



**Comparing methods for measurement error  
correction in survival analysis when the  
variable of interest is change over time**

**Giovanni Veronesi**

**Università degli Studi dell'Insubria - Varese**

V Convegno Nazionale SISMEC

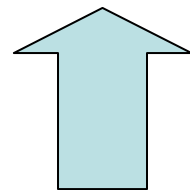
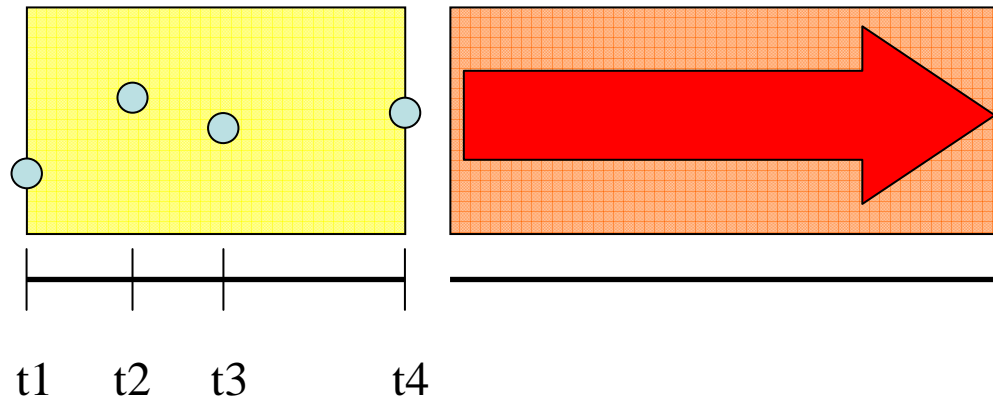
Pavia – 18/09/2009

# *Introduzione*

---

*Dati longitudinali*

*Follow-up*



■ ***Measurement error (ME)***: errore casuale introdotto nella determinazione di una quantità di interesse

## *Revisione della letteratura*

---

■ *Effetti del ME:* in generale, la presenza di ME provoca *sottostima dell'effetto desiderato*

■ *Metodi di correzione:* Regression Calibration, SIMulation-EXtrapolation (SIMEX)

■ *ME e dati longitudinali:* assunti sui parametri (mixed models); contesto di missing data imputation (stima del valore reale sulla base di quello osservato)

## *Scopi della ricerca*

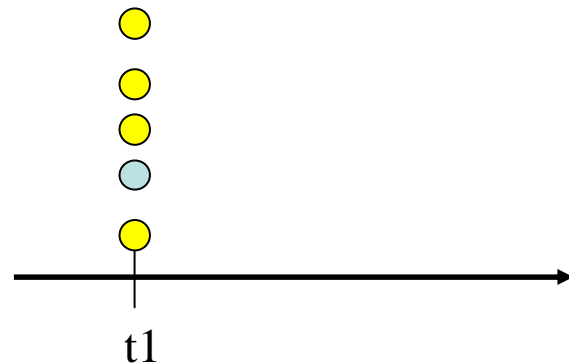
---

- *Generalizzare RC e SIMEX* alla situazione in cui la variabile misurata con errore sia il *tasso di cambiamento nel tempo*
- *Studiare il comportamento di RC e SIMEX* nei modelli di *analisi di sopravvivenza*, con particolare attenzione alla riduzione del bias, *tramite simulazioni*.

## *Errore di Misura - singola rilevazione*

---

i-esimo soggetto:



*Valore “vero”*  $X_i$  (fisso entro soggetto)

*Valore “osservato”*  $W_{i,k}$  (varia entro replicazioni nel breve periodo)

*Errore di Misura: variabilità casuale* introdotta nella misurazione di un fenomeno di interesse (lettura, variabilità biologica, ...)

*Variabilità entro-soggetto (IIV):* il valore osservato  $W_{i,k}$  potrebbe cambiare entro repliche di misura nel breve periodo

## Assunti per la distribuzione dell'errore di misura

**Additive measurement error model:**  $W_{i,k} = X_i + \Gamma_{i,k}$

■  $E(\Gamma_{i,k}) = 0$  e  $\text{Var}(\Gamma_{i,k}) = \sigma_e^2$  costante tra soggetti e nota;

■ Errore di misura indipendente dal “vero” valore e dal valore osservato;

■  $E(W_{i,k}) = E(X_i)$  e  $\text{Var}(W_{i,k}) = \text{Var}(X_i) + \text{Var}(\Gamma_{i,k})$ ;

