UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI

"M.FANNO"

CORSO DI LAUREA MAGISTRALE IN

ECONOMICS AND FINANCE

TESI DI LAUREA

MONTE CARLO SIMULATION IN PRESENCE OF SAMPLING ERRORS IN PANEL DATA MODELS

RELATORE:
CH. MO PROF. LUCA NUNZIATA

LAUREANDA: SANJA JANKOVIC
MATRICOLA N. : 1081719

ANNO ACCADEMICO 2014 - 2015

Il candidato dichiara che il presente lavoro è originale e non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Il candidato dichiara altresì che tutti i materiali utilizzati durante la preparazione dell'elaborato sono stati indicati nel testo e nella sezione "Riferimenti bibliografici" e che le eventuali citazioni testuali sono individuabili attraverso l'esplicito richiamo alla pubblicazione originale.

Firma dello studente

Contents

Abstract

La tecnica di campionamento è importante per risolvere il problema di studiare la popolazione avendo a disposizione dei campioni.
Questo procedimento ha degli svantaggi e il principale è rappresentato dall'errore di campionamento, che nasce proprio dal fatto di selezionare un'unita rappresentativa al posto di prendere la popolazione intera.

Un campione affetto dall'errore porta ad una sovrarappresentazione o sottorappresentazione della popolazione reale; infatti, i ricercatori trarranno conclusioni non corrette sulle caratteristiche della popolazione studiata.
Praticamente ogni campione è affetto dall'errore di campionamento in quanto è difficile ottenere un campione perfettamente randomizzato.

Il metodo di Monte Carlo è importante in questo tipo di analisi perché è in grado di misurare l'errore presente in un campione e la performance degli stimatori utilizzati.
In questo studio è importante avere a disposizione dei dati panel, i quali permettono di osservare gli individui e le loro caratteristiche nel corso del tempo.

Molti studi hanno dimostrato che l'errore è legato alla numerosità del campione. Infatti, campioni piccoli hanno un errore elevato e al crescere del campione l'errore decresce.
Questo è dato dal fatto che un campione più grande comprende un numero elevato di individui e si avvicina di più all'intera popolazione.

Lo scopo di questo studio è quello di trovare il campione ideale per ottenere una rappresentazione il più veritiera possibile della popolazione reale.

Introduction

In many experiments it is impossible to study the entire population, because of the large number of subjects.
Therefore, the unique way to resolve this problem is to sample, so select a unit from the population that is representative of the population and inference its characteristics from it.

The main disadvantage of this process is the sampling error which is the error that arises in a data collection process as a result of taking a sample from a population rather than using the whole population.
The other type of error present in the study is the non-sampling error.
Both of them could be measured, but it is difficult to avoid them in inference process.

A biased sample can lead to an over or underrepresentation of the corresponding parameter in the population.
Almost every sample in practice is biased because it is practically impossible to ensure a perfectly random sample.

Sampling error is correlated with sample size; indeed, a large sample size has less sampling error respect a small sample size, in a study where the population is the same and all the its characteristics are the same.
Bigger sample studies more subjects of the population and could represent more characteristics of it.

In sampling method it is important to have a dataset that observe some characteristics of entities across time. This is called panel data.
Its principle advantage is the fact that it includes time and individual variation which are unobservable in cross sections or aggregate time series.
Panel data is widely used to estimate dynamic econometric models. Many economic issues are dynamic by nature and use the panel data structure to study how these entities change from year to year.

In these type of analysis Monte Carlo simulation is an adequate method to measure the sampling errors in samples and the performance of the estimators.
The experiment starts with the identification of statistical distribution for

each input parameters.
Then, it draws random samples from each distribution, which then represents the values of the input variables. For each set of input parameters, it gets a set of output parameters.
Finally, it performs statistical analysis on the outputs to take the final decisions and inference the characteristics of the population.

Monte Carlo simulation is applicable in many areas, such as mathematics, statistical physics, finance, engineering, reliability analysis, physical sciences and others.
In the study of this paper simple Monte Carlo is used to estimate the population expectation by the corresponding sample expectation.
The sample values usually get a good enough idea of the error.

The following chapters of this paper is organized in this way: the first chapter is about the sampling process and sampling error, the second chapter explains panel data and its characteristics, then, the third chapter is about Monte Carlo method and the last chapter explains specifically simple Monte Carlo, a method used to estimate the sampling error, and presents some case studies which demonstrate the utilization of this method for different experiments.

Chapter 1

Sampling Error

1.1 Definition

Sampling is a process of selection of a sample for studying the inference of the population and, therefore, be able to obtain an adequate description of the population.
In many experiments it is impossible to study the entire population because of the time, expense and a large number of individuals. For these reasons it is necessary to sample, that is, select a unit of the population that is representative of the whole population.

The steps in sampling process are the following:

1. Define the population
2. Identify the sampling frame
3. Select a sampling design
4. Determine the sample size
5. Draw the sample

A sample is expected to mirror the population from which it comes, however, there is no guarantee that any sample will be precisely representative of the population from which it comes.
In sample surveys, since inference is made about the entire population covered by the survey on the basis of data obtained from only a part of the population, the results are likely to be different than if a complete census was taken under the same general survey conditions.

The sampling error is the error that arises in a data collection process as a result of taking a sample from a population rather than using the whole population.
The sampling error depends on factors such as the size of the sample, variability in the population, sampling design and method of estimation.
Further, even for the same sampling design, it can make different calculation to arrive at the most efficient estimation procedure.

The most frequent cause of the sampling error is a biases sampling procedure. Every researcher must seek to establish a sample that is free from bias and is representative of the entire population.
Another possible cause of this error is chance. The process of randomization and probability sampling is done to minimize sampling process error but it is still possible that all the randomized subjects are not representative of the population.

The most common results of sampling error is systematic error wherein the results from the sample differ significantly from the results from the entire population.
Indeed, sampling error is one of the reason for the difference between an estimate of a population parameter and the true, but unknown, value of the population parameter.

Even non-sampling error can cause the results distorted and far from the true value of the population.

Non sampling errors can be attributed to one or more of the following sources:

- *Coverage error:* it results from incomplete listing and inadequate coverage of the population of interest
- *Data response error:* this error may be due to questionnaire design and the characteristics of the question, misinterpretation of the questions and different tendencies of different interviewers in explaining questions or interpreting responses
- *Non-response error:* some respondents may refuse to answer questions, some may be unable to respond, while others may be too late in responding
- *Processing error:* errors that may occur at various stages of processing such as coding, data entry, verification

Non-sampling errors are difficult to measure. Many attempts have been made to minimize the non-sampling errors. It tries to define more precisely the samples and questionnaires are structured to avoid different interpretations.

1.2 Measures of sampling and non-sampling errors

1.2.1 Sampling error measures

Several samples may be subject to the same investigation but every estimate provides is different. The average estimate given by all the possible sample estimates is the expected value.
An estimate of a sample survey is called precise if it is near the expected value. The variability of the sample estimates with respect to its expected value can be measured.

The variance of estimate is a measure of the precision of sample estimate and is defined as the average of the squared difference of the estimates from its expected value.
The standard error, defined as the square root of the variance, is a measure of the sampling error in the same units as the estimate. The standard error is a measure of precision in absolute terms.

The coefficient of variation is a measure of precision in relative terms.
With the use of the coefficient of variation it can compare the sampling error of an estimate with that of another estimate.
The coefficient of variation is given by the following formula:

$$CV(X) = \frac{S(X)}{X} \tag{1.1}$$

where X denotes the estimate and S(X) denotes the standard error of X.
Confidence interval can be constructed around the estimate using the estimate and the coefficient of variation.

1.2.2 Non-sampling error measures

A census sample survey aims to obtain the exact value of the population. Any difference between the expected value and the exact value of the population is defined bias.
The accuracy of a survey estimate is determined by the joint effect of sampling and non-sampling errors.
The response fraction, which is a measure of the data response rate, is the proportion of the sales estimate which is based upon reported data.
The lower the coefficient of variation and the higher the response fraction, the better will be published estimate.

1.3 Types of sampling

1.3.1 Probability sampling

This method is important becuase it guarantees that the selection process is completely randomized and without bias. Indeed, in it each individual has the same chance to be selected for the experiments.

The main advantages of using this technique are the following:

- high accuracy of statistical methods after the simulation
- useful to estimate the population parameters bacause it is representative of the entire population
- reliable method to eliminate sampling bias

1.3.1.1 Simple random sampling

The random sampling is one of the most popular types of probability sampling. All members are included in the list and the entire process is performed in a single step where each subject can be selected independently of the other members of the population.
There are many methods by which to perform random sampling. The simplest method is the lottery. Then, another method may be to leave a computer to execute the random sampling of the population.

This method can be considered as fair as each member has an equal chance of being selected during sampling. Another key feature is the representation of the population, then it is reasonable to make generalization from the results of the sample back to the population.
If the sample is not representative of the population, the differences will be defined random sampling error.

1.3.1.2 Stratified random sampling

This is a technique wherein the subjects are initially grouped into different classifications such as age, socioeconomic status or gender. Researchers usually use this method if they want to study a particular subgroup within the population.
The researcher randomly selects the final list of subjects from the different strata, so it must use the technique of simple random sampling within the different strata.

It warrants more precise statistical outcomes than the simple random sampling.
The layers are not to be overlaid so as not to grant the highest probability of being selected to subjects at the expense of others.

This technique is used most when researchers want to study a specific group or when they want to see the relationships between the different subgroups.
It is equipped with a statistical accuracy higher than simple random sampling. It also requires a small sample size that can save time, money and energy.
With this method the researchers can sample and analyze even the smallest and most inaccessible subgroup in the population.

1.3.1.3 Systematic random sampling

Systematic sampling is a technique of random sampling which it can be linkened to an arithmetic progression where the difference between any two consecutive numbers is the same.

The main advantage of this method, which is frequently chosen by the researchers, is its simplicity and, in addition, can be carried out manually. This allows the researchers to add a degree of system into the random selection of subjects.
Another advantage is the assurance that the population will be evenly samples.

1.3.1.4 Cluster random sampling

When it can not use the simple random sampling due to the size of the population then it is used the technique of cluster random sampling.

The researcher takes several steps in gathering its sample population. First, it identifies the boundaries.
Then, it randomly selects a number of identified areas. That it is important that to all areas within the population is given an equal chance of being selected.
Finally, the researcher can either include all the individuals within the selected areas or it can randomly selects subjects from the identified areas.
The most common use of cluster in research is a geographical cluster.

The principle advantage of this technique is the fact of being economical, simple and fast.
On the other hand, the main disadvantage is the fact of being less representative of the whole population.

The tendency of individuals within a cluster is to have common characteristics, then with a sample cluster the researcher can get a cluster that is overrepresented or underrepresented, so obtaining false results.
Consequently, this technique is subject to a high sampling error.

1.3.1.5 Multi-stage random sampling

This technique includes two different cases of sampling: proportional sampling and disproportional sampling.
The first of them is made from layers that have the same sampling fraction; instead, in the second case the sampling fraction varies from layer to layer.

The disproportional sampling is a technique used to address the difficulties given by the analysis of stratified samples of unequal size.
Disproportional sampling allows the researcher to give a larger representation to one or more subgroups to avoid underrepresentation of the said strata.
Generally, disproportional sample tend to be less accurate and reliable compared to a stratified sample, since mathematical adjustments are done during the analysis of the data.
This process increases the chance of encountering errors in data analysis. It is less accurate in drawing conclusions from the results of such studies.

1.3.2 Non-probability sampling

In non-probability sampling, just the opposite of probability sampling, members of the population do not have the same chance of being selected.
In this case, the sample studied does not fully represent the population of interest and, therefore, the research results can not be used to draw general conclusions and inference about the population.

The subjects, in fact, are selected on the basis of their accessibility and according to the personal judgment of the researcher.
Proceeding in this way, an unknown proportion of the population was not sampled and the sample is not the result of a randomized selection process.

Researchers choose this method when they are interested in certain parameters and not to the entire population.
Also, this technique likes because it is economical, simple and fast.

1.3.2.1 Convenience sampling

In this technique, subjects are selected for their accessibility and proximity to the purpose of the researcher. In fact, the subjects are not representative of the whole population, but were chosen because they are easier to recruit.

This approach has many problems and criticism.
The problems are systematic errors that it gets from this sampling. Very frequently the results significantly differ from the results of the entire population.
A consequence of having systematic bias is obtaining skewed results.
Principle criticism is determined by the fact that the sample is not representative of the entire population then the results of the study can not speak for the whole population and, therefore, it can not generalize and draw conclusions on the population. The results are not valid and truthful.
Some researchers use this technique only for its speed, for cost-effectiveness and ease of finding the topics of study.

