Revista Colombiana de Estadística

Número especial en Bioestadística

Volumen 35. Número 2 - junio - 2012 ISSN 0120 - 1751

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA

Revista Colombiana de Estadística

http://www.estadistica.unal.edu.co/revista
http://es.wikipedia.org/wiki/Revista_Colombiana_de_Estadistica
http://www.emis.de/journals/RCE/
revcoles_fcbog@unal.edu.co

Indexada en: Ulrichsweb, Scopus, Science Citation Index Expanded (SCIE), Web of Science (WoS), SciELO Colombia, Current Index to Statistics, Mathematical Reviews (MathSci), Zentralblatt Für Mathematik, Redalyc, Latindex, Publindex (A_1)

La *Revista Colombiana de Estadística* es una publicación semestral del Departamento de Estadística de la Universidad Nacional de Colombia, sede Bogotá, orientada a difundir conocimientos, resultados, aplicaciones e historia de la estadística. La Revista contempla también la publicación de trabajos sobre la enseñanza de la estadística.

Se invita a los editores de publicaciones periódicas similares a establecer convenios de canje o intercambio.

Dirección Postal:
Revista Colombiana de Estadística
© Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Carrera 30 No. 45-03
Bogotá – Colombia
Tel: 57-1-3165000 ext. 13231
Fax: 57-1-3165327

Adquisiciones:
Punto de venta, Facultad de Ciencias, Bogotá.
Suscripciones:
revcoles_fcbog@unal.edu.co
Solicitud de artículos:
Se pueden solicitar al Editor por correo físico o electrónico; los más recientes se pueden obtener en formato PDF desde la página Web.

Edición en LaTeX: Patricia Chávez R. E-mail: apchavezr@gmail.com

Impresión: Editorial Universidad Nacional de Colombia, Tel. 57-1-3165000 Ext. 19645, Bogotá.

Revista Colombiana de Estadística	Bogotá	Vol. 35	Nº 2
ISSN 0120 - 1751	COLOMBIA	junio-2012	Págs. 185-330

Número especial en Bioestadística

Contenido

Francisco J. Torres-Avilés, Gloria Icaza & Reinaldo B. Arellano-Valle
An Extension to the Scale Mixture of Normals for Bayesian Small-Area Estimation 185-204

Mónica Catalán, M. Purificación Galindo, Javier Martín & Víctor Leiva
Métodos de integración de odds ratio basados en meta-análisis usando modelos de efectos fijos y aleatorios útiles en salud pública 205-222

Diego Alejandro Salazar, Jorge Iván Vélez & Juan Carlos Salazar
Comparison between SVM and Logistic Regression: Which One is Better to Discriminate? 223-237

Carlos Mario Lopera-Gómez, Mario César Jaramillo-Elorza & Natalia Acosta-Baena
¿Cuándo inicia la enfermedad de Alzheimer? Kaplan-Meier versus Turnbull: una aplicación a datos con censura arbitraria 239-254

Eduardo Dávila, Luis Alberto López & Luis Guillermo Díaz
A Statistical Model for Analyzing Interdependent Complex of Plant Pathogens . . 255-270

Minerva Montero, Maria Elena Díaz, Santa Jiménez, Iraida Wong & Vilma Moreno
Modelación de indicadores del estado nutricional de la embarazada desde un enfoque multinivel 271-287

Rajvir Singh, Vrijesh Tripathi, Mani Kalaivani, Kalpana Singh & S.N. Dwivedi
Determinants of Birth Intervals in Tamil Nadu in India: Developing Cox Hazard Models with Validations and Predictions 289-307

Carlos Alberto Martínez, Mauricio Elzo, Carlos Manrique, Luis Fernando Grajales & Ariel Jiménez
Random Regression Models for Estimation of Covariance Functions, Genetic Parameters and Prediction of Breeding Values for Rib Eye Area in a Colombian Bos indicus-Bos *taurus Multibreed Cattle Population* 309-330

Editorial

Número especial de la *Revista Colombiana de Estadística* en Bioestadística

Liliana López-Kleine[a], B. Piedad Urdinola[b]

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

La importancia de la bioestadística es innegable. Sus aplicaciones cubren áreas aparentemente disímiles, como medicina, zoología, demografía, genética, agricultura, epidemiología, veterinaria y biología con un común, el cual son los desarrollos en la estadística que ayudan a resolver problemas teóricos, aplicados y más recientemente computacionales.

Colombia, con todo su potencial en biodiversidad, agrícola, sus recursos humanos y el despegue reciente de la investigación no puede ser el gran ausente en los aportes que cada vez son más y mejores en esta área de la estadística. Con este objetivo en mente decidimos convocar este número especial en Bioestadística para la Revista Colombiana de Estadística.

Con éxito recibimos 18 trabajos que cubrieron todos los sub-temas dentro de la bioestadística. Nueve de ellos, que en efecto cubren preguntas de gran relevancia teórica, aplicaciones en epidemiología, demografía, estudios genómicos y biológicos, y en salud pública fueron seleccionados para publicación luego del proceso arbitral. Esperamos que estos trabajos sean una motivación para muchos investigadores y grupos de investigación multidisciplinarios para el desarrollo de teorías en bioestadística y aplicaciones en diversas áreas de las ciencias biológicas. Igualmente, esperamos que sean de gran utilidad para el caso colombiano que cada día cuenta con mayores y mejores datos en todas estas áreas y para la bioestadística en general.

Agradecemos en particular la participación y apoyo de los miembros del Comité Invitado, cuyo dedicado trabajo y el de los árbitros permitieron que esta empresa fuera posible y a la Asistenta Editorial de la revista. Gracias,

[a] Editora invitada del número especial de la Revista Colombiana de Estadística. Profesora asociada. E-mail: llopezk@unal.edu

[b] Editora invitada del número especial de la Revista Colombiana de Estadística. Profesora asociada. E-mail: bpurdinolac@bt.unal.edu.co

Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 185 a 204

An Extension to the Scale Mixture of Normals for Bayesian Small-Area Estimation

Una extensión a la mezcla de escala de normales para la estimación Bayesiana en pequeñas áreas

Francisco J. Torres-Avilés[1,a], Gloria Icaza[2,b],
Reinaldo B. Arellano-Valle[3,c]

[1] Departamento de Matemática y Ciencia de la Computación, Facultad de Ciencia, Universidad de Santiago de Chile, Santiago, Chile

[2] Instituto de Matemática y Física, Universidad de Talca, Talca, Chile

[3] Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile

Abstract

This work considers distributions obtained as scale mixture of normal densities for correlated random variables, in the context of the Markov random field theory, which is applied in Bayesian spatial intrinsically autoregressive random effect models. Conditions are established in order to guarantee the posterior distribution existence when the random field is assumed as scale mixture of normal densities. Lung, trachea and bronchi cancer relative risks and childhood diabetes incidence in Chilean municipal districts are estimated to illustrate the proposed methods. Results are presented using appropriate thematic maps. Inference over unknown parameters is discussed and some extensions are proposed.

***Key words*:** Disease mapping, Markov random field, Hierarchical model, Incidence rate, Relative risk.

Resumen

Este trabajo aborda las distribuciones obtenidas como mezcla de escala de normales para variables aleatorias correlacionadas, en el contexto de la teoría de los campos markovianos, la cual es aplicada a modelos bayesianos espaciales con efectos aleatorios autoregresivos intrínsecos. Se establecen condiciones para garantizar la existencia de la distribución a posteriori cuando se

[a] Assistant professor. E-mail: francisco.torres@usach.cl
[b] Associate professor. E-mail: gicaza@utalca.cl
[c] Professor. E-mail: reivalle@mat.puc.cl

asume una distribución mezcla de escala de normales para el campo markoviano propuesto. Para ilustrar los métodos propuestos, se estiman los riesgos relativos de cáncer de tráquea, bronquios y pulmón, y tasas de incidencia de diabetes tipo 1 en distritos municipales de Chile. Los resultados son presentados usando mapas temáticos apropiados. Se discute la inferencia sobre los parámetros desconocidos y se proponen algunas extensiones.

***Palabras clave*:** campo aleatorio markoviano, mapeo de enfermedades, modelo jerárquico, riesgo relativo, tasa de incidencia.

1. Introduction

Over the last two decades, Bayesian spatial models have become increasingly popular for epidemiologists and statisticians. In particular, small-area modeling is oriented to illustrate the behavior of rates or relative risks associated to each district that form a region or a country, that is, recognition of spatial patterns through maps is the main aim of these methodologies. The conventional assumption to estimate the standardized mortality ratios (SMRs) or incidence rates is based on the Poisson distribution. This assumption may cause several problems in this class of studies, mainly because of the extra-poisson variation. This extra-Poisson variation generally arises when the observed number of cases on each small-area are more variable than the variation contributed by the standard Poisson model (Mollié 2000). Bayesian models have been developed to solve this problem, introducing random effects to account for unobserved spatial heterogeneity; even more, Markov chain Monte Carlo (MCMC) methods led to an explosive increment of the use of Bayesian analysis in these areas of application.

Important works that develop and use Bayesian theory are mentioned in the following lines. The pioneering work in this direction was done by Clayton & Kaldor (1987) who proposed an empirical Bayes approach with application to lip cancer data in Scotland. In Ghosh, Natarajan, Stroud & Carlin (1998), conditions to demonstrate Bayesian generalized linear model (GLM) integrability are formalized under improper prior assumptions in order to represent lack of knowledge over unknown parameters. Best, Arnold, Thomas, Waller & Collon (1999) investigated several spatial prior distributions based on Markov random field (MRF) theory, and discussed methods for model comparison and diagnostics. Pascutto, Wakefield, Best, Richardson, Bernardinelli, Staines & Elliott (2000) examined some structural and functional assumptions of these models and illustrated their sensitivity through the presentation of results related to informal sensitivity analysis for prior distributions choices. They also explored the effect caused by outlying areas, assuming a Student-t distribution for the nonstructured effect.

Recently, Parent & Lesage (2008) proposed a linear Bayesian hierarchical model to study the knowledge spillovers in European countries, under different specifications of the proximity structures. They also compared this effect through different strategies, for example allowing different prior distributions or Student-t errors, to include heterogeneity in the disturbances.

As it was previously mentioned, the class of spatial models has been related to GLM theory, considering random effects to represent the influence of geography. Besag (1974) presented a pioneering work in the context of the MRF theory, with applications to regular lattice systems when spatial heterogeneity is considered.

Furthermore, the most used structure follows the work developed by Besag (1986), who presents a definition in the context of a MRF: Let $\mathbf{u} = (u_1, u_2, \ldots, u_m)'$ be a set of m random variables and $\mathbf{u}_{-i}$ the random vector without the i-th component, where m represents a number of different and contiguous areas. If the joint distribution of $\mathbf{u}$ can be expressed by each conditional distribution, $u_i \mid \mathbf{u}_{-i}, i = 1, \ldots, m$, then it is called a MRF. Intrinsically conditional autoregressive (CAR) random effects are defined as a particular MRF, initially proposed by Besag, York & Mollié (1991), which name is related to the impropriety of the joint distribution generated by the univariate conditional distributions of $u_i \mid \mathbf{u}_{-i}, i = 1, \ldots, m$ (see details in Banerjee, Carlin & Gelfand 2004).

In this work, an intrinsically Gaussian MRF is considered and its properties are extended to a more general family of continuous distributions. Scale mixture of normal (SMN) distributions have been proposed as robust extensions of the normal model. The genesis of this class of models is presented by Andrews & Mallows (1974). The SMN class of distributions is generated if the vector of interest, $\mathbf{u}$, can be represented as

$$\mathbf{u} = \boldsymbol{\mu} + \psi^{-1/2}\mathbf{z} \tag{1}$$

where $\boldsymbol{\mu}$ is a location vector parameter, and $\mathbf{z}$ and ψ are independent, with $\mathbf{z}$ following a multivariate zero centered normal distribution with covariance matrix Σ and ψ being a non-negative random scale factor with c.d.f. $F_\psi(\cdot \mid \nu)$, so that $F_\psi(0 \mid \nu) = 0$. Here ν is an additional or set of parameters controlling the kurtosis of the distribution of ψ. The SMN distributions have been shown to be a subclass of the elliptical distributions family by Fang, Kotz & Ng (1990). This subfamily presents properties similar to the normal distributions, except that their behavior allows capturing unusual patterns present in the data. In a Bayesian context, robust linear models have been studied since Zellner (1976). The multivariate Student-t and the multivariate Slash distributions are examples of this class of distributions.

Following the above ideas, heavier-tailed models will be assumed instead of working with the usual assumption of normality for $\mathbf{u}$, through the Student-t and Slash distributions developed by Geweke (1993) and Lange & Sinsheimer (1993), respectively. Specifically, the Slash distribution considered in this work corresponds to the distribution of the random vector $\psi^{-1/2}\mathbf{z}$, where $\mathbf{z}$ and ψ are independent, with $\mathbf{z}$ having a multivariate normal distribution as in (1) and $\psi \mid \nu$ following a distribution $Beta(\nu/2, 1)$, $\nu > 0$. The Student-t distribution is obtained through the same representation as the Slash distribution, with the difference that $\psi \mid \nu$ follows a Gamma distribution, where both parameters are equal to $\nu/2$, $\nu > 0$. There are some potential problems with the Slash distribution that probably has resulted in more use of the Student-t. However, the Slash distribution may allow for fatter tails (more extreme values) than the Student-t.

From a MRF context, the class of SMN can be found in papers developed from a geological point of view, where prediction is the main focus. Student-t distributed MRF was treated by Roislien & Omre (2006) using a frequentist approach. Lyu & Simoncelli (2007) made the extension of Gaussian MRF theory to what they called Gaussian scale mixture fields, for image reconstruction modeling.

In this work, Bayesian non-Gaussian spatial models are developed to detect unusual rates or relative risks in a particular area under the following scheme. Standard small-area models are presented in Section 2. SMN theory is applied to extend the Gaussian MRF model (Besag 1974) in Section 3. In Section 4, non-Gaussian models are developed trough extensions of the spatial random effect following a Gaussian MRF to the scale mixture of Normal random field (SMN RF) proposed previously. Three different models are used to estimate the incidence rates of Insulin Dependent Diabetes Mellitus (IDDM) in the Chilean Metropolitan Region, and Respiratory Cancer mortality in the northern regions of Chile. These results are presented in Section 5. Finally, some comments and discussion are made in Section 6.

2. Spatial Models with Random Effects

Let $\mathbf{y} = (y_1, \dots, y_m)'$ be a set of m random variables indexed to a specific region. A general formulation is assumed from the generalized linear mixed model theory (Breslow & Clayton 1993), which includes the following elements:

1. A specification of the likelihood function as member of the exponential family, namely

$$f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^{m} \exp\{\phi_i^{-1}(y_i\theta_i - g(\theta_i)) + \rho(\phi_i; y_i)\} \tag{2}$$

   where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$ is the vector of canonical parameters, $\boldsymbol{\phi} = (\phi_1, \dots, \phi_m)'$ is a vector of known scale parameters, g is a known function that does not depend on the data, and ρ is a known function that does not depend on the unknown parameters.

2. A random specification for the link function, $h(\theta_i) = E(y_i \mid \theta_i)$, is typically represented by the normal linear mixed model

$$h(\theta_i) \mid \mathbf{x}_i, \boldsymbol{\beta}, u_i, \sigma^2 \overset{ind.}{\sim} Normal(\mathbf{x}_i'\boldsymbol{\beta} + u_i, \sigma^2) \tag{3}$$

   where the $\mathbf{x}_i$s are a $p \times 1$ vectors of covariates associated to a $p \times 1$ vector of coefficients $\boldsymbol{\beta}$, the u_is represent spatial random effects, and σ^2 measures the nonstructured variability.

3. A model specification for the spatially structured random effects u_is. Typically, Gaussian assumptions for the u_is are made. In the literature it is recurrent to find that these spatial random effects are influenced by a predefined neighborhood represented by an adjacency matrix $\mathbf{D}_w$, controlling the

local variability. Hence, the mean is smoothed by the information given by its neighbors.

Let $\pi(\mathbf{u} \mid \sigma_u^2, \mathbf{D}_w) = \pi(u_1, \dots, u_m \mid \sigma_u^2, \mathbf{D}_w)$ be the joint probability distribution derived from a MRF given a dispersion parameter σ_u^2 and a $m \times m$ $m \times m$ adjacency matrix $\mathbf{D}_w$. A multivariate Gaussian distribution is then obtained when,

$$\pi(\mathbf{u} \mid \sigma_u^2, \mathbf{D}_w) \propto \frac{1}{(\sigma_u^2)^{m/2}} \exp\left\{-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{D}_w\mathbf{u}\right\} \tag{4}$$

A specific case is considered in this work, where $\mathbf{D}_w$ has diagonal elements w_{i+} representing the number of neighbors of the i-th component, and off-diagonal elements w_{ij} taking values -1 if the elements i and j share boundary, denoted by $i \sim j$, and 0 in other case, i.e.,

$$w_{ij} = \begin{cases} w_{i+} & i = j \\ -1 & i \neq j; i \sim j \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Under 5, equation 4 is reduced to

$$\pi(\mathbf{u} \mid \sigma_u^2, \mathbf{D}_w) \propto \frac{1}{(\sigma_u^2)^{m/2}} \exp\left\{-\frac{1}{2\sigma_u^2}\sum_{i\sim j}(u_i - u_j)^2\right\} \tag{6}$$

A basic discussion and treatment of several proximity matrices can be found in Banerjee et al. (2004). A constraint will be imposed to this expression to guarantee integrability.

4. As a final step of the modeling, prior distributions are required for the unknown parameters to complete the hierarchical model. Usual non-informative prior distributions are represented by

$$\begin{array}{ll} i. & \pi(\boldsymbol{\beta}) \propto constant \\ ii. & \sigma^{-2} \sim Gamma(a/2, b/2) \\ iii. & \sigma_u^{-2} \sim Gamma(c/2, d/2) \end{array} \tag{7}$$

where the improper prior $\pi(\boldsymbol{\beta})$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)' \in \mathbb{R}^p$, is assumed according with Ghosh et al. (1998). The hyperparameters $a, b, c, d > 0$ are known constant. Here, both σ^2 and σ_u^2 represent the dispersion parameters included in the model; σ_u^2 is the local dispersion parameter related to a specific spatial structure. Another useful measure in spatial models is the percentage of spatial aggregation explained by the model, which usually is measured by the ratio

$$\frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} \times 100\% \tag{8}$$

Its interpretation is related to obtain the relative contribution given by the spatial aggregation effect. Here, a common estimation of σ_u^2 is the empirical variance s_u^2, which can be obtained from the estimation of $\mathbf{u}$ for each MCMC iteration.

3. Scale mixture of Intrinsically CAR Models

In this section an extension of the usual multivariate Gaussian MRF is proposed, assuming a multivariate SMN distribution. The next definition will provide an extension of (4) to the SMN random field (SMN RF).

Definition 1. A spatial random vector $\mathbf{u} = (u_1, \dots, u_m)'$ follows an SMN RF, if the kernel distribution can be obtained as

$$\pi(\mathbf{u} \mid \sigma_u^2, \mathbf{D}_w, \nu) \propto \int_0^\infty \left(\frac{\psi}{\sigma_u^2}\right)^{m/2} \exp\left\{-\frac{\psi}{2\sigma_u^2}\mathbf{u}'\mathbf{D}_w\mathbf{u}\right\} dF(\psi \mid \nu) \tag{9}$$

where $F(\cdot \mid \nu)$ is the c.d.f. of $\psi \mid \nu$, σ_u^2 is a dispersion parameter, and $\mathbf{D}_w$ denotes a adjacency matrix. A SMN RF with scale parameter σ_u^2 will be denoted as $SMNRF(\mathbf{0}, \sigma_u^{-2}\mathbf{D}_w, \nu)$.

For the Gaussian case, it is known that specification of $\mathbf{D}_w$ in (5) makes (4) improper (Banerjee et al. 2004), since the matrix $\mathbf{D}_w$ is singular, so that its inverse does not exist, hence

$$\int_{\mathbb{R}^m} \pi(\mathbf{u} \mid \sigma_u^2, \mathbf{D}_w, \nu)\mathbf{du} \propto \int_{\mathbb{R}^m} \frac{1}{(\sigma_u^2)^{m/2}} \exp\left\{-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{D}_w\mathbf{u}\right\}\mathbf{du} = \infty$$

The last equation implies that a density function is available, but not integrable. This result is the intrinsic autoregressive model property, and it is usually relegated to the prior distribution elicitation. If additional assumptions are not considered, the improper condition will imply that if a multivariate SMN RF is assumed with kernel (9), then consistent property (Kano 1994) fails. Therefore, integration theory can not be applied.

In the same way as the joint distribution of the Gaussian MRF treated in the spatial literature, for every SMN RF, the joint distribution will also be improper. In fact, this distribution will be proper only if the associated dispersion matrix is definite positive. Hence, some additional restrictions should be imposed to obtain a proper joint distribution, as discussed in Banerjee et al. (2004) and Assunção, Potter & Cavenaghi (2002). The next proposition establishes conditions to make proper the associated SMN RF. The proof of this proposition can be found in the appendix.

Proposition 1. *Suppose that a set of spatial indexed random variables, represented by the vector* $\mathbf{u} = (u_1, \dots, u_m)'$*, is available. Consider the SMN RF in (9) as the distribution of* $\mathbf{u}$*. Additionally, let us suppose that* $F(\cdot \mid \nu)$ *is a known positive c.d.f. If* $\sum_{i=1}^m u_i = 0$ *and* $\mathbb{E}(\psi^{1/2} \mid \nu) < \infty$*, then (9) is proper.*

Specific choices for $F(\cdot \mid \nu)$ in (9) lead to different scale mixture probability distributions. Student-t and Slash MRFs will be used in this work, which can be obtained using stochastic representations which depend on the selected mixing distribution $F(\cdot \mid \nu)$.

The SMN RF can be represented hierarchically in terms of two stages:

At the first stage of the hierarchy, a Gaussian MRF is specified with an additional random scale factor ψ. At the second stage, a mixing distribution for the scale perturbation ψ is then specified. Specifically:

1. For the Student-t MRF:

$$\text{i)} \quad \mathbf{u} \mid \sigma_u^2, \psi, \mathbf{D}_w \sim Normal\left(\mathbf{0}, \sigma_u^{-2}\psi\mathbf{D}_w\right) \tag{10}$$

$$\text{ii)} \quad \psi \mid \nu \sim Gamma(\nu/2, \nu/2) \tag{11}$$

   In this case, the Student-t MRF with ν degrees of freedom follows, which is denoted by $\mathbf{u} \mid \mathbf{D}_w, \sigma_u^2, \nu \sim t(\mathbf{0}, \sigma_u^{-2}\mathbf{D}_w, \nu)$.

2. For the Slash MRF:

$$\text{i)} \quad \mathbf{u} \mid \sigma_u^2, \psi, \mathbf{D}_w \sim Normal\left(\mathbf{0}, \sigma_u^{-2}\psi\mathbf{D}_w\right) \tag{12}$$

$$\text{ii)} \quad \psi \mid \nu \sim Beta(\nu/2, 1) \tag{13}$$

   In this case, the Slash MRF, denoted by $\mathbf{u} \mid \mathbf{D}_w, \sigma_u^2, \nu \sim Slash(\mathbf{0}, \sigma_u^{-2}\mathbf{D}_w, \nu)$, is obtained.

The model described above is useful to implement the MCMC method. It is important to mention that the distribution of both of the above random fields has the finite condition exposed in Proposition 1.

A prior distribution for ν is required in order to assume a valid Bayesian model. Usually, an exponential distribution prior is considered for this parameter, which is assumed independent of (6), that is,

$$\nu \mid \delta_0 \sim \exp(\delta_0), \quad \delta_0 > 0 \tag{14}$$

Assuming (2), (3), (7), (9) and (14), the full joint posterior distribution is specified as,

$$\begin{aligned}\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma^2, \psi, \nu \mid \mathbf{y}, \mathbf{D}_w, \delta_0) \propto & \prod_{i=1}^{m} \exp\{\phi_i^{-1}(y_i\theta_i - g(\theta_i))\} \\ & \times \prod_{i=1}^{m} \exp\{-(1/2\sigma^2)(h(\theta_i) - \mathbf{x}_i'\boldsymbol{\beta} - u_i)^2\}h'(\theta_i) \\ & \times \exp\{-(\psi/2\sigma_u^2)\mathbf{u}'\mathbf{D}_w\mathbf{u}\}\psi^{m/2}(\sigma^2\sigma_u^2)^{-m/2} \\ & \times \exp\{-a/2\sigma_u^2\}(\sigma^2)^{-(b/2+1)} \\ & \qquad \exp\{-c/2\sigma^2\}(\sigma_u^2)^{-(d/2+1)} \\ & \times f(\psi \mid \nu)\exp\{-\delta_0\nu\}\end{aligned} \tag{15}$$

where $f_\psi(\cdot \mid \nu)$ represents the conditional density or probability function of $\psi \mid \nu$. See item 3 of the appendix for the computational aspects.

4. Proposed Bayesian Small-Area Models

An important point is to demonstrate the integrability of the proposed model. Under the generalized linear model (2), link function (3) and prior assumption given by (7) and (14), Theorem 2 from the work developed by Ghosh et al. (1998) gives the conditions to obtain a proper posterior distribution for $\boldsymbol{\theta} \mid \mathbf{y}$ when $P(\psi = 1 \mid \nu) = 1$ (the Gaussian MRF). Following that theorem, it is possible to find a generalization towards the SMN case.

The next proposition gives conditions when the spatial random effect follows an SMN RF. Its proof is given in the appendix.

Proposition 2. *Consider the model (2), link function (3) and prior assumption given by (7) and (14). Consider also the assumptions of Proposition 1 and the following additional conditions:*

i. $\theta_i \in (\underline{\theta}_i, \overline{\theta}_i)$, *for some* $-\infty < \underline{\theta}_i < \overline{\theta}_i < \infty$, $i = 1, \ldots, m$;

ii. $m - p + a - 1 > 0$;

iii. $b > 0$, $d > 0$, $m + c > 0$

If the condition of integrability

$$\int_{\underline{\theta}_i}^{\overline{\theta}_i} \exp\{\phi_i^{-1}(y_i\theta_i - g(\theta_i))\} h'(\theta_i) d\theta_i < \infty$$

is verified for all $i = 1, \ldots, m$, *then posterior distribution* $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ *is proper.*

The main interest is focused in establishing a non-Gaussian parametric spatial random effect. A MCMC structure seems to be adequate to make inferences from this class of model. Most full conditional distributions computed for this scheme are known distributions, therefore, a hybrid Gibbs sampling - metropolis Hastings algorithm is used to generate samples from the joint posterior distribution. The algorithm given in item 3 of the appendix presents the full conditional distributions for this particular model.

5. Applications

The proposed spatial Bayesian models will be applied assuming SMN random effects for two real data in the epidemiological framework to control for excessive smoothness in small areas with sparse data. One dataset is related to IDDM incidence rates in the Chilean municipal districts from Metropolitan region and the other dataset contains female lung, trachea and bronchi cancer standardized mortality ratios in the municipal districts of the country's northern zone. The municipal district is the smallest administrative area in Chile. In this country there are only few published studies related to spatial epidemiology (Andia, Hsing, Andreotti & Ferreccio 2008, Ferreccio, Rollán, Harris, Serrano, Gederlini, Margozzini, Gonzalez, Aguilera, Venegas & Jara 2007, Icaza, Núñez, Torres, Díaz & Varela 2007, Icaza, Núñez, Díaz & Varela 2006, Torres-Avilés, Icaza, Carrasco &

Pérez-Bravo 2010). Results from non-Gaussian spatial Bayesian modeling related to both diseases are presented in the next subsections.

The specific model that is considered for these two applications is the Poisson hierarchical model given by

$$\begin{aligned} y_i \mid e_i, \lambda_i &\overset{ind.}{\sim} Poisson(e_i \lambda_i) \\ log(\lambda_i) \mid \beta_0, u_i, \sigma^2 &\overset{ind.}{\sim} Normal(\beta_0 + u_i, \sigma^2) \end{aligned}$$

$i = 1, \dots, m$, where $\mathbf{y} = (y_1, \dots, y_m)'$ represents the observed sample vector associated to m different regions under study, $\mathbf{e} = (e_1, \dots, e_m)'$ represents the population at risk or the expected population associated to the m different regions, and $\mathbf{u} = (u_1, \dots, u_m)'$ is the vector of random effects which is assumed to have a SMN distribution constrained to sum zero. Diffuse prior distributions are considered for the location and scale parameters, as those presented in (7). For the variance parameters, σ^2 and σ_u^2, the hyperparameters $a = b = c = d = 0.001$ were assumed.

Posterior estimations are obtained from a single run of the Gibbs sampler, with a burn-in of 1000 iterations followed by 10000 further cycles. Convergence have been checked through trace and autocorrelation plots. Three common ways to measure model assessment are taken into account. The first two are oriented to penalize the observed deviance: The deviance information criterion (DIC) (Spiegelhalter, Best, Carlin & Van der Linde 2002) and a modified BIC (Congdon 2003) will be used. A third model choice criterion is applied, proposed by Gelfand & Ghosh (1998), which is based on a predictive check of the model, and measures the discrepancy between the observed data and predicted observations, taking into account quadratic loss measures. As was described in the introduction, the competing models are related to Gaussian, Student-t and Slash MRF. The percentage of spatial variability is computed using expression (8).

5.1. Insulin Dependent Diabetes Mellitus Incidence, Metropolitan Region, Chile

The objective of this study is to describe spatial patterns of type 1 diabetes in children under 15 of age, diagnosed between 2000 and 2005 with residence in the Metropolitan Region of Chile. The Metropolitan Region is located in the centre of Chile. According to the Chilean National Institute of Statistics (INE), this region represents an area of approximately 15,403 km^2. Total population living at Metropolitan Region was 6,061,185 inhabitants, according to the 2002 census. Metropolitan population represents 40% of the whole country. The region is divided into 52 districts, 18 are considered as rural and 34 as highly urbanized, known as Greater Santiago, in the centre of the region, with the 96.9% of the metropolitan population. With respect to the population at risk, children under 15 years of age represent the 24.9% of the metropolitan region population, which is composed by 1,509,218 children. A population-based registry of type 1 diabetes in children under than 15 years of age has been available in the Metropolitan Region since 2000. See Carrasco, Pérez-Bravo, Dorman, Mondragón & Santos (2006) for details about the registry. Torres-Avilés et al. (2010) show an aggregation on incidence rates in urban areas of the Chilean Metropolitan Region, using the Bayesian methodology proposed by Mollié (2000).

Table 1: IDDM model selection criteria, DIC, BIC and predictive check.

Model	DIC	Dbar	pD	BIC	Predictive (G & G)
Gaussian	846.778	534.886	311.892	1151.067	13240.408
Student-t	852.687	537.371	315.315	1160.315	13335.069
Slash	836.498	529.097	307.401	1136.405	13301.901

Model selection criteria results are presented in Table 1. According to previously mentioned goodness of fit criteria, small values imply better adjustment. Therefore, a spatial model that includes Slash random effects with 7 d.f. is a strong candidate to model geographic dependence. This result seems to be adequate due to those extreme values, which match with the higher socioeconomic areas of the region, as is explained in next paragraphs. The predictive measure *G&G* disagrees with the other methods; this can be interpreted as a “failure of the model for prediction”, pointing out a better performance of the usual Gaussian MRF.

Table 2: Posterior mean, standard deviation and 95% HPD credibility intervals for unknown parameters when a Gaussian MRF, Student-t MRF and Slash MRF are assumed.

	Gaussian MRF	Student-t MRF	Slash MRF
β_0	-9.721 (0.004) (-9.844,-9.634)	-9.760 (0.006) (-9.876,-9.631)	-9.752 (0.002) (-9.841,-9.656)
σ^2	0.346 (0.013) (0.162,0.574)	0.291 (0.016) (0.089,0.537)	0.275 (0.014) (0.090,0.507)
σ_u^2	0.230 (0.016) (0.102,0.547)	0.071 (0.001) (0.035,0.117)	0.067 (0.001) (0.032,0.112)
% Spatial Variability	0.441 (0.011) (0.242,0.649)	0.537 (0.0114) (0.332,0.749)	0.546 (0.012) (0.338,0.749)
ν	- -	10.475 (16.482) (3.958,18.277)	7.346 (6.226) (3.038,12.389)

Robust Bayesian models proposed in the previous section were applied to this problem. Inferences over unknown parameters are displayed in Table 2, when Gaussian MRF, Student-t MRF and Slash MRF are assumed to control spatial variability. Similar values are estimated for β_0 and σ^2, under the three MRF models, showing the models' robustness. In contrast, σ_u^2 presents different values, depending on the distribution assumed for the MRF. The non-Gaussian model (Slash MRF) increases the degree of spatial aggregation from 44.1 % to 54.6 %, that is, the excess of spatial variability presented in these data seems mostly due to a clustering effect. Notice that the estimated degrees of freedom are small, which implies that the excess of variability is better captured by one of the SMN RF model.

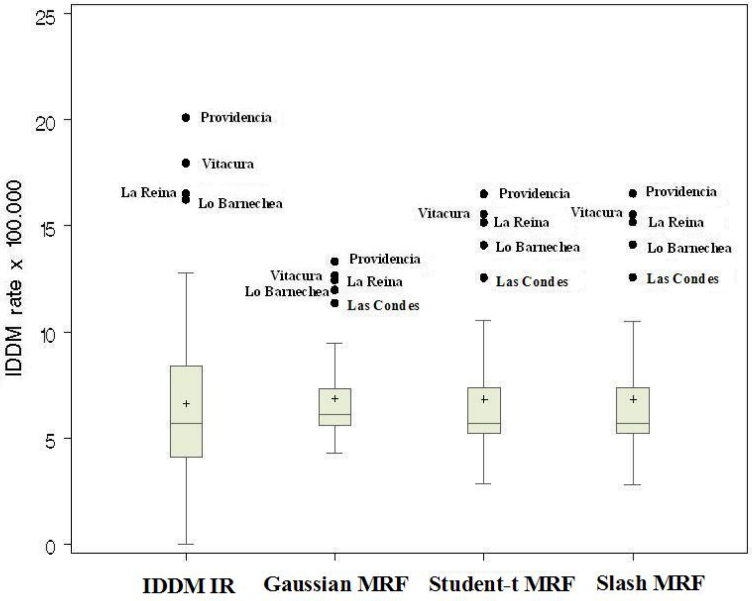


Figure 1: IDDM incidence rate (IR) variability: Raw estimates, Mollié's convolution model (Gaussian MRF), Student-t convolution model (Student-t MRF) and Slash convolution model (Slash MRF).

Figure 1 shows that fully Bayesian estimates of IDDM incidence rates present less variation than raw incidence rate. The three Bayesian variation plots seem to have a similar behavior, due to the presence of several municipal districts with high incidence rates, which are considered as outliers. Comparing the four box-plots, the three fitted models (Gaussian, Student-t and Slash) present and additional municipal district, named Las Condes, as part of the higher incidence group. The normal MRF assumption leads to estimate smoother rates; however, Student-t, and Slash MRF's present slight variability differences. Those differences allow controlling the excess of smoothness, i.e., non-Gaussian shrinkage gives a more adequate estimate of the pattern of underlying risk of disease than that provided by the Mollié's convolution estimates.

From Figure 2, high incidence estimates remain in municipal districts with high socioeconomic level, such as Vitacura and Providencia, located at the northeast side of the map. These results were already found by Torres-Avilés et al. (2010). Slight differences are seen when Slash MRF (d) and Student-t MRF (c) models are assumed, but these differences are clinically important since are in rural municipal districts with zero cases of diabetes located at southwest side of the map.

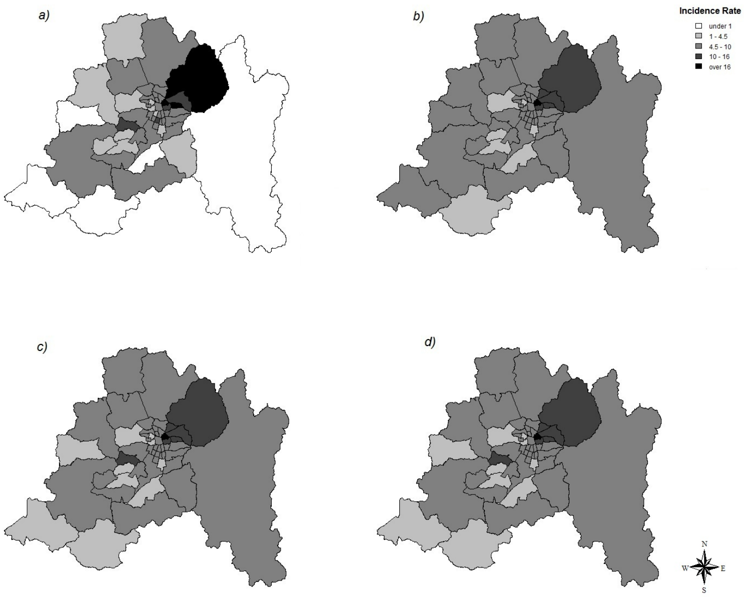


Figure 2: IDDM incidence rate by district: a) Raw incidence rates. b) Mollié's convolution model (Gaussian MRF). c) Student-t convolution model (Student-t MRF). d) Slash convolution model (Slash MRF).

5.2. Female Trachea, Bronchi and Lung Cancer Mortality, Chilean Northern Regions

Bayesian methods that have been applied to several real problems to estimate relative risks of cancer mortality in small-areas can be found in the literature, e.g., Ghosh et al. (1998) and Pascutto et al. (2000), and Mollié (2000). In particular, this application is related to estimate female lung, bronchi and trachea cancer mortality relative risks in the northern regions of Chile. The northern region of Chile represents an area of approximately 300,904 km^2. According to the 2002 census there were 819,177 women inhabitants in this part of the country. The region is divided into 43 districts, many of them (20 or 47%) with less than 10,000 inhabitants. The aim of this study is to describe the geographical distribution of this class of mortality, which has presented smoothness problems in comparison with the usual model.

Mortality statistics for the years 1997-2004 published by the Chilean Ministry of Health were used. The SMR was calculated for 341 districts in the country. Results show an excess of mortality caused by trachea, bronchi and lung cancer in the region. A previous work can be found, where the analysis for both sexes was done for the whole country and published by Icaza et al. (2007). The problem arised when Mollié's model estimates for women cancer mortality risks were too smooth and high in municipal districts where zero cases occurred.

Table 3: Cancer mortality model selection criteria, DIC, BIC and predictive check.

Model	DIC	Dbar	pD	BIC	Predictive (G & G)
Gaussian	4821.381	3064.272	1757.108	8187.896	381675.00
Student-t	4805.212	3058.344	1746.869	8152.110	381671.59
Slash	4792.151	3052.174	1739.977	8125.845	381950.00

Table 4: Posterior mean, standard deviation and 95% HPD credibility intervals for unknown parameters when a Gaussian MRF, Student-t MRF and Slash MRF are assumed.

	Gaussian MRF	Student-t MRF	Slash MRF
β_0	−0.348 (0.001) (−0.409,−0.300)	−0.372 (0.001) (−0.441,−0.313)	−0.391 (0.001) (−0.425,−0.331)
σ^2	0.092 (0.0003) (0.060,0.129)	0.087 (0.0004) (0.054,0.128)	0.085 (0.0003) (0.055,0.128)
σ_u^2	0.197 (0.001) (0.153,0.238)	0.203 (0.001) (0.150,0.253)	0.203 (0.001) (0.153,0.244)
% Spatial Variability	0.770 (0.001) (0.708,0.841)	0.788 (0.001) (0.740,0.848)	0.788 (0.002) (0.715,0.863)
ν	- -	26.406 (116.944) (15.742,53.499)	32.049 (87.516) (15.585,50.462)

For this application, Table 3 shows a better fit for the model that includes the Slash spatial random effect with approximately 32 degrees of freedom, as can be seen in Table 4. Once again, the Slash can not be considered as a predictive alternative, in contrast to a parsimonious model such as the Student-t or the Gaussian MRF. One important result is referred to the 79% estimated proportion of spatial variability associated to this model. Notice that this proportion is almost the same for the three proposed models. This could be related to the estimated degrees of freedom. One important issue is related to the estimation for the other parameters, such as β_0 or baseline risk, which is not affected by the model.

Standardized mortality ratios and Risk estimations are compared in Figure 3. It is important to add that variability estimation is reduced when any of the Bayesian models is considered. All of them show an improvement in contrast to the SMR, and a district called Mejillones is separated from the rest of the distribution, showing the highest risk in the north for this mortality.

Figure 4 displays the cancer mortality relative risk estimation using three different models, with Mollié's convolution model (b), Student-t MRF (c) and Slash MRF (d) as spatial random effects. Models were tested and the best fit was selected among the three different proposed spatial structures.

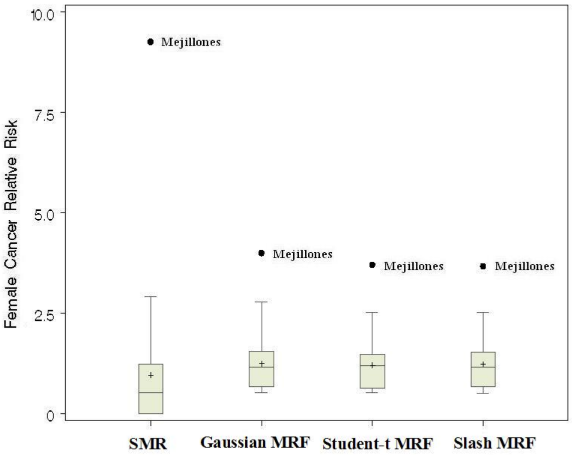


FIGURE 3: Female trachea, bronchi and lung cancer SMR variability: Standardized mortality ratio, Mollié's convolution model (Gaussian MRF), Student-t convolution model (Student-t MRF) and Slash convolution model (Slash MRF).

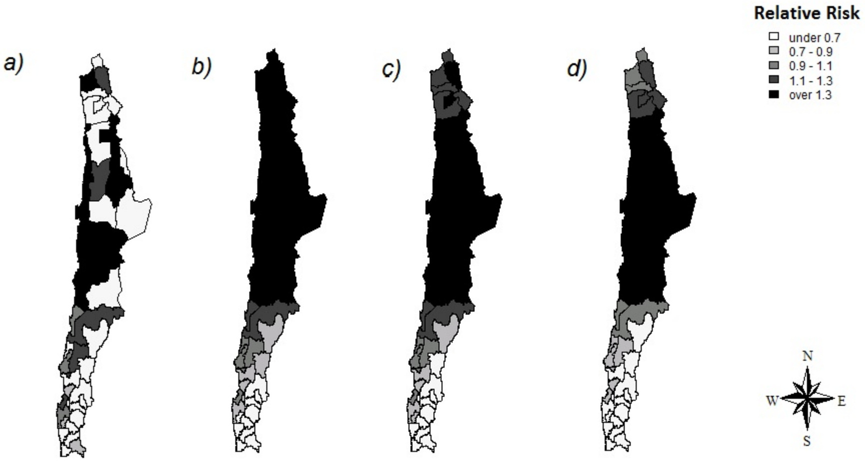


FIGURE 4: Female trachea, bronchi and lung cancer SMR by district: a) Standardized mortality ratio (SMR). b) Mollié's convolution model. c) Student-t convolution model (Student-t MRF). d) Slash convolution model (Slash MRF).

According to the DIC and BIC criteria, the selected Slash MRF model presented better fitted rates, even when Figure 4(d) shows that the first and darkest area in the extreme north, the most populated municipal district (Arica) in that region, presents the highest rates compared to its closer neighbors. It was not possible to reduce the effect produced by the larger areas in the next darkest zones, which correspond to Tarapacá and Antofagasta regions, which are located in the Atacama Desert. The over-smoothing effect lead to flat true variations in risk, even by the selected model.

6. Concluding Remarks

In this work, a non-Gaussian Bayesian-small area estimation is proposed as an alternative to usual parametric models. This approach is particularly useful to obtain estimations of rates or relative risks when subjective geographical dependence is assumed and related results are too smooth for the region under study.

Conditions are required to ensure the propriety of these intrinsic spatial random effect posterior distributions, which must be associated to sum zero constraint and existence of mixing random variable expectations. When spatial correlation structure was available, Proposition 2 provided sufficient conditions to guarantee posterior distribution integrability for Bayesian GLM.

The general methodology is applicable to situations where small area parameters must be estimated. Variability parameters are of interest, since their incorporation in the proposed hierarchical models allowed the computation of the marginal spatial proportion of variability, through the empirical marginal standard deviation function, to quantify excess of variability explained by the spatial effect. This fact has direct relation with the spatial random effect contribution considered for the analysis. As mentioned in Banerjee et al. (2004, p. 166), differences may exist in percentage of variability estimation, when other prior distributions are considered. A prior sensitivity analysis is not studied in this work.

Considering the complex structure of Chilean geography, better results were obtained using our proposed strategy. Both applications were best modeled by Poisson regression with spatial random effects following a joint Slash distribution. It can be seen that β_0 does not produce changes when the three models are fitted to both applications. That is an important consideration that shows the non-Gaussian properties of the Student-t MRF and Slash MRF.

In the future, several topics can be explored in the spatial context. Diagnostic approaches and extensions of model assumptions which include asymmetry in the distribution of the random effects are related topics to be developed. Simulation studies to validate proposed models under different scenarios can also be made.

Bayesian space time models can be proposed, with the subsequent problem of sparseness of data that could affect estimation in municipal districts with low population. Therefore, non-Gaussian models will become more necessary. Temporal trends and geographical patterns are estimated simultaneously, allowing for additional random effects to represent temporal and spatio-temporal interaction variations.

Acknowledgements

The authors acknowledge the helpful and constructive comments of the two anonymous referees, which significantly improved the quality of this paper. Torres-Avilés's research was partially supported by DICYT 040975TA and FONDECYT de Iniciación 11110119 grants. Arellano-Valle's research was partially supported by FONDECYT 1085241-Chile grant. We are very grateful to Ms. Elena Carrasco, Dr. Francisco Pérez-Bravo, the *Fundación de Diabetes Juvenil* and the Department of Statistics of the *Ministerio de Salud de Chile* for providing the data used in this article. We acknowledge the helpful initial discussions with Dr. Pilar Iglesias Z.

[Recibido: septiembre de 2011 — Aceptado: febrero de 2012]

References

Andia, M., Hsing, A. W., Andreotti, G. & Ferreccio, C. (2008), 'Geographic variation of gallbladder cancer mortality and risk factors in Chile: A population-based ecologic study', *International Journal of Cancer* **123**(6), 1411–1416.

Andrews, D. F. & Mallows, C. L. (1974), 'Scale mixture of normal distributions', *Journal of the Royal Statistical Society Series B* **36**(1), 99–102.

Assunção, R. M., Potter, J. E. & Cavenaghi, S. M. (2002), 'A Bayesian space varying parameter model applied to estimating fertility schedules', *Statistics in Medicine* **21**, 2057–2075.

Banerjee, S., Carlin, B. & Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Monographs on Statistics and Applied Probability 101. Chapman and Hall, Boca Ratón, Florida.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society Series B* **36**(2), 192–236.

Besag, J. (1986), 'On the statistical analysis of dirty pictures', *Journal of the Royal Statistical Society Series B* **48**(3), 259–302.

Besag, J., York, J. & Mollié, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

Best, N., Arnold, R., Thomas, A., Waller, L. & Collon, E. (1999), Bayesian models for spatially correlated disease and exposure data, *in* J. Bernardo, A. Smith, A. Dawid & J. Berger, eds, 'Bayesian Statistics 6', Oxford University Press, Oxford, pp. 131–156.

Breslow, N. & Clayton, D. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**, 9–25.

Carrasco, E., Pérez-Bravo, F., Dorman, J., Mondragón, A. & Santos, J. L. (2006), 'Increasing incidence of type 1 diabetes in population from Santiago of Chile: Trends in a period of 18 years (1986-2003)', *Diabetes/Metabolism Research and Reviews* **22**, 34–37.

Clayton, D. & Kaldor, J. (1987), 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics* **43**, 671–681.

Congdon, P. (2003), *Applied Bayesian Modelling*, Wiley & Sons, Chichester.

Damien, P. & Walker, S. (2001), 'Sampling truncated normal, beta and gamma densities', *Journal of Computational and Graphical Statistics* **10**(2), 206–215.

Fang, K. T., Kotz, S. & Ng, K. W. (1990), *Symmetric Multivariate and Related Distributions*, Chapman and Hall, New York.

Ferreccio, C., Rollán, A., Harris, P., Serrano, C., Gederlini, A., Margozzini, P., Gonzalez, C., Aguilera, X., Venegas, A. & Jara, A. (2007), 'Gastric cancer is related to early Helicobacter pylori infection in a high prevalence country', *Cancer Epidemiology, Biomarkers & Prevention* **16**, 662–667.

Gelfand, A. E. & Ghosh, S. K. (1998), 'Model choice: A minimum posterior predictive loss approach', *Biometrika* **85**, 1–11.

Geweke, J. (1993), 'Bayesian treatment of the independent Student-t linear model', *Journal of Applied Econometrics* **8**, 519–540.

