probability π_i to be selected. When performing general inferences using the selected sample, it is desirable to give more weight to the individuals that were less likely to be selected. If the sampling process is repeated many times, individuals with lower selection probability would end up being in the sample less often. Therefore, they should be give more weight.

6.2 Weights

In this context, we have assumed the *weights* to be the inverse of the selection probability:

$$\omega_i = \frac{1}{\pi_i}$$

Among other explanations, compute a weighted average or sum from the sample will provide an unbiased estimate of population totals and means.

6.3 MLE and Pseudo-MLE Estimator

Suppose we have a random variable with unknown parameter θ, then the pseudo-maximum likelihood estimator of it is given by maximizing the weighted pseudo maximum likelihood function:

$$\hat{\theta}_{pseudo} = \arg\max_{\theta} \sum_{i=1}^{n} \omega_i \log(p(y_i|\theta))$$

The only difference between the log-likelihood function and the pseudo log-likelihood function is that the second is a sum of the weighted log probabilities of the observed values.

6.3.1 Overview of Weights

It is interesting to see how the weights vary. If we take an extreme case where all the weights are equal, then the pseudo log-likelihood function is just a scaled version of the original log-likelihood. In that case the arg max would be the same.

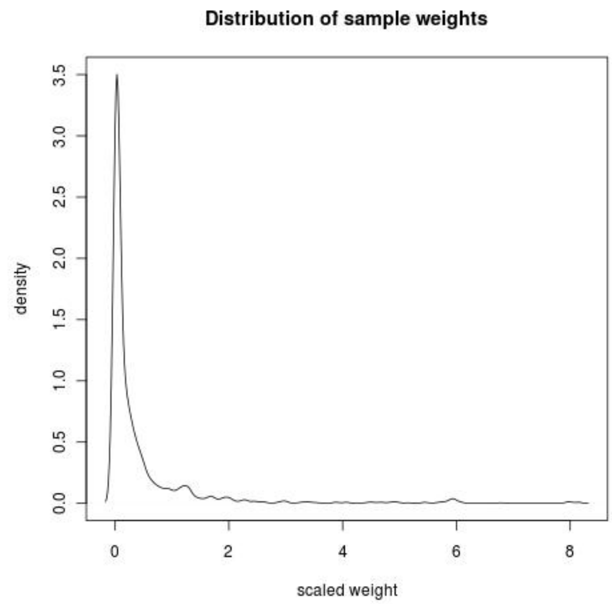

Figure 6.1: Distribution of weights

If the weights vary a lot, we expect the two functions to have a different maximum. Most weights (1,4126 on 1,497) are less than 2; then we expect their distribution on the $[0, 2]$ interval to be as follows: The other 71 weights are relatively big compared to 2 so they have the potential to shift the maximum of the weighted SLL function.

6.4 Variance Estimation

Unfortunately, It is not possible (or at least difficult) to compute the standard errors using the asymptotic theory here so we must rely on re-sampling techniques. We do not compute the standard errors here as we are more interested in the evolution of the coefficients than in the coefficients themselves.

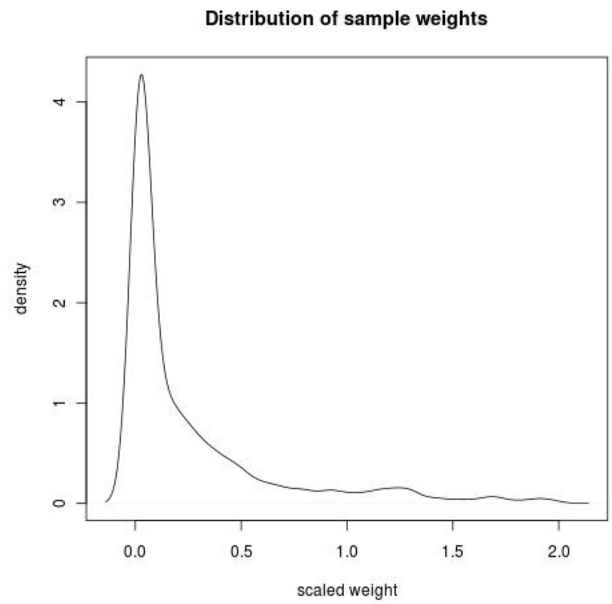

Figure 6.2: Distribution of weights

6.5 New Coefficients

6.5.1 Probit

predictor	alternative	MLE	Pseudo-MLE	difference	% change
constant	1 car	1.33	1.67	0.34	25.80
	2 cars	-0.91	-3.09	-2.18	240.88
	3 cars	-1.49	-2.13	-0.63	42.26
	4+ cars	-0.74	-2.31	-1.58	213.32
income	1 car	0.30	0.43	0.13	41.94
	2 cars	1.06	1.67	0.60	56.63
	3 cars	1.05	0.85	-0.20	-18.75
	4+ cars	0.65	1.23	0.58	88.51
education	1 car	1.75	3.05	1.30	74.11
	2 cars	-0.70	-3.16	-2.46	351.20
	3 cars	-2.07	-2.08	-0.00	0.17
	4+ cars	-2.87	-4.38	-1.51	52.67
num. of drivers	1 car	-2.95	-3.26	-0.32	10.71
	2 cars	1.20	2.22	1.02	85.59
	3 cars	1.58	3.25	1.66	104.42
	4+ cars	0.98	3.38	2.39	244.09

