On verification of a superpopulation model

Małgorzata K. Krzciuk

Department of Statistics, Econometrics and Mathematics
University of Economics in Katowice

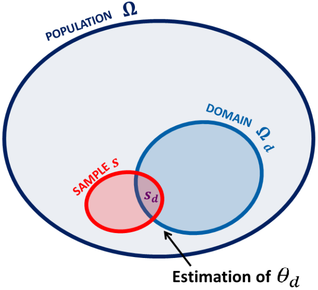
2nd Congress of Polish Statistics
July 10-12 2018

The aims of studies

- An analysis of the possibility of using **parametric bootstrap test** for testing **dependence between random effects**.
- A **comparison** in a simulation study the **properties** of this significance test with one of the **classic test**.
- Analysis taking into account the problem of:
  - **model misspecyfications**,
  - **non-normality** of **random effects**.

Introduction - Small Area Estimation

Small area - domain for which we cannot obtain direct estimates with adequate precision (Rao, 2003, p. 2).



One of the main approache in small area estimation - **model-based approach**. A **superpopultion model** as a **source of randomness** in this approach.

Linear mixed model

The general linear mixed model (cf. Jiang, 2007, pp. 1-2):

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon \tag{1}$$

where:

- $\mathbf{Y}$ – the random vector of values of the dependent variable;
- $\mathbf{X}$, $\mathbf{Z}$ – known matrices of auxiliary variables;
- β – the vector of unknown parameters.

Random effects $\mathbf{v}$ and stochastic disturbance ϵ are independently distributed with variance-covariance matrices denoted by $\mathbf{G}(\delta)$ and $\mathbf{R}(\delta)$, where δ is a vector of variance components.

Linear mixed model with correlation of random effects (1)

We can consider correlation of random effects as two cases:

- **autocorrelation** of random effect
  - simultaneously autoregressive process of random effect in estimation the annual per-capita mean income in Tuscany (Pratesi and Salvati, 2008);
  - normal mixed autoregressive moving average process of random effects (Tiao and Ali, 1971);
  - serial correlation of random effect in estimation of production for Japanes chemical industry (Skoglund and Karlsson, 2001);
- **correlation between** random effects
  - to estimation plasma concentration some drug by nonlinear mixed effects model (Dumont C. et. al., 2014);
  - to analyze the health care costs at the end of life (Menec et al. 2004).

Linear mixed model with correlation of random effects (2)

$$Y_{id} = (\beta_1 + v_1)x_{id} + \beta_0 + v_2 + e_{id} \tag{2}$$

where:

- Y_{id} – values of the dependent variable;
- x_{id} – values of the auxiliary variables;
- β_0, β_1 – unknown parameters

  and
- $\mathbf{G}(\boldsymbol{\delta}) = D^2 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} \sigma_{v_1}^2 & \rho\sigma_{v_1}\sigma_{v_2} \\ \rho\sigma_{v_1}\sigma_{v_2} & \sigma_{v_2}^2 \end{bmatrix}$
- $\mathbf{R}(\boldsymbol{\delta}) = \sigma_e^2 \mathbf{I}_n$

Classic test - Likelihood Ratio Test

We considered the following **hypotheses**:

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

The test statistic is given by:

$$2log\left(\frac{L_2}{L_1}\right) = 2\left[log(L_2) - log(L_1)\right] \tag{3}$$

where:

- L_1 - the likelihood for more general model;
- L_2 - the likelihood for the restricted model.

Under the hypothesis H_0 this test statistic has asymptotic χ^2 distribution with $k_2 - k_1$ degrees of freedom (where k_1 and k_2 are the number of model parameters) (Pinheiro and Bates, 2000, p. 83).

Monte Carlo Test

Idea of the **Monte Carlo test** was first described by Bartlett (1963) and Barnard (1963).
This test based on the comparison of original data with random samples, which are generated under the null hypothesis (Hope, 1968, p. 582).
It should be noted similarity of developed in 1980's **parametric bootstrap approximation** to the procedure of **Monte Carlo test** (Zhu, 2005, p. 2).