Ghosh, M., Natarajan, K., Stroud, T. W. F. & Carlin, B. P. (1998), 'Generalized linear models for small-area estimation', *Journal of the American Statistical Association* **93**(441), 273–282.

Icaza, G., Núñez, L., Díaz, N. & Varela, D. (2006), *Atlas de mortalidad por enfermedades cardiovasculares en Chile, 1997- 2003*, Universidad de Talca y Ministerio de Salud, New York.

Icaza, G., Núñez, L., Torres, F., Díaz, N. & Varela, D. (2007), 'Distribución geográfica de mortalidad por tumores malignos de tráquea, bronquios y pulmón, Chile 1997-2004', *Revista Médica de Chile* **135**(11), 1397–1405.

Kano, Y. (1994), 'Consistency property of elliptical probability density functions', *Journal of Multivariate Analysis* **51**, 139–147.

Lange, K. & Sinsheimer, J. S. (1993), 'Normal/independent distributions and their applications in robust regression', *Journal of Computational and Graphical Statistics* **2**(2), 175–198.

Lyu, S. & Simoncelli, E. P. (2007), Statistical modeling of images with fields of Gaussian scale mixtures, *in* B. Schölkopf, J. Platt & T. Hoffman, eds, 'Advances in Neural Information Processing Systems, 19', MIT Press, Cambridge, pp. 945–952.

Mollié, A. (2000), Bayesian mapping of Hodgkin's disease in France, *in* P. Elliott, J. Wakefield, N. G. Best & D. J. Briggs, eds, 'Spatial Epidemiology: Methods and Applications', Oxford University Press, New York, pp. 267–285.

Parent, O. & Lesage, J. P. (2008), 'Using the variance structure of the conditional autoregressive specification to model knowledge spillovers', *Journal of Applied Economics* **23**, 235–256.

Pascutto, C., Wakefield, J. C., Best, N. G., Richardson, S., Bernardinelli, L., Staines, A. & Elliott, P. (2000), 'Statistical issues in the analysis of disease mapping data', *Statistics in Medicine* **19**(17-18), 2493–519.

Roislien, J. & Omre, O. (2006), 'T-distributed random fields: A parametric model for heavy-tailed well-log data', *Mathematical Geology* **38**(7), 821–849.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society, Series B* **64**, 583–639.

Torres-Avilés, F., Icaza, G., Carrasco, E. & Pérez-Bravo, F. (2010), 'Clustering of cases of type 1 diabetes in high socioeconomic communes in Santiago de Chile: Spatio-temporal and geographical analysis.', *Acta Diabetologica* **47**(3), 251–257.

Zellner, A. (1976), 'Bayesian and non-Bayesian analysis of the regression model with multivariate student-t error terms', *Journal of the American Statistical Association* **71**(354), 400–405.

Appendix

1. **Proof of Proposition 1.** As was showed by Assunção et al. (2002), the $\sum_{i=1}^{m} u_i = 0$ constraint makes the Gaussian kernel (4) proper; i.e., on the set $C = \{\mathbf{u} \in \mathbb{R}^m : \sum_{i=1}^{m} u_i = 0\}$, we have

$$\int_C \frac{1}{(\sigma_u^2)^{m/2}} \exp\left\{-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{D}_w\mathbf{u}\right\}\mathbf{du} < \infty$$

Hence, under the $\sum_{i=1}^{m} u_i = 0$ constraint, by applying the Fubini's theorem and the change variable $\mathbf{y} = \psi^{1/2}\mathbf{x}$, we have in (9) that

$$\begin{aligned}\int_C \pi(\mathbf{u} \mid \sigma_u^2, \mathbf{D}_w, \nu) d\mathbf{u} &= \int_0^\infty \psi^{m/2} \int_C \frac{1}{(\sigma_u^2)^{m/2}} \exp\left\{-\frac{\psi}{2\sigma_u^2}\mathbf{u}'\mathbf{D}_w\mathbf{u}\right\} d\mathbf{u} dF(\psi\nu) \\ &\propto \int_0^\infty \psi^{1/2} dF(\psi \mid \nu) < \infty \qquad ⋇\end{aligned}$$

2. **Proof of Proposition 2.** From (2), (3), (7), (9) and (14) we have for the full joint posterior distribution that

$$
\begin{aligned}
\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma^2, \psi, \nu \mid \mathbf{y}, \mathbf{D}_w, \delta_0) \propto & \prod_{i=1}^{m} \exp\{\phi_i^{-1}(y_i\theta_i - g(\theta_i))\} \\
& \times \prod_{i=1}^{m} \exp\{-(1/2\sigma^2)(h(\theta_i) - \mathbf{x}_i'\boldsymbol{\beta} - u_i)^2\} h'(\theta_i) \\
& \times \exp\{-(\psi/2\sigma_u^2)\mathbf{u}'\mathbf{D}_w\mathbf{u}\}\psi^{m/2}(\sigma^2\sigma_u^2)^{-m/2}, \\
& \times \exp\{-a/2\sigma_u^2\}(\sigma^2)^{-(b/2+1)} \\
& \quad \exp\{-c/2\sigma^2\}(\sigma_u^2)^{-(d/2+1)} \\
& \times f(\psi \mid \nu)\exp\{-\delta_0\nu\}
\end{aligned}
$$

where $f(\cdot \mid \nu)$ is the conditional density (or probability) function of $\psi \mid \nu$. Integrating with respect to $\boldsymbol{\beta}$, σ^2 and σ_u^2, we obtain

$$
\begin{aligned}
\pi(\boldsymbol{\theta}, \mathbf{u}, \psi, \nu \mid \mathbf{y}, \mathbf{D}_w, \delta_0) \propto & \prod_{i=1}^{m} \exp\{\phi_i^{-1}(y_i\theta_i - g(\theta_i))\} h'(\theta_i) \\
& \times \psi^{m/2}(a + \psi\mathbf{u}'\mathbf{D}_w\mathbf{u})^{-(m+b-1)/2} \\
& \times f(\psi \mid \nu)\exp\{-\delta_0\nu\}
\end{aligned}
$$

Notice that this last result has a multivariate Student-t kernel. Now, integrating over $\mathbf{u} \in \mathbb{R}^m$ under the constraint $\sum_{i=1}^m u_i = 0$, the following result is obtained,

$$
\begin{aligned}
\pi(\boldsymbol{\theta}, \psi, \nu \mid \mathbf{y}, \mathbf{D}_w, \delta_0) \leq K & \prod_{i=1}^{m} \exp\{\phi_i^{-1}(y_i\theta_i - g(\theta_i))\} h'(\theta_i) \\
& \times f(\psi \mid \nu)\exp\{-\delta_0\nu\}
\end{aligned}
$$

where K is a constant that does not depend on $\boldsymbol{\theta}$ or any of the parameters previously integrated. Finally, integration over $\boldsymbol{\psi}$ and then over ν leads to the desire result.⋇

3. **Proposed MCMC Algorithm.** To implement the Gibbs sampling, the full conditional distributions associated with the full joint posterior distribution (15) are given in the following, in which $\mathbf{h}(\boldsymbol{\theta}) = (h(\theta_1), \dots, h(\theta_m))'$ denotes the link vector and $\mathbf{X}$ is the $m \times p$ design matrix which has rows $\mathbf{x}_1, \dots, \mathbf{x}_m$.

   a) $\boldsymbol{\beta} \mid \mathbf{X}, \sigma^2, \mathbf{u} \sim Normal(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, where

$$
\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{h}(\boldsymbol{\theta}) - \mathbf{u})
$$

b) $\mathbf{u} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2, \sigma_u^2, \psi, \mathbf{X}, \mathbf{D}_w \sim Normal(\boldsymbol{\mu}_u, \mathbf{V}_u)$, where

$$\boldsymbol{\mu}_u = \frac{1}{\sigma^2}\mathbf{V}_u\left(\mathbf{h}(\boldsymbol{\theta}) - \mathbf{X}\boldsymbol{\beta}\right), \mathbf{V}_u = \left(\frac{1}{\sigma^2}\mathbf{I}_m + \frac{\psi}{\sigma_u^2}\mathbf{D}_w\right)^{-1}$$

and $\mathbf{I}_m$ is the identity matrix of size m

c) $\sigma^{-2} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{X}, \mathbf{u}, c, d \sim Gamma(a^*, b^*)$, where

$$a^* = \frac{1}{2}\left[m + a\right] \text{ and } b^* = \frac{1}{2}\left[(\mathbf{h}(\boldsymbol{\theta}) - \mathbf{X}'\boldsymbol{\beta} - \mathbf{u})'(\mathbf{h}(\boldsymbol{\theta}) - \mathbf{X}'\boldsymbol{\beta} - \mathbf{u}) + b\right]$$

d) $\sigma_u^{-2} \mid \mathbf{u}, \psi, \mathbf{D}_w, c, d \sim Gamma(c^*, d^*)$ where,

$$c^* = \frac{m + c}{2} \text{ and } d^* = \frac{1}{2}\left(\psi(\mathbf{u}'\mathbf{D}_w\mathbf{u}) + d\right)$$

e) Choice of a distribution for the scale random factor ψ:

i. If $\psi \mid \nu \sim Gamma(\nu/2, \nu/2)$, then

$$\psi \mid \mathbf{u}, \sigma_u^2, \mathbf{D}_w, \nu \sim Gamma\left(\frac{1}{2}(\nu + m), \frac{1}{2\sigma_u^2}(\mathbf{u}'\mathbf{D}_w\mathbf{u}) + \nu\right)$$

ii. If $\psi \mid \nu \sim Beta(\nu/2, 1)$, then

$$\psi \mid \mathbf{u}, \sigma_u^2, \mathbf{D}_w, \nu \sim Gamma\left(\frac{1}{2}(\nu + m), \frac{1}{2\sigma_u^2}(\mathbf{u}'\mathbf{D}_w\mathbf{u})\right)\mathbf{1}_{(0,1)}(\psi)$$

where $\mathbf{1}_A$ represents the indicator function. Notice the presence of a truncated Gamma distribution in the $[0, 1]$ interval. To draw from this distribution, the Damien & Walker (2001) algorithm can be performed.

f) Degrees of freedom are estimated from

fa. If $\psi \mid \nu \sim Gamma(\nu/2, \nu/2)$, then

$$\pi(\nu \mid \psi, \delta_0) \propto \Gamma(\nu/2)^{-1}\nu^{\nu/2}\exp\{-\nu(\delta_0 + 0.5(\psi - ln(\psi)))\}$$

fb. If $\psi \mid \nu \sim Beta(\nu/2, 1)$, then

$$\nu \mid \psi, \delta_0 \sim Gamma(2, \delta_0 - ln(\psi/2))\mathbf{1}_{(0,1)}(\psi)$$

g) $\pi(\theta_i \mid \mathbf{y}, \boldsymbol{\beta}, \mathbf{X}, \sigma^2, \mathbf{u}) \propto h'(\theta_i)\exp\{\phi_i^{-1}(y_i\theta_i + g(\theta_i) - \frac{1}{2}(h(\theta_i) - \mathbf{x}_i'\boldsymbol{\beta} - u_i)^2)\}$

The algorithm must be iterated until convergence is detected in order to start to take a sample.

Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 205 a 222

Métodos de integración de odds ratio basados en meta-análisis usando modelos de efectos fijos y aleatorios útiles en salud pública

Integration Methods of Odds Ratio Based on Meta-Analysis Using Fixed and Random Effect Models Useful in Public Health

Mónica Catalán[1,a], M. Purificación Galindo[2,b], Javier Martín[2,c], Víctor Leiva[1,d]

[1]Departamento de Estadística, Universidad de Valparaíso, Valparaíso, Chile
[2]Departamento de Estadística, Universidad de Salamanca, Salamanca, España

Resumen

Un meta-análisis integra información proveniente de varios estudios con el propósito de generar un resultado común para un problema determinado. En la literatura nos encontramos con varios métodos de integración de resultados, siendo el más básico el método de integración de niveles de probabilidad y, con una complejidad mayor, el método de integración del tamaño del efecto. Este último hace uso de modelos de efectos fijos y aleatorios. En este estudio, comparamos los resultados de dos métodos de estimación del tamaño del efecto basados en un meta-análisis usando modelos de efectos fijos y aleatorios. La medida del tamaño del efecto considerada en este estudio es el odds ratio, debido a que esta medida es usada frecuentemente en revisiones sistemáticas de varios temas de interés en salud pública, tales como cáncer cérvico uterino, colecistectomía laparoscópica, enfermedades cardiovasculares, enfermedad de Parkinson y tabaquismo. Las conclusiones de este trabajo indican las condiciones de aplicabilidad de los estimadores analizados del odds ratio en función de la magnitud del efecto poblacional, de la variabilidad entre estudios, del tamaño del meta-análisis y de los tamaños muestrales de tales estudios.

***Palabras clave*:** bioestadística, ensayos clínicos, medicina, tamaño del efecto.

[a]Profesora auxiliar. E-mail: monica.catalan@uv.cl
[b]Profesora titular. E-mail: pgalindo@usal.es
[c]Profesor titular. E-mail: jmv@usal.es
[d]Profesor titular. E-mail: victor.leiva@uv.cl

Abstract

Meta-analysis integrates information from different studies to generate a common response to a determined problem. In the literature, we find several integration methods of results, with the integration method of levels of probability being the more basic and, with a greater complexity, the integration method of the effect size, which uses fixed and random effect models. In this study, we compare the results of two estimation methods of the effect size based on meta-analysis using fixed and random effect models. The measure of the effect size considered here is the odds ratio, due to this measure is frequently used in systematic reviews of several topics of interest in public health, such as heart diseases, laparoscopic colectomy, Parkinson disease, tobacco addiction and uterine cervical cancer. Conclusions of this work indicate the applicability conditions of the analyzed estimators of the odds ratio in function of the size of the population effect, of the variability among studies, of the size of the meta-analysis and of the sample sizes of such studies.

***Key words*:** Biostatistics, Clinical trials, Effect size, Medicine.

1. Introducción

Un meta-análisis integra resultados de varios estudios con el fin de generar una respuesta común frente a un problema de investigación determinado (Glass 1976, Martín, Donaldson, Villarroel, Parmar, Ernst & Higginson 2002, Catalán & Galindo 2003, Burguillo, Martín, Barrera & Bardsley 2010). En general, una revisión sistemática es actualmente reconocida como la búsqueda organizada de literatura de un tema específico, mientras que un meta-análisis estudia de manera estadística esa información que ha sido organizada previamente (Glass 1976, Rosenthal 1984, Vamvakas 2011). La integración de niveles de probabilidad fue uno de los primeros métodos estadísticos usados para sintetizar cuantitativamente los resultados de un conjunto de estudios. Para ese fin, se han desarrollado diversos métodos con la limitación que éstos sólo permiten determinar si se rechaza o no una hipótesis nula, sin indicar cuál es el tamaño del efecto o el grado de influencia de cada estudio en el resultado que se genera (Rosenthal 1984). Los métodos de integración, cuando el tamaño del efecto es el odds ratio (OR), el riesgo relativo o la diferencia de riesgo, presentan un gran desarrollo en la literatura científica, dado que proporcionan mayor información sobre la magnitud del efecto, permitiendo inferir un resultado desde los obtenidos de un conjunto de estudios (Rosenthal 1984, Hedges & Olkin 1985).

El OR es una de las medidas del tamaño del efecto más comúnmente utilizada en ensayos clínicos aleatorizados, donde la variable de respuesta dicotómica se registra para dos conjuntos de sujetos, usualmente llamados grupo tratado y grupo control. Además, el OR se utiliza en otros estudios de interés clínico como las asociaciones con factores de riesgo o en las pruebas de diagnóstico. Sin embargo, para la aplicación de los modelos principales para la integración del tamaño del efecto en meta-análisis, es más apropiado trabajar con el logaritmo del OR estimado, dado que éste cumple con mayor facilidad el supuesto de normalidad (Turner, Omar, Yang, Goldstein & Thompson 2000, Leyland &

Goldstein 2001, Catalán & Galindo 2003). Antecedentes actualizados siguen evidenciando que el OR es una medida del tamaño del efecto presente en revisiones sistemáticas por meta-análisis en varios temas de interés en salud pública, tales como cáncer cérvico uterino (Rydzewska, Tierney, Vale & Symonds 2010), colecistectomía laparoscópica (Claros, Manterota, Vial & Sanhueza 2007, Zhou, Zhang, Wang & Hu 2009), enfermedades cardiovasculares (Moores, Jackson, Shorr & Jackson 2004, Cornelissen 2007, Dentali, Douketis, Lim & Crowther 2007), enfermedad de Parkinson (Allam, Del Castillo & Navajas 2003, Stowe, Ives, Clarke & van Hilten 2008) y tabaquismo (Jiménez-Ruiz, Riesco, Ramos & Barrueco 2008).

En un meta-análisis, los métodos de integración del tamaño del efecto se han analizado desde dos perspectivas, y éstas son:

(i) Considerando un modelo de efectos fijos (M1): en este caso, la hipótesis de partida es la existencia de un único tamaño del efecto poblacional y sólo se considera la variabilidad debido al muestreo, o

(ii) Considerando un modelo de efectos aleatorios (M2): en este otro caso, se parte de una megapoblación de tamaños del efecto y, por tanto, se contempla una nueva variabilidad debido a la diferencia entre estudios. Cada estudio estima un tamaño del efecto de esa población.

No obstante, de acuerdo a lo que se describe en la literatura, la elección entre estos dos modelos (M1 y M2) es un tema de extensa discusión para los investigadores meta-analíticos (Hedges & Vevea 1998, Berlin, Laird, Sacks & Chalmers 1989). Por una parte, el modelo de efectos fijos asume homogeneidad de los parámetros correspondientes a los efectos de los estudios, de modo que el tamaño del efecto es una constante fija desconocida que debe ser estimada. Por otra parte, el modelo de efectos aleatorios supone heterogeneidad de los parámetros correspondientes a los efectos de los estudios, y así cada estudio representa una población. Por consiguiente, este último tipo de modelos (M2) permite descomponer la varianza de los resultados de los estudios en una parte que corresponde a la variación muestral y otra que refleja las diferencias reales entre estos estudios.

Existen varios métodos que se pueden usar para estimar los parámetros en cada tipo de modelo (M1 y M2), haciendo que la decisión de utilizar uno u otro método en el desarrollo de un meta-análisis resulte más compleja. Dos de los estimadores más utilizados en los méta-análisis para la integración del OR bajo el modelo M1 en el campo clínico son el estimador clásico de media ponderada (conocido como DerSimonian-Laird y que llamaremos “clásico”) y el de Peto (que llamaremos “peto”) (Petitti 1994). Bajo el modelo M2, el estimador más utilizado es el de media ponderada, que incluye la estimación de la variabilidad entre estudios propuesta por DerSimonian & Laird (1986). Otro estimador muy utilizado en estudios clínicos es el de Mantel-Haenzel, pero éste tiene problemas en la estimación de su variabilidad. Existen también otros estimadores propuestos en la bibliografía sobre el tema que el lector interesado puede revisar en (Greenland & Salvan 1990). En consecuencia, para elegir el tipo de modelo y el método de estimación que se debe usar en un estudio de meta-análisis, es necesario considerar las características de los estudios que intervienen en el meta-análisis y el problema para el que se pretende obtener un resultado común. Esto quiere decir que, por una parte, si bien los estudios que se integran tratan un problema similar, éstos pueden presentar varias características relacionadas, por ejemplo, al número y al tipo de pacientes en cada uno de ellos, a las diferencias en su diseño y al lugar donde estos estudios se realizan. Por otra parte, los resultados que proporciona un método de integración basados en meta-análisis podrían depender del tamaño del efecto que se pretende estimar, es decir, de un efecto del tratamiento mayor o menor, de la varianza entre los estudios, del número de estudios involucrados en el meta-análisis y del número de individuos considerados.

Frente a las alternativas de elección entre modelos de efectos fijos y de efectos aleatorios para la estimación del tamaño del efecto mediante meta-análisis (Hedges & Vevea 1998), el presente estudio responde a la pregunta de investigación acerca de qué diferencias existen entre los métodos de estimación clásico y peto en los modelos M1 y M2 cuando el tamaño del efecto es el OR. Esto permite valorar el impacto de las distintas condiciones en las diferencias obtenidas por un modelo o por otro. En consecuencia, la hipótesis de investigación es que existen diferencias entre los resultados que proporcionan los métodos de integración del tamaño del efecto en meta-análisis.

El objetivo principal de este artículo es comparar los resultados que proporcionan dos métodos de estimación del tamaño del efecto (clásico y peto) bajo los dos modelos considerados habitualmente en meta-análisis (M1 y M2). Específicamente, se pretende conocer el comportamiento de estos estimadores en función de la magnitud del efecto poblacional, de la variabilidad entre estudios, del tamaño del meta-análisis y de los tamaños muestrales de los estudios. En el caso del modelo de efectos fijos M1, se consideran el estimador clásico y el estimador peto, métodos que llamaremos ef-clásico y ef-peto, respectivamente. En el caso del modelo de efectos fijos M2, se utilizan los estimadores clásico y peto incluyendo además el estimador de la variabilidad entre estudios propuesto por Dersimonian-Laird, métodos que llamaremos ea-clásico y ea-peto, respectivamente.

El resto de este artículo está organizado como sigue. En la sección 2 describimos los materiales y métodos de este estudio. En la sección 3 presentamos los resultados del estudio. En la sección 4 discutimos los resultados obtenidos en la sección 3. En la sección 5 bosquejamos las conclusiones de este trabajo.

2. Materiales y métodos

En esta sección, proporcionamos los materiales y métodos de este estudio que incluyen la definición de las unidades en estudio y las variables a considerar, la generación del meta-análisis y los métodos de integración del tamaño del efecto.

En este trabajo se diseñó un estudio de simulación donde se generaron los datos necesarios para un conjunto de 81 meta-análisis a los que se les aplicaron los dos métodos para estimar los parámetros de los modelos M1 y M2 para integración del tamaño del efecto. Específicamente, en este artículo (i) describimos los resultados generados por los dos métodos de estimación para el conjunto de meta-análisis y (ii) determinamos si existen diferencias entre los métodos de estimación y entre los modelos, en relación al valor estimado del tamaño del efecto, permitiendo valorar el impacto de las distintas condiciones en las diferencias obtenidas por un modelo o por otro. Recalcamos que la hipótesis de investigación es que existen diferencias entre los resultados que proporcionan los métodos de integración del tamaño del efecto en meta-análisis.

2.1. Unidades de análisis y variables

Las unidades en estudio son los meta-análisis. En cada uno de ellos se consideran los modelos M1 y M2 basados en el supuesto distribucional de normalidad para la integración del tamaño del efecto (OR estimado) mediante la aplicación de los métodos ef-clásico, ef-peto, ea-clásico y ea-peto. Las variables en estudio son los resultados para el OR estimado, su logaritmo ($\log(\widehat{\mathrm{OR}})$) y la varianza de $\log(\widehat{\mathrm{OR}})$ estimada.

2.2. Meta-análisis

Debido a que el objetivo principal de este artículo es comparar los resultados que proporcionan los métodos de estimación clásico y peto del tamaño del efecto bajo los modelos M1 y M2 usando meta-análisis, entonces necesitamos un número importante de meta-análisis. Este número nos permite obtener los datos a nivel de cada estudio considerado en cada meta-análisis y los datos específicos de los individuos. Sin embargo, obtener una cantidad grande de meta-análisis basados en estudios reales para lograr el objetivo de este estudio está fuera de nuestro alcance, por lo que se optó por la alternativa de generar los datos individuales para los estudios que intervienen en cada meta-análisis a través de un proceso de simulación. Para esto, se estableció que la variable respuesta dentro de cada estudio corresponde a la presencia o a la ausencia de una enfermedad bajo un factor de exposición o de riesgo como lo es un tratamiento o un control. La medida del efecto considerada aquí en cada estudio es el OR correspondiente al odds ratio del grupo tratado en relación al grupo control y dado por

$$\mathrm{OR} = \frac{p_t/(1-p_t)}{p_c/(1-p_c)}$$

donde p_t y p_c son las probabilidades de presencia de la enfermedad en los grupos tratado y control, respectivamente. De esta manera, $\log(\mathrm{OR}) = \mathrm{logit}(p_t) - \mathrm{logit}(p_c)$, donde $\mathrm{logit}(p_t) = \log(p_t/(1-p_t))$ y $\mathrm{logit}(p_c) = \log(p_c/(1-p_c))$ son las funciones logito correspondientes.

Para hacer inferencias estadísticas para el parámetro OR, es necesario disponer de la distribución del estimador del OR. Sin embargo, ya que el OR está acotado inferiormente en cero, puesto que por definición éste no puede tomar valores negativos, y el OR no tiene una cota superior, su estimador suele seguir una distribución asimétrica que impide asumir una distribución normal. Entonces, para evitar este problema, se suele trabajar con el logaritmo del OR y así suponer una distribución normal para el logaritmo natural del estimador del OR, $\widehat{\mathrm{OR}}$, esto es,

$$\log(\widehat{\mathrm{OR}}) \sim \mathrm{N}(\mathrm{E}[\log(\widehat{\mathrm{OR}})], \mathrm{Var}[\log(\widehat{\mathrm{OR}})])$$

donde $\log(\widehat{\mathrm{OR}}) = \mathrm{logit}(\widehat{p}_t) - \mathrm{logit}(\widehat{p}_c)$, con $\mathrm{logit}(\widehat{p}_t) \sim \mathrm{N}(\theta_t, \sigma_t^2)$ y $\mathrm{logit}(\widehat{p}_c) \sim \mathrm{N}(\theta_c, \sigma_c^2)$. Aquí,

(i) $\mathrm{E}[\log(\widehat{\mathrm{OR}})] = \theta$ es el valor esperado del estimador del logaritmo natural del OR o tamaño del efecto y

(ii) $\mathrm{Var}[\log(\widehat{\mathrm{OR}})] = \tau^2 = \sigma_t^2 + \sigma_c^2$ es la varianza verdadera entre estudios, donde σ_t^2 y σ_c^2 son las varianzas de los grupos tratado y control, respectivamente.

Entonces, para llevar a cabo el proceso de simulación de los datos, se consideran los parámetros siguientes:

(i) Tamaño del efecto (θ);

(ii) Varianza poblacional entre estudios (τ^2);

(iii) Número de estudios del meta-análisis (J);

(iv) Número de individuos dentro de cada estudio (n);

(v) Número de individuos dentro de cada estudio del grupo tratado (n_t) y

(vi) Número de individuos dentro de cada estudio del grupo control (n_c).

Para asignar el número de individuos en cada estudio, se definió un indicador de la proporción de estudios en un meta-análisis (p) con un número de individuos determinado (n). Para cada uno de los parámetros de simulación (θ, τ^2, J, n) y el indicador (p), se seleccionaron tres escenarios distintos con valores considerados como “bajo”, “moderado” y “alto”, basándonos en los valores que utilizan algunos meta-análisis descritos en la literatura (Turner et al. 2000, Coomarasamy, Papaioannou, Gee & Khan 2001). Usando también como referencia estos estudios previos y de acuerdo al número de parámetros establecidos y a los valores de cada uno de ellos, se generaron datos para un total de 81 meta-análisis, que reúnen 1.755 estudios y 739.080 individuos. Estos 739.080 individuos corresponden a la suma de todos los individuos de todos los estudios en todos los meta-análisis. Estos datos se generaron basados en (i) tres valores de tamaño del efecto poblacional (indicado como log(OR): $-0,106$, $-0,714$, $-1,599$), (ii) tres valores de la varianza poblacional (0,015; 0,15 y 0,8), (iii) tres cantidades de estudios dentro de cada meta-análisis (10, 20 y 35), (iv) tres cantidades de individuos dentro de cada estudio (20 para el grupo tratado y 20 para el grupo control, 150 para el grupo tratado y 150 para el grupo control y 500 para el grupo tratado y 500 para el grupo control) y (v) la proporción de estudios con un tamaño específico de individuos dentro de cada meta-análisis. Esto se explica porque generalmente los estudios que forman parte de un meta-análisis tienen un número distinto de individuos. En este trabajo se establecieron los porcentajes siguientes de estudios dentro de un meta-análisis con un número distinto de individuos: (a) 30 % de estudios con 40 individuos en total, 60 % de estudios con 300 individuos en total y 10 % de estudios con 1.000 individuos en total; (b) 10 % de estudios con 40 individuos en total; 70 % de estudios con 300 individuos en total y 20 % de estudios con 1.000 individuos en total; y (c) 10 % de estudios con 40 individuos en total; 50 % de estudios con 300 individuos en total y 40 % de estudios con 1.000 individuos en total. Los 1.755 estudios corresponden a la suma de todos los estudios generados según lo establecido anteriormente. Los 81 meta-análisis resultan a partir de multiplicar 3 tamaños del efecto, 3 valores para las varianzas, 3 cantidades de estudios y 3 proporciones de estudios dentro de cada meta-análisis ($3 \times 3 \times 3 \times 3 = 81$).

De esta manera, sobre la base del supuesto de normalidad y dados los valores de los parámetros de simulación (θ, τ^2, J, n) y el indicador (p) dados en la tabla 1, se obtienen los valores de la media y la varianza del $\text{logit}(\widehat{p}_t)$ para el grupo tratado y del $\text{logit}(\widehat{p}_c)$ para el grupo control. La generación de datos para cada estudio del meta-análisis se realiza usando el algoritmo siguiente de cuatro pasos:

Paso 1. Generar J observaciones de $\text{logit}(\widehat{p}_t)$ y $\text{logit}(\widehat{p}_t)$ desde una distribución normal con media θ y varianza τ^2 establecidas.

Paso 2. Calcular las probabilidades de tener la enfermedad en los grupos tratado y control, p_t y p_c, respectivamente, para cada una de las J observaciones generadas en el Paso 1.

Paso 3. Obtener los datos individuales en los grupos tratado y control para cada uno de los J estudios dentro del meta-análisis desde distribuciones binomiales con parámetros n y p_t, y n y p_c, donde, como se mencionó, n es el número de individuos y p_t y p_c son las probabilidades de presentar la enfermedad en los grupos tratado y control, respectivamente. Para cada uno de los niveles de los parámetros establecidos (bajo, moderado, alto), como se mencionó, se utilizaron 3 opciones para la proporción de estudios en un meta-análisis con un número determinado de individuos. Estas opciones son (ver últimas tres filas de la tabla 1):

(i) 30 % de los estudios con $n = 40$, 60 % con $n = 300$ y 10 % con $n = 1000$;

(ii) 10 % de los estudios con $n = 40$, 70 % con $n = 300$ y 20 % con $n = 1000$;

(iii) 10 % de los estudios con $n = 40$; 50 % con $n = 300$ y 40 % con $n = 1000$.

Paso 4. Resumir las respuestas individuales de los grupos tratado y control en cada uno de los J estudios dentro de un meta-análisis en una tabla de contingencia 2×2, cuyas variables dicotómicas son la enfermedad (presencia/ausencia) y el factor de exposición (tratamiento/control).

El proceso de generación de datos basado en el algoritmo anterior se debe ejecutar para cada uno de los 81 meta-análisis en estudio usando los valores dados en la tabla 1. Una vez generados los datos individuales que corresponden a la

Tabla 1: Escenario del estudio de simulación.

	Valores establecidos		
Parámetro	Bajo	Moderado	Alto
θ	$-0,106$	$-0,714$	$-1,599$
τ^2	0,015	0,15	0,8
J	10	20	35
n	40	300	1000
n_t	20	150	500
n_c	20	150	500
p_t	0,10	0,16	0,06
p_c	0,11	0,28	0,24
p	0,3	0,6	0,1
	0,1	0,7	0,2
	0,1	0,5	0,4

respuesta de los individuos (739.080 en total) en cada estudio (1.755 en total), se deben estimar el OR, su logaritmo (log(OR)) y la varianza del estimador del logaritmo del OR para cada estudio dentro de los 81 meta-análisis.

2.3. Métodos de integración del tamaño del efecto

Considere el modelo de efectos fijos (M1)

$$Y_j = \theta + \varepsilon_j, \quad \varepsilon_j \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

y el modelo de efectos aleatorios (M2)

$$Y_j = \mu_j + \varepsilon_j, \quad \mu_j = \theta + u_j, \quad u_j \sim \mathrm{N}(0, \tau^2), \quad \varepsilon_j \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

donde

(i) Y_j es la variable respuesta en el estudio j-ésimo;

(ii) θ es el tamaño del efecto;

(iii) ε_j es el error aleatorio;

(iv) σ_ε^2 es el varianza del error aleatorio;

(v) μ_j es el tamaño del efecto en el estudio j-ésimo;

(vi) u_j es el error en el estudio j-ésimo y

(vii) τ^2 es la varianza entre estudios.

Para el modelo M1, el estimador de θ y su error estándar están dados por

$$\widehat{\theta} = \frac{\sum_{j=1}^J w_j\, Y_j}{\sum_{j=1}^J w_j} \quad \text{y} \quad \sigma_{\widehat{\theta}} = \frac{1}{\sqrt{\sum_{j=1}^J w_j}}$$

donde $w_j = 1/\text{Var}[Y_j]$, con $\text{Var}[Y_j] = \sigma_\varepsilon^2$ conocida. En este modelo, se consideran los métodos de estimación ef-clásico (Petitti 1994) y ef-peto (Yusuf, Peto, Lewis, Collins & Sleight 1985) para la integración del OR. Para el modelo M2, el estimador de θ y su error estándar están dados por

$$\widehat{\theta} = \frac{\sum_{j=1}^{J} w_j^* Y_j}{\sum_{j=1}^{J} w_j^*} \quad \text{y} \quad \sigma_{\widehat{\theta}} = \frac{1}{\sqrt{\sum_{j=1}^{J} w_j^*}}$$

donde $w_j^* = 1/(\sigma_\varepsilon^2 + \widehat{\tau^2})$. En este modelo, se consideran los métodos de estimación ea-clásico (DerSimonian & Laird 1986) y ea-peto (Martín 1995) para la integración del OR. Así, en general, para M1 y M2, un intervalo de confianza (IC) del $100 \times (1-\alpha)\,\%$ para θ está dado por

$$\text{IC}(\theta)_{100\times(1-\alpha)\,\%} = \left[\widehat{\theta} \pm z_{1-\alpha/2}\,\sigma_{\widehat{\theta}}\right]$$

donde $z_{1-\alpha/2}$ es el percentil $1-\alpha/2$ de la distribución normal estándar. Los métodos de estimación son aplicados a cada meta-análisis en estudio a través de un programa computacional disponible en la literatura. Específicamente, dados el número de integraciones por realizar y la información específica requerida para este estudio, se utilizó un programa computacional desarrollado en Excel por Martín (1995). Previamente, los resultados de este programa fueron contrastados con otros programas comerciales tales como Metawin (Rosenberg, Adams & Gurevitch 2000) y uno de libre acceso como Mix v1.56 (Bax, Yu, Ikeda, Tsuruta & Moons 2006).

Las variables en estudio para los 81 meta-análisis son los resultados alcanzados de la aplicación de los métodos de estimación considerados. Específicamente, el OR estimado, su logaritmo –$\log(\widehat{\text{OR}})$– y la varianza de este logaritmo obtenidos con los métodos de estimación ef-clásico, ea-clásico, ef-peto y ea-peto. Estos 81 meta-análisis fueron divididos en tres conjuntos de 27 meta-análisis cada uno, de acuerdo al tamaño del efecto poblacional definido para el estudio. En el primer grupo se consideró una eficacia baja del tratamiento (OR = 0,90), es decir, que el riesgo de presentar la enfermedad en el grupo control varía muy poco con respecto al grupo que recibe el tratamiento; específicamente, la variación es de 1,1. En el segundo grupo se consideró un efecto moderado del tratamiento (OR = 0,49), donde el riesgo de presentar la enfermedad en el grupo control es el doble que en el grupo tratado. En los 27 meta-análisis restantes correspondientes al tercer grupo se consideró un efecto alto del tratamiento (OR = 0,20), lo que significa que el riesgo de presentar la enfermedad en el grupo control es cinco veces mayor que en el grupo tratado. Para comparar el $\log(\widehat{\text{OR}})$ mediante los métodos ef-clásico, ea-clásico, ef-peto y ea-peto en cada uno de los tres grupos de meta-análisis (OR = 0,90; OR = 0,49 y OR = 0,20), se utilizó la prueba t-Student para diferencia de medias. Habitualmente, la representación gráfica de los resultados del tamaño del efecto de los estudios involucrados en un meta-análisis se hace mediante el forest plot o gráfico de OR o riesgo relativo (Abrams & Jones 1995, Rodríguez 2002, Moores et al. 2004). En este estudio, utilizamos el forest plot con fines prácticos para observar los intervalos de confianza y los OR estimados con los distintos métodos de estimación, representando además el tamaño del efecto poblacional establecido a priori.

3. Resultados

En esta sección proporcionamos los resultados más relevantes de este estudio. Sin embargo, si el lector interesado requiere resultados más específicos, éstos pueden solicitarse a los autores (Catalán 2003). Específicamente, con respecto al OR promedio, los dos métodos de estimación aplicados a los 81 meta-análisis arrojaron los siguientes resultados:

(i) Para los 27 meta-análisis donde el OR poblacional presenta un efecto bajo de tratamiento (OR = 0,90), el OR estimado promedio es igual para todos los métodos, tomando un valor de 0,95. Sin embargo, el método ef-peto para el modelo M1 presenta una variación mayor.

(ii) Para los 27 meta-análisis donde el OR poblacional presenta un efecto moderado del tratamiento (OR = 0,49), el OR estimado promedio es igual tanto para el método ef-clásico como para ea-peto, tomando éste un valor de 0,52. Lo mismo sucede con los métodos ef-peto y ea-clásico, donde el OR estimado promedio toma un valor de 0,51. La variación mayor se observa con el método ef-clásico, y la variación menor con el método ea-peto.

(iii) Para el conjunto de meta-análisis donde el OR poblacional presenta un efecto alto del tratamiento (OR = 0,20), el OR estimado promedio con el método ef-clásico es 0,24, mientras que con el método ea-clásico este valor es 0,20. Los métodos ef-peto y ea-peto proporcionan el mismo resultado promedio, que toma un valor de 0,25. La variación mayor se observa con el método ef-clásico, y la variación menor con el método ea-peto.

En cuanto a la significación estadística del OR estimado, se tiene que:

(i) Cuando el efecto del tratamiento es bajo (OR = 0,90), el modelo M1 presenta el porcentaje mayor de valores significativos (ef-peto = 41 %, ef-clásico = 33 %), a diferencia de los métodos de estimación para el modelo M2, donde el porcentaje de OR estimado significativo es de un 7 % en ambos métodos (ea-clásico y ea-peto).

(ii) Cuando el efecto del tratamiento es moderado (OR = 0,45), los métodos de estimación bajo el modelo M1 entregan el 100 % de valores significativos, y un 93 % con los métodos de estimación para el modelo M2.

(iii) Cuando el efecto del tratamiento es alto (OR = 0,20), todos los métodos de estimación entregan un 100 % de resultados significativos.

Los resultados de las pruebas t-Student muestran que:

(i) Para el conjunto de meta-análisis donde el efecto del tratamiento es bajo (OR = 0,90) y moderado (OR = 0,49), no se observan diferencias significativas para el OR estimado promedio entre los dos métodos de estimación empleados en el análisis (valores-p $> 0,10$).

(ii) Cuando el efecto del tratamiento es alto (OR = 0,20), el método ea-clásico en el modelo M2 difiere significativamente (valor-p $< 0,02$) de los métodos ef-clásico, ef-peto y ea-peto, los que presentan valores mayores para el OR estimado promedio.

En general, el método de estimación de Peto usado para estimar los parámetros del modelo M1 y su adaptación para el modelo M2 proporcionan resultados similares.

Sobre la base de los resultados anteriores, se generó una representación gráfica basada en el forest plot de los IC para el OR sobre el conjunto de meta-análisis, donde se reflejan las diferencias entre los métodos de estimación. En las figuras 1-3 se observa en términos generales que la amplitud de los IC para el OR es diferente a medida que la varianza entre estudios aumenta. Específicamente,

(i) Desde las figuras 1 y 2 se observa que los métodos ef-clásico y ea-clásico dan resultados similares y que lo mismo sucede con los métodos ef-peto y ea-peto. Sin embargo, se observa que los métodos ef-peto y ea-peto sobrestiman el OR, y generalmente el IC no contiene su valor verdadero.

(ii) Desde la figura 3 se observa que los IC para el tamaño del efecto son más amplios, lo que refleja una variabilidad mayor entre los estudios ($\tau^2 = 0,8$). Es en esta figura donde se observa una diferencia mayor entre los métodos de estimación. Específicamente, entre el método clásico para el modelo M2 y el resto, siendo el método ea-clásico el que proporciona resultados más aproximados al tamaño verdadero del efecto.

4. Discusión

En meta-análisis, una de las primeras decisiones que un investigador debe tomar es la elección del modelo bajo el cual va a integrar sus resultados. Esta decisión puede tomarla el investigador a priori en función del conocimiento metodológico del tema revisado. Sin embargo, en la bibliografía hay mucha controversia acerca de la elección del modelo bajo el cual se integran los resultados. Schmidt, Oh & Hayes (2009) consideraron que la aplicación del modelo de efectos fijos está limitada a estudios muy similares entre ellos. Por el contrario, Peto (1987) y Thompson & Pocock (1991) propusieron un modelo de efectos fijos argumentando que la integración debe limitarse a los estudios revisados con hipótesis que hagan referencia a los mismos y rechazar hipótesis inferenciales más ambiciosas que son difíciles de contrastar. Una práctica generalizada es la de partir del modelo de efectos fijos, y ante la presencia de heterogeneidad, el investigador debe utilizar factores elegidos a priori que puedan explicar esta heterogeneidad. En el caso de que el investigador no sea capaz de explicar dicha heterogeneidad, él debería considerar un modelo de efectos aleatorios, aunque Pocock & Hughes (1990) y Greenland & Salvan (1990) concluyeron que ninguna aproximación es buena en presencia de una heterogeneidad fuerte. Una de las conclusiones que se extrae de este trabajo es la diferencia entre los resultados del tamaño del efecto estimado que proporcionan los

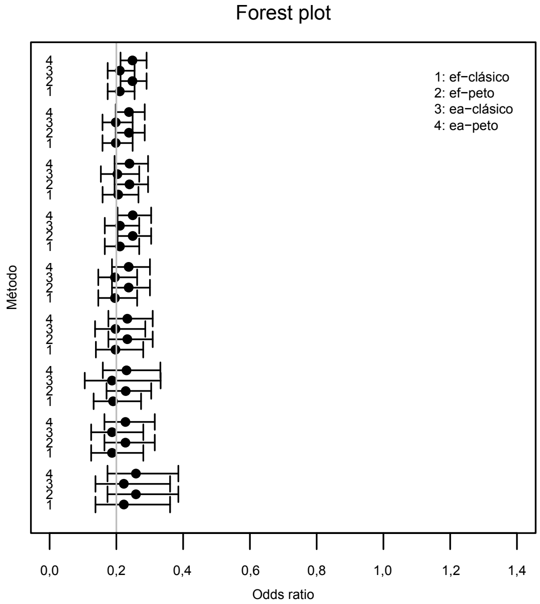


FIGURA 1: IC para el OR con 4 métodos de estimación para 9 meta-análisis, donde el tamaño del efecto verdadero es OR = 0,20 y la varianza entre estudios es pequeña ($\tau^2 = 0,015$).

modelos M1 y M2. Estas diferencias van a depender del tamaño del efecto que se pretende estimar y de la varianza entre estudios. Cuando el efecto de tratamiento en la población es bajo (OR = 0,90) o moderado (OR = 0,49), no se observan diferencias significativas entre los métodos de estimación para modelos de efectos fijos y aleatorios. Las diferencias mayores se producen con la mayor variabilidad considerada ($\tau^2 = 0,8$). Sin embargo, cuando el efecto de tratamiento es alto (OR = 0,20), se detectan diferencias significativas entre los métodos. Esta diferencia refleja tanto el tamaño del efecto en la población como la variabilidad presente entre los estudios. Cuando la varianza entre estudios es grande, el modelo más adecuado es M2, como cabría esperar, lo que se observa al comparar los resultados del método clásico de la media ponderada bajo los modelos M1 y M2. Nuestros resultados coinciden con los obtenidos por Berlin et al. (1989), donde se encuentra que el método ef-peto es similar al método ea-clásico cuando la heterogeneidad no es muy grande. No obstante, entre una gama de procedimientos de estimación descritos en la literatura (DerSimonian & Laird 1986), existen algunos métodos que no reflejan las características de la población en estudio. Uno de éstos es el método de Peto cuando se considera un modelo de efectos aleatorios, ya que este método proporciona resultados similares al método de Peto cuando se considera un modelo de efectos fijos. Esta similitud puede deberse a que el estimador de Peto para el modelo M2 se ha construido de manera artificial a partir del estimador de Peto considerando un modelo de efectos fijos. Por tanto, deducimos que el estimador de Peto para modelos de efectos aleatorios no es una buena aproximación cuando sólo existen efectos fijos. En cuanto a la significación del valor estimado del OR,

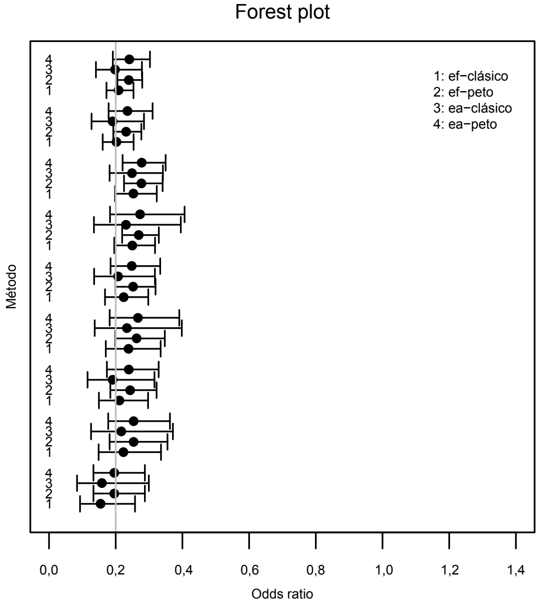


Figura 2: IC para el OR con 4 métodos de estimación para 9 meta-análisis, donde el tamaño del efecto verdadero es OR $=$ 0,20 y la varianza entre estudios es moderada ($\tau^2 = 0,15$).

se tiene que los modelos de efectos fijos son más sensibles para detectar un efecto de tratamiento pequeño. Sin embargo, cuando el efecto de tratamiento en estudio es alto, los dos métodos de estimación empleados entregan resultados significativos, pero se diferencian en la amplitud de los intervalos de confianza y en cómo el OR estimado se aproxima a su valor verdadero. Engels, Schmid, Terrin, Olkin & Lau (2000) analizaron varios meta-análisis que utilizan el OR y riesgos absolutos concluyendo que, al comparar los modelos M1 y M2, existe un incremento lógico de los errores estándar en el modelo M2, lo que produce menos significatividad. El método clásico de la media ponderada es el que proporciona las mejores estimaciones del OR y, según se observa, los valores estimados con el método de Peto usando los modelos M1 y M2 son mayores que el valor verdadero, es decir, hay una sobreestimación del efecto de tratamiento. El estimador de Peto bajo el supuesto de efectos fijos es insesgado bajo la hipótesis de independencia, aunque el sesgo aumenta cuando el OR poblacional se aleja de un valor igual a uno. Además, este sesgo es mayor cuando los estudios individuales son balanceados, situación poco común en los estudios no experimentales (Greenland & Salvan 1990).

Tal como se ha planteado ampliamente en la literatura, para la realización de un meta-análisis se hace necesario conocer en profundidad el problema en estudio, de tal forma que se tengan referencias acerca de la variabilidad de los resultados en cada estudio involucrado y del efecto de tratamiento que se pretende estimar (Egger, Smith & Phillips 1997). El crecimiento en el número de meta-análisis publicados en los últimos años ha llevado a elaborar normas de publicación para

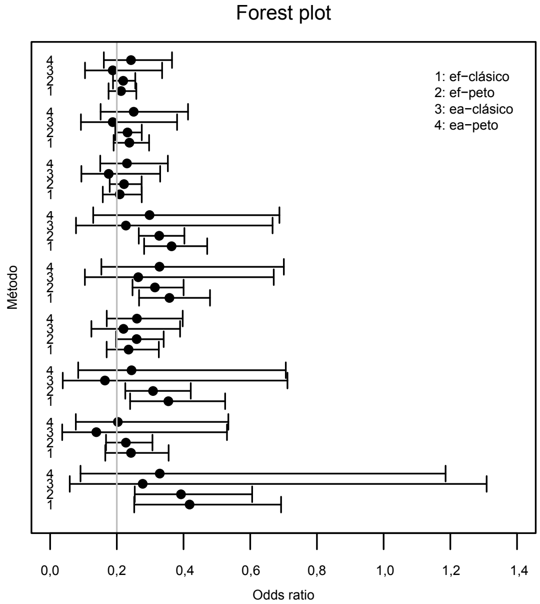


Figura 3: IC para el OR con 4 métodos de estimación para 9 meta-análisis, donde el tamaño del efecto verdadero es OR $= 0{,}20$ y la varianza entre estudios es alta ($\tau^2 = 0,8$).

facilitar su interpretación y utilización. Sobre esto, en la propuesta QUOROM (quality of reporting of meta-analysis) se identifican varios ámbitos por abordar y, específicamente, el punto 14 trata sobre la síntesis cuantitativa de los datos y sobre los métodos de combinación de los resultados (Moher, Cook, Eastwood, Olkin & Rennie 1994, Urrutia, Tort & Bonfill 2005). Moher, Liberati, Tetzlaff & Altman (2009) revisaron y actualizaron las líneas maestras para la realización de meta-análisis que se denomina PRISMA (preferred reporting items for systematic reviews and meta-analysis).