num. of workers	1 car	-1.85	-3.07	-1.22	66.29
	2 cars	-0.15	0.15	0.30	-197.71
	3 cars	0.65	1.89	1.25	192.19
	4+ cars	0.68	-0.01	-0.69	-101.27
urban density	1 car	-0.23	-0.71	-0.48	205.26
	2 cars	-1.68	-1.50	0.18	-10.92
	3 cars	-3.03	-3.74	-0.71	23.43
	4+ cars	-3.40	-3.48	-0.08	2.36
house ownership	1 car	0.63	-1.82	-2.44	-389.58
	2 cars	0.11	1.42	1.30	1100.87
	3 cars	-0.34	0.31	0.66	-190.47
	4+ cars	0.06	-0.41	-0.47	-778.80

First, it should be noted that including the weights in the computation of the SLL function dramatically shifts its maximum. New coefficients are easily twice as big (or as small) when compared to the coefficients obtained using ordinary maximum likelihood estimation.

Meanwhile, the sign of the coefficients rarely change, which is very good news. For instance, it is very natural to assume that living is a dense city is an incentive to own fewer cars and we would be puzzled if this conclusion depended on the estimator we used for the coefficients. In other words, even if we think the MLE estimator is not the most reliable one, we would be uncomfortable if the sign of the coefficients it estimates were different from those obtained with pseudo-MLE estimator gives.

Moreover, the following differences should be stressed:

- constants change drastically; the constant for the 4+ cars goes down by 213%!
- the effect of income does not vary much except for the 4+ cars alternative

where it is almost doubled (88%)

- when using weights, education effect on the utility of owning 2 cars decreases.
- we underestimated the effect of the number of drivers and workers for all alternatives.
- the effect of urban density is not modified much by including weights except for the 1 car alternative (doubled)
- the effect of home ownership drastically increases in the pseudo max-LL estimates.

6.5.2 Regression

predictor	MLE	pseudo-MLE	change	% change
constant	1.14	1.85	0.71	62.3
income	0.69	0.65	-0.04	-5.1
education	-0.44	-1.24	-0.81	185.0
num. of drivers	5.70	6.67	0.97	16.9
num. of workers	3.52	2.62	-0.89	-25.4
urban density	-0.84	-0.79	0.06	-6.8
house ownership	1.36	1.79	0.43	31.3

The regression part of the DC model is not affected much by the sampling weights and certainly less than the probit part (except for the variance). We note that except for education, the coefficients do not shift that much. Education, however, sees its importance in explaining miles traveled multiplied by three. That is, ignoring the weights severely underestimates how education of the household reduces the number of miles traveled by the households.

It is interesting to observe that income (ranged from 1 to 18) and education (ranged from 1 to 5) tend to cancel each other when they are at comparable levels.

Home ownership saw its effect changing in the probit, however the effect of owning a home does not change much when predicting the number of miles traveled. The contribution of owning a home in the predicted number of miles traveled (in thousands) is only increased by 430 miles.

6.5.3 Covariance Matrix

$$\hat{\Sigma}_{pseudo} = \begin{bmatrix} 1.00 & 0.18 & 0.48 & -0.50 & -0.42 & -1.06 \\ 0.18 & 136.70 & 9.02 & 16.73 & 13.21 & -8.11 \\ 0.48 & 9.02 & 96.97 & 1.69 & 2.67 & 31.33 \\ -0.50 & 16.73 & 1.69 & 67.89 & -0.88 & 36.00 \\ -0.42 & 13.21 & 2.67 & -0.88 & 74.11 & 22.97 \\ -1.06 & -8.11 & 31.33 & 36.00 & 22.97 & 129.88 \end{bmatrix}$$

