

Estimation of the parameters in copula models for small areas

Louis-Paul Rivest¹, François Verret and Sophie Baillargeon

ABSTRACT

The goal of this work is to investigate models constructed to predict the mean values of a survey variable Y in small areas, using simple random samples of units drawn in each area and auxiliary variables x known for all the population units. A new class of exchangeable models for the dependency between the regression errors within a small area are considered. Besides a linear regression component, these models involve a copula family, indexed by a one dimensional dependency parameter for the within area dependency, and an arbitrary cumulative distribution function for the marginal error distribution. This work focusses on parameter estimation; it provides large sample approximations to the sampling distributions of estimators of the parameters. A Monte Carlo Study validates the results obtained for the regression parameter.

KEY WORDS: Archimedean copulas, Empirical distribution function, Kendall's tau, Mixed normal linear models.

RÉSUMÉ

L'objectif de ce travail est d'étudier des modèles construits pour prédire la valeur moyenne d'une variable d'intérêt Y dans des petits domaines à l'aide d'échantillons aléatoires d'unités sélectionnées dans ces domaines et de variables auxiliaires x connues pour toutes les unités de la population. On y considère une nouvelle classe de modèles échangeables pour l'association, au sein d'un domaine, entre les erreurs du modèle de régression. En plus d'une régression linéaire, ces modèles comportent une famille de copules, indexées par un paramètre unidimensionnel pour la dépendance intra domaine, et une fonction de répartition quelconque pour la distribution marginale des erreurs. Ce travail porte sur l'estimation des paramètres de ces modèles; il donne des approximations asymptotiques des distributions échantillonnales des estimateurs. Une étude par simulation valide les résultats obtenus pour l'estimateur du paramètre de régression.

MOTS CLÉS : Copules Archimédiennes, Fonction de répartition empirique, Modèle linéaire mixte normal, Tau de Kendall

1. INTRODUCTION

Consider a population divided into m small areas of sizes N_1, N_2, \dots, N_m . The variable of interest is Y and x is a p -dimensional vector of auxiliary variables. The data for the whole population is $\{(Y_{ij}, x_{ij}) : i=1, \dots, m; j=1, \dots, N_i\}$. This paper investigates the following probability model for Y given x ,

$$Y_{ij} = x_{ij}^T \beta + \varepsilon_{ij}, \quad i = 1, \dots, m; j = 1, \dots, N_i.$$

The joint cumulative distribution function (cdf) of the experimental errors $\{\varepsilon_{ij} : j=1, \dots, N_i\}$ in a small area is given by

$$\Pr(\varepsilon_{i1} \leq e_{i1}, \dots, \varepsilon_{iN_i} \leq e_{iN_i}) = C_{\alpha, N_i} \{F_e(e_{i1}), \dots, F_e(e_{iN_i})\}, \quad (1.1)$$

where $C_{\alpha, N}$ belongs to an exchangeable family of copula models and F_e is the marginal cdf for the experimental errors. The only assumption on F_e is that it has a null expectation and a finite variance σ^2 . We are interested in semi-parametric models where F_e is not assumed to be in a predetermined parametric family and where $C_{\alpha, N}$ belongs to a one parameter family of exchangeable copulas indexed by a parameter α . The unknown parameters of this model are the regression coefficients β , the dependency parameters α , and the error cdf F_e . This work investigates the estimation of these three parameters using a data set obtained by drawing independent random samples in each small area. As we are dealing with a semi-parametric model, straightforward maximum likelihood estimation does not apply and ad hoc estimators for the three parameters are proposed and investigated.

¹ Louis-Paul Rivest, Department of Mathematics and Statistics, Université Laval, Quebec city, Qc G1V 0A6, Louis-Paul.Rivest@mat.ulaval.ca

2. ERROR MODELS CONSTRUCTED USING COPULAS

2.1 The normal copula

A copula is a multivariate cdf with uniform margins. If (Z_1, \dots, Z_d) is a random vector and if F_i is the marginal cdf of Z_i , $i=1, \dots, d$, then the marginal distribution of $F_i(Z_i)$ is uniform on $(0,1)$ and the joint distribution of $\{F_1(Z_1), \dots, F_d(Z_d)\}$ is a copula denoted by $C(u_1, \dots, u_d)$. The joint cdf of (Z_1, \dots, Z_d) can be expressed as $C\{F_1(z_1), \dots, F_d(z_d)\}$, see Mai & Scherer (2012) for a more elaborate discussion on multivariate copulas. We now construct the copula behind the multivariate normal distribution.