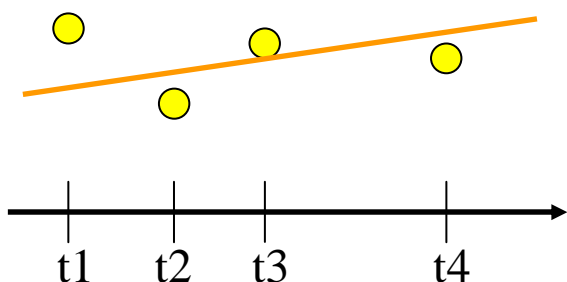
■  $\text{Var}(W_{i,k})$  *costante per soggetti con la stessa IIV*

■  $\text{corr}(W_{i,k}, W_{i,k'}) = \frac{\text{Cov}(W_{i,k}, W_{i,k'})}{\sqrt{\text{Var}(W_{i,k})\text{Var}(W_{i,k'})}} = \frac{\text{Var}(X_i)}{\text{Var}(W_{i,k})} = R$

## Errore di Misura – tasso di cambiamento nel tempo (1)

i-esimo soggetto:

*Additive measurement error model*



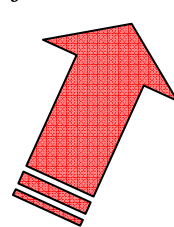
$$W_{ij,k} = X_{ij} + \Gamma_{ij,k}$$

$$\text{Var}(\Gamma_{ij,k}) = \sigma_e^2$$

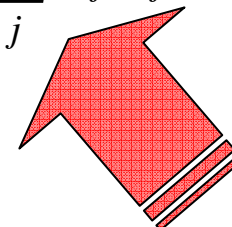
Stima del CA via *regressione lineare entro soggetto*:

$$\hat{B}_{i,k}^W = \sum_j c_{ij} W_{ij,k} = \sum_j c_{ij} X_{ij} + \sum_j c_{ij} \Gamma_{ij} = B_i + \sum_j c_{ij} \Gamma_{ij}$$

dove 
$$c_{ij} = \frac{t_{ij} - \bar{t}}{\sum_j (t_{ij} - \bar{t})^2}$$



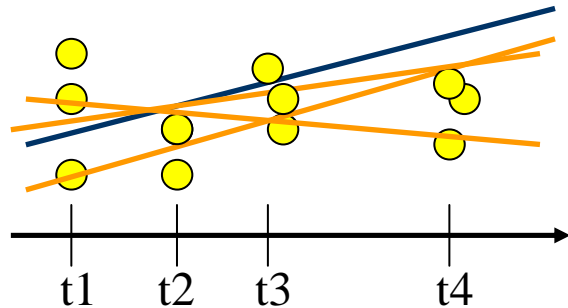
*“Vero” CA*



*Errore di misura*

## Errore di Misura – tasso di cambiamento nel tempo (2)

i-esimo soggetto:



*Additive measurement error model*

$$\hat{B}_{i,k}^W = B_i + \sum_j c_{ij} \Gamma_{ij}$$

■ *Variabilità entro-soggetto per CA:*

$$\text{Var}(\hat{B}_{i,k}^W | i) = \text{Var}\left(\sum_j c_{ij} (X_{ij} + \Gamma_{ij,k}) | i\right) = \sum_j c_{ij}^2 \text{Var}(\Gamma_{ij,k} | i) = \boxed{\frac{\sigma_e^2}{(N_i - 1)S_{t,i}^2}}$$

■ *IIV costante* solo tra soggetti con lo *stesso numero di visite*,  
*a distanza costante* tra le visite

## *Standard regression calibration*

---

Sostituire nel modello il “vero” valore con il suo *miglior predittore, dato quello che ho “osservato”* ed eventuali altre covariate di interesse (error free):

$$E[X|W, Z] = \hat{W}(Z) * (1 - R_Z) + W * R_Z$$

■  $\hat{W}(Z)$  è il *predicted value* di W dato Z

■ *Reliability coefficient:*  $R_Z = \frac{Var(X|Z)}{Var(W|Z)} = 1 - \frac{\sigma_e^2}{MSE(W|Z)}$

■ *RC è stimatore consistente* dell'effetto di interesse

## Regression calibration – tasso di cambiamento nel tempo

*Espressione per la varianza totale di  $\hat{B}_{i,k}^W|Z$  :*

$$\text{Var}(\hat{B}_{i,k}^W|Z) = E[\text{Var}(\hat{B}_{i,k}^W|i, Z)] + \text{Var}[E(\hat{B}_{i,k}^W|i, Z)] = E\left[\frac{\sigma_e^2}{(N_i - 1)S_t^2}\right] + \text{Var}(B_i|Z)$$