The following matrices show the variation (in value and percentage) of the new coefficients compared to the previous ones:

$$\Delta\Sigma = \begin{bmatrix} 0 & -0.112 & 0.179 & 0.0784 & -0.221 & -0.592 \\ -0.112 & 27.0 & 5.44 & 6.9 & 12.4 & 17.2 \\ 0.179 & 5.44 & -12.8 & -6.25 & -0.0312 & 32.9 \\ 0.0784 & 6.9 & -6.25 & -25.6 & -2.50 & 12.6 \\ -0.221 & 12.4 & -0.0312 & -2.50 & -20 & -5.08 \\ -0.592 & 17.2 & 32.9 & 12.6 & -5.08 & -79.3 \end{bmatrix}$$

$$\Delta\Sigma_{\%} = \begin{bmatrix} 0 & -38.6 & 59.4 & -13.6 & 111 & 126 \\ -38.6 & 24.6 & 152 & 70.2 & 1559 & -68 \\ 59.4 & 152 & -11.7 & -78.8 & -1.15 & -2109 \\ -13.6 & 70.2 & -78.8 & -27.4 & -154 & 54.1 \\ 111 & 1559 & -1.15 & -154 & -21.3 & -18.1 \\ 126 & -68 & -2109 & 54.1 & -18.1 & -37.9 \end{bmatrix}$$

We observe that the variance of the regression decreases from 209 to 130. The weights help to "sort out" observations that do not behave well and thus make variance higher in order to reach them.

The variances of the error terms are reduced, except for the 1 car alternative. This result improves the efficiency of the model, given that high variance in the utilities makes it more difficult to distinguish between alternatives, whose utilities are too volatile.

Covariance terms do not change much. We note that the weights boost the covariance of number of miles traveled and utility of owning 2 cars (the most popular alternative). This result is particularly interesting as high error terms for high car ownership alternatives (2, 3, 4+ cars) correspond to high error terms in the number of miles traveled. High error terms in low car ownership alternatives (0 or 1 car) correspond to negative error terms in the number of miles traveled. Previously, we got similar results but we were not able to correlate miles traveled with the utility of owning 2 cars.

Covariance between utilities error are shrunk in the pseudo maximum likelihood. However, it is not easy to give a clear explanation for this result. Covariance terms help to understand how people trade between alternatives. Overestimating them suggests an artificial understanding of how people trade off between different car ownership possibilities.

6.6 Pros and Cons

6.6.1 SLL Computation

The computation of the scaled SLL function is absolutely *not* more difficult than the ordinary SLL function. The only counter-indication has to do with the scale of the weights. Extremely small weights may give a too flat pseudo SLL function, causing the solver to stop earlier than we would expect. Extremely high weights may cause failure in the optimization process since the calibration we use for gradient computations and predictor scales will be canceled by weights.

In addition, ignoring the sampling design does not help in the computation of the standard errors. Due to all numerical problems in the computation of the SLL and its derivatives, we are not comfortable estimating the standard errors with the hessian matrix. It is not more complicated to use bootstrap for the evaluation of the pseudo-SLL than it is for the ordinary SLL. In fact, the generation of new Ys does not involve sampling weights. Assuming that we are able to maximize the pseudo SLL function, then bootstrap can be used for variance estimation.

Chapter 7

Concluding Remarks

7.1 DC Model Review

In this dissertation, we have been able to combine two of the most elementary, yet powerful, models: the multinomial probit and the linear regression. We successfully correlated the total number of miles traveled with not only the number of cars owned, but also with the utilities household have for any car ownership option available. We present in this Section the major contributions from this research work and suggestions for future research.

Our problem involved formulating a simulated likelihood function problem and maximizing it. The deriving likelihood function is not tractable numerically and should be solved trough simulation. However, simulated likelihood suffers from discontinuity problems that prevented us from computing the standard errors of our estimates. As a consequence, we used bootstrap re-sampling technique to calculate confidence intervals of our estimates, to the expense of an increased computational effort.

The Discrete Continuous model was successfully calibrated on data extracted from the 2009 National Household Travel Survey. The model correctly estimates (1) the positive correlation between the 3 and 4 cars ownership options, (2) the error terms of number of miles driven as well as (3) the negative correlations between the 0 and 1 car ownership options and the number of cars driven.

Also, we have been able to incorporate sampling weights in the estimation procedure and their effects were quantified. This resulted into variance reduction for the errors on the estimated coefficients. Thus, the sampling weights helped us to reduce the importance of the observations that were ill-behaved with respect to the model. In particular, by incorporating the weights, the variance of the regression for the number of miles traveled was reduced by nearly 40 %

7.2 Future research

In the optimization process, it was necessary to deal with "meta-parameters" that affect convergence, computing time, etc. We fixed those parameters by setting very strict rules, yet allowing the solver to optimize the SLL function within a reasonable amount of time:

- derivatives' δ as small as possible
- number of simulations as big as possible
- relative tolerance on coefficients as small as possible
- absolute tolerance on coefficients as small as possible

When approaching the solution these strict values might not longer be required. Adaptative algorithm could be developped to control the number of simulations and in particular, the number of draws can be reduced in the initial stages of the optimization process when less precision is required. A higher number of simulations at the end is expected to increase precision on both the log-likelihood function and parameter estimates.