Parametric bootstrap test (1)

The prametric bootstrap method finds application in many areas i.a.:

- estimation of the **MSE** (Gonzales-Manteiga et. al., 2008; Butar and Lahiri, 2003);
- approximation of the distribution **LRT statistic** (Shaw and Geyer, 1997);
- approximation of the distribution **EBLUP** (Chatterje et. al., 2008);
- Bayesian inference (Efron, 2012).

Parametric bootstrap test (2) - testing procedure

1. Estimation ρ for the original data set (denoted by ρ_0);
2. B-times:
   a) generating data under hypothesis H_0;
   b) estimation ρ_{*b} according to the model with correlated random effects;
3. Calculate p-value as $p = \frac{1 + b: |\rho_{*b}| > |\rho_0|}{1+B}$.

Simulation study

- **assumptions**
- **results of part I of the simulation study**
- **results of part II of the simulation study**
- **conclusions**

Simulation study (1)

Data set:

- data Särndal et al (1992);
- population elements – counties in Sweden ($N = 284$);
- y - revenues from municipal taxation (in millions of kronor);
- x - population (in thousands);
- domains according to the cluster indicator ($D = 50$).

Simulation study (2) - choice of the model

$$Y_{id} = \beta_1 x_{id} + \beta_0 + e_{id} \quad (4)$$

$$Y_{id} = \beta_1 x_{id} + \beta_0 + v_d + e_{id} \quad (5)$$

$$Y_{id} = (\beta_1 + v_d) x_{id} + \beta_0 + e_{id} \quad (6)$$

$$Y_{id} = (\beta_1 + v_{1d}) x_{id} + \beta_0 + v_{2d} + e_{id} \quad (7)$$

$$Y_{id} = (\beta_1 + v_{1d*}) x_{id} + \beta_0 + v_{2d*} + e_{id} \quad (8)$$

Simulation study (2) - choice of the model

$$Y_{id} = \beta_1 x_{id} + \beta_0 + e_{id} \tag{4}$$

$$Y_{id} = \beta_1 x_{id} + \beta_0 + v_d + e_{id} \tag{5}$$

$$Y_{id} = (\beta_1 + v_d) x_{id} + \beta_0 + e_{id} \tag{6}$$

$$Y_{id} = (\beta_1 + v_{1d}) x_{id} + \beta_0 + v_{2d} + e_{id} \tag{7}$$

$$Y_{id} = (\beta_1 + v_{1d*}) x_{id} + \beta_0 + v_{2d*} + e_{id} \tag{8}$$

Simulation study (2) - choice of the model

$$Y_{id} = \beta_1 x_{id} + \beta_0 + e_{id} \tag{4}$$

$$Y_{id} = \beta_1 x_{id} + \beta_0 + v_d + e_{id} \tag{5}$$

$$Y_{id} = (\beta_1 + v_d) x_{id} + \beta_0 + e_{id} \tag{6}$$

$$Y_{id} = (\beta_1 + v_{1d}) x_{id} + \beta_0 + v_{2d} + e_{id} \tag{7}$$

$$Y_{id} = (\beta_1 + v_{1d*}) x_{id} + \beta_0 + v_{2d*} + e_{id} \tag{8}$$

Simulation study (2) - choice of the model

$$Y_{id} = \beta_1 x_{id} + \beta_0 + e_{id} \tag{4}$$

$$Y_{id} = \beta_1 x_{id} + \beta_0 + v_d + e_{id} \tag{5}$$

$$Y_{id} = (\beta_1 + v_d) x_{id} + \beta_0 + e_{id} \tag{6}$$