Como en cualquier investigación, en estudios de meta-análisis la elección del método de análisis de datos se fundamenta en las características del problema y en los objetivos e hipótesis que se plantean. Por esta razón, se hace necesario conocer las ventajas y desventajas de los procedimientos estadísticos que se podrían utilizar. En este sentido, el resultado de la estimación del tamaño del efecto en un meta-análisis no depende solamente del diseño de los estudios individuales, sino que también puede depender del tipo de modelo y del procedimiento de estimación que se emplea. Para más antecedentes, el lector interesado puede revisar las guías de Cochrane, las que son ampliamente utilizadas y validadas para meta-análisis en salud (`http://www.cochrane-handbook.org`).

5. Conclusiones

En este estudio hemos comparado los resultados de dos métodos de estimación del tamaño del efecto basados en meta-análisis usando modelos de efectos fijos y aleatorios. La medida del efecto considerada en este estudio fue el odds ratio. Esto se debe a que tal medida del tamaño del efecto está presente en revisiones sistemáticas de temas de interés en el campo clínico procedentes de diseños experimentales, observacionales o pruebas de diagnóstico. Hemos observado diferencias entre los dos modelos analizados en cuanto al porcentaje de estimaciones de odds ratio significativos. Con respecto al tamaño del efecto estimado, cuando el efecto del tratamiento es alto, el resultado del método de DerSimonian-Laird difiere significativamente de los otros métodos, mientras que el método de Peto y su versión adaptada presentan resultados similares. En presencia de heterogeneidad entre los estudios, el método de DerSimonian-Laird es el que más se aproxima a los resultados verdaderos, mientras que el método de Peto no es una buena aproximación.

Agradecimientos

Los autores desean agradecer a las editoras invitadas de este número especial en bioestadística de la Revista Colombiana de Estadística, Dra. Liliana López Kleine y Dra. B. Piedad Urdinola, y a dos árbitros anónimos por sus comentarios constructivos sobre la versión original de este artículo que condujeron a esta versión final mejorada. La investigación de Víctor Leiva fue parcialmente financiada por el proyecto FONDECYT 1120879 del gobierno de Chile.

[Recibido: septiembre de 2011 — Aceptado: febrero de 2012]

Referencias

Abrams, K. & Jones, D. (1995), 'Meta-analysis and the synthesis of evidence', *IMA Journal of Mathematics Applied in Medicine and Biology* **12**, 297–313.

Allam, M. F., Del Castillo, A. S. R. & Navajas, F. (2003), 'Enfermedad de Parkinson temprana y tabaco: metanálisis', *Revista de Neurología* **12**, 1101–1103.

Bax, L., Yu, L. M., Ikeda, N., Tsuruta, H. & Moons, K. G. M. (2006), 'Enfermedad de Parkinson temprana y tabaco: metanálisis', *BMC Medical Research Methodology* **6**.

Berlin, J. A., Laird, N. M., Sacks, H. S. & Chalmers, T. (1989), 'A comparison of statistical methods for combining event rates from clinical trials', *Statistics in Medicine* **8**, 141–151.

Burguillo, F. J., Martín, F. J., Barrera, I. & Bardsley, W. G. (2010), 'Meta-analysis of microarray data: The case of imatinib resistance in chronic myelogenous leukemia', *Computational Biology & Chemistry* **34**, 184–192.

Catalán, M. (2003), Los modelos multinivel como herramienta de análisis de datos biomédicos con estructura jerárquica, Tesis doctoral, Universidad de Salamanca, Departamento de Estadística.

Catalán, M. & Galindo, M. P. (2003), 'Utilización de los modelos multinivel en investigación sanitaria', *Gaceta Sanitaria* **17**(3), 35–52.

Claros, N., Manterota, C., Vial, M. & Sanhueza, A. (2007), 'Efectividad de la profilaxis antibiótica en el curso de la colecistectomía laparoscopica electiva. Revisión sistemática de la literatura', *Revista Chilena de Cirugía* **59**, 353–359.

Coomarasamy, A., Papaioannou, S., Gee, H. & Khan, K. S. (2001), 'Aspirin for the prevention of preeclampsia in women with abnormal uterine artery doppler: A meta-analysis', *Obstetrics and Gynecology* **98**, 861–866.

Cornelissen, V. A. (2007), 'Incidence of cardiovascular events in white-coat, masked and sustained hypertension versus true normotension: a metaanalysis', *Journal of Hypertension* **25**, 2193–2198.

Dentali, F., Douketis, J. D., Lim, W. & Crowther, M. (2007), 'Combined aspirin-oral anticoagulanttherapy compared with oral anticoagulant therapy alone among patients at risk for cardiovascular disease: A meta-analysis of randomized trials', *Archives of Internal Medicine* **167**, 117–124.

DerSimonian, R. & Laird, N. (1986), 'Meta-analysis in clinical trials', *Controlled Clinical Trials* **7**, 177–188.

Egger, M., Smith, G. D. & Phillips, A. N. (1997), 'Meta-analysis: Principles and procedures', *British Medical Journal* **315**, 1533–1537.

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I. & Lau, J. (2000), 'Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses', *Statistics in Medicine* **19**, 1707–1728.

Glass, G. V. (1976), 'Primary, secondary and meta-analysis of research', *Educational Researcher* **6**, 3–8.

Greenland, S. & Salvan, A. (1990), 'Bias in the one-step method for pooling study results', *Statistics in Medicine* **9**, 247–252.

Hedges, L. & Olkin, I. (1985), *Statistical Methods for Meta-Analysis*, Academic Press, New York.

Hedges, L. V. & Vevea, J. L. (1998), 'Fixed and random effects in meta-analysis', *Psychological Methods* **3**, 486–504.

Jiménez-Ruiz, C., Riesco, J. A., Ramos, A. & Barrueco, M. (2008), 'Recomendaciones para el tratamiento farmacológico del tabaquismo. Propuestas de financiación', *Archivos de Bronconeumología* **44**, 213–219.

Leyland, A. H. & Goldstein, H. (2001), *Multilevel Modelling of Health Statistics*, Wiley, New York.

Martín, J. (1995), Métodos estadísticos en meta-análisis, Ph.d. thesis, Universidad de Salamanca, España.

Martín, J., Donaldson, A. N. A., Villarroel, R., Parmar, M. K. B., Ernst, E. & Higginson, I. J. (2002), 'Efficacy of acupuncture in asthma: Systematic review and meta-analysis of published data from 11 randomised controlled trials', *European Respiratory Journal* **20**, 846–852.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I. & Rennie, D. (1994), 'Improving the quality of reporting of meta-analysis of randomized controlled trials: The quorom statement', *Lancet* **354**, 1896–1900.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. J. (2009), 'The prisma group', *PLoS Medicine* **6**, 1–6.

Moores, L., Jackson, W., Shorr, A. & Jackson, J. (2004), 'Meta-analysis: Outcomes in patients with suspected pulmonary embolism managed with computed tomographic pulmonary angiography', *Annals of Internal Medicine* **141**, 866–875.

Petitti, D. B. (1994), *Meta-Analysis, Decision Analysis, and Cost-effectiveness Analysis*, University Press, Oxford.

Peto, R. (1987), 'Why do we need systematic overviews of randomized trials?', *Statistics in Medicine* **6**, 233–240.

Pocock, S. T. & Hughes, M. D. (1990), 'Estimation issues in clinical trials and overviews', *Statistics in Medicine* **9**, 657–671.

Rodríguez, G. (2002), 'Entendiendo los diagramas de odds-ratio de las revisiones sistemáticas', *CES Medicina* **16**, 66–72.

Rosenberg, M., Adams, D. C. & Gurevitch, J. (2000), *MetaWin 2.0: Statistical software for meta-analysis*, Sinauer Associates, Sunderland.

Rosenthal, R. (1984), *Meta-Analytic Procedures for Social Research*, Sage, Beverly Hills.

Rydzewska, L., Tierney, J., Vale, C. L. & Symonds, P. R. (2010), 'Neoadjuvant chemotherapy plus surgery versus surgery for cervical cancer', *Cochrane Database of Systematic Reviews* **1**. CD007406.

Schmidt, F. L., Oh, I. & Hayes, T. L. (2009), 'Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results', *British Journal of Mathematical & Statistical Psychology* **62**, 97–128.

Stowe, R. L., Ives, N. J., Clarke, C. & van Hilten, J. (2008), *Tratamiento con agonistas dopaminérgicos para la enfermedad de Parkinson en sus etapas iniciales*, Wiley, Oxford. (Revisión Cochrane traducida).

Thompson, S. G. & Pocock, S. J. (1991), 'Can meta-analysis be trusted?', *The Lancet* **338**, 1127–1130.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H. & Thompson, S. G. (2000), 'A multilevel model framework for meta-analysis of clinical trials with binary outcomes', *Statistics in Medicine* **19**, 3417–3432.

Urrutia, G., Tort, S. & Bonfill, X. (2005), 'Metaanálisis (quorum)', *Medicina Clínica* **125**(1), 32–37.

Vamvakas, E. C. (2011), Meta-analysis: A statistical method to integrate information provided by different studies, *in* A. M. Marchevsky & M. Wick, eds, 'Evidence Based Pathology and Laboratory Medicine', Springer, New York, pp. 149–171.

Yusuf, S., Peto, R., Lewis, J., Collins, R. & Sleight, P. (1985), 'Beta blockade during and after myocardial infarction: An overview of randomized trials', *Progress in Cardiovascular Diseases* **27**, 335–371.

Zhou, H., Zhang, J., Wang, Q. & Hu, Z. (2009), 'Meta-analysis: Antibiotic prophylaxis in elective laparoscopic cholecystectomy', *Alimentary Pharmacology and Therapeutics* **29**, 1086–1095.

Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 223 a 237

Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?

Comparación entre SVM y regresión logística: ¿cuál es más recomendable para discriminar?

Diego Alejandro Salazar[1,a], Jorge Iván Vélez[2,b], Juan Carlos Salazar[1,2,c]

[1]Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia
[2]Grupo de Investigación en Estadística, Universidad Nacional de Colombia, Medellín, Colombia

Abstract

The classification of individuals is a common problem in applied statistics. If X is a data set corresponding to a sample from an specific population in which observations belong to g different categories, the goal of classification methods is to determine to which of them a *new* observation will belong to. When $g = 2$, logistic regression (LR) is one of the most widely used classification methods. More recently, Support Vector Machines (SVM) has become an important alternative. In this paper, the fundamentals of LR and SVM are described, and the question of which one is better to discriminate is addressed using statistical simulation. An application with real data from a microarray experiment is presented as illustration.

Key words**:** Classification, Genetics, Logistic regression, Simulation, Support vector machines.

Resumen

La clasificación de individuos es un problema muy común en el trabajo estadístico aplicado. Si X es un conjunto de datos de una población en la que sus elementos pertenecen a g clases, el objetivo de los métodos de clasificación es determinar a cuál de ellas pertenecerá una *nueva* observación. Cuando $g = 2$, uno de los métodos más utilizados es la regresión logística. Recientemente, las Máquinas de Soporte Vectorial se han convertido en una alternativa importante. En este trabajo se exponen los principios básicos de ambos métodos y se da respuesta a la pregunta de cuál es más recomendable

[a]MSc student. E-mail: diasalazarbl@unal.edu.co
[b]Researcher. E-mail: jorgeivanvelez@gmail.com
[c]Associate professor. E-mail: jcsalaza@unal.edu.co

para discriminar, vía simulación. Finalmente, se presenta una aplicación con datos provenientes de un experimento con microarreglos.

***Palabras clave*:** clasificación, genética, máquinas de soporte vectorial, regresión logística, simulación.

1. Introduction

In applied statistics, it is common that observations belong to one of two mutually exclusive categories, e.g., presence or absence of a disease. By using a (random) sample from a particular population, classification methods allow researchers to discriminate *new* observations, i.e. assign the group to which this new observation belongs based on discriminant function (Fisher 1936, Anderson 1984) after the assumptions on which it relies on are validated. However, in practice, these assumptions cannot always be validated and, as a consequence, veracity of results is doubtful. Moreover, the implications of wrongly classifying a new observation can be disastrous.

To relax the theoretical assumptions of classical statistical methods, several alternatives have been proposed (Cornfield 1962, Cox 1966, Day & Kerridge 1967, Hosmer & Lemeshow 1989), including logistic regression (LR), one of the most widely used techniques for classification purposes today. More recently, new methodologies based on iterative calculations (algorithms) have emerged, e.g., neural networks (NN) and machine learning. However, pure computational approaches have been seen as "black boxes" in which data sets are throw in and solutions are obtained, without knowing exactly what happens inside. This, in turn, limits their interpretation.

Support Vector Machine (SVM) (Cortes & Vapnik 1995) is a classification and regression method that combines computational algorithms with theoretical results; these two characteristics gave it good reputation and have promoted its use in different areas. Since its appearance, SVM has been compared with other classification methods using real data (Lee, Park & Song 2005, Verplancke, Van Looy, Benoit, Vansteelandt, Depuydt, De Turck & Decruyenaere 2008, Shou, Hsiao & Huang 2009, Westreich, Lessler & Jonsson 2010) and several findings have been reported. In particular, (*i*) SVM required less variables than LR to achieve an equivalent misclassification rate (MCR) (Verplancke et al. 2008), (*ii*) SVM, LR and NN have similar MCRs to diagnose malignant tumors using imaging data (Shou et al. 2009), and (*iii*) NN were much better than LR with sparse binary data (Asparoukhova & Krzanowskib 2001).

In this paper we compare, by statistical simulation, the MCRs for SVM and LR when the data comes from a population in which individuals can be classified in one of two mutually exclusive categories. We consider different scenarios in which the training data set and other functional parameters are controlled. This control allowed us to generate data sets with specific characteristics and further decide whether SVM or LR should be used in that particular situation (Salazar 2012).

2. SVM and Logistic Regression

2.1. SVM for Two Groups

Moguerza & Muñoz (2006) and Tibshirani & Friedman (2008) consider a classification problem in which the discriminant function is nonlinear (Figure 1a), and there exists a kernel function Φ to a *characteristic space* on which the data is linearly separable (Figure 1b). On this new space, each data point corresponds to an abstract point on a p-dimensional space, being p the number of variables in the data set.

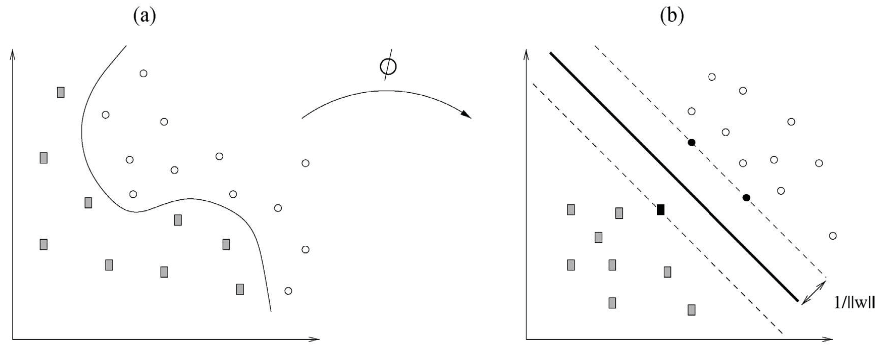


FIGURE 1: An illustration of a SVM model for two groups modified from Moguerza & Muñoz (2006). Panel (a) shows the data and a non-linear discriminant function; (b) how the data becomes separable after a kernel function Φ is applied.

When Φ is applied to the original data, a new data $\{(\Phi(\mathbf{x}_i), y_i)\}_{i=1}^n$ is obtained; $y_i = \{-1, 1\}$ indicates the two possible classes (categories), and any equidistant hyperplane to the closest point of each class on the new space is denoted by $\mathbf{w}^T\Phi(\mathbf{x}) + b = 0$. Under the separability assumption (Cover 1965), it is possible to find $\mathbf{w}$ and b such that $|\mathbf{w}^T\Phi(\mathbf{x}) + b| = 1$ for all points closer to the hyperplane. Thus,

$$\mathbf{w}^T\Phi(\mathbf{x}) + b \begin{cases} \geq 1, & \text{if } y_i = 1 \\ \leq -1, & \text{if } y_i = -1 \end{cases} \qquad i = 1, \ldots, n \tag{1}$$

such that the distance (margin) from the closest point of each class to the hyperplane is $1/||w||$ and the distance between the two groups is $2/||w||$. Maximizing the margin implies to solve

$$\min_{\mathbf{w},b} ||\mathbf{w}||^2 \quad \text{subject to} \quad y_i(\mathbf{w}^T\Phi(\mathbf{x}) + b) \geq 1 \qquad i = 1, \ldots, n \tag{2}$$

Let $\mathbf{w}^*$ and b^* the solution of (2) that defines the hyperplane

$$D^*(\mathbf{x}) = (\mathbf{w}^*)^T\Phi(\mathbf{x}) + b^* = 0$$

on the *characteristic space*. All values of $\Phi(x_i)$ satisfying the equality in (2) are called *support vectors*. From the infinite number of hyperplanes separating the data, SVM gives the optimal margin hyperplane, i.e., the one on which the classes are more distant.

Once the optimal margin hyperplane has been found, it is projected on the data's original space to obtain a discriminant function. For example, Figure 2(a) shows a data set in $\mathbb{R}^2$ in which two groups, linearly separable, are characterized by white and black dots that are not linearly separable. In Figure 2(b), the data is transformed to $\mathbb{R}^3$ where it is separable by a plane and, when it is projected back to the original space, a circular discriminant function is obtained.

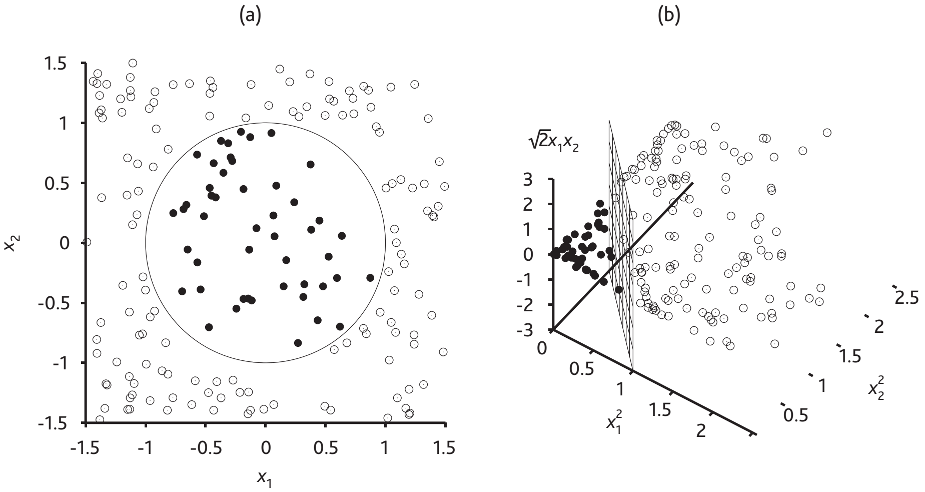


FIGURE 2: An SVM example in which (a) the two-dimensional training data set (black circles represent cases) becomes a linear decision boundary in three dimensions (b). Modified from Verplancke et al. (2008).

2.2. Logistic Regression

Let Y be a random variable such that

$$Y = \begin{cases} 1, & \text{if the condition is present} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and $\mathbf{x} = (x_1, x_2, \dots, x_p)$ be covariates of interest. Define

$$\pi(\mathbf{x}) = E(Y|x_1, \dots, x_p)$$

as the probability that one observation $\mathbf{x}$ belongs to one of the two groups. The Logistic Regression model is given by Hosmer & Lemeshow (1989):

$$\pi(\mathbf{x}) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}} \tag{4}$$

Applying the transformation

$$\text{logit}(y) = \log(y/(1-y)) \tag{5}$$

on (4) yields to a linear model in the parameters. If $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimation of $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$, then the probability that a *new* observation $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_p^*)$ belongs to one of the two groups is

$$\widehat{\pi}(x^*) = \frac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \cdots + \widehat{\beta}_p x_p^*\}}{1 + \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \cdots + \widehat{\beta}_p x_p^*\}} \tag{6}$$

such that a *new* observation $\mathbf{x^*}$ will be classified in the group for which (6) is higher.

3. Simulation and Comparison Strategies

Let $g = \{1, 2\}$ be the group to which an observation belongs to. In our simulation approach, different probability distributions were considered for simulating the *training* and *validation* data sets (see Table 1). As an example, consider the Poisson case in which we generate n_1 and n_2 observations from a Poisson distribution with parameter $\lambda = 1$ (first group, $g = 1$) and $\lambda = d$ (second group, $g = 2$), respectively, with d taking the values 3, 5, 8, and 10. In real-world applications, the Poisson data being generated could be seen as white blood cell counts. Note that the greater the value of d, the greater the separation between groups.

Table 1: Probability distributions in our simulation study.

Distribution	$g = 1$	$g = 2$	d
Poisson	Poisson(1)	Poisson(d)	$\{3, 5, 8, 10\}$
Exponential	Exp(1)	Exp(d)	$\{3, 5, 8, 10\}$
Normal	$N(0, 1)$	$N(d, 1)$	$\{0.5, 1, 2, 2.5\}$
Cauchy-Normal	Cauchy(0, 1)	$N(d, 1)$	$\{1, 2, 4, 5\}$
Normal-Poisson	$N(0, 1)$	Poisson(d)	$\{1, 2, 4, 5\}$
Bivariate Normal[a]	$N_2(\mathbf{0}, \boldsymbol{\Sigma}_1)$	$N_2(d, \boldsymbol{\Sigma}_1)$	[b]

[a] $\boldsymbol{\Sigma}_1$ is a 2×2 correlation matrix whose off-diagonal elements are $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$

[b] d is a bivariate vector with elements $(d_1, d_2) = \{(0, 0), (1, 0), (1, 1.5), (2.5, 0)\}$

Our simulation and comparison strategies involve the following steps:

1. Define a probability distribution to work with.
2. Draw n_g individuals (see Hernández & Correa 2009) to form the D, the *training* data set.
3. On D, fit the LR and SVM models.
4. Draw new observations as in 1 to form D^*, the *validation* data set.
5. On D^*, evaluate the models fitted in 2. Determine their accuracy by estimating the misclassification rate (MCR)[1] calculated as $(g_{1,2} + g_{2,1})/(n_1 + n_2)$,

[1] These tables are available from the authors under request.

where $g_{i,j}$ is the number of individuals belonging to group i being classified in group j, $i, j = 1, 2$.

6. Repeat 3 and 4, $B = 5000$ times and calculate the average MCR.

Steps 2-6 were programmed in R (R Development Core Team 2011) considering several probability distributions (Table 1). Of note, either or both the expected value, variance or correlation parameter were controlled by the simulation parameter d. As samples sizes (i) $n_1 = n_2 = 20,\ 50,\ 100$ and (ii) $n_1 \neq n_2$ were used.

In LR, models were fitted using the `glm()` function from R and individuals were assigned to the group g for which the probability was higher.

SVM models including (i) linear, (ii), polynomial, (iii) radial and (iv) tangential kernels were fitted and tuned using the `e1071` facilities (Dimitriadou, Hornik, Leisch, Meyer, & Weingessel 2011). When tuning these models, the parameters γ, which controls the complexity of the classification function build by the SVM, and C, which controls the penalty paid by the SVM for missclassifying a training point (Karatzoglou, Meyer & Hornik 2006, pp. 3), were determined using the `tune.svm()` function in `e1071`.

4. Results

4.1. Univariate Distributions

Results for the Normal, Poisson and Exponential distributions are reported in Figure 3[2].

In the Normal case, the MCR for the polynomial SVM model is higher (poor performance). On the other hand, the performances of LR and linear, radial and tangential SVM models are equivalent. When the sample sizes differ, the MCR of the tangential and polynomial kernel is lower than when the groups have the same number of individuals. However, the former presents lower MCRs.

When the observations from both groups come from a Poisson distribution and the number of individuals by group is the same, the polynomial SVM kernel performs poorer compared with other methods, which are good competitors to LR. Additionally, the performance of the tangential kernel is not as good as it is for the LR and radial and linear kernels. LR is preferable to SVM methods when the sample sizes are not equal.

In the Exponential case, except for the polynomial kernel, SVM models perform equally well than LR when both groups have the same number of individuals. Conversely, LR performs better than SVM methods when the sample sizes are not the same. As in the Normal and Poisson distributions, the polynomial SVM is not recommended.

[2] Color versions of all figures presented from now on are available from the authors under request.

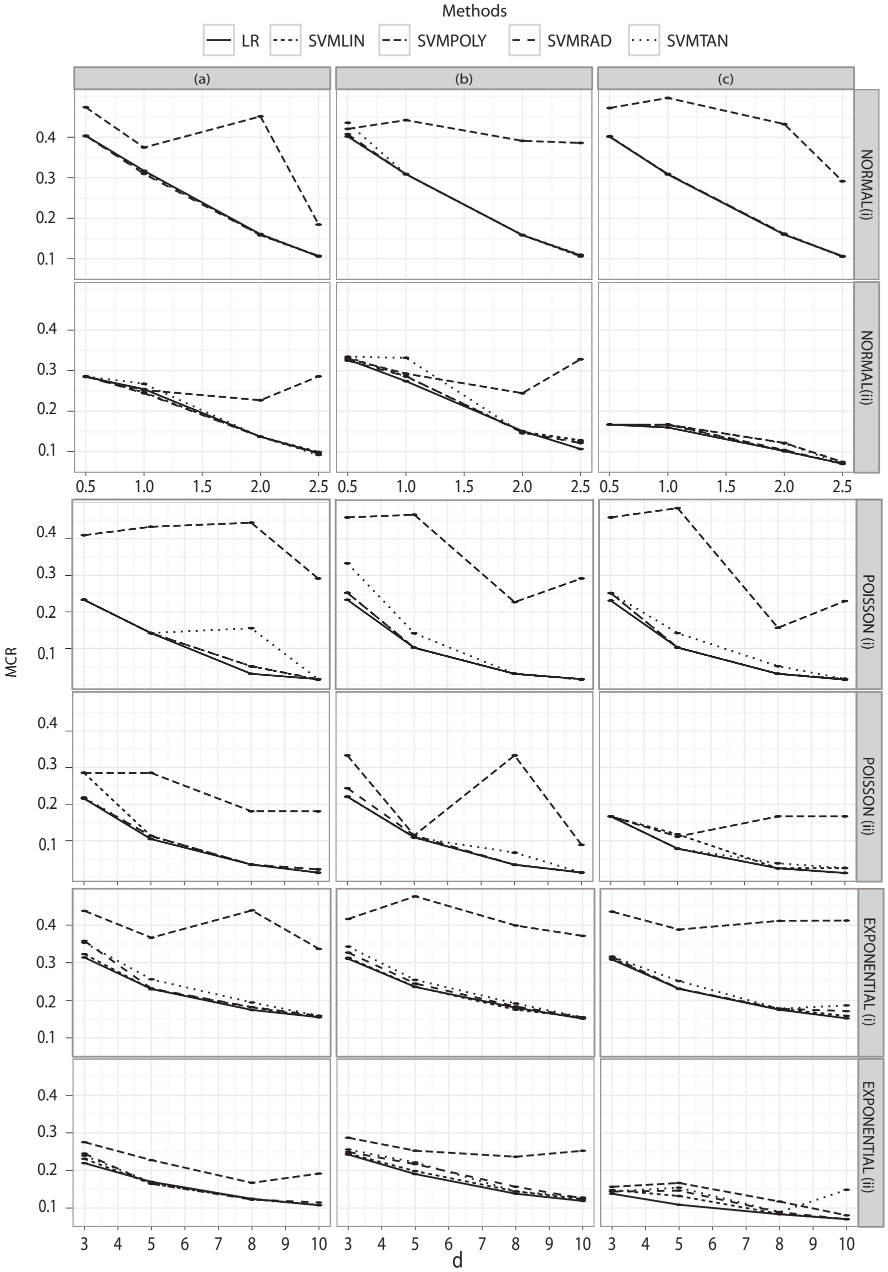


Figure 3: MCR as a function of d for the LR and SVM models when the observations come from the Normal, Poisson and Exponential distributions. Sample sizes in (i) are equal to (a) 20, (b) 50 and (c) 100 individuals per group. In row (ii), (a) $n_1 = 20, n_2 = 50$, (b) $n_1 = 50, n_2 = 100$, (c) $n_1 = 20, n_2 = 100$ individuals. See Table 1 for more details.

4.2. Mixture of Distributions

In Figure 4, the MCR for the Cauchy-Normal and Normal-Poisson mixtures is presented. Regardless the groups' sample sizes, SVM models perform better than LR in a Cauchy-Normal mixture. Interestingly, the polynomial kernel performs poorer when the number of individuals in both groups is the same (upper panel), but its performance improves when they are different (lower panel).

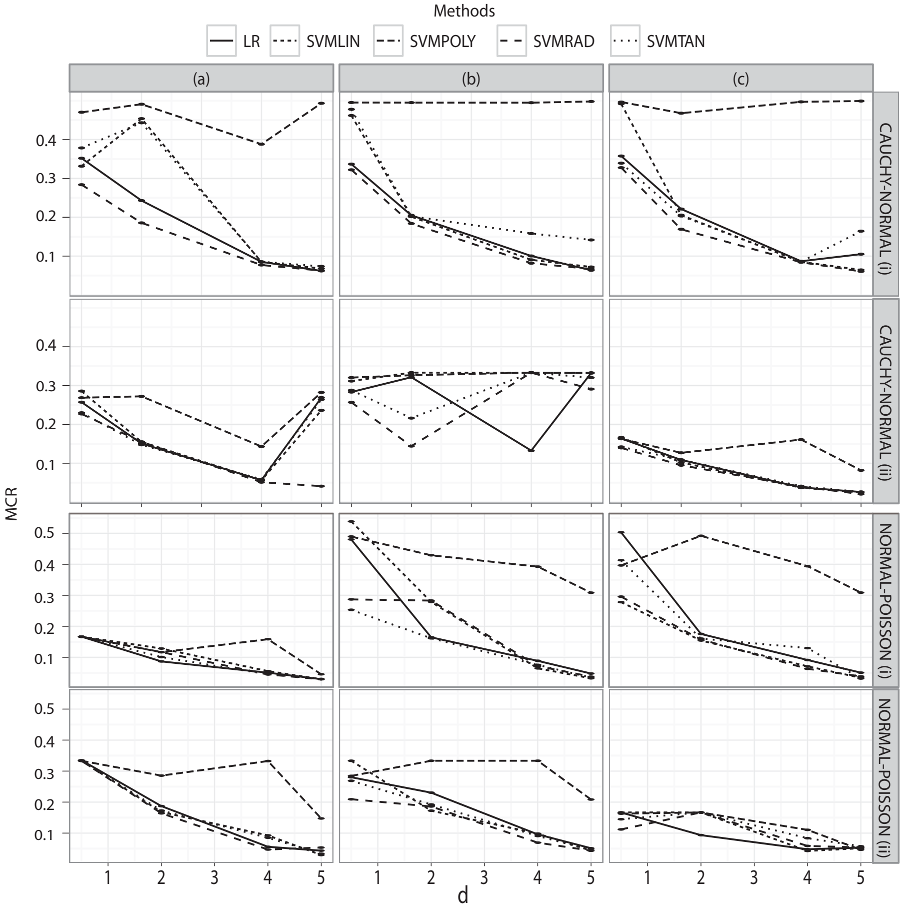


FIGURE 4: MCR as a function of d for the LR and SVM models when the observations come from a Cauchy-Normal and a Normal-Poisson mixture distributions. Conventions as in Figure 3. See Table 1 for more details.

In the Normal-Poisson mixture, the MCRs for SVM are lower than those for LR, especially when d is low, i.e., the expected value of both groups is similar. When $n_1 = n_2$ (upper panel), the linear and radial SVM models present lower MCRs than LR when the sample sizes increase.

Results for the Bivariate Normal distribution are presented in Figure 5. For all methods, the MCR decreases when ρ increases and the sample size is the same for both groups. However, if the number of individuals per group is different and d is low, the MCR for LR is similar regardless ρ. Under this scenario, the radial and tangential SVM models perform as good as LR. Conversely, the linear kernel shows a poor performance.

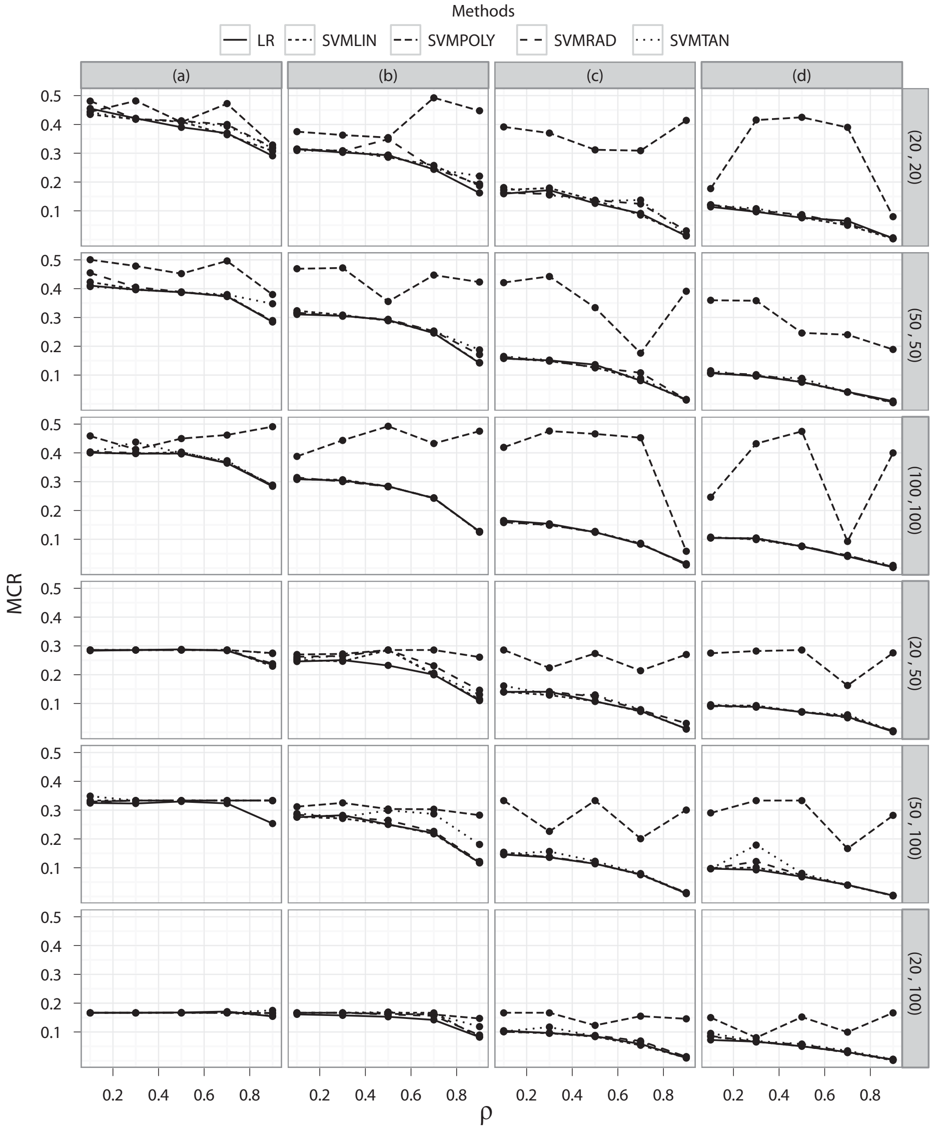


Figure 5: MCR as a function of ρ for the Bivariate Normal distribution when the mean vector is (a) (0,0), (b) (1,0), (c) (1, 1.5) and (d) (2.5, 0). Rows correspond to combinations of n_1 and n_2 of the form (n_1, n_2). Here (20, 50) corresponds to $n_1 = 20$ and $n_1 = 50$.

5. Application

Mootha, Lindgren, Eriksson, Subramanian, Sihag, Lehar, Puigserver, Carlsson, Ridderstrele, Laurila, Houstis, Daly, Patterson, Mesirov, Golub, Tamayo, Spiegelman, Lander, Hirschhorn, Altshuler & Groop (2003) introduce an analytical strategy for detecting modest but coordinate changes in the expression of groups of functionally related genes, and illustrate it with DNA microarrays measuring gene expression levels in 43 age-matched skeletal muscle biopsy samples from males, 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT) and 18 with type 2 diabetes (T2D). As a result, they identified a set of genes involved in oxidative phosphorylation.

Table 2: Statistics for the top 10 differentially expressed genes. No correction by multiple testing was applied.

Gene	t-statistic	$\bar{x}_{\text{NGT}} - \bar{x}_{\text{T2D}}$	P-value
G557	3.8788	0.1632	0.0005
G591	−3.6406	−0.1008	0.0009
G226	3.0621	0.1285	0.0044
G718	−3.0566	−0.1093	0.0044
G45	−2.8978	−0.1275	0.0066
G137	2.8432	0.1255	0.0076
G737	−2.6544	−0.1947	0.0121
G587	−2.5774	−0.2654	0.0146
G232	−2.5607	−0.3213	0.0152
G185	−2.5368	−0.2752	0.0161

For analysis, expression levels were processed as follows. First, a subset of 1000 genes was randomly selected from the original data. Second, the expression levels in samples from NGT (controls, group 1) and T2D individuals (cases, group 2) were compared using a two-sample t-test as implemented in `genefilter` (Gentleman, Carey, Huber & Hahne 2011). Third, only the expression levels for the top 30 differentially expressed (DE) genes were subsequently used to fit the LR and SVM models.

Summary statistics for the top 10 genes found to be DE are presented in Table 2; genes G557, G226 and G137 are down-regulated, i.e., their expression levels are lower in T2D than in NGT samples. Figure 6 depicts a scatterplot for the top 5 genes by disease status. In there, some (expected) correlation structures are observed; these correlations might constitute a potential problem for any classification method.

LR and SVM models were fitted using the disease status as dependent variable and the expression levels of k genes as covariates. Our findings are reported in Figure 7. For predicting the disease status in this data set, (i) SVM models required less variables (genes); (ii) all methods performed similarly when $k <$ 5, but the radial SVM model is more consistent, and (iii) the polynomial and tangential SMVs are definitely not an option. These results may provide important insights in the diagnosis of genetic diseases using this type of models.

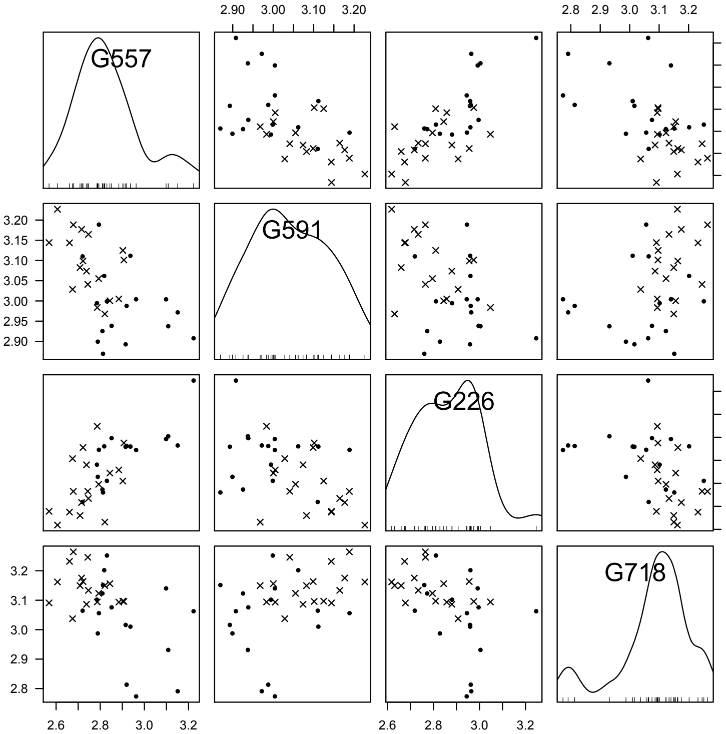


Figure 6: Scatterplot matrix for some of the genes presented in Table 2. Filled dots correspond to NGT samples (controls). In the diagonal panel, density plots are shown.

Although in our application we only used a subset of the genes available in the microarray experiment, it illustrates how a SVM model can be used to predict the (disease) status of a patient using his/her genetic information. Furthermore, we evaluated the possibility of including "one-gene-at-the-time" and determine the MCR of the (full) SVM model as more genetic profiles were added. Using a similar strategy and by combining SVM with other classification methods such as genetic algorithms, several authors have been able to build accurate predictive models that, in the near future, could be used to diagnose patients in the clinical setting. Some examples include the work by David & Lerner (2005) in genetic syndrome diagnosis, and Furey, Cristianini, Duffy, Bednarski, Schummer & Haussler (2000), Peng, Xum, Bruce Ling, Peng, Du & Chen (2003), and Li, Jiang, Li, Moser, Guo, Du, Wang, Topol, Wang & Rao (2005) in cancer.

SVM models have shown to be highly accurate when cancer diagnosis is of interest and either microarray expression data (Furey et al. 2000, Noble 2006) or tumor marker detection (TMD) results for colorectal, gastric and lung cancer (Wang & Huan 2011) are available. For instance, Furey et al. (2000) used 6817 gene expression measurements and fitted a SVM model that achieved near-perfect classification accuracy on the ALL/AML data set (Golub, Slonim, Tamayo, Huard,

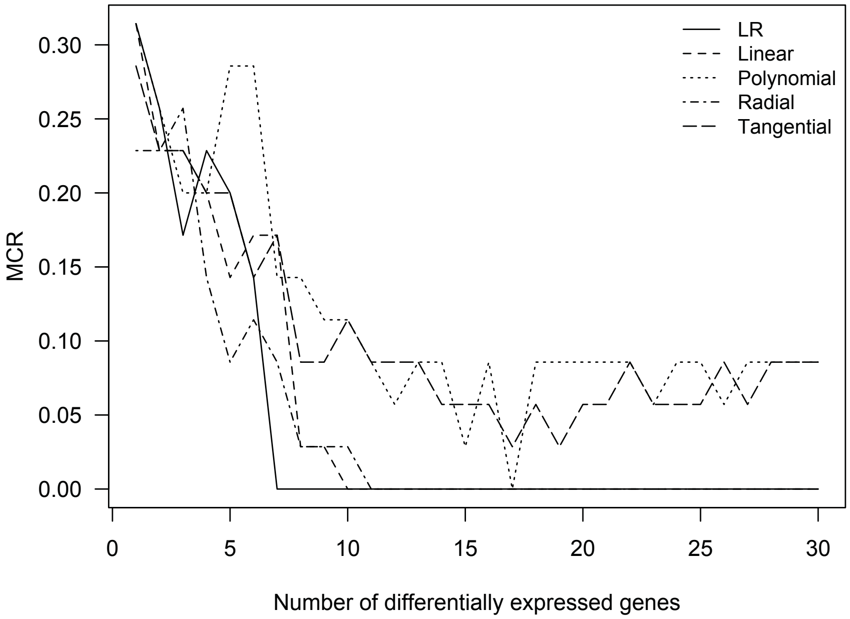


Figure 7: MCR as a function of the number of differentially expressed genes.

Gaasenbeek, Mesirov, Coller, Loh, Downing, Caligiuri, Bloomfield & Lander 1999). For TMD, Wang & Huan (2011) created, trained, optimized and validated SVM models that resulted to be highly accurate compared to others, indicating a potential application of the method as a diagnostic model in cancer. Similarly, Peng et al. (2003) combined genetic algorithms and paired SVM for multiclass cancer identification to narrow a set of genes to a very compact cancer-related predictive gene set; this method outperformed others previously published.

6. Conclusions

We have presented a framework to compare, by statistical simulation, the performance of several classification methods when individuals belong to one of two mutually exclusive categories. As a test case, we compared SVM and LR.

When it is of interest to predict the group to which a *new* observation belongs to based on a single variable, SVM models are a feasible alternative to RL. However, as shown for the Poisson, Exponential and Normal distributions, the polynomial SVM model is not recommended since its MCR is higher.

In the case of multivariate and mixture of distributions, SVM performs better than LR when high correlation structures are observed in the data (as shown in Figure 6). Furthermore, SVM methods required less variables than LR to achieve a better (or equivalent) MCR. This latter result is consistent with Verplancke et al. (2008).

Further work includes the evaluation of the MCR of SVM and LR methods for other probability distributions, different variance-covariance matrices among groups, and high-dimensional (non) correlated data with less variables than observations, e.g., genetic data with up to 5 million genotypes and ~ 1000 cases and controls.

Acknowledgements

DAS and JCS were supported by Programa Nacional de Investigación en Genómica, Bioinformática y Estadística para Estudios en Enfermedades Neurosiquiátricas. Fase I: Enfermedad de Alzheimer, código 9373, Grupo de Neurociencias de la Universidad Nacional de Colombia, Sede Bogotá. We thank two anonymous reviewers for their helpful comments and suggestions.

[Recibido: septiembre de 2011 — Aceptado: febrero de 2012]

References

Anderson, T. (1984), *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York.

Asparoukhova, K. & Krzanowskib, J. (2001), 'A comparison of discriminant procedures for binary variables', *Computational Statistics & Data Analysis* **38**, 139–160.

Cornfield, J. (1962), 'Joint dependence of the risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis', *Proceedings of the Federal American Society of Experimental Biology* **21**, 58–61.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.

Cover, T. M. (1965), 'Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition', *IEEE Transactions on Electronic Computers* **14**, 326–334.

Cox, D. (1966), *Some Procedures Associated with the Logistic Qualitative Response Curve*, John Wiley & Sons, New York.

David, A. & Lerner, B. (2005), 'Support vector machine-based image classification for genetic syndrome diagnosis', *Pattern Recognition Letters* **26**, 1029–1038.

Day, N. & Kerridge, D. (1967), 'A general maximum likelihood discriminant', *Biometrics* **23**, 313–323.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2011), *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.* R package version 1.5-27.
*http://CRAN.R-project.org/package=e1071

Fisher, R. (1936), 'The use of multiple measurements in taxonomic problems', *Annual Eugenics* **7**, 179–188.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. & Haussler, D. (2000), 'Support vector machine classification and validation of cancer tissue samples using microarray expression data', *Bioinformatics* **16**(10), 906–914.

Gentleman, R., Carey, V., Huber, W. & Hahne, F. (2011), *Genefilter: Methods for filtering genes from microarray experiments.* R package version 1.34.0.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Hernández, F. & Correa, J. C. (2009), 'Comparación entre tres técnicas de clasificación', *Revista Colombiana de Estadística* **32**(2), 247–265.

Hosmer, D. & Lemeshow, S. (1989), *Applied Logistic Regression*, John Wiley & Sons, New York.

Karatzoglou, A., Meyer, D. & Hornik, K. (2006), 'Support vector machines in R', *Journal of Statistical Software* **15**(8), 267–73.

Lee, J. B., Park, M. & Song, H. S. (2005), 'An extensive comparison of recent classification tools applied to microarray data', *Computational Statistics & Data Analysis* **48**, 869–885.

Li, L., Jiang, W., Li, X., Moser, K. L., Guo, Z., Du, L., Wang, Q., Topol, E. J., Wang, Q. & Rao, S. (2005), 'A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset', *Genomics* **85**(1), 16–23.

Moguerza, J. & Muñoz, A. (2006), 'Vector machines with applications', *Statistical Science* **21**(3), 322–336.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrele, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. (2003), 'Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes', *Nature Genetics* **34**(3), 267–73.

Noble, W. (2006), ‘What is a support vector machine?’, *Nature Biotechnology* **24**(12), 1565–1567.

Peng, S., Xum, Q., Bruce Ling, X., Peng, X., Du, W. & Chen, L. (2003), ‘Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines’, *FEBS Letters* **555**, 358 – 362.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org/

Salazar, D. (2012), Comparación de Máquinas de Soporte vectorial vs. Regresión Logística: cuál es más recomendable para discriminar?, Tesis de Maestría, Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín.

Shou, T., Hsiao, Y. & Huang, Y. (2009), ‘Comparative analysis of logistic regression, support vector machine and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power doppler’, *Korean Journal of Radiology* **10**, 464–471.

Tibshirani, R. & Friedman, J. (2008), *The Elements of Statistical Learning*, Springer, California.

Verplancke, T., Van Looy, S., Benoit, D., Vansteelandt, S., Depuydt, P., De Turck, F. & Decruyenaere, J. (2008), ‘Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies’, *BMC Medical Informatics and Decision Making* **8**, 56–64.

Wang, G. & Huan, G. (2011), ‘Application of support vector machine in cancer diagnosis’, *Med. Oncol.* **28**(1), 613–618.

