Revista Colombiana de Estadística

Volumen 35. Número 1 - junio - 2012 **ISSN 0120 - 1751**

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

Revista Colombiana de Estadística	Bogotá	Vol. 35	Nº 1
ISSN 0120 - 1751	COLOMBIA	junio-2012	Págs. 1-184

Contenido

Editorial

Leonardo Trujillo[a]

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

Welcome to the first issue of the 35th volume of the Revista Colombiana de Estadistica (Colombian Journal of Statistics). This year we are repeating the success of the previous year by publishing three numbers in the same year. The first number is this one, the regular one and, additional to the traditional one in December, we are publishing a Special Issue about Biostatistics with Professors Liliana Lopez-Kleine and Piedad Urdinola as Guest Editors. We will keep also, as the last number, the characteristic of being an issue entirely published in English language as part of the requirements of being the winners of an Internal Grant for funding at the National University of Colombia (Universidad Nacional de Colombia) among many Journals (see last editorial of December).

The topics in this current issue range over diverse areas of statistics: Two papers in Regression Models by Llinas and Carreno and another one by Ponsot-Balaguer, Sinha, Gotia; two papers in Survey Methodology by Hussain, Shah and Shabbir and another one by Zarate, Sanchez and Marin; one paper in Bayesian Statistics by Usuga and Hernandez; one paper in Categorical Data Analysis by Tahata and Kozai; one paper in Econometrics by Alonso and Serna; one paper in Industrial Statistics by Srinivasa Rao; one paper in Longitudinal Data by Sosa and Diaz and one paper in Nonparametric Statistics by Ojeda, Pulido, Quiroz and Rios.

Last May, there were celebrations in Colombia referring to the Mathematician and the Statistician Day. This is a Colombian celebration known as the Panamerican Day of the Statistics. However, there is not a clear consensus of which one is the agreed date to the Statistician Day in the world. This is good, as we statisticians have many dates to celebrate. Recently, the General Assembly of the United Nations has named the 20th of October as the World Day of Statistics at least until 2015 as this date will be rescheduled every five years. In Argentina, for example, this day is celebrated every 27th of July or in Latin America is well-known the day of "the statistician in health" either in April or September according to the host country. African statisticians celebrate their day on the 18th of November every year and Caribbean statisticians on the 15th of October. What is the purpose of these celebrations? Perhaps for Colombian statisticians could be a good reason to reincorporate the idea of organizing ourselves in a Statistics Society. The lack of this society involving all the academic Statistics departments around the country

[a]General Editor of the Colombian Journal of Statistics, Assistant Professor.
E-mail: ltrujilloo@bt.unal.edu.co

is necessary as the number of alumni in Statistics is increasing exponentially during the last years the number of graduate students as well. Independent of what day you celebrate this date: Happy day for all our statistician readers.

The Colombian Symposium in Statistics has traditionally been, every year, a good way to update statisticians around the country with the last advances in the area and to keep together all the related professionals independently of their city of origin. This year the Symposium will be held at Bucaramanga with important personalities in specialized areas such as Biostatistics, Categorical Data Analysis, Industrial Statistics, Nonparametric Statistics, Quality Control and Survey Sampling (`www.simposioestadistica.unal.edu.co`). Also, Colombia has been designated as the host country for the XIII CLAPEM (Latin American Conference in Probability and Mathematical Statistics) for 2014. Statisticians in Colombia and neighbour countries should take advantage of these opportunities to gather with statisticians around the world but also with local professionals.

This time as the last number in December, I would not like to finish this Editorial without paying a tribute for the 50 years of the death of an eminent statistician: Ronald Fischer (1890-1962). He was a leader scientific of the last XX century: A British biologist, mathematician, and of course, statistician. He was the creator of the inferential statistics in 1920. He introduced the analysis of variance methodology which was considerably superior to the correlation analysis. As being a researcher at the Rothamsted Experimental Station in the UK, he began the study of an extensive collection of data and the results of this study was published under the name of Studies in Crop Variation, a previous essay of all the principles of the Design of Experiments. He was also the founder of the latin squares methodology and his contribution to the Statistics was so huge, it cannot be summarized in this short Editorial. I shall invite the interested readers to follow this excellent web page of Professor John Aldrich at the University of Southampton where almost all Fischer s work is presented as well as biographical notes (`www.economics.soton.ac.uk/staff/aldrich/fischerguide/rafreader.htm`).

Editorial

Leonardo Trujillo[a]

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

---

Bienvenidos al primer número del volumen 35 de la Revista Colombiana de Estadística. Este año estaremos repitiendo el éxito de publicar tres números en un mismo año. El primer número corresponde a este que es el número regular de Junio, y adicional al tradicional en Diciembre, estaremos publicando un número especial en Bioestadística con las Profesoras Liliana López-Kleine y Piedad Urdinola como Editoras Invitadas. También hemos mantenido, al igual que el último número, la condición de ser un número publicado completamente en inglés como parte de los requisitos al ser ganadores de una convocatoria interna para financiación de revistas científicas en la Universidad Nacional de Colombia entre varias revistas de la misma Universidad (ver la última editorial de Diciembre). Para este número, los tópicos varían en diversas áreas de la estadística como son: dos artículos en Metodología de Encuestas por Hussain, Shah and Shabbir y otro de Zárate, Sánchez and Marín; dos artículos en Modelos de Regresión por Llinás y Carreño y otro de Ponsot-Balaguer, Sinha, Gotia; un artículo en Análisis de Datos Categóricos de Tahata y Kozai; uno en Análisis de Datos Longitudinales por Sosa y Díaz; uno en Econometría por Alonso y Serna; uno en Estadística Bayesiana de Usuga y Hernández; uno en Estadística Industrial de Srinivasa Rao; y uno en Estadística no Paramétrica de Ojeda, Pulido, Quiroz and Ríos.

En el último mes de Mayo, el día doce, hubo varias celebraciones del Día del Estadístico y del Matemático en Colombia. Esta es una celebración puramente colombiana en referencia al Día Panamericano de la Estadística. Sin embargo, no hay un claro concenso de cuál es el Día del Estadístico alrededor del mundo. Esto es una buena razón que tenemos los estadísticos para celebrar nuestro día varas veces al año. Recientemente, la Asamblea General de las Naciones Unidas proclamó el día 20 de Octubre como el Día Mundial de la Estadística por lo menos hasta el 2015, pues esta fecha se reevaluará cada cinco años. En Argentina, por ejemplo, el día de los estadísticos se celebra cada 27 de Julio o en Latinoamérica es bien conocido el "Día del Estadístico en la Salud" que bien se celebra en Abril o Septiembre dependiendo del país en que se encuentre. Los estadísticos africanos celebran su día el 18 de Noviembre y los estadísticos de las islas del Caribe el 15 de Octubre. Deberíamos preguntarnos antes que todo: cuál es el propósito de una celebración del Día del Estadístico? Tal vez para los estadísticos en Colombia sería una buena razón para reincorporar la idea de organizarnos en una Sociedad de Estadísticos. La falta de esta sociedad que reúna a todos los departamentos

[a]Editor de la Revista Colombiana de Estadística, Profesor asistente.
E-mail: ltrujilloo@bt.unal.edu.co

de Estadística del país es necesario dado el aumento de programas en Estadística a lo largo de la nación así como del número de estudiantes graduados de ellas. Independiente de que día usted celebre esta fecha, feliz día a nuestros lectores estadísticos.

El Simposio Colombiano de Estadística ha sido tradicionalmente, cada año, una forma de reunir a los estadísticos de todo el país y mantenerlos actualizados en los desarrollos recientes de todas las áreas de la estadística. En este año, 2012, el Simposio tendrá lugar en la ciudad de Bucaramanga con importantes personalidades en áreas especializadas tales como Análisis de Datos Categóricos, Bioestadística, Control de Calidad, Estadística Industrial, Estadística no Paramétrica y Muestreo (`www.simposioestadistica.unal.edu.co`). El Simposio Colombiano de Estadística es organizado en esta oportunidad por la Universidad Nacional de Colombia, la Universidad Industrial de Santander y las Unidades Tecnológicas de Santander. También, es grato anunciar que Colombia ha sido designada como la sede del XIII CLAPEM (Congreso Latinoamericano en Probabilidad y Estadística Matemática) para el año 2014. Los estadísticos en Colombia y en los países cercanos deberían tomar ventaja de estas oportunidades para interactuar con estadísticos locales y provenientes de otras partes del mundo.

Esta vez, como en el último número de Diciembre, no quisiera finalizar esta Editorial sin rendir tributo a los 50 años de la muerte de un eminente estadístico: Ronald Fischer (1890-1962). Fischer fue uno de los científicos líderes del siglo XX: un biólogo, matemático y por supuesto estadístico. Fue el creador de la inferencia estadística hacia 1920. Introdujo la metodología del análisis de varianza la cual se encontró considerablemente superior al análisis de correlación. Mientras era investigador en la Estación Experimental de Rothamsted en el Reino Unido, inició el studio de una extensa colección de datos que le llevaron a publicar sus estudios bajo el nombre de Studies in Crop Variation, el cual fue un ensayo previo a todos los principios del Diseño de Experimentos. También fue el fundador de la metodología de cuadrados latinos en la investigación agrícola y su contribución fue tan extensa que no podría ser resumida en esta corta Editorial. Por esta razón, invito a todos los lectores interesados en conocer el trabajo de Fischer a visitar la excelente página web del Profesor John Aldrich de la Universidad de Southampton donde se resume casi toda la obra de Fischer así como importantes notas biográficas (`www.economics.soton.ac.uk/staff/aldrich/fischerguide/rafreader.htm`).

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 1 a 14

Linearity Measures of the P-P Plot in the Two-Sample Problem

Aplicación de medidas de linealidad del gráfico P-P al problema de dos muestras

Francisco M. Ojeda[1,a], Rosalva L. Pulido[2,b], Adolfo J. Quiroz[2,3,c], Alfredo J. Ríos[1,d]

[1]Departamento de Matemáticas Puras y Aplicadas, Universidad Simón Bolívar, Caracas, Venezuela

[2]Departamento de Cómputo Científico y Estadística, Universidad Simón Bolívar, Caracas, Venezuela

[3]Departamento de Matemáticas, Universidad de Los Andes, Bogotá, Colombia

Abstract

We present a non-parametric statistic based on a linearity measure of the P-P plot for the two-sample problem by adapting a known statistic proposed for goodness of fit to a univariate parametric family. A Monte Carlo comparison is carried out to compare the method proposed with the classical Wilcoxon and Ansari-Bradley statistics and the Kolmogorov-Smirnov and Cramér-von Mises statistics the two-sample problem, showing that, for certain relevant alternatives, the proposed method offers advantages, in terms of power, over its classical counterparts. Theoretically, the consistency of the statistic proposed is studied and a Central Limit Theorem is established for its distribution.

***Key words*:** Nonparametric statistics, P-P plot, Two-sample problem.

Resumen

Se presenta un estadístico no-paramétrico para el problema de dos muestras, basado en una medida de linealidad del gráfico P-P. El estadístico propuesto es la adaptación de una idea bien conocida en la literatura en el contexto de bondad de ajuste a una familia paramétrica. Se lleva a cabo una comparación Monte Carlo con los métodos clásicos de Wilcoxon y Ansari-Bradley, Kolmogorov-Smirnov y Cramér-von Mises para el probelam de dos muestras. Dicha comparación demuestra que el método propuesto ofrece una

[a]Professor. E-mail: fojeda@usb.ve
[b]Professor. E-mail: rosalvaph@gmail.com
[c]Professor. E-mail: aj.quiroz1079@uniandes.edu.co
[d]Professor. E-mail: alfrios@usb.ve

potencia superior frente a ciertas alternativas relevantes. Desde el punto de vista teórico, se estudia la consistencia del método propuesto y se establece un Teorema del Límite Central para su distribución.

***Palabras clave*:** estadísticos no-paramétricos, gráfico P-P, problema de dos muestras.

1. Introduction

Probability plots, usually refered to as P-P plots, are, together with quantile-quantile plots, among the most commonly used tools for informal judgement of the fit of a data set to a hypothesized distribution or parametric family.

Gan & Koehler (1990) propose statistics that can be interpreted as measures of linearity of the P-P plot, for use in goodness of fit testing of univariate data sets to parametric families. They offer, as well, an interesting discussion on how the difference between a distribution and a hypothesized model will be reflected on the corresponding P-P plot. Their discussion is relevant to interpret the results in Section 3 below.

In order to describe the statistic that we will adapt to the two-sample problem, let $X_1, \ldots, X_m$ denote a univariate i.i.d. sample from a distribution that, we believe, might belong in the location-scale parametric family

$$F(\frac{x-\mu}{\sigma}), \qquad \mu \in I\!R, \sigma > 0 \tag{1}$$

for a fixed, continuous distribution F. Let $\hat{\mu}$ and $\hat{\sigma}$ be consistent estimators of μ and σ. Let $p_i = i/(n+1)$ and $Z_{(i)} = F((X_{(i)} - \hat{\mu})/\hat{\sigma})$, $i = 1, \ldots, m$. Let $\overline{Z}$ and $\overline{p}$ denote, respectively, the averages of the $Z_{(i)}$ and the p_i. Except for a squared power irrelevant in our case, one of the statistics proposed by Gan & Koehler (1990) is the following:

$$k(\widehat{X}) = \frac{\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})(p_i - \overline{p})}{\left(\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})^2 \sum_{i=1}^{n}(p_i - \overline{p})^2\right)^{1/2}} \tag{2}$$

Here, $\widehat{X}$ denotes the X sample. The p_i's used above, are the expected values, when we assume that the X_i has a fully specified distribution given by (1), of the transformed order statistics $F((X_{(i)} - \mu)/\sigma)$. Different possibilities for the plotting positions to be used in P-P plots (that is, for the choice of p_i's) are discussed in Kimball (1960). $k(\widehat{X})$ measures the linear correlation between the vectors $(Z_{(i)})_{i \leq n}$ and $(p_i)_{i \leq n}$, which should be high (close to 1) under the null hypothesis. In their paper, Gan & Koehler study some of the properties of $k(\widehat{X})$, obtain approximate (Monte Carlo) quantiles and, by simulation, perform a power comparison with other univariate goodness of fit procedures, including the Anderson-Darling statistic.

In order to adapt the statistic just described to the two-sample problem, one can apply the empirical c.d.f. of one sample to the ordered statistics of the other,

and substitute the values obtained for the Z_i's in formula (2). How this can be done to obtain a fully non-parametric procedure for the univariate two-sample problem is discussed in Section 2, where we consider, as well, the consistency of the proposed statistic and establish a Central Limit Theorem for its distribution. In Section 3, a Monte Carlo study is presented that investigates the convergence of the finite sample quantiles of our statistic to their limiting values and compares, in terms of power, the proposed method with the classical Wilcoxon and Ansari-Bradley statistics for the two-sample problem.

