

Appendice I

PAOLO BERTA, GIANMARIA MARTINI, GIORGIO VITTADINI

SOMMARIO: 1. Metodologie statistiche di misurazione dell'efficacia e dell'efficienza e letteratura scientifica di riferimento, 77 – 2. La valutazione multilevel: caratteristiche e metodologia, 81 – 3. L'efficienza tecnica e allocativa, 82 – 4. La stima dell'efficienza tecnica, 84 – 5. La metodologia della frontiera stocastica, 85.

1. Metodologie statistiche di misurazione dell'efficacia e dell'efficienza e letteratura scientifica di riferimento

Quando si analizza la qualità delle singole strutture sanitarie in termini competitivi il concetto di efficacia che risulta più utile è evidentemente quello di efficacia relativa che confronta diverse strutture sanitarie, permettendo di valutare la loro capacità di modificare in senso migliorativo lo stato di salute del paziente fino ad avvicinarsi alla *best practice* o *effectiveness* teorica massima (Donabedian A. 1990; Pagano A. e Rossi C. 1999).

Quali possono essere gli strumenti statistici più adatti a tali scopi (Rice N, Leyland A. 1996; Leyland A.H., Boddy F.A. 1998; Vittadini G. 2000; Vittadini G. e Sanarico M. 2002; Lillford RJ, Mohammed MA, Spiegelhalter D, Thomson R. 2004)?

Inadeguati sono gli indicatori univariati non ponderati di origine aziendale riferiti alle strutture sanitarie, anziché ai pazienti. Anche quando considerano lo stato di salute dei singoli pazienti dopo la cura in termini dicotomici (morti/vivi) o politomici (grado di salute) non tengono conto dello stato di salute iniziale dell'utente. Questo fatto non solo distorce la valutazione dell'efficacia relativa, ma può portare a situazioni eticamente inaccettabili quando le strutture sanitarie attuano atteggiamenti di selezione avversa e scelgono di curare preferibilmente o esclusivamente i pazienti meno gravi (Goldstein H. e Spiegelhalter D.J. 1996). Occorre, perciò, effettuare confronti a

parità di condizioni (o come spesso si dice *coeteris paribus*), tenendo conto dell'effetto del "case mix" dei pazienti — vale a dire delle caratteristiche cliniche e socio-demografiche — sul valore dell'*outcome* sanitario (Zaslowsky A. 2001) mediante metodologie denominate di *risk adjustment* (Dubois R., Brook R. e Rogers W. 1987; De Long et al. 1997; Iezzoni L. 1998).

In una prima famiglia di procedure di "standardizzazione" (Zaslowsky A. 2001) l'*outcome* sanitario viene ponderato per tener conto che i diversi gruppi di pazienti sono caratterizzati da particolari case mix ¹.

Dati:

x_{kl} = indicatore che definisce il case mix osservato rispetto all' ℓ -esima caratteristica per lo strato k -esimo della popolazione standard $P^{(t)}$ ($k = 1, \dots, K$).

y_{kj} = *outcome* osservato per lo strato k -esimo (rispetto all' ℓ -esima caratteristica) della struttura sanitaria j -esima. $p_{kl}^{(T)}$ = peso relativo dello strato k -esimo individuato dalla caratteristica ℓ -esima della popolazione standard $P^{(t)}$ si ottiene il seguente sistema di pesi:

$$\pi_{kl}^{(t)} = \frac{P_{kl}^{(T)}}{\sum_i P_{kl}^{(T)}} \quad (1)$$

L'*outcome* standardizzato \tilde{Y}_j per la struttura sanitaria j -esima, ($j = 1, \dots, J$) è quindi:

$$\tilde{Y}_j = \sum_k \pi_{kl}^{(t)} y_{kl} \quad (2)$$

La metodologia presenta alcuni problemi: ad esempio non può essere applicata agli *outcomes* che non sono discreti e qualora, a causa di dati mancanti, l'indicatore x_l abbia frequenza nulla rispetto ad un determinato strato la scelta della popolazione standard può essere arbitraria; inoltre, sul piano operativo, risulta arduo ottenere strati con numerosità sufficiente per poter fare confronti (Zaslowsky A. 2001).