1.3.2.2 Consecutive sampling

Sequential sampling is a non-probability sampling technique wherein the researcher picks a single or a group of subjects in a given time interval, conducts his study, analyzes the results and, then, picks another group of subjects if needed and so on.
This method of sampling is repetitive, as to correct analysis and refine the search method was enough to make small changes and repeat the experiment. This method is dependent on the researcher because only after conducting the experiment for the first group of samples is possible to proceed with a second group of samples.

There is very little effort in the part of the researcher when performing this sampling technique. It is not expensive, not time consuming and not workforce extensive.

After the experimets, the researcher can accept the null hypothesis or refuse it and accept the alternative hypothesis. There is another possible choice, that is represented by another experiments with another pool of subjects.

The problems that arise from this technique are the non-representativeness of the population and the sample is not obtained from a randomization. This involves an ultra-low level of representativeness of this sampling technique.
Due to the aforementioned disadvantages, results from this sampling technique cannot be used to create conclusions and interpretations pertaining to the entire population.

1.3.2.3 Quota sampling

Quota sampling is a non-probability sampling technique wherein the assembled sample has the same proportions of individuals as the entire popu-

lation with respect to known characteristics, traits or focused phenomenon.

The first step in non-probability quota sampling is to divide the population into exclusive subgroups. Then, the researcher must identify the proportions of these subgroups in the population; this same proportion will be applied in the sampling process.
Finally, the researcher selects subjects from the various subgroups while taking into consideration the proportions noted in the previous step.
The final step ensures that the sample is representative of the entire population. It also allows the researcher to study traits and characteristics that are noted for each subgroup.

It may appear that this type of sampling technique is totally representative of the population. In some cases it is not real. In fact, while sampling some of the characteristics of the sample may be overrepresented.
In a study examining characteristics such as gender, religion and socioeconomic status, the final results might give a misrepresentation of age, race, level of instruction and other features.

1.3.2.4 Judgmental sampling

In this technique, the criteria by which the researcher takes its decisions have the knowledge of the subjects to be analyzed and its professional judgment.
The major use for this sampling is found in cases where a limited number of individuals possesses the characteristic of interest for which it has to choose a sample composed of a very specific group of people.

There is usually no way to evaluate the reliability of the expert or the authority.
The best way to avoid sampling error brought by the expert is to choose the best and most experienced authority in the field of interest.

From the description above, it can deduce that in this method the subjects have not an equal chance of being selected, so it is not a random sampling.
The main consequence is the fact that it can not generalize and obtain an inference of the entire population by the results of this sampling.

1.3.2.5 Snowball sampling

This method of sampling is used by researchers in cases where the study sample is rare and is composed of a small subset of the population.
Snowball sampling is the proper technique to select the subjects difficult to identify for analysis.

After observing the initial subject, the researcher asks for assistance from the subject to help him to identify people with similar trait of interest.

The chain referral process allows the researcher to reach populations that are difficult to sample when using other sampling methods. The subjects that the researcher can get are based on the topics previously observed.
In fact, the results obtained are distorted because the initial subjects tend to suggest people who know.
Because of this, it is highly possible that the subjects share the same traits and characteristics, thus, it is possible that the sample that the researcher will obtain is only a small subgroup of the entire population.

Proceeding in this way, the researcher has little control over the selected samples, but, on the other hand, someone would choose this technique because it is cheap, simple and convenient.

1.4 Sample size and sampling error

Sampling error is correlated with sample size; indeed, a large sample size has less sampling error respect a small sample size, in a study where the population is the same and all the its characteristics are the same.
Bigger sample studies more subjects of the population and could represent more characteristics of it.
Aydemir and Borjas[1] studied the attenuation bias in measuring the wage impact of immigration because they wanted to demonstrate the role and the importance of the sampling error [1].

The classical economic theory suggests that an increase in supply of immigrant workers has a negative impact on the relative price of labor of the natives.
First, this impact is mitigated by the measurement errors present in the analysis of data.
Also, the sampling error plays an important role in the studies that commonly are used to measure the impact on the wages of natives caused by immigration.
In their analysis they used labor market data drawn from both Canada and the United States to show that the attenuation bias is quite important in empirical context of estimating the wage impact of immigration and adjusting for the attenuation bias can easily double, triple, and sometimes even quadruple the estimated wage impact of immigration.

[1] A. Aydemir and G.J. Borjas, Attenuation bias in measuring the wage impact of immigration, September 2010

The attenuation bias becomes exponentially worse as the size of the sample used to calculate the immigrant share in the typical labor market declines. They used large data files to try to conduct a very accurate and precise analysis. Eventually they were able to verify that also with samples of medium and large is possible to achieve significant sampling errors and, therefore, infer wrong conclusions on the economic impact of immigration.
Consequently, measurement errors have an important role and often contaminate the results of the experiments, so it has to try to reduce them as much as possible and take account of them while it deduces the conclusions of the experiments.
The sense of security that some researchers have in large microdata is not true; in fact, many studies are extremely sensitive and easily influenced by sampling errors.

1.5 Problems caused by sampling bias

A biased sample causes problems because any statistic computed from that sample has the potential to be consistently erroneous.
It is almost impossible to get an unbiased sample and completely random, so it gets an overrepresentation or underrepresentation of the characteristics of the studied population.
If the error is small then the sample can be accepted and treated as a reasonable approximation of a random sample.

Researchers could use a biased sample to produce false results to confirm their studies, but in most cases those who get a biased sample attributes it to a difficulty in obtaining a truly representative sample.
Some samples use a biased statistical design which nevertheless allows the estimation of parameters. Some surveys require the use of sample weights to produce proper estimates across all groups.
Provided that certain conditions are met, that the sample is drawn randomly from the entire sample, these samples permit accurate estimation of population parameters.

There is only one way to eliminate this error. This solution is to eliminate the concept of sample and to test the entire population. In most cases this is not possible.
Consequently, what a researcher must to do is to minimize sampling process error. This can be achieved by a proper and unbiased probability sampling and by using a large sample size.

Chapter 2

Panel data models

The panel data, also known as time series longitudinal or transverse, are a set of data where it is possible to observe the behavior of the same items over time. These may be states, businesses or individuals.

The panel data is used monitor various types of time-varying, as cultural factors, differences in business practices or variables that change over time, but not between individuals.
They allow its to do an analysis at different levels to study multilevel or hierarchical models.

Panel data is called balanced when all individuals are observed in all time periods or unbalanced when the individual are not observed across all time, but it is available data from some years, not all.

The principle advantages are the following:

- Time and individual variation in behavior unobservable in cross sections or aggregate time series
- Observable and unobservable individual heterogeneity
- Rich hierarchical structures
- More complicated models
- Features that cannot be modeled with only cross section or aggregate time series data alone
- Dynamics in economic behavior

2.1 Data structures

Time series and cross-sectional data are special cases of panel data that are in one dimension only (one panel member or individual for the former, one time point for the latter).

Time series data:

- $X_t, t = 1, ..., T$, univariate series, its path over time is modeled. The path may also depend on third variables.
- Multivariate, their individual as well as their common dynamics is modeled. Third variables may be included.

Cross sectional data:

- These data are observed at a single point of time for several individuals, countries, assets. $X_i, i = 1, \ldots, N$.
- The interest lies in modeling the distinction of single individuals, the heterogeneity across individuals.

Pooling data refers to two or more independent data sets of the same type.

Pooled time series:

- Return series of several factors, which are assumed to be independent of each other, togheter with explanatory variables. The numbers of sectors, N, is usually small. Observations are viewed as repeated measures at each point of time. So parameters can be estimated with high precision due to an increased sample size.

Pooled cross sections:

- Mostly these type od data arise in surveys, where people are asked about some arguments. This survey is repeated T times before elections every week. T is usually small. So it has several cross sections, but the persons asked are chosen randomly. Hardly any person of one cross section is member of another one. The cross sections are independent.

Panel data set has both a cross-sectional and a time series dimension, where all cross section units are observed during the whole time period, like in this formula:

$$X_{it}, i = 1, ..., N, t = 1, ..., T$$

where T is usually small.

The panel models analyze the same individual for several periods, instead the pool sections perform repeated random selections of individuals. The pooling model is appropriate if individuals are randomly selected in each period.
Use of panel models in the study of securities observes the same title in several periods, so it can see if a stock with high performance will continue to have the same return in the next period or will change.
For these reasons, the panel model is more efficient and more appropriate model of pooling model.

2.2 Static linear panel data model

The two basic models for the analysis of panel data are the fixed effects model and the random effects model.
Panel data are most useful when it is suspect that the outcome variable depends on explanatory variables. If such omitted variables are constant over time, panel data estimators allow to consistently estimate the effect of the observed explanatory variables.
The standard static model with $i = 1,...,N$, $t = 1,...,T$ is

$$Y_{it} = \beta_0 + x'_{it}\beta + \epsilon_{it} \tag{2.1}$$

where x'_{it} is a k dimensional vector of explanatory variables, without a constant term, β_0 the intercept is independent of i and t, β is independent of i and t, ϵ_{it} the error, varies over i and t.
Individual characteristics (which do not vary over time) z_i may be included

$$Y_{it} = \beta_0 + x'_{it}\beta_1 + z'_i\beta_2 + \epsilon_{it} \tag{2.2}$$

Two main problems of this model are endogeneity and autocorrelation in the errors:

- *Consistency/exogeneity*: assuming iid errors and apppling OLS it gets consistent estimates, if $E(\epsilon_{it}) = 0$ and $E(x_{it}\epsilon_{it}) = 0$, if the x_{it} are weakly exogenous

- *Autocorrelation in the errors*: since individual i is repeatedly observed $Corr(\epsilon_{i,s}, \epsilon_{i,t}) \neq 0$, with $s \neq t$ is very likely

Then, standard errors are misleading and OLS is inefficient.
Unobserved individual factors, if not all z_i variables are available, may be captured by a_i.
It decomposes the error in

$$\epsilon = a_i + u_{it} \tag{2.3}$$

with $u_{it} iid(0, \sigma_u^2)$, where u_{it} has mean 0, is homoscedastic and not serially correlated.
All individual characteristics, including all observed as well as all unobserved ones, which do not vary over time, are summarized in the $a_i's$.

2.2.1 Fixed effects model

In the fixed effects model, the individual-specific effect is a random variable that is allowed to be correlated with the explanatory variables.
The standard fixed effects model is represented by the formula below:

$$Y_{it} = a_i + x'_{it}\beta + u_{it} \tag{2.4}$$

no overall intercept is included in (2.4).
Under FE consistency does not require, that the individual intercepts and x_{it} are uncorrelated. Only $E(x_{it}u_{it}) = 0$ must hold.

The data are often divided into categories such as industries, states, families and, when in this case, it is appropriate to control the characteristics of these categories.

When estimating a linear OLS have to worry about the unobservable characteristics that may correlate with the variables included in the regression.
The phenomenon of omitted variables is very common and cause a problem of endogeneity.
This problem could be solved with regressions with fixed effects because if the unobservable characteristics do not vary over time, that is, are the fixed effects, it will be possible to eliminate omitted variable bias.

In some cases, it might believe that its set of control variables is sufficiently rich that any unobservable are part of regression noise, and therefore omitted variable bias is nonexistent. But it can never be certain about unobservable because they are unobservable.

So fixed effects model are a nice precaution even if it thinks to might not have a problem with omitted variable bias.
If the unobservable is not time-invariant then it still has omitted variable bias. It may never be able to fully rule out this possibility.

A significant problem of this model is the fact of not being able to assess the effect of the variables that have small variations within the group.
In fact, it serves repeated observations for each group and a reasonable amount of variation of the main variables within each group.
If it is crucial that it learns the effect of a variable that does not show much within-group variation, then it will have to forego fixed effects estimation. But this exposes it to potential omitted variable bias.
There is no easy solution to this dilemma.

2.2.1.1 Estimators of fixed effects model

2.2.1.1.1 LSDV estimator

It can write the FE model using N dummy variables indicating the individuals.

$$Y_{it} = \sum_{j=1}^{N} \alpha_j d_{ij} + x_{it}\beta + u_{it}, a_i, \qquad u_{it} \sim iid(0, \sigma_u^2 \tag{2.5}$$

with dummies d_j, where $d_{ij} = 1$ if i=j, and 0 else.
The parameters can be estimated by OLS. The implied estimator for β is called the LS dummy variable estimator, LSDV.
Instead of exploding computer storage by increasing the number of dummy variables for large N the within estimator is used.