2. Measures of Linearity for the Two-sample Problem

We will consider the non-parametric adaptation of the statistic of Gan & Koehler (1990), described above, to the univariate two-sample problem. In this setting we have two i.i.d. samples: $X_1, \ldots, X_m$, produced by the continuous distribution $F(x)$ and $Y_1, \ldots, Y_n$, coming from the continuos distribution $G(y)$. These samples will be denoted $\widehat{X}$ and $\widehat{Y}$, respectively. Our null hypothesis is $F = G$ and the general alternative of interest is $F \neq G$. Let $F_m(\cdot)$ denote the empirical cumulative distribution function (c.d.f.) of the X sample. By the classical Glivenko-Cantelli Theorem, as m grows, F_m becomes an approximation to F and, under our null hypothesis, to G. Therefore, if we apply F_m to the ordered statistics for the Y sample, $Y_{(1)}, \ldots, Y_{(n)}$, we will obtain, approximately (see below), the beta distributed variables whose expected values are the p_i of Gan and Koehler's statistics. Thus, the statistic that we will consider for the two-sample problem is

$$\eta(\widehat{X}, \widehat{Y}) = \frac{\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})(p_i - \overline{p})}{\left(\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})^2 \sum_{i=1}^{n}(p_i - \overline{p})^2\right)^{1/2}} \tag{3}$$

with $Z_{(i)} = F_m(Y_{(i)})$. Our first theoretical result is that $\eta(\cdot, \cdot)$, indeed, produces a non-parametric procedure for the two-sample problem.

Theorem 1. *Under the null hypothesis, the statistic $\eta(\widehat{X}, \widehat{Y})$, just defined, is distribution free (non-parametric), for the two-sample problem, over the class of i.i.d. samples from continuous distributions.*

Proof. The argument follows the idea of the proof of Theorem 11.4.3 in Randles & Wolfe (1979). Since the p_i are constants, $\eta(\widehat{X}, \widehat{Y})$ is a function only of the vector $(F_m(Y_1), F_m(Y_2), \ldots, F_m(Y_n))$ only. Thus, it is enough to show that the distribution of this vector does not depend on F under the null hypothesis. Now, for $i_1, i_2, \ldots, i_n$ in $\{0, 1, \ldots, m\}$, we have, by definition of F_m,

$$\begin{aligned}&\Pr(F_m(Y_1) = i_1/m, F_m(Y_2) = i_2/m, \ldots, F_m(Y_n) = i_n/m) = \\ &\Pr(X_{(i_1)} \leq Y_1 < X_{(i_1+1)}, X_{(i_2)} \leq Y_2 < X_{(i_2+1)}, \ldots X_{(i_n)} \leq Y_n < X_{(i_n+1)}),\end{aligned} \tag{4}$$

where, if $i_j = 0$, $X_{(0)}$ must be taken as $-\infty$ and, similarly, if $i_j = m$, $X_{(m+1)}$ must be understood as $+\infty$. Consider the variables $U_i = F(X_i)$, for $i \leq m$ and

$V_j = F(Y_j)$, for $j \leq n$. Under the null hypothesis, the U_i and V_j are i.i.d. Unif(0,1) and, since F is non-decreasing, the probability in (4) equals

$$\Pr(U_{(i_1)} \leq V_1 < U_{(i_1+1)}, U_{(i_2)} \leq V_2 < U_{(i_2+1)}, \ldots U_{(i_n)} \leq V_n < U_{(i_n+1)})$$

which depends only on i.i.d. uniform variables, finishing the proof. □

Theorem 11.4.4 in Randles & Wolfe (1979) identifies the distribution of $F_m(Y_{(i)})$ as the inverse hypergeometric distribution whose properties were studied in Guenther (1975). The study of these results in Randles & Wolfe (1979) is motivated by the consideration of the exceedance statistics of Mathisen (1943) for the two-sample problem.

Theorem 1 allows us to obtain generally valid approximate null quantiles to the distribution of $\eta(\widehat{X}, \widehat{Y})$, in the two-sample setting, by doing simulations in just one case: $F = G =$ the Unif(0,1) distribution.

We will now study the consistency of $\eta(\widehat{X}, \widehat{Y})$ (and a symmetrized version of it) as a statistic for the two sample problem. We begin by establishing a Strong Law of Large Numbers (SLLN) for $\eta(\widehat{X}, \widehat{Y})$.

Theorem 2. *Suppose that F and G are continuous distributions on $\mathbb{R}$. Then, as m and n go to infinity, $\eta(\widehat{X}, \widehat{Y}) \to cor(F(Y), G(Y))$, almost sure (a.s.), where Y has distribution G and 'cor' stands for 'correlation'.*

Proof. We will only verify that $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(p_i - \overline{p})$ converges, a.s., as $n, m \to \infty$, to $\mathrm{Cov}(F(Y), G(Y))$. The quantities in the denominator of η are studied similarly. Let $G_n(\cdot)$ denote the empirical c.d.f. associated to the Y sample and let, also, $\overline{F}_m = (1/m)\sum F_m(Y_i)$ and $\overline{G}_n = (1/n)\sum G_n(Y_i)$. Observe that $p_i = (n/(n+1))G_n(Y_{(i)})$. It follows that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(p_i - \overline{p}) &= \frac{1}{n+1}\sum_{i=1}^{n}(F_m(Y_{(i)}) - \overline{F}_m)(G_n(Y_{(i)}) - \overline{G}_n) \\ &= \frac{1}{n}\sum_{i=1}^{n}(F(Y_i) - \mathbb{E}\,F(Y_1))(G(Y_i) - \mathbb{E}\,G(Y_1)) + r_{m,n} \end{aligned} \tag{5}$$

Repeated application of the Glivenko-Cantelli Theorem and the SLLN shows that $r_{m,n} \to 0$, a.s., as $m, n \to \infty$, finishing the proof. □

According to Theorem 2, when the null hypothesis: $F = G$ holds, $\eta(\widehat{X}, \widehat{Y})$ will converge to 1. In order to have consistency of the corresponding statistic for the two-sample problem, we would like to have the reciprocal of this statement to hold: If $F \neq G$ then the limit of $\eta(\widehat{X}, \widehat{Y})$ is strictly less than one. Unfortunately, this is not the case, as the following example shows.

Example 1. Let F and G be the Unif(0,2) distribution and the Unif(0,1) distribution, respectively. Then, $\mathrm{cor}(F(Y), G(Y)) = 1$ and, therefore, $\eta(\widehat{X}, \widehat{Y})$ applied to samples from F and G will converge to 1.

The counter-example just given suggests the consideration of a 'symmetrized' version of η in order to attain consistency of the statistic against the general alternative $F \neq G$. For this purpose, one could define

$$\eta_{\text{symm}} = \frac{1}{2}(\eta(\widehat{X}, \widehat{Y}) + \eta(\widehat{Y}, \widehat{X})) \tag{6}$$

For η_{symm}, we have the following result.

Theorem 3. *Let the X and Y samples be obtained from the continuous distributions F and G with densities f and g, respectively, such that the sets $\mathcal{S}_f = \{x : f(x) > 0\}$ and $\mathcal{S}_g = \{x : g(x) > 0\}$ are open. Then, η_{symm} converges to 1, a.s., as $n, m \to \infty$ if, and only if, $F = G$.*

Proof. In view of Theorem 2, we only need to show that, if $F \neq G$, then either $\text{corr}(F(Y), G(Y)) \neq 1$ or $\text{corr}(F(X), G(X)) \neq 1$, where the variables X and Y have distributions F and G, respectively. Let λ denote Lebesgue measure in $\mathbb{R}$. Suppose first that $\lambda(\mathcal{S}_g \setminus \mathcal{S}_f) > 0$. Then, there is an interval $J \subset \mathbb{R}$ such that $g(x) > 0$ for $x \in J$, while $f(x) \equiv 0$ on J. Suppose $\text{corr}(F(Y), G(Y)) = 1$. Then, there are constants a and b, with $a \neq 0$ such that, with probability 1, $G(Y) = a\,F(Y) + b$. By the continuity of the distributions and the fact that g is positive on J, it follows that

$$G(y) = a\,F(y) + b, \text{ for all } y \in J \tag{7}$$

Taking derivatives on both sides, we have, for all $y \in J$,

$$0 < g(y) = a\,f(y) = 0$$

a contradiction. The case $\lambda(\mathcal{S}_f \setminus \mathcal{S}_g) > 0$ is treated similarly.

It only remains to consider the case when $\lambda(\mathcal{S}_f \,\Delta\, \mathcal{S}_g) = 0$, where Δ denotes "symmetric difference" of sets. In this case we will show that $\text{corr}(F(Y), G(Y)) = 1$ implies $F = G$. Suppose that $\text{corr}(F(Y), G(Y)) = 1$. For J any open interval contained in $\mathcal{S}_g$, we have, by the argument of the previous case, $g(x) = a\,f(x)$ in J. Since $\mathcal{S}_g$ is open, it follows that $a\,f$ and g coincide on $\mathcal{S}_g$. Since $\lambda(\mathcal{S}_f \,\Delta\, \mathcal{S}_g) = 0$ and f and g are probability densities, a must be 1 and $F = G$, as desired. □

The result in Theorem 3 establishes the consistency of η_{symm} against general alternatives, and is, therefore, satisfactory from the theoretical viewpoint. According to the results given so far in this section, η would fail to be consistent only in the case when one of the supports of the distributions considered is strictly contained in the other and, in the smaller support, the densities f and g are proportional, which is a very uncommon situation in statistical practice. Therefore, we feel that, in practice, both the statistics η and η_{symm} can be employed with similar expectations for their performances. The results from the power analysis in Section 3 support this belief, since the power numbers for both statistics considered tend to be similar, with a slight superiority of η_{symm} in some instances.

The purpose of next theorem is to show that an appropriate standarization of the statistic η has a limiting Gaussian distribution, as m and n tend to infinite.

This will allow the user to employ the Normal approximation for large enough sample sizes. Of course, for smaller sample sizes the user can always employ Monte Carlo quantiles for η, which are fairly easy to generate according to Theorem 1. Some of these quantiles appear in the tables presented in Section 3.

Theorem 4. *Suppose that the X and Y samples, of size m and n, respectively, are obtained from the continuous distribution F (=G). Let $N = m + n$ and suppose that $N \to \infty$ in such a way that $m/N \to \alpha$, with $0 < \alpha < 1$ (the "standard" conditions in the two-sample setting). Let $\xi_{1,0} = 0.00138\overline{8}$ and $\xi_{0,1} = 0.00\overline{5}/36$, where the bar over a digit means that this digit is to be repeated indefinitely. Let*

$$D = D(\widehat{X}, \widehat{Y}) = \frac{1}{n}\left(\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})^2 \sum_{i=1}^{n}(p_i - \overline{p})^2\right)^{1/2}$$

$D(\widehat{X}, \widehat{Y})$ is the denominator of $\eta(\widehat{X}, \widehat{Y})$ after division by n. Then, as $N \to \infty$, the distribution of

$$W = W(\widehat{X}, \widehat{Y}) = \sqrt{N}\left(\eta(\widehat{X}, \widehat{Y}) - \frac{1}{12D}\right) \tag{8}$$

converges to a Gaussian distribution with mean 0 and variance

$$\sigma_W^2 = 144 \times \left(\frac{\xi_{1,0}}{\alpha} + \frac{9\,\xi_{0,1}}{1-\alpha}\right) \tag{9}$$

Proof. Let $C = C(\widehat{X}, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})(p_i - \overline{p})$. C is the numerator of $\eta(\widehat{X}, \widehat{Y})$ after division by n. The idea of the proof is to show that, essentially, C is a two sample V-statistic of degrees (1,3), and then to use the classical Central Limit Theorem for V-statistics which, in the present case, gives the same limit distribution of the corresponding U-statistic. Then the result will follow by observing that D satisfies a Law of Large Numbers. □

Using, as in Theorem 2, that $p_i = G_n(Y_{(i)})$, we can show that, with probability one (ignoring ties between sample points, which have probability zero)

$$C = \frac{1}{m\,n^2(n+1)} \sum_{j,i,k,r} \mathbb{1}_{\{X_j < Y_i, Y_k < Y_i\}} - \mathbb{1}_{\{X_j < Y_i, Y_k < Y_r\}} \tag{10}$$

where, j goes from 1 to m, while i, k and r range from 1 to n. Thus, except for an irrelevant multiplying factor of $n/(n+1)$, C is the V-statistic associated to the kernel

$$h^*(X; Y_1, Y_2, Y_3) = \mathbb{1}_{\{X < Y_1, Y_2 < Y_1\}} - \mathbb{1}_{\{X < Y_1, Y_2 < Y_3\}} \tag{11}$$

The symmetric version of this kernel is

$$h(X; Y_1, Y_2, Y_3) = \frac{1}{6}\sum_{\tau} \mathbb{1}_{\{X < Y_{\tau(1)}, Y_{\tau(2)} < Y_{\tau(1)}\}} - \mathbb{1}_{\{X < Y_{\tau(1)}, Y_{\tau(2)} < Y_{\tau(3)}\}} \tag{12}$$

where τ runs over the permutations of $\{1, 2, 3\}$. It is easy to see that, under the null hypothesis, the expected value of $h(X; Y_1, Y_2, Y_3)$ is $\gamma = 1/12$. By the two-sample version of the Lemma in Section 5.7.3 of Serfling (1980), it follows that the limiting distribution of C, after standardization, is the same as that for the corresponding U-statistic, for which the sum in (10) runs only over distinct indices i, j and k. Then, according to Theorem 3.4.13 in Randles & Wolfe (1979), $\sqrt{N}(C - \gamma)$ converges, in distribution, to a zero mean Normal distribution, with variance

$$\sigma_C^2 = \frac{\xi_{1,0}}{\alpha} + \frac{9\,\xi_{0,1}}{1-\alpha}$$

where

$$\begin{aligned} \xi_{1,0} = &\ \mathrm{Cov}(h(X; Y_1, Y_2, Y_3), h(X; Y_1', Y_2', Y_3')) \quad \text{while} \\ \xi_{0,1} = &\ \mathrm{Cov}(h(X; Y_1, Y_2, Y_3), h(X'; Y_1, Y_2', Y_3')) \end{aligned}$$

for i.i.d. $X, Y_1, Y_2, Y_3, X', Y_1'$, Y_2' and Y_3' with distribution F. These covariances depend on the probabilities of certain sets of inequalities between the variables involved. Since the vector of ranks of the variables involved has the uniform distribution on the set $\mathcal{S}_7$ of permutations of seven elements, the required probabilities can be computed by inspection on $\mathcal{S}_7$ (with the help of an *ad hoc* computer program), to obtain the numbers given in the statement of the Theorem.

On the other hand, under the null hypothesis, using that $F(Y_i)$ has the U(0,1) distribution, and following the procedure in the proof of Theorem 2, one can check that both $(1/n)\sum_{i=1}^n (Z_{(i)} - \overline{Z})^2$ and $(1/n)\sum_{i=1}^n (p_i - \overline{p})^2$ converge, a.s. to $1/12$. It follows that $D(\widehat{X}, \widehat{Y})$ converges, in probability, to $1/12$. Then, Theorem 2.4 follows from an application of Slutsky's Theorem.

For small values of m and n, the distribution of W in (8) displays a negative skewness, that makes inadequate the use of the Gaussian approximation given by Theorem 4. Figure 1 displays the histogram of a sample of 10,000 values of W obtained from simulated X and Y samples of size 500 ($m = n = 500$) from the Unif(0,1) distribution. We see that for these sample sizes, the distribution of W, displayed in Figure 1, is near the bell shape of the Gaussian family. For this combination of m and n, the asymptotic variance of W, given by (9), is $\sigma_W^2 = 0.8$. Figure 2 shows the P-P plot obtained by applying the N(0,0.8) cummulative distribution function to the order statistics of the W sample and plotting these against the plotting positions, p_i. The closeness to a 45° straight line suggests that the Gaussian approximation is valid for this combination of m and n. We conclude that, when the smaller of m and n is, at least, 500, the Gaussian approximation given by Theorem 4 can be used for the distribution of $\eta(\widehat{X}, \widehat{Y})$, rejecting the null hypothesis when W falls below a prescribed quantile, say 5%, of the N(0,σ_W^2) distribution.