1. Ad esempio l'associazione per le cure mediche dello stato di New York utilizza la standardizzazione diretta per ogni indicatore Healthcare Cost and Utilization Project (HCUP) relativo a complicanze (Jang H.J. e Stayer D. 2003).

Una seconda famiglia di metodi di tipo aritmetico è detta della standardizzazione indiretta² (Zaslowsky A. 2001).

Definiti come in precedenza $x_{\ell i}$, $p_{i\ell}^{(T)}$ e anche $y_i^{(T)}$ = valore dell'*outcome* nello strato i -esimo individuato dalla caratteristica x_{ℓ} nella popolazione standard $P^{(T)}$.

p_{ij} = peso relativo dello strato i -esimo individuato dalla caratteristica ℓ -esima per la struttura sanitaria j -esima si ottiene:

$$\hat{Y}_j = \frac{\sum_i y_{ij}^{(T)} * p_{ij}}{\sum_i p_{ij}} \quad (3)$$

con \hat{Y}_j *outcome* standardizzato in modo indiretto per la struttura sanitaria j -esima, ($j = 1, \dots, J$).

Anche la standardizzazione indiretta presenta problemi: permane l'arbitrarietà legata alla scelta della popolazione e la limitazione ad indicatori discreti.

Nel complesso, gli indicatori utilizzati per la standardizzazione diretta e indiretta presentano i seguenti limiti:

- Sono riferiti agli ospedali e non ai pazienti.
- Sono spesso espressi come frequenza.

I metodi che hanno trovato più consenso³ sono quelli basati su modelli lineari che permettono di ricavare per ogni ospedale *outcomes* “corretti” al netto degli effetti del diverso case mix.

Nella loro più semplice formulazione, in presenza di *outcomes* dicotomici, si hanno regressioni logistiche del tipo

$$\ln \left(\frac{p_{ij}}{1 - p_{ij}} \right) \alpha + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \dots + \beta_k x_{kij} = RS_{ij} \quad (4)$$

esprimibile con opportuni passaggi attraverso la seguente formula

2. Ad esempio nell'ambito del programma di benchmark della Commission for Health Improvement (CHI) del National Health Service (2004) si propone la standardizzazione indiretta di alcuni outcome clinici.

3. Il modello logistico è utilizzato come metodo di risk adjustment nei sistemi di indicatori proposti da ORYX-Joint Commission on Accreditation of Healthcare Organisations (2001), Agency for Healthcare Research Quality (2003), Canadian Institute for Health Information (2003), Solucient (2003).

$$p_{ij} = \frac{e^{a + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \dots + \beta_k x_{kij}}}{1 + e^{a + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \dots + \beta_k x_{kij}}} = \frac{e^{RS_{ij}}}{1 + e^{RS_{ij}}} \quad (5)$$

ove per il paziente i -esimo curato nel j -esimo ospedale, p_{ij} è la probabilità del verificarsi dell'evento favorevole per un *outcome* sanitario dicotomico per l'individuo, x_1, \dots, x_k sono gli indicatori che descrivono il case mix di caratteristiche dei pazienti, la quantità a destra dell'uguale in (4) RS_{ij} è detta "risk score" (De Long E.R. 1997).

Il "risk score atteso" può essere quindi calcolato per ciascun paziente sulla base del valore assunto dalle variabili (X_1, \dots, X_K) in (5). Si ottengono, attraverso il modello (4) stime per i parametri β_1, \dots, β_k e moltiplicate tali stime per i valori osservati $(x_{1ij}, \dots, x_{kij})$ nella formula (5), si ottiene la probabilità \hat{p}_{ij} del verificarsi dell'evento atteso favorevole per il paziente i -esimo ospitato nella j -esima struttura sanitaria dato il suo case mix. Paragonata con la probabilità dell'evento osservato $(0,1)$ \hat{p}_{ij} fornisce una stima dell'efficacia dell'ospedale j -esimo sull'individuo i -esimo.