Westreich, D., Lessler, J. & Jonsson, M. (2010), ‘Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression’, *Journal of Clinical Epidemiology* **63**, 826–833.

Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 239 a 254

¿Cuándo inicia la enfermedad de Alzheimer? Kaplan-Meier versus Turnbull: una aplicación a datos con censura arbitraria

¿When does Alzheimer's Disease Begin? Kaplan-Meier versus Turnbull: An Application to Arbitrary Censoring Data

Carlos Mario Lopera-Gómez[1,a], Mario César Jaramillo-Elorza[1,b], Natalia Acosta-Baena[2,c]

[1] Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

[2] Grupo de Neurociencias de Antioquia y Grupo Académico de Epidemiología Clínica-GRAEPIC, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia

Resumen

La mayoría de los análisis de supervivencia se basan en tiempos de falla exactos y observaciones censuradas a la derecha, utilizándose métodos ampliamente difundidos como el método de Kaplan-Meier (KM). Para estimar la edad de inicio de la Enfermedad de Alzheimer (EA) familiar cuando las censuras son arbitrarias (censura a derecha, a izquierda o en intervalo), ¿cuál es el cambio en los resultados clínicos, si se utiliza el método de KM mediante imputación comparado con el método de Turnbull sugerido para este tipo de datos?

El método de Turnbull se comparó con el método de KM mediante un estudio de simulación y una aplicación con datos reales. Se realizó KM con imputación a través del punto medio del intervalo (PM) y en el extremo derecho (ED). Se analizaron diferentes tamaños de muestra y diferentes tiempos entre visitas.

En todos los escenarios de simulación, las funciones que fueron estimadas, usando imputación de datos, difieren significativamente de la verdadera función de supervivencia $S(t)$.

La edad de inicio de la EA determinada a través de un método de imputación tiene implicaciones clínicas relevantes que afectarían la toma de decisiones a la hora de iniciar una terapia preventiva. El método de Turnbull

[a] Profesor asistente. E-mail: cmlopera@unal.edu.co
[b] Profesor asociado. E-mail: mjarami@unal.edu.co
[c] Profesora. E-mail: natalia.acosta@neurociencias.udea.edu.co

presenta un menor sesgo cuando se necesita realizar un análisis de supervivencia con censuras arbitrarias.

***Palabras clave*:** análisis de supervivencia, censura de intervalo, edad de inicio, enfermedad de Alzheimer familiar.

Abstract

Most of the survival analysis are based on exact failure times and right censored observations, using methods widely known as the Kaplan-Meier (KM). To estimate the onset age of familial Alzheimer's Disease (AD) when the censor times are arbitrary (right, left or interval censor), what is the change in clinical outcomes, using the KM method with data imputation compared with procedure proposed by Turnbull for this kind of data?

Turnbull's method was compared with KM method in a simulation study and an application with real data. KM method was based on data imputation through the midpoint of the interval (MP) and the right side of the interval (RS), considering several sample sizes and different times between visits.

In all simulation scenarios estimated functions using data imputation differ significantly from the actual simulated survival function $S(t)$.

The estimated onset age of AD through data imputation methods has relevant clinical implications that would affect decision-making in initiating preventive therapy. Turnbull's method has fewer bias when was compared with KM with imputation to perform a survival analysis with arbitrary censure data.

***Key words*:** Age of onset, Familial Alzheimer's disease, Interval censoring, Survival analysis.

1. Introducción

El análisis de supervivencia es un conjunto de procedimientos estadísticos para el análisis de datos en los que la variable de resultado es el tiempo hasta que ocurre un evento de interés. La función de supervivencia es quizás la función más importante en los estudios de medicina y salud. Como es usual en el análisis de datos de supervivencia, es de interés estimar la función de supervivencia $S(t)$ y evaluar la importancia de factores potenciales de pronóstico o características individuales sobre este tiempo de supervivencia.

La gran cantidad de estudios epidemiológicos realizados en enfermedades como el cáncer, entre muchas otras, y la cantidad de estudios longitudinales con desenlaces que involucran el tiempo demuestran la importancia del análisis de supervivencia. Alternativamente al desenlace de supervivencia o tiempo hasta la muerte, el tiempo puede hacer referencia al momento en que una persona presenta cualquier otro evento. Si el evento se presenta en todos los individuos, se podrían aplicar muchos métodos. Sin embargo, lo habitual es que al final del seguimiento algunas de las personas no han desarrollado el evento de interés, por lo que el verdadero tiempo trascurrido hasta el evento es no observado. Además, los datos de supervivencia rara vez se distribuyen de forma "normal", y se componen generalmente de muchos eventos al inicio del seguimiento, y los eventos tardíos son relativamente pocos. Estas características de los datos son las que hacen necesario un método especial como el análisis de supervivencia.

Las dificultades específicas relacionadas con el análisis de supervivencia surgen en gran medida por el hecho de que sólo algunas personas han experimentado el evento; por lo tanto, el tiempo de supervivencia se desconoce en un subconjunto de sujetos del estudio. Este fenómeno se llama censura y sus mecanismos pueden deberse a que el individuo no ha experimentado el desenlace en el momento de cierre del estudio; porque se pierde del seguimiento: o porque el sujeto presenta un evento diferente que hace imposible un seguimiento posterior (riesgo competitivo). En este último caso, las censuras deben estimarse de manera distinta y requiere un análisis especial de los datos. Pero al visualizar el proceso de supervivencia de un individuo como una línea de tiempo pueden verse tres tipos de censuras: si el evento (suponiendo que llegara a ocurrir) está más allá del final del período de seguimiento, esta situación se conoce como censura a derecha. Otro caso se presenta cuando se observa el evento de interés antes de la primera evaluación, pero no se sabe exactamente cuándo ocurrió. Este tipo de censura es la censura a izquierda. Y por último, el tiempo trascurrido hasta el evento también puede ser censurado en intervalo; cuando los individuos salen y entran del seguimiento (por ejemplo, cuando los individuos se presentan a controles médicos con cierta frecuencia), el individuo presenta el evento de interés al regreso del seguimiento pero la única información que se tiene en este caso, es que el evento se produce dentro de un intervalo de tiempo dado.

La mayoría de los datos de supervivencia incluyen solo observaciones censuradas a derecha y tiempos de falla exactos, utilizándose métodos ampliamente difundidos como el método de Kaplan-Meier (KM), pruebas de logrank y regresión de Cox (análisis de riesgos proporcionales). Sin embargo, los métodos que soportan datos censurados a izquierda o en intervalo no son tan conocidos. Pocos paquetes estadísticos permiten estos datos, y por esta razón, la práctica común entre los investigadores consiste en simplemente ignorar y descartar las censuras a izquierda de los datos, o realizar una imputación del desenlace para las censuras de intervalo. Es decir, asumir que el evento que ha ocurrido dentro del intervalo $(L_i, U_i]$ ocurrió ya sea en el límite inferior o superior del intervalo o en el punto medio del mismo. Autores como Rucker & Messerer (1988), Odell, Anderson & D'agostinho (1992), Dorey, Little & Schenker (1993) y Iceland (1997) manifiestan que asumir el tiempo de supervivencia de intervalo como si fuera exacto puede conducir a estimadores sesgados, así como a conclusiones y estimaciones parciales que no son completamente fidedignas. Estas afirmaciones motivan, de alguna manera, a propuestas distintas relacionadas con el tratamiento que se debe dar a estas censuras, con el fin de evitar sesgos y que se incorpore mayor información.

Los datos de la Cohorte Antioquia-E280A de 15 años de seguimiento, con sujetos en riesgo de enfermedad de Alzheimer familiar, incluyen los tres tipos de censuras mencionadas previamente. Conocer la edad de inicio de la enfermedad en estos sujetos que inevitablemente van a desarrollar la Enfermedad de Alzheimer (EA) exige métodos alternativos para dicha estimación. En este estudio se pretende difundir tales métodos e ilustrar qué tan erróneas serían las estimaciones en la edad de inicio de la EA, utilizando el método KM comparado con el método de Turnbull para estimación bajo censura arbitraria (Peto 1973, Turnbull 1974, Turnbull 1976). También interesa determinar las implicaciones desde el punto de vista clínico cuando se incurre en un sesgo de medición y la importancia de los resultados para el diagnóstico del inicio de la EA familiar. Inicialmente se realiza un estudio de simulación y posteriormente la aplicación con los datos reales.

En la sección 2 se presenta el problema clínico y la base de datos que servirá para ilustrar los métodos que van a ser comparados. Los métodos estadísticos utilizados y el planteamiento de un estudio de simulación son presentados en la sección 3. La sección 4 recopila los resultados obtenidos a través del estudio de simulación y presenta la aplicación con los datos de enfermedad de Alzheimer. Finalmente, en la sección 5 se dan algunas conclusiones y recomendaciones con base en los hallazgos encontrados.

2. Problema y datos de enfermedad de Alzheimer

Conocer el tiempo hasta el inicio de la enfermedad de Alzheimer sólo es posible gracias a las formas genéticas de la enfermedad, con herencia autosómica dominante. En esta condición, todos los sujetos nacen portando una mutación que predispone a la enfermedad, expresándose en algún momento de la vida. Las manifestaciones consisten en quejas de memoria y deterioro cognitivo evidente en las evaluaciones neurosicológicas alrededor de los 50 años de edad. Conocer la edad más aproximada del inicio de la enfermedad es el primer paso para planear y desarrollar nuevos estudios en busca de terapias preventivas. El Grupo de Neurociencias de la Universidad de Antioquia ha seguido desde 1995 a este conglomerado poblacional, que es el más numeroso del mundo, con 5000 sujetos estimados, con riesgo de desarrollar EA genético mutación E280A en Presenilina 1 (PSEN1). Se identificaron, hasta enero del 2010, 1784 sujetos pertenecientes a 25 familias afectadas. Se detectaron 449 sujetos portadores de la mutación E280A-PSEN1. Los datos de estos últimos sujetos portadores fueron los utilizados para detectar el inicio de la enfermedad de manera retrospectiva (Acosta-Baena, Sepúlveda-Falla, Lopera-Gómez, Jaramillo-Elorza, Moreno, Aguirre-Acevedo, Saldarriaga & Lopera 2011).

3. Métodos

Con base en los datos descritos en la sección 2, se utilizaron los métodos de imputación, para comparar la función de supervivencia de Turnbull con KM.

Para medir la edad de inicio de la enfermedad, se realizó un análisis de supervivencia evaluando el tiempo transcurrido desde la fecha de nacimiento hasta la fecha de aparición del deterioro cognitivo leve o hasta la fecha de la última evaluación. Se utilizó el método de supervivencia desarrollado por Peto (1973), Turnbull (1974) y Turnbull (1976), que incluye los tres tipos de censuras, mediante el algoritmo implementado por Giolo (2004), para el software R versión 2.13.1 (R Development Core Team 2011). El código utilizado hace uso de la librería SURVIVAL del software R, y está disponible bajo pedido a los autores.

3.1. Estimador no paramétrico de Turnbull

En los estudios longitudinales, donde los individuos son monitoreados durante un lapso de tiempo prefijado, o visitados periódicamente un cierto número de veces, el tiempo $T_i,\ i = 1, \dots, n$, hasta que ocurre el evento de interés para cada individuo, se desconoce. Sólo se sabe que está dentro de un intervalo entre dos visitas, es decir, entre la visita en el tiempo L_i y la visita en el tiempo U_i con $L_i < T_i \leq U_i$. Si el evento ocurre exactamente en el momento de una visita, lo cual es muy poco probable, pero puede ocurrir, se tiene un tiempo de supervivencia exacto. En este caso se asume que $L_i = T_i = U_i$.

Por otra parte, se sabe que para los individuos cuyos tiempos están censurados a derecha, el evento de interés no ha ocurrido hasta la última visita, pero puede ocurrir en cualquier instante desde ese momento en adelante. Por consiguiente, se supone en este caso que T_i puede ocurrir dentro del intervalo $(L_i, +\infty)$, con L_i igual al periodo desde el comienzo del estudio hasta la última visita y $U_i = +\infty$.

De modo semejante, para los individuos cuyos tiempos están censurados a izquierda, se sabe que el evento de interés ha ocurrido antes de la primera visita, y, por lo tanto, suponemos que T_i ha ocurrido en el intervalo $(0, U_i]$, con $L_i = 0$ representando el comienzo del estudio, y U_i es el tiempo hasta la primera visita. El método de Turnbull generaliza cualquier situación con combinaciones de tiempos de supervivencia (exacto o intervalo) y censuras a izquierda y derecha como datos de supervivencia de intervalo. Por lo tanto, los tiempos de supervivencia exacta, así como datos de censura a izquierda y derecha, son todos casos especiales de datos de supervivencia con censura de intervalo con $L_i = U_i$ para censuras exactas, $U_i = +\infty$ para las censuras a derecha y $L_i = 0$ para censuras a izquierda.

Como uno de los objetivos principales en análisis de supervivencia es estimar la función de supervivencia e investigar la importancia de factores potenciales de pronóstico bajo tiempos de supervivencia con censura a intervalo, el número de factores bajo estudio debería depender del propósito del estudio. Como lo sugiere Hougaard (1999), la estimación no paramétrica de la función de distribución acumulada $F(t)$, o en su defecto de la función de supervivencia $S(t)$, es preferible a su estimación paramétrica, por varias razones. Por ejemplo, una elección equivocada de la distribución paramétrica de T podría conducir a conclusiones erróneas de $S(t)$. Además, podría ser difícil encontrar una distribución paramétrica apropiada para ajustar los datos. Hougaard da el ejemplo de tiempos de vida de una población cuya función hazard muestra la llamada forma de bañera: la cual en un principio decrece pocos años, luego permanece constante durante muchos años y por último empieza a aumentar. En este caso, el mejor ajuste probablemente se obtendría de una mezcla de distribuciones.

En el caso de censura a derecha, se podría usar el estimador de Kaplan-Meier para obtener $S(t)$ (Kaplan & Meier 1958). Sin embargo, con datos censurados en intervalo, el método de Kaplan-Meier no puede ser aplicado, y han sido Peto (1973), Turnbull (1974) y Turnbull (1976) quienes han desarrollado el estimador no paramétrico de máxima verosimilitud (NPMLE, por su sigla en inglés) para estos datos.

El estimador de Turnbull, se basa en una muestra de intervalos observados $[L_i, R_i]\,,\ i = 1, \dots, n$, los cuales contienen las variables aleatorias independientes $T_1, \dots, T_n$. Como se mencionó antes, una observación exacta de T_i se da sólo si $L_i = R_i$.

Dado este ejemplo, la función de verosimilitud a ser maximizada es la siguiente:

$$L\,(F) = \prod_{i=1}^{n} \left[F\,(R_i{+}) - F\,(L_i{-})\right] \tag{1}$$

Para resolver este problema de maximización, Peto (1973) define dos conjuntos $\gamma = \{L_i,\ i = 1, \dots, n\}$ y $\kappa = \{R_i,\ i = 1, \dots, n\}$ que contienen los extremos izquierdos y derechos de los intervalos, respectivamente. Si se denotan los incrementos de la función F dentro de los intervalos $[q_j, p_j]$ como $s_j, j = 1, \dots, m$, entonces $L\,(F)$ debe ser maximizada como una función de $s_1, s_2, \dots, s_m$, sujeto a las restricciones $s_j \geq 0$ y $s_m = 1 - \sum_{j=1}^{m-1} s_j$. Peto aborda este problema de maximización usando el algoritmo de Newton-Raphson.

Se puede probar que una función que maximice (1) es constante entre los intervalos $[q_j, p_j]$ e indefinida dentro de ellos. Note que esto implica que $\widehat{P}(T \in (p_{j-1}, q_j)) = 0$ para cualquier j. Como la función de distribución es no decreciente, la cual no es constante entre los intervalos, puede no maximizar a $L\,(F)$. Denote los incrementos de F dentro de los intervalos $[q_j, p_j]$ por $s_j\ j = 1, \dots, m$, $L\,(F)$ debe ser maximizada como una función de $s_1, s_2, \dots, s_m$ sujeto a $s_j \geq 0$ y $s_m = 1 - \sum\limits_{j=1}^{m-1} s_j$. Peto aborda este problema de maximización usando el algoritmo de Newton-Raphson. En contraste con Peto, Turnbull (1976) propone el uso del algoritmo de auto-consistencia para el mismo problema de maximización. La idea de este algoritmo fue presentada primero por Efron (1967), y su aplicación para la maximización en (1) es como sigue.

Sea $\alpha_{ij} = I_{\{[q_j, p_j] \in [L_i, R_i]\}},\ i = 1, \dots, n,\ j = 1, \dots, m$, las variables indicadoras que confirman si el intervalo $[q_j, p_j]$ se encuentra dentro o no del intervalo $[L_i, R_i]$; entonces, la probabilidad de que T_i se encuentre dentro del intervalo $[q_j, p_j]$ dado un vector $s = (s_1, s_2, \dots, s_m)'$ está dada por:

$$\mu_{ij}\,(s) = \frac{\alpha_{ij} s_j}{\sum\limits_{k=1}^{m} \alpha_{ik} s_k} \tag{2}$$

puesto que $\widehat{F}$ es constante fuera de los intervalos $[q_j, p_j]$. La proporción de observaciones en el intervalo $[q_j, p_j]$ es igual a:

$$\pi_j\,(s) = \frac{1}{n} \sum_{i=1}^{n} \mu_{ij}\,(s) \tag{3}$$

y un vector $s = (s_1, s_2, \dots, s_m)'$ es llamado auto-consistente, si

$$s_j = \pi_j\,(s)\,,\ j = 1, \dots, m$$

Siguiendo esta definición, el algoritmo de auto-consistencia de Turnbull para el cálculo del estimador no paramétrico de $F(t)$ se puede implementar siguiendo estos pasos:

1. Obtenga estimaciones iniciales de $\mathbf{s}$; por ejemplo, $s_j^{(0)} = \frac{1}{m},\ j = 1, \ldots, m$.
2. Para $i = 1, \ldots, n,\ j = 1, \ldots, m$, calcule $\mu_{ij}\left(\mathbf{s}^{(0)}\right)$ acorde a (2), y luego $\pi_j\left(\mathbf{s}^{(0)}\right)$ de acuerdo a (3).
3. Obtenga estimaciones mejoradas para $\mathbf{s}$ hallando $s_j^{(1)} = \pi_j\left(\mathbf{s}^{(0)}\right)$.
4. Retorne al paso 2, reemplazando $\mathbf{s}^{(0)}$ por $\mathbf{s}^{(1)}$ y continúe hasta que se logre la convergencia.

3.2. Estudio de simulación

Para establecer el efecto de la imputación de fallas exactas cuando en realidad se tiene una censura a intervalo, sobre la estimación de la función de supervivencia se utilizarán datos de falla lognormales con parámetros fijos para la simulación en valores $\mu = 3.78419$ y $\sigma = 0.133$, que se escogieron de tal forma que se emulan las condiciones de falla de los individuos presentes en el estudio de EA descrito en la sección 2 (tales valores son una estimación paramétrica de datos de fallas exactas generados de la función de supervivencia estimada mediante Turnbull, con el método de la transformación inversa de probabilidad integral; Kalbfleisch (1985)).

Se asume un punto de partida aleatorio para que el individuo comience sus visitas al estudio, en donde se registrará si éste tiene o no el evento. Así, se construyen intervalos de tiempo de una de las siguientes formas:

- $(0, U_i]$ un individuo llegó al estudio en el tiempo U_i pero ya tenía el evento de interés (esto constituye una censura a izquierda, la cual se puede ver como una censura a intervalo),
- $(L_i, U_i]$ un individuo llegó al estudio y asistió a visitas regulares, y en el tiempo L_i fue la última visita en la cual no tenía el evento pero al volver en la siguiente visita (al tiempo $U_i = L_i + \text{TEV}$, con TEV: el tiempo entre visitas) el individuo ya tiene el evento de interés (esto también constituye una censura a intervalo), y
- $(L_i, +\infty)$ un individuo llegó al estudio, asistió a varias visitas regulares, y en el tiempo L_i fue la última visita de la que se tiene registro del individuo en el estudio, sin que éste haya presentado el evento (esto constituye una censura a derecha).

Con este esquema de datos, no se tienen tiempos de falla exactos (aunque también las fallas exactas se pueden considerar como censuras a intervalo con $L_i = U_i$) y todos los datos deben entrarse al análisis como intervalos de tiempo.

Los factores de simulación que se van a variar son:

1. Método de imputación (MI): de acuerdo a la literatura se estudiarán los casos en que las censuras de intervalo son imputadas a través del punto medio del intervalo (PM) y utilizando el extremo derecho del mismo (ED). Lo cual lleva a tiempos de falla "exactos" y facilita los análisis, ya que la estimación de Kaplan-Meier (KM) para la curva de supervivencia puede ser estimada. Además, se considera el caso en que ninguna imputación es llevada a cabo (NI), es decir, usando los datos en forma de intervalos de tiempo, lo cual necesariamente lleva a utilizar el estimador de Turnbull (TB) para la función de supervivencia que tiene en cuenta censura arbitraria.
2. Tiempo entre visitas (TEV): indica con qué frecuencia los individuos asisten a los controles en el estudio. Interesan valores de TEV $= 1, 2, 4$ y 6 años.
3. Tamaño de la muestra (n): este factor tiene como objetivo establecer el efecto sobre el proceso de estimación del número de individuos en el estudio. Se tomarán valores de $n = 50, 100, 200, 500$.

Se utilizará como control para comparar el desempeño de las estimaciones el estimador KM, bajo los métodos de imputación "$\widehat{S}(t)_{\text{PM}}$" y "$\widehat{S}(t)_{\text{ED}}$", y el estimador de Turnbull "$\widehat{S}(t)_{\text{TB}}$", a la función de supervivencia real, notada "$S(t)$". Esto permite, a través de las diferencias observadas entre cada una de las curvas "$\widehat{S}(t)_{\text{TB}}$", "$\widehat{S}(t)_{\text{PM}}$" y "$\widehat{S}(t)_{\text{ED}}$", y la curva de supervivencia de referencia "$S(t)$", establecer el efecto de la imputación sobre la estimación.

Para comparar las curvas de supervivencia resultantes de la simulación, se generan $N = 1000$ muestras independientes para cada uno de los 16 escenarios de simulación (resultantes de las combinaciones de los niveles de los factores TEV y n). Luego, en cada escenario se realizan las estimaciones de la función de supervivencia, de acuerdo al factor de imputación: $\widehat{S}(t)_{\text{PM}}$, $\widehat{S}(t)_{\text{ED}}$ y $\widehat{S}(t)_{\text{TB}}$, y se comparan con la función de supervivencia de control $S(t)$. Tal comparación se realiza usando el error cuadrático medio integrado (ECMI) como una medida global de error. Para calcular el ECMI con $N = 1000$ simulaciones en cada escenario, se utiliza la siguiente fórmula:

$$\text{ECMI}_i = \frac{1}{N}\sum_{j=1}^{N}\int\left[\widehat{S}_j\left(t\right)_i - S\left(t\right)\right]^2 dt$$

donde $i = \text{TB}, \text{PM}, \text{ED}$ representa el método de estimación de la función de supervivencia y $S(t)$ es la función de supervivencia real.

Adicionalmente, para establecer dónde se dan las diferencias entre las curvas de supervivencias estimadas con la real, se calculó el error cuadrático medio (ECM) en la estimación de los cuantiles $q_{0.05}$, $q_{0.1}$, $q_{0.25}$, $q_{0.5}$, $q_{0.75}$, $q_{0.9}$, $q_{0.95}$, de manera que se establece el correspondiente sesgo de estimación de los métodos estudiados (TB, ED y PM). El ECM se calculó para $i = \text{TB}, \text{PM}, \text{ED}$ y $h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ como:

$$\text{ECM}_{i,h} = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{q}_{h,i,j} - q_h\right)^2$$

donde $\widehat{q}_{h,i,j}$ son $N = 1000$ estimaciones en cada uno de los métodos estudiados $i = \text{TB}, \text{PM}, \text{ED}$ de los cuantiles reales q_h, $h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ de la distribución lognormal con parámetros $\mu = 3.78419$ y $\sigma = 0.133$.

4. Resultados

4.1. Estudio de simulación

4.1.1. Diferencias en las funciones de supervivencia

La medida de error que se utiliza para comparar las estimaciones basadas en imputación $\widehat{S}(t)_{\text{PM}}$, $\widehat{S}(t)_{\text{ED}}$ y $\widehat{S}(t)_{\text{TB}}$, con la función de supervivencia verdadera $S(t)$, es el ECMI definido en la sección anterior. Un valor pequeño del ECMI indica que el método de estimación correspondiente produce una curva de supervivencia estimada que es muy cercana a la curva de supervivencia real a lo largo del tiempo; por el contrario, valores altos del ECMI indican que las curvas comparadas tienen diferencias a lo largo del tiempo.

La tabla 1 muestra los ECMI obtenidos en cada uno de los 16 escenarios de simulación considerados.

Tabla 1: ECMI estimado con los métodos TB, PM y ED.

n	TEV	ECMI$_{TB}$	ECMI$_{ED}$	ECMI$_{PM}$
50	1	3.04	42.11	32.51
50	2	2.24	30.25	32.04
50	4	2.50	29.69	31.68
50	6	2.95	34.16	32.06
100	1	1.58	41.45	31.85
100	2	1.18	28.82	31.30
100	4	1.35	29.28	31.53
100	6	1.64	33.21	31.24
200	1	0.87	41.17	30.88
200	2	0.71	28.82	30.63
200	4	0.76	28.64	30.86
200	6	0.96	33.15	30.79
500	1	0.45	40.73	30.78
500	2	0.38	28.31	30.52
500	4	0.42	28.36	30.53
500	6	0.50	32.74	30.62

En todos los escenarios de simulación el ECMI muestra que las funciones estimadas $\widehat{S}(t)_{\text{PM}}$ y $\widehat{S}(t)_{\text{ED}}$ difieren significativamente de $S(t)$, lo cual indica que las estimaciones basadas en estas curvas pueden estar muy alejadas de la realidad. Por otro lado, el ECMI asociado a la estimación de Turnbull ($\widehat{S}(t)_{\text{TB}}$) tiene los valores más pequeños en todos los escenarios, lo cual sucede sin importar el tamaño de muestra. Sin embargo, a medida que el tamaño de muestra aumenta, este error disminuye su valor. En el análisis del tiempo entre visitas (TEV) se puede observar que hay un patrón consistente en todos los valores del tamaño de muestra considerados, que indica que TEV= 2 años provoca un ECMI menor que en los demás valores de TEV.

La figura 1 ilustra uno de los escenarios considerados en el estudio de simulación (n = 500, TEV = 2), donde claramente se observan diferencias entre las curvas de supervivencia estimadas usando los diferentes métodos de imputación y la supervivencia real, mientras que la supervivencia estimada mediante Turnbull se ajusta bien a esta última.

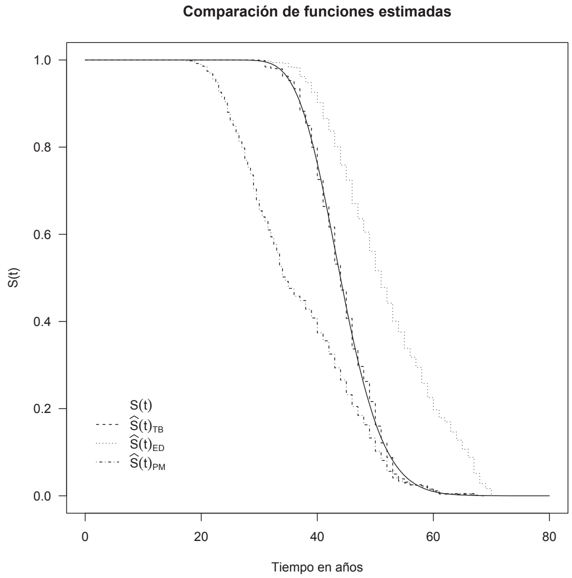


FIGURA 1: Diferencias entre la curva real y las curvas estimadas mediante Turnbull y KM(ED) y KM(PM). Una realización del caso simulado con $n = 500$ y TEV = 2.

4.1.2. Diferencias en las edades de inicio

Para el caso ilustrado en la figura 1, se estimaron la edad de inicio y sus respectivos límites de confianza en cada una de las curvas de supervivencia estimadas, mediante el método bootstrap percentil, lo cual se resume en la tabla 2. Detalles del proceso de estimación bootstrap se encuentran en Acosta-Baena et al. (2011), Meeker & Escobar (1998).

Observe que las edades estimadas de inicio de la EA obtenidas por imputación de datos (PM y ED) difieren significativamente del valor de referencia, mientras que el método de Turnbull estima bien. Esto se repite en todos los demás escenarios considerados.

Tabla 2: Estimaciones de la edad de inicio para datos simulados.

	Mediana	LI95 %	LS95 %
Referencia	44.00000	–	–
TB	44.00006	43.00002	44.99997
KM(ED)	51.00004	51.00002	52.00000
KM(PM)	34.00003	33.99999	38.99993

4.1.3. Sesgos de estimación de algunos cuantiles

A continuación se presentan los ECM calculados en los métodos estudiados.

Tabla 3: ECM para las estimaciones de los cuantiles q_h, $h = 0.05$, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, usando el método TB.

n	TEV	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
50	1	7.34	5.76	4.67	3.96	5.71	9.92	12.96
50	2	6.30	4.54	3.13	2.96	3.72	6.55	9.24
50	4	7.56	5.43	3.72	3.13	3.42	5.86	7.84
50	6	9.24	5.86	4.45	3.96	4.37	6.76	9.67
100	1	3.88	2.92	2.10	2.10	2.66	5.06	7.51
100	2	3.57	2.56	1.61	1.44	1.93	3.42	5.57
100	4	4.16	3.10	2.10	1.66	2.02	3.13	4.84
100	6	5.38	3.65	2.59	2.19	2.37	3.69	5.06
200	1	2.16	1.64	1.14	1.32	1.56	2.59	4.28
200	2	1.99	1.32	0.98	0.98	1.25	1.80	2.89
200	4	2.53	1.72	1.12	1.06	1.25	1.77	2.31
200	6	2.96	2.10	1.49	1.32	1.46	2.07	2.86
500	1	0.94	0.77	0.61	0.66	0.85	1.25	1.72
500	2	0.86	0.74	0.58	0.59	0.59	0.92	1.32
500	4	1.19	0.86	0.64	0.62	0.67	0.94	1.28
500	6	1.44	1.08	0.77	0.69	0.72	1.04	1.37

Note que los sesgos de estimación al utilizar el método TB (tabla 3) son menores que los obtenidos con los métodos de imputación PM y ED (tablas 4 y 5, respectivamente). En particular, los sesgos de estimación asociados al método de imputación PM (tabla 4) son mayores en los cuantiles más pequeños, mientras que para el método de imputación ED (tabla 5) los sesgos mayores se presentan en los cuantiles más grandes.

Ahora, en general (tablas 3, 4, y 5) observe que a medida que el tamaño de muestra aumenta, los sesgos medidos con el ECM disminuyen, y que los resultados señalan que el tiempo óptimo entre visitas sería de dos años, ya que en este caso los ECM resultaron menores que en los demás valores de este factor.

4.2. Aplicación con datos reales

Para los datos de EA, se aplicaron las diferentes técnicas de estimación de la función de supervivencia, y con base en ellas se calculó la mediana como estimador de la edad de inicio de la enfermedad.

Tabla 4: ECM para las estimaciones de los cuantiles q_h, $h = 0.05$, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, usando el método de imputación PM.

n	TEV	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
50	1	166.67	168.48	139.95	88.55	14.29	14.29	17.81
50	2	166.67	169.00	140.19	85.38	11.42	7.56	11.16
50	4	167.18	169.26	139.00	84.64	13.10	7.90	7.90
50	6	168.22	169.26	139.95	86.49	12.89	7.18	7.62
100	1	165.12	163.58	140.42	88.92	8.70	8.64	12.82
100	2	164.61	164.61	139.95	82.08	9.30	4.33	6.60
100	4	165.12	164.61	139.48	83.17	10.30	4.54	4.41
100	6	164.10	164.10	138.06	80.28	9.86	4.00	3.84
200	1	169.26	166.67	139.95	91.58	5.90	3.92	9.42
200	2	168.48	166.41	140.42	85.93	8.29	2.66	3.24
200	4	167.44	165.64	139.71	86.12	9.30	3.24	2.69
200	6	168.74	165.89	139.95	85.93	8.94	2.96	2.28
500	1	169.78	166.41	139.71	94.67	4.93	1.51	3.61
500	2	169.52	166.15	140.66	89.11	7.84	1.90	1.39
500	4	170.04	166.67	140.19	89.49	8.64	2.59	1.56
500	6	169.52	165.89	140.42	89.30	8.24	2.28	1.25

Tabla 5: ECM para las estimaciones de los cuantiles q_h, $h = 0.05$, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, usando el método de imputación ED.

n	TEV	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
50	1	23.33	27.56	45.83	102.01	180.63	199.66	176.62
50	2	16.32	16.65	27.98	58.06	135.02	177.69	159.52
50	4	22.28	23.33	32.95	54.17	114.06	157.25	155.75
50	6	27.88	29.92	40.20	64.32	121.66	160.78	156.50
100	1	18.23	22.75	42.51	97.81	183.60	205.92	178.49
100	2	15.29	16.16	26.01	56.40	133.40	181.44	166.41
100	4	20.98	22.75	31.81	53.44	111.94	163.58	158.26
100	6	26.94	30.03	39.94	62.25	117.07	164.61	158.00
200	1	16.08	21.44	40.83	94.67	181.98	207.65	181.98
200	2	13.76	15.60	24.70	54.61	133.86	184.14	170.82
200	4	19.36	21.90	30.80	53.29	111.72	165.64	160.78
200	6	25.00	29.48	39.44	63.36	116.86	167.96	162.05
500	1	15.37	20.70	39.69	94.09	183.87	209.38	183.33
500	2	13.10	15.21	24.30	53.58	132.94	184.42	171.87
500	4	18.75	21.53	30.47	52.85	111.30	167.44	163.58
500	6	24.11	28.73	39.44	62.73	117.29	169.52	163.84

La tabla 6 muestra cómo es la estimación de la edad de inicio de la enfermedad.

Tabla 6: Estimaciones de la edad de inicio para datos de EA.

	Mediana	LI95 %	LS95 %
TB	44.01006	43.01003	45.01003
KM(ED)	47.00998	46.00002	47.99997
KM(PM)	44.00499	42.00502	45.00498

Los resultados anteriores muestran que las estimaciones que usan TB y PM, estiman la edad de inicio a los 44 años, mientras que el método ED sobrestima tal valor. A nivel de intervalos de confianza, el método de Turnbull es más preciso que el método PM en la estimación de la edad de inicio de la enfermedad.

La figura 2 muestra las diferencias apreciables entre las curvas estimadas.

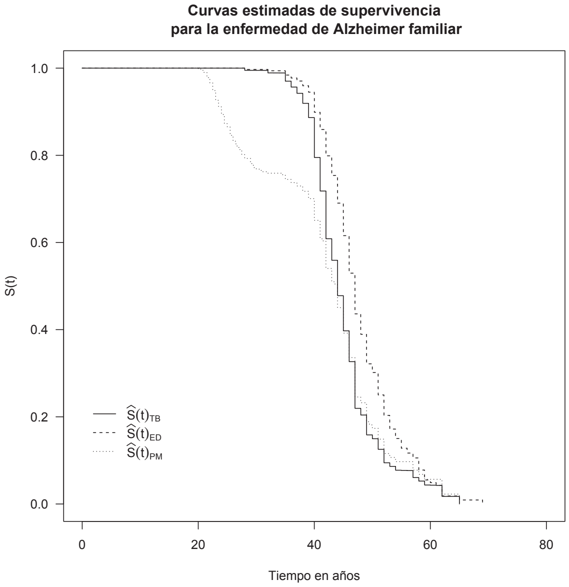


Figura 2: Funciones de supervivencia estimadas para los datos de EA.

Note que aunque las estimaciones de la edad de inicio que usan TB y PM son similares, estimaciones de otros cuantiles, particularmente cuantiles más pequeños, pueden llevar a errores apreciables. Esto puede deberse principalmente a que los datos de EA familiar incluyen un 21 % de datos con censura a izquierda (Acosta-Baena et al. 2011).

5. Conclusiones y recomendaciones

- En las últimas décadas existe gran interés en todo el mundo por definir adecuadamente el inicio de la EA, incluso etapas preclínicas y prodrómicas, con el objetivo de detectar la enfermedad de manera más temprana y ofrecer alternativas de tratamiento más oportuno (Petersen, Stevens, Ganguli, Tangalos, Cummings & DeKosky 2001, Reisberg, Ferris, Kluger, Franssen, Wegiel & de Leon 2008). Conocer adecuadamente la edad de inicio de esta cohorte de portadores de una mutación con irremediable inicio de la enfermedad de Alzheimer tiene utilidad para el diseño de ensayos clínicos dirigidos a tratamientos preventivos (Strobel 2011).

- En los análisis realizados, las edades de inicio de la EA obtenidas por imputación de datos (PM y ED) difieren significativamente de los datos reales en todos los tamaños de muestra y en los diferentes TEV, mientras que el método de Turnbull estima bien en todos los escenarios. También puede concluirse que un tiempo entre visitas igual a 2 años, independiente del tamaño de muestra, es óptimo para estimar la edad de inicio de la EA familiar, ya que en este caso se presentaron diferencias más pequeñas que las obtenidas en los escenarios restantes.
- El análisis de los resultados de la estimación de sesgos para los cuantiles q_h, $h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando los métodos TB, PM y ED (tablas 3, 4 y 5), muestra que en general el método TB presenta menores sesgos en la estimación que los métodos de imputación. También, como es de esperarse a medida que el tamaño de muestra aumenta, los sesgos medidos con el ECM disminuyen. Los resultados de la tabla 4 establecen que en general el método de imputación, usando el punto medio del intervalo, afecta la estimación de los cuantiles más pequeños, mientras que el método de imputación mediante el extremo derecho del intervalo afecta a los cuantiles más grandes (tabla 5).
- Aunque en la aplicación con datos reales se obtuvieron estimaciones de la mediana muy similares mediante Turnbull y usando la imputación PM, no se puede concluir que esto siempre va a ocurrir, de acuerdo a lo que se evidencia en el estudio de simulación. Sin embargo, el interés del investigador puede estar enfocado en otros cuantiles diferentes a la mediana, donde se podrían dar errores apreciables en la estimación, como se evidenció en la sección 4.1.3.
- De acuerdo a las edades de inicio encontradas con los métodos de imputación, el 50 % de los sujetos portadores de la mutación E280A para EA iniciará con deterioro cognitivo leve a los 47 años (según imputación por ED) o a la edad de 44 años (según imputación por TB y PM). La primera estimación, desde el punto de vista clínico, estaría retrasando un tratamiento preventivo.
- Tanto en los datos simulados como en los datos reales, los intervalos de confianza obtenidos usando TB son más estrechos que los calculados mediante KM, lo cual indica que el método de Turnbull es más preciso.
- La imputación de las censuras arbitrarias presentan grandes errores, con impacto clínicamente importante, como en el caso de esta cohorte de sujetos en riesgo de EA familiar, cuyos resultados sesgados implicarían un error en el diagnóstico, en el tratamiento y, por ende, en el pronóstico de la enfermedad.

Agradecimientos

Los autores agradecen de manera especial a los árbitros y a las editoras invitadas por sus valiosos comentarios que enriquecieron el texto.

Se agradece al CODI (Comité para el Desarrollo de la Investigación) de la Universidad de Antioquia y al programa de sostenibilidad 2010-2011.

[Recibido: septiembre de 2011 — Aceptado: febrero de 2012]

Referencias

Acosta-Baena, N., Sepúlveda-Falla, D., Lopera-Gómez, C. M., Jaramillo-Elorza, M. C., Moreno, S., Aguirre-Acevedo, D. C., Saldarriaga, A. & Lopera, F. (2011), 'Pre-dementia clinical stages in presenilin 1 E280A familial early-onset Alzheimer's disease: A retrospective cohort study', *The Lancet Neurology* **10**(3), 213–220.

Dorey, F. J., Little, R. & Schenker, N. (1993), 'Multiple imputation for threshold-crossing data with interval censoring', *Statistics in Medicine* **12**, 1589–1603.

Efron, B. (1967), 'The two sample problem with censored data', *University of California Press* pp. 831–853.

Giolo, S. R. (2004), 'Turnbull's nonparametric estimator for interval-censored data', *Department of Statistics, Federal University of Paraná* pp. 1–10. Consultado en septiembre 6, 2011.
*www.est.ufpr.br/rt/suely04a.pdf

Hougaard, P. (1999), 'Fundamentals of survival data', *Biometrics* **55**, 13–22.

Iceland, J. (1997), *The Dynamics of Poverty Spells and Issues of Left-Censoring*, PSC Research Report Series January 1997. Consultado en septiembre 6, 2011.
*http://www.psc.isr.umich.edu/pubs/pdf/rr97-378.pdf

Kalbfleisch, J. (1985), *Probability and Statistical Inference*, Vol. 1, 2nd edn, Springer-Verlag, New York.

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**(282), 457–481.

Meeker, W. & Escobar, L. (1998), *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.

Odell, P., Anderson, K. & D'agostinho, R. (1992), 'Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model', *Biometrics* **48**, 951–959.

Petersen, R. C., Stevens, J. C., Ganguli, M., Tangalos, E. G., Cummings, J. L. & DeKosky, S. T. (2001), 'Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review)', *Neurology* **56**(9), 1133–1142.

Peto, R. (1973), 'Experimental survival curves for interval-censored data', *Journal of the Royal Statistical Society, Series C* **22**, 86–91.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Consultado en septiembre 6, 2011.
*http://www.R-project.org/

Reisberg, B., Ferris, S. H., Kluger, A., Franssen, E., Wegiel, J. & de Leon, M. J. (2008), 'Mild cognitive impairment (MCI): A historical perspective', *International Psychogeriatrics* **20**(1), 18–31.

Rucker, G. & Messerer, D. (1988), 'Remission duration: An example of interval-censored observation', *Statistics in Medicine* **7**, 1139–1145.

Strobel, G. (2011), Detecting Familial AD Ever Earlier: Subtle Memory Signs 15 Years Before, *in* 'Alzheimer Research Forum'. Consultado en septiembre 6, 2011.
*http://www.alzforum.org/new/detail.asp?id=2725

Turnbull, B. W. (1974), 'Nonparametric estimation of a survivorship function with doubly censored data', *Journal of the American Statistical Association* **69**(345), 169–173.

Turnbull, B. W. (1976), 'The empirical distribution function with arbitrarily grouped censored and truncated data', *Journal of the Royal Statistical Society, Series B* **38**(3), 290–295.

Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 255 a 270

A Statistical Model for Analyzing Interdependent Complex of Plant Pathogens

Un modelo estadístico para analizar complejos interdependientes de patógenos vegetales

Eduardo Dávila[a], Luis Alberto López[b], Luis Guillermo Díaz[c]

Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

We introduce a new approach for modeling multivariate overdispersed binomial data, from a plant pathogen complex. After recalling some theoretical foundations of generalized linear models (GLMs) and Copula functions, we show how the later can be used to model correlated observations and overdispersed data. We illustrate this approach using fungal incidence in vegetables, which we analyzed using Gaussian copula with Beta-binomial margins. Compared to classical and generalized linear models, the model using Gaussian copula function best controls for overdispersion, being less prone to the underestimation of standard errors, the major cause of wrong inference in the statistical analysis of plant pathogen complex.

***Key words*:** Epidemiological methods, Extra-binomial variation, Multivariate data.

Resumen

Se introduce un nuevo enfoque para modelar datos binomiales multivariados con sobredispersión, obtenidos de complejos de patógenos vegetales. Después de revisar los conceptos básicos de los modelos lineales generalizados (GLMs) y las funciones Cópula, se muestra cómo estas últimas pueden usarse para modelar observaciones correlacionadas y datos con sobredispersión. Se ilustra el método usando la incidencia de hongos en hortalizas, analizando el caso por medio de la función cópula Gaussiana con marginales Beta-binomiales. Comparado con los modelos lineales clásicos y generalizados, el modelo construido con la cópula Gaussiana es el que mejor controla la sobredispersión, siendo menos propenso a la subestimación de los errores

[a] Ph.D. student. E-mail: jedavilas@unal.edu.co
[b] Professor. E-mail: lalopezp@unal.edu.co
[c] Professor. E-mail: lgdiazm@unal.edu.co

estándar, la causa más importante de inferencia inapropiada en el análisis estadístico de complejos de patógenos vegetales.

***Palabras clave*:** métodos epidemiológicos, variación extra-binomial, datos multivariados.

1. Introduction

The use of single-parameter family of distributions can sometimes be problematic for statistical inference (Cox 1983). For example, in the binomial distribution the variance is totally determined by the mean, and when this is satisfied there is nominal dispersion, an assumption that cannot be hold in some data analyses. In fact, vector data may display a lack of independence as is commonly the case in experimental trials in plant pathology; in these data, the presence of a fungus often increases the probability of damage in neighboring leaves, leading to marginal dependence in the data. Moreover, the analysis of plant-pathogen complex can also be complicated by the presence of multivariate dependence, as was shown by Dávila (2005).

To get a correct analysis of multivariate binomial data, an overdispersion diagnostic is necessary in order to compare the nominal dispersion against the actual dispersion. To this end, Smith & Heitjan (1993) provided an appropriate statistical tool to detect extra binomial variation. McCullagh & Nelder (1989) maintain that "overdispersion is a common attribute of data arising in many fields, and statistical practitioners shall assume that overdispersion is present in some extent". Accordingly, there are two main approaches to deal with univariate overdispersion: First, the use of full parametric models like dispersion models (Joe 1997), and second, the choice of families of estimating functions (Heyde 1997). In the case of multivariate data, multivariate dispersion models (Jørgensen & Lauritzen 2000) and copula function based models (Song, Li & Yuan 2009) can be used.

The literature on copula model with count data is not abundant, with some references in financial and actuarial sciences. Nikoloulopoulos & Karlis (2010) present a recent review for the use of this methodology with application to discrete data in marketing exchanges. Some applied works have been done in joint modeling of correlated data using Gaussian copulae (Song et al. 2009). Furthermore, a recent approximation to the Gaussian copula likelihood is given in Madsen & Fang (2011), who found that for finite samples the estimator of generalized estimating equations is more efficient than the maximum likelihood estimator (MLE). However, Song, Li & Yuan (2011) maintain that MLE is more efficient.

With respect to applications in the biological sciences, the next are some useful references. Lambert & Vandenhende (2002) propound a model for non-normal longitudinal data with illustration in a dose titration safety study in human medicine. A work in multivariate logistic regression was presented by Li & Wong (2011) and, because of a lack of constraints in the parameters and the admission of a limited range of dependence in the copula, this paper was criticized and corrected (Nikoloulopoulos 2012). A more basic study was carried out by Trégouët, Ducimetière, Bocquet, Visvikis, Soubrier & Tiret (1999), with binary data on nuclear families, in this analysis the response was the presence or the absence of a disease in each member of the family.

In the particular situation of plant-disease complex, the presence of two or more pathogenic fungi can be strongly correlated, thereby violating the assumption of independence amongst observations (Dávila 2005, Dávila & López 2010). In such a situation, it is necessary to use a statistical model with multivariate distributions which include both marginal overdispersion and multivariate dependence (Fischer 2011, Joe 1997, Song 2007).

Ultimately, in relation to the disadvantages of copula-based analysis of count data, two important references shall be mentioned: Genest & Nešlehovà (2007) for details on the danger and limitations of the use of copulae to model discrete data, and Embrechts (2009) who in a personal view gives some review on this theory, recommends some important lectures and analyzes future developments. Additionally, the reader is encouraged to review the controversial article of Mikosch (2006), which is a critical point of view of copula methodology, with discussion and rejoinder. Despite some problems in copula modeling with discrete data, nowadays this model constructions are valid but subject to cautions.

The present paper contains four sections. Section 2 presents the characterization of multivariate vectors, reviews some concepts on overdispersion diagnostics and model selection. Section 3 is dedicated to theoretical details of the proposed model. Section 4 shows an application to empirical data in diseases management on vegetables. Finally, Section 5 presents discussions and conclusion.

2. Material and Methods

In this section we present the characterization of data and parameter vectors, overdispersion diagnostic and a short reminder on copula theory and model selection.

2.1. Structure of data and parameter vectors

In plant pathology studies, data are typically made of binomial observations representing the presence/absence of pathogenic fungi. Data obtained for d fungi are modeled by a d-variate vector:

$$Y = (Y_1, Y_2, \dots, Y_d)^T \tag{1}$$

where Y_i is a binomial random variable associated to the incidence of the ith fungus, $i = 1, 2, \dots, d$. A common assumption is that the probabilistic mechanism that generates marginal data is the binomial law, whose density with respect to the counting measure is given by

$$f_{Y_i}(y_i \mid \pi_i, m_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} \tag{2}$$

where $y_i = 0, 1, \dots, m_i$ and with given probability of success π_i; we write formally that $Y_i \sim bin(m_i, \pi_i)$, with

$$E[Y_i] = m_i \pi_i$$

and

$$Var[Y_i] = m_i \pi_i (1 - \pi_i),\ \ i = 1, 2, \dots, d \tag{3}$$

Provided that multivariate data are generated by the same designed experiment, there is an identical design matrix X associated to any margin Y_i; hence, under the GLM framework, the three components are (see McCullagh & Nelder 1989):

1. The class of densities in (2) with π_i varying in the interval $(0, 1)$, which belongs to the exponential family of distributions,
2. The systematic part $X\theta_i$, where X is a $n \times p$ matrix and $\theta_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{ip})^T$ is a vector of unknown parameters with $\theta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$, and
3. The link function $g_i(\cdot)$.