2.2.1.1.2 Within estimator, FE

The FE estimator forβ is obtained, if it uses the deviations from the individual means as variables.
The model in individual means is the following:

$$\overline{y_i} = a_i + \overline{x_i}\beta + \overline{u_i} \tag{2.6}$$

Subtraction from $y_{it} = a_i + x_{it}\beta + u_{it}$ gives the results below:

$$y_{it} - \overline{y_i} = (x_{it} - \overline{x_i})'\beta + (u_{it} - \overline{u_i}) \tag{2.7}$$

where the intercepts vanish. Here the deviation of y_{it} from $\overline{y_i}$ is explained.
The estimator for β is called the within or FE estimator. Within refers to the variability among observations of individual ii.

2.2.1.1.3 First difference estimator, FD

An alternative way to eliminate the individual effects a_i is to take the first differences, with respect of time, of the FE model:

$$y_{it} - \overline{y_{i,t-1}} = (x_{it} - \overline{x_{i.t-1}})'\beta + (u_{it} - \overline{u_{i,t-1}}) \tag{2.8}$$

FD allows correlation between x_{it} and $u_{i,t-2}$.
The FD estimator is slightly less efficient than the FE. FS looses one time dimensions for each i. FE looses one degree of freedom for each i by using $\overline{y_i}, \overline{x_i}$.

2.2.2 Random effects model

A random effect(s) model, also called a variance components model, is a kind of hierarchical linear model.
It assumes that the dataset being analysed consists of a hierarchy of different populations whose differences relate to that hierarchy.
In econometrics, random effects models are used in the analysis of hierarchical or panel data when one assumes no fixed effects.
The random effects model is a special case of the fixed effects model.

Such models assist in controlling for unobserved heterogeneity when this heterogeneity is constant over time and correlated with independent variables.
This constant can be removed from the data through differencing, for example by taking a first difference which will remove any time invariant components of the model.

There are two common assumptions made about the individual specific effect, the random effects assumption and the fixed effects assumption.
The random effects assumption, made in a random effects model, is that the individual specific effects are uncorrelated with the independent variables.
The fixed effect assumption is that the individual specific effect is correlated with the independent variables.
If the random effects assumption holds, the random effects model is more efficient than the fixed effects model.
However, if this assumption does not hold, the random effects model is not consistent.

In a random effect model the unobserved variable are assumed to be uncorrelated with all the observed variables. That assumption will often be wrong but an RE model may still be desirable under some circumstances.
The model can be estimated by Generalized Least Squares (GLS) which is in general more efficient than OLS.

The formula below represents a random variable, which has the same variance like the others:

$$a_i \sim (0, \sigma_a^2)$$

$$y_{it} = \beta_0 + x'_{it}\beta + a_i + u_{it}, \qquad u_{it} \sim (0, \sigma_a^2) \tag{2.9}$$

The value a_i is specific for individual i. The a's of different individuals are independent, have a mean of zero, and their distribution is assumed to be not too far away from normality. The overall mean is captured in β_0.

As long as $E(x_{it}\epsilon_{it}) = E(x_{it}(a_i + uit)) = 0$ the explanatory variables are exogenous, the estimates are consistent.
In general, random effects are efficient, and should be used if the assumptions underlying them are believed to be satisfied.
For random effects to work in the school example it is necessary that the school-specific effects be uncorrelated to the other covariates of the model.
This can be tested by running fixed effects, then random effects, and doing a Hausman specification test. If the test rejects, then random effects is biased and fixed effects is the correct estimation procedure.

2.2.2.1 Estimation of random effects model, GLS

$$y_{it} = \beta_0 + x'_{it}\beta + a_i + u_{it} \tag{2.10}$$

$$u_{it} \sim (0, \sigma_a^2), \qquad a_i \sim (0, \sigma_a^2)$$

where $(a_i + u_{it})$ is an error of 2 components:

- an individual specific component, which does not vary over time
- a remainder, which is uncorrelated with respect to i and t
- a_i and u_{it} are mutually independent, and independent of all x_{js}

As simple OLS does not take this special error structure into account, GLS is used.
GLS is unbiased, if the x's are independent of all a_i and u_{it}.
The GLS will be more efficient than OLS in general under RE assumptions.

2.3 Dynamic panel data models

Panel data is now widely used to estimate dynamic econometric models. Its advantage over cross-section data in this context is obvious: it cannot estimate dynamic models from observations at a single point in time, and it is rare for single cross-section surveys to provide sufficient information about earlier time periods for dynamic relationship to be investigated.

Its advantage over aggregate time series data include the possibility that underlying microeconomic dynamics may be obscured by aggregation biases, and the scope that panel data offers to investigate heterogeneity in adjustment dynamics between different types of individuals.

Genuine panel data will typically allow more of the variation in the micro data to be used in constructing parameter estimates, as well as permitting the use of relatively simple econometric techniques.
Many economic issues are dynamic by nature and use the panel data structure to understand adjustment. Examples: demand, dynamic wage equation, employment models, investment of firms, etc.

The dynamic model with one lagged dependent without exogenous variables, $|\gamma| < 1$, is explaned in the formula below:

$$y_{it} = \gamma y_{i,t-1} + a_i + uit, \qquad u_{it} \sim (0, \sigma_a^2) \tag{2.11}$$

In (2.11), $\gamma y_{i,t-1}$ depends positevely on a_i.
There is an endogeneity problem. OLS and GLS will be inconsistent for $N \to \infty$ and T fixed, both FE and RE.
The finite sample bias can be substantial for small T. If in addition $T \to \infty$, it obtains a consistent estimator, but T is small for panel data.

2.3.1 Estimators for dynamic panel data

2.3.1.1 The first difference estimator (FD)

Using the first difference estimator (FD), which eliminates the a_i's

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + (u_{it} - u_{i,t-1}) \tag{2.12}$$

is no help, since $y_{i,t-1}$ and $u_{i,t-1}$ are correlated even when $T \to \infty$.
It stays with the FD model, as the exogeneity requirements are less restrictive, and it uses an IV estimator.

2.3.1.2 IV estimator, Anderson-Hsiao

Instrumental variable estimators, IV, have been proposed by Anderson-Hsiao, as they are consistent with $N \to \infty$ and finite T.
Choice of the instruments for $(y_{i,t-1} - y_{i,t-2})$:

- instrument $y_{i,t-2}$ as proxy is correlated with $(y_{i,t-1} - y_{i,t-2})$, but not with $u_{i,t-1}$ or u_{it}, and so $(u_{it} - u_{i,t-1})$
- instrument $(y_{i,t-2}$ - $y_{i,t-3})$ as proxy for $(y_{i,t-1}$ - $y_{i,t-2})$ sacrifies one more sample period

2.3.1.3 GMM estimation, Arellano-Bond

The Arellano Bond (also Arellano-Bover) method of moments estimator is consistent.
The moment conditions use the properties of the instruments

$$y_{i,t-j}, \qquad j \leq 2 \tag{2.13}$$

to be uncorrelated with the future errors u_{it} and $u_{i,t-1}$. It obtains an increasing number of moment conditions for *t = 3, 4, ..., T.*

$$t = 3 : E[(u_{i,3} - u_{i,2})y_{i,1}] = 0$$

$$t = 4 : E[(u_{i,4} - u_{i,3})y_{i,2}] = 0, \qquad E[(u_{i,4} - u_{i,3})y_{i,1}] = 0$$

$$t = 5 : E[(u_{i,5} - u_{i,4})y_{i,3}] = 0, \qquad E[(u_{i,5} - u_{i,4})y_{i,1}] = 0$$

2.3.1.3.1 GMM estimation, Arellano-Bond, without x's

Ignoring exogenous variables, for $\delta y_{it} = \gamma \delta y_{i,t-1} + \delta u_{it}$

$$E[Z_i' \delta u_i] = E[Z_i'(\delta y_i - \gamma \delta y_{i,t-1})] = 0 \tag{2.14}$$

The number of moment conditions are *1 + 2 +.....+ (T-2).*
The optimal matrix yielding as any efficient estimator is the inverse of the covariance matrix of the sample moments.
The matrix can be estimated directly from the data after a first consistent estimation step.
Under weak regularity conditions the GMM estimator is any normal for $N \to \infty$ for fixed *T, T > 2* using our instruments. It is also consistent for $N \to \infty$ and $T \to \infty$, though the number of moment conditions $\to \infty$ as $T \to \infty$.
It is advisable to limit the number of moment conditions.

2.3.1.3.2 GMM estimation, Arellano-Bond, with x's

The dynamic panel data model with exogenous variables is:

$$y_{it} = x_{it}\beta + \gamma y_{i,t-1} + a_i + u_{it}, \qquad u_{it} \sim iid(0, \sigma_u^2) \tag{2.15}$$

As also exogenenous x's are included in the model additional moment conditions can be formulated:

- for strictly exogenous variables, E$[x_{is}u_{it}] = 0$ for all s, t
  E$[x_{is}\delta u_{it}] = 0$
- for predetermined variables, E$[x_{is}u_{it}] = 0$ for $s \leq t$
  E$[x_{i,t-j}\delta u_{it}] = 0$, j = 1, ..., t-1

So there are a lot of possible moment restrictions both for differences as well as for levels, and so variety of GMM estimators.
GMM estimation may be combined with both FE and RE.
Here also, the RE estimator is identical to the FE estimator with $T \to \infty$.

2.4 Panel data from time series of cross-sections

In many countries, there are few or no panel data, but there is a series of independent cross-sections.
Several models that seemingly require the availability of panel data can also be identified with repeated cross-sections under appropriate conditions.
This concerns models with individual dynamics and model with fixed individual-specific effects.

The major limitation of repeated cross-sectional data is that the same individual are not followed over time, so that individual histories are not available for inclusion in a model. On the other hand, repeated cross-sections suffer much less from typical panel data problems like attrition and nonresponse.

In a seminal paper, Deaton [2] (1985)[1] suggests the use of cohorts to estimate a fixed effects model from repeated cross-sections.
A cohort is a group of individuals compounds in a fixed manner and can be identified by the investigation.
The analyzes generate a sequence of random samples composed of people from these cohorts, which must be large enough. From these primary results

[1] Angus Deaton, Panel data from time series of cross-sections

it can get samples that generate a time series as they were panel data and, from these data, it is possible to infer the behavior of cohorts in general.
The starting point of the analysis of Deaton considers linear economic relations that may contain the individual fixed effects, as also not contain them. He argues that the same form of relationships existing in the population is found in the population divided in cohorts.
In addition, if the population have additive individual fixed effects, these same fixed effects will be found for the cohort population.

It is possible to use errors-in-variable estimators to consistently estimate the population relationships.
This methodology was designed in response to a lack of panel data to make an econometric analysis complete. Not for this reason it can be considered a method that provides inferior results; in fact, every year it selects new samples and the representativeness of the population remains and is maintained from year to year.
An important feature is the lack of friction problem in the population cohorts, while in panel data this characteristic represents a huge problem.

Deaton does not support the theory according to which the panel data are free of errors in the variables. In fact, for him the difference between cohorts and panel data model is low.
The technique of cohorts has an important advantage, to recognize the error of measurement from the start and to be able to control it well.

Another scholar, Ashenfelter (1983) has shown in the analysis made by him on the elasticity of labor supply that the measurement errors have a very important role and persistent over time, so it is difficult to come to the truthful conclusions as they are affected by a large error.

Deaton did this first study mainly on linear models, and other scholars have tried to extend his theory even on more complex models, such as non-linear models and dynamic ones. Moffitt (1993) and Collado (1997) have proposed an extension of the model of Deaton. They studied the dynamic linear model.
The starting point of the two versions is the same: both claim that $N \to \infty$.
What changes is the idea of what happens to the cohorts as N increases.
Deaton said that if increases N then increases the number of cohorts, but the size of the cohorts remains constant value.
Instead, Moffitt and Collado claim that increases if N then increases the size of the cohorts, not their numbers.
This new approach can solve the problem of errors in variables which in Deaton's model was present.
The fixed effects estimator based on the pseudo panel of cohort averages may provide an attractive choice, even when a lagged dependent variable is

included in the model (Verbeek and Vella, 2005).
Moffit and Collado did a Monte Carlo experiments and showed that the bias that is present in the within estimator for the dynamic model using genuine panel data is much larger than what is found for similar estimators employed upon cohort aggregates.
Economic theory is very common to find to analyze models containing fixed individual effects correlated with the explanatory variables.
On the other hand, however, often they lack the genuine panel data and, therefore, the theory proposed by Deaton has a significant role in estimating dynamic models and fixed effects models in the absence of the original panel data.
These models are then solved by repeated cross-sections, but they also require the identification conditions quite strong and difficult to estimate.