Since the focus is on error models that are exchangeable within areas, we consider the equicorrelation matrix, $\Sigma(\rho, n)$, with entries 1 on the diagonal and correlation $\rho > 0$ off the diagonal. The normal copula with correlation matrix $\Sigma(\rho, n)$ is

$$C_{\rho, n}(u_1, \dots, u_n) = \int_{-\infty}^{\Phi^{-1}(u_1)} \dots \int_{-\infty}^{\Phi^{-1}(u_n)} \frac{\exp\{z^T \Sigma(\rho, n)^{-1} z / 2\}}{(2\pi)^{n/2} \{1 + (n-1)\rho\}^{1/2} (1-\rho)^{(n-1)/2}} dz_1 \dots dz_n \quad u_i \in (0,1), i=1, \dots, n,$$

as the determinant of $\Sigma(\rho, n)$ is $\{1+(n-1)\rho\}(1-\rho)^{n-1}$. The standard error model in small area estimation, proposed by Battese, Harter & Fuller (1988), has $\varepsilon_{ij} = a_i + e_{ij}$ where a_i , the area effect, has a normal distribution with mean 0 and variance σ_a^2 , abbreviated $N(0, \sigma_a^2)$, while the pure error, e_{ij} has a $N(0, \sigma_e^2)$ distribution and all these variables are independent. Their joint error cdf can then be constructed using $N(0, \sigma_a^2 + \sigma_e^2)$ margins and the normal copula $C_{\rho, n}$ for the dependency where $\rho = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$ is the intra-cluster correlation. Thus the standard normal model is a special case of (1.1). It is obtained by setting the copula family equal to $C_{\rho, n}$, the normal copula defined above, and by having the marginal error cdf F_e equal to the $N(0, \sigma_a^2 + \sigma_e^2)$ distribution. For this model the dependency parameter is $\alpha = \rho$.

A generalization of the basic normal model is obtained by taking the normal copula $C_{\rho, n}$ and an arbitrary error cdf F_e , with a null expectation and a finite variance. For such a model, the dependency parameter α (which is equal to ρ) is not necessarily equal to the intra cluster correlation as this correlation depends on the marginal cdf F_e . Parameter α can be recovered from the joint distribution (1.1) using Kendall's tau as this coefficient depends only on the underlying copula. It is defined using a pair, $(\varepsilon_{ij}, \varepsilon_{ij'})$ and $(\varepsilon_{kl}, \varepsilon_{kl'})$, of bivariate error vectors from two different areas, i and k . It is the probability that these two vectors are concordant minus the probability that they are discordant; it can be expressed as

$$\tau = 2 \Pr\{(\varepsilon_{ij} - \varepsilon_{kl}) \times (\varepsilon_{ij'} - \varepsilon_{kl'}) > 0\} - 1.$$

An equivalent expression is

$$\tau = \int_0^1 \int_0^1 C_{\alpha, 2}(u_1, u_2) c_{\alpha, 2}(u_1, u_2) du_1 du_2,$$

where $c_{\alpha, 2}$ is the density of the bivariate copula. For the normal copula there is a simple relationship between the dependency parameter α and τ , namely $\tau = 2 \arcsin(\alpha) / \pi$ see Hult & Lindskog (2002). This section has focussed on the normal distribution however the proposed copula construction generalizes to an arbitrary elliptical distribution with correlation matrix $\Sigma(\rho, n)$. It is, for instance, possible to construct a t -copula with ν degrees of freedom. The formula $\tau = 2 \arcsin(\alpha) / \pi$ holds for all these copula families where $\alpha = \rho$ is the correlation used in the copula construction.