Definito a questo punto il valore atteso degli eventi favorevoli per i pazienti dell'ospedale j -esimo,

$$E(\hat{p}_{ij}) = \sum_j^n \hat{p}_{ij} \quad (6)$$

e il corrispondente valore della probabilità degli eventi osservati $\Sigma(p_{ij})$ il rapporto

$$\frac{E(p_{ij})}{E(\hat{p}_{ij})} \quad (7)$$

dà una stima dell'efficacia dell'ospedale j -esimo.

I modelli lineari non tengono conto della struttura gerarchica dei dati e questo comporta alcune problematiche nella stima dei modelli e nell'interpretazione dei risultati, in quanto non sussiste più l'indipendenza delle osservazioni. Si parte da un campionamento a stadi, con dipendenza tra le osservazioni appartenenti alla stessa sottopopolazione a differenza del modello classico di regressione che prevede l'indipendenza tra le osservazioni.

Il modello statistico più avanzato per studi di valutazione anche in sanità è però il modello *multilevel* (Hox J.J. 1995; Rodriguez, Goldman 1995; Goldstein 1995; Goldstein, Rasbash 1996; Sneijder, Bosker 1999;

Guo, Zao, 2000; Leyland, Goldstein 2001; Cohen et al. 2003; Hope, Shannon 2005; Gelman, Hill, 2007) proposto in modo organico da Goldstein e Spiegelhalter (1996). Il modello, denominato “a effetti misti”, permette di valutare l’efficacia relativa perché, grazie alla struttura gerarchica dei dati, enuclea la variabilità tra e entro gli ospedali, *coeteris paribus*, al netto delle caratteristiche dei pazienti, superando nello stesso tempo i problemi di mancanza di informazione dovuti a piccoli campioni⁴.

2. La valutazione *multilevel*: caratteristiche e metodologia

Da un punto di vista metodologico, la forma assunta dal modello *multilevel* è la seguente:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij} \quad (1)$$

dove Y_{ij} misura l’*outcome* relativo all’*i*-esimo soggetto inserito nel *j*-esimo ospedale; β_{0j} è l’usuale intercetta di un modello lineare, X_{ij} è il vettore di variabili esplicative con parametri β_{1j} , mentre ε_{ij} sono i residui indipendenti e di media nulla ($\mu = 0$) a livello del soggetto.

L’intercetta β_{0j} fornisce una misura dell’efficacia relativa dell’unità al livello superiore al netto dei fattori di rischio X_{ij} . Nel modello descritto si ipotizza che ogni unità del livello superiore abbia una differente intercetta e un differente coefficiente associato alle variabili esplicative, per questo tali coefficienti sono definiti casuali (random).

Ipotizzato che i parametri β_{0j} e β_{1j} siano variabili casuali con varianza costante e distribuzione nota, è possibile inserire nel modello anche variabili al livello superiore; in questo caso l’equazione (1) si può esprimere nel seguente modo:

$$y_{ij} = [\beta_{00} + \beta_{10}X_{ij} + \beta_{01}Z_j + \beta_{11}Z_jX_{ij}] + [u_{1j}X_{ij} + u_{0j} + \varepsilon_{ij}] \quad (2)$$

$$\text{essendo } \beta_{0j} = \gamma_{00} + \gamma_{01}Z_{0j} + u_{0j} \quad (3)$$

4. Grazie all’introduzione dei “shrinkage residuals”: le stime provenienti dai diversi sottocampioni pesano sulla stima dei parametri in proporzione alla numerosità dei sottocampioni medesimi.

$$\text{ed essendo } \beta_{ij} = \gamma_{10} + \gamma_{11}Z_{ij} + u_{ij} \quad (4)$$

dove le Z_j sono variabili esplicative al livello dell'ospedale.

La distribuzione della variabile dipendente condiziona la scelta del modello lineare: laddove l'*outcome* considerato fosse dicotomico, ad esempio, il modello lineare assumerebbe le caratteristiche di un modello di regressione logistica.