3. Monte Carlo Evaluation of $\eta(\widehat{X}, \widehat{Y})$

All the simulations described here were programmed using the R Statistical Language (see R Development Core Team 2011) on a laptop computer. Tables 1

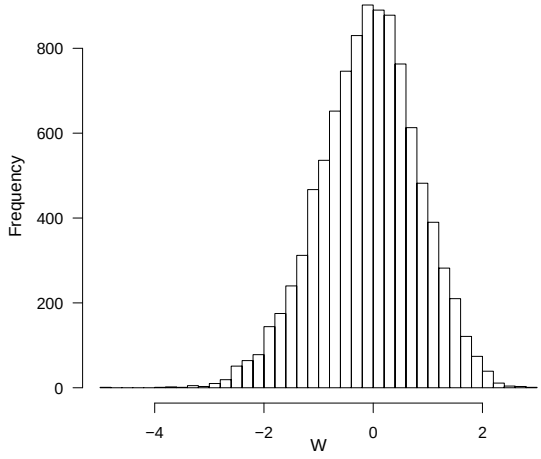

FIGURE 1: Histogram of W for $m = n = 500$.

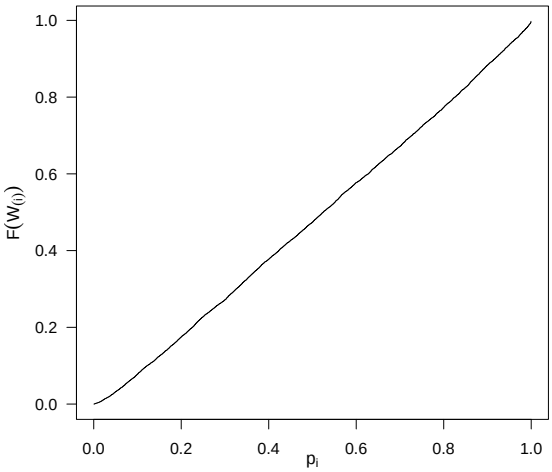

FIGURE 2: P-P plot of W for $m = n = 500$.

and 2 display Monte Carlo null quantiles for the statistics η and η_{symm}, obtained from 10,000 independent pairs of samples for each choice of m and n, using, without loss of generality, data with the Unif(0,1) distribution. Table 2 contains entries for sample size pairs of the form $m \leq n$ only, since, by the symmetry of the statistic, the quantiles are the same when the roles of m and n are interchanged. We see in these tables the convergence towards 1 of all quantiles, as m and n grow, as predicted by Theorem 3. We see, as well, that the quantiles are very similar for both statistics.

In order to evaluate the performance of η and η_{symm} as test statistics for the null hypothesis of equality of distributions, we will consider their power against different alternatives, in comparison to the classical non-parametric tests of Wilcoxon and

TABLE 1: Monte Carlo null quantiles for $\eta(\widehat{X}, \widehat{Y})$.

m	n	1%	2.5%	5%	10%
25	25	0.8956	0.9137	0.9290	0.9436
25	50	0.9203	0.9371	0.9469	0.9576
50	25	0.9235	0.9365	0.9472	0.9578
25	100	0.9363	0.9466	0.9555	0.9646
100	25	0.9360	0.9479	0.9569	0.9656
50	50	0.9471	0.9572	0.9644	0.9715
50	100	0.9624	0.9682	0.9740	0.9788
100	50	0.9598	0.9680	0.9735	0.9786
100	100	0.9744	0.9787	0.9822	0.9858

TABLE 2: Monte Carlo null quantiles for η_{symm}.

m	n	1%	2.5%	5%	10%
25	25	0.8969	0.9171	0.9313	0.9441
25	50	0.9248	0.9374	0.9482	0.9584
25	100	0.9348	0.9474	0.9565	0.9652
50	50	0.9483	0.9581	0.9649	0.9720
50	100	0.9602	0.9682	0.9738	0.9791
100	100	0.9743	0.9790	0.9823	0.9857

Ansari-Bradley, described, for instance, in Hollander & Wolfe (1999). Wilcoxon's test is specifically aimed at detecting differences in location while the statistic of Ansari-Bradley is designed to discover differences in scale. We will also include in our comparison two of the classical tests based on the empirical distribution function (EDF), namely, the two-sample versions of the Kolmogorov-Smirnov and Cramér-von Mises statistics, which are consistent against arbitrary differences in the distribution functions of the samples. These EDF statistics are described in Darling (1957). We will use the particular implementation of the Cramér-von Mises statistic studied by Anderson (1962). As alternatives, we include first the classical scenarios of difference in mean and difference in scale, between Gaussian populations. More precisely, in our first alternative, denoted Δ-mean in the tables below, the sample $\widehat{X}$ has a $N(0,1)$ distribution and $\widehat{Y}$ has the $N(0.4,1)$ distribution, while for our second alternative, denoted Δ-scale in the tables, $\widehat{X}$ has the $N(0,1)$ distribution and $\widehat{Y}$ has a normal distribution with mean zero and variance $\sigma_Y^2 = 3$. Our remaining alternatives seek to explore the advantages of η and η_{symm} when the X and Y distributions have the same mean and variance, but differ in their shape. The Weibull distribution, as described in Johnson, Kotz & Balakrishnan (1995), Chapter 21, with shape parameter $a = 1.45$ and scale parameter $b = 2.23$, has mean and variance both nearly 2.0, and exhibits right skewness. For our third alternative, denoted Gaussian vs. right-skewed, the sample $\widehat{X}$ has the $N(2,2)$ distribution, while $\widehat{Y}$ has the Weibull distribution with parameters (1.45,2.23). In order to produce a distribution with mean and variance equal 2

and left skewness, we take $X = 4 - Z$, where Z has the Gamma distribution with shape parameter $a = 2$ and scale $s = 1$. In our fourth scenario, denoted left-skewed vs. Gaussian, the sample $\widehat{X}$ comes from the distribution just described, while $\widehat{Y}$ has the $N(2, 2)$ distribution. Finally, we consider the situation of right skewness vs. left skewness, in which $\widehat{X}$ comes from the Weibull(1.45,2.23) distribution, while $\widehat{Y}$ is distributed as $4 - Z$, with $Z \sim$ Gamma(2,1).

Tables 3 to 7 display, as percentages, the power against the alternatives just described, of the six statistics compared, namely, Wilcoxon (W), Ansari-Bradley (AB), Kolmogorov-Smirnov (KS), Cramér-von Mises (CvM), η, and η_{symm}, at level 10%. The power is computed based on 1,000 independent pair of samples for each m and n combination with the given alternative distributions, using as reference the 10% quantiles given in Tables 1 and 2 for η and η_{symm}.

Table 3: Monte Carlo power against Δ-mean at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	38.5	8.5	32.4	36.1	22.8	23.5
25	50	47.9	10.0	43.7	45.0	29.5	27.0
50	25	49.3	10.6	42.9	44.1	24.3	28.1
50	50	63.9	10.1	58.3	61.5	36.2	39.8
50	100	73.4	9.8	65.3	70.0	43.2	44.6
100	50	72.2	9.4	63.0	69.7	44.2	43.8
100	100	87.3	10.2	80.7	85.3	55.5	56.1

Table 4: Monte Carlo power against Δ-scale at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	10.9	66.6	25.4	24.0	13.9	22.3
25	50	7.9	77.2	33.2	28.9	13.9	22.1
50	25	14.7	76.1	39.7	32.0	21.8	29.2
50	50	6.3	88.0	47.5	50.0	27.4	35.6
50	100	8.1	96.2	56.4	62.9	36.1	34.9
100	50	13.1	95.1	56.7	64.8	42.7	45.6
100	100	11.5	99.2	77.6	85.5	61.7	56.1

In Table 3 we see, as expected, that for the shift in mean scenario, the Wilcoxon test has the best performance, followed by the KS and CvM statistics. In this case the performances of η and η_{symm} are similar and inferior to that of the EDF statistics, while the Ansari-Bradley statistic has practically no power beyond the test level against the location alternative. The situation depicted in Table 4 (shift in scale) is similar, but now the Ansari-Bradley statistic is the one displaying the best power by far, followed by KS, CvM, η_{symm}, and η, in that order, while the Wilcoxon test shows basically no power against this alternative, as should be expected.

TABLE 5: Monte Carlo power for Gaussian vs. right-skewed at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	9.4	10.3	16.0	14.3	23.5	22.3
25	50	10.9	10.9	18.5	14.8	28.8	29.6
50	25	9.9	12.9	18.8	14.6	25.9	27.6
50	50	11.9	10.8	19.6	19.1	35.3	35.3
50	100	11.8	10.5	24.5	22.5	41.5	42.8
100	50	13.3	13.7	23.0	22.1	41.8	43.9
100	100	14.3	14.1	27.6	24.4	55.8	53.2

TABLE 6: Monte Carlo power for left-skewed vs. Gaussian at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	12.9	13.2	18.2	17.5	23.9	27.4
25	50	15.3	13.5	22.9	18.7	28.1	33.2
50	25	11.5	15.9	20.7	15.8	30.6	33.7
50	50	16.6	16.0	25.1	23.2	39.5	42.0
50	100	18.2	15.8	28.4	25.7	46.7	53.8
100	50	14.9	18.9	30.2	27.5	52.9	53.9
100	100	19.6	18.9	36.4	35.4	66.7	65.4

TABLE 7: Monte Carlo power for right-skewed vs. left-skewed at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	17.7	14.7	31.5	28.7	53.7	54.1
25	50	22.4	15.4	43.3	38.3	69.1	70.5
50	25	20.5	15.2	43.9	38.1	65.4	70.9
50	50	25.9	15.0	50.4	48.4	84.5	85.2
50	100	30.7	15.8	60.2	60.8	92.6	93.0
100	50	27.8	17.7	60.3	61.7	93.2	92.0
100	100	38.2	15.4	80.5	83.2	98.7	98.8

In Tables 5, 6 and 7, in which the distributions considered have the same mean and variance, with differences in their skewness, the results change significantly respect to the previous tables. In these scenarios, the best power clearly corresponds to η_{symm} and η, which for some of the sample sizes nearly double the power of the KS and CvM statistics, which come next in power after η_{symm} and η. In order to understand why the proposed statistics enjoy such good power in the "difference in skewness" scenarios, the reader is advised to see Section 2 in Gan & Koehler (1990), where through several examples (and figures) it is shown the marked departure from linearity that differences in skewness can produce on a P-P plot.

From the power results above, we conclude that η and η_{symm} can be considered a useful non-parametric statistic for the null hypothesis of equality of distributions,

and its application can be recommended specially when differences in shape between F and G are suspected, instead of differences in mean or scale. The power of the two statistics studied here tends to be similar, with η_{symm} being slightly superior in some cases.

We finish this section with the application of our statistic to a real data set. For this purpose, we consider the well known drilling data of Penner & Watts (1991), that has been used as illustrative example of a two-sample data set in Hand, Daly, Lunn, McConway & Ostrowski (1994) and Dekking, Kraaikamp, Lopuhaa & Meester (2005). In these data, the times (in hundreths of a minute) for drilling 5 feet holes in rock were measured under two different conditions: *wet drilling*, in which cuttings are flushed with water, and *dry drilling*, in which cuttings are flushed with compressed air. Each drilling time to be used in our analysis is actually the average of three measures performed at the same depth with the same method, except when some of the three values might be missing, in which case the reported value is the average of the available measurements at the given depth. The sample sizes for these data are $m = n = 80$. Figure 3 shows the P-P plot for the drilling data. In this case, in order to compare the empirical cummulative distribution for the two data sets, the plot consists of the pairs $(F_m(z), G_n(z))$, where z varies over the combined data set and F_m and G_n are, respectively, the EDFs for the dry drilling and wet drilling data. In this figure a strong departure from linearity is evident. This is due to the fact that most of the smallest drilling times correspond to dry drilling, while a majority of the largest drilling times reported correspond to wet drilling, making the plot very flat in the left half and steep in the right half.

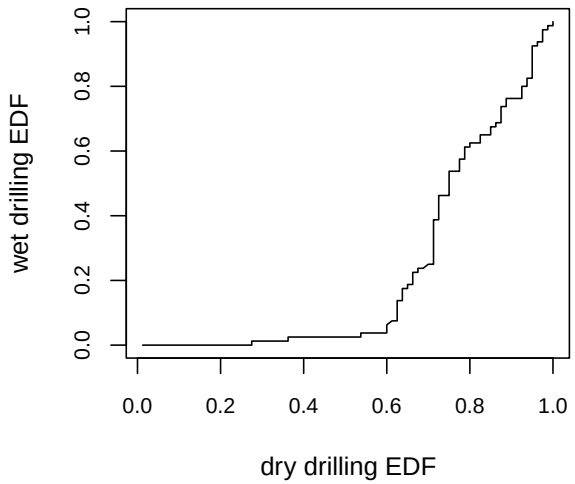

FIGURE 3: P-P Plot for dry drilling vs. wet drilling data.

In order to apply the statistic η to the drilling data, we compute first Monte Carlo null quantiles for η in the case $m = n = 80$, using, as done for Table 1,

10,000 pairs of samples of size 80 from the Unif(0,1) distribution. These quantiles turn out to be the following

1%	2.5%	5%	10%
0.9664	0.9728	0.9777	0.9821

The value of $\eta(\widehat{X}, \widehat{Y})$, taking the dry drilling data as $\widehat{X}$, is 0.9508, which is significant against the null hypothesis of equality of distributions, at the 1% level. Furthermore, comparing the actual value of $\eta(\widehat{X}, \widehat{Y})$ for the drilling data with the 10,000 values calculated for the Monte Carlo null quantile estimation, we obtain an approximate p-value for this data set of 0.0013. Thus, the evidence against equality of distribution is strong in this case.

Statistics based on ideas similar to those leading to $\eta(\widehat{X}, \widehat{Y})$ have been considered in the multivariate case by Liu, Parelius & Singh (1999), who consider statistics based on the Depth-Depth plot. Although generalization of $\eta(\widehat{X}, \widehat{Y})$ to the multivariate case is possible, we do not pursue this line of work, since in the generalization, the full non-parametric character of the statistic is lost and the computation of reference quantiles becomes computationally expensive, thus losing the ease of computation that the statistic enjoys in the univariate case.

4. Conclusions

A modified non-parametric version of the statistic proposed by Gan & Koehler (1990) for the goodness of fit of a univariate parametric family was presented based on a linearity measure of the P-P plot for the two-sample problem. A Monte Carlo comparison was carried out to compare the proposed method with the classical ones of Wilcoxon and Ansari-Bradley for the two-sample problem and the two-sample versions of the Kolmogorov-Smirnov and Cramer-von Mises statistics, showing that, for certain relevant alternatives, the method proposed offers advantages, in terms of power, over its classical counterparts. Theoretically, the consistency of the statistic proposed was studied and a Central Limit Theorem was established for its distribution.