2.2 Archimedean copulas

Archimedean copulas are expressed in terms of $\psi_\alpha(t) = E\{\exp(-ta)\}$, the Laplace transform of a positive latent variable a . They are given by

$$C_{\alpha, n}(u_1, \dots, u_n) = \psi_\alpha^{-1}\{\psi_\alpha^{-1}(u_1) + \dots + \psi_\alpha^{-1}(u_n)\} \quad u_i \in (0,1), i=1, \dots, n$$

where $\psi_\alpha^{-1}(u)$ is the functional inverse of $\psi_\alpha(t)$ and $\alpha > 0$ is the dependency parameter; see Mai & Scherer (2012) for a recent discussion of this class of copulas. The limiting case $\alpha = 0$ gives the independence copula $C_{0, n}(u_1, \dots, u_n) = u_1 \times \dots \times u_n$.

A popular copula family is Clayton's where a has a gamma distribution with shape parameter $1/\alpha$ and scale parameter α , and $\psi_\alpha(t) = (1 + \alpha t)^{-1/\alpha}$. This paper also considers Gumbel's copula with $\psi_\alpha(t) = \exp(-t^{1/(1+\alpha)})$, associated to a latent variable a having a positive stable distribution, and Frank's copula with $\psi_\alpha(t) = -\log[1 + \exp(-t)\{\exp(-\alpha) - 1\}] / \alpha$ whose latent variable a has a logarithmic distribution, see chapter 2 of Mai & Scherer (2012).

Multivariate Archimedean copulas are not symmetrical. Even if the marginal cdf F_e is symmetrical with respect to 0, the cdfs of $(\varepsilon_1, \dots, \varepsilon_N)$ and of $(-\varepsilon_1, \dots, -\varepsilon_N)$ differ. Therefore additional Archimedean models are obtained by assuming that $(-\varepsilon_1, \dots, -\varepsilon_N)$ is distributed according to (1.1). This amounts to using a survival Archimedean copula to model the dependency within a small area, as defined in Mai & Scherer (2012). Just as for the normal copula, Kendall's tau is a function of the dependency parameter α for these copulas. The functional forms for the Clayton, the Gumbel and the Frank copulas are respectively $\tau = \alpha/(\alpha+2)$, $\tau = \alpha/(\alpha+1)$ and $\tau = 1 - 4\{D_1(\alpha) - 1\}/\alpha$, where D_1 is a Debye function of the first kind, see Mai & Scherer (2012) for details.

3. PARAMETER ESTIMATION

3.1 The data

We assume that simple random samples of sizes $\{n_i: i=1, \dots, m\}$ are drawn within each small area. The joint distribution of the sampled errors within small area i is determined by the copula C_{α, n_i} and by the error cdf F_e . This section assumes that the copula family is known; model selection is not considered. Estimators for the parameters (β, α, F_e) are investigated. As the parameter space is infinite dimensional, straightforward maximum likelihood estimation is not possible. This section suggests ad hoc estimators and investigates their large sample properties when m goes to infinity and when the small area sizes N_i are bounded. It presents a sketch of the derivation of the asymptotic properties of the parameters estimators.

3.2 Regression parameter β

To estimate β , we suggest maximizing the likelihood for the normal mixed linear model involving the parameters $(\beta, \sigma_e^2, \sigma_a^2)$. Under error model (1.1), the variance components (σ_e^2, σ_a^2) are defined as follows: σ_a^2 is the covariance between two errors, ε_{ij} and ε_{il} , in area i while $\sigma_e^2 = \sigma^2 - \sigma_a^2$ is the residual error variance. The intra-cluster correlation is then $\rho = \sigma_a^2 / (\sigma_e^2 + \sigma_a^2)$ and the covariance matrix of the sampled errors within area i is $\sigma^2 \Sigma(\rho, n_i)$. Thus under (1.1) the first two error moments are the same as under the standard normal model of Battese et al. (1988) even if, in this general setting, the errors cannot split nicely into independent components for the area effect and the pure error.

