**Università degli Studi di Milano - Bicocca**

FACOLTÀ DI SCIENZE STATISTICHE
Corso di Dottorato di Ricerca in Statistica XXIII ciclo

Tesi di dottorato di Ricerca

# Multilevel and Stochastic Frontier Models: A comparison and a joint approach of their performances when investigating panel data

Discussant:
**Marta Angelici**

Tutor:
**Prof. Giorgio Vittadini**

Co-Tutor:
**Prof. Gianmaria Martini**

**Academic Year 2009 – 2010**

*A Alberto*

# *Acknowledgements*

Vorrei capire, vorrei sapere già quello che devo capire! Per capire bisogna far fatica, dedicare se stessi e avere la pazienza e la disponibilità necessarie.

Ringrazio il Professor Giorgio Vittadini e il Professor Mario Mezzanzanica perchè mi hanno proposto di fare questo dottorato. Mi sono scoraggiata tante volte, ma guardandomi indietro sono profondamente grata di aver avuto questa splendida occasione di dedicare tre anni alla mia formazione.

Ringrazio il Professor Gianmaria Martini con cui ho svolto una parte di tesi, grazie del tempo in cui abbiamo discusso insieme sulle frontiere, compenetrando statistica ed economia. Questo tempo mi ha chiarito che studiare mi piace, quando  fonte di discussione con qualcunaltro.

Ringrazio Paolo Berta, perchè è stato il mio tramite con i dati e con le analisi più statistiche.

Facendo questo lavoro di tesi mi sono proprio resa conto di quanto amo i dati, e non riesco a trattarli male, non considerando quello che sono realmente. Anche se forse non so cosa significa veramente, questo amore per i dati mi fa sentire un po' di pi una statistica!

Ho sempre trovato interessante e necessario studiare gli esami insieme ai miei compagni di corso, e poter discutere delle ricerche. Ringrazio tutti coloro con cui ho studiato, discusso, dialogato, a cui ho posto i miei dubbi.

Ringrazio i miei genitori, che mi hanno sempre sostenuta e mi sono stati vicino alla grande, in modo direttamente proporzionale alla distanza fisica che ci separa. Vivere lontani è difficile, ma se serve ad avere un rapporto così bello vale proprio la pena.

Ringrazio tutti i miei amici, che mi hanno sostenuta in questi anni. Il volto di ognuno di voi è stato per me memoria che siamo fatti per qualcosa di grande, e per tutte le volte che mi avete fatto sentire veramente voluta bene.

Infine ringrazio Alberto, a cui dedico la tesi, perchè la pazienza e la costanza con cui mi sta vicino sono per me un esempio di come affrontare le circostanze della vita, e soprattutto perchè ha creduto in me.

# Contents

# List of Figures

# List of Tables

# *Introduction*

The concept of a joint approach in regard to Multilevel Models and Stochastic Frontiers developed in light of the increasing levels of interest in costs related to health care services, including hospitals, over the last few years. At the same time, both consumers and policy makers prioritize the quality of these services, and a holistic approach is required to identify areas for improvement in this regard. Quality in healthcare services means the ability to meet specific requirements, and it is the result of scientific, technical and technological, organizational, procedural and relational factors , where the human variable plays a primary role, interacting closely in the production process (Vittadini (2006)). In the present day many industrialized countries have healthcare systems which designate resources to hospitals according to predefined tariffs for each pathology. For example, the Lombardy Region in Northern Italy establishes an annual tariff for Diagnostic Related Groups (DRG). Therefore, the structure of the public health system depends on national and regional political decisions, but it becomes self-governing from an operational point of view. Accordingly, hospitals can be compared to firms, thus admitting functional models in an entrepreneurial way and introducing competition between the various healthcare structures. Hospitals try to devise an organizational model which allows the reduction of costs by optimising the use of available resources, while simultaneously increasing patient satisfaction by providing optimal medical assistance. Healthcare is currently following this trend and in many countries it is based on a mixed welfare system made of private profit-oriented agents, private non-profit companies and public hospitals. Consumers are free to choose between them. Within the health care system in Lombardy, consumers have an open choice in deciding where they prefer to receive care. Freedom of choice is a basic principle and it is the basis for all other principles. The purpose of this work is to create decisional models similar to those used in financial planning, with a large number of variables, with the aim of estimating the risks involved in some enterprises, and devising accordingly appropriate allocations of funds.

This study is on the basis of the following structure: the first chapter is dedicated to the multilevel model. As there are several potential developments and applications of the multilevel, this section is effectively a description of its context and principal characteristics. The multilevel model is contextualized in the measure of the health care, so the first aspect considered is the the adjustment of data according to patient-specific and hospital-specific variables. It is also possible to use other techniques, like direct and indirect standardization, linear and logistic models, but the most complete technique is the application of the multilevel model. It is crucial to examine the reasons behind not neglecting the real structure of the data, and the various problems which arise if the structure is ignored. Furthermore it is necessary to introduce a measure which is a descriptive statistic that can be used when quantitative measurements are made on units that are organized into groups. It describes how closely units in the same group resemble each other. In this case we refer to the intra class correlation index. In our case we have developed a three level analysis, so we describe the method behind computing the coefficient for a three level model. At the end of the chapter we also list some critiques addressing the incorrect application of multilevel.

The second chapter is dedicated to the stochastic frontier, and is a theoretic description of this model with a different possible application. In doing this we underline the reasons which led us to use it for hospitals. The interesting aspect is the possibility of seeing the hospital in terms of distance from its frontier of optimality. In this way we take the economic aspects of the structure into account, and we can see its effectiveness in terms of efficiency. It is clear that inside the frontier there is no consideration for the hierarchical structure of the data, but is important to not neglect the real structure of data (as we have explained in the previous chapter). This is an evident reason that allow us to consider both the models together, with all the strength stressed, and all the weaknesses overcome.

Chapter number three is about the characterization of data. The first section describes the data available and the sources, on a general basis. Then we continue with a section analyzing variables of patients and a further section on the hospitals variables. We have decided to describe the variables we have utilized in the model solely, but - as is stated throughout the study - we collated plenty of data, as well as other variables that we excluded from the model as they were either not of interest for our actual research, or because they impeded the convergence of the model. Although we considered and modeled some of the variables which were not used, we concluded they were not material or pertinent for the explanation of our model.

Chapter four comprises a description of the model that we have applied to the data. In the first section we describe in general the concept; then we continue by describing the two step methodology, and the two steps which were implemented. For each step we explained the particular procedure utilized with the peculiar feature which led us to it. For the second step we also describe the two variables we introduced after the first step because are the result just of the first step of the analysis. In addition we have proposed a section with possible future developments that we have elaborated during this work.

Finally chapter five is about the results we obtained through the analysis. An initial section has comments on the first step, namely the multilevel. These comments on the coefficient and the model itself can be utilized on their own to obtain interesting results, but in this section we exploit their utility for the subsequent application of the frontier. Therefore in the subsequent section we comment on the results of the frontier.

The conclusion section contains some considerations we have made in light of the whole work.

# Chapter 1

# Multilevel and Healthcare

We commence our discussion by describing the models that we will use during the analysis: the first being the multilevel model, that takes into account the hierarchical structure of the data and addresses performance; and the second being the stochastic frontier model, which considers the maximum technical efficiency achieved.

Multilevel Analysis is a way to analyze data with complex patterns of variability, focusing the attention on nested source of variability. Classical examples are pupils in schools or patients in hospitals. Considering variability at each level is both useful, and allows us avoid errors in addressing incorrect data.

Multilevel Analysis comprises statistical techniques and methodologies. This type of analysis is mainly used in the fields of social science, like sociology, education, psychology and economics. The Multilevel Analysis is a stream which has two tributaries: contextual analysis and mixed effect models (Snijders and Bosker (1999)).

Contextual analysis is a development in social science that focuses on the effects of social context on individual behaviour. The individual and the context have different sources of variability that have to be modeled as random inferences.

The mixed effects models are all the statistical models that one finds behind an analysis of variance or inside a regression analysis where some coefficients are random and others are fixed.

The principal statistical model of multilevel analysis is the hierarchical linear model. It can be viewed as an extension of the multiple linear regression model to a model that includes nested random coefficients.

After this general description of the multilevel analysis we are going to context the multilevel model inside the set of instrument for the measurement of healthcare services.

## 1.1 Measurement of Healthcare Services

When analyzing Healthcare Institutions it is necessary to purify the data from the effects of case mix variables in order to facilitate comparisons between different healthcare institutions on a like for like basis . Specifically this requires the adjustment of data according to patient-specific and hospital-specific variables. We consider different Risk Adjustment Methodologies. These techniques allow us to measure quality in terms of relative effectiveness, without bias resulting from case mix influence.

### 1.1.1 Direct and indirect standardization

Direct standardization (Zaslavsky (2001)) is computed as follows: given $y_{kj}$ health outcome observed in the $k-eth$ status of the population patient and $\pi_{kj} = w_{kj}/\sum w_{kj}(k = 1, ..., q)$, proportion of the $k-eth$ case mix characteristic in the $j-eth$ health structure, the weights of the outcome. The observed adjusted outcome is the weighted sum: $y_j = \sum_k \pi_{kj} y_{kj}$. This type of standardization has limitations: it is not possible to compute a standardized score for a stratum with no cases or with missing cases, and furthermore it is not adapted to adjusting simultaneously for several variables or for continuous variables. Another approach is the indirect standardization. Given the weighted sum: $\hat{y}_j = \sum_k \hat{\pi}_{kj} y_{kj}$, where the weights $\hat{\pi}_{kj}$ are obtained from a standard theoretical population, in order to evaluate the relative effectiveness of the j-eth health structure, we compare the observed adjusted outcomes with the expected adjusted outcomes: $u_j^* = \hat{y}_j/y_j$. This standardization is not suitable when there are several case mix indicators or when the indicators are not discrete.

### 1.1.2 Linear and logistic models

Below we describe the linear and logistic models used as a method of risk-adjustment. We can define: $y_{ij} = \beta x_{ij} + e_{ij}$, where $y_{ij}$ is the j-eth quantitative outcome for i-eth patient, $x_{ij}$ the corresponding patient characteristics, $e_{ij}$ the error term, and $\beta$ is a vector of coefficients. The first term of the equation captures the effects of individual characteristics $x$ on outcomes for the patients of the same unit. In other cases a covariance analysis is proposed: $y_{ij} = \beta x_{ij} + u_j + e_{ij}$, where $u_j$ captures the hospitals heterogeneity in reaching a given quality in healthcare services of the j-eth agent. In the simplest formulation with dichotomic outcomes (the most commonly used is the hospitals mortality risk), the logistic function models calculate the logit of the outcome $p_{ij}$ as a linear function of the case mix variables $x_k$ $(k = 1, ..., p)$:

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + \beta_1 x_{1ij} + \beta_2 x_{2ij} + ... + \beta_k x_{kij} = RS_{ij}$$

with $RS_{ij}$ risk score associated to $i-eth$ patient of $j-eth$ structure. It can be illustrated in terms of probability:

$$p_{ij} = \frac{e^a + \beta_1 x_{1ij} + \beta_2 x_{2ij} + ... + \beta_k x_{kij}}{1 + e^a + \beta_1 x_{1ij} + \beta_2 x_{2ij} + ... + \beta_k x_{kij}} = \frac{e^{RS_{ij}}}{1 + e^{RS_{ij}}}$$

Where $p_{ij}$ is the observed probability (equal to 0 or 1) regarding the positive or negative occurrence of the dichotomic outcome. By estimating the vector of parameter $\beta$ and substituting $\hat{\beta}$ we obtain the expected probability $\hat{p}_{ij} = \frac{e^{\hat{p}_{ij}}}{1+e^{\hat{p}_{ij}}}$. Comparing $\hat{p}_{ij}$ with $\hat{p}_j$ we arrive at an estimation of the effectiveness of the $j-eth$ health structure for the $i-eth$ patient. Therefore the expected value of $\hat{p}_{ij}$ is obtained as: $E\left(\hat{p}_{ij}\right) = \sum_{i=1}^{n_j} \hat{p}_{ij}$ and the correspondent value of the observed probabilities $\sum_{i=1}^{n_j} p_{ij}$ the ratio $\hat{u}_j = \frac{\sum_{i=1}^{n_j} \hat{p}_{ij}}{\sum p_{ij}}$, estimates the effectiveness of the $j-eth$ health structure.

In some cases logistic or linear models are not adequate for the calculation of the relative effectiveness of health structures for the following reasons: first a sample of agents is chosen in the agents population; and more importantly secondly, the data often shows highly structured hierarchies because patients or episodes of care are nested within health structures and the higher levels of health institutions. In addition measures of quality are likely to differ for subgroups of patients within a single broad category, where the variety in numbers of secondary diagnoses, patients of different ages and sexes, and previous medical histories are all likely to influence outcomes. The characteristics of both patients and health structures need to be taken into account (Normand *et al.* (1995)). The solution to all of these drawbacks is the use of hierarchical models, and in particular we advocate the multilevel model, following the finding of several authors.

### 1.1.3 The Multilevel Model

We now consider the most important element of risk adjustment. It is not a technique, but a model that takes into account the hierarchical structure of the data. This is a global method of adjustment of the data, and has a strong assumption of linearity, but it is more complete. Multilevel models are a particular specification of hierarchical models, that offer solutions for studying the relationship between outcomes (such as mortality, health and quality of life ...) and contextual variables in complex hierarchical structures, considering both individual and aggregate levels of analysis (Goldstein and Rasbash (1996)). During the nineties several authors proposed the study of "relative effectiveness" by means of Multilevel Model (Hox (1995)). Some of these including Carlin *et al.* (2001), Goldstein and Spiegelhalter (1996), Leyland (1995), Verbeke and Molenberghs (2009) and several others, use this basic model.

The Multilevel model specified for the $j-eth$ outcome is:

$$Y_{ij} = \beta_{0j} + \beta_j X_{ij} + \epsilon_{ij}$$

where $Y_{ij}$ measures the outcome relative to the $i-eth$ subject inside the $j-eth$ hospital. Here $\beta_{0j}$ is the random coefficient interpretable as the hospitals-specific heterogeneity with regard to outcome $Y_{ij}$ adjusted for patient characteristics made up of fixed coefficients of patient covariates, a measure of relative effectiveness; $X_{ij}$ is the vector of explicative variables whereas $\epsilon_{ij}$ are the independent residual, with zero mean, at subject level. Casual parameters are usually obtained by means of bayesian inference (Leyland and Boddy (1998)) and they can have prior distribution, above all, in the case of patients. Subsequently, there must be particular focus on estimation parameter procedures as well as the statistical properties of the estimators with respect to single model parameters and different parameter typologies (fixed effects and variance components at different levels of hierarchical structure) as a whole.

Given the hypothesis that the parameters $\beta_{0j}$ and $\beta_{1j}$ are random variables, with constant variance and known distribution, it is also possible to insert into the model variables of superior level, so the multilevel model becomes:

$$Y_{ij} = [\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij}] + [u_{1j} X_{ij} + u_{0j} + \epsilon_{ij}]$$

with $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$ and $\beta_j = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$, where the $Z_j$ represent the explanatory variables at hospitals level. The distribution of the dependent variable influences the choice of linear model -if we have a dichotomous outcome- the linear model will have the characteristics of a logistic regression model. The $u_{0j}$ is a random variable that identifies the relative effectiveness of the $j-eth$ hospital structure net of risk factors. We assume that the residual of the model $\epsilon_{ij}$ are uncorrelated with null expected value: $E(\epsilon_{ij}) = 0; Var(\epsilon_{ij}) = \sigma^2$. We assume that the casual parameter $u_{0j}$ has $\Phi(u_{0j})$ distribution with constant variance : $Var(u_{0j}) = \sigma_u^2$. This model is particularly suited to evaluate relative effectiveness because it allows for the specification of variability *within* and *between* groups.

The basic version of the Multilevel Model used by the authors previously mentioned is subject to some restrictions which limit its utility in practice. The basic version considers only one outcome at a time, like the response variable; it does not consider hospital-specific characteristics as possible covariates; it makes the assumption of binomial distribution of the outcome and of multinomial distribution of random disturbance and the random parameter of effectiveness; and there is no correlation between the expected values of the patients characteristic and the effectiveness of casual parameters, and finally it is a static model with fix effects (multilevel with mix effects).

We can make some generalizations to assure the utility of the model in real cases. In order to obtain solutions we avoid traditional structural equation models, which lead to indeterminacy of latent scores. Instead Partial least Squares and Regression Component Decomposition methods, which approximate latent variables by means of linear combinations of their indicators, can be utilized.

When the indicators are qualitative or mixed, we can intervene in order to obtain quantified observable indicators and their quantitative linear transformations simultaneously, or as an alternative multidimensional scaling methods can be utilized. The Rasch model is suitable for the estimation of latent outcomes as it allows the estimation of objective measurement of performance with the most agreeable properties.

Relative effectiveness needs to be adjusted for hospital characteristics. We can introduce further equations describing hospital characteristics, which a second level equation that expresses the random parameters $u_{jv}, (j = 1, .., p; v = 1, ..., q)$ as a function of hospital characteristics. We can employ families of distribution other than normal or binomial, in order not to be too restrictive.

Mixed effects models assume that the random effects of agents are independent and not correlated with the expected values of patient's characteristics. In complex models this assumption might not be realistic. Individual heterogeneity can be modeled poorly by the available subject-level covariates, and this can affect the estimation of random effects dramatically. In the linear case, possible remedies are based on the use of fixed-effects modeling in place of random effects, correcting the incidental parameter problem caused by using one parameter for each subject. In non-linear cases, fixed-effects estimation is too complicated; we cannot rely on exact conditioning to eliminate the subject-specific parameter, and therefore this is only possible in a generalized linear model with canonical link functions.

### 1.1.4 Why use a probability model

It is informative to justify the use of a probability model when we have all the data of a population available. To do this we have to digress on the sampling theory, where a distinction between design-based inference and model-based inference exists. The design-based inference implies that the researcher draws a probability sample from some finite population. And the way in which the sample is drawn by the researcher implies a particular type of probability model. Instead model-based inference implies that the researcher postulates a probability model, usually aiming at inference to some large and sometimes hypothetical population. At the end of this process it is necessary to check if the probability model is adequate, in this case it is possible to base the inference on it.

One may also use a probability model if no sample is drawn but an entire population is observed. The use of a probability model assumes statistical variability, even though an entire research population was investigated. This fact can be justified by noting that in this way the conclusions can be extended to a wider population. The investigated research population is supposed to be representative for this wider population, also in the applicability is not automatic. The researcher has to argue carefully the reason for which the research population can be considered to be representative for the larger population. The inference is not about individuals, but about mechanisms and processes (which can be social, behavioral or biological). In order to take into account the factors that are not included in the explanatory variable we have the random effect or residuals that play this role. It will be possible to use the model-based inference until the assumptions will be an adequate reflection of the effects that are not explicitly included with the observed variable.

### 1.1.5 Sampling

The simple random sampling is not always a cost-efficient strategy and there is also the possibility to do a multi-stage sample that can be more efficient. In this case the researcher has to consider the clustering of the data during the phase of data analysis. Usually the multi-stage samples are employed to see the relations between variables at different levels in a hierarchical system. The main interest is on the dependency of observations within groups, because in this way is possible to find the aspects in which groups differ.

In standard statistical situation, observations should be sampled independently from each other. Commonly the standard sampling design is the simple random sampling with replacement, which assures a constant possibility of selecting a certain single unit across all units in the population. But there are more cost-efficient sampling designs, based on the idea that probabilities of selection should be known but do not have to be constant. One of these sampling designs is the *multi stage sample*, where the population of interest consists of subpopulations, and selection takes place via those subpopulations. When only one subpopulation level exists, the design is a *two stage sample*. It is possible to notice a mistake which is to ignore the fact that the sampling scheme was a two-stage one, and to pretend that the secondary units were selected independently. The researcher overlooks the fact that the secondary units were not sampled independently from each other and the selection of a primary unit increases the chances of selection of secondary units from that primary unit. The correction is the use of *multi-stage* sampling design, leading to dependent observations. In practice multi-stage samples are also preferred because the costs of interviewing or testing persons are reduced enormously if these persons are geographically or organizationally grouped.

## 1.2 What if we ignore the multilevel structure of the data?

We have introduced the multilevel analysis because we sustain that it is the best way of analyzing data. In this section we seek to support this assertion. In order to prove the efficacy of hierarchical model we can consider some statistical methods for multilevel data that do not use the hierarchical linear model.

### 1.2.1 Aggregation

#### The shift of meaning

A common way in social research of two levels data is the aggregation of the micro-level data to the macro level data. One can normally do this by working with the averages for each macro-unit. The aggregation is not a mistake if the researcher is only interested in the macro-level proposition. But he has to bear in mind that the reliability of an aggregate variable depends on the number of micro level units in a macro-level unit, amongst others factors and that thus this will be larger for larger macro-units than for smaller ones.

It may be the case that the researcher is interested in macro-level or micro-level propositions, and in this case the aggregation may result in gross errors. Under this scenario it is possible to commit a potential error which is *the shift of meaning.* A variable that is aggregated to the macro level refers to the macro-units, but not directly to the micro-units.

### The ecological fallacy

It is also possible to commit a second type of potential error with the aggregation which is *the ecological fallacy.* The problem that can arise is that a correlation between macro-level variables cannot be used to explain something about micro-level relations. An example that Snijders and Bosker (1999) use in their book is that the percentage of black inhabitants in a neighborhood could be related to average political views in the neighborhood. The higher the percentage of blacks in a neighborhood, the higher the proportion of people with extreme right-wing political views may be. This does not give us any clues about the microlevel relation between race and political conviction. In this case the shift of meaning leads to incorrect conclusions: the percentage of black inhabitants is a variable that means something for the neighborhood, and this meaning is distinct from the meaning of ethnicity as an individual-level variable. In general terms we can assert that *the ecological fallacy* is a problem that can arise when the researcher infers characteristics of individuals from aggregate data referring to a population of which they are members. Such aggregate data are frequently used in geographical work, referring to the populations of defined areas (hence they are often termed ecological data), so the problem is potentially serious for some forms of geographical analysis.