7.3 Closed Form

7.3.1 Probit

The probability involved in the computation of the probit likelihood is:

$$P(Y = y) = \int_{\mathbb{R}^k} \mathbb{I}(X_y^T \beta_y + \epsilon_y > X_j^T \beta_j + \epsilon_j \quad \forall j \neq y) \phi(\epsilon) d\epsilon$$

It is possible to rewrite the utilities and error terms into a $(k-1)$-dimensional vector of error terms for which the indicator function will be true over $D \subset \mathbb{R}^k$, thus we can integrate the normal density alone on D *without* the indicator. Furthermore, D's lower bound is minus infinity and its upper bound is some value $(\alpha_1, \ldots, \alpha_{k-1})$, such that the integral is only the normal cdf. [Tra09, p. 98]. The multivariate normal cdf is not analytically tractable but approximations for it have been suggested in the maximum approximate composite marginal likelihood (MACML) estimator. [Bha11]

7.3.2 Conditional Regression

In the earlier stage of our work, we have been using by-products of the probit simulations. As a consequence, any improvement in the probit estimation that removes the simulation must *also* provide a way to estimate the conditional density of the regression:

$$(X_1, \ldots, X_k) \sim N(\mu, \Sigma)$$

It is necessary to be able to derive this distribution, or at least an approximation of it:

$$X_1 | X_2 < x_2, \ldots, X_k < x_k$$

Our empirical observations suggest that this conditional distribution is normal. If we were able to derive this conditional distribution, our DC model would have a *closed form*. The payoff of such a closed form is very high since it would eliminate almost all the numerical issues that we discussed. In particular:

- Computation time would be sensibly reduced;
- Accurate derivative computations would improve the precision of the maximum we compute;
- Accurate mixed derivative computations would help to calculate reliable Hessian-based t-statistics.
- The DC model would handle higher predictor dimension better.

7.3.3 Variable Selection Implications

Also, such a closed form would greatly improve the variable selection involved in the use the DC model. As it is currently possible to estimate coefficient estimates relatively fast, the calculation of their statistical significance is computationally very expensive. However, it is possible to monitor the evolution of the values of the SLL at the maximum by adding predictors, but this method lacks flexibility for the researcher that wants to test different specifications.

Bibliography

[AGM03] Micah Altman, Jeff Gill, and Michael P. McDonald. *Numerical Issues in Statistical Computing for the Social Scientist*. Wiley-Interscience, Hoboken, New-Jersey, first edition, 2003.

[BF05] Richard Burden and J. Douglas Faires. *Numerical Analysis*. Thomson Brooks/Cole, Belmont, California, eighth edition, 2005.

[Bha05] Chandra 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] Chandra 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.

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

[BSE09] Chandra R. Bhat, Sudeshna Sen, and Naveen 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.

[Chi93] Radeep K. Chintagunta. Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Marketing Science*, 12(2):184–208, Spring 1993.

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

[DN89] C. Liu Dong and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

[Efr87] Bradley Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial Mathematics, Philadelphia, Pennsylvania, first edition, 1987.

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

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

[Joh07] Steven G. Johnson. The nlopt nonlinear-optimization package, 2007. `http://ab-initio.mit.edu/nlopt`.

[Law07] Averill M Law. *Simulation Modeling and Analysis*. McGraw Hill, New-York, New-York, fourth edition, 2007.

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

[Noc80] Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[Owe03] A. B. Owen. Quasi-Monte Carlo sampling. In H. W. Jensen, editor, *Monte Carlo Ray Tracing: Siggraph 2003 Course 44*, pages 69–88. SIGGRAPH, 2003.

[Par62] Emanuel Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[Ric07] John A. Rice. *mathematical statistics and data analysis*. Duxbury, Belmont, California, third edition, 2007.

[San10] Conrad Sanderson. Armadillo: An open source c++ linear algebra library for fast prototyping and computationally intensive experiments. software, NICTA, 2010.

[SPPB09] Erika Spissu, Abdul Rawoof Pinjari, Ram M. Pendyala, and Chandra R. Bhat. A copula-based joint multinomial discrete-continuous model of vehicle type choice and miles of travel. *Transportation*, 36(4):403–422, 2009.

[Str97] Bjarne Stroustrup. *The C++ programming language*. Addison-Wesley Professional, Boston, Massachussets, third edition, 1997.

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

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

[UDoT01] Federal Highway Administration U.S. Department of Transportation. National household travel survey, 2001. `http://nhts.ornl.gov`.

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

[Wei05] Sanford Weisberg. *applied linear regression*. John Wiley & Sons, Hoboken, New-Jersey, third edition, 2005.