In GLM modeling, it is supposed that there is independence between any subset of random variables from (1) and that (3) holds.

Because this work is dealing with the lack of independence and overdispersion ($Var[Y_i] \gg m_i \pi_i (1 - \pi_i)$), a natural characteristic of multivariate data arising in plant-disease complex, then a new model shall be considered. Hence a full likelihood inference procedure requires a family of distributions with a great vector of total marginal parameters

$$\Theta = (\theta_1^T, \theta_2^T, \dots, \theta_d^T)^T$$

and an association matrix

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdot & \cdot & \cdot & \gamma_{1d} \\ \gamma_{21} & \gamma_{22} & \cdot & \cdot & \cdot & \gamma_{2d} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot \\ \gamma_{d1} & \gamma_{d2} & \cdot & \cdot & \cdot & \gamma_{dd} \end{pmatrix}$$

where γ_{ii^*}, $i \neq i^*$, $i^*, i = 1, 2, \dots, d$, will be taking in account the bivariate association between each pair of transformed margins; the construction of the desired multivariate distribution is the objective of the Section 3. However, an important prerequisite lies in the detection of extra binomial variation, which we now detail.

2.2. Overdispersion Diagnostic

To test the nominal dispersion in the ith margin, it is important to give an extension of (3), i.e.,

$$Var[Y_i] = \lambda_i m_i \pi_i (1 - \pi_i),$$

and the hypothesis testing problem is formulated for all $i = 1, 2, \dots, d$ as

$$H_{0_i} : \lambda_i = 1 \ versus \ H_{1_i} : \lambda_i > 1 \tag{4}$$

An appropriate procedure to test (4) is the score statistic of Smith and Heitjan (1993), viz.

$$\chi_i^2 = J_i^T A_i^{-1} J_i, \ i = 1, 2, \dots, d \tag{5}$$

where $J_i = (J_{i1.}, J_{i2.}, \dots, J_{ip.})$ is a random vector that registers the difference between actual information and nominal information, in the ith margin with respect to every jth parameter, namely

$$J_{ij.} = \frac{1}{2}\sum_{k=1}^{n}\left[\left(\frac{\partial l_{ijk}}{\partial \theta_{ij}}\right)^2 - \left(\frac{\partial^2 l_{ijk}}{\partial \theta_{ij}^2}\right)\right] j = 1, \dots, p, \ i = 1, 2, \dots, d \tag{6}$$

and A_i is the covariance matrix of J_i corrected for estimation of θ_i, whose explicit expressions are given in the appendix of Smith & Heitjan's (1993) paper.

In equation (6), l_{ijk} is the log-likelihood of the binomial distribution presented in (2). Hence, for each ith margin with respect to the jth parameter and the kth observation, we have

$$l_{ijk} = y_{ijk} \ln\left(\frac{\pi_{ijk}}{1 - \pi_{ijk}}\right) + m_{ijk} \ln(1 - \pi_{ijk})$$

and

$$J_{ij.} = \frac{1}{2}\sum_{k=1}^{n}\left[(y_{ijk} - m_i\pi_i)^2 - m_i\pi_i(1 - \pi_i)\right]x_{ijk}^2, \ j = 1, \dots, p, \ i = 1, 2, \dots, d$$

Under the null hypothesis of nominal dispersion (4), the asymptotic distribution of (5) is the central χ^2-distribution with p degrees of freedom. The eventual reject of (4) will be a clear evidence that $Var[Y_i] \gg m_i\pi_i(1 - \pi_i)$; namely, actual variance is statistically greater than the nominal one.

Hitherto, we have been dealing with marginal overdispersion, whereas the statistical problem in plant-pathogen complex data includes both marginal overdispersion and multivariate dependence. In the following, we show how the latter can be addressed using copulae theory.

2.3. Basics on Copula Modeling

An interesting concept for connecting multivariate cumulative distribution functions and their margins is offered by copulae theory (see Joe 1997, Nelsen 2006). A mapping $C : [0, 1]^d \to [0, 1]$ is called a d-dimensional copula, if it is the distribution of a uniform vector $U = (U_1, U_2, \dots, U_d)$; that is, copulae are joint distribution functions of standard uniform random variates (Cherubini, Luciano

& Vecchiato 2004). Because any marginal distribution function F_i has a uniform distribution, i.e. $F_i(y) \sim U(0,1)$ with $i = 1, 2, \dots, d$, the use of copulae has become evident in the last few years, to construct dependency models (Härdle & Simar 2007).

The application of copulae to statistical modeling is based on Sklar's theorem (Nelsen 2006); this useful theorem states that given marginal distributions, it is possible to couple these margins into a joint distribution whose arguments are the F_i's; provided that the margins are continuous, this kind of representation is unique. Hence, following Grønneberg (2011), there are four basic problems in parametric modeling through copulae theory, namely:

- How to estimate the dependence parameter?
- How should the parametric form of the copula family be chosen?
- How to select among several candidate models on the basis of actual data?
- Is the final model adequate?

The scientific context of plant pathology gives us preliminary responses for the first two items, whereas the two later are pure statistical modeling steps and will be reviewed in the following.

2.4. Model Selection and Goodness of Fit

A usual tool for model selection is the Akaike Information Criterion (AIC), which is not appropriate when dealing with semi-parametric estimation, a common method used in the construction of copulae. A proper generalization of AIC, given in Grønneberg (2011), is the Copula Information Criterion (CIC), viz.

$$CIC = 2l_{N,\max} - 2(\widehat{p}^* + \widehat{q}^* + \widehat{r}^*) \tag{7}$$

where $l_{N,\max}$ is the maximum multivariate pseudo-likelihood. The second term of (7) has a more elaborate formula than in AIC –where it depends only on the length of parameter vector. If the model is correctly specified, then $\widehat{q}^* = 0$. Details for deriving the estimates of $\widehat{p}^*$, $\widehat{q}^*$ and $\widehat{r}^*$ from empirical information, and least false copula derivatives are given in Grønneberg (2011).

Genest, Rémillard & Beaudoin (2009) provide a useful tool to test the final model adequacy. Let H be a joint cumulative distribution function the copula representation of H is

$$H(y_1, y_2, \dots, y_d) = C(F_1(y_1), F_2(y_2), \dots, F_d(y_d)) \tag{8}$$

provided that C is unknown to model $Y = (Y_1, Y_2, \dots, Y_d)^T$, we suppose that C belongs to a class

$$\mathcal{C} = \{C_\omega : \omega \in \Omega\}, \Omega \subseteq \mathbb{R}^d, d \geq 1 \tag{9}$$

so we must test,

$$H_0 : C \in \mathcal{C} \quad \text{versus} \quad H_1 : C \notin \mathcal{C} \tag{10}$$

Genest et al. (2009) advocate the use of "blanket test", based on the empirical copula, viz.,

$$C_N(\mathbf{u}) = \frac{1}{N}\sum_{l=1}^{N} \mathbf{I}(\widehat{\mathbf{U}}_l \leq \mathbf{u}),\ \mathbf{u} \in [0,1]^d \tag{11}$$

where $\widehat{\mathbf{U}}_l$ is a vector of pseudo-observations, whose components are the empirical cumulative distribution functions related to each margin, obtained from actual data, i.e.,

$$\widehat{\mathbf{U}}_l = (\widehat{F}_{l,1}, \dots, \widehat{F}_{l,d}),\ l = 1, 2, \dots, N$$

with N being the size of a random sample from (1); it is important to recall that, under probability transformations, it is expected that $\widehat{F}_{l,i} \sim U(0,1)$ for all $l = 1, 2, \dots, N$ and $i = 1, 2, \dots, d$. The empirical copula (11) is a consistent estimator of C in (8), and the statistic to test H_0 in (10) is

$$S_N = \sum_{l=1}^{N} \{C_N(\widehat{\mathbf{U}}_l) - C_{\omega_N}(\widehat{\mathbf{U}}_l)\}^2 \tag{12}$$

The asymptotic distribution of (12) cannot be directly tabulated, then approximations of p-values shall be obtained via bootstrap-based procedures. Because of its high computational cost, Kojadinovic, Yan & Holmes (2011) recently proposed a fast large-sample testing procedure based on multiplier central limit theorems.

Now that we have recalled the basics of model selection and goodness of fit tests, we can introduce our alternative model for the statistical analysis of plant-pathogen complex.

3. A Model for Multivariate Overdispersed Binomial Data

Here the objective is to present an alternative statistical model to analyze plant-pathogen complex data. More specifically, we shall focus on the analysis of designed experiments to evaluate substances as possible activators of Systemic Acquired Resistance (SAR) (Durrant & Dong 2004). Because SAR is a mechanism which confers a broad spectrum of protection against plant pathogens, it is expected that all fungi in a complex should be affected and that multivariate data should not present independence; additionally, the natural spreading of pathogen inoculum cannot guarantee marginal independency, then marginal overdispersion can be a natural attribute of such data.

We are going to construct the desired model in two steps, first, fitting margins to an appropriate family of distribution, and second, modeling the given margins in a Gaussian copula family framework.

3.1. Marginal Overdispersion Model

In order to model marginal overdispersion, we make use of Beta-binomial hierarchy, a generalization of binomial distribution (Casella & Berger 2002). In this model, it is supposed that $Y_i \mid P_i \sim bin(m_i, P_i)$, whereas $P_i \sim Beta(\alpha_i, \beta_i)$. Then, from now on, we make the assumption that each margin (Y_i) follows a Beta-binomial law. Therefore, unconditionally the compound density, with respect to the counting measure of Y_i, is given by

$$f_{Y_i}(y_i \mid \alpha_i, \beta_i) = \binom{m_i}{y_i} \frac{B(y_i + \alpha_i, m_i - y_i + \beta_i)}{B(\alpha_i, \beta_i)}, \; y_i \in \{0, 1, \dots, m_i\} \tag{13}$$

furthermore, in (13) $B(.,.)$ is the beta function, $\alpha_i > 0$ and $\beta_i > 0$. Conditional to P_i the expectation is given by

$$E(Y_i \mid P_i) = \mu_i = m_i \pi_i = m_i \frac{\alpha_i}{\alpha_i + \beta_i}, \; i = 1, 2, \dots, d$$

the conditional variance is

$$\begin{aligned} Var(Y_i \mid P_i) &= m_i \pi_i (1 - \pi_i) \frac{\alpha_i + \beta_i + m_i}{\alpha_i + \beta_i + 1} \\ &= m_i \pi_i (1 - \pi_i) \{1 + \phi_i (m_i - 1)\}, \quad i = 1, 2, \dots, d \end{aligned} \tag{14}$$

from (14) we can see that the marginal dispersion parameter is

$$\phi_i = \frac{1}{\alpha_i + \beta_i + 1}$$

Comparing (3) with (14) it is noted that the later has a greater variance, whose increment is given by a function of ϕ_i and the marginal binomial index m_i. The R package VGAM and its function `vglm` is actually an alternative to fit marginal responses with Beta-binomial distribution.

3.2. Multivariate Model

Given the marginal distributions $F_1(Y_1), F_2(Y_2), \dots, F_d(Y_d)$ from Beta-binomial hierarchies (13) and using the Sklar's theorem, a new family of d-variate distributions can be obtained and represented by

$$C_\Phi(U_1, U_2, \dots, U_d) = H\left(F_1(Y_1), F_2(Y_2), \dots, F_d(Y_d) \mid \Gamma\right) \tag{15}$$

where H is the d-variate Gaussian distribution with correlation matrix Γ and, in presence of continuous margin, the density is given by

$$f_Y(y; \mu, \phi, \Gamma) = c_\Phi\{F_1(y_1), F_2(y_2), \dots, F_d(y_d) \mid \Gamma\} \prod_{i=1}^{d} f(y_i; \pi_i, \phi_i)$$

where $\pi^T = (\pi_1, \pi_2, \ldots, \pi_d) \in [0,1]^d$ is the main vector of marginal parameters and $\phi^T = (\phi_1, \phi_2, \ldots, \phi_d) \in \mathbb{R}^d$ is the ancillary vector of marginal dispersion parameters. Because (13) is a discrete distribution, then we use the more appropriate expression

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_d = y_d) = \\ \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \cdots \sum_{j_d=1}^{2} (-1)^{j_1+j_2+\ldots+j_d} C_\Phi(u_{1j_1}, u_{2j_2}, \ldots, u_{dj_d} \mid \Gamma) \quad (16)$$

with $u_{i1} = F_i(y_i)$ and $u_{i2} = F_i(y_i - 1)$ $i = 1, 2, \ldots, d$, which is the density with respect to the counting measure, namely the Radon-Nikodym derivative of (15).

3.3. Two Step Inference

To make inference on (16) we use the two parts inference procedure, proposed by Joe (1997). In this methodology, in the first step the margins are fitted from (13) and because it is composed of common functions, both numerical methods or maximum likelihood estimation (MLE) are applicable; see Griffiths (1973) for details. The R package VGAM makes use of Fisher scoring for estimation and it operates quite well for overdispersed binomial data. In a particular situation, to model $g(\pi_i) = X\theta_i$, the score equation, for maximum likelihood estimation from (13), is

$$\frac{\partial l_i}{\partial \theta_i} = (\alpha_i+\beta_i) \sum_{k=1}^{n} \left\{ ddg(y_i, (\alpha_i+\beta_i)\pi_i) - ddg(m_i-y_i, (\alpha_i+\beta_i)(1-\pi_i)) \right\} \frac{1}{g'(\pi_i)} x_{jk}$$

$j = 1, \ldots, p$, where $ddg(a, b) = \log \Gamma(a + b) - \log \Gamma(b)$; additional details can be seen in Hinde & Demetrio (1998).

The second step deals with the selection of an appropriate family of copulae. In the case of Gaussian copula, for the estimation of Γ, can be used some assumptions like the presence of exchangeable Pearson correlation matrix, i.e., $\gamma_{ii^*} = \gamma$, $i \neq i^*$; in any case, from (16) the solution of

$$\frac{\partial C_\Phi(u_{1j_1}, u_{2j_2}, \ldots, u_{dj_d} \mid \Gamma)}{\partial \Gamma} = 0$$

can be obtained using the Gaussian-Hermite quadrature method (see McCulloch, Searly & Neuhaus 2008, pp. 326-331). Finally, consider the vector of marginal and multivariate parameters

$$\eta = (\theta_1, \theta_2, \ldots, \theta_d, \gamma_{12}, \gamma_{13}, \ldots, \gamma_{(d-1)(d)})$$

in order to complete the inference procedure; following Joe (1997), it is necessary to estimate the inverse Godambe information matrix

$$V = D_h^{-1} M_h (D_h^{-1})^T \quad (17)$$

where $D_h = E[\partial h^T(Y,\eta)/\partial\eta]$ and $M_h = E[h^T(Y,\eta)h(Y,\eta)]$, with h being the first derivative of the logarithm of (16) with respect to η. The estimation of $N^{-1}V$, which is the asymptotic covariance matrix of the MLE of η, namely $\widehat{\eta}$, can be done via Jackknife, viz.,

$$\Im = \sum_{l=1}^{N} (\widehat{\eta}^{(l)} - \widehat{\eta})^T (\widehat{\eta}^{(l)} - \widehat{\eta}) \tag{18}$$

In (18), $\widehat{\eta}^{(l)}$ is the estimator of η once the lth observation has been eliminated.

4. Application

Cely (1996) carried out a trial in Colombia in an onion crop, in order to analyze the effect of seven treatments, based on the aspersion of inactive inoculum of the plant pathogen *Peronospora destructor* for cross protection, an approach later included in the SAR methods by Durrant & Dong (2004). The experiment was located under a complete randomized block design, with two blocks (the crop varieties Junca and Monguana). Three responses were captured as binomial data, all of them associated to the incidence of a pathogenic fungus; namely Y_1 represents the downy mildew *Peronospora* sp., Y_2 the leaf blight *Stemphylium* sp. and Y_3 the leaf spot *Cladosporium* sp.; so the dependent response vector to be modeled is

$$Y = (Y_1, Y_2, Y_3)$$

Initially, nominal dispersion was rejected with p-values less than 0.05, for all three margins with respect to the hypothesis testing problem in (4); furthermore, marginal Beta-binomial hierarchy models (13) were fitted; then, given the three CDF's $F_1(y_1), F_2(y_2), F_3(y_3)$ a 3-variate Gaussian copula model was fitted, according to (16). To select the model on the basis of observed data, we use Copula Information Criterion (7), and the goodness of fit was based on Genest et al. (2009); finally, applying Jackknife method (18), the Godambe's asymptotic covariance matrix was estimated. About marginal dispersion parameters, the nominal dispersion (4) was not rejected, under 3-variate framework, for Y_2, i.e., $\phi_2 \simeq 0$ for the random variable associated to *Stemphylium* fungus, an endemic plant pathogen; see Table 1.

Table 1: Estimations, standard errors, and confidence intervals for dispersion parameters(*).

Parameter estimations	Standard error	Lower Limit	Upper Limit
$\widehat{\phi}_1 = 0.01983$	0.0037	0.0125	0.0270
$\widehat{\phi}_2 = 0.00377$	0.1979	−0.3840	0.3915
$\widehat{\phi}_3 = 0.01735$	0.0035	0.0104	0.0242

(*) $\alpha = 0.05$.

The standard errors of the parameter estimators appear on Table 2 for normal linear models (MVN) with Box and Cox transformations –the original model used by Cely (1996)–, generalized linear model (GLM), marginal overdispersion model (ODM), and multivariate overdispersion model with Gaussian copula and Beta-binomial margins (CGB). As it can be seen, ODM and CGB are the models with less significant effects. In fact, both ODM and CGB show a total of six standard errors associated with significant estimations; nevertheless, without differences in relation to the number of significant effects, CGB offers higher values of standard errors.

With respect to the estimation of the association parameters, the correlation matrix, i.e.,

$$\widehat{\Gamma} = \begin{pmatrix} 1.000 & 0.484** & 0.475** \\ 0.484** & 1.000 & 0.688** \\ 0.475** & 0.688** & 1.000 \end{pmatrix}$$

shows a positive dependence between normal scores; all three estimations were highly significant (p-value < 0.0001), leading to the consideration that the appropriate copula, for the analyzed data, is not the independent one.

Table 2: Standard errors for parameter estimators.

Factor(variable)	MVN(1)	GLM(1)	ODM(1)	CGB(1)
T0(y$_1$)	0.0266*	0.103*	0.163*	0.157*
T1(y$_1$)	0.0266*	0.106*	0.166*	0.141*
T2(y$_1$)	0.0266	0.112	0.179	0.221
T3(y$_1$)	0.0266	0.110*	0.174	0.176
T4(y$_1$)	0.0266*	0.107*	0.168*	0.159*
T5(y$_1$)	0.0266	0.112	0.177	0.166
T6(y$_1$)	0.0266	0.113	0.179	0.197
JUNCA(y$_1$)	0.0133*	0.052*	0.082*	0.086*
T0(y$_2$)	0.0186	0.093	0.106	0.113
T1(y$_2$)	0.0186*	0.091*	0.103*	0.136*
T2(y$_2$)	0.0186*	0.094	0.107	0.135
T3(y$_2$)	0.0186*	0.091*	0.104*	0.122*
T4(y$_2$)	0.0186	0.096	0.109	0.121
T5(y$_2$)	0.0186	0.095	0.108	0.142
T6(y$_2$)	0.0186	0.098	0.111	0.107
JUNCA(y$_2$)	0.0092	0.046	0.053	0.061
T0(y$_3$)	0.0265	0.108	0.164	0.169
T1(y$_3$)	0.0265	0.104	0.160	0.196
T2(y$_3$)	0.0265	0.109	0.167	0.202
T3(y$_3$)	0.0265	0.105	0.161	0.210
T4(y$_3$)	0.0265	0.107	0.163	0.176
T5(y$_3$)	0.0265	0.105	0.162	0.208
T6((y$_3$)	0.0265	0.113*	0.170	0.161
JUNCA(y$_3$)	0.0130	0.053	0.081	0.089

(1)*= significative effect (α=0.05).

In relation to the early work of Cely (1996), the author made use of the assumption of independence between the three count variables; hence, let's see that a wrong assumption can lead to an incorrect inference. In the original report of Cely, the SAR-treatment (T2), with respect to the random variable Y_2, was considered a significant one, i.e., it was statistically different from chemical and mixed treatments and its use was not taken account: why shall a small difference lead to significant effect? The answer is an underestimation of standard error, given by lack of independence and marginal overdispersion, that were not considered in the assumed probability model.

In this new data analysis, based on dependence concepts, the treatment T2 does not have differences with respect to chemical and mixed ones, according to response Y_2; therefore, the new position will be that T2 is a good solution to implement an integrated pathogen handing in that crop, because it controls the three pathogens together with statistical significance. In Table 3 we present two inferential situations; first, the analysis under MVN, whose significant effects are represented by "$\ast$"; second, the analysis via CGB, whose significant effects are indicated by "$\diamond$". Because in CGB the treatment T2 is statistically similar to the chemical ones, this new analysis is in favour of T2, the natural SAR-fungicide.

Table 3: Two inferential situations

Treatments	$Y_1(\%)$	$Y_2(\%)$	$Y_3(\%)$
T0= control	18.38$\ast\diamond$	18.37$\ast\diamond$	11.45
T1= SAR low dosage	16.78$\ast\diamond$	20.38$\ast\diamond$	15.10
T2= SAR medium dosage	10.26	16.10$\ast$	10.10
T3= SAR high dosage	12.62	20.17$\ast\diamond$	12.60
T4= Mancozeb	14.88$\ast\diamond$	15.50	11.85
T5= Mancozeb + Cimoxanyl	11.66	15.40	11.80
T6= T2+T4	10.35	14.97	10.3
T7= T2+T5	9.93	15.90	11.15

$\ast$ = significant effect in relation to MVN modeling, $\alpha = 0.05$.
$\diamond$ = significant effect in relation to CGB modeling, $\alpha = 0.05$.

5. Discussion and Conclusion

Gaussian copulae theory is suitable to construct models with given non-normal margins, which is the particular situation in plant diseases control. A very important issue in model selection is the context, i.e., all modeling shall have scientific foundations and clear proposals (Claeskens & Hjort 2008). Because the application of some therapies associates to natural resistance activation (SAR methodologies) on plants against fungi has a broad spectrum, the lack of independence between the incidence of pathogens is evident, and then the use of independent marginal models is out of scientific context.

Also, it is important to stress the difference of the present methodology with respect to the works of Song et al. (2009) and Song (2000), which is the use of margins not belonging to the class of dispersion models (Jørgensen 1997) in our proposal. Here we are using a Beta-binomial hierarchy to deal with marginal overdispersion, a new application to copulae theory in the broad field of plant pathology, a methodology appropriate to modeling SAR-based experiments, the ones that require modern statistical tools.

It is worth to recall some limitations of the proposed model, according to the work of Genest & Nešlehovà (2007). The first limitation is the lack of uniqueness of the copula, once the random variables put their mass on few atoms: it is a crucial aspect in binary data and less important if the binomial index tends to infinity in binomial variables. Accordingly, practitioners may be cautious in the use of the present methodology with sparse data, that is, when the binomial index is small ($m_i < 6$), our model is not appropriate.

A second aspect of copula-based regression for discrete data is that dependence is not only a function of the copula; additionally, Kendall's tau an Spearman's rho may not span the entire interval $[-1, 1]$. About this weakness, the use of Gaussian copula guarantees that the association parameter, i.e., the Pearson correlation coefficient, can reach the Fréchet-Hoeffding bounds (Song 2007). Nevertheless, these dependence parameters are governing the association but they do not have direct interpretation. That is, the correlation between normal scores is not the same that the one between the actual variables; hence, we may interpret Γ as a dependence parameter matrix, all but as a correlation matrix of the original binomial variables. Furthermore, because the margins also characterize the dependence in the copula, when dealing with discrete data we may consider a conditional copula model, where the association parameters are varying with the covariates (see Acar, Craiu & Yao 2011).

Even if the conditions under which the dependence parameters are estimable are not elucidated, hitherto the maximum likelihood estimation is a valid methodology for inference. Hence, no further discussion on this topic are exposed here (Genest & Nešlehovà 2007).

In conclusion, we have that in our example, the model based on Gaussian copula (CGB) displayed the highest standard errors associated to parameter estimators, suggesting that this approach controlled the overdispersion in the data. Additionally, it considers both marginal overdispersion and multivariate dependence, whereas the marginal overdispersion model, based on independent Beta-binomial hierarchy (ODM), assigns multivariate dependence to a marginal overdispersion. Provided that multivariate dependence is present, application shows that normal linear models (MVN) does not differ from modeling via GLM without overdispersion fit, leading to a wrong multivariate inference. The model constructed via Gaussian copula with Beta-binomial margins (CGB) is probably preferable for analyzing overdispersed and non-sparse multivariate binomial data, whereas the classical multivariate normal linear model is not appropriate in such situations.

Acknowledgments

We would like to thank the editors and the two anonymous referees for their constructive comments to improve the quality and presentation of this paper.

[Recibido: septiembre de 2011 — Aceptado: febrero de 2012]

References

Acar, E., Craiu, R. & Yao, F. (2011), 'Dependence calibration in conditional copulas: A nonparametric approach', *Biometrics* **67**, 445–453.

Casella, G. & Berger, R. (2002), *Statistical Inference*, 2 edn, Duxbury Press, Florida, United States.

Cely, B. (1996), Control de mildeo velloso (*Peronospora destructor*) en el cultivo de cebolla de rama mediante protección cruzada, Tesis de grado, Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia.

Cherubini, U., Luciano, E. & Vecchiato, W. (2004), *Copula Methods in Finance*, John Wiley & Sons, England.

Claeskens, G. & Hjort, N. (2008), *Model Selection and Model Averaging*, Cambridge University Press, Cambridge.

Cox, D. R. (1983), 'Some remarks on overdispersion', *Biometrika* **7**(1), 269–274.

Durrant, W. & Dong, X. (2004), 'Systemic acquired resistance', *Annual Review of Phytopathology* **42**, 185–209.

Dávila, E. (2005), Modelación multivariada de la sobredispersión en datos binarios, aplicación en epidemiología vegetal, Tesis de maestría, Universidad Nacional de Colombia, Bogotá, Colombia.

Dávila, E. & López, L. (2010), Modeling multivariate overdispersed binomial data, *in* 'International Biometrics Conference', XXV International Biometric Conference, Florianópolis, Brazil.

Embrechts, P. (2009), 'Copulas: A personal view', *The Journal of Risk and Insurance* **76**(3), 639–650.

Fischer, M. (2011), Multivariate copulas, *in* D. Kurowicka & H. Joe, eds, 'Dependence Modeling Vine Copula Handbook', World Scientific, pp. 19–36,.

Genest, C. & Nešlehovà, J. (2007), 'A primer on copulas for count data', *ASTIN Bulletin* **37**(2), 475–515.

Genest, C., Rémillard, B. & Beaudoin, D. (2009), 'Goodness-of-fit tests for copulas: A review and a power study', *Insurance: Mathematics and Economics* **44**, 199–213.

Griffiths, D. A. (1973), 'Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease', *Biometrics* **29**, 637–648.

Grønneberg, S. (2011), The copula information criterion and its implications for the maximum pseudo-likelihood estimator, *in* Kurovicka & Joe, eds, 'Dependence Modeling Vine Copula Handbook', World Scientific, pp. 113–138.

Härdle, W. & Simar, L. (2007), *Applied Multivariate Statistical Analysis*, Springer-Verlag, Berlin.

Heyde, C. (1997), *Quasi-likelihood And Its Applications: A General Approach To Optimal Methods of Estimation*, Springer, New York.

Hinde, J. & Demetrio, C. (1998), *Overdispersion: Models and estimation*, XIII Sinape, Caxambu, Brazil.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall, London.

Jørgensen, B. (1997), *Dispersion Models*, Chapman and Hall, London.

Jørgensen, B. & Lauritzen, S. (2000), 'Multivariate dispersion models', *Journal of Multivariate Analysis* **74**, 267–281.

Kojadinovic, I., Yan, J. & Holmes, M. (2011), 'Fast large-sample goodness-of-fit for copulas', *Statistica Sinica* **21**, 841–871.

Lambert, P. & Vandenhende, F. (2002), 'A copula-based model for multivariate non-normal longitudinal data: Analysis of a dose titration safety study on a new antidepressant', *Statistics in Medicine* **21**, 3197–3217.

Li, J. & Wong, W. (2011), 'Two-dimensional toxic dose and multivariate logistic regression, with application to decompression sickness', *Biostatistics* **12**, 143–155.

Madsen, L. & Fang, Y. (2011), 'Joint regression analysis for discrete longitudinal data', *Biometrics* **67**(3), 1171–1175.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall/CRC, London.

McCulloch, C., Searly, S. & Neuhaus, J. (2008), *Generalized Linear and Mixed Models*, Wiley, New York.

Mikosch, T. (2006), 'Copulas: Tales and facts (with discussion and rejoinder)', *Extremes* **9**, 3–63.

Nelsen, R. (2006), *An Introduction to Copulas*, 2 edn, Springer, New York.

Nikoloulopoulos, A. (2012), 'Letter to the editor', *Biostatistics* **13**(1), 1–3.

Nikoloulopoulos, A. & Karlis, D. (2010), 'Modeling multivariate count data using copulas', *Statistics in Medicine* **27**, 6393–6406.

Smith, P. & Heitjan, F. (1993), 'Testing and adjusting for departures from nominal dispersion in generalized linear models', *Applied Statistics* **42**(1), 31–34.

Song, P. X. (2000), 'Multivariate dispersion models generated from gaussian copula', *Scandinavian Journal of Statistics* **27**, 305–320.

Song, P. X. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer, New York.

Song, P. X., Li, M. & Yuan, Y. (2009), 'Joint regression analysis of correlated data using gaussian copulas', *Biometrics* **65**, 60–68.

Song, P. X., Li, M. & Yuan, Y. (2011), 'Joint regression analysis for discrete longitudinal data - rejoinder', *Biometrics* **67**(3), 1175–1176.

Trégouët, D., Ducimetière, P., Bocquet, V., Visvikis, S., Soubrier, F. & Tiret, L. (1999), 'A parametric copula model for analysis of familial binary data', *American Journal of Human Genetics* **64**(3), 886–893.

Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 271 a 287

Modelación de indicadores del estado nutricional de la embarazada desde un enfoque multinivel

Modeling of Indicators of Nutritional Status of Pregnant Women from a Multilevel Approach

Minerva Montero[1,a], Maria Elena Díaz[2,b], Santa Jiménez[2,c], Iraida Wong[2,d], Vilma Moreno[2,e]

[1]Departamento de Matemática, Instituto de Cibernética, Matemática y Física, La Habana, Cuba

[2]Departamento de Antropología, Instituto de Nutrición e Higiene de los Alimentos, La Habana, Cuba

Resumen

Se presenta una estrategia para la construcción de indicadores antropométricos empleados para evaluar el estado nutricional de la mujer embarazada. Las referencias del peso por semana de embarazo, según la estatura y el índice de masa corporal de la mujer al inicio de la gestación, se construyen a partir de modelos multinivel para medidas repetidas. Para verificar la consistencia de los valores estimados y ajustar el posible efecto del desbalance de los datos causado por observaciones perdidas, las estimaciones máximo-verosímil se comparan con las obtenidas mediante un método bootstrap. Los resultados obtenidos no sólo evidencian el poder de los modelos multinivel para la construcción de patrones de referencia, sino que además permiten estimar rangos de ganancia de peso recomendados para las embarazadas de la población en estudio.

***Palabras clave*:** antropometría, estudios longitudinales, medidas repetidas, modelo multinivel.

Abstract

A strategy for the construction of anthropometric indicators employed for the evaluation of the pregnant women's nutritional state is presented. The references of weight by pregnancy week, according to height and the

[a]Investigadora auxiliar. E-mail: minerva@icimaf.cu
[b]Investigadora titular. E-mail: maryelen@infomed.sld.cu
[c]Investigadora titular. E-mail: vdninha@infomed.sld.cu
[d]Antropometrista. E-mail: iraida@sisvan.sld.cu
[e]Antropometrista. E-mail: vilma@sisvan.sld.cu

body mass index at the beginning of pregnancy, are constructed by means of multilevel models for repeated measurements. The maximum likelihood estimations are compared to those obtained by the bootstrap method in order to verify the consistency of the estimated values and the fitting of the possible data imbalance effect, caused by missing observations. The obtained results evidence not only the power of the multilevel models for the construction of reference patterns, but they also permit estimate ranks of gain in weight recommended for pregnant women representative of the population under study.

***Key words*:** Anthropometry, Longitudinal studies, Multilevel model, Repeated measurements.

1. Introducción

Una adecuada clasificación nutricional de la mujer durante el embarazo permitirá una mejor identificación de los riesgos adversos asociados a la gestación. Diferentes métodos basados en indicadores antropométricos del estado nutricional materno han sido propuestos como patrones de referencia para realizar la valoración clínica y epidemiológica de la mujer embarazada (Gueri, Jutsum & Sorhaindo 1982, Fescina 1986, Rosso 1995, IOM 1990*a*, Krasovec & Anderson 1991, WHO 1995, Schwarcs, Díaz, Fescina, De Mucio, Belitzky & Delgado 1995, IOM 1990*b*, Lubin, Blot, Berrino, Flamant, Gillis, Kunze, Schmäwhl & Visco 1997, Mardones & Rosso 2005). No obstante, cuando se usan referencias extranjeras deben tenerse en cuenta las diferencias genéticas y ambientales existentes entre áreas de desigual desarrollo económico para evitar una sobrevaloración o subvaloración de los problemas nutricionales. Para corregir tales sesgos es necesario desarrollar herramientas de evaluación conforme al contexto físico y sociocultural de la población de interés.

En este artículo se presentan los principales resultados de una investigación cubana para la construcción de valores de referencias locales (Díaz, Montero, Jiménez, Wong & Moreno 2008*b*, Díaz, Montero, Jiménez, Wong & Moreno 2009), donde el peso materno se utiliza como principal indicador del estado nutricional durante el embarazo. Los datos longitudinales, medidos sobre las mismas embarazadas en diferentes ocasiones, dan lugar a una estructura jerárquica en la que las medidas repetidas se anidan dentro de las gestantes seleccionadas de la población de interés; así, las ocasiones constituyen las unidades de nivel-1, y las embarazadas las unidades de nivel-2. La heterogeneidad y dependencia de los datos se tiene en cuenta modelando el problema desde un enfoque multinivel (Goldstein 1995, Brik & Raudenbush 1992).

Los modelos multinivel se usaron para construir canales de seguimiento del peso de la embarazada a partir de los datos disponibles medidos en las consultas de control prenatal (una vez al mes), como establece el sistema de salud nacional; no obstante, no se desestimó ninguna observación adicional realizada fuera de lo planificado. El resultado es un conjunto de datos con un número desigual de mediciones por embarazadas y donde los intervalos de tiempo varían entre las embarazadas. El enfoque multinivel es capaz de acomodar este tipo de datos desbalanceados y es eficiente aun cuando algunos datos se pierden aleatoriamente, como cuando ocurre si alguna mujer no asiste a la consulta programada. El objetivo de este trabajo es describir el procedimiento utilizado en la construcción de indicadores antropométricos del estado nutricional de la embarazada a partir de una muestra de datos longitudinales con intervalos irregulares entre mediciones y con observaciones perdidas.

En los modelos multinivel propuestos se examinaron los parámetros fijos y los componentes de la varianza de los errores aleatorios. Esto permitió hacer inferencias sobre los efectos de la población utilizando una muestra aleatoria de embarazadas; lo que justifica el desarrollo de un patrón de referencia con el que se puede realizar la valoración nutricional de la mujer en cualquier momento del embarazo.

La metodología presentada en este artículo brinda una útil herramienta que puede conducir al mejoramiento de la eficacia en el diagnóstico del bajo peso y el sobrepeso materno, teniendo en cuenta las características de la población. La elaboración de instrumentos de evaluación adaptados a las condiciones de cada nación puede favorecer las acciones pertinentes para poner en marcha actividades educativas, de intervención y de vigilancia nutricional, entre otras, encaminadas a mejorar la salud de la embarazada.

2. Materiales y métodos

Población: el universo de estudio lo componen embarazadas con edades entre 20 y 39 años, atendidas en los consultorios[1] de la red de policlínicos de los 15 municipios de la capital habanera, en los cuales residen mujeres provenientes de las diferentes regiones del país y que representan todos los estratos socioambientales, según los datos de la encuesta de migraciones internas y estadísticas continuas (Montes, Sanmarful & Lantigua 2003, González-Rego 2003).

Criterios de inclusión y exclusión: las mujeres incluidas en el estudio asistieron a su primera consulta prenatal antes de la semana 13, se encontraban clínicamente sanas, sin anomalías genéticas, ni patologías que pudieran afectar el desarrollo fetal, no eran fumadoras, ni consumían alcohol u otras drogas. Fueron excluidas mujeres con embarazos gemelares o con complicaciones obstétricas que influyeran en el crecimiento del feto (diabetes gestacional, hipertensión inducida por el embarazo u otra patología que debute en el embarazo). También se excluyeron partos con edades gestacionales menores que 37 semanas y mayores de 42 semanas.

Muestra: al comienzo de la investigación se proyectó un tamaño de muestra teniendo en cuenta su representatividad con respecto a las estadísticas de los nacidos vivos en instituciones de salud de los últimos años previos al estudio. Según estas consideraciones, teniendo en cuenta el número de muertes maternas y delimi-

[1] En Cuba existe un programa nacional de atención materno-infantil que prescribe los cuidados prenatales a todas las embarazadas (sanas o enfermas) en cada uno de los consultorios médicos del país.

tando el rango de edad de interés para el estudio, la muestra quedó compuesta por 7000 embarazadas entre 20 y 39 años que asistieron, durante septiembre del 2004 y diciembre del 2006 a las consultas de atención prenatal de todos los consultorios del médico de familia pertenecientes a cada área de salud de los 15 municipios de la capital del país y que cumplían con los criterios de inclusión y exclusión requeridos. Finalmente, después de un análisis exploratorio y un proceso de limpieza de datos, la muestra quedó formada por 6750 embarazadas.

Debe señalarse que las consultas de atención prenatal se realizan invariablemente en los consultorios del médico de familia, de acuerdo con una carpeta metodológica establecida por el Ministerio de Salud Pública de la República de Cuba, va dirigida a todas las gestantes cubanas, independientemente de su condición socioeconómica, educacional, cultural y de salud.

Antropometría: las mediciones antropométricas tomadas en cada embarazada comprendieron la estatura (cm) en la primera consulta prenatal y el peso (kg) en diferentes momentos del embarazo, según las técnicas indicadas (Lohman, Roche & Martorell 1988, Díaz, Montero, Jiménez, Wong & Moreno 2008*a*). El período de observación para cada embarazada estuvo comprendido entre la semana 13 y la consulta de término del embarazo.

Estudio exploratorio: la muestra se dividió en estratos según 12 rangos de estatura (ver tabla 1), determinados de forma tal que se garantizaran más de 200 observaciones por cada semana de embarazo, lo cual corresponde con los criterios recomendados por la Organización Mundial de la Salud (WHO 1995).

Tabla 1: Estratos según rangos de estatura.

Estratos	Estatura (cm)	Estrato	Estatura (cm)	Estrato	Estatura (cm)
1	140.0-150	5	156.1 - 158	9	164.1 - 166
2	150.1-152	6	158.1 - 160	10	166.1 - 168
3	152.1-154	7	160.1 - 162	11	168.1 - 170
4	154.1-156	8	162.1 - 164	12	>170

En la figura 1 se muestra la distribución por estratos de la estatura de las embarazadas.

Se calculó el índice de masa corporal $\left(IMC = peso\,(kg)\,/estatura^2\left(m^2\right)\right)$ de cada mujer en el momento de la captación[2] y se determinaron los percentiles 10, 75 y 90 de este indicador. Estos permiten identificar posibles riesgos de malnutrición de acuerdo con la clasificación presentada en la tabla 2.

En la figura 2.a se muestra la distribución del peso en el momento de la captación para las embarazadas incluidas en cada uno de los estratos establecidos. Como es de esperar, el peso promedio y la varianza muestral son mayores en los estratos con mayor estatura; se observa además que en todos los estratos aparecen casos extremos (por exceso). En la figura 2.b se muestra la distribución del peso en el momento de la captación según el estado nutricional de la mujer al inicio del embarazo. En este gráfico se observan también casos extremos.

[2] Primera consulta prenatal.

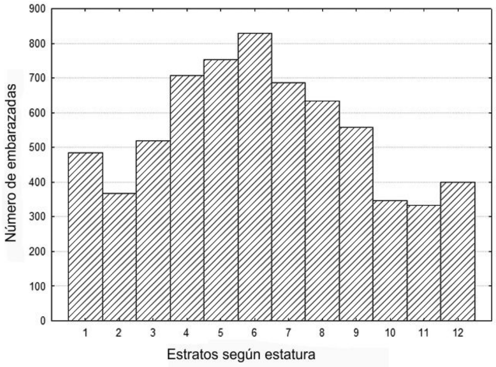
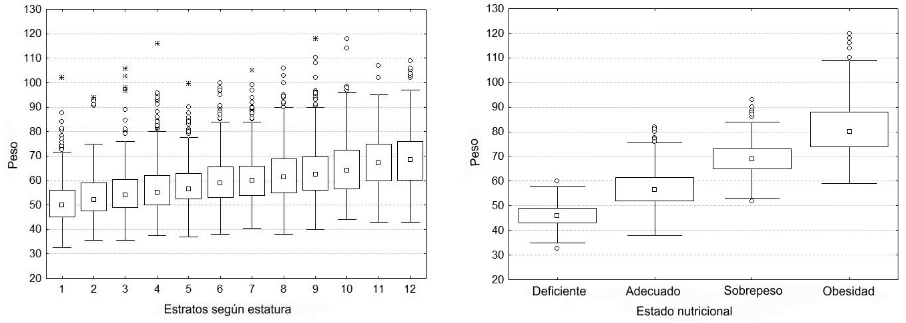


FIGURA 1: Distribución por estratos de la estatura de las embarazadas.

TABLA 2: Estado nutricional según IMC en el momento de la captación.

Estado nutricional	IMC en el momento de la captación
Peso deficiente	IMC $\leq$ 18.8 (*percentil* 10)
Peso adecuado	18.8 (*percentil* 10) $>$ IMC $>$ 25.6 (*percentil* 75)
Sobrepeso	25.6 (*percenti* 75) $\geq$ IMC $\geq$ 28.6 (*percentil* 90)
Obesidad	IMC $\geq$ 28.6 (*percentil* 90)



(a) (b)

FIGURA 2: Distribución del peso por (a): estratos según rangos de estatura y (b): estado nutricional según IMC en el momento de la captación.

En la figura 3 se muestra, para cada uno de los estratos, los perfiles-tiempo del peso (en kg) de cada embarazada medido durante 28 ocasiones (de la semana 13 a la 40 de embarazo). Estos gráficos indican una primera idea de la existencia de la relación lineal entre la respuesta y el tiempo. El análisis señaló además una amplia variación del peso inicial entre las embarazadas y un patrón similar en el comportamiento del peso durante el período de evaluación.

Se observa que en todos los estratos hay perfiles que se separan de la mayoría. Algunos de estos casos podrían pertenecer a una población diferente del cuerpo

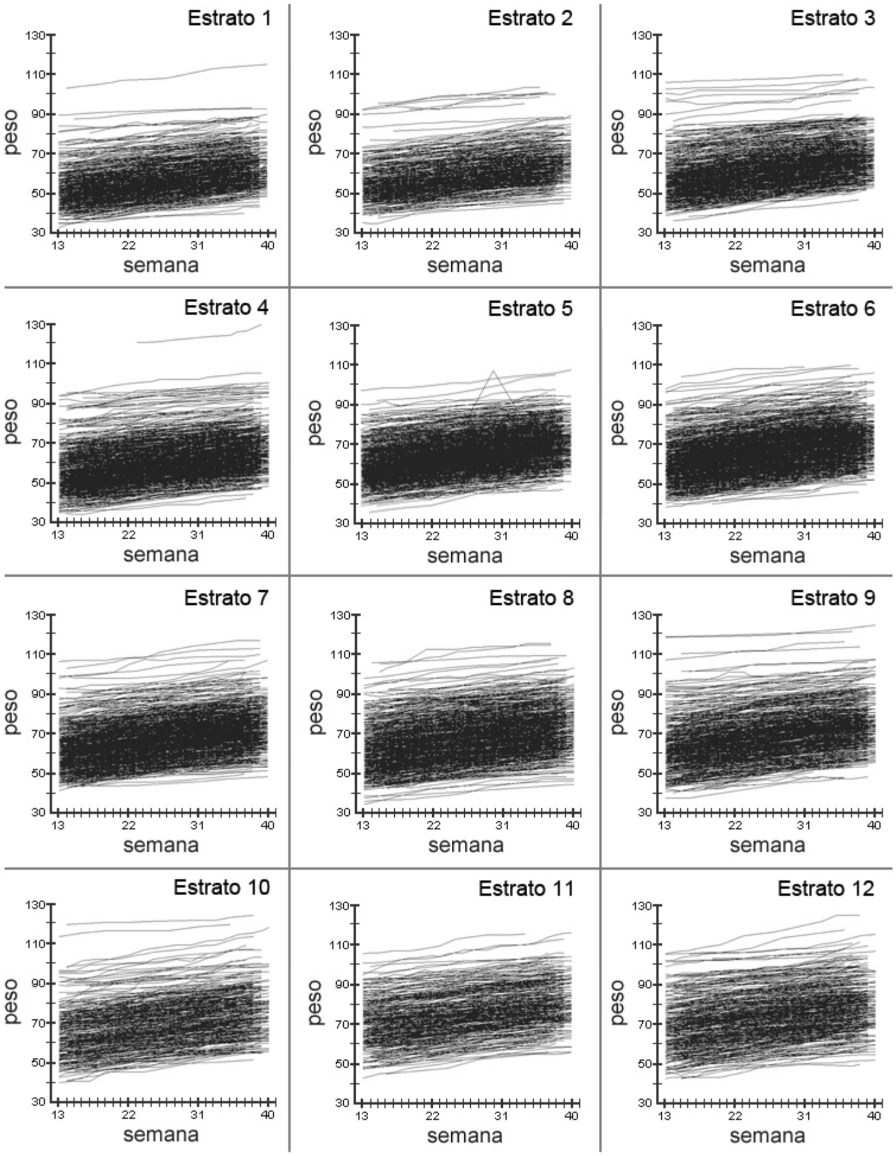


Figura 3: Perfiles-tiempo del peso de las embarazadas en cada estrato.

principal de la población de embarazadas. En el estrato 5 puede observarse un punto aberrante en la semana 30 de una de las embarazadas. Otros patrones atípicos o estructuras comunes son muy difíciles de visualizar debido, al tamaño considerable de las muestras en cada estrato. Para la detección de las observaciones atípicas, se subdividió cada estrato en muestras más pequeñas y se realizó un minucioso análisis gráfico exploratorio para cada submuestra.

En la variabilidad del peso entre las embarazadas de la muestra puede estar influyendo un número considerable de factores (IOM 1990*b*, Rasmussen & Yaktine 2009), en este trabajo se consideran los antropométricos: estatura e IMC.

Modelos multinivel: el análisis estadístico se abordó desde un enfoque de modelación multinivel para medidas repetidas (Beacon & Thompson 1982, Goldstein 1995, Quené & Huub 2004). La variación del peso de cada embarazada a través del tiempo ocurre en el nivel-1 y la variación de los pesos entre las embarazadas ocurre en el nivel-2.

Sea y_{it} el peso de la embarazada i en la ocasión t y x_{it} la variable indicadora del momento en que se hace la medición ($i = 1, ..., n$; $t = 1, \dots, m_i$). Para cada una de las embarazadas el comportamiento de la respuesta observada a través del tiempo se puede investigar mediante el modelo de nivel-1:

$$y_{it} = \beta_{0i} + \beta_{1i} x_{it} + e_{it} \tag{1}$$

donde β_{0i} y β_{1i} son el intercepto y la pendiente para la i-ésima ecuación de nivel-1 y e_{it} captura la variación del peso en el momento t sobre la curva de crecimiento individual i, tal que $E\left[y_{it}\right] = \beta_{0i} + \beta_{1i} x_{it}$, suponiendo que $E\left[e_{it}\right] = 0$ y $var\left[e_{it}\right] = \sigma_e^2$.