The normal log-likelihood for estimating the parameter can be expressed in the terms of y , the $\Sigma n_i \times 1$ data vector, X , the corresponding $\Sigma n_i \times p$ design matrix, and V , the covariance matrix of y which is a $\Sigma n_i \times \Sigma n_i$ block diagonal matrix of $\{\sigma^2 \Sigma(\rho, n_i)\}$. It is given by

$$L(\beta, \sigma_e^2, \sigma_a^2) = -\frac{1}{2} \left[(y - X\beta)^T V^{-1} (y - X\beta) + \log \{|V|\} \right].$$

Let $(\beta_0, \sigma_{0e}^2, \sigma_{0a}^2)$ be the true parameter values and V_0 be the corresponding true covariance matrix for y . We now evaluate the expectation of this log-likelihood and shows that it is maximum at the true parameter values. Since $E\{(y - X\beta)(y - X\beta)^T\} = V_0 + X(\beta_0 - \beta)(\beta_0 - \beta)^T X^T$, one has

$$\begin{aligned} E\{L(\beta, \sigma_e^2, \sigma_a^2)\} &= -\frac{1}{2} \left(\text{tr} \left[V^{-1} E\{(y - X\beta)(y - X\beta)^T\} \right] + \log \{|V|\} \right) \\ &= cte - \frac{1}{2} \left\{ \text{tr}(V^{-1} V_0) + (\beta_0 - \beta)^T X^T V^{-1} X (\beta_0 - \beta) - \log \{|V^{-1} V_0|\} \right\} \end{aligned}$$

For any value of (σ_e^2, σ_a^2) , this expectation is maximum at $\beta = \beta_0$ provided that $X^T V^{-1} X$ is positive definite. In addition, as seen in Watson (1964), on the set of positive definite matrices, $-\text{tr}(\Sigma) + \log |\Sigma|$ is maximum at $\Sigma = I$. This shows that the expectation of the normal log-likelihood is maximized at the true parameter values. Thus, under suitable regularity conditions, the estimators $(\hat{\beta}, \hat{\sigma}_e^2, \hat{\sigma}_a^2)$ maximizing $L(\beta, \sigma_e^2, \sigma_a^2)$ are consistent. The sampling properties of estimators derived from such a misspecified likelihood function are investigated in Huber (1967) who showed that the asymptotic covariance matrix of the estimators have a sandwich format involving the covariance matrix, under the true model, of the score function and the inverse of the matrix of second order partial derivatives of the log-likelihood.

We now evaluate the covariance matrix of $\hat{\beta}$. Our objective is to find out whether the normal based covariance matrix of $\hat{\beta}$ provides an asymptotically unbiased covariance estimator under error model (1.1). The score function for the normal

mixed model, as derived in McCullough et al (2008), has two components, one for β and one for the variance components $\sigma_e^2(j=1)$ and $\sigma_a^2(j=2)$. They are

$$s(\beta, \sigma_e^2, \sigma_a^2) = \begin{pmatrix} X^T V^{-1} (y - X\beta) \\ -(y - X\beta)^T V^{-1} V_j V^{-1} (y - X\beta) / 2 - \text{tr}(V^{-1} V_j) / 2 \end{pmatrix},$$

where V_j is the derivative of V with respect to the j th variance component. These calculations use the fact that the derivatives of V^{-1} with respect to the j th variance component is $-V^{-1} V_j V^{-1}$, while the derivative of $\log|V|$ is $\text{tr}(V^{-1} V_j)$. The covariance matrix of the score function has the following partitioned form:

$$V\{s(\beta, \sigma_e^2, \sigma_a^2)\} = \begin{pmatrix} X^T V^{-1} X & W_{12} \\ W_{21} & W_{22} \end{pmatrix},$$

where the off diagonal term W_{12} involves residual moments of order 3 and W_{22} depends on fourth order error moments. Note that W_{12} is null for a normal model however it is not so in general. Now, minus the partial derivatives of the score function are given by

$$-\frac{\partial s(\beta, \sigma_e^2, \sigma_a^2)}{\partial(\beta, \sigma_e^2, \sigma_a^2)} = \begin{pmatrix} X^T V^{-1} X & -X^T V^{-1} V_j V^{-1} (y - X\beta) \\ -(y - X\beta)^T V^{-1} V_j V^{-1} X & S_{22} \end{pmatrix},$$

