The SIMEX method
bcam Bilbao

Helmut Küchenhoff
Statistical Consulting Unit
Ludwig-Maximilians-Universität München

Bilbao
30-05-2019

2. GENERAL SIMEX IDEA

Linear regression with additive measurement error

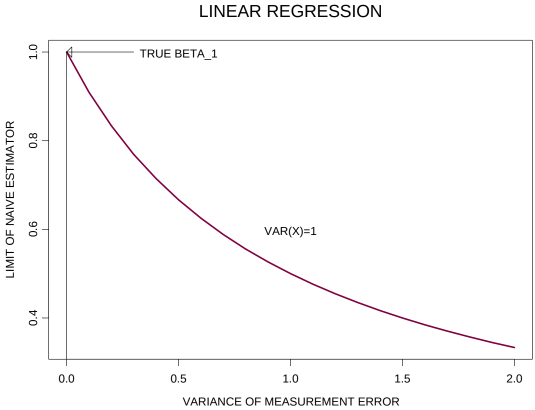
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (i = 1, \ldots, n)$$
$$\mathrm{Var}(X_i) = \sigma_X^2 \ \& \ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

$$X_i^* = X_i + \sigma U_i \ \text{ with } (U_i, X_i, \varepsilon_i) \text{ independent}$$
$$U_i \sim N(0,1)$$

- Ignoring measurement error (U_i) $\Rightarrow$ **naïve estimation** in $Y_i = \beta_0^* + \beta_1^* X_i^* + \varepsilon_i^*$

$$\beta_1^* = \operatorname{plim} \hat{\beta}_{\text{naive}} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma^2} \beta_1$$

$\Rightarrow$ Attenuation increases with measurement error variance


LINEAR REGRESSION
TRUE BETA_1
VAR(X)=1
LIMIT OF NAIVE ESTIMATOR
VARIANCE OF MEASUREMENT ERROR
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0

SIMEX idea (Cook & Stefanski, 1994)

- Assume
  - σ is **known**
  - Observe $\left(Y_i, X_i^*, Z_i\right)_{i=1}^n$ instead of $\left(Y_i, X_i, Z_i\right)_{i=1}^n$

- **SIM**ulation step: generate more measurement error + calculate naïve estimators

  1. **Simulate pseudo-data** $X_{b,i}^*(\lambda) = X_i^* + \sqrt{\lambda}\,\sigma U_{b,i}$ for a fixed grid $\lambda_0 (\equiv 0), \lambda_1, \lambda_2, \ldots, \lambda_m$

     $\Rightarrow \operatorname{Var}\left(X_{b,i}^*(\lambda)\right) = \sigma_X^2 + (1+\lambda)\sigma^2$

  2. **Do this B times** (b=1, …, B)

  3. **Calculate mean**: $\hat{\beta}(\lambda_k) = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}_{NAIVE}\left[\left(Y_i, X_{b,i}^*(\lambda_k), Z_i\right)_{i=1}^n\right]$

- **EX**trapolation step: extrapolate back to λ = -1 to estimate β

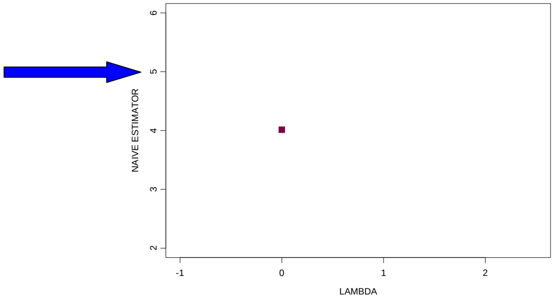
    1. **Fit parametrically** relation $\left(\lambda_k, \hat{\beta}(\lambda_k)\right)$ $(k = 0, \ldots, m)$

    2. **Find** $\hat{\beta}_{SIMEX} \equiv \hat{\beta}(-1)$ for all regression coefficients

Example

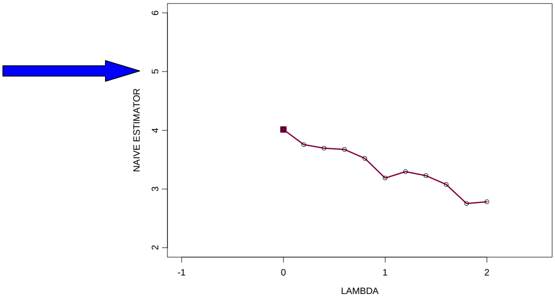
$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \ldots, 200) \qquad X_i \sim N(0, 2^2)\ \&\ \varepsilon_i \sim N(0,1)$$

$$Y_i = \beta_0^* + \beta_1^*(X_i + U_i) + \varepsilon_i \qquad U_i \sim N(0,1)$$