The fallacy was initially highlighted by Robinson (2009). Using 1930 US census data, he obtained a high correlation coefficient of 0.773 from a regression of the proportion of a state's population which was illiterate on the proportion which was black. It could be inferred from this that blacks were much more likely to be illiterate than non-blacks, but using data on individuals from the same source, he found a correlation of only 0.203: there was a higher level of illiteracy among blacks than non-blacks, but much less than the state-level (ecological) analysis suggested. The conclusion was clear: just because blacks were concentrated in the states with the highest levels of illiteracy this did not necessarily mean a much higher level of illiteracy among blacks.

Alker (1969) extended the ecological fallacy (of identifying spurious individual - level correlations from analyses of aggregate data) by identifying five others types.

- The individualistic fallacy which assumes that the whole is no more than the sum of its parts (see regionalism) many societies are more than mere aggregations of their individual members;

- the cross-level fallacy assumes that a relationship observed in one aggregation of a population i.e. one set of spatial units applies to all others, and is thus a universal feature of that population: research on the modifiable areal unit problem has demonstrated that this is invalid;

- the universal fallacy assumes that the pattern observed in a selection of individuals often not randomly selected according to the principles of sampling holds for its population;

- the selective fallacy in which data from carefully chosen cases are used to 'prove' a point;

- the cross-sectional fallacy is the assumption that what is observed at one point in time applies to other times.

Recognition of these fallacies and their associated pitfalls indicates a need for careful interpretation of the results of studies based on aggregate data. An observed relationship may be consistent with a hypothesis, but a causal relationship should never be assumed: as Robinson's example showed, wrong conclusions can be drawn by attempts to move from the particular to the general.

The problem of drawing conclusions about individual - level correlations from aggregate data has long concerned social statisticians. Most attempts to resolve it have failed, because they cannot avoid the possibility of producing "nonsense" answers, such as a population in which 120 % of the members are illiterate. However King (1997) has solved this issue for a particular situation. For example if one has information on the number of black people and the number of illiterates in each sub-area of a larger area for which the inference is to be drawn, then using the "method of bounds" it is possible to produce robust estimates of the number of blacks who are illiterate, the number of non-blacks who are literate etc., in that larger area as well as in each of the sub-areas.

### The neglect of the original data structure

There is another type of potential error which is the neglect of the original data structure, and it happens especially when some kind of analysis of covariance is to be used. It is possible that the averages of all groups are almost perfectly on the regression-line, and this can lead to the incorrect conclusion that there are almost no differences between the group we are studying. But if we go into the details we can find for example that the micro-groups formed have different inclinations. Working with aggregate data "is dangerous at best, and disastrous at worst" (Aitkin and Longford 1986). Dealing with multilevel data without aggregation is possible by distinguishing between the within-groups and the between-groups regressions.

### 1.2.2 Disaggregation

We can treat our data at the micro level, and this is possible if we also have a measure of a variable at the macro level, or if we have only measures of micro-level variables. Inside the study of between-group differences the disaggregation brings a serious risk: to commit the I type error. This error is when, on the basis of the observations that there is a differences between individuals belonging to a particular group, the researcher asserts that there is no such relation. Instead inside the study of within-group differences, disaggregation often leads to unnecessarily conservative tests, with low type I error probabilities.

If we have measures only at the micro-level, analyzing data at the micro level is a correct way to proceed, also if it is important to take into account that the observations are within a macro-unit, and inside the macro-unit they can be correlated. Considering this aspect we are considering the sampling theory, and in particular the two stage samples.

The conclusion of this examination of the method is that a multilevel approach, where the within-group relations are combined, is more difficult to implement but much more productive. The main request is to specify assumptions about the way in which macro and micro effect are put together.

## 1.3 The intraclass correlation

The intraclass correlation coefficient represents the degree of similarity between micro-units belonging to the same macro-unit. We use the term 'class' to refer to the macro-units in the classification system under consideration. It is possible to define and to consider this coefficient in several ways, but we declare now our assumptions on the sampling design. We assume a two-stage sampling design, and infinite populations at either level. The groups are the macro-units. A relevant model is the *random effects ANOVA model*:

$$Y_{ij} = \mu + U_j + \epsilon_{ij} \tag{1.1}$$

Where $\mu$ is the population grand mean, $U_j$ is the specific effect of macro-unit $j$, and $\epsilon_{ij}$ is the residual effect for micro-unit $i$ within this macro unit. We can state that the true mean of the macro unit $j$ is $\mu + U_j$, and each measurement of a micro-unit within this macro-unit deviates from this true mean by some value called $R_{ij}$. The $U_j$ is a random variable, represents the effect of differences of the units from on another and gives the name 'random effects model'. It is assumed that all variables are independent, the group effects $U_i$ having population mean 0 and population variance $\sigma^2_{u_0}$, that is exactly the population between-group variance, and the residual having mean 0 and variance $\sigma^2_\epsilon$, that is the population within group variance. For example we can have as micro-units patients, and macro-units hospitals, then the within group variance is the variance within the hospitals about their true means, while the between-group variance is the variance between the hospitals' true means. This model is also known in the statistical literature as the one-way random effects ANOVA. The total variance of the model is given by the sum of these two variances:

$$var(Y_{ij}) = \sigma^2_{u_0} + \sigma^2_\epsilon.$$

We can write the number of micro - units within the j'th macro - unit with $n_j$, and the number of macro unit in $N$, moreover we have a sample size of $M = \sum_j n_j$. Given these conditions, the intraclass correlation coefficient $\rho_I$ is defined as:

$$\rho_I = \frac{\text{population variance between macro-units}}{\text{total variance}} = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_\epsilon} \tag{1.2}$$

This can be seen like a sort of proportion of variance, at group level. This parameter is called correlation coefficient, because it corresponds to the correlation between values of two randomly drawn micro-units in the same, randomly drawn, macro-unit. The population variance between macro-units is not directly reflected by the observed variance between the means of the macro-units. This is because in a two-stage sample, variation between micro-units will also show up as extra observed variance between macro-units.

At this point it is interesting to show how the intraclass correlation can be estimated and tested.

### Within-group and between-group variance

In this section we consider the macro-units groups. The principal goal is to discover all the information contained in the data about the population between-group variance and the population within -group variance, and for this goal we consider the observed variance of the two type of variance. First of all we provide below the functional form of the means that we need to utilize. The mean of the macro-unit $j$ is:

$$\bar{Y}_{.j} = \frac{1}{n_j} \sum Y_{ij}{}_{i=1}^{n_j}$$

The overall mean is:

$$\bar{Y}_{..} = \frac{1}{M}\sum_{j=1}^{N}\sum_{i=1}^{n_j} Y_{ij} = \frac{1}{M}\sum_{i=1}^{N} n_j \bar{Y}_{.j}.$$

At this point we can compute the observed variance within group j:

$$S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{.j})^2$$

This variance is different in each group. If we are interested in a single value that resume the within variance, we need to have only one parameter that expresses the within-group variability for all group jointly. For this task it is possible to use the observed within group variance, or pooled within group variance. This is a weighted average of the variance within the various macro-units:

$$\begin{aligned} &S^2_{within} = \tfrac{1}{M-N}\textstyle\sum_{j=1}^{N}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{.j})^2 \\ &= \tfrac{1}{M-N}\textstyle\sum_{j=1}^{N}(n_j - 1)S_j^2. \end{aligned} \tag{1.3}$$

If the model 1.1 holds we can affirm that the expected value of the observed within-group variance is exactly equal to the population within-group variance: Expected variance within= $\epsilon S^2_{within} = \sigma^2$. Going to the between-group variance arise some complications, because we have to take into account the group size. For equal group size $n_j$ there are no issues, the observed between-group variance is defined as the variance between the group means:

$$S^2_{between} = \frac{1}{N-1}\sum_{j=1}^{N}(\bar{Y}_{.j} - \bar{Y}_{..})^2 \tag{1.4}$$

But if we have unequal group size we have to take into account the contributions of the various groups, with different weights. We can use as weight:

$$\tilde{n} = \frac{1}{N-1}\left\{M - \frac{\sum_j n_j^2}{M}\right\} = \bar{n} - \frac{s^2(n_j)}{N\bar{n}}$$

where $\bar{n} = M/N$ is the mean sample size, and the variance of the sample size is:

$$s^2(n_j) = \frac{1}{N-1}\sum_{j=1}^{N} n_j - \bar{n}^2$$

The formula that uses weights useful for estimating the population between-group variance is the following:

$$S^2_{between} = \frac{1}{\tilde{n}(N-1)} \sum_{j=1}^{N} n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2. \tag{1.5}$$

The total observed variance is a combination of the within-group and the between-group variances, and we can express it in the following way:

$$observed\ total\ variance = \tfrac{1}{M-1} \textstyle\sum_{j=1}^{N} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2$$

$$= \tfrac{M-N}{M-1} S^2_{within} + \tfrac{\tilde{n}(N-1)}{M-1} S^2_{between}.$$

The complications for the between group variance arise because we have the contribution also at the micro-level of the residuals $R_{ij}$. The theory tells us that the expected between-group variance is:

$$Expected\ observed\ variance\ between =$$

$$true\ variance\ between + expected\ sampling\ error\ variance$$

For the case with constant $n_j$ the formula is:

$$\epsilon S^2_{between} = \sigma^2_{u_0} + \frac{\sigma^2_{\epsilon}}{\tilde{n}}$$

The second term of the formula becomes small when $\tilde{n}$ become large. So we can conclude that for large group sizes the expected observed between variance is practically equal to the true between variance. Instead when we have small group sizes it tend to be larger than the true between variance due to the random differences that also exist between the group means. It is clear that we do not know the population value of the between and within macro-units variances, we need to estimate them from the data. At this point we report the estimation formula without going into the details:

$$\hat{\sigma}^2_{\epsilon} = S^2_{within}$$

$$\hat{\sigma}^2_{u_0} = S^2_{between} - \frac{S^2_{within}}{\tilde{n}}$$

and finally also the estimated intraclass correlation:

$$\hat{\rho} = \frac{\hat{\sigma}^2_{u_0}}{\hat{\sigma}^2_{u_0} + \hat{\sigma}^2_{\epsilon}}$$

We are able to write the standard error of this estimator for the case where all the group sizes are constant, $n_j = n$:

$$S.E.(\hat{\rho}_I) = (1 - \rho_I)(1 + (n-1)\rho_I)\sqrt{\frac{2}{n(n-1)(N-1)}}. \tag{1.6}$$

The estimator that we have given above is the so-called analysis of variance estimators (ANOVA). They can be represented by specific formulas. We can also use maximum likelihood estimation (ML) and residual maximum likelihood (REML). If we have equal sample sizes, ANOVA and REML are the same; if we have unequal group sizes, ML and REML are slightly more efficient estimators than ANOVA. Multilevel software are able to compute ML and REML estimates.

### Utility of intraclass correlation coefficient

The intraclass correlation can assume two types of values: zero or positive. We now present a statistical test which is useful to assess if a positive value for this coefficient has to be attributed to chance. Often it can be possible to assume that the within-group deviations $R_{ij}$ are normally distributed, and in this case is possible to use an exact test for the hypothesis that the intraclass correlation is 0, which means that the between group variance is 0. We are speaking about the F-test for a group effect in the one-way analysis of variance (ANOVA). The test statistic is the following:

$$F = \frac{\tilde{n}S^2_{between}}{S^2_{within}}$$

and has a F-distribution, with $N - 1$ and $M - N$ degrees of freedom if the null hypothesis holds. In light of this consideration the intraclass correlation coefficient becomes:

$$\tilde{\rho} = \frac{F-1}{F + \tilde{n} - 1} \tag{1.7}$$

If we also have covariates it can be relevant to test whether there are group differences in addition to those accounted for by the effect of the covariates. Here we are referring to the analysis of the covariance (ANCOVA). The group effects do emerge when controlling for the covariates.

Finally in order to verify if the multilevel model is useful in a particular situation, and to test whether a given nesting structure in a data set calls for multilevel analysis, one can use standard techniques from the analysis of variance. How is it possible? If there is neither evidence for a main effect nor for interaction effects involving the group structure, then the researcher may leave aside the nesting structure and analyze the data by uni-level methods such as ordinary least squares regression analysis (OLS). Some problems can arise both when the group size is not too large or when we have too many groups, but the use of statistical multilevel software will help.

### Reassuming

In statistics, the intraclass correlation (or the intraclass correlation coefficient, abbreviated ICC) is a descriptive statistic that can be used when quantitative measurements are made on units that are organized into groups. It describes how strongly units in the same group resemble each other. While it is viewed as a type of correlation, unlike most other correlation measures it operates on data structured as groups, rather than data structured as paired observations.

The intraclass correlation is commonly used to quantify the degree to which individuals with a fixed degree of relatedness (e.g. full siblings) resemble each other in terms of a quantitative trait (see heritability). Another prominent application is the assessment of consistency or reproducibility of quantitative measurements made by different observers measuring the same quantity.

**Modern ICCs**

Starting with Ronald Fisher, the intraclass correlation has been regarded within the framework of analysis of variance (ANOVA), and more recently in the framework of random effects models. A number of ICC estimators have been proposed. Most of the estimators can be defined in terms of the random effects model

$$Y_{ij} = \mu + U_i + \epsilon_{ij},$$

where $Y_{ij}$ is the $j^{th}$ observation in the $i^{th}$ group, $\mu$ is an unobserved overall mean, $\alpha_i$ is an unobserved random effect shared by all values in group i, and $\epsilon_{ij}$ is an unobserved noise term. For the model to be identified, the $\alpha_i$ and $\epsilon_{ij}$ are assumed to have expected value zero and to be uncorrelated with each other. Also, the $\alpha_i$ are assumed to be identically distributed, and the $\epsilon_{ij}$ are assumed to be identically distributed. The variance of $\alpha_i$ is denoted $\sigma_\alpha$ and the variance of $\epsilon_{ij}$ is denoted $\sigma_\epsilon$.

The population ICC in this framework is

$$\frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_\epsilon^2}.$$

An advantage of the ANOVA framework is that different groups can have different numbers of data values, which is difficult to handle using the earlier ICC statistics. It should also be noted that this ICC is always non-negative, allowing it to be interpreted as the proportion of total variance that is *between groups*. This ICC can be generalized to allow for covariate effects, in which case the ICC is interpreted as capturing the within-class similarity of the covariate-adjusted data values.

A number of different ICC statistics have been proposed, not all of which estimate the same population parameter. There has been considerable debate about which ICC statistics are appropriate for a given use, since they may produce markedly different results for the same data.

## 1.4 Three-Levels models

Initially is not difficult to extend the two-level regression model to three or more levels. There is an outcome variable at first, the lowest level. In addition there may be explanatory variables at all higher levels. Problems arise for three and more level models which can become complicated very fast. In addition to the usual fixed regression coefficients, we must entertain the possibility that regression coefficients for first-level explanatory variables may vary across units of both the second and the third level. The possible way to explain such variation is to include cross-level interactions in the model. Regression slopes for the cross-level interaction between first-level and second-level variables may themselves vary across third-level units. In order to explain such variation we need a second-order interaction involving variables at all three levels.

The equations for such models can be complicated without the use of the compact summation notation. The resulting models are difficult to follow from a conceptual point of view, as well as difficult to estimate in practice. There are a considerable number of parameters to estimate, and at the same time the highest level sample size tends to get relatively smaller. So we can conclude that three and more level models have their place in multilevel analysis. Intuitively, three-level structures such as pupils in classes in schools, or respondents nested within households, nested within regions, appear to be both conceptually and empirically manageable. The idea that we want apply to our study is to consider patients at the first level, Hospitals at the second and time (years) at the third. With this approach we can analyze a panel and check the effects over time. If the lowest level is repeated measures over time, having repeated measures on pupils nested within schools, again do not appear to be overly difficult Hox and NetLibrary (2002). In such cases, the solution for the conceptual and statistical problems mentioned is to keep models reasonably small, and where it is not possible, to check for the meaningful of the model. Specification of the higher level variances and covariances should be driven by theoretical considerations. A higher-level variance for a specific regression coefficient implies that this regression coefficient is assumed to vary across units at that level. A higher-level covariance between two specific regression coefficients implies that these regression coefficients are assumed to covary across units at that level. Especially when models become large and complicated, it is suggested to avoid higher-order interactions and to only include in the random part those elements for which there is strong theoretical or empirical justification. This implies that an exhaustive search for second-order and higher-orders interactions is not a good approach. In general it is better to seek out higher-order interactions only if there is strong theoretical justification for their importance, or if an unusually large variance component for a regression slope calls for explanation. Random part of the model has more convincing theoretical reasons for the higher-level variance components than for the covariance components. If the covariances are small and insignificant, analysts sometimes do not include all possible covariance in the model. But is important to underline some exceptions: covariances between the intercept and the random slopes are always included; covariances corresponding to slopes of dummy-variables belonging to the same categorical variable have to be included and finally covariances for variables that are involved in an interaction or belong to the same polynomial expression.

### 1.4.1 Three-level intraclass correlations

In order to describe the intraclass correlation for the three-level multilevel models we reiterate that in a two-level model, the intraclass correlation is calculated in the intercept-only model using equation 1.2

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e}$$

The intraclass correlation is an indication of the proportion of variance at the second level, and it can also be interpreted as the expected correlation between two randomly chosen individuals within the same group. If we have a three-level model, for instance pupils nested within classes, nested within schools, there are several ways to calculate the intraclass correlation. The first step is to estimate an intercept-only model for the three level data, for which the single-equation model can be written as follows:

$$Y_{ijk} = \gamma_{000} + v_{0k} + u_{0jk} + e_{ijk} \tag{1.8}$$

The variances at the first, second, and third level are respectively $\sigma_e^2$, $\sigma_{u0}^2$, and $\sigma_{v0}^2$. For the three-level we have available two different methods, that are both correct. The first method (Davis and Scott (1995)) defines the intraclass correlations at the class and school level as:

$$\rho_{class} = \frac{\sigma_{u0}^2}{\sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2} \tag{1.9}$$

and

$$\rho_{school} = \frac{\sigma_{v0}^2}{\sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2} \tag{1.10}$$

This method identifies the proportion of variance at the class and school level. We can be interested in a decomposition of the variance across the available levels, or we can be interested in how much variance is explained at each level.

It is possible to also use a second method Siddiqui *et al.* (1996), that defines the intraclass correlations at the class and school level as:

$$\rho_{class} = \frac{\sigma_{v0}^2 + \sigma_{u0}^2}{\sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2} \tag{1.11}$$

and

$$\rho_{school} = \frac{\sigma_{v0}^2}{\sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2} \tag{1.12}$$

This method represents an estimate of the expected correlation between two randomly chosen elements in the same group. So the $\rho_{class}$ calculated in 1.11 is the expected correlation between two pupils within the same class, and it correctly takes into account that two pupils who are in the same class must also be in the same school. For this reason, the variance components for classes and schools must both be in the numerator of equation 1.11. If the two sets of estimates are different, which may happen if the amount of variance at the school level is large, there is no contradiction involved. We have described two different aspects of the data, which happen to coincide when there are only two levels. We will see in the Results' chapter the particular application to our data

## 1.5 Critique

Two streams of critique exist relative to the multilevel model. Exactly they are critique to the ranking and to the use of mortality rates to judge hospitals performance. In recent years the authors Lilford and Pronovost moved these two critiques to the use of the multilevel, it is incumbent to underline that they are not critiques to the model itself.

### 1.5.1 The use of outcomes

In their paper Lilford *et al.* (2004) criticize the use of outcomes to compare quality of care because this practice implies that the variation due to other causes can be accounted for, such that any residual variation truly indicates quality of care variation. They support the idea that outcomes are influenced by definitions, data quality, patient case-mix, clinical quality of care and chance. This assertion is certainly true, because it is possible to argue about all these types of influence and align them with that are all powerful. For instance ranking of hospitals depends upon the data source used, varying markedly, if the same outcome data were obtained from case records or from administrative data sets. But the authors of the critique strongly advocate that, even if an agreed risk-adjustment method could be derived, outcomes could still vary systematically between providers as one can never be certain that risk adjustment is not hampered by unmeasured prognostic factors. Outcome is neither a sensitive nor a specific marker for quality of care. Sanction and reward should not be applied to the 'worst' 5% of providers on outcome, because these will not be the 5% with the worst quality. We need performance measures which better reflect the quality of care. We would like to reach a continual non-judgment improvement structural factors, those that cannot easily be affected at the organizational level because they depend on the release of substantial resources or changes in policy. Monitoring clinical process has several advantages over outcome monitoring. Clinical process monitoring needs access to information which, although expensive, is likely to be much more cost effective than outcome monitoring process-based monitoring which is subject to potential bias due to the fact that the opportunity for error varies by case-mix. Some process measures are based on management data rather then adherence to clinical standards. These measures include waiting lists, ambulance response times and delays in accident and emergency departments. Such performance data are potentially useful for quality improvement but when used for performance management they often lead to a focus on changing the numbers rather than genuinely improving the systems. The crux of the performance management problem, however, is that you cannot know which of these factors is operating when outcomes differ. Showing people that their outcomes are worse than others does not automatically tell them what to do to improve or even whether they have a greater need than others to improve. They conclude that the use of outcome data and throughput to judge quality of care should be diligently avoided. We overcome these problems with analysis for the single DRG, because in this way we have a less generic correction. Moreover the Ranking is not to use like a judgment that leads to sanctions, but we shoud consider it like a signal that indicate what deepen.

