

Correlation and heterogeneity in commonly used cost-transform distributions: a complex interaction which bad effects should not be underestimated

Baldi Ileana¹, **Pagano Eva**¹, Desideri Alessandro², Berchiolla Paola³, Ferrando Alberto¹, Gregori Dario⁴

1 Unit of Cancer Epidemiology, CeRMS and CPO Piemonte, University of Torino

2 Cardiovascular Research Foundation, San Giacomo Hospital, 31033 Castelfranco Veneto, Italy

3 Department of Public Health and Microbiology, University of Torino

4 Department of Environmental Medicine and Public Health, University of Padova

Background (1): asymmetry

- Asymmetry is a statistical property characterizing many distributions of health economics variables such as length of stay, number of inpatients stays and expenditures
- Transformation of the response has become a popular remedy, but
 - drawbacks in terms of difficult interpretability of coefficients on a different scale
 - poor quality of re-transformed parameter estimates, specially in presence of heteroscedasticity of the error ε in x 's
- Common and subgroups-specific smearing retransformations have been proposed respectively by Duan and Manning.

Standard approaches for independent data (1)

1) **Linear model of log-transformed data**, fitted either via Ordinary Least Squares (OLS) or Maximum Likelihood (ML), assumes the following form for the costs

$$\log(y_i) = \sum \beta_j x_{ij} + \varepsilon_i$$

- Estimates for $E(\log(y)|x)$, not $\log(E(y)|x)$. Usually want arithmetic mean, not geometric mean.
- May be difficult to obtain unbiased estimates of mean response $E(y|x)$ if error ε heteroscedastic in x 's

Standard approaches for independent data (2)

2) To reduce skewness in the residuals, the **Box-Cox transformation** of c_i can be used

$$\frac{c_i^\lambda - 1}{\lambda} = \sum \beta_j x_j + \varepsilon_i \quad \text{if } \lambda \neq 0$$
$$\log(c_i) = \sum \beta_j x_j + \varepsilon_i \quad \text{if } \lambda = 0$$

- Normality of ε_i distribution is still assumed
- When the back-transformation is performed, the eventual bias is function of the variance function

$$\frac{\partial \sigma_\varepsilon^2(x_i)}{\partial x_i}$$

thus, heteroscedasticity, if present, raises additional efficiency and inference problems on the transformed scale

Standard approaches for independent data (3)

3) Generalize Linear Models (GLM):

To avoid bias in transforming the costs directly, since $g^{-1}(E(c_i)) \neq E(g^{-1}(c_i))$

the idea is to model the transformation of the expectation $g(E(c_i)) = \sum \beta_j x_j$

- The distribution for the response is usually taken to be Gamma() and the link function as the log()
- Basu proved by simulations that Gamma regression models with a log link seem to be more robust to alternative data generating mechanisms than either OLS on $\ln(y)$ or Cox proportional hazards regression.

Background (2): correlation

- Data in health economic studies are often clustered (eg. Patients within hospital, general practitioner, etc.) given that patients may share
 - unmeasured socio-demographic or clinical characteristics
 - cluster specific issues
- A within clusters correlation increases the model variance
- Random intercept or even random slope models are usually adopted

Standard approaches for clustered data (2)

1) To fit to the log-transformed outcome a **Linear Mixed Effect (LME)** model accommodating for correlation among observations within the same cluster via a random effect

$$\text{Log-LME: } \log(y_{ij}) = \beta x_{ij} + u_j + \varepsilon_{ij} \quad j=\text{cluster}$$

2) To extend the **Box-Cox transformation** to the LME as described by Gurka

Standard approaches for clustered data (2)

3) **Generalized Linear Mixed Models (GLMM)** with Gamma distribution and log link:

$$y_{ij} \sim \text{Gamma}(\theta_{ij}, \phi_i)$$

$$g(\theta_{ij}) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + u_i \quad u_i \sim \text{normal}(0, \tau^2)$$

- Grieve showed that multilevel Gamma models properly deal with skewed clustered data.