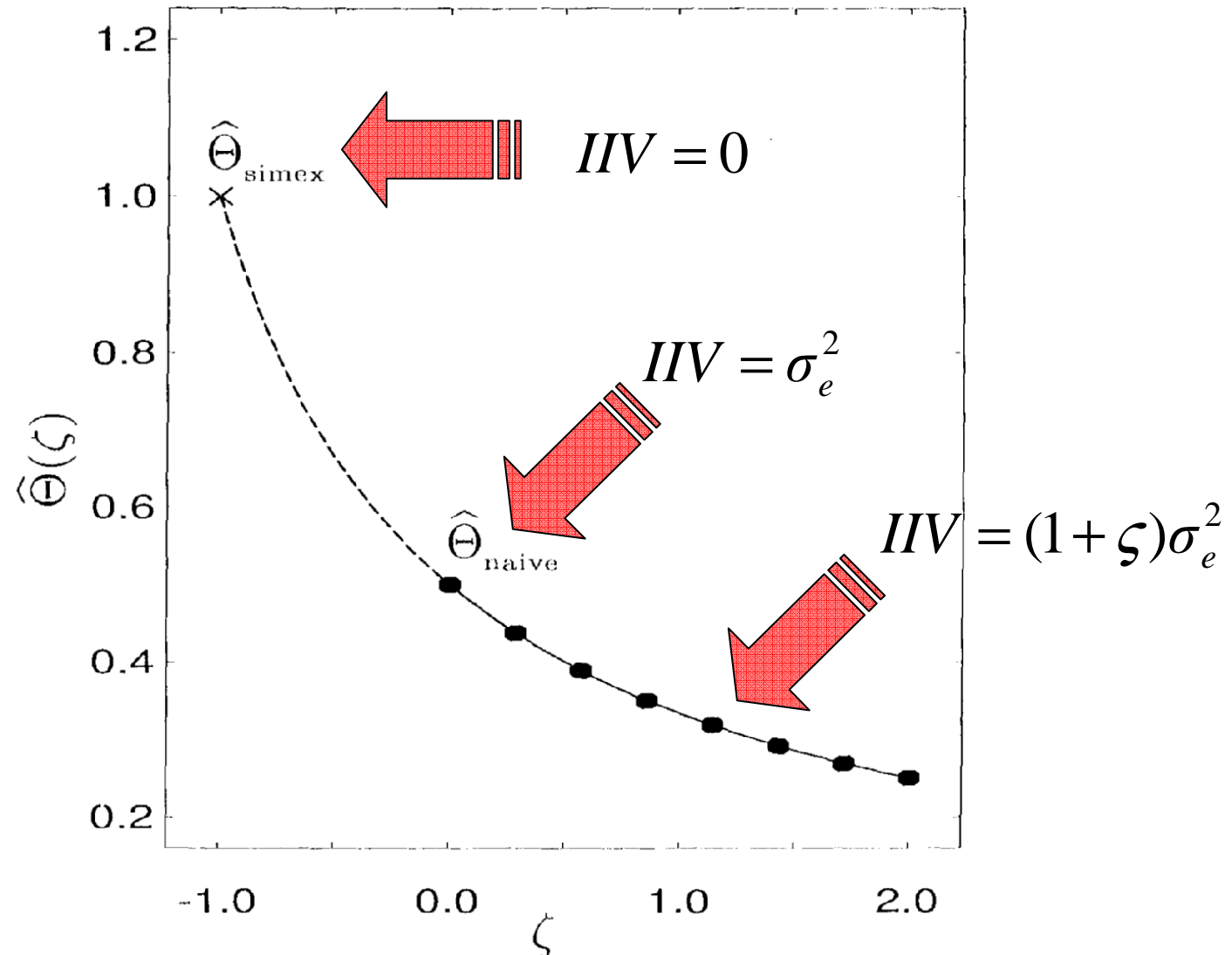
■ **Reliability Coefficient:**  $R_Z = \frac{\text{Var}(B_i|Z)}{\text{Var}(\hat{B}_{i,k}^W|Z)} = 1 - \frac{E[\sigma_e^2 / (N_i - 1)S_{t,i}^2]}{\text{MSE}(\hat{B}_{i,k}^W|Z)}$