The technique of cohorts resembles the technique of instrumental variables, in which instruments are taken as indicators of the group.
An important issue in both the static and dynamic models is the validity and relevance of the instruments that are used to construct the cohorts.
The instruments must be valid and relevant to the analyzed model to get an instrumental variables estimator consistent.
It may happen that even in the presence of valid instruments and theoretically relevant to the model, the latter suffer the problem of weak instruments and, therefore, obtain the estimators with low yields (Bound, Jaeger and Baker, 1995).

A necessary condition for consistency of most estimators is that all exogenous variables exhibit genuine time-varying cohort-specific variation.
The cohorts have the exogenous variables that change differentially over time.
This requirement will be satisfied in empirical applications because estimation error in the reduced form parameters may hide collinearity problems.

Chapter 3

Monte Carlo method

3.1 Definition

It is an econometric technique based on repeated random sampling and, for the fact of relying on expressing random, it can not predict the results that may be obtained from the simulation.
The results from the experiment data, later, are subjected to a statistical analysis to derive the conclusions.

Monte Carlo methods are mainly used in three distinct problem classes: optimization, numerical integration and generating draws from a probability distribution.
It uses mathematical models and mathematical expressions to describe different studies that can be done with this simulation in different fields of study, such as the natural sciences and engineering.
First, it has to define the inputs of the model, which, through the use of various mathematical formulas, leads to obtaining one or more outputs.
The input parameters for the models depend on various external factors. Realistic models are subject to risk from the systematic variation of the input parameters. An effective model should take into consideration the risks associated with various input parameters.
Monte Carlo simulation support the researcher in the analysis of the risks associated with the input, to be able to select the most consistent variables for the model studied.

The order in which the simulation is performed is the following:

1. define a domain of possible inputs
2. generate inputs randomly from a probability distribution over the domain

3. perform a deterministic computation on the inputs

4. aggregate the results

In the first step, it identifies a statistical distribution which will be selected from all the inputs.
Later, in the second step, the inputs are obtained in a random way from the distribution.
In the third step, it analyzes the outputs that have been obtained to be able to reach some conclusions on the initial hypothesis of the study.
In the last part it puts together the results obtained to compare them with other results from different random samples to being able to draw final conclusions.

Monte Carlo simulation does not always require truly random numbers to run, in fact, for many techniques, it can use pseudo-random sequences that make it easier to test and run the simulation.
The main feature is the unpredictability, that is, the results is not known in advance.
Instead, the characteristic necessary to make a enough good simulation is having a pseudo-random sequence that is seems random.
Testing that the numbers are uniformly distributed or follow another desired distribution when a large enough number of elements of the sequence are considered is one of the simplest, and the most common ones. Weak correlations between successive samples is also often desirable.
Pseudo-random number sampling algorithms are used to transform uniformly distributed pseudo-random numbers into numbers that are distributed according to a given probability distribution.
To summarize, the characteristics of a good Monte Carlo simulation are the following:

- the pseudo random number generator has a certain characteristics
- the pseudo- random number generator produces values that pass tests for randomness
- there are enough samples to ensure accurate results
- the proper sampling technique is used
- the algorithm used is valid for what is being modeled
- it simulates the phenomenon in question

A variant of the simulation itself is the quasi - Monte Carlo methods in which random sampling are used for low discrepancy sequences as this ensures a faster convergence than Monte Carlo simulation using random or pseudo-random sequences.
The principle disadvantages of this method are two: first, it might be difficult to evaluate the best and worst case scenarios for each input variable; second, decision making tends to be difficult as well, since it is considering more than one scenario.

3.2 History

The Monte Carlo method has a long history. In statistics it was called model sampling and used to verify the properties of estimates by mimicking the settings for which they were designed.
W.S. Gosset (1908) did some simulations on finger measurement from 3000 criminals and derived some analytical results, that it is what is now called the t-student distributions.
Sampling was also used by physicists. Hammersley and Handscomb (1964) describe some computations done by Kelvin (1901) on the Boltzmann equation.
Primitive idea of random sampling is found in method of Buffon who made an experiment observing a random sequence of needles on a floor and counting the fraction of needles That touch the line between two planks.
One of the most famous early uses of MC simulation was by Enrico Fermi in 1930, when he used a random method to calculate the properties of the newly-discovered neutron.
The Monte Carlo method acquired this name from the famous Casino of Monte Carlo and became an important method in physics for the study of atomic weapons, in the 1940s - 1950s.
Many of the problems studied had a deterministic origin. By now it is standard to use random sampling on problems stated deterministically but early on that this was a major innovation, and was even considered to be part of the definition of a Monte Carlo method.

From the 1950s, there are many papers and many studies were done using the Monte Carlo method to tackle different problems in different fields of study.
Metropolis (1953) showed for the first Markov chain Monte Carlo method to study the relative position of the atoms and calls it the Metropolis algorithm.
Tocher and Owen (1960) describe the GSP software for discrete event simulation of queues and industrial processes. In 1977, Boyle used the Monte

Carlo simulation to solve problems related to financial options and the choice of portfolio.
Gillespie, in the same year, used this method to study completely different topic. He studied the chemical reactions in which the number of molecules is too small to which not even the differential equations are able to analyze them accurately.
Efron's (1979) bootstrap uses Monte Carlo sampling to give statistical answers with few distributional assumptions. Kirkpatrick (1983) utilized Monte Carlo method for optimizing very nonsmooth functions.
Indeed Kajiya (1988) used it for draw a path tracing for graphical rendering.
Tanner and Wong (1987) use Monte Carlo algorithms to cope with problems of missing data.
German and Gelfand (1990) and others researchers used the Monte Carlo method to solve the Bayesian statistical problems.
There are many other studies in which this method has found application and has enabled a solution to many problems before unsolved.

Even the quasi-Monte Carlo method has its own history and was created at about the same time as the real Monte Carlo simulation. In fact, the term was invented by Richtmyer in 1952, which considered proposing a Monte Carlo simulation in which the sequence is more uniform than truly random.

3.3 Methodology

The following steps are typically performed for Monte Carlo simulation:

1. *Static model generation*
   The starting point of the simulation is to develop a model that is as close as possible to a real scenario. Then, the use of mathematical formulas allows starting from the values of the inputs, process them and obtain the outputs.

2. *Input distribution identification*
   In this step it is important to identify the distribution that governs all inputs and, to do this, there are several statistical procedures. The input variables have a risk component that is added to the model, and is important to lower the risk as much as possible.

3. *Random variable generation*
   After identifying the distribution in the second step, now it can get a set of random numbers. One set of random numbers, consisting of one value for each of the input variables, will be used in the deterministic

model, to provide one set of output values. It then repeats this process by generating more sets of random numbers, one for each input distribution, and collect different sets of possible output values. This is the core of Monte Carlo simulation.

4. *Analysis and decision making*
   The simulation provides many outputs, but they are not definitive; in fact, they must be subjected to a statistical analysis. Finally, it is possible to draw final conclusions and have a statistical confidence of the results obtained.

3.3.1 Identification of input distribution

First, it is necessary to discuss the procedure for identifying the input distributions for the simulation model, often called distribution fitting.
The probability distribution determines the outcomes of random variables and, also, the probability with which these can occur. In fact, if the random variables obtained discrete values, then the distribution that governs them is called discrete probability distributions.

Fitting routines provide a way to identify the most suitable probability distribution for a given set of data.
This method uses historical data on particular input parameters and, with mathematical methods, sit the data to discrete or continuous distribution.
The probability distribution is identified by the input parameters, which generate the input data.
For the technique of fitting data to distributions there are several methods by which it can run it.

3.3.1.1 Methods for distribution fitting

3.3.1.1.1 Method of maximum likelihood (ML)

ML estimation (MLE) is a popular statistical method used to make inference about parameters of the underlying probability distribution from a given data set.
If the data drawn from a particular distribution are independent and identically distributed (iid), then this method can be used to find out the parameters of the distribution from which the data are most likely to arise.
Let θ be the parameter vector for f, which can be either a probability mass function for discrete distribution or a probability density function for continuous distributions.

Let the sample drawn from the distribution be $x_1, x_2, \ldots, x_n$ then the likelihood of getting the sample from the distribution is given by

$$L(\theta) = f_\theta(x_1, x_2, ..., x_n|\theta) \tag{3.1}$$

given the parameters of this distribution formula, it can be defined as the probability density function of the date.
In MLE, it tries to find the value of the θ so that the value of L(θ) can be maximized. To achieve this it must consider the log of the function and this is called loglikelihood.
The MLE method can be seen as unconstrained nonlinear optimization problem.
It shall be represented in the following formula:

$$MaxLL(\theta) = \sum_{i=1}^{n} lnf_\theta(x_i|\theta), \qquad \theta \in \Theta \tag{3.2}$$

For some distribution, this optimization problem can be theoretically solved by using differential equations.
MLE method is by far the most used method for estimating the unknown parameters of a distribution.
The most important features of this method are two:

1. The bias in the MLE tends to infinity as the number of samples, then it can define asymptotically unbiased.

2. The MLE has the lowest mean squared error between the unbiased estimators, so it is a method asymptotically efficient.

3.3.1.1.2 Method of moments (ME)

In the method of moments trying to equate sample moments with unobservable population moments, to get some of the estimates of population characteristics such as mean, variance, and the median.
In some cases, ME estimators can be calculated very easily and quickly on the difference of the likelihood equations that are very complex.
ME estimators can be considered as a first approximation to the problem of study, which, then, will be analyzed by the method MLE.
These are two complementary methods. Between the two, MLE method provides a better estimate of distribution parameters, as it has a greater chance to get close to the quantities to be estimated.

3.3.1.1.3 Nonlinear optimization

Another method to estimate the unknown parameters of the distributions is the non-linear optimization.
Different objective functions can be used for this purpose, such as: minimizing one of the goodness-of-fit- statistics, minimizing the sum-squared difference from sample moments or minimizing sum-squared difference from the sample percentiles.
The value of the parameter depends on the algorithm chosen to solve the nonlinear optimization problem.
This method is typically less efficient and often takes more time.

3.3.2 Random variable generation

The second step in the simulation, after identifying the distribution that governs the input parameters, is to generate random numbers. They represent the specific values of the variables.
The most common methods to generate random numbers from discrete and continuous distributions are two: inverse transformation method and bootstrapped Monte Carlo.

3.3.2.1 Generating RV's from a distribution function

3.3.2.1.1 Inverse transformation method

The inverse transformation method provide the most direct route for generating a random sample from a distribution.
This method is described by a mathematical process. It is used in the reverse of probability density function (PDF) for continuous distributions or reverse of probability mass function (PMF) for discrete distribution.
Then, the random numbers between 0 and 1 are converted to random value for the input distribution.
Let X is a continuous random variable that follows the PDF function, defined by f . F denotes the cumulative probability distribution function for the variable X and is continuous and strictly increasing (0, 1). F^{-1} denotes the inverse of the function F.
The two steps from which to get a random number X from the PDF function, f are defined as follow:

- Generate U $\sim U(0,1)$
- Return X = F^{-1}(U)

Since $0 \leq U \leq 1$, F^{-1} (U) always exists.
The inverse transformation method can also be used when X is discrete. For discrete distribution, if $p(x_i)$ is the probability mass function, the cumulative PMF is given by:

$$F(x) = P(X \leq x) = \sum_{x_i<x} p(x_i) \tag{3.3}$$

This method can be applied to different types of functions, not just the one that is continuous or discrete.
Indeed, it may also apply to functions that are formed from a mixture of the two functions above explained.

A major advantage is the fact it can use this method to generate random numbers also from truncated distributions, so it gets the cumulative PMF function with discrete jumps.

One disadvantage, however, is to not be able to implement this method in case of miss in closed-form of inverse CDF for a distribution.
This disadvantage can be overcome by adopting another method proposed by Devroye, in 1986, which proposes an iterative method with which to evaluate the function in the absence of a closed-form.

The method of reverse transformation is the method most used to generate random numbers, but it is not the only one known. In fact, there are other important methods, such as composition method, convolution method and acceptance-rejection method (Law and Kelton, 1995).

3.3.2.2 Generating RV's from a Data set: Bootstrapped Monte Carlo

If it is studying particular distributions, as non-convex or multimodal, or where the problem is represented by scarcity of data, then it can not get an underlying distribution for an input variable.
In this case it has only a few historical value for the input variable.
The method described above can not be fine, in fact, the method bootstrapped Monte Carlo simulation is the right one to be used in these cases to generate random numbers.
Bootstrapped simulation can be a highly effective tool in the absence of a parametric distribution for a set of data.
The implementation of the method consists in repeatedly sample the original dataset to choose one of the data points from the set.
For bootstrapped MC simulation, one has to still use an uniform RNG, specifically an RNG to generate integer random numbers among the indices of an array, which is being used for storing the original datasets.