where S_{22} is a complicated matrix involving variance derivatives. When taking the expectation, the off-diagonal terms vanish, and the asymptotic covariance matrix of $(\hat{\beta}, \hat{\alpha}, \hat{\sigma}^2)$ is equal to

$$V(\hat{\beta}, \hat{\sigma}_e^2, \hat{\sigma}_a^2) = \begin{pmatrix} (X^T V^{-1} X)^{-1} & 0 \\ 0 & ES^{22} \end{pmatrix} \begin{pmatrix} X^T V^{-1} X & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} (X^T V^{-1} X)^{-1} & 0 \\ 0 & ES^{22} \end{pmatrix},$$

where ES^{22} is the inverse of the expectation of S_{22} . The asymptotic covariance matrix of $\hat{\beta}$ is $(X^T V^{-1} X)^{-1}$, the (1,1) block of the above matrix. It is equal to the covariance matrix of $\hat{\beta}$ under a standard normal model. This means that the covariance estimator for $\hat{\beta}$ obtained when fitting a normal mixed linear model is valid under the general copula error model considered in this work.

Note also that the covariance matrix of $\hat{\beta}$ is equal to that of the Best Linear Unbiased Estimator. Thus under a general copula error model, $\hat{\beta}$ is not a maximum likelihood estimator. However, it is the best among all the unbiased linear estimators. In particular, it is better than the simple linear estimator obtained by regressing y on X , without accounting for the within area dependency.

3.3 Error cdf F_e

Let $e_{ij} = y_{ij} - x_{ij}^T \hat{\beta}$ be the regression residual for unit j in area i . As these residuals do not necessarily sum to 0, we use the centered residuals, $e_{ij}^c = e_{ij} - \bar{e}_{\square}$, to estimate F_e , since it is assumed to have a null expectation. The proposed estimator is

$$\hat{F}_e(z) = \frac{1}{(\sum n_i) + 1} \sum_{i=1}^m \sum_{j \in S_i} 1_{\{e_{ij}^c \leq z\}}.$$

Define

$$\tilde{F}_e(z) = \frac{1}{(\sum n_i) + 1} \sum_{i=1}^m \sum_{j \in S_i} 1_{\{e_{ij} \leq z\}}.$$

Clearly the expectation of $\tilde{F}_e(z)$ is $F_e(z)$ and its variance is

$$\text{Var}\{\tilde{F}_e(z)\} = \frac{1}{\{(\sum n_i) + 1\}^2} \sum_{i=1}^m (n_i F_e(z) \{1 - F_e(z)\} + n_i(n_i - 1) [C_{\alpha,2}\{F_e(z), F_e(z)\} - F_e(z)^2]).$$

To derive an approximation to the sampling distribution of $\hat{F}_e(z)$ we approximate the difference $\hat{F}_e(z) - \tilde{F}_e(z)$. When the regression errors are independent, the asymptotic properties of the empirical residual cdf are well known, see for instance Mammen (1996). We provide a sketch of an extension of these results to hierarchical errors.

Define ε as the $\sum n_i \times 1$ vector of experimental errors. Using the approximation, $\hat{\beta} \approx \beta + (X^T V^{-1} X)^{-1} X^T V^{-1} \varepsilon$, one can express e_{ij}^c as

$$e_{ij}^c \approx \varepsilon_{ij} - \bar{\varepsilon}_{\square} - (x_{ij}^T - \bar{x}_{\square})^T (X^T V^{-1} X)^{-1} X^T V^{-1} \varepsilon \approx \varepsilon_{ij} - \theta_1 - (x_{ij}^T - \bar{x}_{\square})^T \theta_2$$