[Recibido: febrero de 2010 — Aceptado: octubre de 2011]

References

Anderson, T. W. (1962), 'On the distribution of the two sample Cramér- von Mises criterion', *Annals of Mathematical Statistics* **33**(3), 1148–1159.

Darling, D. A. (1957), 'The Kolmogorov-Smirnov, Cramér-von Mises tests', *Annals of Mathematical Statistics* **28**(4), 823–838.

Dekking, F. M., Kraaikamp, C., Lopuhaa, H. P. & Meester, L. E. (2005), *A Modern Introduction to Probability and Statistics*, Springer-Verlag, London.

Gan, F. F. & Koehler, K. J. (1990), 'Goodness-of-fit tests based on P-P probability plots', *Technometrics* **32**(3), 289–303.

Guenther, W. C. (1975), 'The inverse hypergeometric - a useful model', *Statistica Neerlandica* **29**, 129–144.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. & Ostrowski, E. (1994), *A Handbook of Small Data Sets*, Chapman & Hall, Boca Raton, Florida.

Hollander, M. & Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, 2 edn, John Wiley & Sons, New York.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, 2 edn, John Wiley & Sons, New York.

Kimball, B. F. (1960), 'On the choice of plotting positions on probability paper', *Journal of the American Statistical Association* **55**, 546–560.

Liu, R. Y., Parelius, J. M. & Singh, K. (1999), 'Multivariate analysis by data depth: Descriptive statistics, graphics and inference', *The Annals of Statistics* **27**(3), 783–858.

Mathisen, H. C. (1943), 'A method for testing the hypothesis that two samples are from the same population', *The Annals of Mathematical Statistics* **14**, 188–194.

Penner, R. & Watts, D. G. (1991), 'Mining information', *The Annals of Statistics* **45**(1), 4–9.

R Development Core Team (2011), 'R: A language and environment for statistical computing'. Vienna, Austria.
*http://www.R-project.org/

Randles, R. H. & Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, Krieger Publishing, Malabar, Florida.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons, New York.

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 15 a 38

Bayesian Analysis for Errors in Variables with Changepoint Models

Análisis bayesiano para modelos con errores en las variables con punto de cambio

Olga Cecilia Usuga[1,2,a], Freddy Hernández[2,b]

[1]Departamento de Ingeniería Industrial, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia

[2]Departamento de Estadística, Instituto de Matemáticas y Estadística, Universidad de São Paulo, São Paulo, Brasil

Abstract

Changepoint regression models have originally been developed in connection with applications in quality control, where a change from the in-control to the out-of-control state has to be detected based on the available random observations. Up to now various changepoint models have been suggested for differents applications like reliability, econometrics or medicine. In many practical situations the covariate cannot be measured precisely and an alternative model are the errors in variable regression models. In this paper we study the regression model with errors in variables with changepoint from a Bayesian approach. From the simulation study we found that the proposed procedure produces estimates suitable for the changepoint and all other model parameters.

***Key words*:** Bayesian analysis, Changepoint models, Errors in variables models.

Resumen

Los modelos de regresión con punto de cambio han sido originalmente desarrollados en el ámbito de control de calidad, donde, basados en un conjunto de observaciones aleatorias, es detectado un cambio de estado en un proceso que se encuentra controlado para un proceso fuera de control. Hasta ahora varios modelos de punto de cambio han sido sugeridos para diferentes aplicaciones en confiabilidad, econometría y medicina. En muchas situaciones prácticas la covariable no puede ser medida de manera precisa, y un modelo alternativo es el de regresión con errores en las variables. En este trabajo estudiamos el modelo de regresión con errores en las variables con

[a]Assistant professor. E-mail: ousuga@udea.edu.co
[b]Ph.D. Student in Statistic. E-mail: fhernanb@ime.usp.br

punto de cambio desde un enfoque bayesiano. Del estudio de simulación se encontró que el procedimiento propuesto genera estimaciones adecuadas para el punto de cambio y todos los demás parámetros del modelo.

***Palabras clave*:** análisis bayesiano, modelos con errores en las variables, modelos con punto de cambio.

1. Introduction

Linear regression is one of the most widely used statistical tools to analyze the relationship between a response variable Y and a covariate x. Under the classic model of simple linear regression the relationship between Y and x is given by

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n \tag{1}$$

where α and β are unknown constants and $e_i \overset{\text{ind}}{\sim} N(0, \sigma_e^2)$, for $i = 1, \ldots, n$, where $N(a, b^2)$ denotes the normal distribution with location parameter a and scale parameter $b > 0$. Usually it is assumed that x_i is measured without error in many practical situations this assumption is violated. Instead of observing x_i is observed

$$X_i = x_i + u_i \quad i = 1, \ldots, n \tag{2}$$

where x_i is the unobservable variable and $u_i \sim N(0, \sigma_u^2)$. Measurements errors (e_i, u_i) are assumed independent and identically distribuited; see, for example, Cheng & Van Ness (1999) and Fuller (1987).

Measurement error (ME) model (also called errors-in-variables model) is a generalization of standard regression models. For the simple linear ME model, the goal is to estimate from bivariate data a straight line fit between X and Y, both of which are measured with error. Applications in which the variables are measured with error are perhaps more common than those in which the variables are measured without error. Many variables in the medical field, such as blood pressure, pulse frequency, temperature, and other blood chemical variables, are measured with error. Variables of agriculture such as rainfalls, content of nitrogen of the soil and degree of infestation of plagues can not be measured accurately. In management sciences, social sciences, and in many other sciences almost all measurable variables are measured with error.

There are three ME models depending on the assumptions about x_i. If the $x_i's$ are unknown constan, then the model is known as a functional ME model; whereas, if the $x_i's$ are independent identically distributed random variables and independent of the errors, the model is known as a structural ME model. A third model, the ultrastructural ME model, assumes that the $x_i's$ are independent random variables but not identically distributed, instead of having possibly different means, μ_i, and common variance σ^2. The ultrastructural model is a generalization of the functional and structural models: if $\mu_1 = \cdots = \mu_n$, then the ultrastructural model reduces to the structural model; whereas if $\sigma^2 = 0$, then the ultrastructural model reduces to the functional model (Cheng & Van Ness 1999).

It is common to assume that all the random variables in the ME model are jointly normal in this case the structural ME model, is not identifiable. This means that different sets of parameters can lead to the same joint distribution of X and Y. For this reason, the statistical literature have considered six assumptions about the parameters which lead to an identifiable structural ME model. The six assumptions have been studied extensively in econometrics; see for example Reiersol (1950), Bowden (1973), Deistler & Seifert (1978) and Aigner, Hsiao, Kapteyn & Wansbeek (1984). They make identifiable the structural ME model.

1. The ratio of the error variances, $\lambda = \sigma_e^2/\sigma_u^2$, is known
2. The ratio $k_x = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ is known
3. σ_u^2 is known
4. σ_e^2 is known
5. The error variances, σ_u^2 and σ_e^2, are known
6. The intercept, α, is known and $E(X) \neq 0$

Assumption 1 is the most popular of these assumptions and is the one with the most published theoretical results; the assumption 2 is commonly found in the social science and psychology literatures; the assumption 3 is a popular assumption when working with nonlinears models; the assumption 4 is less useful and cannot be used to make the equation error model or the measurement error model with more than one explanatory variable identifiable; the assumption 5 frequently leads to the same estimates as those for assumption 1 and also leads to an overidentified model, and finally the assumption 6 does not make the normal model, with more than one identifiable explanatory variable.

In the structural ME model, usually it is assumed that $x_i \sim N(\mu_x, \sigma_x^2)$, $e_i \sim N(0, \sigma_e^2)$ and $u_i \sim N(0, \sigma_u^2)$ with x_i, e_i and u_i independent. A variation of the structural ME model proposed by Chang & Huang (1997) consists in relaxing the assumption of $x_i \sim N(\mu_x, \sigma_x^2)$, so that the $x_i's$ are not identically distributed. Consider an example that can be stated as follows. Let x_i denote some family's true income at time i, let X_i denote the family's measured income, let Y_i denote its measured consumption. During the observations (X_i, Y_i), some new impact on the financial system in the society may occur, for instance, a new economic policy may be announced. The family's true income structure may start to change some time after the announcement; however, the relation between income and consumption remains unchanged. Under this situation Chang & Huang (1997) considered the structural ME model defined by (1) and (2), where the covariate x_i has a change in its distribution given by:

$$\begin{aligned} x_i &\sim N(\mu_1, \sigma_x^2) \qquad i = 1, \ldots, k \\ x_i &\sim N(\mu_2, \sigma_x^2) \qquad i = k+1, \ldots, n \end{aligned}$$

This model with change in the mean of x_i at time k is called structural ME model with changepoint.

The problems with changepoint have been extensively studied. Hinkley (1970) developed a frequentist approach to the changepoint problems and Smith (1975) developed a Bayesian approach. The two works were limited to the inference about the point in a sequence of random variables at which the underlying distribution changes. Carlin, Gelfand & Smith (1992) extended the Smith approach using Markov chain Monte Carlo (MCMC) methods for changepoint with continuos time. Lange, Carlin & Gelfand (1994) and Kiuchi, Hartigan, Holford, Rubinstein & Stevens (1995) used MCMC methods for longitudinal data analysis in AIDS studies. Although there are works in the literature on changepoint problems with Bayesian approach, the Bayesian approach for ME models has not been studied. Hernandez & Usuga (2011) proposed a Bayesian approach for reliability models. The goal of this paper is to propose a Bayesian approach to make inferences in structural ME model with changepoint.

The plan of the paper is as follows. Section 2 presents the Bayesian formulation of the model, Section 3 presents the simulation study and Section 4 presented an application with a real dataset and finally some concluding remarks are presents in Section 5.

2. Structural Errors in Variables Models with Changepoint

The structural ME model with one changepoint that will be studied in this paper is defined by the following equations:

$$\left.\begin{array}{ll} Y_i = \alpha_1 + \beta_1 x_i + e_i & i = 1, \ldots, k \\ Y_i = \alpha_2 + \beta_2 x_i + e_i & i = k+1, \ldots, n \end{array}\right\} \tag{3}$$

and

$$X_i = x_i + u_i \qquad i = 1, \ldots, n\Big\} \tag{4}$$

where X_i and Y_i are observable random variables, x_i is an unobservable random variable, e_i and u_i are random errors with the assumption that $(e_i, u_i, x_i)^T$ are independents for $i = 1, \ldots, n$ with distribution given by:

$$\begin{pmatrix} e_i \\ u_i \\ x_i \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 0 \\ 0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma^2_{e_1} & 0 & 0 \\ 0 & \sigma^2_{u_1} & 0 \\ 0 & 0 & \sigma^2_{x_1} \end{pmatrix} \right), \qquad i = 1, \ldots, k$$

$$\begin{pmatrix} e_i \\ u_i \\ x_i \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 0 \\ 0 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2_{e_2} & 0 & 0 \\ 0 & \sigma^2_{u_2} & 0 \\ 0 & 0 & \sigma^2_{x_2} \end{pmatrix} \right), \qquad i = k+1, \ldots, n$$

The observed data (Y_i, X_i) have the following joint distribution for $i = 1, \ldots, n$.

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \overset{\text{i.i.d}}{\sim} N_2\left(\begin{pmatrix} \alpha_1 + \beta_1\mu_1 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \beta_1^2\sigma_{x_1}^2 + \sigma_{e_1}^2 & \beta_1\sigma_{x_1}^2 \\ \beta_1\sigma_{x_1}^2 & \sigma_{x_1}^2 + \sigma_{u_1}^2 \end{pmatrix}\right), i = 1, \ldots, k$$

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \overset{\text{i.i.d}}{\sim} N_2\left(\begin{pmatrix} \alpha_2 + \beta_2\mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \beta_2^2\sigma_{x_2}^2 + \sigma_{e_2}^2 & \beta_2\sigma_{x_2}^2 \\ \beta_2\sigma_{x_2}^2 & \sigma_{x_2}^2 + \sigma_{u_2}^2 \end{pmatrix}\right), i = k+1, \ldots, n$$

The likelihood function $L(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{\theta} = (k, \alpha_1, \beta_1, \mu_1, \sigma_{x_1}^2, \sigma_{e_1}^2, \sigma_{u_1}^2, \alpha_2, \beta_2, \mu_2, \sigma_{x_2}^2, \sigma_{e_2}^2, \sigma_{u_2}^2)^T$, $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ can be written as:

$$\begin{aligned} L(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y}) \propto\ & (\beta_1^2\sigma_{u_1}^2\sigma_{x_1}^2 + \sigma_{e_1}^2\sigma_{x_1}^2 + \sigma_{u_1}^2\sigma_{e_1}^2)^{-k/2} \exp\left\{-\frac{A}{C}\right\} \\ & \times (\beta_2^2\sigma_{u_2}^2\sigma_{x_2}^2 + \sigma_{e_2}^2\sigma_{x_2}^2 + \sigma_{u_2}^2\sigma_{e_2}^2)^{-(n-k)/2} \exp\left\{-\frac{B}{D}\right\} \end{aligned} \tag{5}$$

where

$$\begin{aligned} A =& (\beta_1^2\sigma_{x_1}^2 + \sigma_{e_1}^2) \sum_{i=1}^{k} (X_i - \mu_1)^2 - 2\beta_1\sigma_{x_1}^2 \sum_{i=1}^{k} (Y_i - \alpha_1 - \beta_1\mu_1)(X_i - \mu_1) \\ & + (\sigma_{x_1}^2 + \sigma_{u_1}^2) \sum_{i=1}^{k} (Y_i - \alpha_1 - \beta_1\mu_1)^2 \\ B =& (\beta_2^2\sigma_{x_2}^2 + \sigma_{e_2}^2) \sum_{i=k+1}^{n} (X_i - \mu_2)^2 - 2\beta_2\sigma_{x_2}^2 \sum_{i=k+1}^{n} (Y_i - \alpha_2 - \beta_2\mu_2)(X_i - \mu_2) \\ & + (\sigma_{x_2}^2 + \sigma_{u_2}^2) \sum_{i=k+1}^{n} (Y_i - \alpha_2 - \beta_2\mu_2)^2 \\ C =& 2(\beta_1^2\sigma_{u_1}^2\sigma_{x_1}^2 + \sigma_{e_1}^2\sigma_{x_1}^2 + \sigma_{u_1}^2\sigma_{e_1}^2) \\ D =& 2(\beta_2^2\sigma_{u_2}^2\sigma_{x_2}^2 + \sigma_{e_2}^2\sigma_{x_2}^2 + \sigma_{u_2}^2\sigma_{e_2}^2) \end{aligned}$$

2.1. Prior and Posterior Distributions

It was considered the discrete uniform distribution for k in the range $1, \ldots, n$ allowing values of $k = 1$ or $k = n$, which would indicate the absence of changepoint. Also, it was considered inverse Gamma distribution for each of the variances and normal distributions for the remaining parameters to obtain posterior distributions. The above distributions with their hyperparameters are given below.

$$p(k) = \begin{cases} P(K = k) = \frac{1}{n}, & k = 1, \ldots, n, \\ 0, & \text{otherwise}, \end{cases}$$

$$\sigma_{e_1}^2 \sim GI(a_{e_1}, b_{e_1}) \quad \sigma_{e_2}^2 \sim GI(a_{e_2}, b_{e_2})$$

$$\begin{aligned}
\sigma^2_{u_1} &\sim GI(a_{u_1}, b_{u_1}) \quad \sigma^2_{u_2} \sim GI(a_{u_2}, b_{u_2}) \\
\sigma^2_{x_1} &\sim GI(a_{x_1}, b_{x_1}) \quad \sigma^2_{x_2} \sim GI(a_{x_2}, b_{x_2}) \\
\alpha_1 &\sim N(\alpha_{01}, \sigma^2_{\alpha_1}) \quad \alpha_2 \sim N(\alpha_{02}, \sigma^2_{\alpha_2}) \\
\beta_1 &\sim N(\beta_{01}, \sigma^2_{\beta_1}) \quad \beta_2 \sim N(\beta_{02}, \sigma^2_{\beta_2}) \\
\mu_1 &\sim N(\mu_{01}, \sigma^2_{\mu_1}) \quad \mu_2 \sim N(\mu_{02}, \sigma^2_{\mu_2})
\end{aligned}$$

where $GI(a, b)$ denotes the inverse Gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$. The hyperparameters a_{e_1}, b_{e_1}, a_{e_2}, b_{e_2}, a_{u_1}, b_{u_1}, a_{u_2}, b_{u_2}, a_{x_1}, b_{x_1}, a_{x_2} y b_{x_2}, α_{01}, $\sigma^2_{\alpha_1}$, α_{02}, $\sigma^2_{\alpha_2}$, β_{01}, $\sigma^2_{\beta_1}$, β_{02}, $\sigma^2_{\beta_2}$, μ_{01}, $\sigma^2_{\mu_1}$, μ_{02} and $\sigma^2_{\mu_2}$ are considered as known. The prior distribution for the vector $\boldsymbol{x}$ is denoted by $\pi(\boldsymbol{x})$ and it is based on the assumption of independence and normality of the model.