Para modelar la variación entre las embarazadas se formularon ecuaciones adicionales de nivel-2, donde uno o ambos de los parámetros de regresión de nivel-1 se modelan como la suma de una media general más una desviación aleatoria de la media. Durante el desarrollo de la modelación se pueden incluir, además, variables explicativas de nivel-2 invariantes en el tiempo. Por ejemplo, en el caso de un único predictor continuo z_i, los modelos de nivel-2 se componen de las siguientes ecuaciones:

$$\begin{aligned} \beta_{0i} &= \gamma_{00} + \gamma_{01} z_i + u_{0i} \\ \beta_{1i} &= \gamma_{10} + \gamma_{11} z_i + u_{1i} \end{aligned} \tag{2}$$

donde u_{0i} y u_{1i} son errores aleatorios con esperanza y varianza:

$$E\left(u_{0i}\right) = E\left(u_{1i}\right) = 0$$

$$var\left(u_{0i}\right) = \sigma_{u_0}^2, \qquad var\left(u_{1i}\right) = \sigma_{u_1}^2, \qquad cov\left(u_{0i}, u_{1i}\right) = \sigma_{u_0 u_1} \tag{3}$$

Este modelo de dos niveles puede escribirse como una única ecuación sustituyendo las ecuaciones (2) en la ecuación (1). Reordenando términos se obtiene el modelo combinado:

$$y_{it} = \gamma_{00} + \gamma_{10} x_{it} + \gamma_{01} z_i + \gamma_{11} x_{it} z_i + u_{0i} + u_{1i} x_{it} + e_{it} \tag{4}$$

El término $x_{it}z_i$ representa el efecto de interacción entre los predictores de los dos niveles. Los errores de nivel-2 se interpretan como las desviaciones del intercepto y la pendiente de la embarazada i con respecto a los valores medios de la población, después de haber controlado el predictor de nivel-2. El segmento $(\gamma_{00} + \gamma_{10}x_{it} + \gamma_{01}z_i + \gamma_{11}x_{it}z_i)$ en la ecuación (4) contiene todos los coeficientes fijos y se le conoce como parte fija. El segmento $(u_{0i} + u_{1i}x_{it} + e_{it})$ contiene todos los términos que representan los errores aleatorios y se le conoce como parte aleatoria. La parte fija no varía entre las embarazadas y la parte aleatoria es susceptible de variar entre las embarazadas. Dentro de este contexto, los términos $\gamma_{00}, \gamma_{10}, \gamma_{01}, \gamma_{11}$ se denominan parámetros fijos y las varianzas/covarianzas de los errores aleatorios se denominan parámetros aleatorios.

Estimación de los parámetros: para estimar los parámetros desconocidos se utilizó el procedimiento Mínimos Cuadrados Generalizados Iterativos Restringidos (RIGLS, por su sigla en inglés), el cual supone que la distribución de los errores es normal multivariada. En la práctica, sin embargo, las muestras son finitas y la suposición de normalidad puede ser violada, posiblemente dando lugar a estimaciones sesgadas y errores estándar inapropiados. Para corregir los errores estándar y las estimaciones de los intervalos de confianza de forma tal que sean menos dependientes de la distribución supuesta, se usa como procedimiento alternativo de estimación un método Bootstrap (Efron & Gong 1983, Laird & Louis 1989), donde los límites de confianza se calculan a partir de los percentiles de las estimaciones bootstrap.

Después del ajuste de cada modelo se realizó un análisis del cumplimiento de las hipótesis utilizando métodos gráficos de diagnóstico, se estudiaron las observaciones atípicas y se analizaron las que tenían una mayor influencia en el modelo. En los casos en que se consideró apropiado se excluyeron las observaciones anómalas y se hizo un reajuste del modelo.

3. Resultados y discusión

Análisis multinivel: en esta sección se presenta un resumen del desarrollo de los cuatro modelos propuestos, ajustados para todos los conjuntos de datos. A modo de ilustración, en la tabla 2 se muestran las estimaciones de los parámetros asociados al estrato 5, dentro de cuyos límites se ubica el valor estimado de la estatura de la población cubana femenina.

Para cada embarazada (unidades de nivel-2) seleccionada de la población de interés se tienen hasta 28 mediciones (unidades de nivel-1) de su peso. Sea *Sem* la variable indicadora de la semana de gestación en que se efectúa la medición. Para iniciar el proceso de ajustar el efecto del tiempo sobre el peso de la embarazada se usó el siguiente modelo:

$$y_{it} = \gamma_{00} + \gamma_{10}\,(Sem)_{it} + u_{0i} + e_{it} \qquad \text{(modelo A)}$$

En el modelo A se considera una variabilidad del peso materno en la semana 13 de embarazo; sin embargo, el efecto del tiempo de gestación (semana de embarazo), se modela como constante, o sea, se supone que el ritmo de crecimiento es el mismo para todas las embarazadas. La desviación del peso de la i-ésima embarazada al inicio de la gestación, con respecto a los valores promedios de la población, queda expresada mediante el término aleatorio u_{0i}. Con el objetivo de permitir la variación del efecto del tiempo entre las diferentes embarazadas, se introduce en la ecuación el término de error u_{1i}, de modo que el nuevo modelo se exprese como:

$$y_{it} = \gamma_{00} + \gamma_{10}(Sem)_{it} + u_{0i} + u_{1i}(Sem)_{it} + e_{it} \qquad \text{(modelo B)}$$

Como se deduce de la tabla 3, existe una fuerte evidencia de la supuesta variación del efecto asociado al tiempo de embarazo, ya que el cambio del modelo A al B en la log-verosimilitud (35786.53-29495.1=6291.43), comparado con una distribución con 2 gl, es significativo. Correspondientemente, se reduce el valor estimado de la varianza, “dentro” de las embarazas (de 2.269 a 0.781). Para explicar la variación del peso de las embarazadas en la semana 13 se incluye en el modelo la variable IMC, que representa el IMC en el momento de la capacitación, centrado con respecto a su mediana muestral (22.9) para todas las embarazadas; así, cuando esta variable toma el valor cero, la respuesta en el modelo se interpreta como el peso corporal en la semana t para las embarazadas con un IMC correspondiente al percentil 50. Ahora, el modelo toma la forma:

$$y_{it} = \gamma_{00} + \gamma_{10}\,(Sem)_{it} + \gamma_{01}\,(IMC)_{i} + u_{0i} + u_{1i}\,(Sem)_{it} + e_{it} \qquad \text{(modelo C)}$$

La importante disminución (de 84.029 a 2.721) del valor estimado de $\sigma^2_{u_0}$ es un indicador de que la variación entre las embarazadas en el parámetro intercepto depende, como es de esperar, del estado nutricional al comienzo de su gestación. Finalmente, se introduce la variable IMC como un predictor que también podría influir en las divergencias del comportamiento del peso durante el embarazo; así, el modelo queda expresado mediante la siguiente ecuación:

$$\begin{aligned} y_{it} = \gamma_{00} + \gamma_{10}\,(Sem)_{it} + \gamma_{01}\,(IMC)_{i} + \gamma_{11}(Sem \times IMC)_{it} + u_{0i} \\ + u_{1i}\,(Sem)_{it} + e_{it} \quad \text{(modelo D)} \end{aligned}$$

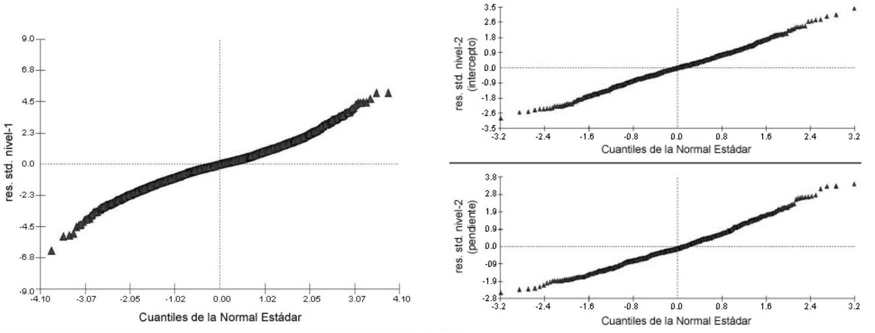
La introducción de γ_{11}, que representa el impacto medio por cada unidad de cambio en la variable IMC sobre el efecto de la semana de embarazo, no provoca cambios importantes en los valores del resto de los parámetros; sin embargo, el valor estimado es significativo (tabla 3) y se destaca su signo negativo, que corresponde con la interpretación biológica esperada.

En la figura 4 se presentan los gráficos para el diagnóstico de las diferencias entre la distribución normal y la distribución de probabilidad de la población de la que se ha extraído la muestra aleatoria de embarazadas pertenecientes al estrato 5. Los gráficos muestran que la distribución asintótica de los residuos de nivel-1 y la de los residuos de nivel-2 son aproximadamente normales. No se observan residuos correspondientes a observaciones marcadamente atípicas.

El proceso de modelación y análisis para el resto de los estratos fue similar al realizado para la muestra perteneciente al estrato 5. En todos los casos se comprobó que la variable respuesta tiene una distribución normal y que los tamaños de

Tabla 3: Estimación (errores estándar) de los parámetros de los modelos A, B, C y D para los datos correspondientes al estrato 5.

Parámetros	Modelos			
	A	B	C	D
Fijos				
γ_{00} $(const)$	58.099 (0.335)	58.123 (0.345)	56.761 (0.066)	56.744 (0.066)
γ_{10} (Sem)	0.481 (0.002)	0.478 (0.006)	0.478 (0.006)	0.483 (0.006)
γ_{01} (IMC)			2.445 (0.017)	2.476 (0.018)
γ_{11} $(Sem \times IMC)$				−0.009 (0.002)
Aleatorios				
nivel 2				
$\sigma^2_{u_0}$	78.278 (4.174)	84.029 (4.488)	2.721 (0.163)	2.711 (0.164)
σ^2_{u1}		0.023 (0.001)	0.023 (0.001)	0.022 (0.001)
$\sigma_{u_0 u_1}$		-0.365 (0.056)	−0.003 (0.011)	−0.059 (0.010)
nivel 1				
σ^2_e	2.269 (0.036)	0.781(0.013)	0.781(0.013)	0.781(0.013)
−2(log-verosimilitud)	35786.530	29495.100	27139.770	27106.390



(a) (b)

Figura 4: Normales para los residuos estandarizados de (a): nivel-1 y (b): nivel-2.

muestra por estrato (entre 334 y 830 embarazadas) en el nivel-2 pueden considerarse adecuados (Kreft 1996, Monk 1995). En el nivel-1, sin embargo, el tamaño de muestra puede ser muy pequeño, debido a observaciones perdidas (algunas embarazadas sólo tienen dos mediciones). Para ajustar los resultados de las estimaciones se aplicó el método bootstrap paramétrico (Rasbash, Browne, Goldstein, Yang, Plewis, Healy, Woodhouse, Draper, Langford & Lewis 2000). En cada caso se generaron 1000 muestras bootstrap y los parámetros desconocidos se estimaron utilizando el algoritmo RIGLS. Las estimaciones bootstrap y los errores estándar son las medias y las desviaciones estándar de las 1000 muestras bootstrap. El método bootstrap paramétrico también se empleó para obtener intervalos de confianza que se basan en los percentiles suavizados de las 1000 réplicas. En las tablas 4, 5, 6 y 7, se muestran, para propósitos de comparación, los resultados de las estimaciones RIGLS y bootstrap, correspondientes al modelo D en cada uno de los doce estratos. Las semejanzas en la amplitud de los intervalos de confianza Wald y los intervalos bootstrap reflejan la calidad de las estimaciones de todos los parámetros; por tanto, las inferencias derivadas de las estimaciones RIGLS pueden considerarse realistas.

Para la predicción del peso esperado sólo se consideran las estimaciones de los parámetros fijos; no obstante, en cada modelo deben tenerse en cuenta las estimaciones de los parámetros aleatorios, ya que son un indicador de cuánta varianza residual queda como un potencial para ser "explicado" por variables de los dos niveles.

Tabla 4: Estimaciones, errores estándar e intervalos de confianza para el parámetro γ_{00} del modelo D según los métodos RIGLS y bootstrap.

Estrato	estimación		E. S.		I.C. 95 %			
	IGLS	Boot.	IGLS	Boot.	IGLS	Boot.	IGLS	Boot.
1	49.77	49.77	0.11	0.12	49.54	49.55	49.50	50.00
2	52.55	52.55	0.08	0.08	52.39	52.40	52.70	52.70
3	53.70	53.69	0.08	0.08	53.55	53.53	53.85	53.85
4	55.00	55.00	0.07	0.06	54.87	54.87	55.13	55.13
5	56.74	56.74	0.07	0.07	56.61	56.61	56.87	56.88
6	58.18	58.17	0.06	0.06	58.06	58.05	58.30	58.30
7	59.58	59.59	0.08	0.07	59.43	59.44	59.73	59.71
8	60.91	60.91	0.08	0.07	60.76	60.76	61.06	61.06
9	62.36	62.35	0.08	0.08	62.20	62.19	62.52	62.51
10	63.91	63.91	0.10	0.10	63.71	63.70	64.11	64.11
11	65.82	65.82	0.10	0.10	65.63	65.62	66.01	66.01
12	68.74	68.74	0.15	0.15	68.44	68.44	69.03	69.04

Tabla 5: Estimaciones, errores estándar e intervalos de confianza para el parámetro γ_{10} del modelo D según los métodos RIGLS y bootstrap.

Estrato	estimación		E. S.		I. C. 95 %			
	IGLS	Boot.	IGLS	Boot.	IGLS	Boot.	IGLS	Boot.
1	0.46	0.46	0.01	0.01	0.45	0.45	0.47	0.48
2	0.47	0.47	0.01	0.01	0.46	0.45	0.48	0.48
3	0.49	0.49	0.01	0.01	0.48	0.47	0.50	0.50
4	0.47	0.47	0.01	0.01	0.46	0.46	0.48	0.48
5	0.48	0.48	0.01	0.01	0.47	0.47	0.49	0.49
6	0.49	0.49	0.01	0.01	0.48	0.48	0.50	0.50
7	0.49	0-49	0.01	0.01	0.48	0.48	0.50	0.50
8	0.49	0.49	0.01	0.01	0.48	0.48	0.50	0.50
9	0.49	0.49	0.01	0.01	0.48	0.47	0.50	0.50
10	0.50	0.50	0.01	0.01	0.48	0.48	0.52	0.52
11	0.49	0.49	0.01	0.01	0.47	0.47	0.51	0.51
12	0.48	0.48	0.01	0.01	0.46	0.46	0.50	0.50

En la tabla 8 se muestran las estimaciones RIGLS (y errores estándar) de los parámetros aleatorios del modelo D para todos los estratos. Puede observarse que en los estratos extremos, donde los rangos de estatura son más amplios, los valores de las estimaciones de la varianza del intercepto son mayores. Esto es un indicador de que en estos estratos el peso inicial de al menos una de las embarazadas se desvía considerablemente del intercepto de la curva de regresión media. Sin embargo, los valores de las estimaciones de las varianzas de la pendiente media, aunque significativos, son casi constantes para todos los estratos y representan una proporción muy pequeña de la varianza total, interpretándose que las mayores di-

Tabla 6: Estimaciones, errores estándar e intervalos de confianza para el parámetro γ_{01} del modelo D según los métodos RIGLS y bootstrap.

Estrato	estimación		E. S.		I. C. 95 %			
	IGLS	Boot.	IGLS	Boot.	IGLS	Boot.	IGLS	Boot.
1	2.09	2.09	0.03	0.03	2.04	2.04	2.14	2.14
2	2.29	2.29	0.02	0.02	2.55	2.55	2.33	2.32
3	2.37	2.37	0.02	0.02	2.34	2.34	2.40	2.40
4	2.43	2.43	0.01	0.02	2.40	2.39	2.46	2.45
5	2.48	2.47	0.02	0.02	2.44	2.44	2.51	2.51
6	2.55	2.55	0.02	0.02	2.52	2.52	2.58	2.59
7	2.61	2.61	0.02	0.02	2.57	2.57	2.65	2.65
8	2.69	2.69	0.02	0.02	2.65	2.66	2.72	2.73
9	2.72	2.72	0.02	0.02	2.68	2.69	2.76	2.76
10	2.80	2.80	0.02	0.02	2.75	2.75	2.85	2.85
11	2.88	2.88	0.02	0.02	2.83	2.84	2.93	2.93
12	2.98	2.98	0.04	0.02	2.91	2.91	3.05	3.06

Tabla 7: Estimaciones, errores estándar e intervalos de confianza para el parámetro γ_{11} del modelo D según los métodos RIGLS y bootstrap.

Estrato	estimación		E. S.		I. C. 95 %			
	IGLS	Boot.	IGLS	Boot.	IGLS	Boot.	IGLS	Boot.
1	−0.01	−0.01	0.00	0.00	−0.01	−0.01	−0.01	−0.01
2	−0.01	−0.01	0.00	0.01	−0.01	−0.01	−0.01	−0.00
3	−0.01	−0.01	0.00	0.00	−0.01	−0.02	−0.01	−0.01
4	−0.01	−0.01	0.00	0.00	−0.01	−0.01	−0.01	−0.01
5	−0.01	−0.01	0.00	0.00	−0.01	−0.01	−0.01	−0.01
6	−0.01	−0.01	0.00	0.00	−0.01	−0.01	−0.01	−0.01
7	−0.01	−0.01	0.01	0.00	−0.01	−0.01	−0.01	−0.00
8	−0.01	−0.01	0.01	0.01	−0.01	−0.01	−0.01	−0.01
9	−0.01	−0.01	0.00	0.01	−0.01	−0.02	−0.01	−0.01
10	−0.01	−0.01	0.00	0.00	−0.01	−0.01	−0.01	−0.00
11	−0.01	−0.01	0.00	0.00	−0.01	−0.02	−0.00	−0.01
12	−0.01	−0.01	0.00	0.00	−0.02	−0.02	−0.00	−0.01

ferencias en los patrones de crecimientos se deben principalmente al IMC inicial de las embarazadas. La covarianza negativa entre los errores de los dos niveles en todos los estratos confirma que la relación entre el IMC inicial y la ganancia de peso durante el embarazo se mantiene siempre inversa, independientemente de la estatura materna. Las estimaciones de las varianzas de nivel-1 no se presentan por no tener interpretación útil para el problema que se discute.

Interpretación de los parámetros: el desarrollo de los modelos propuestos para cada uno de los doce conjuntos de datos debe dar respuesta a tres aspectos fundamentales: 1. ¿En promedio, existe alguna tendencia del peso de la embarazada a través del tiempo? 2. ¿Existen diferencias entre las embarazadas con respecto a la tendencia del peso a través del tiempo? 3. ¿En caso de diferencias pueden encontrarse razones que las expliquen? A continuación se presenta un reporte de la interpretación realizada para los parámetros del modelo D en todos los estratos.

El parámetro intercepto representa la condición para la cual las variables explicativas en el nivel-1 y nivel-2 son cero. Por tanto, los valores en la primera columna de la tabla 4 estiman el peso esperado en la semana 13 para mujeres con un IMC correspondiente al percentil 50. Se observa que, en promedio, cuanto mayor es la estatura, mayor es el peso esperado al comienzo del embarazo.

El coeficiente asociado a la variable SEM (tabla 5) representa el incremento del peso por cada semana de gestación para embarazadas con un IMC en el percentil 50 al comienzo del embarazo. Se destaca el primer estrato, cuyas mujeres tienen una estatura menor de 150 cm, con el menor incremento ponderal.

El efecto positivo asociado a la variable IMC (tabla 6) representa la contribución del IMC en el momento de la captación en la evolución del peso de las embarazadas. Por tanto, mujeres con IMC en el momento de la captación por debajo de la mediana (IMC negativo) tendrán en promedio menor peso por semana de embarazo que las mujeres clasificadas en el mismo estrato según rangos de estatura pero con mayor masa corporal. Adicionalmente, se puede observar (tabla 6) que el efecto del IMC en el momento de la captación sobre el peso durante el embarazo es mayor cuanto mayor es la estatura de la mujer.

En la tabla 7 se muestra que el efecto de la interacción entre niveles es negativo y el mismo para todos los estratos. Esto significa, por ejemplo, que independientemente de la estatura, las mujeres que comienzan su gestación con un peso deficiente, alcanzan, en promedio, al final del embarazo, una ganancia de peso superior a las clasificadas con un estado nutricional adecuado, en sobrepeso u obesidad, lo que justifica que estas últimas sean las que logren menor ganancia ponderal.

En la tabla 8 se observa que, en todos los estratos, los valores estimados de la varianza residual, que queda como un potencial para ser explicado, son en general pequeños. Los mayores valores se encuentran en los estratos extremos, donde hay una mayor amplitud de los rangos de estatura, y como se deduce, una mayor variabilidad en el peso.

En general, los resultados obtenidos brindan un escenario de confianza para predecir, a partir de las variables explicativas consideradas en el modelo, las curvas de crecimiento del peso de las embarazadas representativas de la distribución de la población.

Para construir los indicadores antropométricos se ignoran los términos aleatorios del modelo y la predicción se obtiene evaluando valores específicos de las variables explicativas en el siguiente modelo marginal:

$$y_{it} = \gamma_{00} \ + \gamma_{10}\left(Sem\right)_{it} + \gamma_{01}\left(IMC\right)_{i} + \gamma_{11}\left(Sem \times IMC\right)_{it} \tag{5}$$

En la presente investigación se construyeron tablas antropométricas (Díaz, Montero, Jiménez, Wong & Moreno 2010*a*, Díaz, Montero, Jiménez, Wong & Moreno 2010*b*) para los percentiles 3, 10, 25, 50, 75, 90 y 97 del IMC. Como ilustración, en la tabla 9 se presentan los valores del peso corporal de referencia, correspondientes a mujeres ubicadas en el percentil 50 del IMC muestral (22.9 kg/m^2), estimados mediante el modelo predictivo para las embarazadas de todos los estratos a partir de la semana 13 y hasta la semana 40.

Los diferentes percentiles del IMC representan puntos de corte para la evaluación del estado nutricional de la embarazada en su primara consulta prenatal. El significado biológico de los puntos de corte del IMC se validó mediante un estudio sobre la correspondencia entre los niveles de riesgo de la antropometría materna y el peso del recién nacido (Montero & Díaz 2008).

Tabla 8: Estimaciones y errores estándar de los parámetros aleatorios del modelo D según el método RIGLS.

Estrato	$\sigma^2_{u_0}$		σ^2_{u1}		$\sigma_{u_0 u_1}$	
	IGLS	E. S.	IGLS	E. S.	IGLS	E. S.
1	5.62	0.39	0.02	0.00	-0.02	0.00
2	1.84	0.16	0.02	0.00	-0.03	0.01
3	2.50	0.18	0.02	0.00	-0.05	0.01
4	2.58	0.16	0.02	0.00	-0.06	0.01
5	2.71	0.16	0.02	0.00	-0.06	0.01
6	2.67	0.15	0.02	0.00	-0.04	0.01
7	3.27	0.20	0.02	0.00	-0.06	0.01
8	3.27	0.19	0.02	0.00	-0.04	0.01
9	2.99	0.21	0.02	0.00	-0.00	0.01
10	2.81	0.25	0.02	0.00	-0.04	0.02
11	2.53	0.23	0.02	0.00	-0.01	0.01
12	8.02	0.62	0.02	0.00	-0.09	0.03

Tabla 9: Peso corporal esperado por semana gestacional y por rangos de talla para las embarazadas con IMC en el momento de la captación en el percentil 50.

Semana	Estratos según rangos de estatura											
	1	2	3	4	5	6	7	8	9	10	11	12
13	50.2	53.0	54.2	55.5	57.2	58.7	60.1	61.4	62.8	64.4	66.3	69.2
14	50.7	53.5	54.7	55.9	57.7	59.2	60.6	61.9	63.3	64.9	66.8	69.7
15	51.2	53.9	55.2	56.4	58.2	59.7	61.1	62.4	63.8	65.4	67.3	70.2
16	51.7	54.4	55.7	56.9	58.7	60.2	61.5	62.9	64.3	65.9	67.8	70.7
17	52.1	54.8	56.2	57.3	59.2	60.6	62.0	63.4	64.7	66.4	68.3	71.1
18	52.6	55.3	56.7	57.8	59.7	61.1	62.5	63.9	65.2	66.9	68.8	71.6
19	53.1	55.7	57.2	58.3	60.2	61.6	63.0	64.4	65.7	67.4	69.3	72.1
20	53.6	56.2	57.7	58.7	60.6	62.1	63.5	64.9	66.2	67.9	69.8	72.6
21	54.0	56.7	58.2	59.2	61.1	62.6	64.0	65.4	66.6	68.4	70.3	73.1
22	54.5	57.1	58.7	59.7	61.6	63.1	64.5	65.9	67.1	68.9	70.8	73.5
23	55.0	57.6	59.2	60.2	62.1	63.6	65.0	66.4	67.6	69.4	71.3	74.0
24	55.4	58.0	59.7	60.6	62.6	64.1	65.5	66.9	68.1	69.9	71.8	74.5
25	55.9	58.5	60.2	61.1	63.1	64.6	66.0	67.4	68.6	70.4	72.3	75.0
26	56.4	58.9	60.7	61.6	63.6	65.1	66.5	67.9	69.0	70.9	72.8	75.5
27	56.9	59.4	61.2	62.0	64.1	65.6	67.0	68.4	69.5	71.4	73.3	75.9
28	57.3	59.9	61.7	62.5	64.6	66.1	67.5	68.9	70.0	71.9	73.8	76.4
29	57.8	60.3	62.2	63.0	65.0	66.6	67.9	69.4	70.5	72.4	74.3	76.9
30	58.3	60.8	62.7	63.4	65.5	67.1	68.4	69.9	70.9	72.9	74.8	77.4
31	58.8	61.2	63.2	63.9	66.0	67.6	68.9	70.4	71.4	73.4	75.3	77.9
32	59.2	61.7	63.7	64.4	66.5	68.1	69.4	70.9	71.9	73.9	75.8	78.3
33	59.7	62.1	64.2	64.8	67.0	68.5	69.9	71.4	72.4	74.4	76.3	78.8
34	60.2	62.6	64.7	65.3	67.5	69.0	70.4	71.9	72.8	74.9	76.8	79.3
35	60.6	63.1	65.2	65.8	68.0	69.5	70.9	72.4	73.3	75.4	77.3	79.8
36	61.1	63.5	65.7	66.2	68.5	70.0	71.4	72.9	73.8	75.9	77.8	80.3
37	61.6	64.0	66.2	66.7	68.9	70.5	71.9	73.4	74.3	76.4	78.3	80.7
38	62.1	64.4	66.6	67.2	69.4	71.0	72.4	73.9	74.7	77.0	78.8	81.2
39	62.5	64.9	67.1	67.6	69.9	71.5	72.9	74.4	75.2	77.5	79.3	81.7
40	63.0	65.3	67.6	68.1	70.4	72.0	73.4	74.9	75.7	78.0	79.8	82.2

Finalmente, la estimación de la varianza de los errores de nivel-2 con respecto a la pendiente aleatoria puede utilizarse para calcular los rangos de ganancia de peso semanal esperado en cada percentil del IMC. Por ejemplo, considerando las estimaciones de los parámetros del modelo D, un intervalo del 95 % de confianza para la ganancia promedio de una embarazada con un peso adecuado al comienzo de la gestación es de $0.4833 \pm 1.96 \times \sqrt{0.02} = [0.44, 0.52]$ unidades. Luego, la ganancia de peso “total” (de la semana 13 a la 40) recomendada para una embarazada con un patrón de comportamiento normal debe estar dentro del rango de 12.32 y 14.56 kg. Cuando el facultativo en salud observe ganancias bruscas de peso semanal podría recomendar a la embarazada modificar su peso hasta alcanzar ganancias ponderadas dentro del rango esperado.

4. Conclusiones

Se elabora, desde un enfoque multinivel, una metodología para la construcción de indicadores antropométricos del estado nutricional de la embarazada. A partir del procedimiento propuesto se obtuvieron las primeras referencias cubanas. Estas tienen la ventaja de relacionar el inicio con la evolución del embarazo, teniendo en consideración los diferentes rangos de estatura de la población.

Los resultados obtenidos en el estudio responden a las características propias de la población cubana, pero el procedimiento propuesto para la construcción de las tablas antropométricas puede ser también una estrategia favorable en investigaciones de otras regiones interesadas en construir sus propias referencias, conforme al contexto físico y sociocultural de la población de interés.

Los modelos propuestos logran explicar claramente el efecto del tiempo de gestación y el IMC inicial sobre el peso de la mujer durante el embarazo, además de que permiten una descripción de las diferencias entre las mujeres. Según los parámetros estimados, las mujeres que comienzan su gestación con un estado nutricional deficiente alcanzan en promedio una ganancia de peso superior a las que lo hacen estando en sobrepeso u obesidad.

Las tablas construidas a partir de los modelos multinivel predictivos proporcionan canales de seguimiento que permiten identificar con claridad las posibles desviaciones en la trayectoria ponderal a través del embarazo. Es posible, además, calcular los rangos de ganancia de peso semanal esperado según el estado nutricional inicial.

Los indicadores antropométricos resultantes de los modelos propuestos se construyeron tratando de mantener un diseño simple que convierte la estrategia en una herramienta prácticamente útil; no obstante, la flexibilidad del enfoque multinivel permite la construcción de modelos alternativos considerando otros efectos.

Agradecimientos

Esta investigación fue financiada por el programa de cooperación República de Cuba-UNICEF: Atención integral en edades tempranas (proyecto: tablas antropométricas para la evaluación nutricional de la mujer embarazada).

[Recibido: septiembre de 2010 — Aceptado: febrero de 2011]

Referencias

Beacon, H. J. & Thompson, S. G. (1982), 'Multi-level models for repeated measurement data: Application to quality of life data in clinical trial', *Statistics in Medicine* **15**, 2717–2732.

Brik, A. S. & Raudenbush, S. W. (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods.*, Sage Publications, California, USA.

Díaz, M. E., Montero, M., Jiménez, S., Wong, I. & Moreno, V. (2008*a*), *Diseño y confección de las tablas antropométricas de la embarazada cubana*, UNICEF-INHA, La Habana, Cuba.

Díaz, M. E., Montero, M., Jiménez, S., Wong, I. & Moreno, V. (2008*b*), Tablas antropométricas para la evaluación nutricional de la mujer embarazada, Informe, Programa de cooperación República de Cuba-Unicef, La Habana.

Díaz, M. E., Montero, M., Jiménez, S., Wong, I. & Moreno, V. (2009), 'Tablas antropométricas para la evaluación nutricional de la gestante', *Revista Chilena de Nutrición* **36**(suplemento 1), 382.

Díaz, M. E., Montero, M., Jiménez, S., Wong, I. & Moreno, V. (2010*a*), 'A weight gain chart for cuban pregnant women', *Public Health Nutrition* **13**(9), 138.

Díaz, M. E., Montero, M., Jiménez, S., Wong, I. & Moreno, V. (2010*b*), 'Tablas de referencias para el monitoreo del estado nutricional de la mujer embarazada', *Revista Española de Nutrición Comunitaria* **25**(1), 157.

Efron, B. & Gong, E. (1983), 'A leisurely look at the Bootstrap, the Jacknife and Cross-validation', *The American Statistician* **37**(1), 36–48.

Fescina, R. H. (1986), 'Aumento de peso durante el embarazo. Método para su cálculo cuando se desconoce el peso habitual', *American Journal of Clinical Nutrition* **90**, 156–162.

Goldstein, H. (1995), *Multilevel Statistical Models*, 2 edn, Halsted Press, New York.

González-Rego, R. A. (2003), 'Diferenciación socioambiental en áreas urbanas. El caso de La Habana', *Cuadernos Geográficos* (33), 105–132.

Gueri, M., Jutsum, P. & Sorhaindo, B. (1982), 'Anthropometric assessment of nutritional status in pregnant women: A reference table of weight-for-height by week of pregnancy.', *American Journal of Clinical Nutrition* (35), 609–611.

IOM (1990*a*), *Nutrition during pregnancy*, National Academy Press, Washington, DC.

IOM (1990*b*), *Nutrition RiskŨcriteria: A Scientific Assessment*, National Academy Press, Washington, DC.

Krasovec, K. & Anderson, M. A. (1991), 'Maternal nutrition and pregnancy outcomes. Anthropometric assessment', *PAHO Scientific Publication* **529**, 24.

Kreft, I. G. (1996), *Are Multilevel Techniques Necessary? An Overview, including Simulation Studies*, California State University Press, Los Angeles.

Laird, N. M. & Louis, T. L. (1989), 'Empirical Bayes confidence intervals for a series of related experiments', *Biometrics* **45**, 481–495.

Lohman, T. G., Roche, A. F. & Martorell, R. (1988), *Anthropometric Standardization Reference Manual*, Human Kinetics Books, A division of Human Kinetics Publishers, Illinois.

Lubin, J. H., Blot, W. J., Berrino, F., Flamant, R., Gillis, C. R., Kunze, M., Schmäwhl, D. & Visco, G. (1997), 'Design of a weight gain chart for pregnant women', *Revista Médica de Chile* **125**, 1437–1448.

Mardones, F. & Rosso, P. (2005), 'A weight gain chart for pregnant women designed in Chile', *Maternal and Child Nutrition* **7**(2), 77–90.

Monk, M. (1995), 'Sample size requirements for 2-level designs in educational research', *Multilevel Modelling Newsletter* **7**(2), 11–15.

Montero, M. & Díaz, M. E. (2008), *Antropometría materna y su relación con el peso del recién nacido*, Jornada Científica ICIMAF, La Habana.

Montes, N., Sanmarful, E. & Lantigua, G. (2003), 'Exploración sobre las migraciones internas en las provincias y los municipios de Cuba', *Cuadernos Geográficos* **33**, 43–53.

Quené, H. & Huub, v. d. B. (2004), 'On multilevel modeling of data from repeated measures designs: A tutorial', *Speech Communication* **43**, 103–121.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. & Lewis, T. (2000), *A user's guide to MlwiN*, Multilevel Models Proyect, University of London).

Rasmussen, K. M. & Yaktine, A. L. (2009), *Weight Gain During Pregnancy: Reexamining the Guidelines*, National Academies Press, Washington, DC.

Rosso, P. (1995), 'A new chart to monitor weight gain during pregnancy', *American Journal of Clinical Nutrition* **41**, 544–552.

Schwarcs, R., Díaz, A. G., Fescina, R. H., De Mucio, B., Belitzky, R. & Delgado, L. (1995), 'Salud reproductiva materna perinatal. Atención prenatal y del parto de bajo riesgo', *Publicación Científica* **1321.01**, 231.

WHO (1995), Physical status: the use and interpretation of anthropometry, Technical Report Series 854, World Heath Organisation (WHO), Geneva.

Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 289 a 307

Determinants of Birth Intervals in Tamil Nadu in India: Developing Cox Hazard Models with Validations and Predictions

Determinantes de los intervalos genésicos en Tamil Nadu (India): desarrollando modelos de riesgos de Cox con validaciones y predicciones

Rajvir Singh[1,a], Vrijesh Tripathi[2,b], Mani Kalaivani[3,c],
Kalpana Singh[4,d], S.N. Dwivedi[3,e]

[1]Medical Research Centre- Cardiology, CCS Department, Hamad Medical Corporation, Doha, Qatar

[2]Department of Mathematics & Statistics, Faculty of Science and Agriculture, The University of the West Indies, Trinidad & Tobago, West Indies

[3]Department of Biostatistics, All India Institute of Medical Sciences, New Delhi, India

[4]Clinical Research Department, Max Healthcare Institute Ltd., New Delhi, India

Abstract

The present study uses data from National Family Health Survey (NFHS-1) 1992-93 (International Institute for Population Sciences 1995) conducted in the state of Tamil Nadu, India. Cox models were developed to analyze the effect of breastfeeding as time varying and time dependent factor on birth intervals. Breastfeeding alone improved the log likelihood up to a higher level in each birth interval. Other factors that entered into the models were: at first birth interval, women's education (high school & above) and working status of women; at second birth interval, survival status of index child alive and husband's education (high school & above), and at third birth interval, breastfeeding more than 22 month were found to be protective factors for next births. Validation of the developed models was done through bootstrapping to predict birth intervals.

Key words: Cox model, Multivariate analysis, Validation, Predictions.

[a] Senior consultant. E-mail: rajvir.aiims@gmail.com
[b] Lecturer. E-mail: Vrijesh.Tripathi@sta.uwi.edu
[c] Scientist. E-mail: manikalaivani@gmail.com
[d] Researcher. E-mail: nkalpanasingh@gmail.com
[e] Professor. E-mail: dwivedi7@hotmail.com

Resumen

Este estudio utiliza datos de la Encuesta Nacional de Salud Familiar (International Institute for Population Sciences 1995) realizada en el estado de Tamil Nadu, India. Se desarrollaron modelos de Cox para analizar el efecto de la lactancia materna cuando varía en el tiempo y el factor tiempo depende de los intervalos genésicos. La lactancia materna sólo mejora la probabilidad de acceder a un nivel más alto en cada intervalo de nacimiento. Otros factores que entraron en los modelos fueron en el intervalo del primer parto: nivel educativo de la madre (secundaria y superior) y trabajo de la madre; en el intervalo del segundo parto: nivel de supervivencia en el índice de vida infantil y nivel educativo del padre (secundaria y superior), y en el intervalo del tercer parto: lactancia materna más 22 meses. Cada uno de los anteriores es un factor protector para ampliar el intervalo entre nacimientos en el estudio. Además, este estudio confirma los modelos desarrollados en los servicios públicos de predicción para los intervalos genésicos.

Palabras clave**:** análisis multivariado, modelo de Cox, predicciones, validación.

1. Introduction

Population change is a global phenomenon. This varies significantly among regions and even among countries within the same region. This also varies significantly among states within the same country. Cognizant of inherent problems in rapid population growth in developing countries like India, epidemiologists including biostatisticians, demographers and social scientists have given high priority to a thorough understanding of the differentials and determinants of this phenomenon. Rates of population change and various aspects of reproductive health need to be understood to understand this phenomenon. Birth interval is defined as interval between termination of one completed pregnancy and the termination of the next. The intrinsic growth rate as well as the mean generational length of any population may get affected by the birth interval pattern (Srinivasan 1980). Thus birth interval can be viewed as a major determinant of population change. The mechanism of reproductive process can be assessed through the analysis of birth interval. This is possible because the disaggregating of the reproductive process is possible into a series of stages, beginning with marriage followed by first birth, second birth, third birth and so on, provides an insight into the fertility behavior of the population which is principally responsible for population change. Emphasis has often been laid on delaying the first birth, interval births, avoiding too many births, and on stopping child bearing in time (UNFPA 1997). An appropriate epidemiological understanding of birth intervals in a region may be helpful to policy planners for an appropriate public health program for the region in the belief that such an attempt is likely to provide more accurate results, which would lead to more appropriate intervention.

2. Material and Methods

The National Family Health Survey (NFHS-I) is a state representative survey of ever-married women aged 13-49 years. Survey period was from 18th April, 1992 to 7th July, 1992 in Tamil Nadu (TN) (International Institute for Population Sciences 1995). Data were collected in the form of systematic, stratified sample of households with two stages in rural areas (selection of villages followed by selection of households) and three stages in urban areas (selection of cities/towns, followed by urban blocks, and finally households) in self weighting fashion. The number of households surveyed was 4,287 having 3,948 ever-married women. Out of them 66.3% were non-sterilized and currently married. The detailed reports covering sampling methods and all other aspects mentioned above were prepared and documented in Population Research Centre, The Gandhigram Institute of Rural Health and Family Welfare Trust, Ambathurai R.S., and International Institute for Population Sciences (1994).

The parity (birth order) specific hazards models for birth intervals in TN have been worked out utilizing available data on 627 women of parity-I, 566 women of parity-II, and 310 women of parity-III. The results provide information on factors associated with experiencing next live birth. The order of the interval was the parity of a woman; e.g., the first birth interval is time interval between effective age at marriage to first parity of women; the second birth interval is time interval between first parity and second parity; and so on. In other words (Trussell, Martin, Fledman, Palmore, Concepcion & Abu Bakar 1985), the order of the interval is the order of birth that would close the interval; e.g., the first birth interval extends from the effective age at marriage to the first birth; the second birth interval extends from the first birth to the second birth; the third birth interval extends from the second birth to third birth; and so on. Birth intervals were considered in months as interval variables in the analysis (Trussell et al. 1985).

Data regarding children from multiple births (including twins) were considered as single birth and included for the analysis. Further, it was decided to exclude birth interval during which there was no possibility of conception because the woman or her husband had been sterilized. As a result of preliminary analysis, the decision to exclude the higher order birth intervals (fourth onwards) from the analysis was taken, mainly because of insufficient number of records/events. Incomplete records, almost negligible in number, were not considered in the analysis.

All the variables in the study satisfied proportional hazard (PH) assumption except breastfeeding and were considered as fixed covariates with fixed effect. Age at index child was taken as continuous variable and did not satisfy linear assumption. Age was considered as time varying with fixed effect and age^2 was added to overcome the non-linearity. Breastfeeding was considered as time varying time dependant factor (Dwivedi & Rajvir 2003). As per the method followed by Trussell et al. (1985), the interval for first birth interval was divided into five categories: $\leq$ 15 months/16-21 months/22-27 month/28-33 month/$\geq$ 34 months. On account of lesser proportion of women experiencing next live birth under extreme categories, the live birth interval related to the second birth interval was divided into three categories: $\leq$ 21 months/22-27 months/$\geq$ 28 months. Similarly, live birth interval under third birth interval was categorized as $\leq$ 21 months and $\geq$ 22 months for meaningful analysis. All the variables having ($p < 0.25$) at univariate analysis were selected for multivariate Cox analysis and the variables ($p < 0.10$) at multivariate analysis were considered in the models.

Being breastfeeding as time varying covariate with time dependent effect, an extended Cox hazards model suggested by Trussell et al. (1985) was used. If birth interval categorized into k categories, general form of the extended Cox Hazards considered as under:

$$\lambda_k(t, X(t)) =$$
$$\lambda_{0k} t \exp\left[\sum_{i=1}^{p_{11}} \beta_{1i} X_{1i} + \sum_{i=1}^{p_{12}} \beta_{2ik}(t) X_{2i} + \sum_{j=1}^{p_{21}} \beta_{1j} X_{1jk}(t) + \sum_{j=1}^{p_{22}} \beta_{2jk}(t) X_{2jk}(t)\right]$$

where, $\lambda_{0k}(t)$ is category-specific baseline hazard; β_{1i} are respective fixed (time-independent) effects of fixed covariates X_{1i}; $\beta_{2ik}(t)$ are respective category-specific (time-dependent) effects of fixed covariates X_{2i}; β_{1j} are respective fixed effects of time-varying covariates $X_{1jk}(t)$; and $\beta_{2jk}(t)$ are respective category-specific (time-dependent) effects of time-varying covariates $X_{2jk}(t)$.

Maximum likelihood functions for extended Cox model were calculated to produce estimates of the coefficients and their standard errors (Trussell & Charles 1983). Using regression coefficients and respective standard errors, Risk Ratio or Hazard Ratio (HR) in the form of $\exp(\beta)$ related to an exposure variable and its 95% confidence interval were calculated and interpreted using standard convention followed in the case of Cox Proportional Hazard model (Kleinbaum 1996).

To satisfy the linearity assumption in the Cox PH models, at each time t, $\log \lambda(t)$ and equivalently $\log[-\log(S(t))]$ were linearly related to covariates, where $\lambda(t)$ was the hazard function or instantaneous event rate at time t and $S(t)$ was the probability of surviving until time t (not having next birth in the study).

Log-log survival curves (Cox 1972, Namboodiri & Suchindran 1987, Kleinbaum 1996) were assessed to check PH assumption of proportionality for each fixed effect with fixed covariate whereas; for continuous covariate i.e. woman's age, birth spacing was categorized as $\leq$ 15 months, 16-21 months, 22-27 months, 28-33 months, and $\geq$ 34 months, based on an exploratory analysis. For a procedure involving time dependent variable, presence of breastfeeding for these categories was specified as > 0, ≥ 16, ≥ 22, ≥ 28 and ≥ 34 months, respectively.

First order interactions between covariates were tested using stratified analysis and no interaction was found. Collinearity among the covariates was checked through correlation analysis (Fox 2008, pp. 307-331).

All covariates considered in the multivariate analysis were followed by stepwise method to select variables for inclusion or exclusion from the model in a sequential fashion. For this, a forward selection with a test for backward elimination was used with probability levels for entry and removal as 0.15 and 0.10, respectively. This was done in view of the fact that early deletion of covariates with little chance of being measured reliably or of being predictive would result in models with less overfitting and more generalization.

In order to test validation of developed models, bootstrapping was applied (Efron & Tibshirani 1993). Calibration curve with 200 re-samples were used to estimate the optimism between predicted survival probability estimates from the developed Cox model and the corresponding Kaplan Meier survival probability (Kaplan & Meier 1958). Shrinkage coefficient was calculated to check for overfitting of the model (Van Houwelingen & Cessie 1990), and discrimination aspect of the model was measured through Somer's D_{xy} rank correlation between predicted log hazard and observed survival time (Harrell, Lee & Mark 1996, Harrell 2001).

Predictive probabilities for a woman not attaining next live birth for a particular variable or combination of variables by holding other variables at their mean levels were estimated. The exponential expression of the Cox model, also known as "Risk score" and generally denoted by R, may be defined as follows: $R = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ (Dickson, Grambsch, Fleming, Fisher & Langworthy 1989, Singh, Begum, Ahuja, Chandra & Dwivedi 2007), where, $X_1, X_2, \ldots, X_p$ are the considered levels of p predictor variables and $\beta_1, \beta_2, \ldots, \beta_p$ are respective unknown regression coefficients.

Thus, using maximum likelihood estimates of regression coefficients for the model being used and substituting the observed values of the covariates for each individual, risk score is obtained for every person (woman) included in the data analysis. The arithmetic mean of these risk scores provides an average risk score R_1 and hence R_1 is constant for a given data set. Risk score R_2 is obtained again by using the equation substituting again the estimated values of the regression coefficients and changed levels of selected variable/set of variables (same level for every woman) but retaining other variables at their mean level. $S_0(t)$, the baseline survival probabilities at different points of time for a person with average risk score R_1 may be worked out using Kaplan Meier method. Thus, $S_0(t)$ at a given point of time is nothing but the survival probability obtained through Kaplan Meier method at that point of time. Gain in survival probability after adjustment in relation to considered levels of selected covariates has been obtained by $S(t) = S_0(t)^{\exp(R_2 - R_1)}$.

In the present study, for each model, survival probabilities in relation to R_1 are listed under first row of the concerned table, whereas those related to R_2 are listed in successive rows. Thus, differences between these two probabilities provide gain/loss as a result of proposed change in the levels of selected variable/set of variables. BMDP 7.0, University of California, 1992; S-Plus 4.0, 1988-97, Mathsoft Inc. Seatle , WA 98109-3044, USA and Excel 2000 Statistical Software were employed for the analysis.

3. Results

There is similarity in distribution of women for each birth interval in terms of religion/caste, place of residence, (ever) contraceptive use, (ever) fetal loss, sex of index child, survival status of index child, husband's occupation, type of house, media exposure, and distance of primary health center (Table 1). As the parity increased, there was an increase in the proportion of women in categories characterized by illiteracy, working status of women, illiteracy of husband and breastfeeding for 22 and more months. Women with high school (and above) education were significantly less likely to experience the next live birth in comparison to illiterate women in case of first and second birth interval, whereas education was not a significant factor in case of the third birth Interval (HR: 0.79; C.I.: 0.24-2.63). In addition, women with middle education were also significantly less likely to the experience the next live birth in case of second birth interval (HR: 0.50; C.I.: 0.27-0.92). Ever contraceptive use was a significant protective factor in case of first and second birth interval in contrast to third birth interval (HR: 0.83; C.I.: 0.38-1.81). Ever fetal loss was a significant protective factor only in case of first birth interval (HR: 0.70; C.I.: 0.50-0.97). Previous birth interval was not used in case of first birth interval. But, women with more than 36 months of previous birth interval were significantly less likely to experience the next live birth in case of second birth interval (HR: 0.54; C.I.: 0.33-0.88) but previous birth interval was not a significant factor for third birth interval (HR: 0.50; C.I.: 0.22-1.10). Further, women with surviving index child were significantly less likely to experience next live birth in case of first and second birth interval, but this was not the case for third birth interval (HR: 0.45; C.I.: 0.16-1.26). Husband's education (high school and above) was a significant protective factor against the next live birth only in the case of second birth interval (HR: 0.41; C.I.: 0.24-0.72). Shorter distance from primary health center was also a significant protective factor only in the case of second birth interval (HR: 0.68; C.I.: 0.47-0.99). Surprisingly, breastfeeding did not emerge as a significant predictor of birth interval in most cases. But, the period 1-15 months of breastfeeding did predict the first birth interval where it was noticed to be a significant risk factor (HR: 1.92; C.I.: 1.04-3.58).

Table 1: Covariates associated with parity specific birth intervals: definitions and means.