where $\theta_1 = \bar{\varepsilon}_{\square}$ and $\theta_2 = (X^T V^{-1} X)^{-1} X^T V^{-1} \varepsilon$ have $O_p(m^{-1/2})$ entries. Now let θ_1 and θ_2 be respectively an $O(m^{-1/2})$ constant and a $p \times 1$ vector of $O(m^{-1/2})$ nonrandom terms and consider the difference $\frac{1}{(\sum n_i) + 1} \sum_{i=1}^m \sum_{j \in s_i} \left[1_{\{\varepsilon_{ij} - \theta_1 - \theta_2^T (x_{ij} - \bar{x}_{\square}) \leq z\}} - 1_{\{\varepsilon_{ij} \leq z\}} \right]$. The proof consists in showing that this quantity is approximately equal to its expectation,

$$\frac{1}{(\sum n_i) + 1} \sum_{i=1}^m \sum_{j \in s_i} \left[F_e \{z + \theta_1 + \theta_2^T (x_{ij} - \bar{x}_{\square})\} - F_e(z) \right] \approx f_e(z) \theta_1,$$

where $f_e(z)$ is the density of $F_e(z)$. If this holds uniformly in θ_1 and θ_2 , we get the approximation $\hat{F}_e(z) \approx \tilde{F}_e(z) + f_e(z) \bar{\varepsilon}_{\square}$ which characterizes the large sample distribution of $\hat{F}_e(z)$.

The proof uses Hoeffding's inequality (see Mammen, 1996): if the $\{A_i\}$ are independent bounded random variables with a null expectation such that $|A_i| < M$, for some positive M , then

$$\Pr(|A_1 + \dots + A_m| > c) \leq 2 \exp \left(- \frac{c^2}{2 \sum \text{Var}(A_i) + 2Mc/3} \right).$$

In our context the random variables A_i are

$$A_i = \frac{\sum_{j \in s_i} \left[1_{\{\varepsilon_{ij} - \theta_1 - \theta_2^T (x_{ij} - \bar{x}_{\square}) \leq x\}} - 1_{\{\varepsilon_{ij} \leq x\}} - F_e \{x + \theta_1 + \theta_2^T (x_{ij} - \bar{x}_{\square})\} + F_e(x) \right]}{(\sum n_i) + 1} = \frac{\sum_{j \in s_i} \delta_{ij}}{(\sum n_i) + 1},$$

where the entries of θ_1 and θ_2 are $O(m^{-1/2})$. The bound M is $3 \times \max(n_i) / \sum n_i$ and is $O(1/m)$. The constant c is equal to $\eta/m^{1/2}$, for an arbitrarily small $\eta > 0$. We want to show that as m goes to infinity, $\Pr(|A_1 + \dots + A_m| > c)$ converges to 0, proving that $\sum A_i$ is asymptotically negligible. Using Hoeffding's inequality, this holds as long as $\text{Var}(A_i)$ is $o(1/m^2)$. One has

$$\text{Var}(A_i) = \frac{n_i \text{Var}(\delta_{i1}) + n_i(n_i - 1) \text{Cov}(\delta_{i1}, \delta_{i2})}{\{(\sum n_i) + 1\}^2} \leq \frac{n_i^2 \text{Var}(\delta_{i1})}{\{(\sum n_i) + 1\}^2}$$

and this goes to 0 at a $o(1/m^2)$ rate since $\text{Var}(\delta_{ij})$ goes to 0 as m becomes large. This completes a sketch of the proof.