Study aim

- to understand how common models, used for accommodating asymmetry, perform in the case of clustered data, under heteroscedasticity in the variance function...
- ...since interaction of variability due to the distribution of the transformed data and that induced by correlation in data is expected to worsen efficiency and to bias the estimates

Methods

- Assessment of the performance of the Gamma GLMM with log link and log-LME model under different data generating mechanisms
- Comparison of the two approaches on two case studies:
 - the COSTAMI trials
 - the EPUAP study

Simulation Study (1)

The simulation considered:

$j = 6, 12, 24$ clusters

$m = 10, 15, 30$ measures per cluster

A Gamma model with random mean equal to $\exp(\beta_0 + \beta_1 X + u_j)$
and scale (c) equal to 0.8 or $1+X$ or $1+X^2$

- $X \sim \text{Uniform}(1, 10)$
- fixed effects $\beta_0 = 0.5, \beta_1 = 0.8$
- $u \sim \text{Normal}$ distributed with mean and variance (τ^2)
set at 0.6 and 1.2

Simulation Study (2)

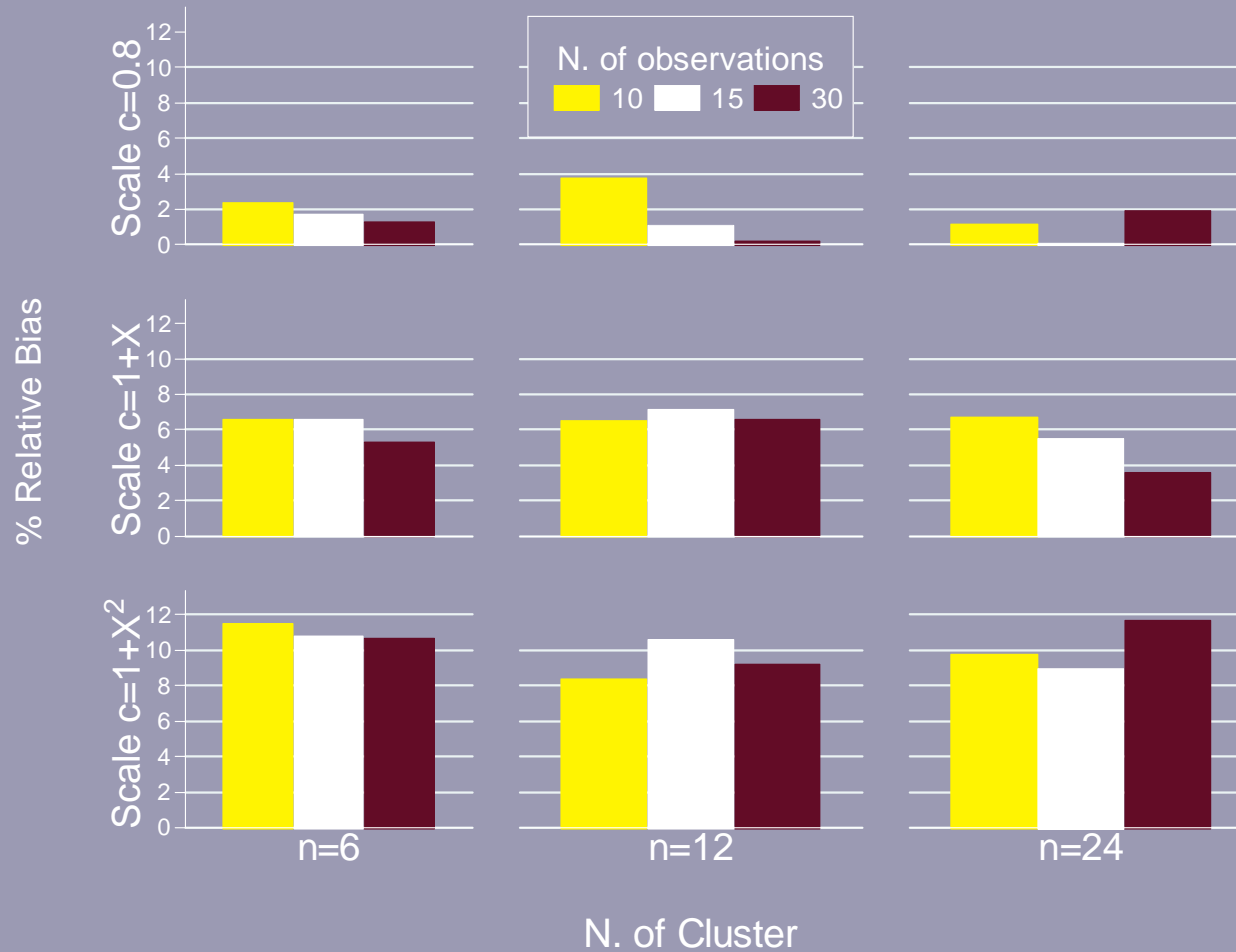
- For each setting of (j, m, τ, c) , 1000 sets of data were simulated. Each dataset was analyzed by both LME and Gamma models. LME model was fit on the log-transformed outcome
- Relative bias, coverage probability and median square error (MSE) were calculated to summarize consistency and precision in each simulation