For many datasets, this method provides good result for simulation purposes. However, it does not provide general finite sample guarantees and has a tendency to be overly optimistic.
A criticism of this method is its simplicity, as many method using assumptions made formally, however in the bootstrapped method assumptions are formed when undertaking the bootstrap analysis.

The results obtained by this method can run into a problem of endogeneity, because there is a correlation in repeated observations and, because of this, false statistical inference could be drawn.

3.3.3 Monte Carlo simulation output analysis

The results of Monte Carlo simulation, to be meaningful, must be subjected to statistical analysis.
There are several model formula to be applied for each set of random numbers generated for each of the random variable to reach a final result for the output variables.
Aggregating the output values into groups by size and displaying the values as a frequency histogram provides the approximate shape of the probability density function of an output variable.

The output values can be used in different ways: they can be aggregated into empirical distribution or can be fitted to a probability distribution, which, then it can calculate the theoretical statistics.
To increase the accuracy of the output is important to compute many simulations, because the higher the number, the better then the approximations of distributional shape and the expected value of the variable.

3.3.3.1 Formulas for basic statistical analysis

The main formulas used to make a basic statistical analysis of the output values are listed below.
They are used to draw the final conclusions and infer the characteristics of the real population, starting from simulated samples. For this they are formulas which belong to sample statistics.
Let assume that it has N values for each of the output parameters, each value represented as $x_i, i = N$.

The most important and meaningful formulas to interpret the output variable are the following:

$$Mean(\overline{x}) \qquad \overline{x} = \frac{1}{n}\sum_i x_i$$

$$Standard\ deviation(s) \qquad s = \sqrt{\frac{1}{N-1}\sum(x_i - \overline{x})^2}$$

$$Variance(s^2) \qquad s^2 = \frac{1}{N-1}\sum(x_i - \overline{x})^2$$

$$Skewness \qquad Sk = \sum \frac{(x_i - \overline{x})^3}{(N-1)s^3}$$

$$Kurtosis \qquad Ku = \sum \frac{(x_i - \overline{x})^4}{(N-1)s^4} - 3$$

$$Coefficient of variability \qquad CF = \frac{s}{\overline{x}}$$

$$Minimum(x_{min}) \qquad x_{min} = min_i x_i$$

$$Maximum(x_{max}) \qquad x_{max} = max_i x_i$$

$$Range\ Width \qquad RW = x_{max} - x_{min}$$

$$Mean\ standard\ error \qquad MSE = \frac{s}{\sqrt{n}}$$

3.4 Application areas for Monte Carlo simulation

3.4.1 Monte Carlo simulation in mathematics and statistical physics

Perhaps the most important use of the Monte Carlo method is in mathematics and physics. It has found a solution for complex multi-dimensional partial differentiation and integration problems.
In the context of solving integration problems, MC method is used for simulating quantum systems, which allows a direct representation of many-body effects in the quantum domain.

Monte Carlo methods are mainly used in three problem classes, which are integration, optimization and inverse problems. All these techniques are explained below.

3.4.1.1 Monte Carlo integration

In mathematics, Monte Carlo integration is a technique for numerical integration using random numbers.
It is a particular Monte Carlo method that numerically computes a definite integral and is particularly useful for higher-dimensional integrals.
The special feature of this method that differentiates it from the others is the fact that the points at which the integrand is evaluated are chosen in a random way; indeed, in other methods, the integrand are evaluated at a regular grid.

There are different methods to perform a Monte Carlo integration, such as uniform sampling, stratified sampling, importance sampling, sequential Monte Carlo and mean field particle methods.
Monte Carlo integration employs a non-deterministic approach: each realization provides a different outcome.
The final outcome is given by an approximation of the correct value which is within the errors bars obtained by outcome.
The problem Monte Carlo integration addresses is the computation of a multidimensional definite integral, explains in the following formula:

$$I = \int_{\Omega} f(\overline{x}) d\overline{x} \tag{3.4}$$

in (3.4) Ω, a subset of R^m has volume expressed by the formula:

$$V = \int_{\Omega} d\overline{x} \tag{3.5}$$

The naive Monte Carlo approach is to sample points uniformly on Ω: given N uniform samples, I can be approximated by the following formula:

$$I \approx Q_N \equiv V \frac{1}{N} \sum_{i=1}^{N} f(\overline{x_i}) = V(f) \tag{3.6}$$

This is because the law of large numbers ensures that:

$$lim_{N \to \infty} Q_N = I \tag{3.7}$$

Given the estimation of I from Q_N, the error bars of Q_N can be estimated by the sample variance using the unbiased estimate of the variance:

$$Var(f) \equiv \sigma_N^2 = \frac{1}{N-1} \sum_{i=1}^{N} (f(\overline{x_i}) - (f))^2 \tag{3.8}$$

which leads to

$$Var(Q_N) = \frac{V^2}{N^2} \sum_{i=1}^{N} Var(f) = V^2 \frac{Var(f)}{N} = V^2 \frac{\sigma_N^2}{N} \tag{3.9}$$

As long as the sequence $(\sigma_1^2, \sigma_2^2, \sigma_3^2, ...)$ is bounded, this variance decreases asymptotically to zero as 1/N.
The estimation of the error of Q_N is thus given by:

$$\delta Q_N \approx \sqrt{Var(Q_N} = V \frac{\sigma_N}{\sqrt{N}} \tag{3.10}$$

which decreases as $\frac{1}{\sqrt{N}}$.
The biggest advantage of this technique is the fact that the result does not depend on the numbers of dimensions of the integral, while, in many deterministic methods, the results depends on the dimension.
A feature common to the MC method and deterministic methods is the fact that the estimate of the error is not a strict error bound.
The random sampling, therefore, can lead to an underestimate of the error since it may not include all important features.

The literature has discussed very much the cases in which Monte Carlo simulation is used to improve the error estimates.
The two techniques that best address this problem are the stratified sampling, where the regions are divided into sub-domains, and importance sampling, where sampling is made from non-uniform distributions.
In the next two paragraphs these two techniques are best explained.

3.4.1.1.1 Recursive stratified sampling

Recursive stratified sampling is used for the analysis of a multidimensional integrals.
In fact, it is divided into several recursion steps and, for each step, Monte Carlo algorithm is used for calculating the integral and error.

If the error estimate is larger than the required accuracy the integration volume is divided into sub-volumes and the procedure is recursively applied to sub-volumes.
In multidimensional integral the numbers of sub-volumes grow too quickly, so it is not possible to use the simple strategy called dividing by two.

The stratified sampling algorithm wants to be able to reduce the error and, therefore, get more effective sampling.
It can obtain this by focusing on regions where the variance of the function is largest, so to reduce the wide variance proceed dividing the volume and subdivision should bring most dividends.

3.4.1.2 Importance sampling

Importance sampling is another technique that is used to reduce the variance.
The starting point is to emphasize the values of the input random variables that have more impact on the parameters.
Then, it samples more these more important values and, by doing so, it gets to reduce the variance of the estimator.

The fundamental issue in implementing importance sampling simulation is the choice of the biased distribution which encourages the important values of the input variables.
Choosing a good biased distribution is the core part of importance sampling. It must pay attention to the choice of the distribution, because if it chooses a biased distributions then it will get a biased estimator by simulation.
However, the simulation outputs are weighted to correct for the use of the biased distribution, and this ensures that the new importance sampling estimator is unbiased. These weights are given by the likelihood ratio, such as derivative of the underlying distribution with respect to the biased distribution (Radon and Nikodym).
Consider X to be the sample and $\frac{f(X)}{g(X)}$ to be the likelihood ratio, where f is the probability density (mass) function of the desired distribution and g is the probability density (mass) function of the biased/proposal/sample distribution.
Then the problem can be characterize by choosing the sample distribution g that minimizes the variance of the scaled sample:

$$g^* = min_g Var_g \left(X \frac{f(X)}{g(X)} \right) \tag{3.11}$$

It can be shown that the following distribution minimizes the above variance:

$$g^*(X) = \frac{|X| f(X)}{f|x| f(x) dx} \tag{3.12}$$

In the (3.22),It is easy to see that when $X \geq 0$, this variance becomes 0.

3.4.1.3 Optimization

A very common method used for random numbers in Monte Carlo simulation is optimization.
It is used to minimize or maximize functions that have a large numbers of dimensions in their respective vectors.

There are many problems which could be resolved with this technique: minimize the number of moves in a computer chess program, solve particle problems by exploring large configuration space, minimize distance travel in the traveling salesman problem, etc.

This last problem is maybe the most famous problem of optimization and is defined like a conventional optimization problem.
Assuming to know all the distances between each destination point, the purpose of the issue is to find the possible travel choice with the lowest total distance.
However, let's assume that instead of wanting to minimize the total distance traveled to visit each desired destination, it wanted to minimize the total time needed to reach each destination.
This assumption is not acceptable because travel time is uncertain (traffic, time of day, etc. . .), so it is necessary to proceed with simulation-optimization to achieve the solution to this problem.
First, through the use of a probability distribution, it has to understand time it could spend to go from one point to another.
Then, taking account the uncertainty linked to time travel, it optimizes its travel decisions to identify the best path to follow.

3.4.1.4 Inverse problems

The definition of inverse problem is supported by the definition of probability distribution.
The last model obtained by measuring the observable parameters and combines the information submitted previously obtained with those just obtained.

In the general case, the theory that links the data to the model parameters, defined posterior probability, is nonlinear, it may by multimodal or some moments may not be defined.
Accepting a maximum likelihood model like final outcome is not possible, because in inverse problems it wish to have information on the resolution power of the data.

Due to a large number of model parameters is difficult to apply an inspection of the marginal probability densities.
This can be accomplished using an efficient Monte Carlo method, even in cases when there is not available an explicit formula for the distribution.
So, the solution to this problem is given by the possibility to generate in a pseudo-randomly way a large collection of models according to functional the posterior probability distribution and, then, to display the models so that the properties of the model likelihood conveyed to the researchers.

Based on the importance sampling method it can get to do an analysis of inverse problems with complex information and data obtained in advance and distributed according to an arbitrary noise distribution. This is possible starting by the algorithm of Metropolis and generalize it for analysis possibly highly nonlinear.

3.4.2 Monte Carlo simulation in finance

Financial analysts use Monte Carlo simulation quite often to model various scenarios.

3.4.2.1 Real options analysis

In the analysis of real options, Monte Carlo simulation is important for the calculation of net present value (NPV) of projects.
The input variables , that are characterized by uncertainty, are used in stochastic model to run Monte Carlo.
Then, from the analysis of the outputs, it is possible to deduce the average NPV of the project, its volatility and other sensitivities.

3.4.2.2 Portfolio Analysis

It is possible to used Monte Carlo simulation also in problems of portfolio evaluation.
For each simulation, it obtained the value of the instruments components the portfolio and the value of the portfolio.
Then, after collecting the values of many simulations, it is possible to combined them in a histogram and deduce the characteristics of the portfolio from this representation.

3.4.2.3 Option analysis

Monte Carlo simulation can be used for analyzing the prices of different types of option.
From a simulation, it obtained various price path for the underlying share for options on equity.
Then, these paths are subjected to statistical analysis to deduce final conclusions.

This simulation can be used, also, for studying some characteristics of bonds and bond options.
In this case, Monte Carlo simulation is useful to analysis the uncertainty of the annual interest rate.

3.4.2.4 Personal financial planning

MC methods are used for personal financial planning, simulating the overall market to find the probability of attaining a particular target balance for the retirement savings account.

3.4.3 Monte Carlo simulation in reliability analysis and six sigma

Reliability analysis is useful for evaluating cycle costs, cost-effectiveness of the products and many others problems.
In the reliability analysis, the starting point is the evaluating of failure distribution.
Then, random numbers are generated for these distribution.
Finally, the output results are statistically analyzed to calculate the probability of these failure events.

Six sigma, indeed, is a business management strategy, which has the goal to identify and remove the causes of defects and errors in business processes. Six-sigma principles can be applied to various industries, including manufacturing, financial and software.
Monte Carlo simulation is used to analyzing many problems in this areas: to identify optimal strategy in selecting projects, providing probabilistic estimates project cost benefits, creating virtual testing grounds in later phases for proposed process and product changes, predicting quality of business processes, identifying defect-producing process steps driving unwanted variation.

3.4.4 Monte Carlo simulation in engineering

Monte Carlo methods are used in many engineering studies because it could solve the interactive, co-linear and non-linear behavior of the processes. More precisely, it is widely used for sensitivity analysis and quantitative probabilistic analysis in process design.