Si considera che u_{oj} sia una variabile casuale che individua l'efficacia relativa della struttura ospedaliera j -esima al netto dei fattori di rischio introdotti come predittori nel modello. Si assume che i residui del modello ε_{ij} siano incorrelati con valore atteso nullo:

$$E(\varepsilon_{ij}) = 0; \text{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2 \quad (5)$$

Si ipotizza inoltre che il parametro casuale u_{oj} abbia distribuzione $\varphi(u_{oj})$ con varianza costante:

$$\text{Var}(u_{oj}) = \sigma_u^2 \quad (6)$$

Tale modello, come abbiamo detto, è particolarmente appropriato per valutare l'efficacia relativa perché permette di enucleare la variabilità tra e entro gli ospedali.

La distribuzione a posteriori del parametro u_{oj} fornisce il tasso relativo all'*outcome* considerato per l'ospedale j -esimo che tiene conto delle differenze tra pazienti e tra strutture sanitarie.

3. L'efficienza tecnica e allocativa

Le definizioni di efficienza tecnica ed allocativa riportate nel testo sono illustrate graficamente nella figura seguente, che rappresenta un esempio di funzione di produzione in cui due input, x_1 e x_2 , sono utilizzati per produrre un output, y . L'isoquanto, denominato con y_0 , indica tutte le combinazioni di x_1 e x_2 che danno luogo allo stesso livello di produzione. La pendenza dell'isoquanto misura il tasso marginale di sostituzione tra i fattori della produzione, ossia di quanto deve essere ridotto l'impiego di x_2 se aumenta di un'unità l'impiego di x_1 , per mantenere costante il livello di output. La retta di isocosto,

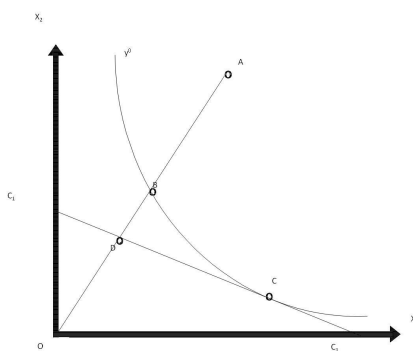


Figura 4. Efficienza tecnica e allocativa.

denominata con C_1 , indica le combinazioni di x_1 e x_2 che determinano lo stesso livello di spesa. La pendenza dell'isocosto è pari al rapporto tra i prezzi dei fattori, $-p_1/p_2$, dove p_i indica il prezzo del fattore produttivo i .

L'efficienza tecnica si ottiene quando non vi è un utilizzo eccessivo dei fattori produttivi. Si supponga che A e B siano due combinazioni input-output cui è associato il medesimo livello produttivo y_0 . La combinazione B è caratterizzata da efficienza tecnica, in quanto è posizionata sull'isoquante; A, al contrario, è caratterizzata da inefficienza tecnica, in quanto è più lontana dall'origine rispetto a B.

Si ha, invece, efficienza allocativa quando la combinazione degli input è tale da minimizzare il costo di produzione per un dato livello di output, o, viceversa, è tale da massimizzare il livello di output per un dato livello di spesa. Nella figura 4 ciò si verifica nel punto C, dove le pendenze dell'isoquante e dell'isocosto coincidono.

Se la situazione di una struttura è rappresentata dal punto A, essa opera in una situazione di inefficienza, sia tecnica che allocativa. Il suo grado di efficienza tecnica è misurato dal rapporto OB/OA , che indica la proporzione degli input utilizzati che è effettivamente necessaria per la produzione del livello di output osservato. L'inefficienza tecnica è data da $1 - (OB/OA)$, valore che indica di quanto potrebbe essere ridotto il costo di produzione del livello di output osservato, mantenendo invariato il rapporto tra i fattori. Il livello di efficienza allocativa è misurato dal rapporto OD/OB , che rappresenta il costo che si sarebbe sostenuto in presenza di efficienza allocativa in rapporto a quello

effettivo. L'inefficienza allocativa è data da $1 - (OD/OB)$. L'efficienza produttiva è definita come il rapporto OD/OA ed è il prodotto tra efficienza tecnica ed allocativa. L'inefficienza complessiva è data da $1 - (OD/OA)$ e misura di quanto i costi eccedono il loro minimo potenziale; essa può essere scomposta nella somma dell'inefficienza allocativa e quella tecnica.