■ *R di nuovo costante* tra i soggetti ed esprime ancora la *correlazione tra ripetizioni di misura entro soggetto*

■ *Stimatore consistente* (modelli di regressione lineare)

## *SIMulation EXtrapolation (1)*

---



## *SIMulation EXtrapolation (2)*

---

■ ***SIMulation***: un nuovo valore per CA a partire da quello osservato, con 
$$\text{IIV} = (1 + \zeta) \left[ \sigma_e^2 / (N_i - 1) S_{t,i}^2 \right] \quad \zeta \geq 0$$

■ ***Stimare*** l'effetto di CA sull'outcome di interesse  $\hat{\beta}_X(\zeta)$

■ ***EXtrapolation***: modello di regressione per i valori di  $\hat{\beta}_X(\zeta)$

lo stimatore SIMEX è il *predicted value*  $\hat{\beta}_X(-1)$

■ ***Rifiniture***: ripetere le simulazioni S volte per ciascun valore di  $\zeta$ , e prendere il valore medio per  $\hat{\beta}_{X,s}(\zeta)$

## *Scopi della ricerca*

---

- *Generalizzare RC e SIMEX* alla situazione in cui la variabile misurata con errore sia il *tasso di cambiamento nel tempo*
- *Studiare il comportamento di RC e SIMEX* nei modelli di *analisi di sopravvivenza*, con particolare attenzione alla riduzione del bias, *tramite simulazioni*.

## *Simulazione – settings (1)*

---

- *“Vero” CA e una covariata continua (error-free)* simulate tramite *multi-normal distribution* (correlazione diversa da zero)
- *Simulazione di 3 visite per soggetto*, a  $365 \pm 90$  giorni, da una distribuzione normale. Media 5% delle visite fuori dal range considerate missing (soggetto manca la visita)
- *Follow-up*: a partire dalla visita 3. *Time-to-event e censura* simulati da due distribuzioni esponenziali indipendenti. *Censura amministrativa* applicata.

## *Simulazione – settings (2)*

---

**Modello Target:** Modello di Cox

$$\lambda_i(t) = \lambda_0(t) * \exp(\beta_X \boxed{B_i} + \beta_Z Z_i) \quad \textit{True slope}$$

$$\lambda_i(t) = \lambda_0(t) * \exp(\beta_X \boxed{\hat{B}_i^W} + \beta_Z Z_i) \quad \textit{Naïve}$$

$$\lambda_i(t) = \lambda_0(t) * \exp(\beta_X \boxed{E[B_i | \hat{B}_i^W, Z]} + \beta_Z Z_i) \quad \textit{RC}$$

$$\lambda_i(t) = \lambda_0(t) * \exp(\beta_X \boxed{\hat{B}_{i,s}^W (\zeta = -1)} + \beta_Z Z_i) \quad \textit{SIMEX*}$$

\*Estrapolazione tramite modello di regressione quadratico e per S = 50

## *Simulazione – settings (3)*

---

### ■ *Diverse combinazioni di:*

*Sample Size:*  $n=500$  e  $n=1000$

*Event Rate:* 15% e 30%

*Valore di beta per CA:* 0.10 e 0.35

*Reliability coefficient (cross-sectional):* 0.90 e 0.85

### ■ *Indicatori di performance:*

*Bias:* Media della differenza tra il valore vero e quello stimato;

*SD:* SD della distribuzione dello stimatore

*SE:* Media degli SE stimati dal modello di Cox

*Coverage:* 95% CI include il valore vero (%)

## *Simulazione – risultati (1)*

---

### *Confronto per diversi valori del parametro da stimare*

	True Beta: 0.10			True Beta: 0.35		
	NAIVE	RC	SIMEX	NAIVE	RC	SIMEX
Bias	-0.03	-0.001	-0.008	-0.08	-0.027	-0.015
SD	0.012	0.016	0.015	0.017	0.021	0.025
SE	0.011	0.016	0.014	0.017	0.020	0.023
Coverage	25.2	94.7	88.6	1.0	71.4	82.4

■ *RC migliore di SIMEX* quando l'effetto da stimare è contenuto

■ Performance peggiorano *all'aumentare del valore del parametro* da stimare (in particolare RC)

Sample Size: n=1000; Event Rate: 30% R: 0.90

## *Simulazione – risultati (2)*

---

### *Confronto per diversi valori di R*

	Reliability: 0.90			Reliability: 0.85		
	NAIVE	RC	SIMEX	NAIVE	RC	SIMEX
Bias	-0.075	-0.020	-0.013	-0.123	-0.033	-0.044
SD	0.025	0.030	0.034	0.023	0.033	0.036
SE	0.022	0.027	0.029	0.020	0.028	0.026
Coverage	10.8	84.4	86.2	0.4	73.6	56.6

■ *RC migliore di SIMEX* quando R peggiora

■ *SE < SD e scarso coverage:* necessità di derivare formulazione opportuna per SE.