### 1.5.2 The use of mortality rate

Another critique is the one to the use of mortality rates to judge hospital performance, Lilford and Pronovost (2010). We are conscious that differences in quality only explain a minimal part of the observed variance of mortality, and we overcome this problem with the use of also others outcome. Death is an outcome of care of undisputed importance, so mortality rates, especially overall hospital mortality rates, have therefore become the natural focus for the measurement of clinical quality. Usually it is possible to use risk adjustment techniques, that model preventable from inevitable deaths separately, but the authors sustain that is incorrect for two reasons. Firstly, risk adjustment can only adjust for factors that can be identified and measured accurately, Lilford *et al.* (2007). The error of attributing differences in risk adjusted mortality to differences in quality of care is the “case-mix adjustment fallacy”, Lilford *et al.* (2004). Secondly, risk adjustment can exaggerate the very bias that it is intended to reduce. This counterintuitive effect is called the “constant risk fallacy” and it arises when the risk associated with the variable on which adjustment is made varies across the units being compared, Nicholl (2007). Moreover, the mortality rates can be very different from hospital to hospital. The proposal that variance of this magnitude can be attributed to differences in the quality of care is not clinically intuitive and does not respect some standards. For example Mant and Hicks (Mant and Hicks (1995)) showed that differences in the quality of care could explain only half the observed variance in heart attack mortality. The famous Harvard malpractice study found that quality of care accounts for only a small proportion of the observed variance in mortality between hospitals. Little or no correlation exists between how well a hospital performs on one standard of safe and effective care and how well it performs on another; differences in the quality of care within hospitals are much greater than differences between hospitals, Jha *et al.* (2005). All these features lead to the affirmation that hospital mortality rates are a poor diagnostic test for quality and do not identify preventable deaths. Mortality rates are neutral, it is the use to which they are put that has moral salience and that will determine the balance of benefits and harms. The authors believe that the collection of mortality itself can be useful, but not using it like a criterion for performance management, or as the basis for the imposition of penalties. It is important to take into account other additional factors, because hospital mortality rates alone are silent about where problems may lie.

The correct use of mortality rates is as signal to identify where further investigation is necessary. But the investigation is also a sanction, so the use of mortality rate in this way is not neutral. But the use of mortality rate endure. This fact can be due to some decision makers, that believe mortality reflects quality. Clearly preventing people from dying is a positive result, but there also needs to be a search for robust measurements so as to avoid not fixing prematurely on a parameter that offers false hope. A possible solution to this problem should be to consider several outcomes (other than hospital mortality) and clinical processes. But a few outcome measures appear to be sensitive to quality. They ultimately assert that the use of mortality can be considered a sub-optimal solution.

In conclusion although the critique could be considered as a borderline source of proof, it reinforces the outcome we arrived at previously, namely the insertion of the mortality rate inside a multilevel model. This fact also leads us to consider the characteristics of the patients, and to take into account the condition of departure.

# Chapter 2

# The Stochastic Frontier

## 2.1 The stochastic frontier

The concept of stochastic frontier is introduced as a measure of the productive efficiency. Indeed typically the microeconomic models develop costs models and profit models, but not all the producers are successful in solving their optimization problems. The point is that not all the producers succeed in utilizing the minimum inputs required to produce the output they choose to produce, given the technology their disposal. We can say that not all the producers are technically efficient, we talk about technical efficiency. We can repeat the same speech for the allocation of costs (cost efficiency) and for the allocation of the output (not all producers are profit efficient). Results evident the failure of some producers in optimize, and it is desirable to recast the analysis of production, cost and profit going away from the traditional way, and always much more near the frontiers.

A production frontier is characterized by the minimum input bundles required, given the technology, to produce various outputs. In this way we can locate producers operating on their production frontier, that are labeled technically efficient, and we can also locate producers operating beneath their production frontier, that are labeled technically inefficient. The same discussion is for cost, revenue and profit. The minimum expenditure required to produce a given bundle of outputs is characterized by a dual cost frontier; producers operating on their cost frontier are labeled 'cost efficient'. The econometric implication of this re-formulation -from functions to frontiers-, is that symmetrically distributed error terms with zero means are no longer appropriate when analyzing producer behavior. It is still possible that the producer will end up above the deterministic kernel of an estimates production, revenue, cost and profit frontier, due to an unusually favorable operating environment. Nevertheless is much more probable that a producer will end up beneath an estimated production, revenue or profit frontier because there are two factors that work in this direction.

Firstly it is interesting to consider the environmental effects. Typically is assumed that they are random, in this case is possible to have an unfavorable operating environment like a favorable one, and this cause a producer to end up beneath an estimated production, revenue or profit frontier.

The second possibility that we take into account is the failure in optimizing. This also cause a producer to end up beneath an estimated production, revenue, or profit frontier. We can conclude from these observations that the error terms associated with frontiers are *composed* error terms, composed of a traditional symmetric random noise component and a new one-sided inefficiency component. These composed error terms are the focal point of the dissertation, because they bring the major change to the traditional considerations. They cannot be symmetric and they cannot have zero means, they must be skewed and they must have non zero means. This re-formulation make stochastic the production, cost, revenue and profit frontiers,due to random variation in the operating environment. Deviations from these stochastic frontiers are one-sided, due to various types of inefficiency. The error components are symmetric, and are designed to capture the effects of random variation and it is possible to conserve them in keeping with the older least squares-based approach to the estimation of production, cost, revenue, profit functions. So we introduce the one-sided error components designed to capture the effects of inefficiency, like a new econometric contribution to the estimation of production, cost, revenue and profit frontier. Consequently we will refer to this part of work like to the '*Stochastic Frontier Analysis*'.

## 2.2 Technical efficiency

At first we can make an excursus on several models of analysis that the stochastic frontier will use. The econometric analysis of productive efficiency is make exploring various econometric models designed to provide estimates of technical efficiency. The principal purpose of this section will be the estimation of technical efficiency, under the assumption that producers produce only a single output, both because they really produce a single output, and because it is possible to aggregate their multiple outputs into a single one.

Production frontiers provide the standards against which producer performance is evaluated, by means of an output oriented measure of technical efficiency. The innovative aspect of the frontiers, differently from the functions, is that frontiers explore stochastic distance functions. As first attempt we limit our discussion to single equation models. In these type of models the parameters describing the structure of a production frontier are estimated, and estimates of the output oriented technical efficiency of each producer are obtained like a byproduct of the exercise. The data are the observations on quantities of inputs employed and the output produced by each producer, without price information or behavioral objective on the producers. We have to discuss the estimation techniques, but they depend on the richness of the quantity of data available.

### 2.2.1 Cross-sectional production frontier models

Let's we start with cross sectional data. First of all a definition of cross sectional data : Cross-sectional data or cross section (of a study population) in statistics and econometrics is a type of one-dimensional data set. cross-sectional data refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects.

In our case we assume that cross-sectional data on the quantities of N inputs used to produce a single output are available for each of $I$ producers. We obtain the following production frontier model:

$$y_i = f(x_i;\beta)TE_i \tag{2.1}$$

with $y_i$ scalar output of producer $i$, $i = 1,...,I$; $x_i$, a vector of $N$ inputs used by producer $i$; $f(x_i;\beta)$, production frontier; $\beta$ a vector of technology parameters to be estimated; finally $TE_i$ is the output oriented technical efficiency of a producer $i$. Since we are interested in technical efficiency, we can obtain it from the previous equation:

$$TE_i = \frac{y_i}{f(x_i;\beta)} \tag{2.2}$$

which define the technical efficiency as the ratio of observed output to maximum feasible output. $y_i$ is the observed output, and achieves its maximum feasible value of $f(x_i;\beta)$ if, and only if, $TE_i = 1$. If this value is different, for example $TE_i < 1$, we have a decrease of the observed output from maximum feasible output. The definition of the production frontier $f(x_i;\beta)$ is such way is the definition of a *deterministic* production frontier. In this case if we have a shortfall of observed output $y_i$ from maximum feasible output $f(x_i;\beta)$, it is attributed to technical inefficiency. This kind of specification ignores the fact that output can be affected by random shocks that are not under the control of a producer. A clear example can be that a farmer cannot control the rain! So it is interesting to incorporate producer-specific random shocks into the analysis, and this process requires the specification of a stochastic production frontier. So we can update our equation in the following way:

$$y_i = f(x_i;\beta)exp\{v_i\}TE_i \tag{2.3}$$

where $[f(x_i;\beta)exp\{v_i\}]$ is the stochastic production frontier. The first part is the deterministic part, common to all producers; while the $exp\{v_i\}$ is a producer-specific part, which captures the effect of random shocks on each producer.

If the production frontier is specified as being *stochastic*, the equation assume the following form:

$$TE_i = \frac{y_i}{f(x_i;\beta)exp\{v_i\}} \tag{2.4}$$

that represent the technical efficiency as the ratio of observed output to maximum feasible output in an environment characterized by $exp\{v_i\}$. In this case the $y_i$ assume the maximum feasible value of $[f(x_i;\beta)exp\{v_i\}]$ if, and only if, $TE_i = 1$, and we can notice again a shortfall of observed output with $TE_i < 1$, in an environment characterized by $exp\{v_i\}$, which vary across producers.

The estimation of the technical efficiency can be made using both the deterministic production frontier model and the stochastic production frontier model. Obviously is preferred the last one because takes into account the effect of random shocks. However also the deterministic production frontier is wide spread.

### Deterministic production frontiers

$$y_i = f(x_i;\beta)exp\{-u_i\} \tag{2.5}$$

in this formulation the $TE_i$ is equal to $exp\{-u_i\}$. We have chanced the name of the variable because we require that $TE_i \leq 1$, and this can happen in the exponential function when we are in the negative quadrant so it has to be that $u_i \geq 0$.

Now we assume a particular form for the generic function $f(x_i;\beta)$, that for easy of calculate, can be the Cobb-Douglas form. In this case the deterministic production frontier will become

$$lny_i = \beta_0 + \sum_n \beta_n lnx_{ni} - u_i \tag{2.6}$$

where the condition of $u_i \geq 0$ guarantees that $y_i \leq f(x_i; \beta)$. The previous equation is a linear regression model with a non positive disturbance. We want to obtain the estimates of the parameters vector $\beta$, which contain the description of the structure of the production frontier, and we want also to obtain the estimates of the $u_i$. These last estimate are used to obtain the estimates of $TE_i$ for each producer by means of $TE_i = exp\{-u_i\}$. We can use several method of estimation, but in all cases it must somehow incorporate the restriction $u_i \geq 0$.

At this point three methods of estimation are proposed.

**Goal Programming** is an optimization program. It can be thought as an extension or generalization of linear programming to handle multiple, normally conflicting objective measures. Each of these measures is given a goal or target value to be achieved. Unwanted deviations from this set of target values are then minimized in an achievement function. This can be a vector or a weighted sum dependent on the goal programming variant used. As satisfaction of the target is deemed to satisfy the decision makers, an underlying satisfaction philosophy is assumed.

Aigner and Chu (1968) showed that the deterministic production frontier model can be converted to either of a pair of mathematical programming models. The model is a *linear* programming model , in which the goal is to calculate a parameter vector $\beta$ for which the sum of the proportionate deviations of the observed output of each producer beneath maximum feasible output is minimized. The subsequent step is the conversion of the deviations into measures of technical efficiency for each producer. Linear programming model:

$$min \sum_i u_i \tag{2.7}$$

subject to

$$[\beta_0 + \sum_n \beta_n lnx_{ni}] \geq lny_i, i = 1, ..., I \tag{2.8}$$

We can later on consider another type of goal programming, that is the *quadratic* programming model. In this type of model the goal is to calculate a parameter vector $\beta$ for which the sum of squared proportionate deviations of the observed output of each producer beneath maximum feasible output is minimized. The *quadratic* programming model:

$$min \sum_i u_i^2 \tag{2.9}$$

subject to

$$[\beta_0 + \sum_n \beta_n lnx_{ni}] \geq lny_i, i = 1, ..., I \tag{2.10}$$

The constraints that are appended to either model are non negativity constraints on the parameters $\beta_n$, n=1, ..., N. After the computation of the parameter values, the technical efficiency of each producer can be calculated from the slacks in the functional constraints.

The most problematic drawback of the goal programming is that the parameters are *calculated* rather than *estimated*, which create problems with statistical inference. It is possible solve this problem, and give a statistical interpretation to the goal programming models, if a distributional assumption is imposed on the $u_i$.

The estimates are maximum likelihood estimates of the parameters of the deterministic production frontier if the $u_i \geq 0$ follow an exponential distribution:

$$f(u) = \frac{1}{\sigma_u} exp\{-\frac{u}{\sigma_u}\} \tag{2.11}$$

than we can calculate the log likelihood

$$lnL = -Iln\sigma_u - \frac{1}{\sigma_u}\sum_i |u_i| \tag{2.12}$$

where $I$ is the total number of entrepreneurs. The exponential distribution is a single-parameter distribution, and it is easily portrayed graphically.

The estimates of the quadratic programming are maximum likelihood estimates of the parameters of the deterministic production frontier if the $u_i \geq 0$ follow a half normal distribution

$$f(u) = \frac{2}{\sqrt{2\pi\sigma_u}} exp\{-\frac{u^2}{2\sigma_u^2}\} \tag{2.13}$$

than we can calculate the log likelihood distribution also in this case:

$$lnL = constant - \frac{1}{2}ln_u^2 - \frac{1}{2\sigma_u^2}\sum_i u_i^2 \tag{2.14}$$

Also the half normal distribution is a single-parameter distribution. The values computed with the goal programming techniques are calculated rather than estimated like we have said prematurely, so the parameter vector $\beta$ do not come with standard error attached. The solution is to establish a linkage between production frontiers calculated by goal programming methods and production frontiers estimated by maximum likelihood methods. This connection supply the statistical foundation that we are looking for.

Schmidt noted that the statistical properties of the maximum likelihood estimators cannot be obtained by traditional methods, due to the dependence of $lny_i$ on $\beta$, which violates one of the regularity conditions for maximum likelihood estimation.

Greene observed in a second instance that the Hessian of the log-likelihood functions are singular under both exponential and half-normal distributions, and this make impossible to estimate the precision of the maximum likelihood estimators using an Hessian. So Greene suggested the use of an alternative model, in which $u_i \geq 0$ follows a gamma distribution, and which satisfies all regularity conditions for obtaining asymptotic properties of the maximum likelihood estimators. This is a possible solution, especially if one usually use MLE, but the real problem is that there are no goal programming problems that have the inefficiency distributed like a gamma. Finally these two goal programming problems have two different drawbacks: if they have known MLE counterparts, they also have statistical properties; whereas if the MLE problem has desirable statistical properties, it can also has no known goal programming counterpart.

**Corrected Ordinary Least Squares (COLS)**

A possible way of estimation of the deterministic production frontier model is suggested by Farrell (1957): it could be an estimation in two steps. In the first step ordinary least square (OLS) is used to obtain consistent unbiased estimates of the slope parameters, and a consistent but biased estimate of the intercept parameter. In the second step the biased OLS intercept $\beta_0$ is corrected shifting up. This correction is made to ensure that the estimated frontier bounds the data from above. At this point the COLS intercept can be estimated in a consistent way, and the estimation can be written in the following way:

$$\hat{\beta}_0^* = \hat{\beta}_0 + max_i\{\hat{u}_i\} \tag{2.15}$$

where the $\hat{u}_i$ are the OLS residuals and the $\hat{\beta}_0$ are the OLS estimation. The intercept is moved up, whereas the new residual are modified in the opposite direction.

$$\hat{u_i^*} = max\{\hat{u}_i\} - \hat{u}_i \tag{2.16}$$

$$-\hat{u_i^*} = \hat{u}_i - max\{\hat{u}_i\} \tag{2.17}$$

The COLS residual $\hat{u_i^*}$ are nonnegative, with at least one being zero, and can be used to provide consistent estimates of the technical efficiency of each producer by means of $TE_i = exp\{-\hat{u_i^*}\}$. This technique is of easy computation, and generates an estimated production frontier that lies on or above the data. The cons of this easy of calculation is that the estimated production frontier is parallel to the OLS regression, because we correct only the intercept. This is a restriction, because in this way the structure of the 'best practice' production technology is the same as the structure of the 'central tendency' production technology. This is an undesirable property of the COLS procedure, because the structure of best practice production technology is permitted to differ from that of production technology just in the middle of the data, where producers are less efficient than best practice producers. Moreover the COLS frontier is not as closely as possible to the data, since it is required to be parallel to OLS regression.

### Modified Ordinary Least Squares (MOLS)

It can be interesting to take into account a variation in the COLS. This variation has been proposed by Afriat (1972) and Richmond (1974). They suggested to estimate the production frontier model with OLS, but this time under the assumption that the disturbances follow an explicit one-sided distribution, such as exponential or half-normal. It is reasonable to expect that the technical efficiency follows one of these distribution. The MOLS procedure is identical to COLS one, with the unique difference that after the OLS, the estimated intercept is shifted up by the mean of the assumed one-sided distribution , instead of a shifting with the maximum value. The equation become:

$$\hat{\beta_0^*}* = \hat{\beta}_0 + E(\hat{u}_i) \tag{2.18}$$

and

$$-\hat{u_i^*}* = \hat{u}_i - E(\hat{u}_i) \tag{2.19}$$

and the OLS residual are used also here to provide consistent estimates of the technical efficiency of each producer, exactly like in the COLS. The procedure is so similar that also in this case is easy to implement. The cons are that there are no guarantee that the modification of OLS shifts the estimated intercept up by enough to ensure that all producers are really bounded from above by the estimated production frontier, since it is possible that $[\hat{u}_i - E(\hat{u}_i)] > 0$ for a producer. In this case you have to explain a technical efficiency score greater than unity. Another possibility is that MOLS shift the estimated intercept so far up that no producer is technically efficient.

The last cons that we consider is that also in the MOLS case the production frontier is parallel to the OLS regression, since also in this case we modify only the intercept. The discussion is identical to the COLS case.

All these three techniques share the virtue of simplicity. But again all these three techniques have a great deficiency: they all measure technical efficiency relative to a deterministic production frontier. If we have a variation in output not associated with variation in inputs, it will be attribute to technical inefficiency. None of these techniques considers the possibility that there could be random shocks, that might contribute to variation in output. On the other hand OLS estimation attribute all variation in output not associated with variation in inputs, to random shocks, without consider the technical inefficiency. Finally what is required is a model that is able to attribute variations in output not associated with inputs both to random shocks and to technical inefficiency, a combination.

### Stochastic Production Frontiers

The stochastic production frontier models are introduced simultaneously in Aigner and Lovell (1977) and by Meeusen and van Den Broeck (1977). The pioneering work of Farrell (1957) took in serious consideration the possibility of estimating so-called frontier production functions, in an effort to bridge the gap between theory and empirical work. But his attempt have not been completely successful,so Aigner and Lovell (1977) and Meeusen and van Den Broeck (1977) suggested a new approach to the estimation of frontier production functions, that allow for technical inefficiency, but also consider the fact that exist random shocks outside the control of producers.

In the stochastic production frontier models the impact on output of shocks due to variation in labor and machinery performance, changing of the weather and also simply luck, can be separated from the contribution of variation in technical efficiency. Also in this section we assume that the function $f(x_i, \beta)$ tales the log-linear Cobb Douglas form, and the stochastic production frontier model can be written in the following way:

$$lny_i = \beta_0 + \sum \beta_n lnx_{ni} + v_i - u_i \tag{2.20}$$

how we can notice the error term is composed by two components, the $v_i$ represent the two-sided “noise”component, and $u_i$ is the nonnegative technical inefficiency component. For this reason usually the stochastic production frontier model is referred as a “composed error” model. Let’s comment on the features of the error term: $v_i$ is assumed to be *iid* and symmetric, distributed independently from $u_i$. The global error term $\epsilon_i = v_i - u_i$ is asymmetric, because $u_i \geq 0$. Moreover we can assume without lose of meaning that $v_i$ and $u_i$ are distributed independently of $x_i$.

At this point arise a problem: with the OLS estimation we obtain consistent estimates of the $\beta_n s$ but not of $\beta_0$ since $E(\epsilon_i) = -E(u_i) \leq 0$, and also we notice that OLS does not provide estimates of producer-specific technical efficiency. Nevertheless OLS provide a simple test for the presence of technical inefficiency in the data. This test consist in checking if $u_i > 0$, and in this case the global error $\epsilon = v_i - u_i$ is negatively skewed, and there is evidence of technical inefficiency in the data. This observation is useful, since conduct to construct a test for the presence of technical inefficiency based directly on the OLS residuals. Schmidt and Lin (1984) proposed a test statistic:

$$(b_1)^{1/2} = \frac{m_3}{(m_2^{3/2})} \tag{2.21}$$

where $m_2$ and $m_3$ are the second and third sample moments of the OLS residuals. It is possible to notice that $m_3$ is simply the third sample moment of the $u_i$, because $v_i$ is symmetrically distributed. Under the light of this consideration if $m_3$ has a negative value $m_3 < 0$, than the OLS residuals will be negatively skewed, and would suggest the presence of technical inefficiency. If the $m_3$ assume a positive value $m_3 > 0$, the OLS residuals will be positively skewed, but this case make no sense in this context, in a certain sense we can think that positive skewness in the OLS residuals provides an indication that the model is mispecified. The distribution of $(b_1^{1/2})$ is not so much used, thus Coelli (1995) proposed an alternative test statistic that has the characteristic of being asymptotically distributed as $N(0,1)$. Resume, negative skewness occurs when $m_3 < 3$, a test of the hypothesis that $m_3 \geq 0$ is appropriate. Considering the null hypothesis of zero skewness of the errors, the test statistic $m_3/(6m_2^3/I)^{1/2}$ is asymptotically distributed as $N(0,1)$. These test are based on simple OLS results, and it is a strong advantage, but they are also based on asymptotic theory, that is a disadvantage because usually the samples are relatively small. Fortunately Coelli (1995) presented really encouraging Monte Carlo results concerning the power of his OLS-based test. Under these considerations we assume that there is negative skewness in the OLS residuals, since this consideration bring evidence of technical inefficiency in the data, and is does make sense to proceed to the estimation of a stochastic production frontier. The objectives that we want to achieve are mainly two: to obtain estimates of the production technology parameters $\beta$ in $f(x;\beta)$; and to obtain estimates of the technical efficiency of each producer. A problem arise, it is necessary to know the two separated estimates of statistical noise $v_i$ and technical inefficiency $u_i$, instead of the global error $\epsilon_i$ for each producer that we are able to know. We can assume that the $u_i$ are distributed independently from the inputs, and OLS provides consistent estimates of all production technology parameters except for the intercept. But is not enough. Additional assumption are required, and also a different estimation technique, in order to obtain a consistent estimate of the intercept and estimates of the technical efficiency of each producer. The firs method that we consider is a maximum likelihood method that can be used to estimate $\beta$ and the $u_i$. This method derive from the two-step procedure that we have previously described, in which the first step involves the use of OLS to estimate the slope parameters, and the second step involves the use of maximum likelihood to estimate the intercept and the variances of the two error components. The assumptions that we can make conduct to several models.