Below, it represents the main use of Monte Carlo in different engineering areas:

- In microelectronics engineering, Monte Carlo methods are applied to analyze correlated and uncorrelated variations in analog and digital integrated circuits.
- In geostatistics and geometallurgy, Monte Carlo methods underpin the design of mineral processing flowsheets and contribute to quantitative risk analysis.

- In wind energy yield analysis, the predicted energy output of a wind farm during its lifetime is calculated giving different levels of uncertainty.
- Impacts of pollution are simulated and diesel compared with petrol.
- In Fluid Dynamics where the Boltzmann equation is solved for finite Knudsen number fluid flows using the Direct Simulation Monte Carlo method in combination with highly efficient computational algorithms.
- In autonomous robotics, Monte Carlo localization can determine the position of a robot. It is often applied to stochastic filters such as the Kalman filter or Particle filter that forms the heart of the SLAM (Simultaneous Localization and Mapping) algorithm.
- In telecommunications, when planning a wireless network, design must be proved to work for a wide variety of scenarios that depend mainly on the number of users, their locations and the services they want to use. Monte Carlo methods are typically used to generate these users and their states. The network performance is then evaluated and, if results are not satisfactory, the network design goes through an optimization process.
- In reliability engineering, one can use Monte Carlo simulation to generate mean time between failures and mean time to repair for components.
- In signal processing and Bayesian inference, particle filters and sequential Monte Carlo techniques are a class of mean field particle methods for sampling and computing the posterior distribution of a signal process given some noisy and partial observations using interacting empirical measures.

3.4.5 Monte Carlo in Physical sciences

Monte Carlo methods are very important in computational physics, physical chemistry, and related applied fields, and have diverse applications from complicated quantum chromodynamics calculations to designing heat shields and aerodynamic forms as well as in modeling radiation transport for radiation dosimetry calculations.

Below, there are a list of the main areas of uses of this method for different physical sciences:

- In statistical physics Monte Carlo molecular modeling is an alternative to computational molecular dynamics, and Monte Carlo methods are used to compute statistical field theories of simple particle and polymer systems.
- Quantum Monte Carlo methods solve the many-body problem for quantum systems.
- In experimental particle physics, Monte Carlo methods are used for designing detectors, understanding their behavior and comparing experimental data to theory.
- In astrophysics, they are used in such diverse manners as to model both galaxy evolution and microwave radiation transmission through a rough planetary surface.
- Monte Carlo methods are also used in the ensemble models that form the basis of modern weather forecasting.

3.4.6 Monte Carlo in computational biology

Monte Carlo methods are used in various fields of computational biology and for studying biological systems such as genomes, proteins, or membranes. Computer simulations allow us to monitor the local environment of a particular molecule to see if some chemical reaction is happening for instance.

In this case, Monte Carlo is very important because it can conduct some physical experiments that it was not possible to execute before, such as breaking bonds, introducing impurities at specific sites, changing the local/global structure, or introducing external fields.

Monte Carlo methods are used in various fields of computational biology and for studying biological systems such as genomes, proteins, or membranes.

3.4.7 Monte Carlo in applied statistics

In applied statistics, Monte Carlo methods are generally used for solve two types of situations:

1. Real data often do not have classical distributions, so they could not be analyzed with simple statistics analysis. In this case, Monte Carlo simulation is a technique that under realistic data is able to compare competing statistics for small samples.

2. In presence of very efficient and impossible to compute tests such as permutation test, only Monte Carlo simulation is able to provide an implementation.

Monte Carlo methods can also be defined like a compromise between approximate randomization and permutation tests, where the first one is based on a specific subset of all permutations while, the second one, is based on a specific numbers of randomly drawn permutations.

Chapter 4

Monte Carlo for estimation of sampling errors

4.1 Theory of simple Monte Carlo

A problem, which is studied a lot in statistics, is to infer the characteristics of a population from analyzing samples.
Simple Monte Carlo tries to do exactly that, in fact, it aims to estimate a population expectation when it has sample expectation.

Below it is presented the accuracy of this method, so, it is used the laws of large numbers and the central limit theorem for derive confidence intervals of the sample mean from the sample data values.

4.1.1 Accuracy of simple Monte Carlo

First, it starts with a variable Y and, from its distribution, it is necessary to generate random and independent values defined as Y_1, ...,Y_n.
To obtain the average of these values is applied the following formula:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{4.1}$$

as the estimates of μ.
From this calculation it obtains the estimate of the mean, then it gets to know the expected value of the random variable Y, as it observed that from the following formula: $\mu = E(Y)$.
The random variable Y is defined in this way: *Y=f(X)*, where the random variable X has a probability density function *p(x)* or it can be considered as

a discrete random variable with a probability mass function called p and f is a function defined by real values.

For some problems it is easier to work with expectations while for other tasks it is simpler to work directly with the integrals.
Here it is possible to apply the simple Monte Carlo because the Y is defined by $Y = f\ (X)$ and it is a quantity expressed by a real number or by a vector.
The laws of large numbers is one of the theorem used to explain simple Monte Carlo.

The assumptions from which is then possible to construct a mathematical model are the following:

- Y is a random variable for which $\mu = E(Y)$ exists
- the generated values of Y, $Y_1, \ldots, Y_n$, are iid (independent and identically distributed) have the same distribution of Y

Then, the law of large numbers has two versions: weak law of large numbers and the strong law of large numbers.
The first one tells that the chance of missing by more than E goes to 0 and is represented by the formula below:

$$lim_{n\to\infty} P(|\mu_n - \mu| \leq \epsilon) = 1 \tag{4.2}$$

holds for any $\epsilon > 0$.
Instead, the strong law of large numbers, is more complex and tells more respect the first one.

$$P(lim_{n\to\infty} |\mu_n - \mu| = 0) = 1 \tag{4.3}$$

Both the laws of large numbers shows that Monte Carlo will be able to achieve its principle purpose, to get the error as small as possible.
None of the laws, however, indicates how large n must be to achieve this, and does not indicate whether the error is low in the given samples $Y_1, ..., Y_n$.

To improve the analysis just described, it must place the assumption that Y has a finite variance, for which $Var(Y) = \sigma^2 < \infty$
In IID sampling, μ_n is a random variable and it has its own mean and variance.
Monte Carlo is unbiased if the expected value of μ_n is equal to μ.
This is verify in the formula below:

$$E(\mu_n) = \frac{1}{n} \sum E(Y_i) = \mu \tag{4.4}$$

The variance is represented by the formula below:

$$E((\mu_n - \mu)^2) = \frac{\sigma^2}{n} \tag{4.5}$$

From it, it has to deduce that the result is better with increased sample size and worse with increased variance.
So, the error of the samples is lower when the samples become larger.
The root mean squared error (RMSE) is given from the formula below:

$$\sqrt{E((\mu_n - \mu)^2) = \frac{\sigma}{\sqrt{n}}} \tag{4.6}$$

A disadvantage of the Monte Carlo method is that it can not be used for problems that require a high precision.
This weakness in certain cases does not represent a big problem because they require only a rough estimate of μ in order to decide what action to take.
In other cases, however, it represents a bigger problem and for prevents this obstacle is necessary to put some idealized assumptions, such as specific distributional forms.

The principle advantage of Monte Carlo is that it is greater than closed form estimates because it can put more real world complexity into the computations and it is useful when closed forms are unavailable.
Simple Monte Carlo is most competitive in high dimensional non-uniform problems.

4.1.2 Error estimation

Monte Carlo method is very important in studies of error estimation. The sample values obtained from the simulation expressed a very good idea of the error $\mu_n - \mu$.

From the central limit theorem (CLT), it also know that the $\mu_n - \mu$ has approximately a normal distribution with mean 0 and variance $\frac{\sigma^2}{n}$.
The average squares error is given by the variance.
It is easy to estimate sigma from the sample values and the formula used for the estimation are the following two:

$$s^2 = \frac{1}{n}\sum(Y_i - \mu_n)^2 \tag{4.7}$$

$$\sigma^2 = \frac{1}{n}\sum(Y_i - \mu_n) \tag{4.8}$$

From this estimation, it is evident that μ_n has mean μ and variance $\frac{s}{\sqrt{n}}$.

A variance estimate s gives the order of the error which is $\frac{s}{\sqrt{n}}$.
IID central limit theorem (Chung, 1974): *"Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with mean and finite variance $\sigma^2 > 0$. Let $\mu_n = \frac{1}{n}\sum Y_i$. Then for all $z \ni R$*

$$P\left(\sqrt{n}\frac{\mu_n - \mu}{\sigma} \leq z\right) \to \Phi(z) \tag{4.9}$$

as $n \to \infty$".
This theorem is useful in Monte Carlo simulation because it can produce an approximate confidence interval that is almost always used.
It had the necessity to know the value of σ but, when it is not available, it can be substitute by the value of s, the estimate of σ.

4.1.3 Random sample size

Simple Monte Carlo is based on random sample size.
The initial assumptions for the simulation is that it samples $X_1, \ldots, X_n$ independently. Its interest is in X that satisfy the following condition: *E(f(X))*.
Another assumption for the model is that it focus only on those X that satisfy the condition $X_i \in A$ for some set A.
This is represented in the system below:

$$N_A = \sum A_i \tag{4.10}$$

where

$$A_i = \begin{cases} 1 & X_i \in A \\ 0 & \text{else} \end{cases}$$

The observations, n_a, are obtained from the distribution of X given X=A, so the density distribution of n_a is $p_A(x) = p(x)1_{X \in A}/\int_A p(x)dx$.
The objective of this study is to get the value of $\mu_A = E(f(X)|X \in A)$.
The estimate of its is given by the following formula:

$$\mu_A = \frac{1}{n_A}\sum_{i=1}^{n} A_i Y_i \tag{4.11}$$

The estimate of its variance is expressed by the formula below:

$$s_A^2 = \frac{1}{n_A - 1}\sum_{i=1}^{n} A_i (Y_i - \mu_A)^2 \tag{4.12}$$

In order to estimate these values it is necessary to assume that n_a is large enough to be able to get reasonable estimates.

4.1.4 Estimating ratios

Sometimes it is useful to evaluate and focus on the ratio of two jointly distributed random variables X and Y.
The ratio estimator is defined in this way: $\theta = \frac{E(Y)}{E(X)}$, may be more accurate than an ordinary estimator sampled from p.
The most natural way to estimate θ is to be sampled several pairs of random variables *(X, Y)* from their distributions and apply the following formula:

$$\hat{\theta} = \frac{\overline{Y}}{\overline{X}} \tag{4.13}$$

In (4.13) $\overline{X} = \frac{1}{n}\sum X_i$ and $\overline{Y} = \frac{1}{n}\sum Y_i$.

The main problem of using Monte Carlo in this estimation is given from the fact that $E(\hat{\theta}) \neq \theta$, with the consequence of having to estimate a confidence interval for θ to obtain an estimate of the variance of θ.
The confidence interval for *f(E(X), E(Y))* is centered on $f(\overline{XY})$, where *f(x,y) = y/x*.
This problem becomes unimportant for large n.

The delta method solves this problem because it can approximate the mean and the variance of θ.
This method is based on Taylor expansion of *f*, which can be a smooth function of one, two or more means.
It is possible to simplify the formula for estimation of ratio estimator arriving to the following formula:

$$\frac{1}{n}\frac{E((Y-\theta X)^2}{\mu_x^2} \tag{4.14}$$

Instead, the estimation of $f(x,y) = \frac{y}{x}$ s simplify in this formula below:

$$E(\hat{\theta} - \theta) = \frac{1}{n\mu_x^2}(\theta\sigma_x^2 - \rho\sigma_x\sigma_y) \tag{4.15}$$

The confidence interval for θ ignores the bias, so the bias is 0 while the RMSE is of order $\frac{1}{\sqrt{n}}$.

Another method that can be used, in place of delta method, is the Fieller solutions (Fieller, 1954).
He started the analysis from a different definition of ratio estimator where $\theta = \frac{E(Y)}{E(X)}$ as $E(Y - \theta X) = 0$.

For any candidate value θ it can make a confidence interval I_θ for $E(Y-\theta X)$. Then the confidence interval for θ is $(\theta|0 \in I_\theta)$.

Between the two methods, the delta method is simpler so many researcher use it for the ratio estimation.

4.1.5 When Monte Carlo fails

Monte Carlo method is very robust, but there are some cases where also it could fail.

There is no one problem until $\mu = E(Y)$ exists and the expected value is finite, so $E(|Y|) < \infty$.

The problems start when mu might not exist. In a problem with $\mu = \infty$, it has $P(\mu_n \to \infty) = 1$ by the law of large numbers. But when all of the x_i are always finite, then $P(\mu_n = \infty) = 0$ for all n.