3.4 Dependency parameter α

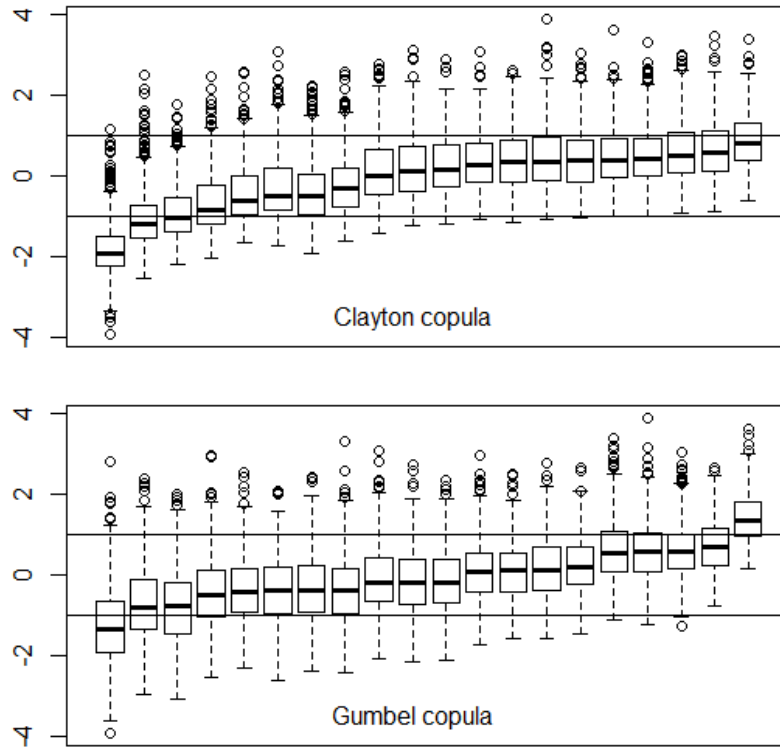
This parameter is estimated by solving the equation $\hat{\tau} = g(\alpha)$ where $g(\alpha)$ is the function expressing Kendall's tau as a function of α . The statistic $\hat{\tau}$ is calculated with the regression residuals e_{ij} . First the $\sum_{i > j} n_i(n_i - 1)n_j(n_j - 1)$ pairs of ordered bivariate vectors of residuals coming from two different areas are examined to determine whether they are concordant or not. The estimate $\hat{\tau}$ is the proportion of concordant pairs; its asymptotic distribution is investigated in Rhomdani et al. (2014) who propose an estimator for its sampling variance.

4. MONTE CARLO INVESTIGATION OF THE SAMPLING PROPERTIES OF $\hat{\beta}$

The goal of this section is to investigate whether $(X^T \hat{V}^{-1} X)^{-1}$ provides a good approximation to the sampling covariance matrix of $\hat{\beta}$ when the errors come from the general copula error model (1.1) studied in this work. We considered three copulas, Normal, Clayton and Gumbel, and two marginal distributions, either normal or a centered exponential both with a variance equal to 1. In the simulations we used a Kendall tau of 1/3, corresponding to α equal to 1/2, 1 and 0.5 respectively for the three copula families. The simulated populations have $m=20$ areas with two sample sizes, either $n_i=3$ in each one or $n_i=5$ in each one. A single explanatory variable with a $N(2, .35^2)$ distribution was used and both the intercept and the slope were set to 1; $R=10,000$ Monte Carlo replications were run for each scenario.

To help visualize the differences between the copulas considered in the simulations, Figure 1 presents the normal score boxplots for the Clayton and the Gumbel copulas. These boxplots are constructed by simulating $m=20$ random vectors from a copula $C_{\alpha,500}$. The $20 \times 500 = 10^4$ random variables are replaced by their normal scores, $\Phi^{-1}\{(R_{ij}-1/2)/10^4\}$, where R_{ij} is the rank of the j th entry of the i th vector in the combined data set. Then the normal score medians are calculated for each vector and the 20 vectors are ordered by increasing median. The graphs in Figure 1 present the boxplots of the normal scores for the 20 ordered vectors. The marginal distribution of the normal scores is obviously normal however the conditional distributions within areas are not. For instance, Gumbel copula has an extreme area where most of the large errors are found. Note also the positive skewness of the within area distributions, especially for the Clayton copula. This contrasts with a normal copula where the within area distributions of the normal scores are normal.

Figure1 – Normal score boxplots for two copula models with $\tau=1/3$.



The simulations focus on the slope parameter β_1 . Each run reports a slope estimate $\hat{\beta}_1$, its variance estimate $v(\hat{\beta}_1)$ calculated by the R-function `lme` from the package `n1me`, and an indicator function that takes the value 1 if the confidence interval $\hat{\beta}_1 \pm 1.96\sqrt{v(\hat{\beta}_1)}$ covers the true value $\beta_1=1$. The Monte Carlo relative biases $B(\hat{\beta}_1)$ and $B\{v(\hat{\beta}_1)\}$ are evaluated using the formulae,

$$B(\hat{\beta}_1) = \frac{1}{R} \sum_{r=1}^R \hat{\beta}_1^{(r)} - 1 \quad \text{and} \quad B\{v(\hat{\beta}_1)\} = \frac{\sum_{r=1}^R v(\hat{\beta}_1)^{(r)}}{\sum_{r=1}^R (\hat{\beta}_1^{(r)} - 1)^2} - 1.$$

The coverage of the confidence interval is the proportion of 1 among the indicator functions. The results are presented in Table 1.