The likelihood function based on complete data $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is denoted by $L^*(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y})$ and can be expressed as

$$L^*(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y}) \propto \prod_{i=1}^{k} (\sigma^2_{e_1} \sigma^2_{u_1} \sigma^2_{x_1})^{-\frac{1}{2}} e^E \prod_{i=k+1}^{n} (\sigma^2_{e_2} \sigma^2_{u_2} \sigma^2_{x_2})^{-\frac{1}{2}} e^F \tag{6}$$

where

$$\begin{aligned}
E &= -\frac{(Y_i - \alpha_1 - \beta_1 x_i)^2}{2\sigma^2_{e_1}} - \frac{(X_i - x_i)^2}{2\sigma^2_{u_1}} - \frac{(x_i - \mu_1)^2}{2\sigma^2_{x_1}} \\
F &= -\frac{(Y_i - \alpha_2 - \beta_2 x_i)^2}{2\sigma^2_{e_2}} - \frac{(X_i - x_i)^2}{2\sigma^2_{u_2}} - \frac{(x_i - \mu_2)^2}{2\sigma^2_{x_2}}
\end{aligned}$$

Based on the prior distributions for each parameter the posterior distribution for $\boldsymbol{\theta}$ can be written as

$$\begin{aligned}
\pi(\boldsymbol{\theta}, \boldsymbol{x} \mid \boldsymbol{X}, \boldsymbol{Y}) \propto & \prod_{i=1}^{k} (\sigma^2_{e_1} \sigma^2_{u_1})^{-\frac{1}{2}} e^G \prod_{i=k+1}^{n} (\sigma^2_{e_2} \sigma^2_{u_2})^{-\frac{1}{2}} e^H \, p(k) \\
& \times \pi(\alpha_1)\, \pi(\alpha_2)\, \pi(\beta_1)\, \pi(\beta_2)\, \pi(\mu_1)\, \pi(\mu_2) \\
& \times \pi(\sigma^2_{e_1})\, \pi(\sigma^2_{e_2})\, \pi(\sigma^2_{u_1})\, \pi(\sigma^2_{u_2})\, \pi(\sigma^2_{x_1})\, \pi(\sigma^2_{x_2})\, \pi(\boldsymbol{x})
\end{aligned} \tag{7}$$

where

$$\begin{aligned}
G &= -\frac{(Y_i - \alpha_1 - \beta_1 x_i)^2}{2\sigma^2_{e_1}} - \frac{(X_i - x_i)^2}{2\sigma^2_{u_1}} \\
H &= -\frac{(Y_i - \alpha_2 - \beta_2 x_i)^2}{2\sigma^2_{e_2}} - \frac{(X_i - x_i)^2}{2\sigma^2_{u_2}}
\end{aligned}$$

The conditional posterior distributions for the parameters obtained from the previous posterior distribution are given in Appendix. For almost all the parameters posterior distributions with pdf known were obtained, except for the parameter k. The conditional posterior distribution of the changepoint k in the model

has not pdf known, making it necessary to use the Gibbs sampler, introduced by Geman & Geman (1984) to approximate this distribution. The sampler Gibbs is an iterative algorithm that constructs a dependent sequence of parameter values whose distribution converges to the target joint posterior distribution (Hoff 2009).

The procedure used to implement the Gibbs sampler to the problem was:

1. Generate appropriate initial values for each of the 13 parameters to create the initial parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{13})^T$.

2. Update the component $j = 1, \ldots, 13$ of $\boldsymbol{\theta}$ generating a random observation for the parameter θ_j using the corresponding posterior distribution of Appendix and the subset of parameters of $\boldsymbol{\theta}$ present in the posterior distribution of θ_j.

3. Repeat step 2 a number of times until obtaining convergence in all the parameters.

3. Simulation Study

In this section we present the results of implementation of the Gibbs sampler for the model given in equations (3) and (4) under three different assumptions of the parameters. In the first case we analyze the model with a simulated dataset considering $\lambda = \sigma^2_{e_i}/\sigma^2_{u_i}$ known; in the second case we consider the variances $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ known and equal, and in the third case we consider the variances $\sigma^2_{e_1}$ and $\sigma^2_{e_2}$ known and equals. In addition to the above cases we also analized the changepoint estimate of the model for different n values with the aim of observing the behavior of the estimate of k with respect to its true value.

3.1. λ Known

In Table 1 we present a dataset of $n = 60$ observations generated in R Development Core Team (2011) from the model given in equations (3) and (4) with the assumption of $\lambda = 1$ considering the following set of parameters: $k = 20$, $\alpha_1 = 2$, $\beta_1 = 2$, $\mu_1 = 1$, $\sigma^2_{x_1} = 1$, $\sigma^2_{e_1} = 1.5$, $\sigma^2_{u_1} = \sigma^2_{e_1}/\lambda$, $\alpha_2 = -1$, $\beta_2 = 4$, $\mu_2 = 5$, $\sigma^2_{x_2} = 2$, $\sigma^2_{e_2} = 2.5$ and $\sigma^2_{u_2} = \sigma^2_{e_2}/\lambda$. Figure 1 shows the scatter plot for the data generated and there is not clear indication of the changepoint in the model structure.

We use the Gibbs sampler to obtain estimates of the parameters. The prior distributions used to run the Gibbs sampler were as follows: $\alpha_1 \sim N(2, 15)$, $\beta_1 \sim N(2, 15)$, $\mu_1 \sim N(1, 15)$, $\sigma^2_{x_1} \sim GI(2, 5)$, $\sigma^2_{e_1} \sim GI(2, 5)$, $\sigma^2_{u_1} \sim GI(2, 5)$, $\alpha_2 \sim N(-1, 15)$, $\beta_2 \sim N(4, 15)$, $\mu_2 \sim N(5, 15)$, $\sigma^2_{x_2} \sim GI(2, 5)$, $\sigma^2_{e_2} \sim GI(2, 5)$ and $\sigma^2_{u_2} \sim GI(2, 5)$.

TABLE 1: Random sample of simulated data with $\lambda = 1$.

X	0.05	4.34	1.18	0.65	1.47	−0.57	2.42	0.78	1.63	−1.02	−0.78	0.48
Y	0.03	4.92	4.42	1.62	5.91	3.81	6.60	4.97	2.79	2.37	1.45	5.24
X	1.05	2.50	3.76	0.61	1.46	0.98	1.59	−0.95	4.89	2.28	7.09	7.18
Y	1.74	6.07	6.96	6.65	2.08	2.63	5.55	1.39	17.84	9.15	16.82	25.40
X	6.58	4.85	6.23	5.30	7.29	6.73	6.78	7.46	2.86	3.33	3.80	9.66
Y	18.51	18.86	13.12	19.57	17.16	22.71	22.16	26.90	21.30	12.82	27.43	27.96
X	3.63	3.66	5.70	5.64	2.15	3.13	9.10	9.88	4.73	7.48	2.55	11.11
Y	9.53	12.68	18.54	19.15	20.43	21.36	32.97	27.38	13.67	18.73	14.54	23.11
X	7.10	4.95	9.17	2.03	9.54	5.08	7.36	6.37	4.35	1.45	7.67	2.97
Y	25.54	21.00	27.77	9.75	27.62	22.29	19.45	17.34	23.68	13.99	25.15	11.38

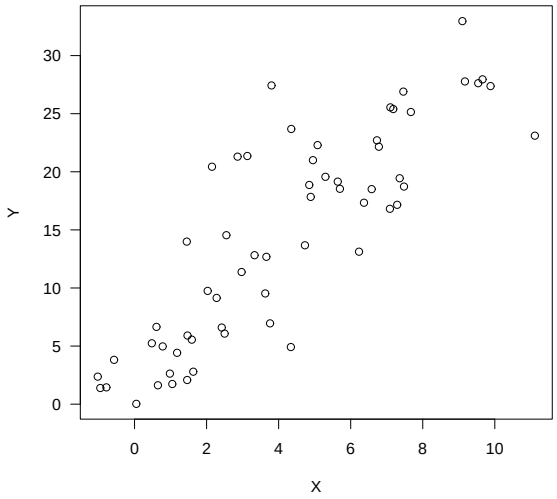

FIGURE 1: Scatter plot for simulated data with $\lambda = 1$.

We ran five chains of the Gibbs sampler. Each sequence was run for 11000 iterations with a burn-in of 1000 samples. The vectors of initial values for each of the chains were:

$$\boldsymbol{\theta}_1^{(0)} = (5, 1.886, 1.827, 2.4, 0.942, 1.134, 1.015, -1.5, 2.100, 1.3, 0.6, 0.893, 1.8)$$
$$\boldsymbol{\theta}_2^{(0)} = (10, 2.537, 1.225, 2.2, 1.404, 2.171, 0.552, 0.2, 3.500, 4.3, 1.1, 0.903, 3.4)$$
$$\boldsymbol{\theta}_3^{(0)} = (30, 1.856, 1.855, 2.6, 0.928, 1.087, 1.029, -0.3, 3.829, 4.5, 2.0, 0.900, 2.1)$$
$$\boldsymbol{\theta}_4^{(0)} = (40, 2.518, 1.242, 2.8, 1.386, 2.142, 0.571, -2.0, 3.829, 3.5, 2.8, 0.901, 1.4)$$
$$\boldsymbol{\theta}_5^{(0)} = (50, 2.516, 1.244, 1.8, 1.383, 2.138, 0.573, -1.3, 3.829, 2.5, 3.5, 0.899, 2.4)$$

The above vectors were obtained by the following procedure. For fixed values of $k = 5, 10, 30, 40, 50$ numerical methods were used to determine the values of $\boldsymbol{\theta}$ that maximize the likelihood function given in (5). These estimates were obtained using the function `optim` of R Development Core Team (2011), which uses optimization

methods quasi-Newton such as the bounded limited-memory algorithm L-BFGS-B (Limited memory, Broyden- Fletcher-Goldfarb-Shanno, Bounded) proposed by Byrd, Lu, Nocedal & Zhu (1995).

In order to verify the convergence of the chains we use the diagnostic indicator proposed by Brooks & Gelman (1998). The diagnostic value of R found in this case was 1.04; values close to 1 indicate convergence of the chains. Additionally, for each parameter, the posterior distribution was examined visually by monitoring the density estimates, the sample traces, and the autocorrelation function. We found not evidence of trends or high correlations. Figures 2 and 3 show the Highest Density Region (HDR) graphics for the parameters of the chain 1. These graphics show that the true values of the model parameters are in the Highest Density Regions.

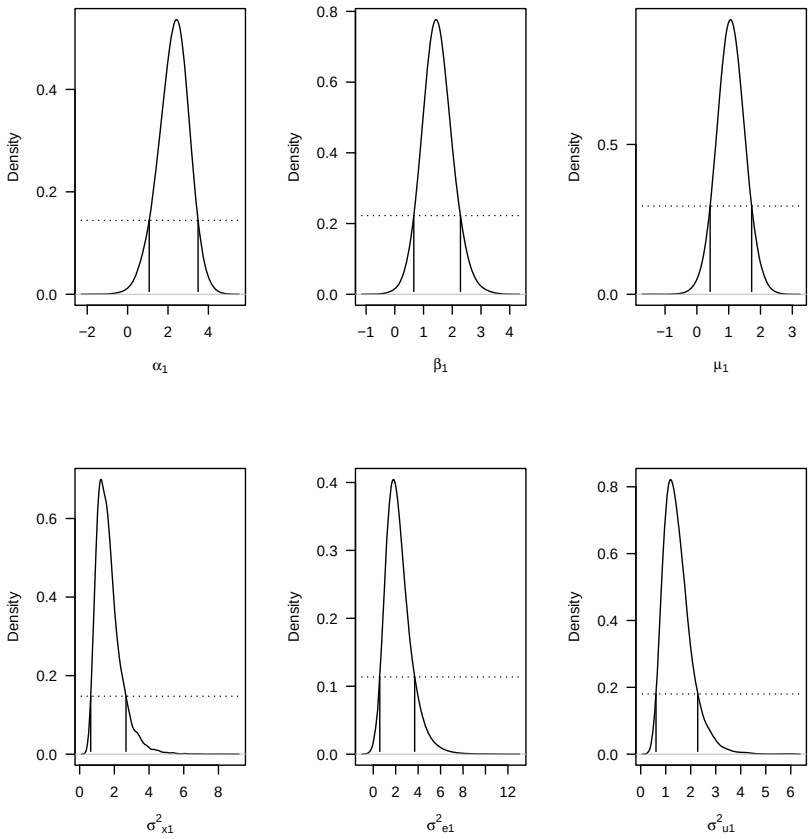

FIGURE 2: HDR plot for α_1, β_1, μ_1, $\sigma^2_{x_1}$,$\sigma^2_{e_1}$ and $\sigma^2_{u_1}$.

Table 2 presents the posterior mean and standard deviation (SD) for the model parameters and the 90% HDR interval. Note that the true values parameters are close to mean and are within the HDR interval.