In the other case, when $E(|Y|) = \infty$ then it is possible that $E(Y) = +\infty$ or $E(Y) = -\infty$ or that *E(Y)* is not even defined as a member of $[-\infty, \infty]$.

The latter case arises when $E(max(Y, 0)) = E(max(-Y, 0)) = \infty$.

A infinite means could often arises in situations of ratio estimation, because $\mu = E(\frac{Y}{X})$ might not exist because $\frac{(}{Y}X$ can become larger for small X, not just large Y.

In ratios, for μ to be finite it ordinarily need the mean of the numerator to be finite and the denominator should not have a positive density at 0.

Small changes in the distribution of the denominator can turn a problem with finite expected ratio into one with infinite expected ratio.

In some cases μ could be infinite without show it, but in the St. Petersburg paradox it is important to show from the beginning that μ is infinite.

Another situation in which Monte Carlo fails is when the variance is infinite, but the μ is finite.

If $E(|Y|) < \infty$ and $Var(Y) = \infty$ then it is still possible to estimate $\mu = E(Y)$ and get a confidence for it.

With Monte Carlo methods it has many ways to reformulate the problem, preserving the finite expectation while obtaining a finite variance. Importance sampling is one such method.

Only mathematical analysis can determine if some moments exist. Monte Carlo could give an indication on μ, but μ could fail to exist because of small region of space. This happens when importance sampling is poorly applied.

4.2 Case studies

4.2.1 Measuring the extent of small sample biases

Many models, unfortunately, are potentially biased because of sampling error when group sizes are small (Deaton, 1985).
Other results from literature suggests that sampling error is not a problem in practice when there are at least 100 or 200 observations per group (Verbeek and Nijman, 1992, 1993).

Paul J Devereux [1], in this paper, shows that these conclusions are not necessarily correct. He uses synthetic cohorts to solve problems within unobservable factors, that are time-invariant.
This approach is often uses in labor supply estimation and other areas of labor economics.
In fact, he investigates small sample biases in the context of two synthetic cohort applications: intertemporal labor supply model for men and a female labor supply model.

The purpose of his studies is to compare the estimates of samples randomly selected for examining the extent of small sample biases.
For this kind of problem he used Monte Carlo simulation, which quantify biases in a precisely way and permit him to obtained an estimate of group sizes necessary to make biases negligible.

A further goal of this paper is to examine the performance of possible indicators of small sample bias in synthetic cohort models.
He has used and compared different estimators, which are EWALD (efficient Wald estimator), EVE (errors in variables estimator), UEVE (unbiased errors in variables estimator), LIML (limited information maximum likelihood estimator).

4.2.1.1 Case 1: Intertemporal male labor supply

The analysis involves starting with very large group sizes. The men are divided into 6 evenly divided 5-year birth cohorts.
There are in total 90 groups and on average 9818 observations per group.
He reports results for the following percentages of the sample: 1%, 2%, 5%, 10% and 20%. He carries out 1000 replications.

He carries out a Monte Carlo simulation based on the above application. All estimators considered are consistent as group sizes go to infinity with the number of groups fixed.

[1] P. J. Devereux, *Small sample bias in synthetic cohort models of labor supply*, University college Dublin, May 2006

With such large groups, one would expect the sample means to be close to the population means and small sample bias to be small.
Indeed, there is very small sample bias in EWALD. This estimator is sensitive to sample size, so with smaller samples the bias increases.
The other estimators, instead, are more robust to small sample bias. EVE is biases in finite sample, so its estimates tend to decrease as the group sizes increase.

The output of the simulation confirms that huge group sizes are sufficient to get approximately unbiased estimates.
Indeed, EWALD estimator is almost unbiased in the samples of 10000.
The estimates for the smaller samples imply average biases of 88% for the group of 100, 47% for the groups of 200, 19% for the groups of 500, 10% for the groups of 1000 and 5% for the groups of 2000.

In contrast to EWALD, LIML and UEVE perform quite well in terms of bias in both small and large samples.

4.2.1.2 Case 2: Female labor supply model

In the studies of female labor supply, many researches treat wages as exogenous or use age or education as instruments for wages.
In cases where wages are related to taste for work, then cross-sectional estimations produce inconsistent estimates of wages and income.
He grouped married working women aged between 20 and 50 by birth cohort and by education into 8 groups.
For education groups, he split the sample between women with high school or less, and women with more than high school. There are in total 110 groups and on average 4533 observations per group.

He carries out a Monte Carlo simulation for this application.
The results from randomly generating average group sizes of 200, 500, 1000 and 2000 suggests that there is clear evidence of small sample bias in the EWALD estimator.
Indeed, in groups of 200 the bias is 28%, in groups of 500 it is 23%, in groups of 1000 it is 18% and in groups of 2000 it is 13%.
Even with 4533 observations per group there is substantial small sample bias in the EWALD estimator. Some results also suggest that the EWALD estimate may be biased in the full sample.
The latter estimator has coverage rate equal to zero, which mislead the researcher into believing that the income elasticity is very small.

The others estimators, LIML and UEVE perform well in terms of median bias, particularly LIML is better when group sizes are very small.

UEVE is superior to EVE and EVE2, in terms of both median bias and median absolute error.

From the results of these two applications, it is evident that it is required thousands of observations per group before small sample biases can be ignored in estimation.
Sampling error leads one to underestimate intertemporal labor supply elasticities for men, and conclude that the income response of female labor supply is zero or tiny when in fact it is quite large.

4.2.2 Measuring sampling error from an artificial population

The purpose of this paper [2] is to examine the relationship and correlation between immigration and crime, analyzing European recent immigration waves.
The data used for studies derived from surveys, so they are affected by sampling errors which are responsible of attenuation bias in empirical estimates. It is proposed some methods to decrease the attenuation bias; indeed, models with fixed effects have small sampling error and models with instrumental variables have practically zero errors.

Monte Carlo simulation is used to demonstrate and calculate the size of samples necessary for eliminate sampling errors and to obtained an unbiased estimation of the parameters.
The starting point of the simulation is to define population model. In this paper, it studies an artificial population of 10 million of individuals, divided in 100 regions.
It wants to study panel model, so this population is analyzed for 4 years, and, for each year, the immigration shares increases because of a random positive immigration shock.
Then, it draws 500 random sampling with different sampling rates and for both models, fixed effect and instrumental variables, it estimates the coefficient of interest Beta and standard error.

The aim of Monte Carlo is to measure the presence of attenuation bias in samples selected randomly and of different sizes, for region/year cells of different sizes.
From fixed effects estimation, it is possible to observe that with low sampling rate the attenuation bias can be large.

[2] Luca Nunziata, *Immigration and crime: evidence from victimization data*, March 2015

Precisely, with sampling rate of 5/1000 the bias is 52%, with sampling rate of 1/100 the bias is 35% and with sampling rate of 3/100 the bias is 15%.

The SSIV model gives better results than the fixed effects model; indeed, even for very low sampling rate the attenuation bias is very small.
The results show that with sampling rate equal to 1/1000 the bias is lower than 5%.

This study confirms that the fixed effects estimates may be subjected to an attenuation bias.
However, even if the data are affected by sampling error, it will be possible to measure the effects of immigration on crime victimization or perception.
The empirical results show that an increase in immigration share involves an increase of crime perception and of fear of crime of natives.

4.2.3 Monte Carlo for proving the consistency of nonparametric poolability tests

Jin and Su (2013)[3] propose a nonparametric poolability test for large dimensional semiparametric panel data models with cross-section dependence. This test requires an estimation of heterogeneous regression relationship and the test statistics have an asymptotic normal distributions under both hypothesis, such as poolability and a sequence of Pitman local alternatives.

In their paper, they prove the consistency of the test with Monte Carlo simulation. Indeed, they show that the test performs well in finite sample.
In addition, they suggest a bootstrap method as an alternative way to obtain the critical values.

Economic theory cannot tell if the regression is homogenous, so it is useful to conduct a test to verify this property.
In the case when it could accept this hypothesis, then it could be possible to estimate a single homogenous relationship more efficiently because the cross section data are pool together.
Indeed, a large literature has been developed to test structural stability of economic relationships over time or equality of regression functions over individuals.
The lack of these test consists in the number of nonparametric regression curves; in fact, this test verify only the case of fixed number, but it is not known what could happen when the number increases over the sample size.

In this paper they try to test all possible cases and resolve some questions. So, they analyze both heterogeneous and homogenous nonparametric

[3]S. Jin and L. Su, *A nonparametric poolability test for panle data models with cross section dependence*, Singapore management university, 2013

regressions when both the cross-section dimension and the time dimension are large. Indeed, they consider a nonparametric test for poolability in the model.
They find that it is possible to obtain large gains when the regression relationship is homogenous and use it in the following estimation procedure.

There are many differences between their test and others, because their test have some particular characteristics:

- their test is a nonparametric test for homogeneity or poolability of nonparametric regression relationships
- their test is designed to test for poolability in large dimensional panel data models with cross-section dependence
- in their test the number of regression curves must tend to infinity sufficiently fast to ensure that the proxy error is asymptotically negligible in the tests

They conduct a Monte Carlo simulation because they want to compare the sample performance of their tests with some other tests.
They analyze three tests of poolability:

1. Test which assumes unobservable common factors and unknown functional relationship
2. Test which assumes unknown functional relationship, but does not assume unobservable common factors
3. Test which assumes linear functional relationship, but does not assume unobservable common factors

For the normal critical values based tests, they consider two sample sizes for n=50 and 100.
They consider T=20, 30, 40, 50 when n =50 and T = 25, 50, 75, 100 when n = 100. In each scenario the number of replications is 1000 for the size study and 500 for the power study.

For the bootstrap version of the test, they suggest using a conditional bootstrap method to obtain the bootstrap p-values.
In this case, they use B0200 bootstrap resamples for each replications.

From the first analyze, the tests perform reasonably well, in fact the results from the empirical levels of 5% and 10% tend to be similar when n and T are large.

Test 1 can be oversized for smaller value of T, while test 3 can be undersized for some value of n and T.
The last test always rejects the null hypothesis of homogenous regression functional relationship.

For the bootstrap version of the three tests, they find that the bootstrap p-values based test outperforms the normal critical values based test in that the empirical level of the bootstrapped test is quite close to the nominal level for both test, 1 and 2.
The bootstrapped test is severely oversized for all combinations of n and T. In fact, the empirical level of the bootstrapped test seems fine for test 1 when n = 50, but it is identically zero when n=100.

The main findings are the following:

- the bootstrap version of the test tends to be more powerful than the normal-critical-values-based test
- as T increases, the power of the test tends to increase; for fixed T, but increasing n, the power is not necessarily increasing, so the larger is n, the more heterogenous is the regression relationship
- as the degree of heterogeneity increases the power of the test increases rapidly

4.2.4 Monte Carlo for estimation of optimal IV estimators

Mandy and Martins-Filho [4] focus their studies on the structure of the error covariance matrix and on the instrument design because they want to demonstrate the asymptotic equivalence between FGLS IV (feasible generalized least squares instrumental variable) and GLS IV (generalized least squares instrumental variable).

They provide sufficient conditions that permit this equivalence and apply them to stationary dynamic systems with stationary VAR errors.
The sufficient conditions allow them to expand the class of IV estimators, that enable the use of lagged endogenous variables, despite the presence of of VAR errors in the dynamic system.
The use of new instrumental variables improved the asymptotic efficiency, so also small-sample efficiency is improved as well.

[4] D.M. Mandy and C. Martins-Filho, *Optimal IV estimation of systems with stochastic regressors and VAR disturbances with applications to dynamic systems*, Econometric review, 2001

Monte Carlo experiments compares optimal FGLS IV estimators with each other and with other proposed by other researchers in the literature.
From the simulation it is possible to observe the asymptotic properties of the estimators, such as T increases from 20 to 40 and to 80, they obtain less disperse and better centered estimates for all parametric specifications.
The major distinction between the estimators is with smallest T, so T=20.
Then, more the sample size increase, more the estimators become similar in performance.

With larger first-step of instrumental variable set, also, they obtain better dispersion and centrality of all estimators, but the improvement is modest and is more important for small sample size.

Finally, from the analyses of their outputs, they note that introducing additional IVs when constructing FGLS IV estimators is beneficial both asymptotically and in small samples.
Indeed, the benefits of additional IVs diminish as the number of IVs expands.

4.2.5 Monte Carlo for measuring the powerful of tests in small and medium sized samples

In this paper[5], it is studied a generalized panel data model with random effects and first-order spatially correlated residuals.
These models are tested using Lagrange multiplier and likelihood ratio tests.
Monte Carlo simulation measures the powerful of tests for these restricted specifications even in small and medium sized samples.