Results of the simulation study

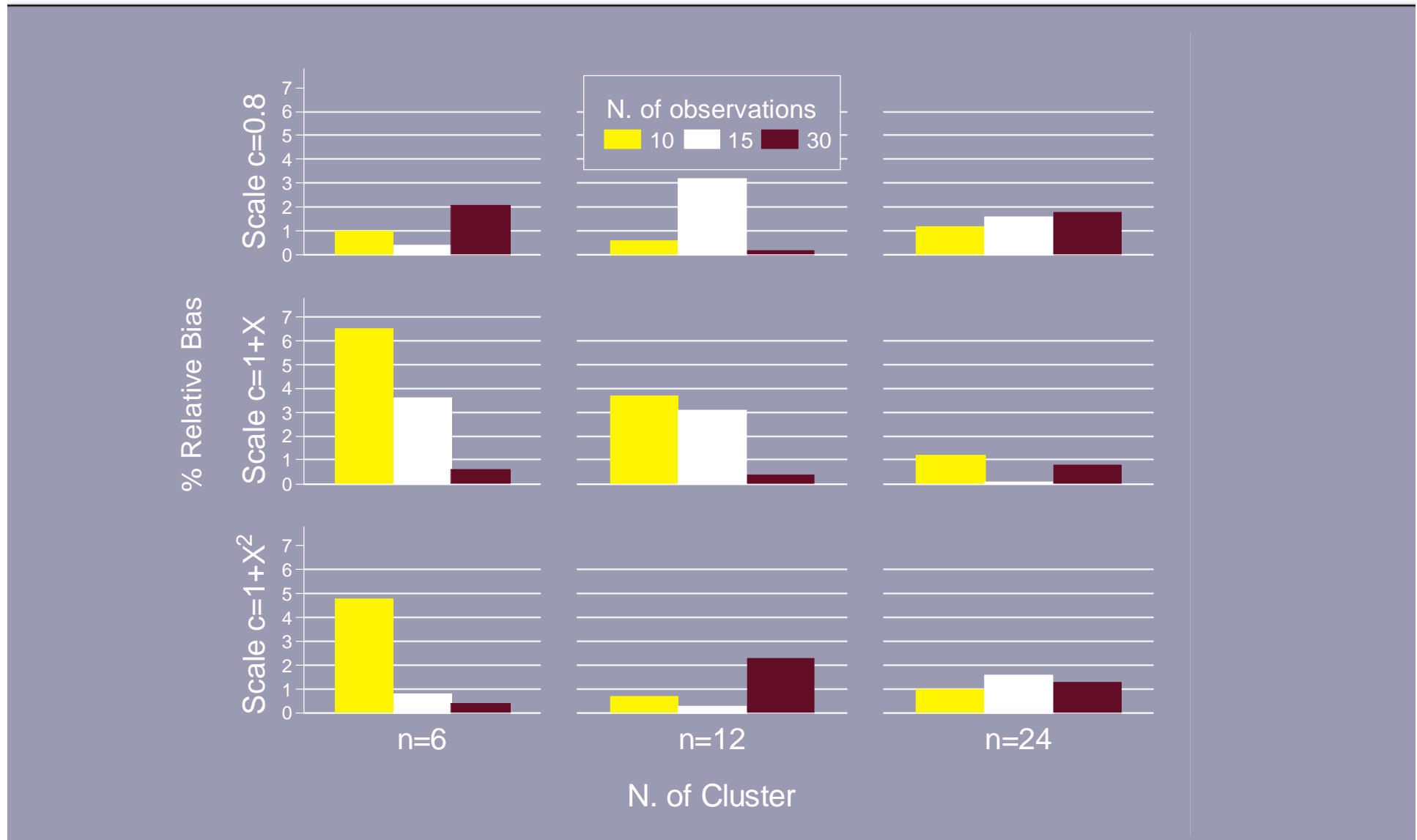
- When the error variance is constant:
 - log-LME and Gamma GLMM produce similar estimates with a negligible relative bias
 - confidence intervals show a coverage probability of about 90% under most conditions