**The Normal-Half Normal Model**

Let's we start introducing the distributional assumptions that we make using this model:

- $v_i \sim iidN(0, \sigma_v^2)$, conventional assumption
- $u_i \sim N^+(0, \sigma_u^2)$, nonnegative half normal, based on the plausible proposition that the modal value of $TE$ is equal to zero, and increasing values of technical inefficiency becoming increasingly less likely.
- $v_i$ and $u_i$ are distribute independently of each other, and of the regressors. This assumption is crucial, because if producers know something about their technical efficiency, this may influence their choice of inputs.

In order to compute the density function of the global error $\epsilon$, we remember the functional form of the density function of $u \geq 0$.
The density function of $u_i \geq 0$ is:

$$f(u) = \frac{2}{\sqrt{2\pi}\sigma_u} exp\left\{-\frac{u^2}{2\sigma_u^2}\right\} \tag{2.22}$$

The density function of $v$ is:

$$f(v) = \frac{2}{\sqrt{2\pi}\sigma_v} exp\left\{-\frac{v^2}{2\sigma_v^2}\right\} \tag{2.23}$$

We can obtain the form of the joint density function of $u$ and $v$ remembering that the independence assumption allow to make the product of the individual density functions:

$$f(u, v) = \frac{2}{2\pi\sigma_u\sigma_v} exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\} \tag{2.24}$$

We know that $\epsilon = v - u$, so it can be useful to rewrite the conjoint density function substituting this quantity, for example we can get the value $v = \epsilon + u$ and:

$$f(u,\epsilon) = \frac{2}{2\pi\sigma_u\sigma_v} exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{(\epsilon+u)^2}{2\sigma_v^2}\right\} \tag{2.25}$$

At this point we can recall some statistical property and obtain the marginal density function integrating $u$ out of $f(u,\epsilon)$:

$$f(\epsilon) = \textstyle\int_0^\infty f(u,\epsilon)du$$

$$= \tfrac{2}{\sqrt{2\pi}\sigma}\left[1 - \Phi(\tfrac{\epsilon\lambda}{\sigma})\right] exp\{-\tfrac{\epsilon^2}{2\sigma^2}\}$$

$$= \tfrac{2}{\sigma}\phi(\tfrac{\epsilon}{\sigma})\Phi(-\tfrac{\epsilon\lambda}{\sigma})$$

We have applied a convenient re-parameterization, the value of $\sigma$ is $\sigma = (\sigma_u^2 + \sigma_v^2)$, $\lambda = \sigma_u/\sigma_v$, and $\Phi(.)$ and $\phi(.)$ are the standard normal cumulative distribution and density functions. In this way $\lambda$ gives indications about the relative contribution of $u$ e $v$ to $\epsilon$. As $\lambda \rightarrow 0$ , either $\sigma_v^2 \rightarrow \infty$ or $\sigma_u^2 \rightarrow 0$, and the symmetric error component dominates the one-sided error component in the determination of $\epsilon$. In this case we are again in an OLS production function model with no technical inefficiency. As $\lambda \rightarrow \infty$ , either $\sigma_u^2 \rightarrow \infty$ or $\sigma_v^2 \rightarrow 0$ one-sided error component dominates the symmetric error component in the determination of $\epsilon$. This time we are in the case of a deterministic production frontier model with no noise. The normal-half normal models contains two different parameters, that can be $\sigma_u$ and $\sigma_v$, but can also be $\sigma$ and $\lambda$.

The distribution parameters $\sigma$ and $\lambda$ are estimated along with the technology parameters $\beta$. It is desirable to conduct a statistical test of the hypothesis that $\lambda = 0$, based on the maximum likelihood estimate of $\lambda$. The first test proposed is the Wald test, the second one a likelihood ratio test. But it is important to notice that the value of $\lambda$ lies on the boundary of the parameter space, and this made it difficult to interpret the test statistic. Coelli has shown has found a different test, and has shown that in this case is appropriate the utilize of a one-sided likelihood test statistic that is asymptotically distributed as a mixture of $\chi^2$ distribution rather than as a single distribution. The marginal density function $f(\epsilon)$ is asymmetrically distributed, with mean and variance

$$E(\epsilon) = -E(u) = -\sigma_u\sqrt{\frac{2}{\pi}} \tag{2.26}$$

$$V(\epsilon) = \frac{\pi - 2}{\pi}\sigma_u^2 + \sigma_v^2. \tag{2.27}$$

Considering the various attempts of the researchers, we can assert that Aigner and Lovell (1977) suggested $[1 - E(u)]$ as estimator of the mean technical efficiency of all producers. Instead Lee and Tyler (1978) proposed a more complex form for the mean:

$$E(exp\{-u\}) = 2[1 - \Phi(\sigma_u)]exp\left\{\frac{\sigma_u^2}{2}\right\} \tag{2.28}$$

that is a better form because $[1-u]$ includes only the first term in the power series expansion of $exp\{-u\}$, and the proposal of Lee and Tyler is consistent with definition of technical efficiency. Using the marginal density function of $\eta$, we can write the log likelihood function for a sample of $I$ producers:

$$lnL = constant - Iln\sigma + \sum_i ln\Phi\left(-\frac{\epsilon_i\lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_i \epsilon_i^2 \tag{2.29}$$

the previous formulation can be maximized with respect to the parameters to obtain maximum likelihood estimates of all parameters. These estimates are consistent as $I \to \infty$. At this point is desirable to be able to obtain the estimates of the technical efficiency of each producer. We can make some considerations on the estimates of $\epsilon_i = v_i - u_i$ that we have, and this quantity contain information on $u_i$. If $\epsilon_i > 0$, chances are that $u_i$ is not large, and consequently the producer is relatively efficient. If $\epsilon_i < 0$ chances are that $u_i$ is large, which suggests that this producer is relatively inefficient. Arise a problem: to extract the information that $\epsilon_i$ contains on $u_i$. An helpful consideration is take into account the conditional distribution of $u_i$ given $\epsilon_i$, which contains whatever information $\epsilon_i$ contains concerning $u_i$. Jondrow *et al.* (1982) have considered the conditional distribution of $u$ given $\epsilon$ if $u_i \sim N^+(0, \sigma_u^2)$, and assume the following form:

$$f(u|\epsilon) = \frac{f(u,\epsilon)}{f(\epsilon)}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_*}exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}/[1 - \Phi(-\frac{\mu_*}{\sigma_*})]$$

where we have to explain the content of the letter with the '*', $\mu_* = -\epsilon\sigma_u^2/\sigma^2$ and $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2$. $f(u|\epsilon)$ is distributed as a $N^+(\mu_*, \sigma_*^2)$ also the mean and the mode distribution are useful to estimate the $u_i$.

$$E(u_i|\epsilon_i) = \mu_{*i} + \sigma_*\left[\frac{\phi(-\mu_{*i}/\sigma_*)}{1-\Phi(-\mu_{*i}/\sigma_*)}\right]$$

$$= \sigma_*\left[\frac{\phi(\epsilon_i\lambda/\sigma)}{1-\Phi(\epsilon_i\lambda/\sigma)} - \left(\frac{\epsilon_i\lambda}{\sigma}\right)\right]$$

and

$$M(u_i|\epsilon_i) = \begin{cases} -\epsilon_i\left(\frac{\sigma_u^2}{\sigma}\right) & \text{if } \epsilon_i \leq 0 \\ \\ 0 & otherwise \end{cases}$$

The use of the mean value is more frequent than the mode. Once we obtain the estimates of $u_i$, the estimates of technical efficiency of each producer can be obtained from:

$$TE_i = exp\{-\hat{u}_i\}. \tag{2.30}$$

where $-\hat{u}_i$ can be both $E(u_i|\epsilon_i)$ and $M(u_i|\epsilon_i)$. It is widespread a lot the proposal of Battese and Coelli ):

$$TE_i = E(exp{-}u_i|\epsilon_i) = \left[\frac{1-\Phi(\sigma_* - \mu_*/\sigma_*)}{1-\Phi(-\mu_*/\sigma_*)}\right] exp\left\{-\mu_* + \frac{1}{2}\sigma_*^2\right\} \tag{2.31}$$

We can chose several method of estimation but in each case the estimates of technical efficiency are inconsistent because the variation associated with the distribution of $(u_i|\epsilon_i)$ is independent from $i$. It is possible to counter this problem obtaining confidence intervals for the point estimates of technical efficiency. Horrace and Schmidt (1996) have identified upper and lower bounds on $(u_i|\epsilon_i)$, and also many others researchers have worked on this subject. Finally they found negative bias in the estimated inefficiencies, and a mean empirical coverage accuracy of the confidence intervals to be significantly below the corresponding theoretical confidence levels for all values of $\lambda$ and for sample size less than 200. So we have based our analysis of stochastic production frontiers on the assumption that $u \sim N^+(0, \sigma_u^2)$, since this distributional assumption is both plausible and tractable.

**The Normal-Exponential Model**

We now make some different distributional assumptions:

- $v_i \sim$ iid $N(0, \sigma_v^2)$
- $u_i \sim$ iid exponential
- $u_i$ and $v_i$ are distributed independently of each other, and of the regressors.

The remarks are the same that for the normal-half normal model. We have already mentioned the density functions for $u_i$ and for $v_i$, and as consequence of the independence assumption, the joint density function of $u$ and $v$ is the product of their individual density functions:

$$f(u,v) = \frac{1}{\sqrt{2\pi}\sigma_u \sigma_v} exp\left\{ -\frac{u}{\sigma_u} - \frac{v^2}{2\sigma_v^2} \right\} \tag{2.32}$$

Joint density function of $u$ and $\epsilon$:

$$f(u,\epsilon) = \frac{1}{\sqrt{2\pi}\sigma_u \sigma_v} exp\left\{ -\frac{u}{\sigma_u} - \frac{(u+\epsilon)^2}{2\sigma_v^2} \right\} \tag{2.33}$$

Finally the marginal density function of $\epsilon$ is:

$$f(\epsilon) = \int_0^\infty f(u,\epsilon)du = \left(\frac{1}{\sigma_u}\right) \Phi \left( -\frac{\epsilon}{\sigma_v} - \frac{\sigma_v}{\sigma_u} \right) exp\left\{ \frac{\epsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} \right\} \tag{2.34}$$

The marginal density function $f(\epsilon)$ is asymmetrically distributed, with mean and variance

$$E(\epsilon) = -E(u) = -\sigma_u$$

$$V(\epsilon) = \sigma_u^2 + \sigma_v^2.$$

As $\sigma_u/\sigma_v$ increases, the distribution looks more and more like a negative exponential distribution, as $\sigma_v/\sigma_u$ increases, the distribution looks more and more like a normal distribution.

We can write the log likelihood function for a sample of $I$ producers in the following way:

$$lnL = constant - Iln\sigma_u + I\left(\frac{\sigma_v^2}{2\sigma_u^2}\right) + \sum_i ln\Phi\left(-A\right) + \sum_i \frac{\epsilon_i}{\sigma_u} \tag{2.35}$$

where $A = -\tilde{\mu}/\sigma_v$ and $\tilde{\mu} = -\epsilon - (\sigma_v^2/\sigma_u)$

In this model, point estimates of technical efficiency can be obtained from either the mean or the mode of the conditional distribution of $u$ given $\epsilon$. These estimates will be unbiased but not consistent.

**The Normal-Truncated Normal Model**

It is a generalization of the Normal-half normal model, where we allow $u$ to follow a truncated normal distribution. In this case the distributional assumptions will be:

- $v_i \sim$ iid $N(0,\sigma_v^2)$

- $u_i \sim$ iid $N^+(\mu, \sigma_u^2)$
- $u_i$ and $v_i$ are distributed independently of each other, and of the regressors.

The considerations are the same that for the normal-half normal, with the unique exception that the truncated normal distribution assumed for $u$ generalizes the one-parameter half normal distribution, by allowing the normal distribution, which is truncated below zero, to have a nonzero mode. The truncated normal distribution contains an additional parameter $\mu$ to be estimated, and for this reason is a more flexible representation of the data. The density function of $v$ is the same that in the other cases, while the density function of the truncated normal is:

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma_u \Phi(-\mu/\sigma_u)} exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2}\right\} \tag{2.36}$$

$\mu$ is the mode of the normal distribution, truncated below zero; $\Phi(.)$ is the standard normal cumulative distribution function. Definitely $f(u)$ is the density function of a normally distributed variable, with possibly nonzero mean $\mu$, truncated below at zero.. If $\mu = 0$ the density function collapses to the half normal density function. The truncated normal distribution is a two parameter distribution depending on placement and spread parameters $\mu$ and $\sigma_u$. The way of proceed is the same that for the previous models.

We have to write a joint density function of $u$ and $v$, that is the product of their individual density functions:

$$f(u,v) = \frac{1}{2\pi\sigma_u\sigma_v\Phi(-\mu/\sigma_u)} exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\}. \tag{2.37}$$

We continue writing the joint density function of $u$ and $\epsilon$:

$$f(u,\epsilon) = \frac{1}{2\pi\sigma_u\sigma_v\Phi(-\mu/\sigma_u)} exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2} - \frac{(\epsilon+u)^2}{2\sigma_v^2}\right\}. \tag{2.38}$$

The marginal density function of $\epsilon$:

$$\begin{aligned} f(\epsilon) &= \int_0^\infty f(u,\epsilon)du \\ &= \frac{1}{\sqrt{2\pi}\sigma\Phi(-\mu/\sigma_u)} \Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\epsilon\lambda}{\sigma}\right) exp\left\{-\frac{(\epsilon+\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sigma}\phi\left(\frac{\epsilon+\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\epsilon\lambda}{\sigma}\right)\left[\Phi\left(-\frac{\mu}{\sigma_u}\right)\right]^{-1} \end{aligned} \tag{2.39}$$

where $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$ and $\lambda = \sigma_u/\sigma_v$ as in the normal-half model, and $\phi()$ is the standard normal density function. If $\mu = 0$ the previous marginal density of $\epsilon$ collapses to the half normal marginal density function. The normal-truncated normal distribution has three parameters, a placement parameter $\mu$ and two spread parameters $\sigma_u$ and $\sigma_v$.

The log likelihood function for a sample of $I$ producers is:

$$lnL = constant - Iln\sigma - Iln\Phi\left(-\frac{\mu}{\sigma_u}\right) + \sum_i ln\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\epsilon_i\lambda}{\sigma}\right) - \frac{1}{2}\sum_i\left(\frac{\epsilon_i + \mu}{\sigma}\right)^2 \quad (2.40)$$

where $\sigma_u = \lambda\sigma/\sqrt{1+\lambda^2}$. In order to obtain the estimates of all the parameters we can maximize the log likelihood function with respect to the parameters.

Now it can be interesting to proceed with the description of the use of conditional distribution, since we will use it several times and for others models but in the same way. The conditional distribution $f(u|\epsilon)$ is:

$$f(u|\epsilon) = \frac{f(u|\epsilon)}{f(\epsilon)} = \frac{1}{\sqrt{2\pi}\sigma_*\left[1 - \Phi\left(-\tilde{\mu}/\sigma_*\right)\right]} exp\left\{-\frac{\left(u - \tilde{\mu}\right)^2}{2\sigma_*^2}\right\} \quad (2.41)$$

we can notice that $f(u|\epsilon)$ is distributed as $N^+\left(\tilde{\mu}_i, \sigma_*^2\right)$, where the specification of the parameters are: $\tilde{\mu}_i = \left(-\sigma_u^2\epsilon_i + \mu\sigma_v^2\right)/\sigma^2$ and $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2$. At this point is possible to use both the mean or the mode of $f(u|\epsilon)$ to estimate technical efficiency of each producer. We have:

$$E(u_i|\epsilon_i) = \sigma_*\left[\frac{\tilde{\mu}_i}{\sigma_*} + \frac{\phi(\tilde{\mu}_i/\sigma_*)}{1 - \Phi(-\tilde{\mu}_i/\sigma_*)}\right] \quad (2.42)$$

and

$$M(u_i|\epsilon_i) = \begin{cases} \tilde{\mu}_i & \text{if } \tilde{\mu}_i \geq 0 \\ 0 & otherwise \end{cases} \quad (2.43)$$

We can obtain the technical efficiency by substituting either $E(u_i|\epsilon_i)$ or $M(u_i|\epsilon_i)$ into equation (2.31), or utilizing:

$$\begin{aligned} TE_i &= E(exp\left\{-u_i|\epsilon_i\right\}|\epsilon_i) \\ &= \tfrac{1-\Phi[\sigma_* - (\tilde{\mu}_i/\sigma_*)]}{1-\Phi(-\tilde{\mu}_i/\sigma_*)} exp\left\{-\tilde{\mu}_i + \tfrac{1}{2}\sigma_*^2\right\}, \end{aligned} \quad (2.44)$$

which is the same thing that (2.31) if $\mu = 0$. The use of (2.31)and (2.44) produces unbiased but inconsistent estimates of technical efficiency.

### The Normal Gamma Model

We can consider also a generalization of the normal-exponential model, assuming the $u$ follows a gamma distribution. The normal gamma formulation was introduced by Greene (1980) and Stevenson (1980) and extended again by Greene (1990). In this case the distributional assumptions are:

- $v_i \sim$ iid $N(0, \sigma_v^2)$
- $u_i \sim$ iid gamma
- $u_i$ and $v_i$ are distributed independently of each other, and of the regressors.

The exception of the Gamma distribution is that generalizes the one-parameter exponential distribution by introducing an additional parameter to be estimated, and this provides a more flexible representation of the pattern of technical efficiency in the data. The Gamma density function $f(u)$, for $u \geq 0$ is:

$$f(u) = \frac{u^m}{\Gamma(m+1)\sigma_u^{m+1}} exp\left\{-\frac{u}{\sigma_u}\right\}, \quad m > -1. \tag{2.45}$$

We can comment on the different values that $m$ can assume. If we have $m = 0$ the gamma density function collapses into an exponential distribution. If we have $-1 < m < 0$, the gamma density has the shape of an exponential density, the mass of the density remains concentrated near zero. If $m > 0$ the density is concentrated at a point farther away from zero as $m$ increases.

At this point is interesting to understand the importance of the distributional assumptions. The sample mean efficiencies are apt to be sensitive to the distribution assigned to the one-sided error component $u$. It is evident that there are a lot of examples of this sensitivity. But is still not clear if a ranking of producers made by their individual efficiency scores, or a composition of the top and bottom efficiency score deciles, is sensitive to distributional assumptions. There are some example that show the non-sensibility.

### Method of Moments Approach

Resuming the estimation strategy that we have considered until now, it consist in two steps: the first step is the maxim likelihood method used to estimate all parameters of the model; the second step follows from the first one, and considers all these maximum likelihood estimates to obtain technical efficiency. The procedure consist in the decomposition of the maximum likelihood residual term into a noise component and a technical inefficiency

component. In this paragraph we try to supply an alternative to the first step. The first step can be break into two parts. The first part use the OLS to generate consistent estimates of all the parameters describing the structure of the production frontier, apart from the intercept. This part is independent of distributional assumptions on either error component. The second part use the distributional assumptions to obtain consistent estimates of the intercept and the parameters describing the structure of the error components. If we divide in two part the estimation procedure we apply a sort of MOLS to a stochastic production frontier model. At this point we can repeat the second step, in which the JLMS technique is used to estimate $u_i$ for each producer. The method of moments approach has been discussed for normal-exponential, normal-gamma and normal-truncated model. For the application we'll see here what happen in the normal-half normal model.
Stochastic production frontier model:

$$lny_i = [\beta_0 - E(u_i)] + \sum_n \beta_n lnx_{ni} + v_i - [u_i - E(u_i)] \tag{2.46}$$

We assume that $v_i$ is symmetrically distributed with zero mean and that $u_i \geq 0$. The error term $\{v_i - [u_i - E(u_i)]\}$ has zero mean and constant variance. Let's trace the various step that we have enunciated before. In the first part of the estimation procedure OLS can be used to obtain consistent estimates of the $\beta_n s$. In the second part of the estimation procedure it is necessary to proceed with the estimation of $\beta_0$, $\sigma_u^2$ and $\sigma_v^2$. At this point we need distributional assumptions on the error components $v_i$ and $u_i$. We assume that $v_i \sim N(0, \sigma_v^2)$. Moreover we can assume that $u_i$ follows an half-normal distribution, so $E(u_i) = \sqrt{2/\pi}\sigma_u$, $V(u_i) = [\pi-2)/\pi]\sigma_u^2$, and we are also interested in third moment, $E(u_i^3) = -\sqrt{27\pi}(1-4/\pi)\sigma_u^3$. We can now compute the moments of $\epsilon = v_i - u_i$, $\mu_2 = \sigma_v^2 + [(\pi-2)/\pi]\sigma_u^2$ and $\mu_3 = \sqrt{2/\pi}(1-4/\pi)\sigma_u^3$. This two moments are the same also for $\epsilon_i^* = \{v_i - [u_i - E(u_i)]\}$, since $E(u_i)$ is a constant. Continuing the estimates of $\sigma_u^2$ and $\sigma_v^2$ are:

$$\hat{\sigma}_u^2 = \left(\frac{m_3}{\sqrt{2/\pi}(1-4/\pi)}\right)^{2/3} \tag{2.47}$$

and

$$\hat{\sigma}_v^2 = m_2 - \left(1 - \frac{2}{\pi}\right)\hat{\sigma}_u^2 \tag{2.48}$$

After all these passages we obtain a consistent estimate of $\beta_0$ from

$$\hat{\beta}_0 = [\beta_0 - E(\hat{u}_i)] + \sqrt{\tfrac{2}{\pi}}\hat{\sigma}_u$$

$$= OLS\ intercept + \sqrt{\tfrac{2}{\pi}}\hat{\sigma}_u$$

We now have consistent estimates of all parameters in the model. Finally we can apply the JLMS technique to obtain either $E(u_i|\epsilon_i)$ or $M(u_i|\epsilon_i)$.

Olson *et al.* (1980) noticed two potential problems with the method of moments approach. First of all the third central moment $\mu_3$ of the OLS disturbances must be negative, but it is possible for the third central moment $\mu_3$ to be positive. This imply that $\hat{\sigma}_u$ is negative and the model is misspecified. It is necessary to respecified the model, a possible way is changing the functional form or the variables of $f(x;\beta)$. When these changing are not possible it is natural to set $\hat{\sigma}_u = 0$ and this will lead to the conclusion that there is no technical inefficiency in the data. This firs problem $(\sigma_u^2 < 0)$ arise when the true but unknown $\lambda = \sigma_u/\sigma_v$ is small.