4. La stima dell'efficienza tecnica

Per stimare una frontiera di produzione deterministica, occorre specificare una funzione di produzione $y_i = f(x_i) + e_i$, dove l'indice i identifica la struttura sanitaria, y_i è l'*output* prodotto, x_i sono gli *input* utilizzati ed il residuo e_i rappresenta il termine d'errore. Se si vuole identificare da tale termine di errore l'inefficienza della struttura, occorre identificare una frontiera. L'inefficienza è infatti definita come distanza da tale frontiera. Tale misura deve essere per definizione non negativa. Per stimare l'inefficienza con una frontiera deterministica si adotta il seguente approccio: una volta stimata la frontiera con il metodo dei Minimi Quadrati Ordinari (OLS), occorre calcolare il valore massimo tra gli scarti e traslare verso l'alto la linea di interpolazione OLS di una quantità pari al massimo residuo. Il metodo è noto in letteratura come Corrected Ordinary Least Squares (COLS) ed è illustrato nella figura 5.

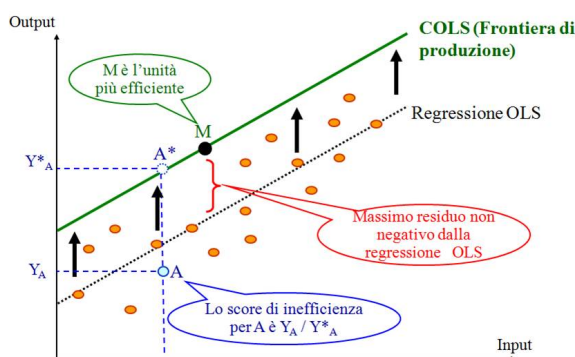


Figura 5. Stima della funzione di produzione con il metodo COLS.

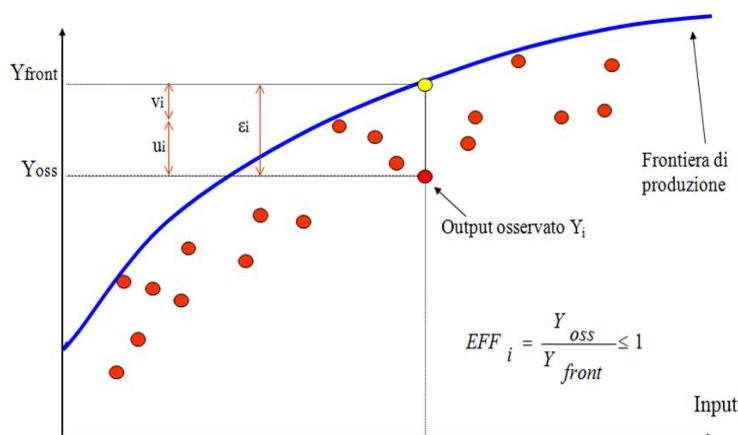


Figura 6. Stima della funzione di produzione con il metodo SFA.

Qualora, invece, si voglia stimare una frontiera di produzione stocastica, la tecnica comunemente utilizzata è la Frontiera Stocastica (SFA). A differenza del caso deterministico, in questo caso bisogna specificare una funzione di produzione $y_i = f(x_i) + v_i + u_i$, dove y_i è l'*output* prodotto dall'ospedale i -esimo e x_i sono gli *input* utilizzati. Il termine v_i è una variabile casuale con distribuzione simmetrica e cattura gli effetti casuali di errori di misurazione e di *shock* esogeni, che fanno sì che la posizione degli ospedali vari rispetto alla parte deterministica della frontiera di produzione $f(x_i)$. L'inefficienza tecnica rispetto alla frontiera stocastica è misurata dalla variabile casuale u_i ; la condizione $u_i \leq 0$ assicura che tutte le osservazioni si collochino in corrispondenza della frontiera stocastica o sotto di essa (figura 6).

5. La metodologia della frontiera stocastica

Una struttura sanitaria può essere rappresentata approssimativamente come una unità produttiva che trasforma lavoro e capitale (*input*) in servizi di assistenza sanitaria per gli utenti (*output*).