$$Y_{id} = (\beta_1 + v_{1d}) x_{id} + \beta_0 + v_{2d} + e_{id} \tag{7}$$

$$Y_{id} = (\beta_1 + v_{1d*}) x_{id} + \beta_0 + v_{2d*} + e_{id} \tag{8}$$

Simulation study (2) - choice of the model

$$Y_{id} = \beta_1 x_{id} + \beta_0 + e_{id} \tag{4}$$

$$Y_{id} = \beta_1 x_{id} + \beta_0 + v_d + e_{id} \tag{5}$$

$$Y_{id} = (\beta_1 + v_d) x_{id} + \beta_0 + e_{id} \tag{6}$$

$$Y_{id} = (\beta_1 + v_{1d}) x_{id} + \beta_0 + v_{2d} + e_{id} \tag{7}$$

$$Y_{id} = (\beta_1 + v_{1d*}) x_{id} + \beta_0 + v_{2d*} + e_{id} \tag{8}$$

Simulation study (3)

- We compare the properties of two test - **Likelihood Ratio Test** and **parametric bootstrap test**:
- We considered the problem of **model misspecification** and **non-normality** of random effects;
- We divide simulation study into part:
  - **part I** - probability of **type I** errors;
  - **part II** - probability of **type II** errors;
- The number of MC iterations equals 1.000 and the number of bootstrap iterations – 200;
- The simulation study was prepared using R language (R Development Core Team, 2018);

The problem of non-normality of random effects (1)

In part I of the simulation:

- **normal** distribution with expected value equal 0;
- shifted **exponential** distribution with expected value equal 0;
- shifted **gamma** distribution with expected value equal 0 and coefficient of asymmetry equal 4

and variance computed based on real data.

The problem of non-normality of random effects (2)

In part II of the simulation:

- **multivariate normal** distribution with expected values equal 0 and $\rho = \{-0.9; -0.6, -0.3; 0.3; 0.6; 0.9\}$;
- **normal copula** with $\rho = \{-0.9; -0.6, -0.3; 0.3; 0.6; 0.9\}$ and marginal distributions:
  - shifted **exponential** distribution;
  - shifted **gamma** distribution.
- **t copula** with $\rho = \{-0.9; -0.6, -0.3; 0.3; 0.6; 0.9\}$, $df = 3$ and marginal distributions:
  - shifted **exponential** distribution;
  - shifted **gamma** distribution.

The problem of non-normality of random effects - copula function (1)

$$C(u) : [0,1]^n \to [0,1]$$

$$H(x,y) = C\left(F_1(x), F_2(x)\right)$$

(Sklar, 1959)

Gaussian copula

$$C_\rho(u_1, v_2) = \Phi_\rho\left(F_1(u_1), F_2(u_2)\right)$$

t copula

$$C_{\rho,v}(u_1, v_2) = T_{\rho,v}\left(F_1(u_1), F_2(u_2)\right)$$

(cf. Nelsen, 1999, pp. 46-47)

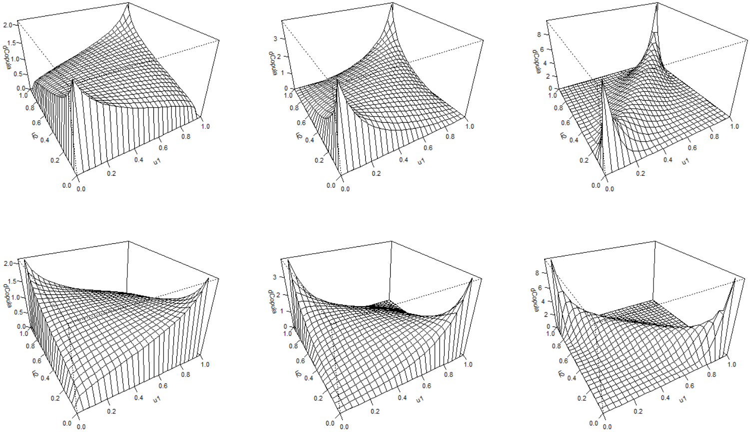
The problem of non-normality of random effects - copula function (2)