Example $Y_i = 1 + 5X_i + \varepsilon_i \ \ (i = 1, \ldots, 200)$ $X_i \sim N(0, 2^2)$ & $\varepsilon_i \sim N(0,1)$

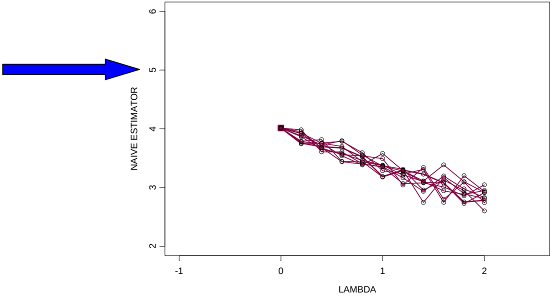
$Y_i = \beta_0^* + \beta_1^* (X_i + U_i) + \varepsilon_i$ $U_i \sim N(0,1)$



Example

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \ldots, 200) \qquad X_i \sim N(0, 2^2)\ \&\ \varepsilon_i \sim N(0,1)$$

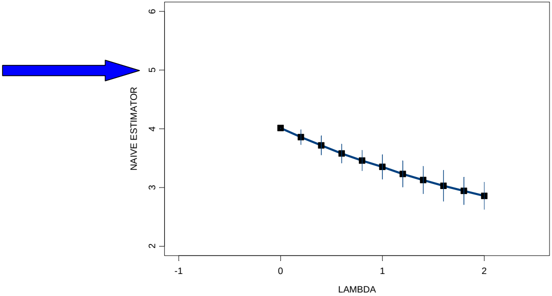
$$Y_i = \beta_0^* + \beta_1^* (X_i + U_i) + \varepsilon_i \qquad U_i \sim N(0,1)$$



Example

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \ldots, 200) \qquad X_i \sim N(0, 2^2) \; \& \; \varepsilon_i \sim N(0,1)$$

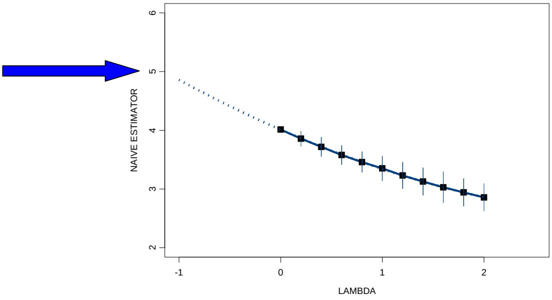
$$Y_i = \beta_0^* + \beta_1^* (X_i + U_i) + \varepsilon_i \qquad U_i \sim N(0,1)$$



Example

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \ldots, 200) \qquad X_i \sim N(0, 2^2) \,\&\, \varepsilon_i \sim N(0,1)$$

$$Y_i = \beta_0^* + \beta_1^* (X_i + U_i) + \varepsilon_i \qquad U_i \sim N(0,1)$$



Example $Y_i = 1 + 5X_i + \varepsilon_i \ (i = 1, \ldots, 200)$ $X_i \sim N(0, 2^2)$ & $\varepsilon_i \sim N(0,1)$

$Y_i = \beta_0^* + \beta_1^* (X_i + U_i) + \varepsilon_i$ $U_i \sim N(0,1)$

Average of extrapolated estimate = $\hat{\beta}_{1,SIMEX} = 4.86$

Extrapolation functions

$$
\begin{aligned}
\text{Linear} &: \quad g(\lambda) = \gamma_0 + \gamma_1 \lambda \\
\text{Quadratic} &: \quad g(\lambda) = \gamma_0 + \gamma_1 \lambda + \gamma_2 * \lambda^2 \\
\text{Nonlinear} &: \quad g(\lambda) = \gamma_1 + \frac{\gamma_2}{\gamma_3 * \lambda}
\end{aligned}
$$

- Nonlinear is motivated by linear regression
- Quadratic works fine in many examples
- Motivation by Taylor Series expansions

Variance estimation

- Delta method (Carroll et al.(1996))
- For known error variance the variance can be also be estimated by extrapolation, Stefanski and Cook (1995)
- Bootstrap (computer intensive)