Nella definizione del modello, oltre agli *input* del processo produttivo, sono state considerate anche alcune caratteristiche degli ospedali ed alcuni comportamenti adottati dagli stessi in grado di spiegare l'ete-

Tabella 5. Le variabili del modello

Tipologia di variabile	Descrizione
Output dell'ospedale	Numero equivalente di ricoveri corretto per la complessità
Input dell'ospedale	Letti Medici Infermieri Altro personale
Caratteristiche dell'ospedale	Presenza PS Tipologia di Ownership Monospecialistico Universitario Appartenenza ad un gruppo
Comportamenti opportunistici	Upcoding Cream skinning Ricoveri ripetuti
Trend temporale	Trend temporale

rogenità nell'output prodotto. La frontiera di produzione specificata può essere rappresentata dalla seguente funzione di produzione:

$$Y = f(I, CAR, COMP, T)$$

dove Y rappresenta l'output prodotto, I il vettore degli input utilizzati per produrlo, CAR un insieme di caratteristiche dell'ospedale, $COMP$ un vettore di variabili che misurano il ricorso ad alcuni comportamenti opportunistici e T è uno scalare che rappresenta il trend temporale e cattura l'impatto di progresso tecnico.

L'elenco delle variabili considerate per stimare l'efficienza tecnica degli ospedali è mostrato nella tabella 5.

La variabile dipendente è costituita dall'output prodotto dagli ospedali. La natura multi-output del processo produttivo degli ospedali rende necessaria la costruzione di un indice sintetico che consenta di aggregare i diversi output del processo produttivo, riconducibili essenzialmente a ricoveri ordinari, ricoveri diurni e prestazioni di specialistica ambulatoriale. L'indice viene costruito calcolando il numero equivalente di ricoveri ordinari attraverso la normalizzazione dei ricavi per attività di day hospital ed ambulatoriale. L'indice così ottenuto viene successivamente corretto per il peso medio dell'attività di degenza ordinaria (peso dei DRG) in modo tale da tenere in considerazione

differenze nella complessità della casistica trattata dai diversi ospedali. Pertanto, l'output degli ospedali è dato dalla seguente espressione:

$$y_{it}^* = y_{it}^{DO} \times \left(1 + \frac{R_{it}^{DH} + R_{it}^{AMB}}{R_{it}^{DO}} \right) \times PM_{it}^{DO}$$

dove y_{it}^* è il numero equivalente di ricoveri corretto per la complessità dell'ospedale i all'anno t e per l'intensità della tipologia di attività svolta, y_{it}^{DO} è il numero di ricoveri ordinari, R_{it}^{DO} sono i ricavi per attività di degenza ordinaria, R_{it}^{DH} i ricavi per attività di day hospital, R_{it}^{AMB} i ricavi per attività ambulatoriale e PM_{it}^{DO} è il peso medio dell'attività di degenza ordinaria.

Le variabili relative agli input riguardano il numero di posti letto, quale proxy dello stock di capitale dell'azienda, ed il *full time equivalent* del personale, classificato sinteticamente in medici, infermieri e altro personale.

Inoltre, sono state incluse nell'analisi alcune caratteristiche degli ospedali ritenute potenzialmente rilevanti per spiegare differenze nel livello di efficienza raggiunto: la presenza di un dipartimento per la gestione delle emergenze, la vocazione monospecialistica o plurispecialistica degli ospedali, la presenza di dipartimenti a gestione universitaria e l'appartenenza a gruppi di strutture gestiti dalla stessa autorità.

Da ultimo è stata inclusa una variabile temporale per cogliere l'effetto del progresso tecnologico.

Una volta definito il modello, si procede alla sua stima. Il modello di frontiera di produzione è basato su un modello ad effetti casuali e per valutare la robustezza dei risultati, sono considerate due forme funzionali, una funzione di produzione Cobb–Douglas e una funzione di produzione Translog (Christensen L., Jorgenson D., e Lau L., 1973; Meeusen D., Van den Broeck, 1977; Kim H. Y., 1992; Grant J. H., 1993).