Moreover arise a second potential problem when $m_3$ is negative, $m_3$ is sufficiently large to cause $\hat{\sigma}_v < 0$. In this event it is appropriate to set $\hat{\sigma}_v^2 = 0$ and conclude that there is no noise in the data. this second problem instead arise when $\lambda$ is large.

OSM made a Monte Carlo experiment and decide to base the type of estimator on the value of $\lambda$ and on the sample size. For example, if we have the sample size below 400 and $\lambda < 3.16$ the method of moments estimator outperforms MLE, instead Coelli found in a subsequent Monte Carlo study that MLE outperform the method of moments when $\lambda$ is large. We have seen that MOLS procedure generates consistent estimators for all parameters in the model, but that they are inefficient compared to the MLE that are based on distributional assumptions.

### Stochastic Distance Functions

So far we have considered the single-output case in the estimation of a stochastic production frontier, but it is possible to estimate a stochastic output distance function in the multiple-output case. We notice the presence of two main complications: there is no natural choice for a dependent variable in the multiple-output case; the endogeneity of regressors is apt to pose a problem. After solving these problems the estimation procedure will be the same as the single-output case.

The procedure in the single-output case start writing a stochastic production frontier:

$$y_i = f(x_i;\beta)exp\{v_i - u_i\}, \tag{2.49}$$

and in a different and useful form we can rewrite the same quantity:

$$\frac{y_i}{f(x_i;\beta)} = exp\{v_i - u_i\}. \tag{2.50}$$

In the single-output case we had this formulation:

$$D_0(x_i, y_i; \beta) = \frac{y_i}{f(x_i; \beta)}, \tag{2.51}$$

consequently in the multiple-output version we have:

$$D_0(x_i, y_i; \beta) = exp\{v_i - u_i\} \tag{2.52}$$

and we can write it as a stochastic distance function model

$$1 = D_0(x_i, y_i; \beta) exp\{u_i - v_i\} \tag{2.53}$$

The previous equation has to be converted into an estimable regression model. This can be carry out remembering the property of homogeneity of degree one of output distance functions:

$$D_0(x_i, \lambda y_i; \beta) = \lambda D_0(x_i, y_i; \beta), \qquad \lambda > 0 \tag{2.54}$$

Moreover we have to set $\lambda = |y_i|^{-1} = (\sum_m y_{mi}^2)^{-1/2}$, the reciprocal of the Euclidean norm of the output vector, that generates

$$D_0(x_i, y_i/|y_i|; \beta) = |y_i|^{-1} D_0(x_i, y_i; \beta) \tag{2.55}$$

from this we can obtain:

$$D_0(x_i, y_i; \beta) = |y_i| D_0(x_i, y_i/|y_i|; \beta) \tag{2.56}$$

The substitution of this last equality into the (2.53), and the division of both members by $|y_i|$ generate an estimable composed error regression model

$$|y_i|^{-1} = D_0\left(x_i, \frac{y_i}{|y_i|}; \beta\right) exp\{u_i - v_i\}. \tag{2.57}$$

We can analyze all the components of this regression model, let's start from the dependent variable, that is the reciprocal of the norm of the output vector; the regressors are the inputs and the normalized outputs; the $v_i$ is the symmetric error component apt to capture the effects of random noise; $u_i$ is the nonnegative error component, one-sided, and provides the basis for a reciprocal measure of output oriented technical efficiency. The entire analysis of the Stochastic Production Function can be applied to the last equation, with a change in the sign of the $u_i$, from + to − and with an appropriate flexible functional form selected for $D_0(x_i, y_i/|y_i|; \beta)$. There can be a serious problem with the outputs that can appear as regressors, and also the normalized outputs appear as regressors and may not be exogenous.

Finally it is interesting to stress the attention of the reader on the drawback of the cross sectional stochastic production frontier. First of all the MLE of the stochastic production function model and the subsequent separation of technical inefficiency from statistical noise, both require distributional assumptions on each error component. The robustness of inferences is not proved , also if we are able to made some observations on it. A second drawback is that MLE requires the assumption that the technical inefficiency might be correlated with the input vectors producers select. The last observation is that the technical inefficiency of producers can be estimated using the JLMS technique, but not consistently, because the variance of the conditional mean or the conditional mode of $(u_i|\epsilon_i)$ for each individual producer, does not go to zero as the size of the cross section increase.

### 2.2.2 Panel data production frontier models

The strength of a panel data model is that all the limitations of the cross-sectional models can be avoided. The possibility to know repeated observations on each producer can be equivalent of the strong distributional assumption used with cross sectional data, that can - in this case - be relaxed.

A panel contains more information than does a single cross section. The access to panel data enables us to adapt conventional panel data estimation techniques to the technical efficiency measurement problems. The interesting aspect is that there is not dependence from strong distributional assumption, and moreover not all panel data estimation techniques require the assumption of independence of the technical inefficiency error component from the regressors. Repeated observations on a sample of producers serve indeed as a substitute for the independence assumption. We can notice that adding more observations on each producer generates information not provided by adding more producers to a cross section. In short the technical efficiency of each producer in the sample can be estimated consistently as $T \to \infty$, $T$ is the number of observations on each producer. There is another benefit: repeated observations on a sample of producers resolves the inconsistency problem with the JLMS technique, but it is not so realistic because many panels are relatively short.

In the first step of analysis we consider panel data production frontier models in which technical efficiency is allowed to vary across producers, but is assumed to be constant through time for each producer. We can find several conventional panel data models that can be adapted to the problem of estimating technical efficiency. However it can be considered restrictive to assume the time invariance of technical efficiency, so in a second moment we will consider panel data production frontier models in which technical efficiency vary across producer and across time for each producer. The use of these model is not so widespread and rely on extending maximum likelihood cross-sectional production frontier models. We assume throughout this section that the panel is *balanced* - each producer is observed $T$ times. Panels can be also *unbalanced*, producer $i$ is observed $T_i < T$ times, with not all $T_i$ equal, and can be accommodate by each of the panel data models we discus.

### Time-Invariant Technical Efficiency

The assumption that we made at first is that we have observations on $I$ producers, indexed by $i = 1, ...I$, through $T$ time periods, indexed by $t = 1, ..., T$. We can write like simplified functional form for the production frontier a Cobb-Douglas, with time-invariant technical efficiency:

$$lny_{it} = \beta_0 + \sum_n \beta_n lnx_{nit} + v_{it} - u_i, \tag{2.58}$$

where $v_{it}$ is the random statistical noise and $u_i \geq 0$ represents technical inefficiency.The structure of production technology is assumed to be constant through time, there is no consideration for technical change. This model differs from cross-sectional production frontier model only for the time subscripts to the output, to the inputs and to statistical noise. This model can also be associated - for similarity of characteristics- to a conventional panel data model, with producer effects but without time effects, but with the difference that producer effects are required to be nonnegative, because they represent the technical inefficiency. It is possible to estimate the parameters and the technical efficiency in several ways. Let's go into the detail with a list of some possible types of models.

**The Fixed-Effects Model**

The fixed effects model is the simplest panel data model. It is necessary to modify one of the assumption in order to adapt such a model to the efficiency measurement context. It is required now that $u_i \geq 0$, and we can also assume that the $v_{it}$ are iid$(0, \sigma_v^2)$ and are uncorrelated with the regressors. It is not necessary to make distributional assumptions on the $u_i$ and we allow the $u_i$ to be correlated with the regressors or with the $v_{it}$. We can assert that the $u_i$ are treated as fixed effects, and they became producer specific intercept parameters to be estimated along with the $\beta_n s$. A way to estimate the model is by applying OLS to

$$lny_{it} = \beta_{0i} + \sum_n \beta_n lnx_{nit} + v_{it} \tag{2.59}$$

where $\beta_{0i} = (\beta_0 - u_i)$ are producer specific intercepts. The first step is the estimation of the coefficients. There are three equivalent ways of estimation: (i) by suppressing $\beta_0$ and estimating $I$ producer specific intercepts; (ii) by retaining $\beta_0$ and estimating $(I - 1)$ producer specific intercepts; or (iii) by applying the within transformation, in which all data are expressed in terms of deviations from producer means and the $I$ intercepts are recovered as means of producer residuals. We can have several variants as least squares with dummy variables, and we can express this variant with an abbreviation of this type: LSDV. The second step is the employment of the following normalization:

$$\hat{\beta}_0 = max_i \left\{ \hat{\beta}_{0i} \right\}, \tag{2.60}$$

after that the $u_i$ are estimated from

$$\hat{u}_i = \hat{\beta}_0 - \hat{\beta}_{0i} \tag{2.61}$$

and this formulation ensures that all $\hat{u}_i \geq 0$.

At this point we have all the elements to obtain the producer-specific estimates of technical efficiency:

$$TE_i = exp\{-\hat{u}_i\}. \tag{2.62}$$

In the fixed effects model at least one producer is assumed to be 100% technically efficient, and the technical efficiencies of other producers are measured relative to the technically efficient producer(s). The fixed effects model is similar to the COLS model based on cross-sectional data because the transformation is the same.

The LSDV estimates of the $\beta_n s$ are consistent as either $I \to \infty$ or $T \to +\infty$, and the consistency property does not require that the $u_i$ be uncorrelated with the regressors. We can comment also the LSDV estimate of the $\beta_{0i}$. They are consistent as $T \to +\infty$, although consistency of the LSDV estimates of $u_i$ requires both $I \to \infty$ and $T \to +\infty$. Neither consistency property requires the assumption that the $v_{it}$ be normally distributed.

The fixed effects model has the virtue of simplicity and it has good consistency properties, in particular the fixed effects panel data model provides consistent estimates of producer-specific technical efficiency. We have also to consider the drawback. The fixed effect $u_i$ capture the effects of all the phenomena that vary across producers but are time invariant, but the intent is to capture variation in time-invariant technical efficiency. So we have a confounding effect that make not clear if other effects are included as regressors or not. The solution is to use another panel data model, that we take into account here later on.

### The Random Effects Model

In this model we assumed that the $u_i$ are randomly distributed with constant mean and variance, and are also uncorrelated with the regressors and with the $v_{it}$. We don't make distributional assumption also in this case and we still require that the $u_i$ be nonnegative. The assumption are the same that before also for the $v_{it}$, they have zero expectation and constant variance. With these modifications in the assumptions is possible to include time-invariant regressors in the model. We can rewrite the model, adding and subtracting the quantity $E\left(u_i\right)$:

$$lny_{it} = \left[\beta_0 - E\left(u_i\right)\right] + \textstyle\sum_n \beta_n lnx_{nit} + v_{it} - \left[u_i - E\left(u_i\right)\right]$$

$$= \beta_0^* + \textstyle\sum_n \beta_n lnx_{nit} + v_{it} - u_i^*$$

The assumption that the $u_i$ are random allows some of the $x_{nit}$ to be time invariant. If we observe the model we can assert now that the random effect model fits exactly into the one-way error components model in the panel data literature. The method we can use to estimate this model is the standard two-step generalized least squares (GLS). In the first step OLS is used to obtain estimates of all parameters; it is possible to estimate the two variance components with several methods. In the second step $\beta_0^*$ and the $\beta_n s$ are estimated again using the feasible GLS. It is interesting to notice that $\beta_0^*$ does not depend on $i$, since $E\left(u_i\right)$ is a positive constant, so there is only one intercept term to be estimated. After all these estimation we continue estimating the $u_i$ from the residual in the following way:

$$\hat{u}_i^* = \frac{1}{T} \sum_i \left( lny_{it} - \hat{\beta}_0^* - \sum_n \hat{\beta}_n lnx_{nit} \right). \tag{2.63}$$

This is a temporal mean of the residuals. In order to obtain the one-sided we make a normalization of this type:

$$\hat{u}_i = max_i \left\{\hat{u}_i^*\right\} - \hat{u}_i^*. \tag{2.64}$$

We can speak of consistency if both $I$ and $T$ go to infinity, $I \to \infty$ and $T \to \infty$. We obtain the estimates of producer-specific technical efficiency by substituting $\hat{u}_i^*$ into equation (2.62), the same process than in the fixed-effects model. This bring to the conclusion that it is possible to obtain consistent estimates of producer-specific technical efficiency using a random-effects panel data model.

We can consider also an alternative estimator of $u_i^*$: the best linear unbiased predictor (BLUP). It assumes the following form:

$$\tilde{u}_i^* = -\left[ \frac{\hat{\sigma}_u^2}{T\hat{\sigma}_u^2 + \hat{\sigma}_v^2} \right] \sum_t \left( lny_{it} - \hat{\beta}_0^* - \sum_n \hat{\beta}_n lnx_{nit} \right), \tag{2.65}$$

from which we obtain the estimator

$$\tilde{u}_i = max_i \{\tilde{u}_i^*\} - \tilde{u}_i^*, \tag{2.66}$$

that can also be substituted into (2.62) in order to generate another estimate of the technical efficiency. For $T$ large the different estimates became equivalent, and we can affirm that both are consistent for $I \to \infty$ and $T \to \infty$. The general meaning is the same that for the fixed-effects model. It is required that at least one producer is 100% technical efficient. Which method we have to chose? GLS is appropriate when $I$ is large, because consistent estimator of $\sigma_u^2$ requires $I \to \infty$. and when the effects are uncorrelated with the regressors, because the effect of uncorrelatedness is the increasing of efficiency in estimation.

It is interesting also to introduce the Hausman test, on the uncorrelatedness hypothesis significance of the difference between the fixed-effects estimator and the GLS estimator.

Finally we can affirm that GLS allows the presence of time-invariant regressors, and the impact of these regressors can be confounded with the impact of variation in technical efficiency in the fixed-effects model, moreover GLS require that the $u_i$ be uncorrelated with the regressors, whereas the fixed-effects approach does not.

**Maximum likelihood**

Until this moment we have considered the panel data models because they have the strength to avoid distributional assumptions and independence assumptions. Nonetheless in some cases it is possible that such assumptions are tenable in a panel data context, in these cases the maximum likelihood estimation is feasible.

We begin now a close examination of the possible assumptions that we can make to implement a maximum likelihood estimation. The maximum likelihood estimation of a stochastic production frontier panel data model with time-invariant technical efficiency is structurally similar to the same procedure applied to cross-sectional data.

The first step of our examination is the introduction of a first set of distributional assumption:

- $v_{it} \sim iidN(0, \sigma_v^2)$
- $u_i \sim N^+(0, \sigma_u^2)$
- $u_i$ and $v_{it}$ are distributed independently of each other, and of the regressors.

These assumptions are parallel to the ones in the normal-half normal model based on cross-sectional data, the only difference is that now the noise component varies through time as well as across producers.

Pitt and Lee (1981) used the production frontier models and estimated them with these assumptions, to obtain technical efficiency using panel data. They used the dependence and independence assumptions, the density functions and log-likelihood functions and the conditional distribution to obtain the estimates.

### Sensitivity of Results to the Estimation Method

We have showed three different approaches to the estimation of a production frontier model when we have panel data: fixed-effects approach, based on LSDV; random-effects approach based on GLS; maximum likelihood approach.

The selection of one approach rather than another, depend on the different properties of the data, and on the requirements imposed by the approach.

With large $I$ and small $T$ or in presence of time-invariant regressors, a random-effects approach based on GLS is clearly preferred to a fixed-effects approach based on LSVD. Instead if we have as plausible independence assumptions of effects and regressors, MLE is generally more efficient than either LSDV or GLS, because utilize distributional information that the other two do not do. In the literature all these approaches are used, the focal point is to discover if they generate different results or if they are different ways to reach the same conclusion. Gong and Sickles (1989) have implemented a series of Monte Carlo experiments using all the three approach, and have found that the three approach generate similar estimates of efficiency, that are similar in terms of correlation and of rank correlation. Having equal results is preferable to use the approach that requires less assumptions and is computationally easy. This lead to a preference for the fixed-effects model. After this first finding Gong and Sickles have also noticed that as the complexity of the underlying technology increases, the performance of all three approaches deteriorate. Across the literature is possible to find several application of these approaches, that find different result. However it is possible to conclude that also if the evidence conflict sometimes, the approaches are mostly similar, and generate similar efficiency ranking, in particular at the bottom and at the top of the distribution, where the managerial interest is concentrated.

### Technical Improvement

The panel is useful if is long, but is important to observe an improvement of the technology, that has not to remain constant. For this reason is important to include time among the regressors, because time can be seen as a proxy of the technical change. This practice is common in the estimation of production functions based on panel data, but not so common in the estimation of production frontiers using panel data. The possible reason that one can individuate is that in the panel data frontier models there are time-varying technical efficiency specifications, and can be difficult to separate the effects of technical change from the effects of technical efficiency change, because both the effects are influenced from the passage of time. The next step will be to consider the time-varying technical efficiency.

### Time-Varying Technical Efficiency

We now relax the previous assumption that technical efficiency is constant through time, because is a too strong assumption, especially if we operate in a competitive environment. It is not so realistic to consider that technical inefficiency remain constant through many periods of time. The longer the panel, the more desirable it is to relax this assumption. It is possible to take into account also the time, but the cost is of additional parameters to estimate. Also with time is possible to use three different approaches: fixed or random effects model and maximum likelihood.

**Fixed-effects Models and Random-Effects Models**

The model with time-varying technical efficiency was proposed first by Cornwell *et al.* (1990) (CSS) and assume the following form:

$$\begin{aligned} lny_{it} &= \beta_{0t} + \textstyle\sum_n \beta_n lnx_{nit} + v_{it} - u_{it} \\ &= \beta_{it} + \textstyle\sum_n \beta_n lnx_{nit} + v_{it}, \end{aligned} \tag{2.67}$$

where $\beta_{0t}$ is the production frontier intercept common to all producers in period $t$, $\beta_{it} = \beta_{0t} - u_{it}$ is the intercept for producer $i$ in period $t$ and all the others variables are as previously defined. We can proceed as usual, beginning with the estimation of the parameters describing the structure of production technology; then we have to obtain producer-specific estimates of technical efficiency.

In this case we have a $I$x$T$ panel, and it is hard to obtain all the $IT$ intercepts $\beta_{it}$, the $N$ slope parameters $\beta_n$ and $\sigma_v^2$. CSS solve this problem through a quadratic specification

$$\beta_{it} = \Omega_{i1} + \Omega_{i2}t + \Omega_{i3}t^2, \tag{2.68}$$

this formulation reduces the number of intercepts and parameters to $I3$, however the parameters to be estimated are still a lot, especially if the ratio

$(I/T)$ is large. The quadratic specification allows technical efficiency to vary through time. Commenting on the values assumed by the $\Omega$ it is possible to affirm that if $\Omega_{i2} = \Omega_{i3} = 0 \ \forall i$, this model collapses to time-invariant technical efficiency model; if $\Omega_{i2} = \Omega_2$ and $\Omega_{i3} = \Omega_3 \ \forall i$ this model collapses to a fixed-effects model. We can interpret this restricted version of the model such as the technical efficiency is producer-specific and varies through time , but in the same way far all producers. The interpretation can also be that technical efficiency is producer-specific and time invariant, with the quadratic time term capturing the effects of technical change. The problem is that it is not possible to distinguish between the two different interpretation.

CSS describe different estimation strategies. Let's start with the *fixed-effects approach.* Considering the model with time-varying technical efficiency, first of all we have to delete $u_{it}$ from the equation and estimate the $\beta_n s$ from the residuals, then regress the residual on a constant $t$ and $t^2$ to obtain estimates of $(\Omega_{i1}, \Omega_{i2}, \Omega_{i3})$ for each producer. Instead of this process if $I/T$ is relatively small we can include $u_{it}$ in the model with time-varying technical efficiency and estimate the $\Omega_{i1}$ as coefficient of producer dummies, and estimates $\Omega_{i2}$ and $\Omega_{i3}$ as coefficients of producer dummies interacted with $t$ and $t^2$. The following step is the creation of estimates of $\beta_{it}$ and the definition of $\hat{\beta}_{0t} = max_i \left\{\hat{\beta}_{it}\right\}$, the estimated intercept of the production frontier in period $t$. The resulting technical efficiency is $TE_{it} = exp\left\{-\hat{u}_{it}\right\}$ where $\hat{u}_{it} = (\hat{\beta}_{0t} - \hat{\beta}_{it})$. The request is also in this case that in each period at least one producer has to be estimated to be 100% technically efficient.

CSS developed a GLS random-effects estimator to overcome the problem that time-invariant regressors cannot be included in the fixed-effects model with time-invariant technical efficiency. For a fixed $T$, GLS remain more efficient than the fixed-effects estimator in the time-varying efficiency context. But is necessary to stress the fact that GLS remain inconsistent if technical efficiencies are correlated with the regressors. CSS developed an efficient instrumental variables (EIV) estimator again to overcome this other problem. EIV is consistent when efficiencies are correlated with regressors and allows for the inclusion of time-invariant regressors. GLS and EIV proceed in the estimation in the same way as before, with the only difference of the sets of residuals that are used.

We can consider also an alternative formulation proposed by Lee and Schmidt (1993), with a different and more generic formulation for the $u_{it}$, that are specified as:

$$u_{it} = \beta(t)u_i \tag{2.69}$$

where $\beta(t)$ is specified as a set of time dummy variables $\beta_t$, for all the producers of that year. This model is more flexible than the previous, because does not restrict the temporal pattern of the $u_{it}$ to any parametric form. However LS is at the same time less flexible than CSS, because it restricts the temporal pattern of the $u_{it}$ to be the same ($\beta_t$) for all producers. This model is appropriate for short panels, since it requires estimation of $T-1$ additional parameters. LS consider fixed-effects and random-effects within which time-varying technical efficiency can be estimated, and in both approaches the $\beta_t s$ are treated as coefficients of the effects of $u_i$, that can be fixed or random. Having the $\beta_t s$ and $u_i$ it is possible to compute:

$$u_{it} = max_i \left\{ \hat{\beta}_t \hat{u}_i \right\} - \left( \hat{\beta}_t \hat{u}_i \right), \tag{2.70}$$

and consequently $TE_{it} = exp\left\{ -\hat{u}_{it} \right\}$ can be calculated.