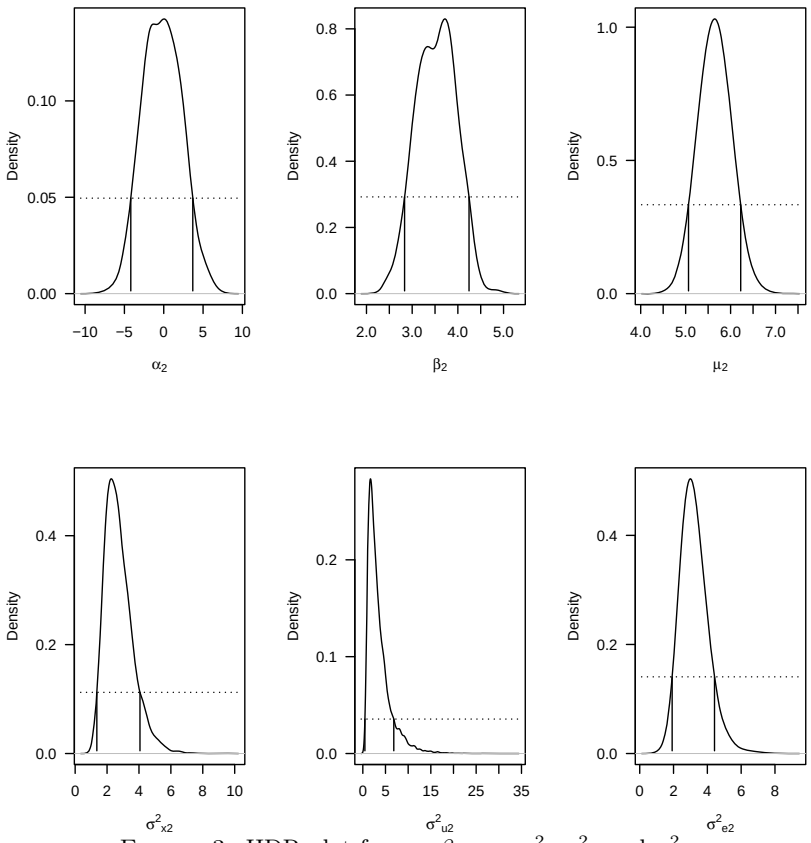

FIGURE 3: HDR plot for α_2, β_2, μ_2, $\sigma^2_{x_2}$,$\sigma^2_{e_2}$ and $\sigma^2_{u_2}$.

TABLE 2: Posterior mean, standard deviation (SD), HDRlower and HDRupper of parameters with $\lambda = 1$.

Parameter	Mean	SD	HDRlower	HDRupper
k	19.99	0.10	-	-
α_1	2.21	0.83	0.94	3.57
β_1	1.50	0.54	0.64	2.38
μ_1	1.09	0.40	0.44	1.74
$\sigma^2_{x_1}$	1.64	0.72	0.60	2.61
$\sigma^2_{e_1}$	2.28	1.14	0.62	3.77
$\sigma^2_{u_1}$	1.47	0.59	0.56	2.24
α_2	0.17	2.54	−3.97	4.46
β_2	3.44	0.45	2.71	4.20
μ_2	5.72	0.38	4.10	5.34
$\sigma^2_{x_2}$	2.86	0.94	1.39	4.22
$\sigma^2_{e_2}$	3.77	2.68	0.53	7.32
$\sigma^2_{u_2}$	3.18	0.81	1.86	4.42

3.2. $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ Known

In this case we consider the structural ME model with $\sigma^2_{u_1} = \sigma^2_{u_2} = 2$. Table 3 shows a dataset of size $n = 60$ generated from the model given in equations (3) and (4) with the following set of parameters: $\alpha_1 = 2$, $\beta_1 = 2$, $\mu_1 = 1$, $\sigma^2_{x_1} = 1$, $\sigma^2_{e_1} = 1.5$, $\alpha_2 = -1$, $\beta_2 = 4$, $\mu_2 = 5$, $\sigma^2_{x_2} = 2$ and $\sigma^2_{e_2} = 2.5$. Figure 4 shows the scatter plot for the simulated data.

TABLE 3: Random sample of data simulated with $\sigma^2_{u_1} = \sigma^2_{u_2} = 2$.

X	2.03	2.16	1.68	−0.07	1.00	−0.82	1.42	−0.42	3.36	0.88	0.12	0.80
Y	5.11	4.31	5.33	2.73	0.33	1.69	7.48	3.06	2.65	0.48	1.82	5.37
X	2.84	4.15	−1.54	0.84	1.55	0.99	−0.27	4.16	6.13	5.01	5.09	1.12
Y	0.42	5.20	3.75	4.88	3.87	0.73	6.01	8.41	20.03	18.82	15.29	8.10
X	5.40	3.28	8.06	5.78	5.68	3.26	2.48	3.72	2.85	6.13	2.85	8.47
Y	8.55	16.47	26.13	15.54	16.11	14.48	11.52	21.86	9.55	24.49	14.44	24.76
X	3.18	3.90	2.58	7.58	5.59	6.79	7.20	4.01	6.10	5.73	1.82	7.95
Y	20.74	15.84	4.54	20.84	20.96	24.59	23.95	11.74	18.99	15.13	9.98	29.01
X	4.42	4.01	7.72	9.25	4.60	4.73	0.52	0.46	2.76	5.44	7.22	3.33
Y	13.82	14.92	23.08	32.71	10.53	22.03	11.28	14.74	8.30	15.60	30.96	17.54

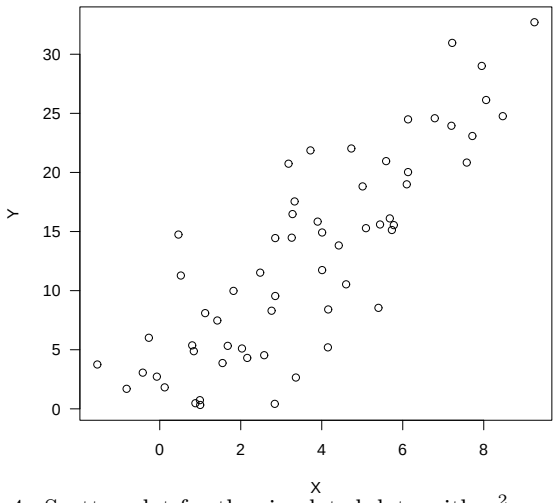

FIGURE 4: Scatter plot for the simulated data with $\sigma^2_{u_1} = \sigma^2_{u_2} = 2$.

The prior distributions for $\alpha_1, \beta_1, \mu_1, \sigma^2_{x_1}, \sigma^2_{e_1}, \alpha_2, \beta_2, \mu_2, \sigma^2_{x_2}$, and $\sigma^2_{e_2}$ were the same considered in the case of λ known. The vectors of initial values for the model parameters in each of the five Markov chain were as follows:

$$\boldsymbol{\theta}_1^{(0)} = (5, 1.0, 2.305, 1.063, 0.189, 3.0, 2, -2.0, 4.071, 4.727, 2.0, 1.984, 2)$$
$$\boldsymbol{\theta}_2^{(0)} = (10, 1.5, 2.800, 1.063, 0.189, 1.7, 2, -2.0, 4.071, 3.300, 3.0, 1.980, 2)$$
$$\boldsymbol{\theta}_3^{(0)} = (30, 3.1, 2.305, 0.300, 2.000, 2.0, 2, -0.5, 1.500, 3.200, 2.0, 2.700, 2)$$
$$\boldsymbol{\theta}_4^{(0)} = (40, 3.1, 1.900, 1.063, 1.400, 2.1, 2, -2.0, 4.071, 4.727, 1.4, 1.981, 2)$$
$$\boldsymbol{\theta}_5^{(0)} = (50, 2.9, 0.500, 1.063, 0.189, 2.6, 2, 0.7, 2.400, 1.500, 3.1, 1.981, 2).$$

The diagnostic value of convergence R was 1.01 indicating the convergence of the chains. A visual monitoring of the density estimates, the sample traces, and the correlation function for each parameter in each of the chains did not show any problem. In Figures 5 and 6 we present the HDR graphics for the parameters in the chain 1. The graphics show that true values of the parameters model are within the Highest Density Regions.

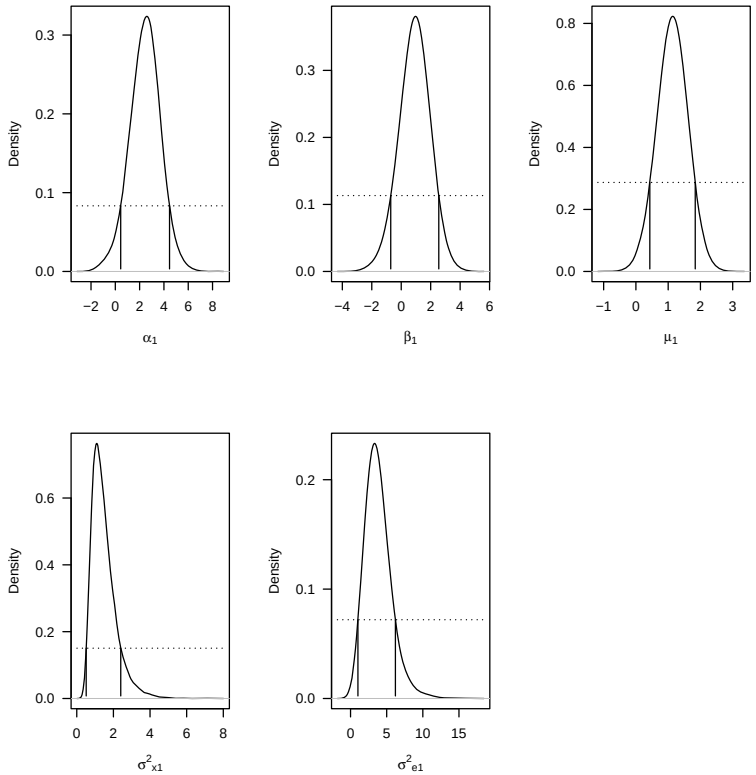

FIGURE 5: HDR plot for α_1, β_1, μ_1, $\sigma^2_{x_1}$ and $\sigma^2_{e_1}$.

Table 4 shows the posterior mean and standard deviation (SD) for each of the parameters model and the 90% HDR interval. It is noted again that the true values of the parameters are close to the mean and within the HDR intervals.

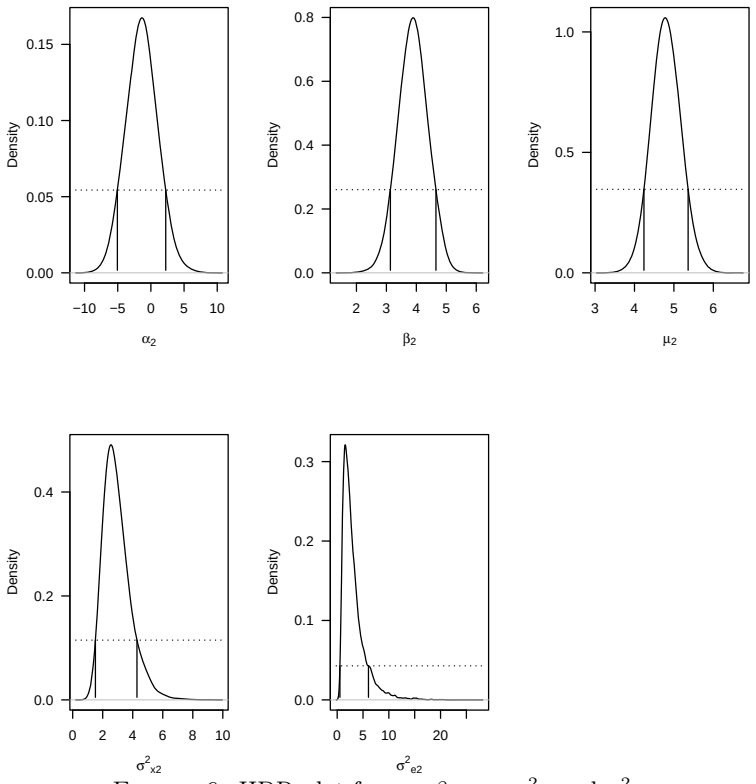

FIGURE 6: HDR plot for α_2, β_2, μ_2, $\sigma^2_{x_2}$ and $\sigma^2_{e_2}$.

TABLE 4: Posterior mean, standard deviation (SD), HDRlower and HDRupper of parameters with $\sigma^2_{u_1} = \sigma^2_{u_2}$.

Parameter	Mean	SD	HDRlower	HDRupper
k	19.25	0.61	-	-
α_1	2.48	1.24	0.48	4.47
β_1	0.89	0.99	−0.69	2.46
μ_1	1.15	0.43	0.44	1.86
$\sigma^2_{x_1}$	1.51	0.73	0.52	2.47
$\sigma^2_{e_1}$	3.85	1.72	1.05	6.18
α_2	−1.00	2.31	−4.62	2.91
β_2	3.82	0.46	3.04	4.55
μ_2	4.79	0.36	4.20	5.36
$\sigma^2_{x_2}$	3.02	0.94	1.53	4.39
$\sigma^2_{e_2}$	3.21	2.16	0.53	6.07

3.3. $\sigma^2_{e_1}$ and $\sigma^2_{e_2}$ Known

In this case we consider the structural ME model with $\sigma^2_{e_1} = \sigma^2_{e_2} = 2$. Table 5 shows a dataset of size $n = 60$ generated from the model given in equations (3) and (4) considering the same parameters values of the case λ known. Figure 7 presents the scatter plot of the simulated data.

TABLE 5: Random sample of the simulated data with $\sigma^2_{e_1} = \sigma^2_{e_2} = 2$.

X	−0.82	1.06	−2.25	0.76	1.71	−0.93	0.50	−0.08	1.12	−0.14	1.59	0.99
Y	2.72	2.75	1.88	8.66	5.12	2.51	7.00	7.64	4.55	7.50	2.89	2.55
X	−1.69	−0.14	2.32	−1.42	2.99	0.70	0.99	0.06	5.20	2.18	6.69	8.79
Y	2.47	0.43	5.32	3.10	6.63	1.99	5.57	3.39	11.10	13.80	17.71	32.28
X	4.98	8.05	6.29	5.99	5.32	5.80	7.73	5.45	4.27	2.84	7.69	10.61
Y	16.11	19.64	16.59	17.86	16.14	19.01	34.01	19.65	19.77	21.84	23.58	30.02
X	5.90	3.51	2.15	3.10	9.42	3.31	4.06	1.44	7.18	1.72	6.61	5.28
Y	21.45	11.51	17.44	17.75	29.13	18.25	20.95	8.24	24.11	8.12	29.31	19.54
X	4.68	1.88	5.02	1.76	7.67	5.31	6.42	7.79	4.32	1.20	3.22	3.37
Y	17.90	17.81	16.27	14.20	23.72	27.51	28.17	18.41	18.28	11.89	15.10	19.96

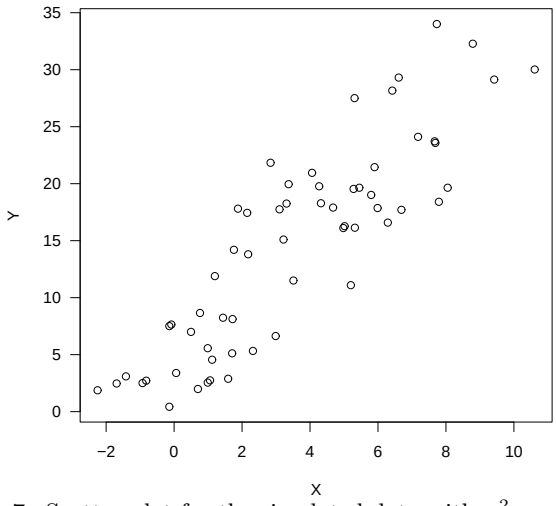

FIGURE 7: Scatter plot for the simulated data with $\sigma^2_{e_1} = \sigma^2_{e_2} = 2$.