- In presence of heteroschedasticity, the log-LME provides a substantially biased estimates for as much as 10% of the true value, increasing as the error variance increases

Simulation study results: Log-LME relative bias



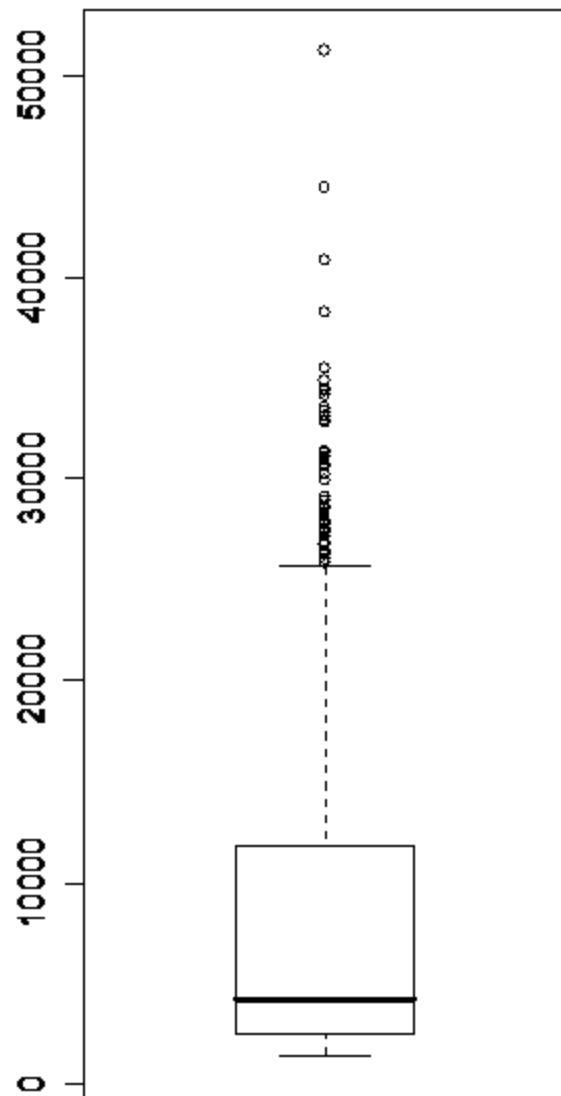
Simulation study results: Gamma GLMM relative bias



The COSTAMI trial

- Compared an early discharge strategy based on stress echocardiography to standard care based on clinical evaluation and post-discharge exercise electrocardiography (ECG) for post-infarction uncomplicated patients
- Total medical costs per patients were measured as follow-up hospital and outpatients costs at 1 year (euros).
- A total of 720 patients from 17 participating centers were recruited during the period January 1998 and August 2000