**Maximum Likelihood**

Ultimately we consider the maximum likelihood method of estimation, we can use this approach when independence and distributional assumption are tenable. The departure is from the production frontier model (2.67), with $u_{it} = \beta_t \cdot u_i$, where $v_{it} \sim iidN(0, \sigma_v^2)$ and $u_i \sim iidN^+(0, \sigma_u^2)$. We define $\epsilon_{it} = v_{it} - u_{it} = v_{it} - \beta_i \cdot u_i$, where the vector $\epsilon_i$ assumes the form $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{iT})'$. The density function for $\epsilon_i$ is:

$$\begin{aligned}
& f(\epsilon_i) = \textstyle\int_0^\infty f(\epsilon_i, u_i) du_i \\
& = \textstyle\int_0^\infty \prod_i f(\epsilon_{it} - \beta_t \cdot u_i) f(u_i) du_i \\
& = \tfrac{2}{(2\pi)^{(T+1)/2} \sigma_v^T \sigma_u} \textstyle\int_0^\infty exp\left\{ -\tfrac{1}{2} \left[ \tfrac{\sum_t (\epsilon_{it} - \beta_t \cdot u_i)^2}{\sigma_v^2} + \tfrac{u_i^2}{\sigma_u^2} \right] \right\} du_i \\
& = \tfrac{2\sigma_* exp\left\{ -\frac{1}{2} a_{*i} \right\}}{(2\pi)^{(T+1)/2} \sigma_v^T \sigma_u} \textstyle\int_0^\infty \tfrac{1}{\sqrt{2\pi}\sigma_*} exp\left\{ -\tfrac{1}{2\sigma_*^2} (u_i - \mu_{*i})^2 \right\} du_i,
\end{aligned} \tag{2.71}$$

where

$$\int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_*} exp\left\{ -\frac{1}{2\sigma_*^2} (u_i - \mu_{*i})^2 \right\} du_i = 1 - \Phi\left( -\frac{\mu_{*i}}{\sigma_*} \right),$$

and in detail the elements with the '*'assume the following values:

$$\mu{*}i = \frac{(\sum_t \beta_t \cdot \epsilon_{it}) \sigma_v^2}{(\sigma_v^2 + \sigma_u^2 \sum_t \beta_t^2)},$$

$$\sigma_*^2 = \frac{\sigma_v^2 \sigma_u^2}{\sigma_v^2 + \sigma_u^2 \sum_t \beta_t^2},$$

$$a_{*i} = \frac{1}{\sigma_v^2} \left[ \sum_t \epsilon_{it}^2 - \frac{\sigma_u^2 (\sum_t \beta_t \cdot \epsilon_{it})^2}{\sigma_v^2 + \sigma_u^2 \sum_t \beta_t^2} \right].$$

Since we are facing maximum likelihood, now we have to write the log-likelihood function, in order to obtain the maximum likelihood estimates of $\beta$, $\beta_t$, $\sigma_u^2$ and $\sigma_v^2$:

$$lnL = constant - \frac{I}{2} ln\sigma_*^2 - \frac{1}{2} \sum_i a_{*i} - \frac{I \cdot T}{2} ln\sigma_v^2 - \frac{I}{2} ln\sigma_u^2 + \sum_i ln \left[ 1 - \Phi \left( -\frac{\mu_{*i}}{\sigma_*} \right) \right], \tag{2.72}$$

It is possible to show that $u_i|\epsilon_i \sim N^+(\mu_{*i}, \sigma_*^2)$, and obtain an estimator fo $u_i$ from the mean or the mode of $u_i|\epsilon_i$:

$$E(u_i|\epsilon_i) = \mu_{*i} + \sigma_* \left[ \frac{\phi(-\mu_{*i}/\sigma_*)}{1 - \Phi(-\mu_{*i}/\sigma_*)} \right],$$

$$M(u_i|\epsilon_i) = \begin{cases} u_{*i} & \text{if} \quad \sum_t \beta_t \epsilon_{it} \geq 0 \\ \\ 0 & otherwise \end{cases} \tag{2.73}$$

After the estimation of $u_i$ it is possible to estimate $u_{it}$ from $\hat{u}_{it} = \hat{u}_i \cdot \hat{\beta}_t$, with $\hat{u}_i$ can be either from $E(u_i|\epsilon_i)$ and from $M(u_i|\epsilon_i)$, while the $\hat{\beta}_t$ are maximum likelihood estimates of $\beta_t$, $t = 1, ..., T$ that can be normalized. Minimum squared error predictor of technical efficiency is:

$$\begin{aligned} &E(exp\{-u_{it}\}|\epsilon_i) = E(exp\{-u_i \cdot \beta_t\}|\epsilon_i) \\ &= \tfrac{1-\Phi(\beta_t \cdot \sigma_* - \mu_{*i}/\sigma_*)}{1-\Phi(-\mu_{*i}/\sigma_*)} \cdot exp\left\{-\beta_t \cdot \mu_{*i} + \tfrac{1}{2}\beta_t^2 \cdot \sigma_*^2\right\}. \end{aligned} \tag{2.74}$$

Two different specification of the model have been considered in the literature, the first one is by Kumbhakar (1990) with $\beta(y) = \left[1 + exp\left\{\gamma t + \delta t^2\right\}\right]^{-1}$, that requires the estimation of the two additional parameters $\gamma$ and $\delta$. The second alternative time-varying technical efficiency model was proposed by Battese and Coelli (1992) and the specification for the $\beta$ is:

$$\beta(t) = exp\left\{-\gamma\left(t - T\right)\right\},$$

that has only one additional parameter $\gamma$ to be estimated. We can notice that the function $\beta(t)$ satisfies the properties $(i)\beta(t) \geq 0$ and $(ii)\beta(t)$ decreases at an increasing rate if $\gamma > 0$, increases at an increasing rate if $\gamma < 0$ or finally remain constant if $\gamma = 0$. The specification of Battese and Coelli continue with distributional assumptions: normal for $v_{it}$, truncated-normal for $u_i$. They use maximum likelihood to obtain estimates of all parameters in the model. Under the light of these assumptions, the authors showed that:

$$u_i|\epsilon_i \sim iidN^+(\mu_{**i}, \sigma_*^2),$$

where $\epsilon_i = v_i - \beta \cdot u_i$; $\mu_{**i} = \frac{\mu\sigma_v^2 - \beta'\epsilon_i\sigma_u^2}{\sigma_v^2 + \beta'\beta\sigma_u^2}$; $\sigma_*^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma_v^2 + \beta'\beta\sigma_u^2}$ and $\beta' = (\beta(1), ..., \beta(T))$. If technical efficiency is time invariant, $\gamma = 0 \Rightarrow \beta(t) = 1$ and $\beta'\beta = T$, and the expressions for $\mu_{**i}$ and $\sigma_*^2$ collapse to their time-invariant versions. Lastly we write the minimum squared error predictor of technical efficiency:

$$\begin{aligned} &E(exp{-}uit|\epsilon_i) = E(exp\beta(t) \cdot u_i|\epsilon_i) \\ &= \tfrac{1-\Phi(\beta(t)\sigma_* - \mu_{*i}/\sigma_*)}{1-\Phi(-\mu_{*i}/\sigma_*)} \cdot exp\left\{-\beta(t)\mu_{*i} + \tfrac{1}{2}\beta(t)^2\sigma_*^2\right\} \end{aligned} \tag{2.75}$$

**Method of Moments**

Also in this case is possible to estimate the production frontier model using a method of moments approach. We rewrite the equation in a more convenient form:

$$\begin{aligned} &lny_{it} = \beta_0 - \beta_t \cdot \sqrt{\tfrac{2}{\pi}}\sigma_u + \textstyle\sum_n \beta_n lnx_{nit} + v_{it} - (u_{it} - E(u_{it})) \\ &= \beta_t^* + \textstyle\sum_n \beta_n lnx_{nit} + v_{it} - u_{it}^*, \end{aligned} \tag{2.76}$$

where $u_{it} = u_i \cdot \beta_t$ and $E(u_{it}) = \beta_t \cdot \sqrt{2/\pi} \cdot \sigma_u$. The analysis start with a first step, where OLS is performed on equation 2.76, with time dummies added. The coefficients of the time dummies are the $\beta_t^*s$. The analysis continue with a second step where the residuals of the first-step OLS regression (estimates of $\epsilon_{it}^* = v_{it} - u_{it}^*$) are used to compute third moments for each $t$. The third moments for each $t$ assumes the following form:

$$m_{3t} = \beta_t^3 \cdot E(u_i - E(u_i))^3 = \beta_t^3 \cdot \sigma_t^3 \cdot \left[\sqrt{2/\pi}(1 - 4/\pi)\right]$$

and through the third moment we can compute :

$$\beta_t\sigma_u = \left[\frac{m_{3t}}{\sqrt{2/\pi}}\left(1 - \frac{4}{\pi}\right)^{-1}\right]^{1/3}, t = 1, ..., T.$$

In order to obtain the estimates of $\sigma_u$ and $\beta_t$ we make the normalization $\beta_1 = 1$. For instance we can write an estimation of $\beta_0$ from $\hat{\beta}_0 = (1/T)\sum_t \hat{\beta}_t^* + \hat{\beta}_t\hat{\sigma}_u\sqrt{2/\pi}$. The analysis carry on with a third step: the estimation of the $\sigma_v^2$.

We know that the variance of $\epsilon^*_{it}$ is $[\sigma_v^2 + \beta_t^2\sigma_u^2(1-2/\pi)]$, $\sigma_v^2$ can be estimated in the following way:

$$\sigma_v^2 = \frac{1}{I \cdot T}\sum_i \sum_t \hat{\epsilon}^{*2}_{it} - \frac{1}{T}\hat{\sigma}_u^2\left(1-\frac{2}{\pi}\right)\sum_t \hat{\beta}_t^2. \tag{2.77}$$

We have a further step, the fourth step, where the estimated values of $\beta$, $\beta_t$, $\sigma_v^2$ and $\sigma_u^2$ are used to obtain estimates of $u_{it}$ from $E(u_i|\epsilon_i)$ or from $M(u_i|\epsilon_i)$. It is also possible to calculate the minimum squared error predicted like seen previously in equation 2.75.

It is possible to make several different specifications, for instance Kumbhakar and Hjalmarsson (1993) proposed a specification of the time-varying technical efficiency where $u_{it}$ is broken down into two components: a producer specific component capturing producer heterogeneity, and a producer and time specific component representing technical inefficiency. This approach avoid to impose distributional assumptions until the second step, but has also a problem: any time-invariant component of technical inefficiency is captured by the fixed effects, rather than by the one-sided error component, where it really belongs. This subject is wide, and goes beyond the needs of our actual analysis.

# Chapter 3

# The Data

## 3.1 Generality

We have available a large administrative dataset covering the full population of patients and hospitals operating in the Lombardy Region, one of the most populated and richest regions in Italy. The dataset regards about the entire population of hospitals within the Italian Lombardy Region. We have omitted some hospitals because the information about them were not exhaustive. We are speaking about 126 hospitals, and these hospitals have public and private (profit and not-for-profit) ownership structures. We observed these hospitals over a period of five years, from 2003 to 2007. We are speaking about 4819606 admissions in this period. Given the fact that we use administrative data to investigate the entire population and not a census of it, the sample selection error component of causal estimation error vanishes, as stated by Imai *et al.* (2008). The model is population based, at both Hospital and Patient level.

In order to compute the quality of health care received by a single patient, we started from the Hospital Discharge Chart (HDC), where all the information regarding patients' characteristics and the treatments received during the admission period are recorded. We have applied a few exclusion criteria to avoid bias. These do not influence the correctness of the results. For instance we have considered all the ordinary discharges, excluding the day-hospitals and rehabilitations. In addition, to avoid bias connected with the peculiarity of the case study, we have also excluded children under 2 years old. Moreover we have keep inside the analysis only the hospitals for which we have information along all the time period taken into account. This lead to exclude roughly 8 hospitals, but is not source of error thanks to the big number of data.

## 3.2 The variables

We have available two type of data: a type about the Hospitals, and a type is about the patients. In order to have a clear map of all the information, we give here a short description of the variables.

### 3.2.1 Patients' variables

About patients we know the *age*, but is too much dispersive to take into account this data. So after some attempts we have divided the population into five classes of age. The classes are the following:

- *class* 1 include patients under 18;
- *class* 2 include patients from 18 to 35 included;
- *class* 3 include patients from 36 to 50 included;
- *class* 4 include patients from 51 to 70 included:
- *class* 5 include patients aged 71 and over.

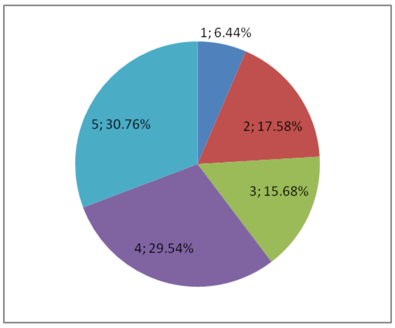

Figure 3.1: Age classes

Another important variable is the *DRG weight*. DRG (diagnosis-related groups) are prospective payment system. Medicare pays hospitals a flat rate per case for inpatient hospital care so that efficient hospitals are rewarded for their efficiency and inefficient hospitals have an incentive to become more efficient. The DRGs classify all human diseases according to the affected organ system, surgical procedures performed on patients, morbidity, and sex of the patient. The Centers for Medicare & Medicaid Services (CMS) assigns a unique weight to each DRG. The weight reflects the average level of resources for an average Medicare patient in the DRG, relative to the average level of resources for all Medicare patients. The weights are intended to account for cost variations between different types of treatments. More costly conditions are assigned higher DRG weights.

About the personal information of the patients it should also be interesting to know the *gender*, this variable assumes value 1 for the male and value 2 for the female, like the ISTAT coding. We have the 46.66% of male, and the remaining 53.34% of female.

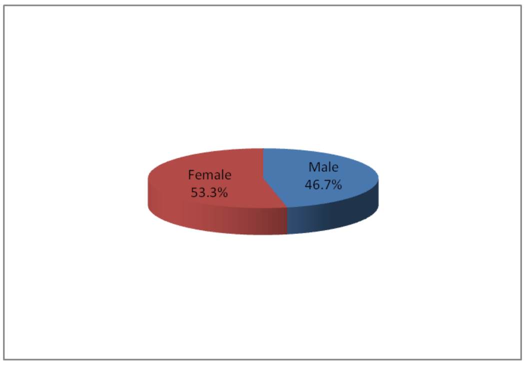

Figure 3.2: Gender

We can continue with the variable description, and we have now *length of stay* in hospital. This variable collects the number of days that a patient stay in hospital. We have decided to drop the permanence over 365 days because is not representative for the purpose of our analysis. Patients that stay so long in hospital are almost always psychiatric, and the kind of care for psychiatric patients are very peculiar and don't influence the quality of a hospitals. The most of the stay are less then 28 days, it is about the 97.76% of the stays.

*Urgent* is a dichotomous variable that assumes value 1 if the patients was *urgent* at the moment of the hospitalize, and assumes value 0 in the opposite case. The 94% of the case are not urgent, we have only a 6% of urgent cases, that maybe are more serious.

*Comorbility* is concomitant but unrelated pathological or disease process. This variable describes the number of diseases in the patient present at the same time. It assumes value 0 if there is no *comorbility*; it assumes value 1 if there is one *comorbility*; it assumes value 2 if there is the presence of two *comorbility*; it assumes value 3 if there is the presence of three *comorbility*; it assumes value 4 if there is the presence of four *comorbility* and finally it assumes value 5 if there is the presence of five *comorbility* in the patient. We can see the distribution in the following figure 3.4. We have computed the values of this variable through the Elixhauser Comorbidity Index that is

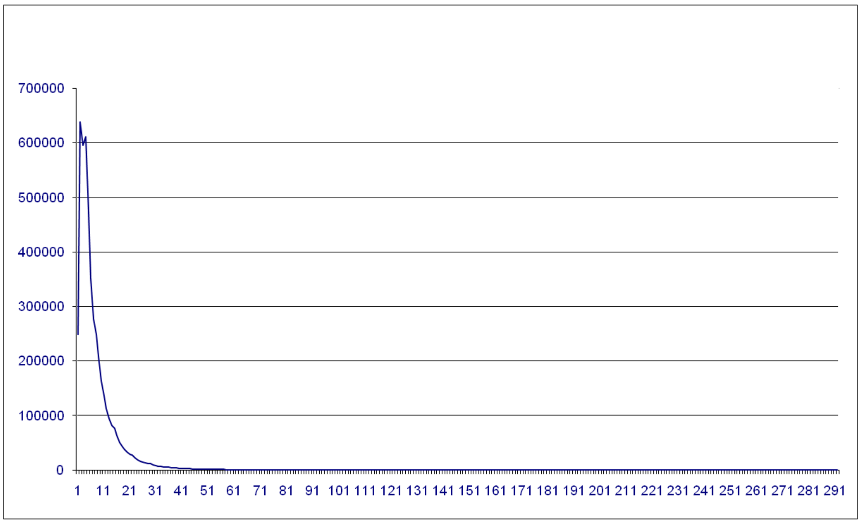

Figure 3.3: Length of stay

defined like "*A measurement tool that defines 30 comorbid (i.e., co-existing) conditions using ICD-9-CM codes. This instrument is meant to be used with large administrative data sets. The index has been updated for use with ICD-10 coding* "(Quan *et al.* (2005)).

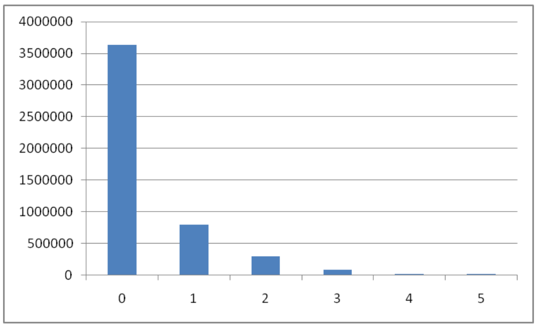

Figure 3.4: Comorbility

Finally we have decided to use two additional variables that are *oncological* and *cardiology*. They are dichotomous variables that assume value 1 if the disease is of that type, and 0 otherwise. It can be interesting to consider these two variables because they represent the majority of diseases. Cardiology is the disease of the 17.87% of the population of patients, and oncological is the disease of the 8.43% of the population.

We have considered also the variable that indicate the year of survey, *year*; and a variable that give a codification to the particular hospital that we are analyzing, *COSP*.

### 3.2.2 Hospitals' variables

For the second phase of the analysis we have the possibility of exploit the data about the hospitals. The variables that we find again are *Year* and *COSP*. Moreover we have *emergency* that indicates the presence (1) or the absence (0) of the emergency inside the hospital. In particular for the 126 hospitals that we have analyzed, we record that 79 hospitals have the emergency, and this data correspond to the 62.7% of the whole population. The remaining 47 hospitals don't have an emergency and they represent the 37.3% of the population.

We can continue with the description of the variables, and we find the variable *mono-specialist* that indicates if the hospital has only one specialization or more. In detail we have only 8 hospitals mono-specialist, and the other are pluri.

The subsequent variable is *university*, and indicates if the hospital has also a teaching part, and is connected with a university of medicine. Also in this case the variable is dichotomous and indicates with a 1 the presence of the characteristic and with a 0 the absence. In our data set we have 13 hospitals connected to a university and 113 no.

Until here the characteristic of the hospitals were unchanged during the years, but the following one are different, maybe a little, from year to year.

The variable *ward* indicates the number of wards every hospital has. There are some changes during the years. The number of wards indicate how much the hospital is general. It will be interesting to check inside the model if is better to have a lot of wards or a few. In figure 3.5 we can see in the x-axis the number of wards in the hospitals, that goes from 1 to 31, and with different colors of the columns we indicate the different years. In the y-axis there is the number of hospitals that have that number of wards in that year.

Then we can consider the number of *beds*, in which we resume all the type of beds inside the hospital, the beds for the day hospital plus the bed for ordinary discharge. The peculiarity that is important to notice is that if we look inside the dataset the number of beds is not an integer. This characteristic depend on the fact that the survey is made on the base of particular standard of accreditation of beds during the years.

We have then three variables about the personnel of the hospital. First of all we consider the *phys*, the number of physicians present in the hospital. Also in this case is curious to notice that is not an integer number, and this time is clear the reason: we have like reference the full-time physicians, and we have to take into account also the part-time.

The same hold for the *nurse* and for the *adm*. The last one indicate the administrative personnel.

Ultimately we have also the data about the ownership of the hospital. We have chosen to classify the hospitals with two variables: the first one divide the public (indicated with a 1) from the private one (indicated with a 0), this variable is named *own*. We have created also a second variable *profit*, that divide the private - profit from all the others.

## 3.3 The choice of the outcome

To start the process we have to select the outcome to implement the multilevel model. There are several possibilities available to compute a proxy for the quality of care in hospital services. These are: in-hospital mortality, voluntary discharges, transfers between hospitals, repeated admissions for the same MDC (Major Diagnostic Categories), return to the operating room, mortality within 30 days from discharge and total mortality. The latter is given by the sum of in-hospital mortality and mortality within 30 days from discharge and also others. The latter is a measure of quality in healthcare services received from hospitals that we adopted in our investigation, and that we choose as the proxy for quality, in line with the desired outcome of our study. It is a dichotomous variable that assumes only two values: 0 in case of no event, 1 in case of death event. It is a not ambiguous variable, because the researcher has only to observe the event of death, that is objective, there is not need of interpretation. There is a branch of literature that criticizes the use of mortality rate like outcome.We have illustrated it at the end of chapter one and now we are conscious of all the possible problems and error that we can meet. But we have consider all the aspects and the way in which we use the mortality is not incorrect.

After a series of considerations about the interpretation of the results we have decided to reverse this variable, and instead of *mortality* we call it *hope of life*. So we have 1 if the patients stay alive during the hospitalize and after 30 days, and 0 if the patients die. The dichotomy nature of the variable creates a necessity for a logistic multilevel model.

The data have a hierarchical structure as patients are classified within different structures, and patients from the same hospital could have similarities in terms of particular characteristics compared to patients from different hospitals. Therefore it seems reasonable to take into account a set of characteristics, covariates of the model, at both levels of the analysis.