Case study : Occupational Dust and chronic bronchitis

HK/Carroll (1997) and Goessl /HK(2001)
Research question: Relationship between occupational dust and chronic bronchitis

Data form N=1246 workers:

X:	log(1+average occupational dust exposure)
Y:	Chronic bronchitis (CBR)
X^*:	Measurements and expert ratings
Z_1:	Smoking
Z_2:	Duration of exposure

Results for the TLV

Method	TLV-τ_0	Nom s. e.	boot s.e.
Naive	1.27	.41	.24
Pseudo-MLE	1.76	.17	.21
Regression Calibration	1.75	.12	.19
simex: linear	1.37	.23	.23
simex: quadratic	1.40	.23	.34
simex: nonlinear	1.40	.23	.86

Misclassification SIMEX

General Regression model with misclassification matrix Π

$$\begin{aligned}\beta^*(\Pi) &:= p\lim \hat{\beta}_{naive} \\ \beta^*(I_{k\times k}) &= \beta\end{aligned}$$

Problem: $\beta^*(\Pi)$ is a function of a matrix.
We define:

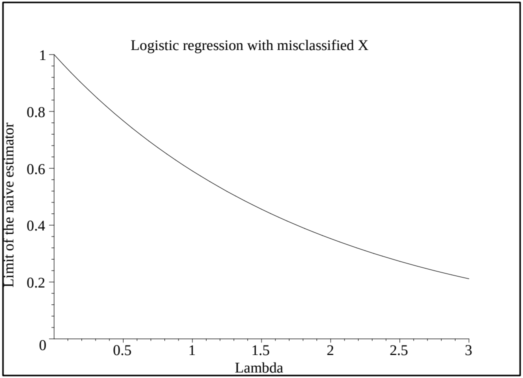
$$\lambda \to \beta^*(\Pi^\lambda)$$

Π^λ is defined by $\Pi^0 = I_{kxk}$, $\Pi^{n+1} = \Pi^n * \Pi$ for $\lambda = 0, 1, 2...$

$$\begin{aligned}\Pi^\lambda &:= E\Lambda^\lambda E^{-1} \\ E &:= \text{Matrix of eigenvectors} \\ \Lambda &:= \text{Diagonal matrix of eigenvalues}\end{aligned}$$

Parameter Estimation in Relationship to the amount of misclassification

Logistic regression with misclassified X ($\pi_{11} = \pi_{00} = 0.8$)



Properties of the function $\beta^*(\Pi^\lambda)$

- $\beta^*(\Pi^0) = \beta$
- differentiable

If X^* is misclassified in relation to X by MC-matrix Π
$X^*(\lambda)$, is misclassified in relation to X^* by MC-matrix Π^λ,
$\Rightarrow$
$X^*(\lambda)$ is misclassified in relation to X by MC-matrix $\Pi^{\lambda+1}$

The MC-SIMEX Procedure

Data $(Y_i, X_i^*, Z_i)_{i=1}^n$,
X^* is observed instead of X with MC-matrix Π
Naive estimator: $\hat{\beta}_{naive}[(Y_i, X_i^*, Z_i)_{i=1}^n]$.
Simulation
For a fixed grid $\lambda_1 \ldots \lambda_m$ B new pseudo data are generated by

$$X_{b,i}^*(\lambda_k) \quad := \quad MC[\Pi_k^\lambda](X_i^*), \;\; i = 1, \ldots, n; \; b = 1, \ldots B;$$

There $MC[M](X_i^*)$ is simulated from X_i^* using the misclassification matrix M.

$$\hat{\beta}(\lambda_k) := B^{-1} \sum_{b=1}^{B} \hat{\beta}_{na} \left[(Y_i, X_{b,i}^*(\lambda_k), Z_i)_{i=1}^n \right], k = 1, \ldots m.$$

Extrapolation function

Parametric model:

$$\beta(\Pi^{\lambda}) \quad = \quad \mathcal{G}(\lambda, \Gamma) = \gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2$$

Fit by least squares from data $[\lambda_k, \hat{\beta}(\lambda_k)]_{k=0}^{m}$.

$$\hat{\beta}_{SIMEX} := \mathcal{G}(-1, \hat{\Gamma})$$

Calculate true function $\beta(\Pi)$ in several examples and simulation studies

- Funktion monotonic
- Quadratische Extrapolation suitable
- Loglinear Extrapolant

Remarks

- Existence of Π^{λ} for $\lambda < 1$ has to be checked
- If β vector then use MC-SIMEX for every component
- The procedure also works for misclassified Y or more general cases
- $\hat{\beta}_{SIMEX}$ is consistent, if the extrapolating function is correctly specified.
- In general MC-SIMEX is approximately consistent, if $\mathcal{G}(\lambda, \Gamma)$ is a good approximation of $\beta^{*}(\Pi^{\lambda})$.