Copula functions have applications i.a. in:

- modeling the relationship between mortality and losses of the insurer (Frees et al., 1996; Frees et al., 2005);
- modeling insolvency risk and risk management (Embrechts et al., 2003);
- portfolio optimization (Patton, 2004);
- the processes of multidimensional control and hydrological modeling (Yan, 2006; Genest and Faure, 2007);
- twodimensional storm surges analysis (Ciupak and Rokiciński, 2011);
- modeling the dependence of the prices of futures contracts on agricultural products (Hołda and Malik, 2013).

The problem of non-normality of random effects - copula function (3)

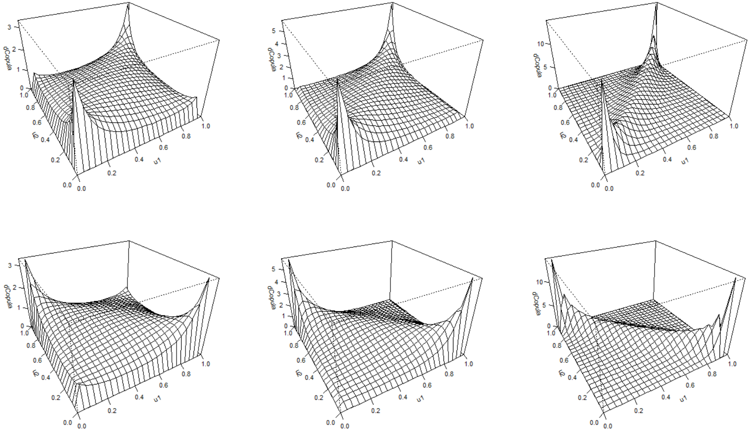
Gaussian copula



(Source: own elaboration)

The problem of non-normality of random effects - copula function (4)

t copula



(Source: own elaboration)

Results - part I

Table 1. The values of probabilities of type I errors.

norm		exp		gam	
LRT	**PB**	**LRT**	**PB**	**LRT**	**PB**
0.042	0.036	0.053	0.042	0.050	0.044

(Source: own elaboration)

Table 2. The values of probabilities of type II errors.

	mvnorm		normcop.exp		tcop.exp		normcop.gam		tcop.gam	
ρ	**LRT**	**PB**	**LRT**	**PB**	**LRT**	**PB**	**LRT**	**PB**	**LRT**	**PB**
-0.9	0.000	0.000	0.000	0.000	0.002	0.001	0.476	0.516	0.521	0.556
-0.6	0.006	0.006	0.029	0.032	0.122	0.127	0.717	0.744	0.800	0.831
-0.3	0.442	0.447	0.626	0.644	0.653	0.662	0.952	0.957	0.901	0.898
0.3	0.397	0.414	0.549	0.553	0.459	0.467	0.691	0.693	0.460	0.461
0.6	0.001	0.004	0.025	0.022	0.041	0.043	0.151	0.156	0.144	0.147
0.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000

(Source: own elaboration)

Conclusions - part I

In the simulation study we show that for the considered real data and model (6):

- **parametric bootstrap test** gives lower values of the probability of **type I** errors than **Likelihood Ratio Test**;
- considered tests are **quite robust** on **non-normality** of random effects if assumption about the lack of correlation is fulfilled;

Conclusions - part II

Obtained results show that:

- for **moderate** and **strong** correlation considered tests have good properties - low values of the probability of **type II** errors, except some cases with marginal gamma distribution of the random effects;
- if random effects have **multivariate normal distribution** the power of the considered tests is higher than **0.5** even for the weak correlation.