La metodologia della Frontiera Stocastica è stata sviluppata da Aigner, Lovell e Schmidt nel 1977 e si fonda sulla seguente formulazione canonica (Aigner D. J., Lovell C. A. e Schmidt P., 1977):

$$y = \beta'x + v - u$$

dove y è l'*outcome* osservato, $\beta'x + v$ è l'obiettivo ottimo (ad esempio la massima quantità producibile o il minimo costo) perseguito

dall'impresa o dall'organizzazione (nel nostro caso la struttura ospedaliera), $\beta'x$ è la parte deterministica della frontiera e $v[0, \sigma_v^2]$ è la parte stocastica. L'unione delle due parti costituisce nel suo complesso la frontiera stocastica. La quantità per cui l'osservato si discosta dall'ottimo, ovvero dalla frontiera, è u , che è definita nel seguente modo:

$$u = |U| \quad \text{e} \quad U \sim N[0, \sigma_u^2]$$

(Nelle frontiere di costo e in tutti gli ambiti in cui l'ottimo sia un minimo la formulazione della frontiera cambia in $y = \beta'x + v + u$.)

In questo contesto, u è l'inefficienza. La variabile casuale u si distribuisce come una semi normale.

Molte varietà di modelli di frontiera stocastica sono stati presentati in letteratura nel corso degli anni. Per una review sistematica della letteratura si vedano Kumbhakar e Lovell (2000) e Greene (2005). Di seguito vengono presentati i due modelli econometrici utilizzati per dati panel, ovvero per osservazioni ripetute nel corso degli anni, impiegati negli studi di efficienza dalla Regione Lombardia.

Pitt e Lee nel 1981 propongono una specificazione ad effetti casuali per dati panel basata sul seguente modello (Pitt M., Lee L., 1981):

$$y_{it} = \alpha + \beta'x_{it} + v_{it} - u_i$$

dove i indica l'ospedale e $t = 1, \dots, T$ denota l'anno. La variabile dipendente y_{it} è l'output osservato dell'ospedale i al tempo t , α è una costante, β un vettore di parametri e x_{it} un vettore di covariate osservate per l'ospedale i al tempo t . Il termine di errore è scomposto in due componenti: il termine v_{it} rappresenta il residuo *white noise*, mentre il termine u_i rappresenta lo score di inefficienza degli ospedali. Il modello viene stimato con tecnica della massima verosimiglianza (MLE). Il caso base è il modello in cui v_{it} è distribuito normalmente mentre u_i ha una distribuzione seminormale con $u_i = |U_i|$, $U_i \sim N[0, \sigma^2]$.

Il modello assume che la componente di inefficienza u_i sia specifica dell'ospedale ma non cambia nel tempo, questo significa che la regressione individua il livello di inefficienza di ogni ospedale, e che tale livello rimane costante lungo tutto il periodo di analisi.

Nel 2005 Greene propone un modello ad effetti casuali denominato True Random Effects Model (Greene W., 2005a; 2005b) basato sulla seguente specificazione:

$$y_{it} = \alpha + \beta' x_{it} + w_i + v_{it} - u_{it}$$

dove w_i rappresenta l'eterogeneità non osservata specifica dell'ospedale i (distribuita normalmente), v_{it} è la componente white noise dell'errore (con distribuzione normale) e $u_{it} \geq 0$ è l'inefficienza dell'ospedale i (con distribuzione semi normale). Il modello è stimabile con la massima verosimiglianza simulata. In questo modello, è possibile stimare l'inefficienza degli ospedali per ogni anno (non è costante nel tempo) ed è possibile distinguerla dalla eterogeneità latente dell'ospedale.

Una volta stimata la frontiera efficiente, si studiano i coefficienti della regressione e la loro significatività statistica. Ciò fornisce informazioni sull'impatto, positivo o negativo, delle diverse variabili considerate sull'efficienza degli ospedali, e sull'importanza relativa dei vari input e sulle elasticità di impiego.

Paolo Berta, Gianmaria Martini, Giorgio Vittadini