Delta Method Variance Estimation

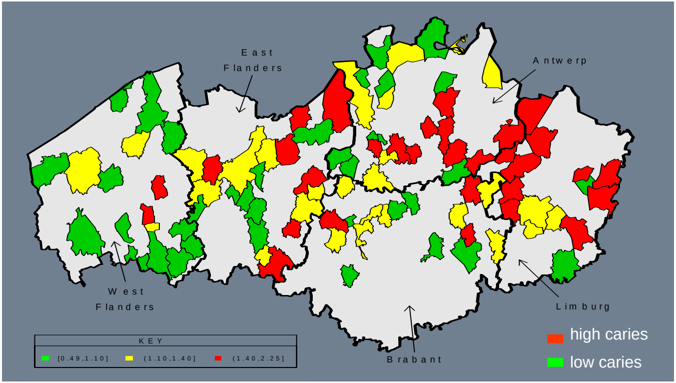
- All Estimators are solution of (biased) estimating equations
- Asymptotic expansions on averages of different estimating equations
- Extrapolation is a differentiable transformation
- Estimation of misclassification matrix can be included

7. APPLICATION TO THE SIGNAL TANDMOBIEL STUDY®

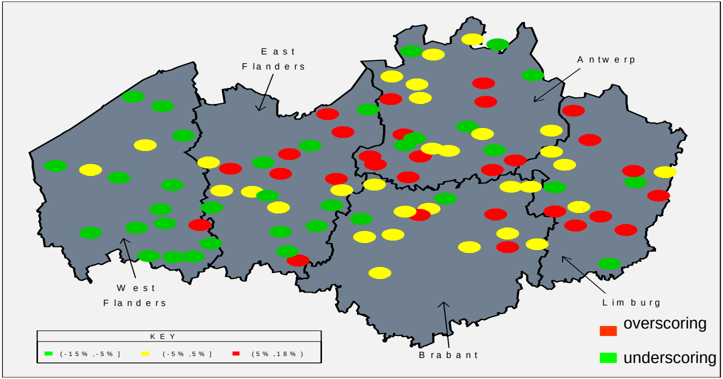
- Oral health study involving 4468 children in Flanders (Belgium)
- Children were examined annually by one of 16 dental examiners
- Binary response Y=1 if tooth is decayed, filled or extracted due to caries
- GEE analysis for caries (combined response & individual teeth) on 4 first molars as a function of covariates
- **Questions**:
    - **East-West gradient in caries experience on the first 4 molars?**
    - **Does the trend remain the same in time?**

But: dental examiners showed high & different misclassification ⇔ benchmark scorer

STS: East-West trend in caries experience (1st year's cross-sectional results)



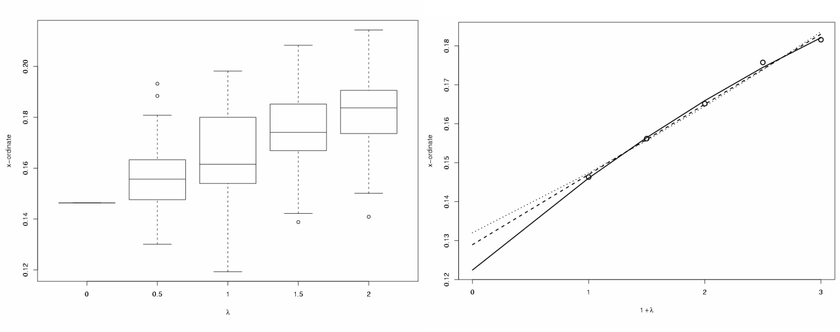
STS: Dental examiners are active in restricted geographical areas



⇒ East-West gradient?

Results SIMEX approach (individual teeth)

- X-coordinate



Results SIMEX approach (individual teeth)

- East-West gradient confirmed
- East-West gradient increases over the years
- ….

Software

- R-Package available (W. Lederer)
- Flexible statement for the main model
- Misclassification and additive measurement error
- Graphic display for the results

Lederer, HK R-news (2006)

Summary

- Very general computer intensive method
- Illustration of the effect of misclassification
- MC in X,Y or both, differential MC etc. can be handled
- Misclassification known or can be estimated by validation data