The recent literature on spatial panels distinguishes between two different spatial autoregressive error processes.
One is based on fact that the spatial correlation occurs only in the error term and does not take place in individual effects.
The other one is based on fact that the same spatial error process is applicable to both individual effects and error term.

The model studied in this paper is a generalized spatial error model that allows the spatial correlation for both individual effects and error term, which may have different spatial autoregressive parameters.
They used a MLE (maximum likelihood estimator), where the assumption is to consider the individual effects random.
They want to test three different hypothesis given by some restrictions on their generalized model; indeed, they are interested to obtain the Anselin model, the KKP model and simple random effects model.

[5]B.H. Baltagi, P. Egger. And M. Pfaffermayr, *A generalized spatial panel data model with random effects*, 2013

For each of them, they derive the corresponding LM and LR tests.

Monte Carlo simulation is used here for measure and compare the size and power performance of these three hypothesis.

The cross-sectional and time dimensions are N =50, 100 and T = 3, 5, 10. The proportion of the variance due to the random individual effects takes the value $\theta = 0.25, 0.50, 0.75$. In total this gives 882 experiments.

For each experiment, they calculate the three LM and LR tests, using 2000 replications.

They testing the simple random effects model without spatial correlation and, instead, Anselin model and KKP model are testing in small and medium sized samples.

The three LM and LR tests perform reasonably well and things improves is the number of observations increases.

Both the size and the power of the LM test improve as the sample size increases, especially as N becomes larger.

With small samples and small signal to noise ratio, there is no gain using robust LM than the non-robust ones.

The robust test size is more off the nominal size than this is the case for the nonrobust test size.

The correction factors of the LM statistics deflate the nonrobust test statistics.

With oversized LM tests, the corresponding correction factors would adjust the test size towards the nominal size.

There is no systematic over-rejection in the samples considered so that the correction factors lead to even more undersized tests.

Problems with such correction factors in small samples also accrue to the use of higher moments of the disturbances which can not be estimated without bias in small samples (Teuscher, 1994).

They find, from this study, that the LM tests are easy to calculate and their power is reasonably high for all three tests considered.

Under normal disturbances the LM tests are properly sized and powerful even in small samples.

Furthermore, under normal disturbances, the power of LM tests matches that of the corresponding LR tests.

In conclusion, the power of the tests increases with the relative importance of the individual effects' variance as a proportion of the total variance, as well as with increasing N and T.

They are robust to non-normality of the error term and sensitive to the specification of the weight matrix.

4.2.6 Monte Carlo for verifing the finite-sample properties of estimators in small samples

Drukker, Egger and Prucha (2013) [6] study a spatial-autoregressive model with autoregressive disturbances which is useful for endogeneous regressors. They propose a joint test of zero spatial interactions in the dependent variable, the exogenous variables and the disturbances.
For these analysis they used two estimation approach, such as a two-step generalized method of moments (GMM) and instrumental variable approach (IV).

Spatial models, in previous literature, are based only on spatial spillovers in the dependent variable, whtch has also an endogenous weighted average. This model is commonly referred to as spatial-autoregressive model, SAR (Cliff and Ord, 1973, 1981).
The combined spatial autoregressive model with spatial autoregressive residuals is often referred to as SARAR (Anselin and Florax, 1995).

In this paper, they started their analysis from a generalized method of moments (GMM) and instrumental variables (IV) estimation for systems of linear equation developed by Kelejian and Prucha.
Then, they obtain the distribution of the regression parameters and the consistent estimator; indeed, they derive the joint limiting distribution of the IV estimators and of GMM estimators and the consistent estimators for the variance-covariance matrix.

In their study Monte Carlo is used to verify the finite-sample properties of the estimators, IV and GMM estimators, in small samples.
The experiments is organized that from each observation of each variable is subtracting the corresponding sample average and, then, dividing that results by sample standard deviation.
For sample size, n, each vector of normalized observations are twice underneath each other and is drawn the first n values of these normalized variables. The set of normalized observations on these variables is fixed in repeated samples in their Monte Carlo runs.

From the results of Monte Carlo simulation they could deduce that a good approximation of small-samples distributions is given by derived large-sample distribution of their estimators.
Furthermore, small-sample biases of the estimators are small for each of the parameters, the means of the estimated standard deviations of the parameter estimators over the Monte Carlo repetitions are close to the actual standard

[6] D. M. Drukker, P.Egger and I.R. Prucha, *On two-step estimation of spatial autoregressive model with autoregressive disturbances and endogenous regressors*, 2013

deviations and their estimators and their large-sample approximation to its distribution work well for the considered experiments.

4.2.7 Monte Carlo for studying the performance in finite samples of instruments selection procedure

This study considers the instrumental variable (IV) estimation of spatial autoregressive (SAR) models with endogenous regressors in the presence of many instruments [7].
This kind of analysis, where the number of instruments increases with the sample size, has attracted a lot of attention in the IV estimation of instruments.
From this research it is possible to obtain the asymptotic distribution of the two-stage least squares estimator and to deduce a procedure to correct the bias, using an leading-order of many instruments bias.

Monte Carlo simulation is useful here to study the performance in finite samples of the instruments selection procedure.

Analyzing the outputs of the experiments they could deduce the following conclusions:

- the instrument selection reduces median absolute deviation and dispersion
- the instrument selection improves coverage probability
- the instruments selection reduces median bias of the 2SLS; in fact, the bias-correction procedure could reduces many-instruments bias. Choosing the number of instruments tends to raise precision and lower dispersion of the 2SLS, but when instruments are equally important, instrument selection may be less useful for the 2SLS (Donald and Newey, 2001)
- the instrument selection improves precision of the 2SLS estimator with only moderate sample size such as n=490
- the instruments selection leads to smaller bias, better precision and more reliable inference when there are some instruments more important than others

[7] X. Liu and L. Lee, *Two-stage least squares estimation of spatial autoregressive models with endogenous regressors and many instruments*, 2013

4.2.8 Monte Carlo for for studying finite sample properties of estimators

Several recent studies have focused on the estimation of panel data under cross-sectional dependence [8], which is common in economic data.
In panel data modeling it is also important to have a correct specification in the conditional mean.

This study extends the method to nonparametric estimation in large panels under multifactor cross-sectional dependence that are based on random-effects specification.
The estimator is applicable to both static and dynamic panels.

Monte Carlo simulation studies the finite sample properties of the proposed estimator.

They compare three estimators in the simulation:

1. local constant regression
2. local linear estimation
3. the MGCCE, where the slope estimator is obtained by averaging estimates from all cross-sectional units (Pesaron, 2006)

In simulations, they discard the first 100 observations, take different values with n = 50, 100, 200 and T = 50, 100, 200.
The number of replications is 1000.

The efficiency gain of using nonparametric estimator increases with sample size.
The local linear estimator also has a clear advantage over the local constant estimator, while MGCCE is outperformed by two nonparametric estimators in all sample sizes.
The MGCCE is almost unbiased, while the bias of the local linear estimator is small and decreases as sample size increases.

Monte Carlo simulation demonstrate that the estimators proposed here produce very good results for models with high degrees of heterogeneity and dynamics.
Furthermore, it is evident that the proposed method has good finite sample properties and that the efficiency loss of this nonparametric method decreases as sample size increases.

[8] Xiao Huang, *Nonparametric estimation in large panels with cross-sectional dependence*, Kannesaw state university, 2013

Conclusions

This paper explains the importance and the role of sampling error in the inference process, because a biased sample leads researchers to take wrong deductions on real population.
It is evident that the sampling error is inversely correlated with the sample size. In fact, with increasing sample size, the sampling error decreasing.

The case studies presented in this paper discuss different types of analysis; in fact, in some papers, Monte Carlo is used to estimate the extent of small sample biases in samples randomly selected and, in other papers, it is used for testing the performance of the estimators and the consistency of the tests in finite samples.

From the empirical results it is possible to deduce some conslusions:

1. it is required thousands of obervations per group before small sample biases can be ignored

2. in fixed effect models are present small sampling error so they may be subjected to an attenuation bias, while instrumental variables models have a bias practically zero

3. nonparametric poolability tests (Jin and Su, 2013) for large panel data performs well in finite samples; in fact, the larger is the numerosity of the sample, the more heterogenous is the regression relationship and as the degree of heterogeneity increases then the power of the test increases rapidly

4. the power of tests increases with increasing N and T

5. the instruments selection reduces the bias in finite samples, indeed choosing the number of instruments tends to raise precision and lower the disperion of the estimators

It is impossible to eliminate the sampling error completely. The only way is to sampling whole the population, but this is impossible.

It is important to measure this bias and consider it when take some decisions and deduce the characteristics of the population.
Monte Carlo permits to understand the importance of the error and to decide which estimarors are better to obtain more reliable estimates.

Monte Carlo method is used in many other areas of study. In this paper are presented the other uses, but it is analyzed only the estimation of errore. It permits the analysis and resolve many difficult probems that other methods can not solve.

Bibliography

[1] A. Aydemir and G.J. Borjas, *Attenuation bias in measuring the wage impact of immigraztion*, September 2010

[2] Angus Deaton, *Panel data from time series of cross-sections*, Journal of Economics, 1985

[3] Michael Hauser, *Financial Econometrics*, 2014/2015

[4] Paul J Devereux, *Small Sample Bias in Synthetic Cohort Models of Labor Supply*, University College Dublin, 2006

[5] Sainan Jin and Liangjun Su, *A nonparametric poolability test for panel data models with cross section dependence*, Singapore Management University, Singapore, 2013

[6] Thomas T. Semon, *Non-response bias affects all survey research*, July 2004

[7] Martin N. Marshall, *Sampling for qualitative research*, Oxford University, 1996

[8] Jianguo Lu and Dingding Li, *Bias correction in a small sample from big data*, November 2012

[9] Good and Hardin, *Common errors in statistics*, 2006

[10] David E. Freedman, *Statistical models and causal inference*

[11] H.F. Weisberg, *The total survey error approach: a guide to the new science of survey research*, University of Chicago, 2005

[12] William Greene, *Econometric analysis of panel data*, Stern School of business

[13] M. Verbeek and T. Nijman, *Minumum MSE estimation of a regression model with fixed effects from a series of cross-sections*, Journal of Econometrics, 1993

[14] Paul J. Devereux, *Improved errors-in-variables estimators for grouped data*, School of Economics Dublin, January 2006

[15] V. Sarafidis and D. Robertson, *On the impact of error-sectional dependence in short dynamic panel estimation*, The Econometrics Journal, 2009

[16] Samik Raychaudhuri, *Introduction to Monte Carlo simulation*, 2008

[17] David W. Gerbing and James C. Anderson, *Monte Carlo evaluations of goodness of fit indices for structural equation models*, November 1992

[18] Liangjun Su and Zhenlin Yang, *QML estimation of dynamic panel data models with spatial errors*, Journal of Econometrics, November 2012

[19] Gabor Kezdi, *Robust standard error estimation in fixed-effects panel models*, Budapest University of Economics, October 2003

[20] Stephen Nickell, *Biases in dynamic models with fixed effects*, Econometrica, November 1981

[21] Peter M. Robinson, Carlos Velasco, *Efficient inference on fractionally integrated panel data models with fixed effects*, Journal of econometrics, March 2013

[22] R.A. Judson and A. L. Owen, *Estimating dynamic panel data models: a practical guide for macroeconomists*, January 1996

[23] T.W. Hall, A.W. Higson, B.J. Pierce, K.H. Price, C.J. Skousen, *Haphazard sampling: selection biases induced by control listing properties and the estimation consequences of these biases*, Behavioral research in accounting, 2012

[24] P. J. Devereux, *Small sample bias in synthetic cohort models of labor supply*, University college Dublin, May 2006

[25] Luca Nunziata, *Immigration and crime: evidence from victimization data*, March 2015

[26] S. Jin and L. Su, *A nonparametric poolability test for panle data models with cross section dependence*, Singapore management university, 2013

[27] B.H. Baltagi, P. Egger. And M. Pfaffermayr, *A generalized spatial panel data model with random effects*, 2013

[28] D. M. Drukker, P.Egger and I.R. Prucha, *On two-step estimation of spatial autoregressive model with autoregressive disturbances and endogenous regressors*, 2013

[29] X. Liu and L. Lee, *Two-stage least squares estimation of spatial autoregressive models with endogenous regressors and many instruments*, 2013

[30] Xiao Huang, *Nonparametric estimation in large panels with cross-sectional dependence*, Kannesaw state university, 2013

[31] Kalos J.T and Whitlock P.A., *Monte Carlo methods*, New York

[32] D.M. Mandy and C. Martins-Filho, *Optimal IV estimation of systems with stochastic regressors and VAR disturbances with applications to dynamic systems*, Econometric review, 2001