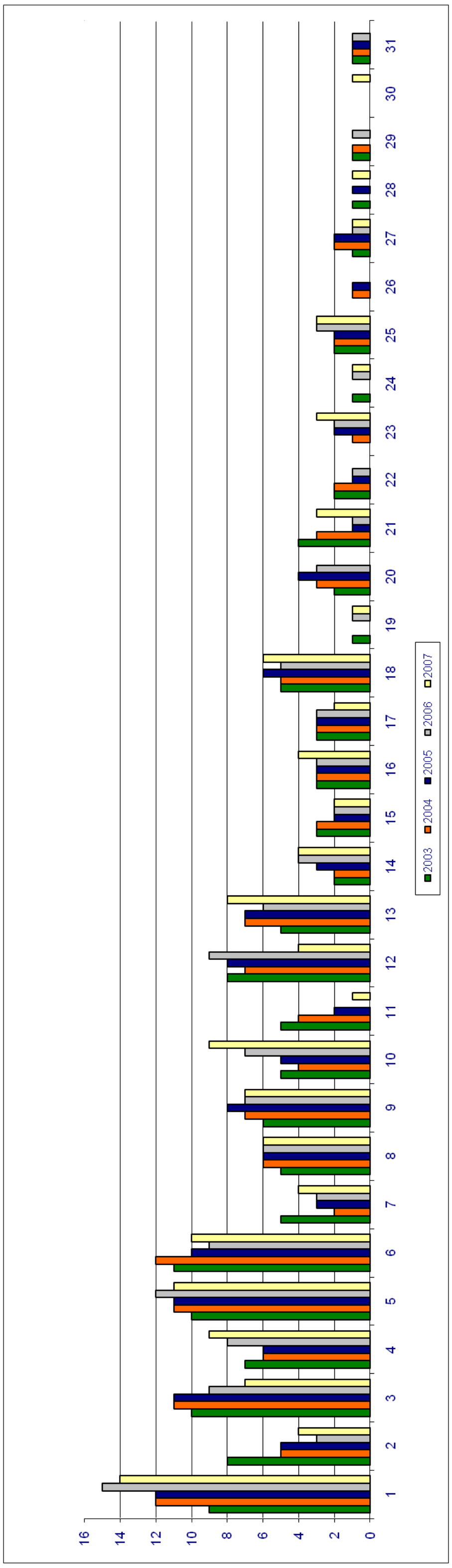

Figure 3.5: Wards

# Chapter 4

# The Model

Our study considers the Health care system like a branch of economic management. There is a current of thought that consider the working of Hospitals like the working of firms. Under the light of this consideration it is interesting to evaluate Hospitals from the economic point of view. Nevertheless this evaluation became useful when is based not only on costs but also on the performance.

## 4.1 The idea

The idea of the model born from the intuition that can be very interesting to be able to take into account both the hierarchical structure of the data and the distance of hospitals from their *frontier* of productivity. We can say that our intent is to consider the effectiveness like an efficiency, with all the economic and performance aspects correlated. Going into the detail we underline again the new aspect from the economic point of view: the hierarchical structure of the data is not neglect. We have widely explained all the reasons and the strength inside the first chapter of this work. The innovative way to look at the necessity of the accountability of hierarchical structure is to not isolate it, and to use multilevel model together with others models. It is possible to implement several different approaches, but the wide availability of data have lead us to the utilize of two different models in two subsequent moments. This choice supports the great possibility to use two different statistical software, and above all to have not to consider restrictive statistical conditions, and to calculate from the beginning algorithm and technique of estimation. We use what yet exist, but in a different way, with different applications and with a particular sequence of step. We have plenty of punctual data, and it is a pity to not exploit all the capability. In light of the main purpose, we have to say that the better way could have been the construction of a statistical model that keep both the features of the models we have used in the two stage. But with a unique model it is necessary to program a new procedure and is more difficult and maybe impossible to utilize the about five millions of data. So in this first analysis we have chosen to keep all the data and to see what happen, if the joint approach of multilevel and stochastic frontier can be useful. We leave as future aim a unique analysis, in which we have a unique multilevel model with three source of error. In this case it means that we introduce the technical inefficiency inside the model that considers the hierarchical structure of the data. Moreover no piece of software will solves the challenging statistical issue underlying decisions about model specification with complex data structures. In our case the structure is not so complex but the plenty of data and the importance of the possible results have conduct us to choose a two stage model. It can be seen like a very hazarded decision, but is the one that we have thought to verify eaven if with a major component of error.

## 4.2 Two step Method

The use of two methods in subsequent moments is not a new technique. There is a current of thought that support the idea that the two step procedure is less time consuming and more flexible. This is because leave the possibility to use different statistical software and drop the necessity to verify contemporary conditions. Maybe is a simplification, but is well used in all the social science, like scientific instrument available for everybody.

### 4.2.1 The first step: Multilevel model

During the first part of the analysis we have done a multilevel study of the population. We start from a wide quantity of data, with the objective to reach a good dataset containing all the information we desire to use. The variables and the records available are even more then the ones we have used, but we have to consider the coherence of the information,and the completeness of the panel of the hospitals during the years.

In this step of the model we use SAS to implement the multilevel model. SAS (statistical analysis software) cover from traditional statistical analysis of variance and predictive modeling to exact methods and statistical visualization techniques. This software is designed for both specialized and enterprise wide analytical needs, because provides a complete, comprehensive set of tools that can meet the data analysis needs of the entire organization. The strength of this software is that is capable of keep and elaborate an enormous quantity of data, in a reasonable time.

### Glimmix

The GLIMMIX procedure fits statistical models to data with correlations or non-constant variability and where response is not necessarily normally distributed. These models are known as generalized linear mixed models (GLMM). The GLMMs, like linear mixed models, assume normal (Gaussian) random effects. Conditional on these random effects, data can have any distribution in the exponential family. The exponential family comprises many of the elementary discrete and continuous distributions. The binary, binomial, Poisson, and negative binomial distributions, for example, are discrete members of this family. The normal, beta, gamma, and chi-square distributions are representatives of the continuous distributions in this family. In the absence of random effects, the GLIMMIX procedure fits generalized linear models (fit by the GENMOD procedure). GLMMs can be useful for different applications, for instance they allow to estimate trends in disease rates, or modeling the proportion of infected plants on experimental units in a design with randomly selected treatments or randomly selected blocks. Others applications can be the prediction of the probability of high ozone levels in counties, modeling skewed data over time, analyzing customer preference and finally joint modeling of multivariate outcomes.

Such data often display correlations among some or all observations as well as non-normality. The correlations can arise from repeated observations of the same sampling units, shared random effects in an experimental design, spatial (temporal) proximity, multivariate observations, and so on. The GLIMMIX procedure does not fit hierarchical models with non-normal random effects. With the GLIMMIX procedure you select the distribution of the response variable conditional on normally distributed random effects.

### Multilevel flexible specification of the production function in health economics

Previous studies on hospitals efficiency often refer to quite restrictive functional forms for the technology. In this work, referring to a study about the hospitals in Lombardy Region, we formulate a convenient way to use statistical model together with economics models. More specifically, in order to take into consideration the hierarchical structure of the data we propose a multilevel model in a first stage. Then the analysis will continue with a second stage, in order to not ignoring the one-side error specification, typical of stochastic frontier analysis. Given this simplification, the use of these two models in two different moments, however, we have to to take into account some typical econometric problems as, e.g. heteroschedasticity or the fact that in the second stage we use estimated values. The estimated production function can be used to identify the technical inefficiency of hospitals (as already seen in previous works), but also to draw some economic considerations without ignoring the real structure of the data. We formulate convenient correctives to a statistical model based on the trans logarithmic function, the most widely used flexible functional form.

The interpretation of the results is surely an interesting administrative instrument for decision makers in order to analyze the productive conditions of each hospital and also to decide the preferable interventions. We want to analyze the effectiveness from an economic point of view but also considering the outcome of the hospitalize.

### 4.2.2 The second step: Stochastic Frontier Analysis

Among the parametric techniques we chose to use the *Stochastic Frontier Model* because takes into account the measuring errors and others stochastic factors. This assumption is realistic in health care system, because several qualitative factors are not catch by the utilized variable. The *Stochastic Frontier Model* itself presents some problem, the principal problem is the neglect of multidimensional and hierarchical structure of the data, Siciliani (2006).

The better way we have thought to overcome this problem is to utilize a frontier after have implemented a multilevel model. The Hospitals have to reach the frontier of optimality, and we believe that we can consider the distance between each hospital and the frontier. The utilization of data from a multilevel implementation allows to take into account the hierarchical structure of the data, since in the previous step we use a hierarchical model. At this point we focus our attention on the hospitals and their performance, but we have not neglected that inside a certain hospital we have exactly that patient.

#### Further Variables

We have implemented a frontier, with the data available about hospitals. In this second step, the data we are interested in are data about hospitals, and not hospitalizes or patients. The data are along the five years, something remain equal during all the period of observation, something else change from year to year. For this reason we have a dataset of 630 records, one per hospital per year. We have reversed the variable of the multilevel model in positive term, so the frontier is the optimality to reach, when the patients has a wide probability to stay alive. We can obtain information about the hospitals' variable from chapter 4, but we have also to introduce two new variables that we have derived from the multilevel.

The first variable *derived* that we introduce in this section is *mherror*. We take initially the residuals of the multilevel model per hospital. This error is distributed like a variable with mean zero and constant variance, and can be used to make a sort of ranking of the hospitals. The major is this residual the better is the performance of that hospital, positive value indicate a performance over the mean. This type of classification has been made a lot in previous works, also if it is not completely correct. We advocate the fact that we use it only as link between the two model, and we correct with the second model in some way the problems that can arise using an imperfect measure. We have another problem to solve: the residuals assume both positive and negative values, but we want to use it as production outcome of the frontier. A production, by definition, has to be positive, so we need to transform our ranking. The solution we have applied is that we have shifted the value up, in order to have only positive values. At each value we have added the minimum value ever assumed. At this point we have obtained a positive variable that we utilized like output of the frontier. *Mherror* correspond to the error of the hospital along the time from the multilevel, and is a ranking translated. Since is important to keep an order, with a shifting we maintain the same ranking.

The second variable *derived* is the *meanp*, that correspond to the mean probability of stay alive after an hospitalize. The probability of survival is derived also from the multilevel model, but this probability is a data available for each patients. It is a very interesting data that tells us just the probability that a particular individual, with his particular characteristics, has to stay alive. In this section we are analyzing hospitals data, so we need the data aggregated. For this reason we make a mean of the probability for each hospitals. We have tried to use it also for the implementation of a frontier, but the entity of this variable dose not allow a lot manipulation because all the records are near to the same value, always over 0.9 (luckily!).

### Frontier 4.1

FRONTIER Version 4.1 was written by Tim Coelli. This program is used to obtain maximum likelihood estimates of the parameters of a variety of stochastic production and cost frontiers; and estimates of mean and individual technical or cost efficiencies. The program can accommodate:

- cross-sectional or panel data;
- half-normal or truncated normal distributions;
- any functional form which is linear in parameters;
- time-invariant or time-varying efficiencies;
- inefficiency effects can be explicitly influenced by a number of firm-specific factors.

This step of the analysis has been implemented with *R*, a software environment for statistical computing and graphics. We have utilized the package *frontier*, implemented by Coelli and Henningsen for the Stochastic Frontier Analysis. It gives the Maximum Likelihood Estimation of Stochastic Frontier Production and Cost Functions. With this package two specifications are available: the error components specification with time-varying efficiencies Battese and Coelli (1992) and a model specification in which the firm effects are directly influenced by a number of variables Battese and Coelli (1995).

Our case is well suitable to Stochastic Frontier Production Functions, because we have no data about cost for now. We have chosen the interface *frontierQuad*, because is convenient for estimating quadratic or translog stochastic frontier functions. Among all the possible functions we have adopted the trans log (transcendental logarithmic) because in economics and econometric is the most flexible specification of the utility, production and cost function. This class of function is named flexible because allow for the analysis of effects that depend on second derivative, like elasticity of substitution, that usually are assumed given and constant in classical functional forms like Cobb-Douglas and CES.

### Battese Coelli frontier

The frontier we have chose to implement is the one proposed by Battese and Coelli (1992). Once decided that the best functional form is the translog, we have made the logarithmic transformations of all the variables, except that for the endogenous one. We proceed specifying all the elements that play a role in the frontier. In “yName”we have to put the name of the endogenous variable, that in our case is *mherror*. With “xNames” we indicate a vector of strings containing the names of the X variables (exogenous variables of the production function) that should be included as linear, quadratic, and interaction terms. It is possible to insert also the specification *shifterNames*, a vector of strings containing the names of the X variables that should be included as shifters only (not in quadratic or interaction terms). Moreover we have also "zNames" a vector of strings containing the names of the Z variables (variables explaining the efficiency level). In our model a possible variable of this type is *own*, that indicate the ownership of the hospital. The data is a panel data frame, created with "plm.data", and it is assumed that these are panel data.

Function frontier is a wrapper function that calls *sfa* for the estimation. The two functions differ only in the user interface; function frontier has the old user interface and is kept to maintain compatibility with older versions of the frontier package. One can use functions *sfa* and *frontier* to calculate the log likelihood value for a given model, a given data set, and given parameters by using the argument "startVal" to specify the parameters and using the other arguments to specify the model and the data. The log likelihood value can then be retrieved by the "logLik" method with argument which set to "start". Setting argument "maxit" to 0 avoids the (eventually time-consuming) ML estimation and allows to retrieve the log likelihood value with the "logLik" method without further arguments.

## 4.3 Possible future development

During our research we have had way to think again to the choice of the two step method. A deeper analysis have give light to another possible specification of the same idea. The future application will be a multilevel model with a further error component. The major problem with this specification is that we have to re-write all the integral and computation of the algorithm of estimation. The computational weight is not indifferent, so we have thought to verify if worthy with the simpler process of two stage analysis. This is only the basis for a detailed research, that we leave like next stage of our analysis.

In order to model the probability of survive, function of aspects connected to patient, hospital and inefficiency (of the hospital) it could be necessary to reconsider the way of modeling the positive replay probability from the patient's side. Rather then a frontier with the probability of survive, we can make a *logit*. The frontier so will assume a form that would contain covariates, a-symmetric error and normal error. The better result that we wish have from the previous model is to obtain the functional form of the probability of survive, and estimate it directly. This approach would bring to specify stochastic frontiers for dichotomous variables.

# Chapter 5

# Results

## 5.1 Multilevel Model

We now write comments to multilevel model. The three-level model reach the convergence after 11 iteration, we allow a maximum number of iteration of 30 in order to reach the convergence.

We are analyzing a model where the first level is the time, with a panel of five years. After this level we have the hospital level, 126 Hospital along all the five years, that are among the totality of the hospital of the Lombardy Region. First of all have a look to the p-value: they are all meaningful. This fact is not so reliable, because there are evidence that the significance of the p-value is not crucial. Some variables can be reliable also without these significance. But if we have it, is only a major confirmation that we underline. The values that we have are all significantly different from zero, it is reasonable to think that the model adequately fit the data.

Age classes are considered instead of the punctual data that are anyway available. We have made this choice because the data are too much dispersive with the punctual data and the model not converge. Accordingly to what table 5.1 displays, the reference class is the number 2, (people aged from 18 to 35), people inside this class are stationary, they have not influence on *hope of life*. The class number 1, (people aged from 1 to 17) has a positive estimate of the coefficient, equal to 0.36. This means that the belonging to this class has a positive influence on the *hope of life* , although not so high. If we continue with this description following the natural order of the age, we arrive at the third class, that include people aged from 36 to 50. This class begins to be negative, meaning that the *hope of life* decrease belonging to this class. The estimate of the coefficient is exactly $-1.32$. The value has a certain importance also if it is designed to grow up with the increasing of

| Effect | Estimate | StdErr | tValue | Probt |
|---:|---:|---:|---:|---:|
| Intercept | 6.45 | 0.094 | 68.39 | <.000 |
| class3 | -1.32 | 0.022 | -59.0047 | <.000 |
| class5 | -3.18 | 0.020 | -156.76 | <.000 |
| class1 | 0.36 | 0.044 | 8.12 | <.000 |
| class4 | -2.23 | 0.021 | -108.33 | <.000 |
| class2 | 0 | | | |
| DRG weight | -0.13 | 0.0013 | -106.58 | <.000 |
| length of stay | -0.011 | 0.00020 | -52.39 | <.000 |
| urgent | -1.34 | 0.0058 | -229.36 | <.000 |
| comorbility | -0.23 | 0.0024 | -96.96 | <.000 |
| genderF | 0.15 | 0.0043 | 34.86 | <.000 |
| genderM | 0 | | | |
| cardiology | 0.13 | 0.0054 | 23.52 | <.000 |
| oncological | -1.10 | 0.0057 | -192.59 | <.000 |

Table 5.1: *Estimate of the coefficients*

the age. In fact the class number 4 (people aged from 51 to 70) has also a negative value of the coefficient and has also a greater value, equal to $-2.23$. Ultimately we arrive at the last class - number 5 - that comprises all the patients over 71. For this class the influence is again negative, the coefficient is equal to $-3.18$, one can conclude that the belonging to this class has a very consistent negative influence on the *hope of life*.

It should also be noted the influence of the others variable. The *drg weight* has a coefficient of $-0.13$, this bring us to consider the gravity of the disease like negative correlated with the *hope of life*. And it is a reassuring result, because corresponds to our expectation. Indeed is natural to think that the more is serious the illness the more the *hope of life* decreases. We can switch to the subsequent variable: *length of stay*. Also this variable has a negative coefficient, but of small entity, because it is equal to $-0.011$. This leads us to think that the longness of stay in the hospital does not bring to a succeed of the care. Going on we have the variable *urgent*, that also has a negative coefficient with value $-1.34$. This means that if a case is more urgent, the probability of remain alive decreases. We have also for this variable an inverse proportionality. *Comorbility* is another variable with inverse proportionality. Its coefficient assumes value $-0.23$, negative but not so wide. Once again we have found an aspect with a bad influence on the *hope of life*, and it is natural, because having more than one disease is for sure a bad way to implement *hope of life.* With the following variable instead we underline a characteristic that has a positive influence on *hope of life.* This variable with a positive coefficient is the *gender.* To be a female has a positive influence, 0.15. Proceeding with the last two variable we have a positive coefficient for *cardiology*, equal to 0.13. This indicates that this type of disease has more *hope of life* because has a well developed way of care and nowadays can be solved easy. Finally we have considered the variable *oncological*, that has a negative coefficient with value −1.10. We can expect this result, because a cancer can be also not treatable.

At this point we can affirm to be satisfied by the results of this first step of the analysis. First of all because the model reaches the convergence, then is a good result to have all the coefficients meaningful, and with a correct values and sign (correct because reasonable). Ultimately it not so usual that the time is considered like a superior level of the hierarchy, and we notice that his meaningful is not so high. This fact can be justified because Singer affirms that the time starts to be meaningful like level of hierarchy if we ave available almost ten years. We have only five years, bud we would anyway stress the importance to be able to analyze a panel.

### 5.1.1 Computation of the intraclass correlation index

We have seen in the first chapter that to check the significance of the model it useful to compute the intraclass correlation index. Going over the two-level theory we need the three-level specification. In particular we recall the formulas that we are going to use, 5.1 with the peculiarity that the major level is time, so we change the subscript '*k*' with '*t*':

$$Y_{ijt} = \gamma_{000} + v_{0t} + u_{0jt} + e_{ijt} \tag{5.1}$$

After that we take also the equation for the hospital level with the first method mentioned:

$$\rho_{hospital} = \frac{\sigma^2_{u_0}}{\sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e} \tag{5.2}$$

The complete model that we have implemented is the one with all the variables of interest that we have mentioned before. For this model the *glimmix* procedure give the following coefficients, that correspond to the value of the variances that we need for the computation of the intraclass correlation index.

The estimate of the intercept for the year corresponds to $\sigma^2_{v_0} = 0.03788889$; the second value that we obtain is the estimate of the intercept of the influence of both year and hospital (remember that *cosp* is the code to identify the

| CovParm | Subject | oldest | StdErr | ZValue | ProbZ |
|---|---|---|---|---|---|
| Intercept | year | **0.03788889** | 0.03007 | 1.260054 | 0.104 |
| Intercept | cosp(year) | **0.55183864** | 0.03569 | 15.46032 | <.000 |
| Residual | | **0.79169558** | 0.00051 | 1552.232 | <.000 |

Table 5.2: *Variances of the complete model*

different hospitals), $\sigma^2_{u_0} = 0.55183864$ and finally the estimate of the residual corresponds to the total variance of the residual, $\sigma^2_e = 0.79169558$. Having all the data available is possible to compute the coefficient that assumes value $\rho_{hospital} = 0.39947$.

There is a second possible approach to compute the same index, the second method exposed in chapter 1:

$$\rho_{hospital} = \frac{\sigma^2_{v_0} + \sigma^2_{u_0}}{\sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e} \tag{5.3}$$

We have to take the same value for the variances and now the intraclass coefficient will be: $\rho_{hospital} = 0.4239$. Both are correct, with a nuance of different meaning.

At this point we continue calculating the intraclass coefficient for the year. The two method coincide in this case. We recall the formula 1.12:

$$\rho_{year} = \frac{\sigma^2_{v_0}}{\sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e} \tag{5.4}$$

and we obtain a value equal to $\rho_{year} = 0.02742$.

It can be interest to compare this model to the only-intercept model, that is the model without any explicative variable. So we repeat these operation on it.

| CovParm | Subject | Estimate | StdErr | ZValue | ProbZ |
|---|---|---|---|---|---|
| Intercept | year | 0.0255 | 0.02426 | 1.051 | 0.14663003 |
| Intercept | cosp(year) | 1.06738 | 0.0672 | 15.88 | <.000 |
| Residual | | 0.98742 | 0.00064 | 1552 | <.000 |

Table 5.3: *Variances of the only intercept model*

We can see that the variances assume value $\sigma^2_{v_0} = 0.0255$ for the year; $\sigma^2_{u_0} = 1.06738$ for the common term to hospital and year and finally the estimate of the residual correspond to the total variance of the residual, $\sigma_e^2 = 0.98742$. The intraclass correlation index for the hospital for the first method assumes value $\rho_{hospital(1)} = 0.5131$, and with the second method assumes value $\rho_{hospital(2)} = 0.5254$.