Bibliography

- Barnard G.A. (1963), Discussion of "The spectral analysis of point processes"by M. S. Bartlett. Journal of the royal statistical society, B, 25, 264-96.
- Bartlett M.S. (1963), The spectral analysis of point processes. Journal of the royal statistical society B, 25, pp. 264-296.
- Butar F.B., Lahiri P. (2003), On Measures of Uncertainty of Empirical Bayes Small-Area Estimators, Journal of Statistical Planning and Inference, 112, pp. 635–76.
- Chatterje S. et. al.(2008), Parametric bootstrap aproximation to the distribution of EBLUP and related prediction intervals in Linar Mixed Models. The Annals of Statistics, 36, 3, pp. 1221–1245.
- Ciupak M., Rokiciński K.(2011), Zastosowanie metody kopuli do dwuwymiarowej analizy wezbrań sztormowych w profilach wodowskazowych Świnoujście i Kołobrzeg, Zeszyty Naukowe Akademii Marynarki Wojennej, 187, 4, p. 15-34.

Bibliography

- Dumont C. et. al. (2014), Influence of covariance between random effects in design for nonlinear mixed-effect models with an illustration in pediatric pharmacokinetics. Journal of Biopharmaceutical Statistics, 24 (3), pp.471-92.
- Efron B. (2012), Bayesian inference and the parametric bootstrap. The Annals of Applied Statistics, 6, 4, pp. 1971-1997.
- Embrechts P., et al. (2003), Modelling Dependence with Copulas and Applications to Risk Management. In: Handbook of Heavy Tailed Distribution in Finance. Ed. S. Rachev.
- Frees E.W., et al. (2005), Credibility Using Copulas. North American Actuarial Journal, 9, 2.
- Frees E.W., et al. (1996), Annuity Valuation with Dependent Mortality. Journal of Risk and Insurance, 63.
- Genest C., Favre A.C. (2007), Everything You Always Wanted to Know about Copula. Modeling abut Were Afraid to Ask. Journal of Hydrologic Engineering, 12.

Bibliography

- Gonzales-Manteiga W., et. al. (2008) Bootstrap Mean Squared Error of Small-Area EBLUP, Journal of Statistical Computation and Simulation, 78, pp. 433–462.
- Hołda, Malik (2013), Modelowanie zależności cen kontraktów terminowych na produkty rolne notowanych na giełdzie towarowej w Chicago z wykorzystaniem funkcji kopuli. Studia Ekonomiczne. Zeszyty naukowe Uniwersytetu Ekonomicznego w Katowicach, 135, p. 64-78
- Hope A. (1968), A Simplified Monte Carlo Significance Test Procedure. Journal of the Royal Statistical Society. Series B (Methodological),30, 3, pp. 582-598.
- Jiang J. (2007), Linear and Generalized Linear Mixed Models and Their Applications, Springer Science+Business Media, New York.
- Menec V, et. al.(2004), Patterns of Health Care Use and Cost at the End of Life. Winnipeg, MB: Manitoba Centre for Health Policy.
- Nelsen R.B. (1999), An Introduction to Copulas. Springer, New York.
- Pinheiro J.C., Bates D.M. (2000), Mixed-Effects Models in S and S-PLUS, Springer Verlag, New York.

Bibliography

- Patton A. (2004), On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation. Journal of Financial Econometrics, 2, 1.
- Pratesi M., Salvati N. (2008), Small Area Estimation: The EBLUP Estimator Based on Spatially Correlated Random Area Effects. Statistical Methods and Applications, vol. 17, p. 113-141.
- Shaw F.H., Geyer C.J. (1997), Estimation and testing in constrained covariance component models.Biometrika , 84, 1, pp. 95-102.
- Skoglund J., Karlsson S. (2001), Specifiation and estimation of random effects model with serial correlation of general form. SSE/EFI Working paper series in Economics and Finance, 433.
- Yan J. (2006), Multivariate Modeling with Copulas and Engineering Applications. in: Handbook in Engineering Statistics. Red. H. Pham.
- Zhu L. (2005), Nonparametric Monte Carlo Tests and Their Applications. Springer, Berlin, Heidelberg New York.

Thank you
for Your attention