The COSTAMI trial



Homoscedasticity was suggested by a formal evaluation of the residuals on the log-scale via the Breusch-Pagan test
(Chi-square= 0.08;
p-value=0.78)

The COSTAMI trial: results

	GLMM - Gamma			Log - LME		
	Effect	SE	P-value	Effect	SE	P-value
(Intercept)	9.03	0.11	<0.001	8.65	0.12	<0.001
Strategy 1 vs 4	0.09	0.13	0.52	0.14	0.13	0.29
Strategy 2 vs 4	0.17	0.13	0.19	0.21	0.13	0.10
Strategy 3 vs 4	-0.17	0.08	0.03	-0.16	0.08	0.06
Gender (Male vs Female)	-0.08	0.09	0.38	-0.10	0.09	0.27
Age >65 vs <=65	-0.09	0.08	0.22	-0.11	0.08	0.15
Hypertension (Pres vs Abs)	0.21	0.07	<0.001	0.17	0.07	0.01
Previous AMI (Pres vs Abs)	-0.07	0.13	0.57	-0.03	0.13	0.81
Diabetes (Pres vs Abs)	0.16	0.09	0.08	0.16	0.09	0.07
Random-effect variance τ^2		0.02			0.02	

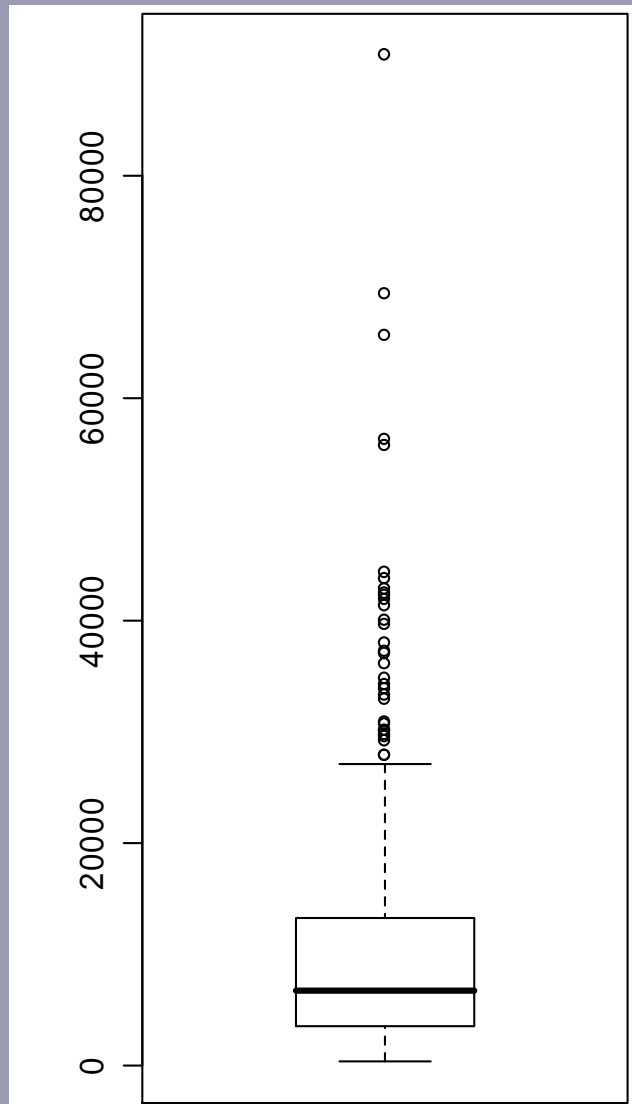
The COSTAMI trial: results

	GLMM - Gamma			Log - LME		
	Effect	SE	P-value	Effect	SE	P-value
(Intercept)	9.03	0.11	<0.001	8.65	0.12	<0.001
Strategy 1 vs 4	0.09	0.13	0.52	0.14	0.13	0.29
Strategy 2 vs 4	0.17	0.13	0.19	0.21	0.13	0.10
Strategy 3 vs 4	-0.17	0.08	0.03	-0.16	0.08	0.06
Gender (Male vs Female)	-0.08	0.09	0.38	-0.10	0.09	0.27
Age >65 vs <=65	-0.09	0.08	0.22	-0.11	0.08	0.15
Hypertension (Pres vs Abs)	0.21	0.07	<0.001	0.17	0.07	0.01
Previous AMI (Pres vs Abs)	-0.07	0.13	0.57	-0.03	0.13	0.81
Diabetes (Pres vs Abs)	0.16	0.09	0.08	0.16	0.09	0.07
Random-effect variance τ^2		0.02			0.02	