We can compute also for the only intercept model the intraclass coefficient for the year. Recalling the previous formula 5.4 we obtain a value of $\rho_{year} = 0.01226$. The *year* is more meaningful for the complete model also if its importance is very marginal and we can definitively say that there is no variability along the years.

It is not totally satisfactory that the base model has better result then the complete. But having so complex pattern of analysis is not so easy to implement the structure of the data. Anyway the results are only a few worst for the complete model, so it is not so worrying. Moreover the intraclass correlation index for the year is better for the complete model rather then for the only intercept one, and this fact give evidence of the good development of our model. At this point we can affirm to have developed a good multilevel model and we can switch to the model of the second stage.

## 5.2 Stochastic Frontier Model

The implementation of the stochastic frontier model is done through the *frontier* package of R. We have charged also other libraries: *nlme*, *car*, *lmtest* and *dse*. After that we have charged the data. We have decided to implement a translog function, so we have to transform the data in logarithmic form. The subsequent step is to create the matrix where the frontier works, with the *plm.data* function. At this point we are ready to make the frontier. Like first attempt we use as dependent variable *mherror*. This is a variable that we have create, derived from the multilevel implementation. In order to obtain the best estimate possible we have used the last version of R 2.12.1 and moreover the improved package of frontier, that the author himself: Arne Henningsen, Coelli *et al.* (2009), had provided us. Recently he has found out that NAs in the efficiencies can be caused by numerical instability, which often occurs on digital computers, particularly when numbers get very large or small. In the calculation of efficiencies, the problem can occur, because a denominator of a fraction consisted of the normal density function, which returns zero if it is applied to numbers smaller than minus 37. Henningsen solved this problem in the *frontier* package by replacing "$pnorm(a)/pnorm(b)$" by something like "$exp(log(pnorm(a)) - log(pnorm(b)))$", which is algebraically the same but is numerically more stable. The results we obtain with this implemented package are clearly better. The principal objective of this analysis is to be able to say something about the efficiency of the hospitals. We have implemented several frontiers and now we are going to show the most meaningful one.

### 5.2.1 Dependent *mherror*, without Z

The frontier that gave us better results is the one with output variable *mherror*, and with input variables *ward*, *beds*, *physicians*, *nurses* and *administrative personnel*. Unfortunately we can not keep the external variables for the ownership, because in that case the model would not be meaningful. This attempt give us a warning message that advice us of the possibility that the residuals of the OLS estimates are right-skewed; this might indicate that there is no inefficiency or that the model is misspecified. We have applied an Error Components Frontier by Battese and Coelli (1992). In this frontier the inefficiency decreases the endogenous variable (as in a production function) and the dependent variable is logged. Iterative ML estimation terminate after 37 iterations: log likelihood values and parameters of two successive iterations are within the tolerance limit.

The table 5.4 report the results of the final maximum likelihood estimates for the coefficient of the model. We have reported also the significance codes: if the level of significance is very high and near to 0, three stars are displayed on the table; if the level of significance is around 0.001, two stars are displayed on the table; if we obtain a level of significance of 0.01 the table display only a star; we can obtain also variables with level of significance of 0.05, and in this case the table displays a point "."; finally we can obtain also values with a significance of 0.1 and in this case in the table will be an empty space. After this technical explanation we switch our attention on the variables.

The first variable that enters in the model is *wards*, with a positive coefficient, meaningful and of a certain entity. The coefficient of the single variable is equal to 2.15, and has a meaningful of two stars. We continue analyzing the effect of wards on the frontier, and we can look at the interaction of *wards* with the other variables. We have to say that we can look at the factor of second level, but is non meaningful, and among the interaction is interesting to underline the one with *physicians*, that is still meaningful (also if with only a star) and positive, also if of small entity. Finally for the *wards* we have to take into account another interaction: the one with *administrative personnel*. This time the effect is negative, since the coefficient has value −0.36 and a significance of a star. This is not a great result because if the sign of all the coefficients for that variable is the same we can be sure of the effect of that variable, in case of contrasting sign we are less certain.

We can switch to the second variable: *beds*. This variable has a quite

|  | Estimate | Std. Error | z value | Pr(>\|z\|) |  |
|---|---|---|---|---|---|
| a_0 | 9.57 | 1.87 | 5.12 | 0.00 | *** |
| a_1 | 2.15 | 0.74 | 2.91 | 0.00 | ** |
| a_2 | -2.52 | 0.85 | -2.95 | 0.00 | ** |
| a_3 | 0.08 | 0.61 | 0.13 | 0.90 |  |
| a_4 | 0.45 | 0.74 | 0.60 | 0.55 |  |
| a_5 | -0.71 | 0.70 | -1.02 | 0.31 |  |
| b_1_1 | 0.05 | 0.21 | 0.24 | 0.81 |  |
| b_1_2 | -0.22 | 0.16 | -1.36 | 0.17 |  |
| b_1_3 | 0.22 | 0.10 | 2.10 | 0.04 | * |
| b_1_4 | -0.05 | 0.14 | -0.38 | 0.71 |  |
| b_1_5 | -0.36 | 0.15 | -2.36 | 0.02 | * |
| b_2_2 | 0.74 | 0.28 | 2.59 | 0.01 | ** |
| b_2_3 | -0.16 | 0.17 | -0.94 | 0.35 |  |
| b_2_4 | 0.17 | 0.21 | 0.83 | 0.41 |  |
| b_2_5 | -0.15 | 0.19 | -0.77 | 0.44 |  |
| b_3_3 | -0.20 | 0.12 | -1.70 | 0.09 | . |
| b_3_4 | 0.27 | 0.16 | 1.66 | 0.10 | . |
| b_3_5 | -0.04 | 0.16 | -0.28 | 0.78 |  |
| b_4_4 | -0.92 | 0.32 | -2.88 | 0.00 | ** |
| b_4_5 | 0.39 | 0.20 | 2.00 | 0.05 | * |
| b_5_5 | 0.04 | 0.14 | 0.30 | 0.76 |  |
| sigmaSq | 2.11 | 0.30 | 7.11 | 0.00 | *** |
| gamma | 0.97 | 0.00 | 225.93 | 0.00 | *** |
| time | 0.01 | 0.00 | 1.44 | 0.15 |  |

Table 5.4: *Final Maximum Likelihood Estimates*

high level of significance (two stars) but a negative influence on the good efficiency of hospitals. Her coefficient assumes value equal to $-2.52$, and this mean that hospitals has too much capacity of beds, that influence negatively the performance. If we consider the second order of influence of *beds* we have the surprise that this time it has a positive coefficient, with also a good level of significance (two stars).

Going ahead we can see that the others variables are not meaningful alone, but are meaningful the iterations or the second order factors. For example the second order factor for *physicians* has a significance of 0.05, the coefficient is negative and of small entity, but this result alone mean that more *physicians* make the hospital less efficient. Also the interaction between physicians and nurses is a bit meaningful, with only a point, but the coefficient this time is positive. This indicate that the cooperation between physicians and nurses is positive and bring good results. We continue saying that *nurses* become significant as factor of second order, nevertheless they have a negative influence on the hospitals' performance, since nurses have a negative coefficient. The influence becomes positive also if less meaningful for the interaction with administrative personnel. Finally is crucial to notice that the sigma square has the maximum level of meaningful and also gamma. Moreover the value that gamma assume: 0.97, indicate that our model is good. The log likelihood value is equal to −219.3866, that is a negative value but if we transform the log likelihood in likelihood we will obtain a value between zero and one, that is not so high but positive.

At this point the program remember us that we have implemented a panel data model, where we have 126 cross-section and 5 time periods, with a corresponding total number of observations of 630.

### 5.2.2 Efficiency estimates

Since the whole model is meaningful in this case we are able to compute also the efficiency estimates for all the panel and we arrive to have mean efficiency of each year:

| 1 | 2 | 3 | 4 | 5 |
|---:|---:|---:|---:|---:|
| 0.3315149 | 0.3335679 | 0.3356250 | 0.3376861 | 0.3397511 |

Table 5.5: *Mean efficiency of each year*

These mean efficiency are very low, and this fact indicates that there is a great part of inefficiency. If we look into the efficiency estimates of all the hospitals along the years we can notice that actually hospitals are different one from each other. For example we can individuate the optimal hospitals of the study, the ones with a very high level of efficiency, around 0.9.

The first hospital that we individuate is the number 22, it has an efficiency estimate of 0.95. It is interesting to go into details in order to understand which kind of hospital presents a good performance. This hospital is a public hospital, with an emergency, and without connection with university. It has 11 wards in 2003 and 2004 that become 12 in the subsequent years. The mean number of beds is 164.864, one can finds the details in table **??** the number of physician grow along the years and happen the same for the number of nurses, instead the administrative personnel start with a very high number until arrive to an inferior one. This is not a typical profile, because going ahead we find different typologies.

| year | beds | phys | nurs | adm |
|---|---|---|---|---|
| 2003 | 146.9 | 137.5037 | 287.3876 | 153.9007 |
| 2004 | 170.03 | 134.6132 | 309.794 | 125.5982 |
| 2005 | 173.3 | 122.8551 | 301.9282 | 118.1608 |
| 2006 | 171.1 | 146.4698 | 359.9635 | 140.8731 |
| 2007 | 162.99 | 146.4698 | 359.9635 | 140.8731 |

Table 5.6: *Hospital 22*

Another hospital with high efficiency is the number 41. It is a private profit hospital, with three wards for all the periods of observation. This hospital has a small number of wards but a lot of beds, a strange aspect is that the number of physicians is very small, it has one physician for four beds. The detailed numbers are displayed in table 5.7.

| year | beds | phys | nurs | adm |
|---|---|---|---|---|
| 2003 | 239 | 64 | 157.5 | 123.5 |
| 2004 | 239 | 61 | 157.15 | 124.85 |
| 2005 | 236 | 58.6 | 158.8 | 126.45 |
| 2006 | 234 | 55.95 | 160.25 | 117.8 |
| 2007 | 248 | 55.95 | 160.25 | 117.8 |

Table 5.7: *Hospital 41*

Another hospital with a good efficiency is the number 46. The efficiency grows throughout the years, from 0.89344522 to 0.89607815. It is a private profit hospital, with 4 wards in the 2003 and three in the following years. It has not emergency, it is not mono-specialized and not university. The strong point can be that it has among one physician for each 2 beds and a small number of administrative personnel. The detailed numbers are displayed in table 5.8.

The subsequent hospital that we comment is the number 50, again a private profit hospital, with five wards until 2006 and four in 2007. The efficiency is really high, plus than 0.953 for each year, and growing along the years. Again this hospital has not emergency, is not mono-specialized and not university. The major change along the years is about the beds, how is possible to see in table 5.9 that are much less in the last years, and

| year | beds | phys | nurs | adm |
|---|---|---|---|---|
| 2003 | 45.83 | 59.65 | 38 | 8.65 |
| 2004 | 50 | 64.65 | 40 | 10 |
| 2005 | 47.5 | 51.65 | 34 | 10.65 |
| 2006 | 50 | 51.65 | 37 | 9.65 |
| 2007 | 100 | 51.65 | 37 | 9.65 |

Table 5.8: *Hospital 46*

this is for sure due to the cut of a ward. Notwithstanding the cut of the ward, the personnel remain about the same, both for physicians, nurses and administrative. This means that the good work of the hospital is maintained by an high number of worker.

| year | beds | phys | nurs | adm |
|---|---|---|---|---|
| 2003 | 162.33 | 96.65 | 116.55 | 94.25 |
| 2004 | 153.39 | 86.65 | 112.55 | 93.25 |
| 2005 | 158 | 90.65 | 101.85 | 91.85 |
| 2006 | 128.1 | 92.65 | 102.2 | 86.5 |
| 2007 | 108 | 92.65 | 102.2 | 86.5 |

Table 5.9: *Hospital 50*

Hospital number 52 has only one ward, is private profit and like the previous has not emergency, is not mono-specialized and has not connection with university. Table 5.10 Also shows the beds remain the same along all the years. It is possible to observe a cut off of the number of administrative personnel. For this hospital the efficiency of each year is around 0.94.

| year | beds | phys | nurs | adm |
|---|---|---|---|---|
| 2003 | 60 | 20.9 | 23 | 50 |
| 2004 | 60 | 18.9 | 24 | 47 |
| 2005 | 60 | 22.9 | 21 | 42 |
| 2006 | 60 | 22.9 | 20 | 24 |
| 2007 | 60 | 22.9 | 20 | 24 |

Table 5.10: *Hospital 52*

Finally we can comment the characteristics of the last hospital with high values of efficiency estimates. We are speaking about hospital number 113, with an efficiency estimate of 0.94 for all the years. This hospital is public, with 27 wards in the first year, 28 in 2004 and 2005, and again 27 wards in 2006 and 2007. This hospital has an emergency, is pluri- specialized and university. It a large number of nurses and administrative personnel.

At the end of this description of optima hospitals we notice that we have find out two different kind of optimal hospitals: big public, and small private. It can be interesting to deepen this aspect, and we consider the ownership in the following subsection.

### 5.2.3 Test for ownership

An aspect of great interest is to distinguish the influence of the ownership on the good performance. The way to achieve this objective can be make a comparison between private profit hospitals versus public and private non for profit hospitals. We Make a T-test on the mean, we would verify if the public hospitals have a better performance than the private profit one. It is desirable this result because would say that the public health is well developed. At this point we take all the efficiency estimates and we divide the population of the hospitals in two group: public and private. Then we make the following mean test:

$$\frac{\hat{\mu}_{pub} - \hat{\mu}_{priv}}{\sqrt{\frac{s^2_{pub}(n_{pub}-1)+s^2_{priv}(n_{priv}-1)}{n_{pub}+n_{priv}}}} \sim T_{n_{pub}+n_{priv}-2} \tag{5.5}$$

The hypothesis that we are going to test is that public hospital are more efficient than private, $H_0 : \mu_{pub} > \mu_{priv}$. We make the test at level $\alpha = 0.05$, and the test has one tail. The refuse region is the right tail, so we refuse the hypothesis if the value that we find is superior to the critique value. In our case we are in the region of acceptation, so we have an evidence that public hospitals are better. This is a reassuring result, because supports that the health care system provided by the state is good, for the majority of cases, also if we observe also a few point of excellence among the private institutions.

# *Conclusions*

We analyzed in a combined way how the patients' characteristics influence the performance of health care at hospital level.

In the first level of our analysis we considered only the patients' variable, and with this variable we formulated a ranking of hospitals through a ranking of the residuals from the multilevel model. This practice was criticized, but it is still one of the most common. The error at hospital level is the most interesting and useful variable we observed, and we did obtain some results from this first step. But we went further on, because we wanted to provide a new approach to this ranking: so we decided to make a second step ahead, and analyze the relationship between the outcome and the input given, in order to revise the ranking of hospitals.

It was interesting to see that inadequacy cannot be explained, but can only be charged to technical inefficiency. This practice, used in Economics with the stochastic frontier analysis, was integrated with the common practice used in Statistics to analyze hierarchical data.

The results showed that there are a lot of hospitals with poor performance, while few are very good. Among these hospitals we identified two types of excellence: large public hospitals and small private profit hospitals. In general terms, with a t-test for the Mean we found out that public structures are better than private ones.

This work is only a first attempt to combine two models or better two scientific subjects: the statistical science and the social science of Economics. We lack the study of the conditions in which we formulate the model, and also the possibility to cooperate in a simultaneous model, as we propose in a subsection for future development.

What we wish to keep as a result of the whole work, and as an innovative contribution, is the concept of a model that takes into account both the hierarchical structure of data and the distance of hospitals from their productivity level. We can say that our intention was to consider effectiveness as efficiency, along with the related economic and performance aspects.

# Bibliography

Afriat, S. (1972). Efficiency estimation of production functions. *International Economic Review*, **13**(3), 568–598.

Aigner, D. and Chu, S. (1968). On estimating the industry production function. *The American Economic Review*, **58**(4), 826–839.

Aigner, D. and Lovell, C. (1977). P. Schmidt, 1977,Formulation and estimation of stochastic frontier production function models,. *Journal of Econometrics*, **6**, 21–37.

Alker, H. (1969). A typology of fallacies. *Quantitative ecological analysis in the social sciences*, pages 69–86.

Battese, G. and Coelli, T. (1992). Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. *Journal of productivity analysis*, **3**(1), 153–169.

Battese, G. and Coelli, T. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical economics*, **20**(2), 325–332.

Carlin, J., Wolfe, R., Brown, C., and Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics*, **2**(4), 397.

Coelli, T. (1995). Estimators and hypothesis tests for a stochastic frontier function: a Monte Carlo analysis. *Journal of Productivity Analysis*, **6**(3), 247–268.

Coelli, T., Henningsen, A., and Henningsen, M. (2009). Version 0.996-0 Date 2009-11-11 Title Stochastic Frontier Analysis.

Cornwell, C., Schmidt, P., and Sickles, R. (1990). Production frontiers with cross-sectional and time-series variation in efficiency levels.

Davis, P. and Scott, A. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, **21**, 99–106.

Farrell, M. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, **120**(3), 253–290.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **159**(3), 505–513.

Goldstein, H. and Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **159**(3), 385–443.

Gong, B. and Sickles, R. (1989). Finite sample evidence on the performance of stochastic frontier models using panel data. *Journal of Productivity Analysis*, **1**(3), 229–261.

Greene, W. (1980). On the estimation of a flexible frontier production model. *Journal of Econometrics*, **13**(1), 101–115.

Greene, W. (1990). A gamma-distributed stochastic frontier model. *Journal of econometrics*, **46**(1-2), 141–163.

Horrace, W. and Schmidt, P. (1996). Confidence statements for efficiency estimates from stochastic frontier models. *Journal of Productivity Analysis*, **7**(2), 257–282.

Hox, J. (1995). *Applied multilevel analysis*. Citeseer.

Hox, J. and NetLibrary, I. (2002). *Multilevel analysis: Techniques and applications*. Hogrefe & Huber.

Imai, K., King, G., and Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, **171**(2), 481–502.

Jha, A., Li, Z., Orav, E., and Epstein, A. (2005). Care in US hospitals-the hospital quality alliance program.

Jondrow, C. *et al.* (1982). On the estimation of technical inefficiency in the stochastic frontier production function model* 1. *Journal of econometrics*, **19**(2-3), 233–238.

King, G. (1997). Reconstructing individual behavior from aggregate data: a solution to the ecological inference problem.

Kumbhakar, S. (1990). Production frontiers, panel data, and time-varying technical inefficiency. *Journal of Econometrics*, **46**(1-2), 201–211.

Kumbhakar, S. and Hjalmarsson, L. (1993). Technical efficiency and technical progress in Swedish dairy farms. *The Measurement of Productive Efficiency Techniques and Applications*, pages 257–270.

Lee, L. and Tyler, W. (1978). The stochastic frontier production function and average efficiency: An empirical analysis. *Journal of Econometrics*, **7**(3), 385–389.

Lee, Y. and Schmidt, P. (1993). A production frontier model with flexible temporal variation in technical efficiency. *The Measurement of Productive Efficiency: Techniques and Applications*, pages 237–255.

Leyland, A. (1995). Examining the relationship between length of stay and readmission rates for selected diagnoses in Scottish hospitals. *Mathematical Medicine and Biology*, **12**(3-4), 175.

Leyland, A. and Boddy, F. (1998). League tables and acute myocardial infarction. *The Lancet*, **351**(9102), 555–558.

Lilford, R. and Pronovost, P. (2010). Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *British Medical Journal*, **340**(apr19 2), c2016.

Lilford, R., Mohammed, M., Spiegelhalter, D., and Thomson, R. (2004). Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *The Lancet*, **363**(9415), 1147–1154.

Lilford, R., Brown, C., and Nicholl, J. (2007). Use of process measures to monitor the quality of clinical practice. *BMJ*, **335**(7621), 648.

Mant, J. and Hicks, N. (1995). Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *British Medical Journal*, **311**(7008), 793.

Meeusen, W. and van Den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, pages 435–444.

Nicholl, J. (2007). Case-mix adjustment in non-randomised observational evaluations: the constant risk fallacy. *British Medical Journal*, **61**(11), 1010.

Normand, S., Glickman, M., and Ryan, T. (1995). Modelling mortality rates for elderly heart attack patients: profiling hospitals. *The Co-Operative Cardiovascular Project. Case Studies in Bayesian Statistics, Springer Verlag, New York*, pages 435–456.

Olson, J., Schmidt, P., and Waldman, D. (1980). A Monte Carlo study of estimators of stochastic frontier production functions. *Journal of Econometrics*, **13**(1), 67–82.

Pitt, M. and Lee, L. (1981). The measurement and sources of technical inefficiency in the Indonesian weaving industry. *Journal of Development Economics*, **9**(1), 43–64.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J., Saunders, L., Beck, C., Feasby, T., and Ghali, W. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, **43**(11), 1130.

Richmond, J. (1974). Estimating the efficiency of production. *International economic review*, **15**(2), 515–521.

Robinson, W. (2009). Ecological correlations and the behavior of individuals. *International journal of epidemiology*, **38**(2), 337.

Schmidt, P. and Lin, T. (1984). Simple tests of alternative specifications in stochastic frontier models. *Journal of Econometrics*, **24**(3), 349–361.

Siciliani, L. (2006). Estimating technical efficiency in the hospital sector with panel data: A comparison of parametric and non-parametric techniques. *Applied Health Economics and Health Policy*, **5**(2), 99–116.

Siddiqui, O., Hedeker, D., Flay, B., and Hu, F. (1996). Intraclass correlation estimates in a school-based smoking prevention study: outcome and mediating variables, by sex and ethnicity. *American Journal of Epidemiology*, **144**(4), 425.

Snijders, T. and Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE publications Ltd.

Stevenson, R. (1980). Likelihood functions for generalized stochastic frontier estimation. *Journal of econometrics*, **13**(1), 57–66.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Verlag.

Vittadini, G. (2006). Ex-post Evaluation and Relative Effectiveness of Health Structures: an Overview. In *Towards Quality of Life Improvement, Proceedings of the Third International Conference*, pages 132–151.

Zaslavsky, A. (2001). Statistical issues in reporting quality data: small samples and casemix variation. *International Journal for Quality in Health Care*, **13**(6), 481.