Sample Size: n=1000; Event Rate: 15% Beta: 0.35

## *Simulazione – risultati (3)*

---

### *Altri risultati:*

- *Risultati simili* al variare della *numerosità campionaria*, *mentre tendono a peggiorare all'aumento dell'event rate*
- *I due metodi danno risultati molto simili* per quanto riguarda la stima dell'effetto della covariata error-free
- *Tempi di implementazione e calcolo:* RC semplice, SIMEX molto più complesso (dipende da numero di simulazioni, numero di punti e da modello di regressione scelto per EXtrapolation)

## *Conclusioni*

---

■ La metodologia descritta può essere applicata nelle situazioni in cui si desidera *stimare l'effetto prognostico del cambiamento nel tempo* di un fattore di esposizione.

■ In particolare, *negli studi osservazionali*, il numero di visite e il tempo tra visite successive sono *raramente costanti tra i soggetti; questo costituisce ambito applicativo preferenziale*

## *Conclusioni*

---

- Sia *RC* e *SIMEX* sono applicabili alla situazione in cui *l'IIV non è costante tra i soggetti*
- Entrambi i metodi possono essere *generalizzati per modelli di sopravvivenza più complessi* (ex.: aggiunta della misura basale)
- I principali fattori che devono guidare nella *scelta del metodo* di ME correction sono: *l'ampiezza dell'effetto in studio, la quantità di ME e il tempo di implementazione sono i principali*

## *Ringraziamenti*

---

### *Coautori*

■ ***Lloyd E Chambless, Michael Hudgens***

*Department of Biostatistics, University of North Carolina at  
Chapel Hill, NC, US*

■ ***Marco M Ferrario***

*Dipartimento di Scienze Cliniche e Biologiche, Università  
degli Studi di Varese, Varese, Italy*

---

## *ARIC Study (1)*

---

- *Atherosclerosis Risk in Communities (ARIC)*: studio osservazionale prospettico di 4 U.S. Communities allo scopo di studiare l'eziologia e la storia naturale dell'aterosclerosi
- Baseline: 1987-89. Partecipanti ri-esaminati ogni 3 anni: 1990-92, 1993-95, 1996-98. *Max 4 visite per soggetto*
- *Campione considerato: N=7,462 donne*, libere da CVD al basale, età media  $53.8 \pm 5.7$ , 73% Caucasiche
- *Follow-up mediano: 6 anni* successivi alla visita 4, *197 primi eventi CHD* (tasso di evento: 4.8 per 1,000 anni-persona)

## ARIC Study (2)

---

	LDL- Cholesterol	Systolic Blood Pressure
N subjects included	6,630	6,648
Intra-individual variability original variable	153.6*	94.5**
Intra-individual variability slope§	6.14 (8.10)	3.67 (4.85)
Reliability coefficient original variable***	0.90	0.71
Reliability coefficient slope§,***	0.77	0.38
Subjects with complete data (%)	77.7	80.4
Number of events	193 (2.9%)	197 (3.0%)
Error-free covariate	Age, Race	Age, Race

\* From internal replication study

\*\* Estimated as  $R \times \text{Total Variance of SBP}$ ,  $R = 0.75$

\*\*\* Relative to the case with no other variable in the model.

§: Mean and SD

## ARIC Study (3)

		<b>LDL-Cholesterol</b>					
		NAIVE		RC		SIMEX	
Change	Estimate	0.018		0.018		0.019	
over time	95% CI	-0.007	0.044	-0.014	0.050	-0.012	0.050
Error-	Estimate	0.043		0.042		0.042	
Free Z	95% CI	0.018	0.068	0.016	0.067	0.016	0.067
		<b>Systolic Blood Pressure</b>					
		NAIVE		RC		SIMEX	
Change	Estimate	0.070		0.095		0.081	
over time	95% CI	0.014	0.126	-0.043	0.232	0.004	0.157
Error-	Estimate	0.039		0.035		0.037	
Free Z	95% CI	0.014	0.064	0.009	0.061	0.011	0.062

■ **Risultati confermano simulazioni:** RC migliore di SIMEX quando R è più bassa, come nel caso della SBP