In Table 1, the biases are in general smaller than 5% and the coverage of the confidence intervals are close to the nominal 95% level. The Gumbel model with exponential errors stands out as it has larger biases and poorer coverage than the other error models. Considering Figure 1, the Gumbel copula has a positive skewness in the errors of the more extreme regions. Combined with the positive skewness of the exponential distribution, this might create areas with outlying values that affect the sampling properties of $\hat{\beta}_1$. In general, the sampling properties reported in Table 1 are quite good and they provide an empirical validation of the findings of Section 3.2: the inference procedures for $\hat{\beta}$ derived from a mixed normal linear model are valid when the true error distribution belongs the family presented in equation (1.1).

Table 1 –Relative biases, in percentage, of $\hat{\beta}_1$ and of $v(\hat{\beta}_1)$ and confidence interval coverage.

copula	Margin	$B(\hat{\beta}_1)$		$B\{v(\hat{\beta}_1)\}$		Coverage	
		$n_i=5$	$n_i=3$	$n_i=5$	$n_i=3$	$n_i=5$	$n_i=3$
Normal	Normal	0.0	0.0	-1.4	-5.2	94.4	93.6
Normal	Exponential	-0.2	-0.4	-1.5	-5.8	94.5	93.6
Clayton	Normal	0.2	-0.2	2.1	-0.8	94.1	94.4
Clayton	Exponential	-0.1	-0.3	-0.8	-0.7	94.4	94.5
Gumbel	Normal	-0.1	0.1	-1.8	-1.8	94.5	94.1
Gumbel	Exponential	0.2	-0.2	-3.3	-4.4	89.1	88.5

5. CONCLUSIONS

Estimators of the parameters of a semi-parametric model for small area estimation have been proposed and asymptotic approximations to their sampling distributions have been derived. An important finding of this work is that inference procedures for regression parameter β , constructed using a standard normal mixed model, are valid under error model (1.1). This has been demonstrated using asymptotic expansions and by Monte Carlo simulations.

Acknowledgements

Louis-Paul Rivest thanks, for their hospitality, the Department of Mathematics and Statistics of the University of Western Australia where this work was carried out. The financial supports of the Natural Sciences and Engineering Research Council of Canada, of the Canadian Statistical Sciences Institute, and of the Canada Research Chair in Statistical Sampling and Data Analysis are gratefully acknowledged.

REFERENCES

- Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). "An error-components model for prediction of county crop areas using survey and satellite data". *Journal of the American Statistical Association*, **83**, 28-36
- Huber, P. J. (1967). "The behavior of maximum likelihood estimates under non-standard conditions". *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol 1, 221-233, University of California Press: Berkeley
- Hult, H. & Lindskog, F. (2002). "Multivariate extremes, aggregation and dependence in elliptical distributions". *Advances in Applied Probability*, **34**, 587-608.
- McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models, 2nd Edition*. John Wiley: New York
- Mai, J.-M. & Scherer, M. (2012). *Simulating Copulas; Stochastic Models, Sampling Algorithms and Applications. Series in Quantitative Finance: Volume 4*. World Scientific Publishing Company. Imperial College Press: London
- Mammen, E. (1996). "Empirical process of residuals for high-dimensional linear models". *The Annals of Statistics*, **24**, 307-335
- Romdhani, H., Lakhel-Chaieb, L. & Rivest, L.-P. (2014). "An exchangeable Kendall's tau for clustered data". *Canadian Journal of Statistics*, **42**, 384-403
- Watson, G. S. (1964). "A note on maximum likelihood". *Sankhya A*, **26**, 302-03