Variables	Category	Parity		
		1st Interval	2nd Interval	3rd Interval
Religion/caste	SC/ST Hindu	0.15	0.18	0.23
	Other Hindu	0.70	0.67	0.65
	Non-Hindu	0.15	0.15	0.12
Place of residence	Rural	0.61	0.61	0.67
	Urban	0.39	0.39	0.33
Women's education	Illiterate	0.35	0.43	0.58
	Primary	0.29	0.25	0.23
	Middle	0.17	0.15	0.08
	≥High school	0.19	0.17	0.11
Ever contraceptive use	No	0.69	0.66	0.72
	Yes	0.31	0.34	0.28
Ever fetal loss	No	0.79	0.76	0.72
	Yes	0.21	0.24	0.28
Previous birth interval	< 24 Months	- -	0.28	0.34
	24-36 Month	- -	0.39	0.26
	> 36 Months	- -	0.33	0.40

Table 1: Continue.

Variables	Category	Parity		
		1^{st} Interval	2^{nd} Interval	3^{rd} Interval
Sex of index child	Male	0.49	0.49	0.45
	Female	0.51	0.51	0.55
Survival status of index child	Alive	0.94	0.94	0.96
	Dead	0.06	0.06	0.04
Women's occupation	Not Working	0.72	0.68	0.54
	Working	0.28	0.32	0.46
Husband's occupation	Not working	0.03	0.03	0.02
	Working	0.97	0.97	0.98
Husband's education	Illiterate	0.19	0.20	0.27
	Primary	0.31	0.34	0.36
	Middle	0.18	0.16	0.16
	≥High School	0.32	0.30	0.21
Type of house	Kuchha	0.32	0.37	0.39
	SemiPucca+Pucca	0.68	0.63	0.61
Media exposure	No	0.16	0.19	0.22
	Yes	0.84	0.81	0.78
Distance of primary health	≥ 2 km	0.54	0.54	0.60
Center	< 2 km	0.46	0.46	0.40
Breastfeeding (months)	≥ 1	0.92		
	≥ 16	0.29	0.38*	
	≥ 22	0.11	0.17	0.22***
	≥ 28	0.04	0.05**	
	≥ 34	0.02		
Age of women at index child (years) (X SD)	Continuous	20.41 ± 3.51	22.36 ± 3.72	24.0 ± 4.01

*: for period of birth interval ≤ 21 months; **: for period of birth interval ≥ 28 months
***: for period of birth interval ≥ 22 months

Table 2: Univariate analysis and multivariate with extended cox model of first birth interval according to different variables in TN.

Variables	Categories	Univariate		Multivariate	
		$\exp(\beta)$	C.I. 95%	$\exp(\beta)$	C.I. 95%
Women's age at index child	Continuous			1.33	0.88 - 2.04
Women's age^2 at index child	Continuous			0.99	0.98 - 1.00
Religion/caste[a]	Non-Hindu	0.86	0.59 - 1.25		
	Other Hindu	1.21	0.76 - 1.91		
Place of residence[b]	Urban	0.96	0.73 - 1.27		
Women's education[c]	Primary	1.24	0.89 - 1.72	0.96	0.65 - 1.42
	Middle	1.05	0.71 - 1.54	0.92	0.58 - 1.45
	≥High school	0.58	0.38 - 0.90	0.40	0.22 - 0.75
Ever contraceptive use[d]	Yes	0.70	0.52 - 0.93		
Ever fetal loss[e]	Yes	0.70	0.50 - 0.97	0.63	0.44 - 0.88
Sex of index child[f]	Male	0.95	0.73 - 1.25	0.52	0.27 - 1.02
Survival status of index child[g]	Alive	1.86	1.18 - 2.96		
Women's occupation[h]	Working	0.76	0.56 - 1.03	0.72	0.52 - 1.00

Table 2: Continue.

Variables	Categories	Univariate		Multivariate	
		exp(β)	C.I. 95%	exp(β)	C.I. 95%
Husband's occupation[i]	Working	0.56	0.30 - 1.06		
Husband's education[j]	Primary	1.05	0.73 - 1.51	1.08	0.72 - 1.61
	Middle	0.86	0.56 - 1.32	0.66	0.41 - 1.08
	≥High school	0.70	0.48 - 1.03	0.75	0.45 - 1.26
Type of house[k]	Pucca+Semi Pucca	0.79	0.60 - 1.04		
Media exposure[l]	Yes	1.00	0.70 - 1.44		
Distance primary health centre[m]	< 2 km	0.90	0.69 - 1.19		
Birth interval	1-15 months	0.07	0.02 - 0.18	0.07	0.03 - 0.19
Breastfeeding[n]	$\geq$ 1 months	1.92	1.04 - 3.58	3.08	1.44 - 6.60
Birth interval	16-21 months	0.29	0.13 - 0.62	0.30	0.14 - 0.66
Breastfeeding[n]	$\geq$ 16 months	0.82	0.56 - 1.19	0.74	0.50 - 1.09
Birth interval	22-27 months	0.39	0.13 - 1.16	0.40	0.13 - 1.19
Breastfeeding[n]	$\geq$ 22 months	0.98	0.52 - 1.86	0.90	0.46 - 1.73
Birth interval	28-33 months	0.46	0.04 - 4.86	0.51	0.05 - 5.51
Breastfeeding[n]	$\geq$ 28 months	0.54	0.14 - 2.07	0.45	0.11 - 1.74
Birth interval	$\geq$ 34 months	0.85	0.04 - 16.4	0.74	0.04 - 14.70
Breastfeeding[n]	$\geq$ 34 months	0.47	0.05 - 4.70	0.45	0.04 - 4.69
Reference Categories:	a) SC/ST Hindu, b) Rural, c) Illiterate, d) No, e) No, f) Female, g) Dead, h) Not working, i) Not working, j) Illiterate, k) Kuccha, l) No, m) >= 2 km , n) Less than the given.				

Table 3: Univariate analysis and multivariate with extended cox model of second birth interval according to different variables in TN.

Variables	Categories	Univariate		Multivariate	
		exp(β)	C.I. 95%	exp(β)	C.I. 95%
Women's age at index child	Continuous			0.80	0.49 - 1.29
Women's age^2 at index child	Continuous			1.00	0.99 - 1.01
Religion/caste[a]	Non-Hindu	0.68	0.44 - 1.05		
	Other Hindu	0.67	0.36 - 1.26		
Place of residence[b]	Urban	0.70	0.48 - 1.03		
Women's education[c]	Primary	0.72	0.47 - 1.11		
	Middle	0.50	0.27 - 0.92		
	≥High school	0.46	0.25 - 0.83		
Ever contraceptive use[d]	Yes	0.65	0.44 - 0.97		
Ever fetal loss[e]	Yes	1.00	0.67 - 1.50		
Previous birth interval[f]	24-36 Months	0.87	0.58 - 1.31		
	$\geq$ 36 Months	0.54	0.34 - 0.88		
Sex of index child[g]	Male	1.11	0.78 - 1.58		
Survival status of index child[h]	Alive	4.30	2.45 - 7.54	0.40	0.22 - 0.72
Women's occupation[i]	Working	1.01	0.70 - 1.47		
Husband's occupation[j]	Working	1.22	0.38 - 3.85		
Husband's education[k]	Primary	0.87	0.56 - 1.37	0.72	0.46 - 1.13
	Middle	0.90	0.52 - 1.56	0.62	0.35 - 1.11
	≥High school	0.41	0.24 - 0.72	0.34	0.19 - 0.61

Table 3: Continue.

Variables	Categories	Univariate		Multivariate	
		exp(β)	C.I. 95%	exp(β)	C.I. 95%
Type of house[l]	Pucca+Semi Pucca	0.94	0.65 - 1.36		
Media exposure[m]	Yes	0.82	0.53 - 1.25		
Distance primary health centre[n]	< 2 km	0.68	0.47 - 0.99		
Birth interval	1-21 month	0.10	0.02 - 0.47	0.06	0.02 - 0.23
Breastfeeding[o]	≥ 1 month	0.86	0.53 - 1.39	0.99	0.59 - 1.68
Birth interval	22-27 month	0.07	0.01 - 0.57	0.04	0.01 - 0.57
Breastfeeding[o]	≥ 22 month	0.99	0.47 - 2.08	1.17	0.54 - 2.52
Birth interval	≥ 28 month	1.61	0.17 - 15.4	0.48	0.04 - 5.94
Breastfeeding[o]	≥ 28 month	0.55	0.07 - 4.47	1.07	0.10 - 11.0
Reference Categories:	a) SC/ST Hindu, b) Rural, c) Illiterate, d) No, e) No, f) < 24 Month, g) Female, h) Dead, i) Not working, j) Not working, k) Illiterate, l) Kuccha, m) No, n) ≥ 2 km, o) Less than the given.				

Table 4: Univariate analysis and multivariate with extended cox model of third birth interval according to different variables in TN.

Variables	Categories	Univariate		Multivariate	
		exp(β)	C.I. 95%	exp(β)	C.I. 95%
Women's age at index child	Continuous			0.74	0.43 - 1.26
Women's age^2 at index child	Continuous			1.00	0.99 - 1.02
Religion/caste[a]	Non-Hindu	1.72	0.66 - 4.45		
	Other Hindu	1.52	0.41 - 5.66		
Place of residence[b]	Urban	1.20	0.60 - 2.38		
Women's education[c]	Primary	0.76	0.33 - 1.75		
	Middle	0.63	0.15 - 2.65		
	$\geq$High school	0.79	0.24 - 2.63		
Ever contraceptive use[d]	Yes	0.83	0.38 - 1.81		
Ever fetal loss[e]	Yes	0.50	0.22 - 1.14		
Previous birth interval[f]	24-36 Months	0.90	0.42 - 1.95		
	≥ 36 Months	0.50	0.22 - 1.10		
Sex of index child[g]	Male	1.47	0.75 - 2.86		
Survival status of index child[h]	Alive	2.24	0.79 - 6.33		
Women's occupation[i]	Working	0.81	0.42 - 1.54		
Husband's occupation[j]	Working	0.49	0.07 - 3.61		
Husband's education[k]	Primary	1.45	0.62 - 3.40		
	Middle	0.91	0.30 - 2.79		
	$\geq$High school	1.13	0.42 - 3.02		
Type of house[l]	Pucca+Semi Pucca	1.10	0.57 - 2.13		
Media exposure[m]	Yes	0.53	0.26 - 1.07	0.50	0.24 - 1.02
Distance primary health centre[n]	< 2 km	0.99	0.51 - 1.92		

Table 4: Continue.

Variables	Categories	Univariate		Multivariate	
		exp(β)	C.I. 95%	exp(β)	C.I. 95%
Birth interval	1-21 months	0.73	0.04 - 12.8	0.79	0.04 - 13.98
Breastfeeding[o]	$\geq$ 1 months	1.73	0.23 -13.2	2.15	0.25 - 18.14
Birth interval	$\geq$ 22 months	1.63	0.15 - 17.4	1.50	0.14 - 16.16
Breastfeeding[o]	$\geq$ 22 months	0.13	0.02 - 1.00	0.13	0.05 - 1.00
Reference Categories:	a) SC/ST Hindu, b) Rural, c) Illiterate, d) No, e) No, f) < 24 Month, g) Female, h) Dead, i) Not working, j) Not working, k) Illiterate, l) Kuccha, m) No, n) $\geq$ 2 km, o) Less than the given.				

4. Multivariate Analysis

The final models consisted of varying subsets of variables for first, second, and third birth intervals. Variables that entered partially are considered fully in the presentation of final models for a meaningful presentation. In order to account for age which is a well-known confounder, woman's age at index child was forced into the model. Square of woman's age at index child was also considered in order to overcome the problem of non-linear relationship.

The first variable to enter in the model for each birth interval was breastfeeding. Also, for each birth interval, breastfeeding alone improved the log likelihood up to a higher level, clearly showed the inclusion of breastfeeding even partially at first step itself significantly improved the model. High improvement in chi-square with one degree of freedom was seen for each birth interval, the improvement being 28.3 for first birth interval, 39.9 for second birth interval and 17.7 for third birth interval. Surprisingly, under the first birth interval, the effect of breastfeeding persisted only during the period 0-15 months. Also, effect of breastfeeding under the second birth interval disappeared during each of the periods considered in the analysis. However, its effect again persisted under the third birth interval during the period 22 and more months. It may be worth reporting that the role of breastfeeding fell in line with that reported based on univariate analysis.

Before comparison of variables entered in the final extended Cox models related to various birth intervals, it may be noted that subsets of variables considered in the data analysis vary from first birth interval to third birth interval because varying periods of classification of breastfeeding were considered. Strictly speaking, this may prohibit a comparison among the models. However, a qualitative comparison of results presented in Tables 2 to 4 reveals that high school (and above) education of women was a significant protective factor under the first birth interval analysis (HR: 0.40; C.I.: 0.22-0.75). On the other hand, high school (and above) education of father (HR: 0.34; C.I.: 0.19-0.61) and survival status of index child (HR: 0.40; C.I.: 0.22-0.72) were significant protective factors under the second birth interval analysis. Media exposure entered into the model for the third birth interval. Ever-fetal loss, survival status of index child, occupation of woman and husband's education also entered into the model for the first birth interval. Hence, variables that entered into the models varied from the first birth interval to the third birth interval.

5. Validation of the Models

Calibration curves for extended Cox models for the birth intervals are shown in Figures 1 to 3. Except for one group with extremely bad prognosis in each figure, bias corrected calibrations are very good. Shrinkage coefficients related to first to third birth interval are 0.90, 0.92 and 0.78, respectively (Table 5). This clearly reveals that 10%, 8% and 22% of the model fitting will be noisy in relation to first to third birth intervals, respectively. Thus, especially in case of third birth interval, the shrinkage coefficient could easily be used to shrink predictions to yield better calibration. Table 5 also shows that the discrimination accuracy in terms of the calculated Somer's D_{xy} rank correlation related to first to third birth interval are -0.56, -0.62 and -0.68, respectively. This index provides good predictive accuracy especially in case of third birth interval. In summary, these models are good enough to describe the parity specific birth intervals in Tamil Nadu.

Table 5: Validity indices of extended cox hazard models developed for parity specific birth intervals.

Shrinkage Coefficient and D_{xy}	Index Original	Training	Test	Optimism	Index Corrected	Resample
Parity-I	1.00	1.00	0.90	0.10	0.90	200
	-0.58	-0.59	-0.57	-0.02	-0.56	200
Parity-II	1.00	1.00	0.92	0.08	0.92	200
	-0.63	-0.64	-0.62	-0.01	-0.62	200
Parity-III	1.00	1.00	0.78	0.22	0.78	200
	-0.72	-0.74	-0.70	-0.04	-0.68	200

D_{xy}: Somer's D-rank correlation.

6. Prediction from the Final Models

Prediction from the final model may be used to provide important clues to policy planners through predicted survival probabilities at considered level of a variable by holding all other variables at their average level in the model. In the present prediction analysis, the possible selected variables(s) and some combination of variables are: women's primary education; women's middle education; women's high school (and above) education; survival of index child; working women; husband's education of high school (and above); media exposure; women's high school (and above) education and survival index of child; and women's high school (and above) education and husband's high school (and above) education. On account of varying subsets of variables in the models, only results possible under each model are presented in the Tables 6 to 8 that deal with first to third birth intervals.

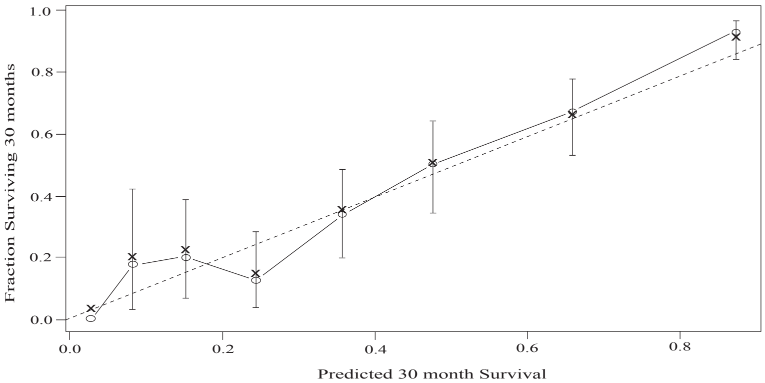


Figure 1: Bootstrap estimates of calibration accuracy for 30 months estimates from the final extended Cox model for 1st birth interval. Dots correspond to apparent predictive accuracy. X marks the bootstrap-corrected estimates.

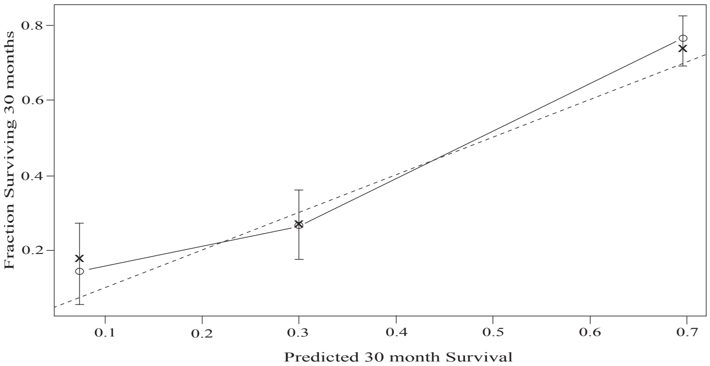


Figure 2: Bootstrap estimates of calibration accuracy for 30 months estimates from the final extended Cox model for 2nd birth interval. Dots correspond to apparent predictive accuracy. X marks the bootstrap-corrected estimates.

There is a decreasing trend in probability of not having next child over a period of time related to each birth interval. This is more evident in relation to first birth interval. There is no specific trend with increasing period of breastfeeding. However, within each category, there is an increasing trend in not having next child probability in relation to increasing education of women. Women's high school (and above) education was noticed to provide maximum benefit. This is in further evidence if women have a surviving index child.

Very few predictions were possible in relation to second and third birth intervals (Tables 7-8). High school (and above) education of husband provided the maximum benefits up to the category 22-27 months under the second birth interval. Similar results were obtained in relation to survival of index child. Surprisingly, these probabilities were lower during the period of 28 and more months. Under third birth interval, prediction was possible only in relation to media exposure.

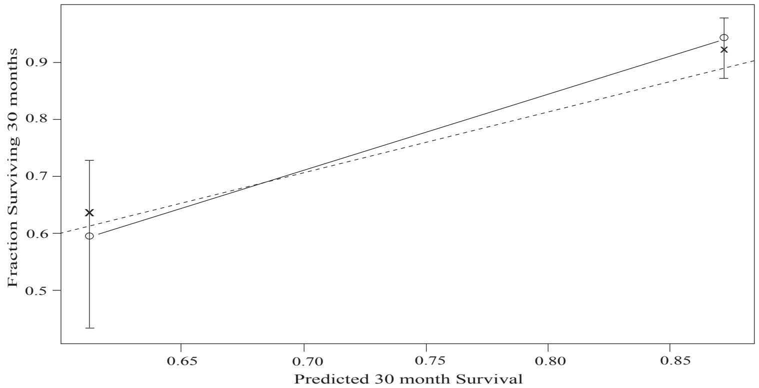


FIGURE 3: Bootstrap estimates of calibration accuracy for 30 months from the final extended Cox model for 3rd birth interval. Dots correspond to apparent predictive accuracy. X marks the bootstrap-corrected estimates.

This was not possible under the first and second birth intervals. Interestingly, media exposure showed maximum benefit during 22 and more months of breastfeeding (Table 8).

Table 6: Estimated probabilities of not having second live birth at specific months after first live birth, by selected characteristics, according to model (Ist Birth Interval).

Characteristics	Probability of not having births at months						
	12	18	24	30	36	42	48
Breastfeeding (0-15 months)							
Average	0.99	0.97	0.92	0.87	0.82	0.73	0.68
Primary educated women	0.99	0.96	0.91	0.85	0.79	0.70	0.64
Middle educated women	0.99	0.96	0.91	0.85	0.80	0.71	0.65
High school and above educated women	1.00	0.98	0.96	0.93	0.90	0.86	0.83
Index child alive	0.99	0.97	0.92	0.87	0.82	0.74	0.69
Working women	0.99	0.97	0.94	0.89	0.85	0.78	0.74
High school and above educated husband	0.99	0.97	0.93	0.88	0.84	0.77	0.72
High school and above educated women + index child alive	0.99	0.99	0.97	0.94	0.92	0.88	0.85
High school and above educated women & husband	1.00	0.98	0.96	0.94	0.91	0.86	0.83
Breastfeeding (16-21 months)							
Average	0.99	0.96	0.92	0.86	0.81	0.73	0.67
Primary educated women	0.99	0.96	0.90	0.84	0.78	0.69	0.63
Middle educated women	0.99	0.96	0.91	0.85	0.79	0.70	0.64
High school and above educated women	1.00	0.98	0.96	0.93	0.90	0.86	0.82
Index child alive	0.99	0.97	0.92	0.87	0.82	0.74	0.69
Working women	0.99	0.97	0.93	0.89	0.85	0.79	0.73
High school and above educated husband	0.99	0.97	0.93	0.88	0.83	0.76	0.71
High school and above educated women + index child alive	1.00	0.98	0.96	0.93	0.91	0.86	0.83
High school and above educated women & husband	1.00	0.98	0.96	0.94	0.92	0.87	0.85

Table 6: Continue.

Characteristics	Probability of not having births at months						
	12	18	24	30	36	42	48
Breastfeeding (22-27 months)							
Average	0.99	0.96	0.87	0.79	0.72	0.61	0.67
Primary educated women	0.98	0.94	0.85	0.76	0.68	0.56	0.48
Middle educated women	0.98	0.94	0.86	0.77	0.69	0.57	0.50
High school and above educated women	1.00	0.97	0.94	0.90	0.85	0.79	0.83
index child alive	0.99	0.95	0.88	0.80	0.73	0.62	0.55
Working women	0.99	0.96	0.90	0.83	0.77	0.67	0.61
High school and above educated husband	0.99	0.95	0.89	0.82	0.75	0.65	0.58
High school and above educated women + index child alive	0.99	0.97	0.94	0.90	0.86	0.79	0.75
High school and above educated women & husband	0.99	0.98	0.94	0.91	0.87	0.81	0.77
Breastfeeding (28-33 months)							
Average	0.99	0.96	0.92	0.86	0.81	0.73	0.67
Primary educated women	0.99	0.96	0.90	0.84	0.78	0.69	0.63
Middle educated women	0.99	0.96	0.91	0.85	0.79	0.70	0.64
High school and above educated women	1.00	0.98	0.96	0.93	0.90	0.86	0.83
Index child alive	0.99	0.97	0.92	0.87	0.82	0.73	0.68
Working women	0.99	0.97	0.93	0.89	0.85	0.78	0.73
High school and above educated husband	0.99	0.97	0.93	0.88	0.83	0.76	0.71
High school and above educated women + index child alive	1.00	0.98	0.96	0.94	0.91	0.87	0.84
High school and above educated women & husband	1.00	0.98	0.96	0.93	0.90	0.86	0.83
Breastfeeding ($\geq$ 34 months)							
Average	0.99	0.95	0.88	0.81	0.73	0.63	0.56
Primary educated women	0.98	0.94	0.86	0.78	0.70	0.58	0.51
Middle educated women	0.99	0.94	0.87	0.79	0.71	0.59	0.52
High school educated women	0.99	0.97	0.94	0.90	0.86	0.80	0.76
Index child alive	0.99	0.95	0.89	0.81	0.74	0.64	0.57
Working women	0.99	0.96	0.90	0.84	0.78	0.69	0.63
High school and above educated husband	0.99	0.96	0.90	0.73	0.77	0.67	0.61
High school and above educated women + index child alive	0.99	0.98	0.94	0.90	0.86	0.80	0.76
High school and above educated women & husband	0.99	0.98	0.94	0.90	0.86	0.80	0.76

Table 7: Estimated probabilities of not having third live birth at specific months after second live birth in TN, by selected characteristics, according to model (II^{nd} Birth Spacing).

Characteristics	Probability of not having births at months						
	12	18	24	30	36	42	48
Breastfeeding (16-21 months)							
Average	1.00	1.00	0.99	0.98	0.98	0.97	0.95
High school and above educated women	1.00	1.00	0.99	0.99	0.99	0.98	0.97
Index child alive	1.00	1.00	0.99	0.98	0.98	0.97	0.95
Breastfeeding(22-27 months)							
Average	1.00	1.00	1.00	0.99	0.99	0.99	0.98
High school and above husband	1.00	1.00	1.00	1.00	1.00	0.99	0.99
Index child alive	1.00	1.00	1.00	0.99	0.99	0.99	0.98
Breastfeeding ($\geq$ 28 months)							
Average	0.99	0.91	0.79	0.68	0.62	0.50	0.33
High school and above husband	0.99	0.95	0.87	0.80	0.76	0.68	0.53
Index child alive	0.99	0.91	0.80	0.69	0.63	0.52	0.35

Table 8: Estimated probabilities of not having fourth live birth at specific months after third live birth in TN, by selected characteristics, according to model (III^{rd} Birth Spacing).

Characteristics	Probability of not having births at months						
	12	18	24	30	36	42	48
Breastfeeding (0-21 months)							
Average	0.99	0.95	0.90	0.83	0.78	0.71	0.71
Media exposure	0.99	0.96	0.91	0.85	0.81	0.75	0.75
Breastfeeding ($\geq$ 22 months)							
Average	1.00	0.99	0.97	0.95	0.94	0.91	0.91
Media exposure	1.00	0.99	0.98	0.96	0.95	0.93	0.93

7. Discussion

To our knowledge, there is no study on birth interval, in which an exercise related to the validity of the developed Cox hazards models has been carried out. Therefore, there is no scope to compare the developed models in the present study with those reported under other studies, especially with regard to validity of the models. However, if necessary, one could examine the reported likelihood values for models under other studies to provide for a comparison with models developed in the present study.

Breastfeeding is the only covariate, which is noticed to be a significant protective factor associated with each birth interval. Education of women was significantly associated with first birth interval only while husband's education was significantly associated with first and second birth interval. Survival status of index child emerged as an important associated factor at second birth interval only while fetal loss was associated at the first birth interval. However, contraceptive use did not emerge as a significant associated factor at any birth interval.

Breastfeeding is the most important and significant factor for extending the birth interval at all the parities in TN. Trussell et al. (1985) in a study done in the Philippines, Malaysia, and Indonesia also found breastfeeding beyond 11 months to be a significant protective factor on birth interval. Anderson & Bean (1985) also support the relation between ever breastfeeding and exclusive breastfeeding and birth interval. Thus, though the relation between breastfeeding and birth interval is already documented, this study is able to predict the precise nature of this effect.

Education of woman high school (and above) was a protective factor for the first birth interval. This finding is supported by Rajaram, Rao & Pandey (1994) and Gandotra, Retherford, Pandey, Luther & Mishra (1998), who found that education of woman led to reduction in fertility, probably due to increase in awareness and choice. In contrast, Rodriguez, Hobcraft, McDonald, Menken & Trussell (1984) found little association between education and birth interval except at higher parities. Ojha (1998) and Richter, Podhisita, Chamratrithirong & Soonthorndhada (1994) support Rodriguez's findings. A similar reasoning can be attributed to the fact that media exposure had a protective impact on higher order birth interval. This finding is supported by Gandotra et al. (1998).

This study has clearly indicated that working status of women was a significant protective factor specifically for the first birth interval in TN. While this is not supported by Trussell et al. (1985), Richter et al. (1994) noticed that women employed as salesgirls and manufacturing laborer and in self-employment were significantly less likely to go for the next birth.

Ojha (1998) and Blanchard & Bogaert (1997) reported that birth intervals are comparatively longer following the birth of a male in comparison to female child. However, sex of index child did not emerge as a significant associated factor. The present study did not indicate the likelihood of it being a protective factor at any birth interval.

Survival of index child emerged as a significant protective factor for the first and second birth intervals. This was in line with many other studies like Oheneba-Sakyi & Heaton (1993); Rehman & DaVanzo (1993) Rajaram et al. (1994); Ojha (1998); and Palloni & Hantamala (1999). This shows that this factor is not a country or region specific determinant.

Other documented factors such as contraceptive use (Rajaram et al. 1994, Mahmud & Islam 1995), place of residence (Swenson & Thang 1993) and importance of previous birth interval in extending succeeding birth intervals (Rodriguez et al. 1984, DaVanzo & Starbird 1991, Miller, Trussell, Pabley & Vaughan 1992, Swenson & Thang 1993, Trussell et al. 1985) were not supported in this study.

8. Limitations

First National Health and Family Survey (NFHS) was conducted in 1992-93 in India and data was available to use in 1995. However, no study is available till now on breastfeeding as a time varying covariate with time dependent effect using bootstrap technique for validations and predictions. These techniques have been used for the first time on birth interval data. Our internet search has not revealed any similar study. Therefore, we felt that the study has valuable information for strategic and policy planners and gives more occasion for readership.

9. Conclusion

This study showed that subsets of important covariates, which entered into the final models, varied among the birth intervals within the state. However, the assessment of predictive accuracy clearly established the suitability of the parity specific developed models in describing respective birth interval. Thus, the present study emphasizes the need for regional studies in planning public health programs as per needs of the region. Further, this study also demonstrates the importance of parity specific analysis of birth interval and may assist in working out parity specific strategies in the considered region. Breastfeeding emerged as an important protective covariate that extended the birth interval irrespective of parity. Further, education of women, sex of index child, husband's education, and media exposure also demonstrated an important protective role for extending birth interval in the study.

[Recibido: agosto de 2011 — Aceptado: marzo de 2012]

References

Anderson, D. L. & Bean, L. L. (1985), 'Birth spacing and fertility limitation: A behavioral analysis of nineteenth century populationl', *Demography* **22**, 169–183.

Blanchard, R. & Bogaert, A. F. (1997), 'Additive effects of older brothers and homosexual brothers in the prediction of marriage and cohabitation', *Behavior Genetics* **27**, 45–54.

Cox, D. R. (1972), 'Regression models and life tables (with Discussion)', *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

DaVanzo, J. & Starbird, E. (1991), 'Correlates of short interbirth intervals in peninsular Malaysia: Their pathways of influence through breastfeeding and contraceptive use', *Studies in Family Planning* **22**(4).

Dickson, E., Grambsch, P., Fleming, T., Fisher, L. & Langworthy, A. (1989), 'Prognosis in primary biliary cirrohsis: Model for decision making', *Hepatology* **10**(1), 1–7.

Dwivedi, S. & Rajvir, S. (2003), 'On assessing the child spacing effect of breastfeeding using cox proportional hazards model with nfhs data', *Demography India* **32**(2), 215–224.

Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Fox, J. (2008), *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publications, Inc., London.

Gandotra, M., Retherford, R., Pandey, A., Luther, N. & Mishra, V. (1998), Fertility in India, National Family Health Survey Subject Reports 9, Mumbai: International Institute for Population Sciences; and Honolulu.

Harrell, F. E. (2001), *Regression Modeling Strategies with Application to Linear Models, Logistic Regression, and Survival Analysis*, Springer-Verlag, Berlin.

Harrell, F. E., Lee, K. L. & Mark, D. B. (1996), 'Tutorial in Biostatistics Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and meassuring and reducing errors', *Statistics in Medicine* **15**, 361–387.

International Institute for Population Sciences (1994), National Family Health Survey (MCH and Family Planning), Tamil Nadu 1992, Summary report, Population Research Centre, The Gandhigram Institute of Rural Health and Family Welfare Trust, Ambathurai R.S. (PRC, Gandhigram), and International Institute for Population Sciences (IIPS), Bombay, India.

International Institute for Population Sciences (1995), National Family Health Survey (MCH and Family Planning): India 1992-93, Summary report, International Institute for Population Sciences (IIPS), Bombay, India.

Kaplan, E. & Meier, P. (1958), ' Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**, 457–481.

Kleinbaum, D. G. (1996), *Survival Analysis, A Self Learning Text*, 1 edn, Springer-Verlag, New York.

Mahmud, M. & Islam, M. (1995), 'Adolescent contraceptive use and its determinants in Bangladesh: Evidence from Bangladesh Fertility Survey 1989.', *Contraception* **52**(3), 181–186.

Miller, J. E., Trussell, J., Pabley, A. R. & Vaughan, B. (1992), 'Birth spacing and child mortality in Bangladesh and the Philippines', *Demography* **29**(2), 305–316.

Namboodiri, K. & Suchindran, C. M. (1987), *Life Table Techniques and Their Applications Studies in Population*, Academic Press, Orlando, Florida.

Oheneba-Sakyi, Y. & Heaton, T. B. (1993), 'Effects of socio-demographic variables on birth intervals in Ghana', *Journal of Comparative Family Studies* **24**(1), 113–135.

Ojha, A. (1998), 'The effect of sex preference on fertility in selected states of India', *The Journal of Family Welfare* **44**(1), 42–48.

Palloni, A. & Hantamala, R. (1999), 'The effects of infant mortality on fertility revisited: New evidence from Latin America', *Demography* **36**(1), 41–75.

Rajaram, S., Rao, S. & Pandey, A. (1994), 'Birth interval dynamics in Goa: A parity specific analysis', *Demography India* **23**(1), 67–81.

Rehman, M. & DaVanzo, J. (1993), ' Gender preference and birth spacing in Matlab, Bangladesh', *Demography* **30**(3), 315–332.

Richter, K., Podhisita, C., Chamratrithirong, A. & Soonthorndhada, K. (1994), 'The impact of child care on fertility in urban Thailand', *Demography* **31**(4), 651–662.

Rodriguez, G., Hobcraft, J., McDonald, J., Menken, J. & Trussell, J. A. (1984), 'Comparative analysis of the determinants of birth intervals', *WFS Comparative Studies* (30).

Singh, R., Begum, S., Ahuja, R. K., Chandra, P. & Dwivedi, S. N. (2007), 'Prediction of child survival in India using developed Cox PH model: a utility for health policy programmers', *Statistics in Transition* **8**(1), 97–110.

Srinivasan, K. (1980), Birth interval analysis in fertility surveys, *in* 'World Fertility Survey Scientific Reports', number 7:19, Voorburg and London.

Swenson, I. & Thang, N. M. (1993), 'Determinants of birth intervals in Vietnam: A hazard model analysis', *Journal of Tropical pediatrics* **39**, 163–167.

Trussell, J. & Charles, H. (1983), 'A hazards model analysis of the covariates of infant and child mortality in Sri Lanka', *Demography* **20**(1), 1–24.

Trussell, J., Martin, L., Fledman, R., Palmore, J., Concepcion, M. & Abu Bakar, D. (1985), 'Determinants of birth interval length in the Phillipines, Malaysia and Indonesia: A hazard model analysis', *Demography* **22**(2).

UNFPA (1997), Reproductive Rights, Reproductive Health and Family Planning, Population issues, United Nations Fund for Population Activities (UNFPA).

Van Houwelingen, J. C. & Cessie, S. (1990), 'Predictive value of statistical models', *Statistics in Medicine* **8**, 1303–1325.

Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 309 a 330

Random Regression Models for Estimation of Covariance Functions, Genetic Parameters and Prediction of Breeding Values for Rib Eye Area in a Colombian *Bos indicus-Bos taurus* Multibreed Cattle Population

Modelos de regresión aleatoria para la estimación de funciones de covarianza, parámetros genéticos y predicción de valores genéticos en una población bovina multirracial *Bos indicus-Bos taurus* en Colombia

Carlos Alberto Martínez[1,2,a], Mauricio Elzo[2,b], Carlos Manrique[1,c], Luis Fernando Grajales[4,d], Ariel Jiménez[1,3,e]

[1] Grupo de Estudio en Mejoramiento y Modelación Animal GEMA, Departamento de Producción Animal, Universidad Nacional de Colombia, Bogotá, Colombia

[2] Department Animal Sciences, University of Florida, Florida, United States

[3] Asociación Colombiana de Criadores de Ganado Cebu ASOCEBU, Bogota, Colombia

[4] Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogota, Colombia

Abstract

In this paper we present an application of random regression models (RRM) to obtain restricted maximum likelihood estimates of covariance functions and predictions of breeding values for longitudinal records of rib eye area measured by ultrasound (REA) in a Colombian multibreed cattle population. The dataset contained 708 records from 340 calves progeny of 37 sires from nine breeds mated to Gray Brahman Cows. The mixed model was a RRM that used Legendre polynomials (LP) of order 1 to 3. Fixed effects were age of animal, dam parity, contemporary group (herd*year*season*sex), breed additive genetic and heterosis, whereas direct and maternal additive genetic and maternal permanent environment were random effects. Residual

[a] MSc in Quantitative Genetics. E-mail: camartinezn@unal.edu.co
[b] Professor. E-mail: maelzo@ufl.edu
[c] Professor. E-mail: cmanriquep@unal.edu.co
[d] Assistant professor. E-mail: lfgrajalesh@unal.edu.co
[e] MSc in Quantitative Genetics. E-mail: jimenezariel@hotmail.com

variances were modeled either as constant or changing across the growth trajectory. Models were compared with two Information Criteria, the corrected Akaike's and the Schwartz's Bayesian. According to these criteria the best model was the one with first order LP and constant residual variance. Given that with this model estimated maternal additive genetic and permanent environment covariance functions showed that these effects were not accurately disentangled, a parsimonious model without maternal additive genetic effects was used to obtain genetic parameters and breeding values. Direct additive genetic variance decreased until 150 days and then increased. Maternal permanent environment variance increased with age. Direct heritability estimates for REA at 4 months, weaning, 12 and 15 months (considered as target ages), were 0.003, 0.007, 0.034 and 0.058, respectively. Direct additive correlations ranged from −0.7 to 1. Maternal permanent environmental correlations were close to unity across the entire range of ages. Estimates of (co)variance components showed the need to validate results with larger multigenerational multibreed populations before implement RRM in regional or national genetic evaluation procedures in Colombia.

***Key words*:** Animal population, Covariance functions, Mixed model.

Resumen

En este trabajo presentamos una aplicación de modelos de regresión aleatoria (RRM) para obtener estimadores de máxima verosimilitud restringida de funciones de covarianza y predicciones del valor genético para datos longitudinales de área de ojo del lomo medidos por ultrasonido (REA) en una población bovina multirracial en Colombia. El conjunto de datos contenía 708 registros de 340 animales descendientes de 37 toros de 9 razas apareados con hembras Brahman Gris. Los modelos mixtos empleados fueron RRM que usaron polinomios de Legendre (LP) de orden 1 a 3. Los efectos fijos fueron edad del animal, número de partos de la madre, grupo contemporáneo (hacienda*año*época*sexo), efectos genéticos aditivos de raza y heterosis, mientras que los efectos genéticos aditivos directos y maternos y de ambiente permanente materno fueron aleatorios. Las varianzas residuales se modelaron como constantes o cambiantes a través de la trayectoria de crecimiento. Los modelos fueron comparados mediante el criterio de información de Akaike corregido y el de información bayesiana de Schwartz. Según esos criterios, el mejor modelo fue aquel con LP de orden 1 y varianza residual constante. Dado que con este modelo las estimaciones de las funciones de covarianza genética aditiva materna y de ambiente permanente materno indicaron que estos dos efectos no se separaron adecuadamente, un modelo más parsimonioso sin los efectos genéticos aditivos maternos fue empleado para obtener parámetros y valores genéticos. La varianza genética aditiva directa decreció hasta 150 días y luego aumentó. La varianza de ambiente permanente materno aumentó con la edad. Las estimaciones de heredabilidad directa para REA a los 4 meses, destete, 12 y 15 meses (consideradas como edades de referencia) fueron 0.003, 0.007, 0.034 y 0.058, respectivamente. Las correlaciones aditivas directas variaron de −0.7 a 1. Las correlaciones de ambiente permanente materno fueron cercanas a la unidad a través de todo el rango de edades. Las estimaciones de componentes de (co)varianza mostraron la necesidad de validar los resultados con poblaciones multirraciales multigeneracionales mayores antes de implementar RRM en procedimientos de evaluación genética regionales o nacionales en Colombia.

***Palabras clave*:** modelo mixto, funciones de covarianza, población animal.

1. Introduction

Modeling of longitudinal records with Legendre polynomials (LP) was proposed by Kirkpatrick, Lofsvold & Bulmer (1990) to describe direct additive genetic covariances among records at any pair of ages in a continuous form. The LP are solutions to the Legendre's differential equation and they are orthogonal. This property allows describing patterns of genetic variation through a growth trajectory. Continuous functions representing covariances among records are called covariance functions (Kirkpatrick et al. 1990). Meyer (1998) suggested that coefficients of covariance functions could be estimated as covariances among random regression coefficients by fitting linear mixed models. Advantages of random regression over multiple trait models (MTM) involve the inclusion of all available data without pre-adjustment to particular ages, no lose of records taken outside certain age ranges, and reduction in the number of parameters to be estimated by fitting parsimonious models (Kirkpatrick et al. 1990, Meyer & Hill 1997). Until today, these models have not been implemented for genetic analysis in Colombia. Carcass quality is important in the current beef market. Thus, there exists great interest in carcass traits measured by ultrasound like the rib eye area (REA), because they are closely related to the true carcass values and meat yields (Hougton & Turlington 1992). Genetic evaluation of carcass traits has been implemented in animal breeding programs in different countries and species (Wilson 1992, Hassen, Wilson & Rouse 2003, Fischer, van der Werf, Banks, Ball & Gilmour 2006, Choy, Lee, Kim, Choi, Choi & Hwang 2008). However, few genetic studies have considered ultrasound carcass traits in a longitudinal manner either in purebred or crossbred cattle (Fischer et al. 2006, Speidel, Enns, Brigham & Keeman 2007, Mercadante, El Faro, Pinheiro, Cyrillo, Bonilha & Branco 2010). Jiménez, Manrique & Martínez (2010) conducted the only study in Colombia on ultrasound carcass traits in cattle under pasture conditions using purebred Brahman. In low tropical areas of Colombia, there are limiting environmental conditions for livestock production. Consequently, crossbreeding between native Creole or European (*Bos taurus*) with Zebu (*Bos indicus*) breeds is frequently used as a strategy to increase beef production while maintaining adaptability (FEDEGAN 2006). This mating strategy has created a need to establish genetic evaluation programs involving animals from temperate and tropically adapted breeds for carcass traits. These programs must take into consideration that 72% of the Colombia's cattle population is Zebu (mainly Brahman) (FEDEGAN 2006). Thus, the objective of this research was to show how to apply the RRM to obtain restricted maximum likelihood estimates of covariance functions and predictions of breeding values for longitudinal records of rib eye area measured by ultrasound (REA) in a Colombian multibreed cattle population.

2. Materials and Methods

All of the practices involving manipulation of animals that were performed to obtain records in this research were approved by the Animal Bio-ethics Committee of the National University of Colombia (Approval letter number: CBE-FMVZ-012, July, 2010).

2.1. Breeds, Matings and Animal's Management

To construct the multibreed population, 37 bulls from 9 breeds were mated to third-parity Gray Brahman (GB) cows and heifers. Sire breeds were Gray Brahman (GB; $n = 12$), Red Brahman (RB; $n = 4$), Guzerat (GUZ; $n = 3$), Romosinuano (ROM; $n = 3$), Blanco Orejinegro (BON; $n = 3$), Simmental (SIM; $n = 3$), Braunvieh (BVH; $n = 3$), Normand (NOR; $n = 3$) and Limousin (LIM; $n = 3$). These *Bos taurus* breeds (Creole and temperate) were chosen because they are frequently used for crossbreeding programs with zebu cattle in Colombia's low tropical beef production systems. Brahman was included because it has the largest cattle population in the country (Jiménez et al. 2010), and GUZ is a *Bos indicus* breed with increasingly higher representation in Colombia that has not been studied as a single breed or in crosses with Brahman. Females were chosen on the basis of a normal reproductive cycle and a healthy reproductive system. Subsequently, cows and heifers were randomly allocated to males, and artificially inseminated using a fixed-time protocol. Firstly, females received a progesterone implant (CIDR, Pfizer, NY, USA) and 2 mg of estradiol benzoate. Eight days later, the CIDR implants were removed, and 1 cm^3 of F2 α prostaglandin (Estrumate, Schering Plough S.A., Kenilworth, NJ, USA) was applied, followed by an injection of 1 mg of estradiol benzoate 24 hours later. Females were artificially inseminated 54 hours after progesterone implant removal. Calves were born in 2008 and 2009. Table 1 shows the number of sires per breed and the number of calves per breed group by year and total.

Table 1: Number of sires per breed and number of calves per breed group by year of birth.

Sire breed	Number of sires	Calf breed group	Number of calves		
			2008	2009	Total
BON	3	BON X GB	21	12	33
BVH	3	BVH X GB	13	8	21
GB	12	BG X GB	63	34	97
GUZ	3	GUZ X GB	18	9	27
LIM	3	LIM X GB	20	13	33
NOR	3	NOR X GB	22	14	36
RB	4	BR X GB	26	8	34
ROM	3	ROM X GB	18	10	28
SIM	3	SIM X GB	21	10	31
Total	37		222	118	340

BON = Blanco Orejinegro; BVH = Braunvieh; GB = Gray Brahman;
GUZ = Guzerat; LIM = Limousin; NOR = Normand;
RB = Red Brahman; ROM = Romosinuano; SIM = Simmental.

Animals were kept in two herds located in Southern Cesar, municipality of Aguachica, Colombia. The ecosystem in this micro region is a very dry tropical forest. This region has a mean annual temperature of 28 °C, a height above sea level of 50 m, a relative humidity of 80% and sandy-loam soils. Because of its environmental conditions, Southern Cesar is considered to be better suited for beef cattle production than other regions in Colombia. The feeding system was based on pastures. Grass species were Brachipará (*Brachiaria plantaginea*), Guinea (*Panicum máximum*) and Angleton (*Dichantium aristatum*). Pastures were not fertilized. Animals were provided with an 8% phosphorus mineral supplement (GANASAL$^{®}$, Colombia). Mineral supplement consumption was *ad libitum.* The grazing system was rotational with a rotation period of 60 days. All calves were weaned between 7 and 8 months of age and males were castrated at 12 months of age.

2.2. Records

The REA records were taken by a certified technician of the Colombian Zebu Cattle Breeders Association (ASOCEBU, Bogotá D.C., Colombia) using an Aquila Esaote model device (Pie Medical Equipment B.V., Maastricht, Limburg, The Netherlands). Once ultrasound images were collected, they were analyzed to check quality and to obtain the REA values (cm^2) using the Echo Image Viewer software of Pie Medical (Pie Medical Equipment B.V., Maastricht, Limburg, The Netherlands). The total number of REA records was 708. Age of animals ranged from 70 to 492 days. Records were intended to be taken approximately at four, eight (weaning), twelve and fifteen months. Mean ages at each of these data collection points were: 120, 233, 332 and 445 days. At 4 months of age, calves are more dependent on the cow's milk production that at weaning. This is due to the fact that at this stage the calf has not finished its transition from pre-ruminant to ruminant (Van Soest 1994). Thus, REA measurements taken at this age are useful to evaluate maternal effects (both genetic and non genetic).

2.3. Genetic Analysis

Mixed models procedures were carried out to obtain restricted maximum likelihood (REML) estimates of covariance components and best linear unbiased predictors (BLUP) of animal breeding values (BV). The following effects were assumed to be fixed in the mixed model: Contemporary group (herd*year*season*sex subclass), breed group additive effects, non additive effects (individual heterosis), dam parity (heifer or third parity cow) and age of the animal (linear and quadratic effects). In a first approach, the random effects were: Direct additive genetic, maternal additive genetic, maternal permanent environment, and residual. Seasons within years were defined as rainy or dry. The first season was a rainy season from mid April to mid August of 2009, the second was a dry season from mid August to mid December of 2009, the third was a dry season from mid December of 2009 to mid April of 2010, and the fourth was a rainy season from mid April to mid August of 2010. The GB and RB bulls were grouped as a single breed (BR). Thus, there were 8 breed groups for calves: BR x GB, BON X GB, BVH X GB, GUZ X GB, LIM X GB, NOR X GB, ROM X GB and SIM X GB. Breed group effects were modeled as a continuous function of breeds over time. This function was a linear LP. Additive genetic breed group effects were modeled in such a way because individual random deviations and breed group solutions are required to obtain BV at a particular age in a multibreed population (Elzo & Wakeman 1998). In addition, because of the orthogonality of LP, the block of the mixed model equations corresponding to breed group effects was an identity matrix, thus, multicollinearity and confounding problems that are commonly present among genetic fixed effects in multibreed populations (Elzo & Famula 1985) could be alleviated at least partially. To estimate covariance functions (CF) for the following effects: Direct additive genetic (DAGCF), maternal additive genetic (MAGCF) and maternal permanent environment (MPECF) and to compute BV, the regression variables used were normalized LP (LP with norm 1), evaluated at age of animal when records were collected. Orders of LP ranged from 1 to 3. The following combinations of LP to describe direct additive, maternal additive and maternal permanent environment CF were used: one (LP1), 2(LP2) and 3(LP3) for the 3 covariance components, and 3 for direct additive genetic covariances and 2 for maternal additive genetic and permanent environment covariances (LP32). The orders of LP were defined taking into account data set size and literature reports (Fischer et al. 2006, Mercadante et al. 2010). The residual variance was modeled in two ways. The first one assumed that the residual variance was the same along the entire growth trajectory (LP1HOM, LP2HOM, LP3HOM, LP32HOM), and the second one assumed a step function (LP1HET, LP2HET, LP3HET, LP32HET) across 3 age intervals ($70 \leq age \leq 230$ days, $230 < age \leq 365$ days, and $365 < age \leq 492$ days). Residuals were assumed to be independent and normally distributed. Thus, there were a total of 8 random regression models to compare: LP1HET, LP2HET, LP3HET, LP32HET, LP1HOM, LP2HOM, LP3HOM, and LP32HOM. Models comparison was made through the Schwartz's Bayesian Information Criterion (BIC) and the Corrected Akaike's Information Criterion (AICC):

$$BIC = -2\log L + K\log(N-r)$$
$$AICC = AIC + \frac{(2(K+1)(K+2))}{(N-K-2)}$$

Where AIC is the Akaike's information criterion, K is the number of parameters, N is the number of records, logL is the natural logarithm of the likelihood function and r is the rank of the fixed part of the model, that is, the rank of the incidence matrix for all fixed effects in the model. The AICC was preferred over the AIC in our study because of the small data set size, which is suggested by Littell, Milliken, Stroup, Wolfinger & Schabenberger (2006). However, estimated covariance functions showed a strong negative correlation among maternal additive genetic and maternal environmental effects, which indicated that these effects were not accurately separated. Thus, a parsimonious version of the model selected in the first approach (LP1HOM) considering only maternal permanent environmental effects and denoted as LP1HOMS was used to compute variance-covariance components, genetic parameters and BV. The number of variance-covariance parameters ranged from 7 for the most parsimonious model (LP1HOMS) to 33 for model LP4HET (Table 2). In matrix notation the RRM used was as follows:

$$y = X\beta + Q_{ga}g_a + Q_h h + \Phi_a a + \Phi_p p + e$$

$$Var\begin{bmatrix} a \\ p \\ e \end{bmatrix} = \begin{bmatrix} A \otimes K_a & & \\ & I \otimes K_p & \\ & & R \end{bmatrix}$$

$$E[y] = X\beta + Q_{ga}g_a + Q_h h$$

$$Var(y) = \Phi_a(A \otimes K_a)\Phi'_a + \Phi_p(I \otimes K_p)\Phi'_p + R$$

Table 2: Akaike's corrected information criterion (AICC), Schwartz's Bayesian information criterion (BIC), residual analysis and number of parameters for each model.