The prior distributions for $\alpha_1, \beta_1, \mu_1, \sigma^2_{x_1}, \sigma^2_{u_1}, \alpha_2, \beta_2, \mu_2, \sigma^2_{x_2}$ and $\sigma^2_{u_2}$ were the same considered in the case of λ known. The vectors of the initial values for each of the five Markov chains were as follows:

$$\boldsymbol{\theta}_1^{(0)} = (5, 3.264, 2.650, 0.366, 0.437, 2.001, 1.276, 4.287, 3.0, 5.106, 3.793, 2.001, 2.394)$$

$$\boldsymbol{\theta}_2^{(0)} = (10, 1.000, 1.904, 1.500, 1.370, 2.001, 0.500, 4.512, 5.0, 2.000, 2.700, 2.001, 1.000)$$

$$\boldsymbol{\theta}_3^{(0)} = (30, 2.500, 2.650, 1.000, 0.437, 2.001, 1.500, 4.286, 2.0, 4.000, 3.792, 2.001, 0.700)$$

$$\boldsymbol{\theta}_4^{(0)} = (40, 0.500, 1.904, 0.900, 1.369, 2.000, 2.000, 4.512, 1.8, 5.500, 4.000, 2.001, 2.100)$$

$$\boldsymbol{\theta}_5^{(0)} = (50, 3.600, 2.652, 1.900, 0.437, 2.000, 0.700, 4.287, 4.1, 4.200, 3.792, 2.001, 2.000)$$

The diagnostic value of convergence was of 1.04 indicating the convergence of the chains. In Figures 8 and 9 we present the HDR graphics of the parameters for the chain 1. Note that the true values of the parameters model are within the Highest Density Regions.

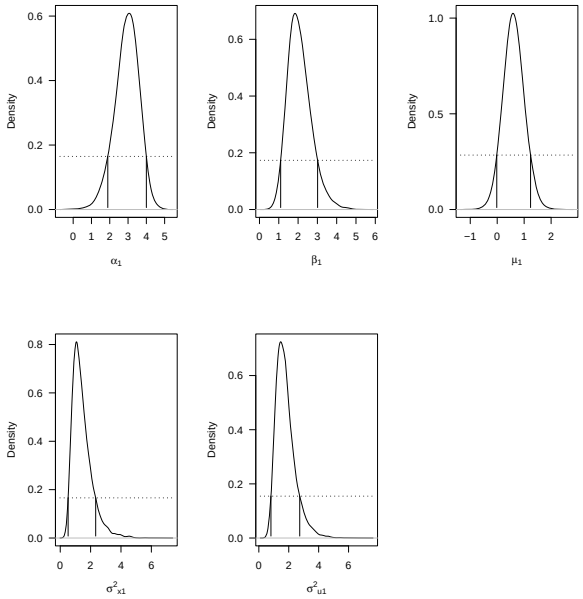

FIGURE 8: HDR plot for α_1, β_1, μ_1, $\sigma^2_{x_1}$ and $\sigma^2_{u_1}$.

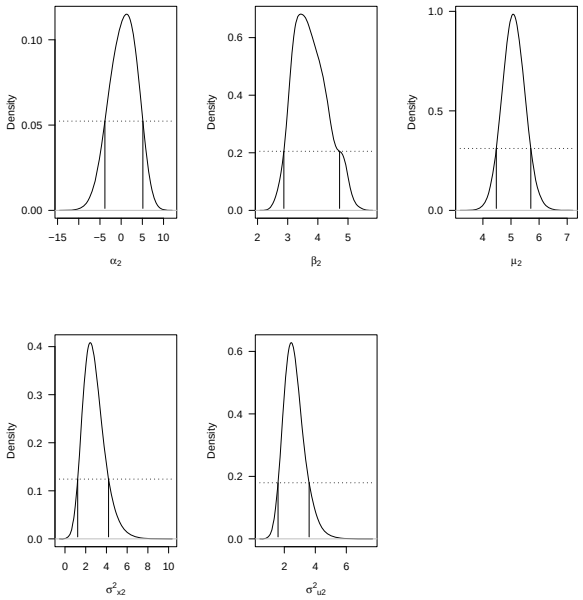

FIGURE 9: HDR plot for α_2, β_2, μ_2, $\sigma^2_{x_2}$ and $\sigma^2_{u_2}$.

Table 6 shows the posterior mean and the standard deviation (SD) for the model parameters and the 90% HDR interval. As in the previous cases the posterior means are close to the true values and are within the HDR interval.

Table 6: Posterior mean, standard deviation (SD), HDRlower and HDRupper of parameters with $\sigma^2_{e_1} = \sigma^2_{e_2}$.

Parameter	Mean	SD	HDRlower	HDRupper
k	20.22	0.47	-	-
α_1	2.95	0.62	0.96	2.98
β_1	2.09	0.61	1.09	3.02
μ_1	0.59	0.37	−0.01	1.20
$\sigma^2_{x_1}$	1.41	0.64	0.53	2.26
$\sigma^2_{u_1}$	1.76	0.64	0.80	2.64
α_2	0.60	2.62	−2.01	0.97
β_2	3.70	0.49	2.93	4.49
μ_2	5.16	0.35	4.57	5.73
$\sigma^2_{x_2}$	2.85	0.93	1.41	4.19
$\sigma^2_{u_2}$	2.62	0.63	1.59	3.53

3.4. Constant Sample Size and Variable Changepoint

In this case our objective was determine if the estimated changepoint of the model given in equations (3) and (4) differs from its true value when $n = 60$ is fixed. We generated nine random samples of size $n = 60$ based on the structure considered in Section 3.1; the values of the parameters were the same ones used in this section. The changepoint k for each nine random samples had different values, and the values were $k = 3, 5, 10, 20, 30, 40, 50, 55$ and 58. For each of the random samples were run five Markov chains of size 150000 with a burn in of 15000. Table 7 presents for the estimated changepoint k the posterior mean, standard deviation and percentiles of 10% and 90% when $n = 60$. Note that posterior mean of the changepoint is very close to the true value and the standard deviation tends to increase as the changepoint approach to the extreme values. Also the Table 7 shows that the distance between the percentiles 10% and 90% is at most 1%, which indicates that the posterior distribution for the parameter k is highly concentrated in one or two possible values and they match with the true value of k.

3.5. Sample Size and Changepoint Variable

In this case our objective was to determine if the estimated changepoint of the model given in equations (3) and (4) differs from its true value for differents values of n. As in the previous case we generated nine dataset with the structure of the Section 3.1. Each of the dataset had samples sizes of $n = 20, 30, 40, 50, 60, 70, 80, 90$ and 100. The true value for k in each of the nine set was $k = n/2$. Table 8 presents the posterior mean, standard deviation and 10% and 90% percentiles for the estimated changepoint and the true values of k. Again we see that the posterior mean of k is very close to the true values of k; it is also noted that the standard deviation tends to increase as the size sample n decreases; this means that if we have fewer observations the posterior distribution for k tends to have

greater variability. As in the previous case the distance between the percentiles 10% and 90% is at most 1%, which means that the posterior distribution for the parameter k is highly concentrated in one or two possible values and they match the true value of k.

TABLE 7: Posterior mean, standard deviation (SD) and 10% and 90% percentiles of k estimated when $n = 60$.

k	Mean	SD	10%	90%
3	3.16	0.67	3.00	4.00
5	5.03	0.43	5.00	5.00
10	9.95	0.30	10.00	10.00
20	19.89	0.31	19.00	20.00
30	29.97	0.19	30.00	30.00
40	39.99	0.12	40.00	40.00
50	49.98	0.13	50.00	50.00
55	54.98	0.14	55.00	55.00
58	57.97	0.18	58.00	58.00

TABLE 8: Posterior mean, standard deviation (SD) and 10% and 90% percentiles of k.

n	k	Mean	SD	10%	90%
20	10	9.94	0.29	10.00	11.00
30	15	14.96	0.27	15.00	16.00
40	20	19.98	0.21	19.00	20.00
50	25	24.98	0.16	24.00	24.00
60	30	30.17	0.12	30.00	30.00
70	35	34.99	0.10	35.00	35.00
80	40	39.99	0.10	39.00	39.00
90	45	44.99	0.11	44.00	45.00
100	50	50.00	0.06	49.00	49.00

4. Application

This section illustrates the proposed procedure for the structural ME model with changepoint using a dataset of imports in the French economy.

Malinvaud (1968) provided the data of imports, gross domestic product (GDP), and other variables in France from 1949-1966. The main interest is to forecast the imports given the gross domestic product of the country. Chatterjee & Brockwell (1991) analyzed these data by the principal component method and found two patterns in the data; they argued that the models before and after 1960 must be different due to the fact the European Common Market began operations in 1960. Maddala (1992) considered a functional ME model; however, he ignored

the possibility that some changes in the data may arise. Chang & Huang (1997) considered a structural ME model with changepoint using the likelihood ratio test based on the maximum Hotelling T^2 for the test of no change against the alternative of exactly one change and concluded that the changepoint ocurred in 1962. Table 9 presents the import data (Y) and gross domestic product (X).

Table 9: Imports and gross domestic product data from January 1949 to November 1966.

Year	1949	1950	1951	1952	1953	1954
GDP	149.30	161.20	171.50	175.50	180.80	190.70
Imports	15.90	16.40	19.00	19.10	18.80	20.40
Year	1955	1956	1957	1958	1959	1960
GDP	202.10	212.40	226.10	231.90	239.00	258.00
Imports	22.70	26.50	28.10	27.60	26.30	31.10
Year	1961	1962	1963	1964	1965	1966
GDP	269.80	288.40	304.50	323.40	336.80	353.90
Imports	33.30	37.00	43.30	49.00	50.30	56.60

The data were reanalized under a Bayesian perspective by adopting the structural ME model with changepoint. We considered non informative prior distributions for all parameters. Again as in the previous cases, we built five chains with different initial values of size 11000 with a burn in of 1000 samples to avoid correlations problems. We found the value $R = 1.03$, indicating the convergence of the chains.

Figure 10 shows the high concentration in the value 14 for the posterior distribution for the parameter k. The mean for this distribution is 13.92, which is the same obtained by Chang & Huang (1997), indicating that the data present a changepoint for the year 1962. Table 10 presents estimates for the remaining parameters of the model. The values are also close to the results obtained by Chang & Huang (1997). It is also noted that the means for β_1 and β_2 were 0.14 and 0.16, which indicates no significant changes in the slope for the trend lines before and after $k = 14$. The means obtained for the parameters α_1 and α_2 were -5.53 and -2.23, not being to close these values, which indicate that the trend lines before and after the change have different changepoints; this can be seen clearly in the Figure 11.

5. Conclusions

This paper proposes the Bayesian approach to study the structural ME model with changepoint. Through the simulation study was shown that the proposed procedure identifies correctly the point where the change comes to structure; note also that the variability in the posterior distribution of k decreases as the number of observations in the dataset increases. Another important aspect is that the variability of the posterior distribution for k increases as the true value of k is

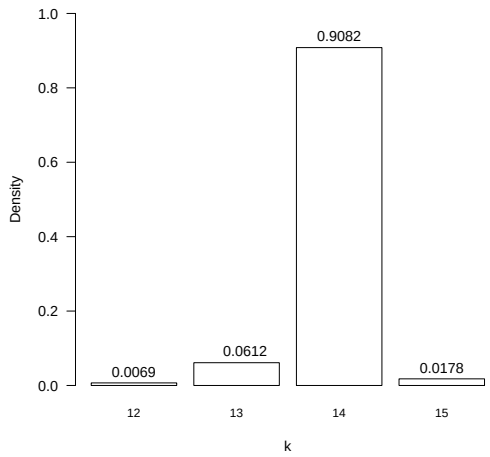

FIGURE 10: Posterior density for k.

TABLE 10: Posterior summary results.

Parameter	Mean	SD	HDRlower	HDRupper
α_1	-5.53	1.95	8.77	-2.50
β_1	0.14	0.01	0.13	0.16
μ_1	16.33	3.88	9.99	22.67
$\sigma^2_{x_1}$	34601.81	13191.57	13641.42	51588.36
$\sigma^2_{e_1}$	1.93	0.89	0.72	3.07
$\sigma^2_{u_1}$	4.37	5.18	0.37	8.66
α_2	-2.23	3.78	-8.35	3.96
β_2	0.16	0.01	0.14	0.18
μ_2	5.39	3.89	-1.02	11.82
$\sigma^2_{x_2}$	70266.22	48727.80	1141.85	120582.77
$\sigma^2_{e_2}$	5.43	4.61	0.72	10.16
$\sigma^2_{u_2}$	5.14	13.70	-	-

close to 1. For the other parameters the proposed procedure generated posterior distributions with means very close to the real parameters in all cases considered. The proposed procedure generates chains that converge to the true parameters, regardless of whether or not identifiability assumptions.

Possible future works could consider other prior distributions such as non informative and skew normal and also introduce multiple changepoints in Y and X.

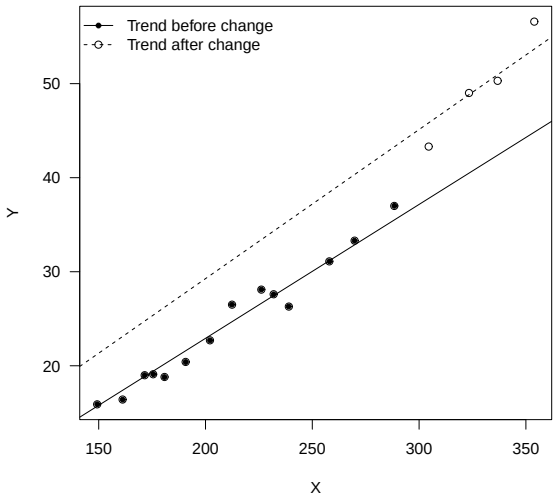

FIGURE 11: Scatter plot for the application.

6. Acknowledgements

The authors are grateful to the editor and referees for their valuable comments and helpful suggestions which led to improvements in the paper, and the professor Carlos Alberto de Bragança Pereira (IME-USP) for the motivation to write this work.

During the course of this work the authors received financial support from Coordenação de aperfeiçoamento de pessoal de nível superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Universidad de Antioquia.

[Recibido: octubre de 2010 — Aceptado: enero de 2011]

References

Aigner, D. J., Hsiao, C., Kapteyn, A. & Wansbeek, T. J. (1984), Latent variable models in econometrics, *in* Z. Griliches & M. D. Intriligator, eds, 'Handbook of Econometrics', Vol. 2, Elsevier Science, Amsterdam, pp. 1321–1393.

Bowden, R. J. (1973), 'The theory of parametric identification', *Econometrica* **41**, 1069–1174.

Brooks, S. P. & Gelman, A. (1998), 'General methods for monitoring convergence of iterative simulations', *Journal of Computational and Graphical Statistics* **7**(4), 434–455.

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995), 'A limited memory algorithm for bound constrained optimization', *SIAM Journal on Scientific Computing* **16**, 1190–1208.

Carlin, B. P., Gelfand, A. E. & Smith, A. F. M. (1992), 'Hierarchical Bayesian analysis of changepoint problems', *Applied of Statistics* **41**, 389–405.