The EPUAP study

- From a prevalence survey performed by the European Pressure Ulcer Advisory Panel, within several Italian hospitals.
- Hospital costs of patients (N=389) with pressure ulcers from San Giovanni Battista Hospital were analyzed
- The following predictors were considered: gender, age, length of stay at index day and PU clinical staging in four categories (the higher the worse) according to EPUAP staging system

The EPUAP study



Heteroscedasticity was suggested by formal evaluation of the residuals on the log-scale via the Breusch-Pagan test (Chi-square= 45.34; p-value<0.0001)

The EPUAP study: results

	GLMM - Gamma			Log - LME		
	Effect	SE	P-value	Effect	SE	P-value
(Intercept)	8.60	0.09	<0.001	8.40	0.10	<0.001
Stage II vs I	0.15	0.09	0.12	0.18	0.10	0.09
Stage III	0.19	0.11	0.10	0.26	0.13	0.05
Stage IV	0.26	0.10	0.008	0.36	0.11	0.001
Gender (Male vs Female)	0.02	0.07	0.73	0.03	0.08	0.69
Length of stay at index day	0.02	0.01	<0.001	0.02	0.01	<0.001
Age <40 (vs <60-69)	0.04	0.18	0.81	-0.16	0.20	0.44
Age 40-59	-0.01	0.10	0.98	-0.01	0.12	0.94
Age 70-79	-0.10	0.09	0.26	-0.08	0.10	0.42
Age >90	-0.03	0.18	0.87	-0.17	0.20	0.38
Random-effect variance τ^2		0.04			0.01	

The EPUAP study: results

	GLMM - Gamma			Log - LME		
	Effect	SE	P-value	Effect	SE	P-value
(Intercept)	8.60	0.09	<0.001	8.40	0.10	<0.001
Stage II vs I	0.15	0.09	0.12	0.18	0.10	0.09
Stage III	0.19	0.11	0.10	0.26	0.13	0.05
Stage IV	0.26	0.10	0.008	0.36	0.11	0.001
Gender (Male vs Female)	0.02	0.07	0.73	0.03	0.08	0.69
Length of stay at index day	0.02	0.01	<0.001	0.02	0.01	<0.001
Age <40 (vs <60-69)	0.04	0.18	0.81	-0.16	0.20	0.44
Age 40-59	-0.01	0.10	0.98	-0.01	0.12	0.94
Age 70-79	-0.10	0.09	0.26	-0.08	0.10	0.42
Age >90	-0.03	0.18	0.87	-0.17	0.20	0.38
Random-effect variance τ^2		0.04			0.01	

Final remarks (1)

The co-presence of clustering and asymmetry has an interacting adverse effect on the quality of model estimates.

When the error variance is constant, log-LME and Gamma GLMM produce similar estimates with a negligible relative bias.

The presence of heteroschedasticity translates into different model performance as shown with the difference in variance estimates (EPUAP study) and consistently with the simulations.

Both in case of asymmetry only (COSTAMI) and heteroscedasticity (EPUAP), the consequences of the choice of the modeling approach also impact the interpretation of the results.

Final remarks (2)

The combination of heterogeneity and asymmetry in clustered data stresses the need to develop appropriate methods for the identification of such situations in practical settings, particularly with reference to specific covariates

The issue of which remedial could be used in approaching the issue of heteroscedasticity in presence of asymmetry in clustered data was beyond our study purpose ...

... nevertheless, common techniques, like adding random slopes for the most sensitive set of covariates seems attractive