Model	AICC	BIC	Number of variance covariance parameters	Log L[1]
LP1HET	3394.74	3448.66	12	−1685.15
LP2HET	3412.64	3506.42	21	−1684.65
LP3HET	3435.54	3581.68	33	−1683.10
LP32HET	3426.62	3555.42	29	−1683.03
LP1HOM	3392.82	3437.8	10	−1686.25
LP2HOM	3410.66	3495.62	19	−1685.78
LP3HOM	3508.34	3645.82	31	−1721.70
LP32HOM	3426.66	3546.72	27	−1685.22
LP1HOMS	3386.66	3418.21	7	−1686.25

[1] Natural logarithm of the restricted likelihood function.

Where y = vector containing the REA records, β = vector of unknown fixed effects of contemporary group, dam parity and age of animal, g_a = vector of fixed additive genetic group effects (modeled as a continuous function of time) which correspond to the mean effects of genes from a given breed (Elzo 2010), h = vector of fixed non additive genetic effects (individual heterosis) these are the effects due to the presence of alleles from different breeds in one locus (Elzo 2010), a = vector of random regression coefficients for direct additive genetic effects, which are the sum of effects of individual genes affecting REA (Kempthorne 1957, Lynch & Walsh 1998), p = vector containing random regression coefficients for maternal permanent environmental effects, which correspond to those effects explained by the environment proportioned to the calf by its dam, maternal effects are genetic to the dam and environmental to the calf, e = random vector of residuals, X, Q_{ga}, Q_h, Φ_a, Φ_p were known incidence matrices respectively relating vectors β, g_a, h, a, p to REA records and super index "′" denotes transposition. Columns in X relating records to fixed effects of age contained second order LP evaluated at each age; columns for the other fixed effects contained zeroes and ones. Matrix Q_{ga} contained linear LP evaluated at the expected fraction of each breed in an animal times the age of the animal, and matrix Q_h contained probabilities of alleles of different breeds occurring at one locus in an animal (Elzo & Famula 1985) and it was calculated as: $H_I = 1 - \sum_{i=1}^{b}(Rp \times Rm)_i$, where R_p and R_m are the expected fractions of each breed in sire and dam of the animal and b is the number of breeds, matrices Φ_a, Φ_m and Φ_p contained LP evaluated at the ages of the animals when records were taken (Meyer 1998); matrices K_a and K_p contained the coefficients for additive genetic, and maternal permanent environmental covariance functions, A was the additive relationship matrix, $\otimes$ represents the Kronecker product, and R was the residual covariance matrix which had the form $R = I\sigma_e^2$. The mixed models analyses were performed with software WOMBAT (Meyer 2007) using an average information (AI) algorithm. Different starting values were used to ensure that estimates corresponded to global maximums. Convergence was declared when change of value of the natural logarithm of the restricted likelihood function in two consecutive iterations was lower than 5×10^{-4}. Model effects were estimated by solving the mixed model equations:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Q_{ga} & X'R^{-1}Q_n & X'R^{-1}\Phi_a & X'R^{-1}\Phi_p \\ & Q'_{ga}R^{-1}Q_{ga} & Q'_{ga}R^{-1}Q_n & Q'_{ga}R^{-1}\Phi_a & Q'_{ga}R^{-1}\Phi_p \\ & & Q'_nR^{-1}Q_n & Q'_nR^{-1}\Phi_a & Q'_nR^{-1}\Phi_p \\ & & & \Phi'_aR^{-1}\Phi_a + A^{-1} \otimes K_a^{-1} & \Phi'_aR^{-1}\Phi_p \\ & Symmetric & & & \Phi'_pR^{-1}\Phi_p + I \otimes K_p^{-1} \end{bmatrix}$$

$$\begin{bmatrix} \beta \\ g_a \\ h \\ a \\ p \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Q'_{ga}R^{-1}y \\ Q'_nR^{-1}y \\ \Phi'_aR^{-1}y \\ \Phi'_pR^{-1}y \end{bmatrix}$$

The eigenfunctions (EF) of a CF are continuous smooth functions representing a possible deformation in the mean growth trajectory (Kirkpatrick et al. 1990). Thus, the EF were calculated to study variation patterns throughout the REA growth curve. Each EF has a correspondent eigenvalue. Only EF whose eigenvalues together explained at least 80% of the respective variance component were computed. The EF were computed for direct additive genetic CF from eigenvectors of K_a matrix as:

$$\psi_i(t) =< c_{\psi_i}, \phi_{t^*} >$$

where c_{ψ_i} is the i^{th} eigenvector of the matrix K_a and ϕ_{t^*} is a vector with LP evaluated at t^* (age t standardized to the real interval $[-1, 1]$) and the operator $< \cdot, \cdot >$ represents the internal or dot product between vectors. The age t was standardized to the real interval $[-1, 1]$ by using the following expression (Kirkpatrick et al. 1990):

$$t^* = \frac{2(t - t_{\min})}{t_{\max} - t_{\min}} - 1$$

where $t_{\min}$ and $t_{\max}$ are the minimum and maximum ages at which records were taken. Matrices of covariance components for additive direct genetic effects and maternal permanent environmental effects as well as BV for REA at 4 target ages were obtained using the REML estimates of covariance matrices among random regression coefficients obtained at convergence which are equal to the coefficient matrices of corresponding CF (Meyer 1998). Target ages were 120, 230, 365 and 450 days, and the corresponding REA values were denoted as REA4, REAW, REAY and REAF.

Covariance matrices for REA at target ages were computed using the CF which were obtained as the product of a matrix containing LP evaluated at those ages (Φ), the correspondent coefficients matrix (K_a for direct additive covariance, and K_p for maternal permanent environmental covariance) and the transpose of matrix Φ (Kirkpatrick et al. 1990, Meyer 1998):

$$cov_j = \Phi K_j \Phi'$$

where, Cov_j is the covariance matrix for the j^{th} covariance component (additive genetic or maternal permanent environment). The matrix Φ was obtained as the product of two matrices. The first is matrix $M = (m_{ij})_{dxk} = t_i^{*j-1}$, where t_i^* is the i^{th} age standardized to the real interval $[-1, 1]$, d is the number of ages considered (4 in this case) and $k - 1$ is the order of the LP. The second matrix was $\Lambda_{k\times k}$, which contained the coefficients of the LP. Thus, $\Phi = M\Lambda$ (Kirkpatrick et al. 1990). Consequently,

$$Cov_j = \Phi K_j \Phi' = M\Lambda K_j \Lambda' M' = MC_j M', \quad \text{where} \quad C_j = \Lambda K_j \Lambda'$$

By using matrix C_j instead of matrix K_j for representing the j^{th} CF, cov_j is calculated directly as a function of the age standardized to the interval $[-1, 1]$ (i.e., t^*). This equivalent form was used to compute critical points of CF. The extremes of the CF were also assessed in order to detect the global maximum and minimum values of each CF.

The BV were computed for REA4, REAW, REAY and REAF for all individuals in the population (sires, dams, and offspring). The additive breeding value for animal i at age t (BV_{it}) was computed by adding two terms. The first term was a weighted sum of probabilities of alleles of breed b in animal i times the generalized least squares estimate of breed b (deviated from BR) at time t, $b = 1, 2, \dots, 7$. The second term was the BLUP of the random solution for each individual. This value was computed as the internal (or dot) product between a vector containing LP evaluated at age t and a vector whose entries were the BLUP for random regression coefficients of animal i. Thus, BV_{it} was computed as:

$$BV_{it} =< \phi_{bt}, \hat{g}_a > + < \phi_t, \hat{a}_i >$$

where ϕ_{bt} is a vector of LP evaluated at the product of the fraction of breed b ($b = 1, 2, \dots, 7$) in animal i times calf age t standardized to real interval $[-1, 1]$, $\hat{g}_a$ is the generalized least squares solution of the fixed coefficient for breed additive genetic effects, ϕ_t is a vector of LP evaluated at calf age t standardized at real interval $[-1, 1]$, and $\hat{a}_i$ is the BLUP vector of the random coefficients for animal i.

3. Results

3.1. Model Selection

As stated before, estimated covariance functions, covariance components, genetic parameters and breeding values were computed using model LP1HOMS. Although this model was selected given the evidence of correlation among maternal additive genetic and environmental effects, according to AICC and BIC values, this was the best model since it had the smallest AICC and BIC values (Table 2).

3.2. REML Estimates of Covariance Functions and Covariance Components

Direct additive genetic (DAGC) and maternal permanent environment (MPEC) covariances between pairs of ages t_1 and t_2 such that $70 \leq t_1, t_2 \leq 492$, were described by the following CF (DAGCF, and MPECF, respectively) obtained with model LP1HOMS using $cov_j = \Phi K_j \Phi'$:

$$DAGC(t_1,t_2) = \begin{bmatrix} \phi_0(t_1^*) & \phi_1(t_1^*) \end{bmatrix} \begin{bmatrix} 1.5900 & 1.2435 \\ 1.2435 & 1.1589 \end{bmatrix} \begin{bmatrix} \phi_0(t_2^*) \\ \phi_1(t_2^*) \end{bmatrix}$$

$$MPEC(t_1,t_2) = \begin{bmatrix} \phi_0(t_1^*) & \phi_1(t_1^*) \end{bmatrix} \begin{bmatrix} 53.482 & 4.5003 \\ 4.5003 & 0.3787 \end{bmatrix} \begin{bmatrix} \phi_0(t_2^*) \\ \phi_1(t_2^*) \end{bmatrix}$$

where t_i^* is the i^{th} age standardized in the real interval $[-1,1]$, and $\phi_j(t_i^*)$, $j = 0,1$, is the j^{th} LP evaluated at i^{th} age. The equivalent forms of these 2 CF, using $cov_j = MC_jM'$, were as follows:

$$DAGC(t_1,t_2) = \begin{bmatrix} 1 & t_1^* \end{bmatrix} \begin{bmatrix} 0.7950 & 1.0769 \\ 1.0769 & 1.7382 \end{bmatrix} \begin{bmatrix} 1 \\ t_2^* \end{bmatrix}$$

$$MPEC(t_1,t_2) = \begin{bmatrix} 1 & t_1^* \end{bmatrix} \begin{bmatrix} 26.7405 & 3.8972 \\ 3.8972 & 0.5680 \end{bmatrix} \begin{bmatrix} 1 \\ t_2^* \end{bmatrix}$$

These functions are defined (domain) for the following set: $D = [70, 492] \times [70, 492]$. The partial derivatives were:

$$\frac{\partial CF_j}{\partial t_1^*} = c_{12} + c_{22}t_2^*; \quad \frac{\partial CF_j}{\partial t_2^*} = c_{12} + c_{22}t_1^*$$

where c_{ij} is the $(i-j)^{th}$ entry of the matrix C and CF_j is the j^{th} CF ($j =$ DAGCF or MPECF), and t_i^* are standardized calf ages at $[-1,1]$. By equating these expressions to zero yielded that the critical arguments of the CF were $\frac{-c_{12}}{c_{22}}$ for both t_1^* and t_2^* (because the 2 CF were symmetric).

To determine if the critical points obtained from the last expression were maximums, minimums or saddle points the determinant of the Hessian matrix was computed. Because these functions are polynomials, the Clairaut's theorem (Stewart 2008) applies making the Hessian matrix to be symmetric. This matrix was:

$$H = \begin{bmatrix} \frac{\partial^2 CF_j}{\partial t_1^{*2}} & \frac{\partial^2 CF_j}{\partial t_1^* \partial t_2^*} \\ \frac{\partial^2 CF_j}{\partial t_2^* \partial t_1^*} & \frac{\partial^2 CF_j}{\partial t_2^{*2}} \end{bmatrix} = \begin{bmatrix} 0 & c_{22} \\ c_{22} & 0 \end{bmatrix}$$

Thus: $|H| = -(c_{22}^2)$, and the critical point is a saddle point. Variance functions (VF) are special cases of CF when $t_1^* = t_2^*$. Because there is a single age, VF are univariate. Critical points computed for CF and VF could be outside the range of calf ages (i.e., outside their domain). If this happens, these critical points should be ignored because in regression analysis values outside the domain (range of calf ages) would have no valid interpretation (Draper & Smith 1981).

The DAGCF had a saddle point located at 150 days. Thus, covariances before 150 days tended to decrease with age. After 150 days, the pattern was more complex. Covariances among ages lower than 150 days and ages greater than 150 days tended to decrease as distance among them increased (Figure 1). On the other hand, covariances among ages greater than 150 days tended to increase as the animals grew older. The MPECF was positive throughout the entire domain (Figure 1). The minimum value of MPECF (19.51 cm^4) was located at coordinates (in days): (70, 70) while the maximum (35.10 cm^4) was located at (492, 492). The analysis of derivatives showed that MPECF had a critical point outside the range of calf ages in this study. As indicated before, VF are special cases of CF, because by definition they are the covariance of a random variable with itself. Consequently, the diagonals of the CF correspond to VF. According to the analysis of first and second derivatives of the direct additive genetic variance function, direct additive genetic variance (DAGV) had a global minimum located at 150 days (0.13 cm^4). The largest value of DAGV was 4.69 cm^4 at 492 days. As shown in Table 3 for the target ages, DAV was 0.16 for REA4, 0.38 for REAW, 1.93 for REAY and 3.64 cm^4 for REAF. The DAGC were negative for REA4-REAY (-0.13 cm^4) and REA4-REAF (-0.23 cm^4) and the biggest covariance value was among REAY and REAF (2.64 cm^4). For the target ages, maternal permanent environment variance (MPEV) ranged from 21.12 (REA4) to 33.35 cm^4 (REAF). MPEC had its lowest value (22.93 cm^4) among REA4 and REAW and the largest (31.60 cm^4) for REAY-REAF (Table 3). Considering the entire range of ages, MPEV had its maximum value at 492 days (35.10 cm^4) and the minimum (19.51 cm^4) at 70 days.

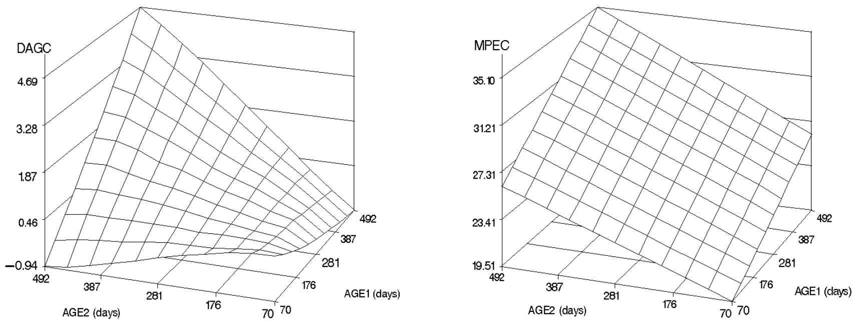


Figure 1: Plots of direct additive genetic (DAGC (cm^4); left), and maternal permanent environment (MPEC (cm^4); right) covariances.

REML estimate of residual variance was 25.55 cm^4. Because phenotypic variance (PhV) is the sum of genetic and environmental variance components, it also increased as animals grew older. Its minimum value was 45.45 cm^4 at 70 days and its maximum was 65.35 cm^4 at 492 days. Plots of DAGV and MPEV are shown in Figure 2.

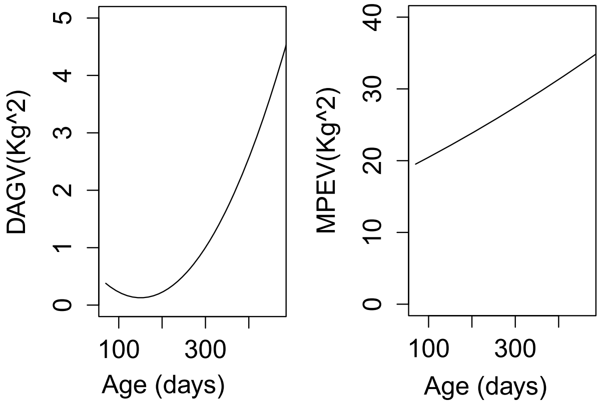


Figure 2: Direct additive (DAGV; left), and maternal permanent environmental (MPEV; right) variances.

Table 3: Estimates of covariance components, (cm^4), genetic parameters, and variance ratios for five target ages.

Pair of traits	DAGC	MPEC	DAGR/Dh[1]	MPER/MPr[2]	PhR
REA4,REA4	0.164	21.124	0.003	0.451	1
REA4,REAW	0.033	22.930	0.133	1	0.471
REA4,REAY	−0.126	25.146	−0.224	1	0.483
REA4,REAF	−0.227	26.542	−0.294	1	0.486
REAW,REAW	0.376	24.890	0.007	0.490	1
REAW,REAY	0.796	27.296	0.935	1	0.520
REAW,REAF	1.061	28.811	0.908	1	0.530
REAY,REAY	1.928	29.934	0.034	0.521	1
REAY,REAF	2.641	31.595	0.998	1	0.571
REAF,REAF	3.635	33.349	0.058	0.533	1

REA4 = rib eye area at 4 months; REAW = rib eye area at weaning (230 days);
REAY = rib eye area at year; REAF = rib eye area at 15 months;
DAGC = direct additive genetic covariance;
MPEC = maternal permanent environmental covariance;
DAGR = direct additive genetic correlation; Dh = direct heritability;
MPER = maternal permanent environmental correlation;
MPr = ratio of maternal permanent environmental variance to phenotypic variance;
PhR = phenotypic correlation.

[1] When both ages are the same, the value is heritability; when ages are different is a correlation.

[2] When both ages are the same, the value is the corresponding variances ratio; when ages are different is a correlation.

3.3. Heritability and Ratio of MPEV to PhV

The direct heritability (the ratio of DAGV to PhV) estimates (Dh), were low at the entire trajectory. The Dh reached a global minimum at 150 days (0.003) and its maximum at 492 days (0.072). The estimate of Dh at 70 days was 0.008. The Dh estimates at the 4 target age points were 0.003 (REA4), 0.007 (REAW), 0.034 (REAY) and 0.058 (REAF) (Table 3). The trend of Dh across the range of calf ages is shown in Figure 3. The ratio of MPEV to phenotypic variance (MPr) ranged from 0.43 at 70 days to 0.54 at 492 days. The MPr had an upward trend trough the REA trajectory (Figure 3). The MPr estimates for the target ages were 0.45 for REA4, 0.49 for REAW, 0.52 for REAY and 0.53 for REAF (Table 3).

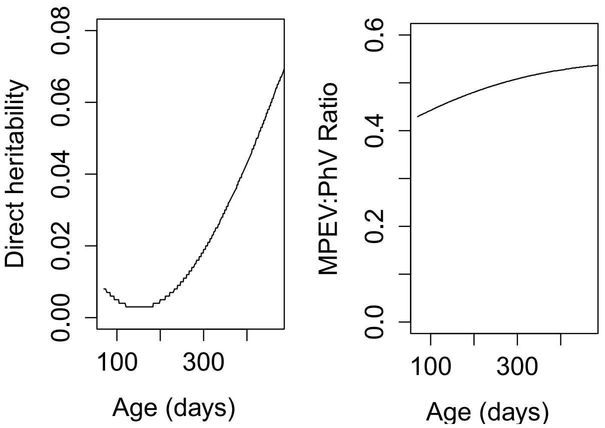


FIGURE 3: Graphics of continuous functions describing direct heritability (left), and ratio of maternal permanent environmental variance (MPEV) to phenotypic variance (PhV) (right).

3.4. Correlations

The estimates of direct additive genetic (DAGR), maternal permanent environment (MPER) and Phenotypic (PhR) correlations at the 4 target ages are shown in Table 3. Estimates of DAGR formed a plateau close to unity approximately after 240 days. The DAGR between REA at 70 days and REA at other ages were negative after 193 days and had its lowest value at 492 days (−0.71). For target ages, DAGR estimates ranged from -0.29 among REA4 and REAF to 0.99 among REAY and REAF (Table 3). The MPER estimates were close to unity throughout the entire range of ages considered. The PhR estimates were always positive and ranged from moderate to high. For the 4 selected age points, PhR values ranged from 0.47 (REA4-REAW) to 0.57 (REAY-REAF).

3.5. Eigenfunctions

The first eigenvalue for DAGCF was 2.64 and it accounted for 95.9% of total DAGV. Thus, for DAGCF only the first EF (DAGEF1) was computed. The first eigenvector of the coefficient matrix associated with DAGCF was $\begin{pmatrix}0.7651 & 0.6439\end{pmatrix}'$, and the DAGEF1 was:

$$DAGEF1 = 0.5358 + 0.7991t^*$$

Figure 4 shows a graph of this function across the entire range of calf ages. The DAGEF1 was an increasing function, but it was not positive at the entire range trajectory. The point where this function crossed the age axis was 136 days. The behavior of the EF was a consequence of the estimates obtained here for DAGR. As described previously, there were negative DAGR between early and late calf ages.

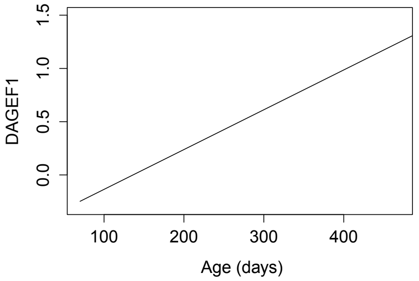


Figure 4: First eigenfunction of the direct additive genetic (DAGEF1) covariance function.

3.6. Breeding values

Descriptive statistics for BLUP of BV in general and discriminated by sire breed are shown in Table 4. Values for sire breeds were obtained using information from the bulls and the overall values were obtained from BV of all animals. Overall mean BV were 0.41 for REA4, 0.72 for REAW, 1.26 for REAY and 1.55 cm^2 for REAF. Values presented in Table 4 indicate that on average LIM bulls had the highest BV for REA. Sires of BVH and NOR breeds had the smallest BV at the 4 target age points. In the Creole cattle group, ROM sires had greater mean BV than BON sires. Finally, for the Bos indicus breeds, GUZ bulls had the greatest mean BV.

4. Discussion

4.1. Model Selection

Selection of the most parsimonious model (LP1HOMS) as the best model by BIC and AICC implies that the larger log likelihood values obtained with other models was insufficient to counterbalance BIC and AICC penalties due to the higher number of required parameters. Consequently, the BIC and AICC values of those other models were larger than the values for model LP1HOMS (Table 2).

Table 4: Descriptive statistics for breeding values at the selected age points according to the breed of sire and in general.

Sire breed[1]	Statistic[2]	REA4	REAW	REAY	REAF
BON	Min	0.17	0.35	0.53	0.63
	Max	0.27	0.56	1.14	1.47
	Mean	0.22	0.42	0.76	0.95
BR	Min	−0.07	−0.17	−0.40	−0.54
	Max	0.02	0.22	0.53	0.73
	Mean	−0.02	0.05	0.13	0.19
BVH	Min	−0.34	−0.83	−1.61	−2.06
	Max	−0.30	−0.57	−0.99	−1.21
	Mean	−0.32	−0.69	−1.26	−1.59
GUZ	Min	1.22	2.18	3.73	4.58
	Max	1.29	2.37	4.24	5.29
	Mean	1.25	2.29	4.03	4.99
LIM	Min	4.49	7.32	12.36	15.06
	Max	4.56	7.57	12.97	15.91
	Mean	4.52	7.48	12.76	15.62
NOR	Min	−0.09	−0.11	−0.20	−0.30
	Max	0.06	−0.03	0.03	0.07
	Mean	−0.04	−0.07	−0.11	−0.14
ROM	Min	1.15	1.86	3.04	3.65
	Max	1.25	2.16	3.83	4.76
	Mean	1.22	1.98	3.35	4.09
SIM	Min	1.17	1.95	3.25	3.95
	Max	1.25	2.43	4.42	5.55
	Mean	1.20	2.21	3.89	4.83
	Min	−0.34	−0.83	−1.60	−2.06
Overall	Max	4.56	7.57	13.00	15.91
	Mean	0.41	0.72	1.26	1.55

BON = Blanco Orejinegro; BR = Brahman (gray and red);
BVH = Braunvieh; GUZ = Guzerat; LIM = Limousin;
NOR = Normand; ROM = Romosinuano; SIM = Simmental;
Min = minimum predicted value; Max = maximum predicted value;
REA4= rib eye area at 4 months; REAW = rib eye area at weaning (230 days);
REAY= rib eye area at year; REAF = rib eye area at 15 months.
[1] Descriptive statistics by breed were computed using sires breeding values; overall: descriptive statistics were constructed using all animals' breeding values
[2] All units in cm^2.

The use of heterogeneous error structures was reported for Nellore cattle in tropical conditions (Mercadante et al. 2010), for crossbred Australian cattle under pasture and feedlot conditions (Mirzaei, Verbyla & Pitchford 2011), and for lambs (Fischer et al. 2006). However, heterogeneous error structure models in these studies were not compared with models fitting a homogeneous residual variance structure. For Colombian Buffaloes, it was found that a model fitting within animal homogeneous variance structure described better REA data (Bolívar, Cerón-Muñoz, Elzo, Ramírez & Agudelo 2011). Meyer (2000), suggested that seasonal variations could be responsible for the heterogeneity in the measurement error. Given that the heterogeneous error variance approach did not show a better fit here, it indicates that environmental factors such as weaning and castration of bulls were not important sources of environmental variation in this multibreed population.

The order of LP used to estimate DAGCF was in agreement with the results found by Mercadante et al. (2010) who compared orders 1, 2 and 3 using AIC and BIC as model selection criteria. However, they did not consider LP of order 1 to model random non genetic effects. In that study, orders of LP to model those effects were either 2 or 3. Mercadante et al. (2010) found that the model considering the lower orders of fit for both direct additive genetic and permanent environmental effects was the best 1. The LP of order one were also reported to be sufficient to explain direct additive genetic effects for weight data in crossbred cattle cows (Arango, Cundiff & Van Vleck 2004).

Considering the small size of the dataset in this study and that a model with only 7 parameters that permitted the use of all records was selected, RRM seem to be a good option to model longitudinal ultrasound data. If a four-trait model assuming zero covariance between direct and maternal additive effects had been fitted here, the number of parameters needed would have been $4 \times (4 \times (4+1)/2) = 40$, which is more than 4 times greater than the number of parameters estimated with the LP1HOMS model. Even if two-trait models had been utilized, a total of 6 two-trait analysis would have had to be performed to estimate the full covariance matrix for REA at the 4 target ages. In addition, because each analysis would be performed separately, there would have been no certainty for the estimated six-trait covariance matrix to be positive definite.

4.2. REML Estimates of Covariance Functions and Covariance Components

The direct additive genetic variance function corresponding to the DAGCF when $t_1 = t_2$ (Figure 2) was concave up with a global minimum at 150 days of age. Thus, the increase in the magnitude of the variance after the minimum point was always positive and greater as the animals grew older. Among the few literature reports using RRM to model ultrasound longitudinal data, a smoother pattern for DAGV (in the age interval 60 to 360 days) was reported for eye muscle depth (a ultrasonic measure at the same point where REA is taken, but measuring depth not area) in lambs (Fischer et al. 2006). Although they found that additive genetic variance did not have great changes, it had a concave up shape. A Nellore cattle study under pasture and feedlot conditions in a tropical region was conducted by Mercadante et al. (2010) in Brazil. However they did not discuss the covariance tendencies. The very low values of DAGV around 150 days here may have been due to computing artifacts rather than biology. Numerical problems have been reported for RRM using LP as base functions (Nobre, Misztal, Tsuruta, Bertrand, Silva & Lopes 2003, Bohmanova, Misztal & Bertrand 2005, Bertrand, Misztal, Robins, Bohmanova & Tsuruta 2006).

The DAGV did not decrease after weaning but it increased with the calf's age. Maternal effects have been found to be important for REA and other ultrasound traits (Speidel et al. 2007). These results suggested that maternal effects would need to be considered in models for genetic analysis of postweaning growth traits. No other literature reports were found for longitudinal REA data considering maternal effects in cattle.

4.3. Heritability and Ratio of MPEV to PhV

The Dh values followed the same trajectory as DAGV. Low values of Dh (particularly at 150 days) could be due to numerical problems related to the population structure and small size of dataset. The only literature report found for Dh of REA in cattle using RRM showed higher values than those reported in the current study. That study considered a range of ages from 323 to 773 days in a Brazilian Nellore cattle population and Dh estimates ranged from 0.31 to 0.42 (Mercadante et al. 2010). The Dh for REA at slaughter for Australian crossbred cattle in pasture conditions until 18 months of age and then placed in feedlot conditions was estimated to be 0.40 (Mirzaei et al. 2011). In a Colombian purebred Brahman population under similar management conditions (pastures and mineral supplementation) to those in this study, Dh for REAF was 0.37 (Jiménez et al. 2010). For Red Angus animals of ages between 300 and 480 days and with a single ultrasonic REA measurement, Speidel et al. (2007) found a Dh estimate of 0.35. Crews & Kemp (1999) suggested that maternal effects were unimportant for the genetic evaluation of carcass traits (including REA) in a multibreed population. However, they did not use RRM because they considered REA data only at slaughter. Thus, differences in the data structure (longitudinal vs. simple), the model used, and the fact that presumably maternal effects have a small effect on traits measured at slaughter could explain the different results. In agreement with results here, for Red Angus cattle, Speidel et al. (2007) concluded (based on a likelihood ratio test) that inclusion of maternal effects improved the ability of genetic models to account for variability on carcass traits. The MPr estimates increased smoothly with age. The MPr had medium to high values across all ages and had a total (maximum value - minimum value) change of 10.8 percentage units. For live weight, under similar conditions and for a *Bos indicus* (Nellore) beef cattle population, Albuquerque & Meyer (2001) found a similar pattern for MPr. No research including maternal permanent environmental effects for REA data in cattle was found in the literature. The MPr values did not decrease after weaning, thus, the permanent maternal environmental effects were important for post weaning development phases. This suggests that remnants of pre-weaning permanent environmental cow effects continued to influence calf REA until 492 days of age. Maternal effects are mainly explained for cow's milk production (genetic to the dam and environmental to the calf). Considering the values of MPr (0.43 to 0.54), it seems that a key point to obtain animals with greater REA, which are expected to have a greater meat production, would be to implement an adequate selection program that includes both direct growth and maternal milk production. It has to be taken into account that although maternal additive genetic effects were not included in the model due to estimation problems, they are still present. On the other hand, the unique maternal effect term in the model is possibly accounting for both: Additive genetic and permanent environment maternal effects.

4.4. Correlations

As the DAGR formed a plateau after approximately 240 days, for genetic evaluation purposes, when considering REA data with ages greater than 240 days (for example, from weaning to greater ages), it will be possible to use a repeatability model. The simplicity of this model will make it desirable, especially for small data sets as present one. For live weight records, a similar conclusion was found by Arango et al. (2004) for crossbred beef cows in a temperate region.

The negative DAGR between ages at the beginning of the trajectory and final ages indicated that those genes controlling REA at ages near to 70 days are antagonist to genes controlling this trait at ages near to 492 days. Taking into account that what matters is REA at ages near slaughter, animals could be selected for REA at ages after 240 days (because of the plateau formed by DAGR occurred after that point). Because MPER values were medium to high across calf ages, it appears that maternal permanent environmental effects exerted a positive effect on REA preweaning, and this effect persisted until 492 days of age. As a general observation taking into account, MPr and MPER values for this population, maternal effects appeared to be important to obtain greater REA.

4.5. Eigenfunctions

The proportion of DAGV explained by the first eigenvalue (95.9%) was in the range of proportions found by Mercadante et al. (2010). Such range was 84% to 99% depending on the model used. A similar proportion (90%) was described for Longissimus muscle depth at the same point where REA was taken in lambs (Fischer et al. 2006). As the DAGEF1 crossed the age axis at 136 days, this is a critical age because selection for greater REA values before this trajectory point will tend to negatively deform the mean population REA growth curve for later ages. Considering only ages after that point, selection for direct additive genetic effects will increase REA mean population growth curve. Thus, selection for REA could be performed after 136 days, i.e., roughly 4 months of age under field conditions. However, considering the high DAGR between 136 days and 240 days of age, a practical age to perform selection for REA would be at weaning.

4.6. Breeding Values

Given the small number of sires considered in the current study (especially for *Bos taurus* breeds) results should be viewed with caution. As expected, all genetic additive direct breed effects were estimable. Thus, the use of orthogonal functions to describe fixed genetic effects when modeling longitudinal data could be useful in order to prevent estimability problems. No research that considered breed effects as a continuous function of age of calf was found in the literature.

Range of BV for REAF of BR sires (Table 4) was smaller than the range reported by Jiménez et al. (2010) for purebred Brahman cattle under pasture conditions in Colombia. They reported EPD values ranging from -2.84 to 3.47 cm^2, thus, the BV (twice the EPD) ranged from -5.68 to 6.94 cm^2. As in the current study, BV were deviated from BR. The range of BV for purebred BR animals (non parents; -0.82 to 1.12 cm^2) was smaller than those reported by Jiménez et al. (2010) suggesting that the amount of genetic variability in the dataset here was smaller than in the Brahman population analyzed by these authors.

The BLUP of BV suggested that among the tested sires and under the conditions of the study LIM bulls had the greatest mean genetic merit for REA at all target ages (Table 4). When all of the sires were ranked according to individual BV, LIM sires were always those with the greatest values. Consequently, the LIM breed would have to be considered for crossbreeding programs with Brahman cows under pasture conditions in the Southern Cesar region of Colombia. The LIM breed had been reported to have greater additive genetic effects for REA at different ages when compared to *Bos indicus* and *Bos taurus* breeds in temperate areas under feedlot or high supplement conditions (Ríos-Utrera, Cundiff, Gregory, Koch, Dikeman, Koohmaraie & Van Vleck 2006, Williams, Aguilar, Rekaya & Bertrand 2010). According to the results of this research, in tropical regions and under pasture conditions, LIM animals also showed a good performance for this trait.

5. Final Remarks

It should be mentioned that genetic parameters and breeding values were estimated with limited accuracy due to the structure and small size of the available multibreed population. Estimates of (co)variance components showed that it is necessary to validate the results of this research with substantially larger multigenerational populations before implement RRM in regional or national genetic evaluation procedures. Thus, there is a need to continue obtaining longitudinal ultrasound information from different beef cattle herds where the breeds studied here are represented. Results suggested that maternal effects were important, both preweaning and postweaning. Thus, maternal effects (genetic and non-genetic) appeared to be relevant effects to be included in models for genetic evaluation of REA pre and postweaning under pasture conditions in Colombia.

Acknowledgments

We sincerely thank two referees for helpful comments and suggestions which led to improve this paper.

[Recibido: agosto de 2011 — Aceptado: abril de 2012]

References

Albuquerque, L. G. & Meyer, K. (2001), ‘Estimates of covariance functions for growth from birth to 630 days of age in nellore cattle’, *Journal of Animal Science* **79**(1), 2776–2789.

Arango, J. A., Cundiff, L. V. & Van Vleck, L. (2004), ‘Covariance functions and random regression models for cow weight in beef cattle’, *Journal of Animal Science* **82**(1), 54–67.

Bertrand, J. K., Misztal, I., Robins, K. R., Bohmanova, J. & Tsuruta, S. (2006), Implementation of random regression models for large scale evaluations for growth in beef cattle, *in* ‘Proceedings of the 8th World Congress on Genetic Applied to Livestock Production’, Minas Gerais: Sociedade Brasileira de Melhoramiento Animal, Belo Horizonte.

Bohmanova, J., Misztal, I. & Bertrand, J. K. (2005), ‘Studies on multiple trait and random regression models for genetic evaluation of beef cattle for growth’, *Journal of Animal Science* **83**(1), 62–67.

Bolívar, D. M., Cerón-Muñoz, M. F., Elzo, M. A., Ramírez, E. J. & Agudelo, D. A. (2011), ‘Growth curves for buffaloes (*Bubalus bubalis*) using random regression mixed models with different structures of residual variances’, *Journal of Animal Science* **89**(1), 62–67. Suppl E1: 530.

Choy, Y. H., Lee, C. W., Kim, H. C., Choi, S. B., Choi, J. G. & Hwang, J. M. (2008), ‘Genetic models for carcass traits with different slaughter endpoints in selected hanwoo herds I. linear covariance models’, *Journal of Animal Science* **21**, 1227–1232.

Crews, D. H. & Kemp, R. A. (1999), ‘Contributions of preweaning growth information and maternal effects for prediction of carcass trait breeding values among crossbred beef cattle’, *Journal of Animal Science* **79**, 17–25.

Draper, N. R. & Smith, H. (1981), *Applied regression analysis*, 2 edn, John Wiley & Sons Inc., New York.

Elzo, M. A. (2010), *Animal breeding notes*, University of Florida, Gainesville.

Elzo, M. A. & Famula, T. R. (1985), ‘Multibreed sire evaluation procedures within a country’, *Journal of Animal Science* **60**, 942–952.

Elzo, M. A. & Wakeman, D. L. (1998), ‘Covariance components and prediction for additive and nonadditive preweaning growth genetic effects in an angus-brahman multibreed herd’, *Journal of Animal Science* **76**, 1290–1302.

FEDEGAN (2006), *Plan estratégico de la ganadería colombiana 2019*, San Martin Obregon y Cía, Bogotá, D.C.

Fischer, T. M., van der Werf, J. H. J., Banks, R. G., Ball, A. J. & Gilmour, A. R. (2006), 'Genetic analysis of weight, fat and muscle depth in growing lambs using random regression models', *Journal of Animal Science* **82**, 13–22.

Hassen, A., Wilson, D. E. & Rouse, G. H. (2003), 'Estimation of genetic parameters for ultrasound-predicted percentage of intramuscular fat in angus cattle using random regression models', *Journal of Animal Science* **81**, 35–45.

Hougton, P. L. & Turlington, L. M. (1992), 'Application of ultrasound for feeding and finishing animals: A review', *Journal of Animal Science* **70**, 930–941.

Jiménez, A., Manrique, C. & Martínez, C. A. (2010), 'Parámetros y valores genéticos para características de composición corporal, área de ojo del lomo y grasa dorsal medidos mediante ultrasonido en la raza brahman', *Revista Medicina Veterinaria y Zootecnica* **57**, 178–190.

Kempthorne, O. (1957), *An Introduction to Genetic Statistics*, John Wiley.

Kirkpatrick, M., Lofsvold, D. & Bulmer, M. (1990), 'Analysis of the inheritance, selection and evolution of growth trajectories', *Genetics* **124**, 979–993.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. & Schabenberger, O. (2006), *SAS for Mixed Models*, Cary (NC): SAS Institute Inc.

Lynch, M. & Walsh, B. (1998), *Genetic and Analysis of Quantitative Traits*, Sinauer Associates, Inc., Arizona.

Mercadante, M. E. Z., El Faro, L., Pinheiro, T. R., Cyrillo, J. N. S. G., Bonilha, S. F. M. & Branco, R. H. (2010), Estimation of heritabality and repeatability for ultrasound carcass traits in nelore cattle using random regression models, *in* 'Proceedings of the 9th World Congress on Genetic Applied to Livestock Production', Leipzig.

Meyer, K. (1998), 'Estimating covariance functions for longitudinal data using a random regression model', *Genetics Selection Evolution* **38**, 221–240.

Meyer, K. (2000), 'Random regression to model phenotypic variation in monthly weights of australian beef cattle', *Livestock Production Science* **65**, 19–38.

Meyer, K. (2007), 'WOMBAT -A program for mixed models analyses in quantitative genetics by REML', *Journal of Zhejiang University Science B* **8**, 815–821.

Meyer, K. & Hill, W. G. (1997), 'Estimation of genetic and phenotypic covariance functions for longitudinal or "repeated" records by restricted maximum likelihood', *Livestock Production Science* **47**, 185–200.

Mirzaei, H. R., Verbyla, A. P. & Pitchford, W. S. (2011), 'Joint analysis of beef growth and carcass quality traits through calculation of co-variance components and correlations', *Genetics and Molecular Research* **10**, 433–447.

Nobre, P. R. C., Misztal, I., Tsuruta, S., Bertrand, J. K., Silva, L. O. C. & Lopes, P. S. (2003), 'Analysis of growth curves of nellore cattle by multiple-trait and random regression models', *Journal of Animal Science* **81**, 918–926.

Ríos-Utrera, A., Cundiff, L. V., Gregory, K. E., Koch, R. M., Dikeman, M. E., Koohmaraie, M. & Van Vleck, L. D. (2006), 'Effects of age, weight, and fact slaughter end points on estimates of breed and retained heterosis effects for carcass traits', *Journal of Animal Science* **84**, 63–87.

Speidel, S. E., Enns, R. M., Brigham, B. W. & Keeman, L. D. (2007), 'Genetic parameter estimates for ultrasound indicators of carcass', *Journal of Animal Science* **58**, 39–42.

Stewart, J. (2008), *Cálculo en varias variables. Trascendentes tempranas*, 6 edn, Cengage Learning, México DF.

Van Soest, P. J. (1994), *Nutritional Ecology of the Ruminant*, 2 edn, Comstock Publishing Sssociates, New York.

Williams, J. L., Aguilar, I., Rekaya, R. & Bertrand, J. K. (2010), 'Estimation of breed and heterosis effects for growth and carcass traits in cattle using published crossbreeding studies', *Journal of Animal Science* **88**, 460–466.

Wilson, D. E. (1992), 'Application of ultrasound for genetic improvement', *Journal of Animal Science* **70**, 973–983.

Información para los autores

La **Revista Colombiana de Estadística** publica artículos originales de carácter teórico o aplicado en cualquiera de las ramas de la estadística. Los artículos puramente teóricos deberán incluir la ilustración de las técnicas presentadas con datos reales o por lo menos con experimentos de simulación, que permitan verificar la utilidad de los contenidos presentados. Se consideran también artículos divulgativos de gran calidad de exposición sobre metodologías o técnicas estadísticas aplicadas en diferentes campos del saber. Únicamente se publican artículos en español e inglés, si el autor escribe en una lengua diferente a la nativa debe enviar un certificado de un traductor oficial o de un corrector de estilo que haya revisado el texto.

El Comité Editor únicamente acepta trabajos para evaluación que no han sido publicados previamente y que no están siendo propuestos simultáneamente para publicación en otros medios, ni lo serán sin previo consentimiento del Comité, a menos que, como resultado de la evaluación, se decida no publicarlos en la Revista. Se supone además que cuando los autores hacen entrega de un documento con fines de publicación en la **Revista Colombiana de Estadística**, conocen las condiciones anteriores y que están de acuerdo con ellas.

Material

Los artículos remitidos a la **Revista Colombiana de Estadística** deben ser presentados en archivo PDF o PS, con textos, gráficas y tablas en color negro y, además, los autores deben agregar una versión del artículo sin nombres ni información de los autores, que se utilizará para el arbitraje. Se debe enviar una carta firmada por cada uno de los autores, donde manifiesten estar de acuerdo con someter el artículo y con las condiciones de la Revista. Si un artículo es aceptado, los autores deben poner a disposición del Comité Editorial los archivos: fuente en LaTeX y de gráficas en formato EPS en blanco y negro.

Para facilitar la preparación del material publicado se recomienda utilizar MiKTeX[1], usando los archivos de la plantilla y del estilo *revcoles* disponibles en la página Web de la Revista[2] y siguiendo las instrucciones allí incorporadas.

Todo artículo debe incluir:

- Título en español y su traducción al inglés.
- Los nombres completos y el primer apellido, la dirección postal o electrónica y la afiliación institucional de cada autor.
- Un resumen con su versión en inglés (*abstract*). El resumen en español no debe pasar de 200 palabras y su contenido debe destacar el aporte del trabajo en el tema tratado.

[1] http://www.ctan.org/tex-archive/systems/win32/miktex/

[2] http://www.estadistica.unal.edu.co/revista

- Palabras clave (*Key words*) en número entre 3 y 6, con su respectiva traducción al inglés, siguiendo las recomendaciones del *Current Index to Statistics* (CIS)[3].

- Cuando el artículo se deriva de una tesis o trabajo de grado debe indicarse e incluirse como una referencia.

- Si se deriva de un proyecto de investigación, se debe indicar el título del proyecto y la entidad que lo patrocina.

- Referencias bibliográficas, incluyendo solamente las que se hayan citado en el texto.

Referencias y notas al pie de página

Para las referencias bibliográficas dentro del texto se debe utilizar el formato autor-año, dando el nombre del autor seguido por el año de la publicación dentro de un paréntesis. La plantilla LaTeX suministrada utiliza, para las referencias, los paquetes BibTeX y Harvard[4]. Se recomienda reducir el número de notas de pie de página, especialmente las que hacen referencia a otras notas dentro del mismo documento y no utilizarlas para hacer referencias bibliográficas.

Tablas y gráficas

Las tablas y las gráficas, con numeración arábiga, deben aparecer referenciadas dentro del texto mediante el número correspondiente. Las tablas deben ser diseñadas en forma que se facilite su presentación dentro del área de impresión de la Revista. En este sentido, los autores deben considerar en particular la extensión de las tablas, los dígitos representativos, los títulos y los encabezados. Las gráficas deben ser visualmente claras y debe ser posible modificar su tamaño. Cuando el artículo sea aceptado para su publicación, los autores deben poner la versión definitiva a disposición del Comité Editorial. Todos los elementos como barras, segmentos, palabras, símbolos y números deben estar impresos en color negro.

Responsabilidad legal

Los autores se hacen responsables por el uso de material con propiedad intelectual registrada como figuras, tablas, fotografías, etc.

Arbitraje

Los artículos recibidos serán revisados por el Comité Editorial y sometidos a arbitraje por pares especializados en el tema respectivo. El arbitraje es “doble ciego” (árbitros anónimos para los autores y viceversa). El Comité Editorial decide aceptar, rechazar o solicitar modificaciones a los artículos con base en las recomendaciones de los árbitros.

[3] http://www.statindex.org/CIS/homepage/keywords.html

[4] http://tug.ctan.org/tex-archive/macros/latex/contrib/harvard

La Revista Colombiana de Estadística agradece a las siguientes personas, por su colaboración en el número especial en aplicaciones en Bioestadística volumen 35 No. 2 (2012).

Alex Sánchez Pla, Ph.D.
Ali Faisal, M.Sc.
Alexander de Leon, Ph.D.
Aurélien Latouche, Ph.D.
Adin Cristian-Andrei, Ph.D.
Adriana Rojas, Especialista
Carla J. Machado, Ph.D.
Carlos Henriquez, M.Sc.
César Sánchez Sellero, Ph.D.
Fabio Alexander Sierra Matamoros, M.Sc.
Flavio H. Freire, Ph.D.
Francisco Díaz, Ph.D.
Gilles Renand, Ph.D.
Guillermo Martinez, Ph.D.
Juan Carlos Garcia Díaz, Ph.D.
Julio A. Di Rienzo, M.Sc.
Laura R. Wong, Ph.D.
Luigi Spezia, Ph.D.
Mª Carmen Iglesias Pérez, Ph.D.
Mario Piscoya, Ph.D.
Moises Alberto Calle Aguirre, Ph.D.
Olivier David, Ph.D.
Ricardo Cifuentes, M.Sc.
Roxana de las Salas Martínez, M.Sc.
Sergio Muñoz, Ph.D.
Zhigang Zhang, Ph.D.