Chang, Y. P. & Huang, W. T. (1997), 'Inferences for the linear errors-in-variables with changepoint models', *Journal of the American Statistical Association* **92**(437), 171–178.

Chatterjee, E. & Brockwell, P. J. (1991), *Regression Analysis by Example*, Wiley, New York.

Cheng, C. L. & Van Ness, J. W. (1999), *Statistical Regression with Measurement Error*, Arnold, London.

Deistler, M. & Seifert, H. G. (1978), 'Identifiability and consistent estimability in econometrics models', *Econometrica* **46**, 969–980.

Fuller, W. A. (1987), *Measurement Error Models*, Wiley, New York.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Hernandez, F. & Usuga, O. C. (2011), 'Analisis bayesiano para la distribucion log-normal generalizada aplicada a modelos de falla con censura', *Revista Colombiana de Estadística* **34**(1), 95–100.

Hinkley, D. V. (1970), 'Inference about the change-point in a sequence of random variables', *Biometrika* **57**, 1–17.

Hoff, P. (2009), *A First Course in Bayesian Statistical Methods*, Springer, New York.

Kiuchi, A. S., Hartigan, J. A., Holford, T. R., Rubinstein, P. & Stevens, C. E. (1995), 'Change points in the series of T4 counts prior to AIDS', *Biometrics* **51**, 236–248.

Lange, N., Carlin, B. & Gelfand, A. E. (1994), 'Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion)', *Journal of the American Statistical Association* **87**, 615–632.

Maddala, G. S. (1992), *Introduction to Econometrics*, Macmillan, New York.

Malinvaud, E. (1968), *Statistical Methods of Econometrics*, Rand-McNally, Chicago.

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

Reiersol, O. (1950), 'Identifiability of a linear relation between variables which are subject to errors', *Econometrica* **18**, 375–389.

Smith, A. F. M. (1975), ' A Bayesian approach to inference about a change-point in a sequence of random variables', *Biometrika* **62**, 407–416.

Appendix. Conditional Posterior Distributions

1. Conditional posterior distribution of k

$$P(K = k \mid \boldsymbol{\theta}_{-k}, \boldsymbol{x}, \boldsymbol{X}, \boldsymbol{Y}) = \frac{L^*(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y})}{\sum_{k=1}^{n} L^*(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y})}$$

   where $L^*(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y})$ is given in equation (6).

2. Conditional posterior distribution of α_1

$$\pi\left(\alpha_1 \mid \boldsymbol{\theta}_{\{-\alpha_1\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim N\left(\frac{\sigma^2_{\alpha_1} \sum_{i=1}^{k} (Y_i - \beta_1 x_i) + \alpha_{01}\sigma^2_{e_1}}{k\sigma^2_{\alpha_1} + \sigma^2_{e_1}}, \frac{\sigma^2_{e_1}\sigma^2_{\alpha_1}}{k\sigma^2_{\alpha_1} + \sigma^2_{e_1}}\right)$$

   where $\boldsymbol{\theta}_{\{-\theta_i\}}$ is the vector $\boldsymbol{\theta}$ without considering the parameter θ_i.

3. Conditional posterior distribution of α_2

$$\pi\left(\alpha_2 \mid \boldsymbol{\theta}_{\{-\alpha_2\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim N\left(\frac{\sigma^2_{\alpha_2} \sum_{i=k+1}^{n} (Y_i - \beta_2 x_i) + \alpha_{02}\sigma^2_{e_2}}{(n-k)\sigma^2_{\alpha_2} + \sigma^2_{e_2}}, \frac{\sigma^2_{e_2}\sigma^2_{\alpha_2}}{(n-k)\sigma^2_{\alpha_2} + \sigma^2_{e_2}}\right)$$

4. Conditional posterior distribution of β_1

$$\pi\left(\beta_1 \mid \boldsymbol{\theta}_{\{-\beta_1\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim N\left(\frac{\sigma^2_{\beta_1} \sum_{i=1}^{k} (Y_i - \alpha_1) x_i + \beta_{01}\sigma^2_{e_1}}{\sigma^2_{\beta_1} \sum_{i=1}^{k} x_i^2 + \sigma^2_{e_1}}, \frac{\sigma^2_{e_1}\sigma^2_{\beta_1}}{\sigma^2_{\beta_1} \sum_{i=1}^{k} x_i^2 + \sigma^2_{e_1}}\right)$$

5. Conditional posterior distribution of β_2

$$\pi\left(\beta_2 \mid \boldsymbol{\theta}_{\{-\beta_2\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim N\left(\frac{\sigma^2_{\beta_2} \sum_{i=k+1}^{n} (Y_i - \alpha_2) x_i + \beta_{02}\sigma^2_{e_2}}{\sigma^2_{\beta_2} \sum_{i=k+1}^{n} x_i^2 + \sigma^2_{e_2}}, \frac{\sigma^2_{e_2}\sigma^2_{\beta_2}}{\sigma^2_{\beta_2} \sum_{i=k+1}^{n} x_i^2 + \sigma^2_{e_2}}\right)$$

6. Conditional posterior distribution of μ_1

$$\pi\left(\mu_1 \mid \boldsymbol{\theta}_{\{-\mu_1\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim N\left(\frac{\sigma^2_{\mu_1}\sum_{i=1}^{k} x_i + \mu_{01}\sigma^2_{x_1}}{k\sigma^2_{\mu_1}+\sigma^2_{x_1}}, \frac{\sigma^2_{x_1}\sigma^2_{\mu_1}}{k\sigma^2_{\mu_1}+\sigma^2_{x_1}}\right)$$

7. Conditional posterior distribution of μ_2

$$\pi\left(\mu_2 \mid \boldsymbol{\theta}_{\{-\mu_2\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim N\left(\frac{\sigma^2_{\mu_2}\sum_{i=k+1}^{n} x_i + \mu_{02}\sigma^2_{x_2}}{(n-k)\sigma^2_{\mu_2}+\sigma^2_{x_2}}, \frac{\sigma^2_{x_2}\sigma^2_{\mu_2}}{(n-k)\sigma^2_{\mu_2}+\sigma^2_{x_2}}\right)$$

8. Conditional posterior distribution of $\sigma^2_{u_1}$

$$\pi\left(\sigma^2_{u_1} \mid \boldsymbol{\theta}_{\{-\sigma^2_{u_1}\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim GI\left(\frac{k}{2}+a_{u_1}, \frac{1}{2}\sum_{i=1}^{k}(X_i-x_i)^2+b_{u_1}\right)$$

9. Conditional posterior distribution of $\sigma^2_{u_2}$

$$\pi\left(\sigma^2_{u_2} \mid \boldsymbol{\theta}_{\{-\sigma^2_{u_2}\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim GI\left(\frac{(n-k)}{2}+a_{u_2}, \frac{1}{2}\sum_{i=k+1}^{n}(X_i-x_i)^2+b_{u_2}\right)$$

10. Conditional posterior distribution of $\sigma^2_{e_1}$

$$\pi\left(\sigma^2_{e_1} \mid \boldsymbol{\theta}_{\{-\sigma^2_{e_1}\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim GI\left(\frac{k}{2}+a_{e_1}, \frac{1}{2}\sum_{i=1}^{k}(Y_i-\alpha_1-\beta_1 x_i)^2+b_{e_1}\right)$$

11. Conditional posterior distribution of $\sigma^2_{e_2}$

$$\pi\left(\sigma^2_{e_2} \mid \boldsymbol{\theta}_{\{-\sigma^2_{e_2}\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim GI\left(\frac{(n-k)}{2}+a_{e_2}, \frac{1}{2}\sum_{i=k+1}^{n}(Y_i-\alpha_2-\beta_2 x_i)^2+b_{e_2}\right)$$

12. Conditional posterior distribution of $\sigma^2_{x_1}$

$$\pi\left(\sigma^2_{x_1} \mid \boldsymbol{\theta}_{\{-\sigma^2_{x_1}\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim GI\left(\frac{k}{2}+a_{x_1}, \frac{1}{2}\sum_{i=1}^{k}(x_i-\mu_1)^2+b_{x_1}\right)$$

13. Conditional posterior distribution of $\sigma^2_{x_2}$

$$\pi\left(\sigma^2_{x_2} \mid \boldsymbol{\theta}_{\{-\sigma^2_{x_2}\}}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\right) \sim GI\left(\frac{(n-k)}{2} + a_{x_2}, \frac{1}{2}\sum_{i=k+1}^{n}(x_i - \mu_2)^2 + b_{x_2}\right)$$

14. Conditional posterior distribution of x_i, with $\boldsymbol{x}_{-i} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$

$$\pi\left(x_i \mid \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}_{-i}\right) \sim N\left(\mu_{x_i}, Var(\mu_{x_i})\right)$$

where,

$$\mu_{x_i} = \frac{(Y_i - \alpha_1)\beta_1\sigma^2_{u_1}\sigma^2_{x_1} + X_i\sigma^2_{e_1}\sigma^2_{x_1} + \mu_1\sigma^2_{e_1}\sigma^2_{u_1}}{\beta_1^2\sigma^2_{u_1}\sigma^2_{x_1} + \sigma^2_{e_1}\sigma^2_{x_1} + \sigma^2_{e_1}\sigma^2_{u_1}}$$

and

$$Var(\mu_{x_i}) = \frac{\sigma^2_{e_1}\sigma^2_{u_1}\sigma^2_{x_1}}{\beta_1^2\sigma^2_{u_1}\sigma^2_{x_1} + \sigma^2_{e_1}\sigma^2_{x_1} + \sigma^2_{e_1}\sigma^2_{u_1}}$$

15. Conditional posterior distribution of x_i, with $\boldsymbol{x}_{-i} = (x_{k+1}, x_{k+2}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$

$$\pi\left(x_i \mid \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}_{-i}\right) \sim N\left(\mu_{x_i}, Var(\mu_{x_i})\right)$$

where,

$$\mu_{x_i} = \frac{(Y_i - \alpha_2)\beta_2\sigma^2_{u_2}\sigma^2_{x_2} + X_i\sigma^2_{e_2}\sigma^2_{x_2} + \mu_2\sigma^2_{e_2}\sigma^2_{u_2}}{\beta_2^2\sigma^2_{u_2}\sigma^2_{x_2} + \sigma^2_{e_2}\sigma^2_{x_2} + \sigma^2_{e_2}\sigma^2_{u_2}}$$

and

$$Var(\mu_{x_i}) = \frac{\sigma^2_{e_2}\sigma^2_{u_2}\sigma^2_{x_2}}{\beta_2^2\sigma^2_{u_2}\sigma^2_{x_2} + \sigma^2_{e_2}\sigma^2_{x_2} + \sigma^2_{e_2}\sigma^2_{u_2}}$$

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 39 a 54

An Alternative Item Count Technique in Sensitive Surveys

Una técnica alternativa de conteo de ítems en encuestas sensitivas

Zawar Hussain[1,2,a], Ejaz Ali Shah[2,b], Javid Shabbir[2,c]

[1]Department of Statistics, Faculty of Sciences, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

[2]Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan

Abstract

The present study is basically meant to propose an improved item count technique which will mainly have an impact on sensitive fields such as health care. It is attempted to highlight the scope of the proposal relative to the usual and existing methods serving the same purpose. The proposed improved Item Count Technique (ICT) has the major advantage that it does not require two subsamples (as is the case in usual ICT) and there is no need of finding optimum subsample sizes. The proposed ICT has been observed performing well, as compared to the usual ICT, in terms of relative efficiency. The innovative method of Randomized Response (RR) technique has also been compared with the proposed ICT and it is found that the proposed technique uniformly performs better when the number of innocuous items is greater than 3.

***Key words*:** Health surveys, Privacy, Proportion estimation, Randomized response, Sensitive question.

Resumen

El presente articulo propone una técnica de conteo de items con aplicaciones principalmente en el campo de la salud. Se muestran las ventajas de nuestra propuesta y de otros métodos que sirven con el mismo fin. La técnica de conteo de ítems propuesta (ICT, por su sigla en inglés) tiene la ventaja de que no requiere dos submuestras (como es el caso en el ICT clásico) y no es necesario de encontrar los tamaños de las submuestras óptimos. El ICT propuesto tiene un mejor comportamiento en términos de eficiencia relativa. El método de la técnica de respuesta aleatorizada (RR, por su sigla en inglés) es también comparado con el ICT propuesto y se encuentra que la técnica

[a]Professor. E-mail: zhlangah@yahoo.com
[b]Professor. E-mail: alishah_ejaz@yahoo.com
[c]Professor. E-mail: jsqau@yahoo.com

propuesta se desempeña mejor cuando el número de ítems inocuos es mayor de 3.

***Palabras clave*:** encuestas de salud, estimación de la proporción, preguntas sensibles, privacidad, respuesta al azar.

1. Introduction

In estimating the population proportion of a sensitive characteristic (induced abortion, shoplifting, tax evasion) through direct questioning, truthfulness of the answers may be suspected due to various reasons, namely, social stigma, embarrassment, monetary penalty, and many others. These and similar other factors are directly related to the health issues and some improved/alternative techniques to hit these areas are indispensable to address the complications involved in them. There are a number of papers showing such concerns. Some literature in this regard may be seen in Bjorner, Kosinski & Ware (2003) and Martin, Kosinski, Bjorner, Ware & MacLean (2007), and the references therein.

An ingenious alternative to direct questioning introduced by Warner (1965), known as Randomized Response Technique (RRT), has been developed rapidly. For a good review of developments on RRTs we would refer the reader to Tracy & Mangat (1996) and Chaudhuri & Mukherjee (1988). The RRT has been used in many studies including Liu & Chow (1976), Reinmuth & Geurts (1975), Geurts (1980), Larkins, Hume & Garcha (1997), etc. Geurts (1980) reported that RRT had financial limitations since it requires larger sample sizes to obtain the confidence intervals comparable to the direct questioning technique. More time is needed to administer and explain the procedure to the survey respondents. In addition, tabulation and calculation of the results are comparatively laborious. Larkins et al. (1997) found that RRT was not a good alternative for estimating the proportion of tax payers/non-payers. Dalton & Metzger (1992) were of the view that RRT might not be effective through a mailed or telephonic survey. Hubbard, Casper & Lessler (1989) stated that the main technical problem for RRTs is making the decision about what kind of the randomization device would be the best in a given situation, and that the most crucial aspect of the RRT is about the respondent's acceptance of the technique. Chaudhuri & Christofides (2007) also gave a criticism on the RRT in the sense that it demands the respondent's skill of handling the device and also asks respondents to report the information which may be useless or tricky. A clever respondent may also think that his/her reported response can be traced back to his/her actual status if he/she does not understand the mathematical logic behind the randomization device. Some of the alternatives to the RR technique include the Item Count Technique (Droitcour, Caspar, Hubbard, Parsley, Visscher & Ezzati 1991), the Three card method (Droitcour, Larson & Scheuren 2001), and the Nominative technique (Miller 1985). These alternatives are designed because, in general, respondent evade sensitive questions especially regarding personal issues, socially deviant behaviors or illegal acts. Chaudhuri & Christofides (2007) also added that in these three alternatives to RRT respondents know that what they are revealing about themselves and they do not need to know

about any special estimation technique. Also respondents provide answers which make sense to them.

2. Item Count Techniques

In order to estimate the proportion of people with a stigmatizing attribute a promising indirect questioning technique called Item Count Technique (ICT), was introduced by Droitcour et al. (1991). It consists of taking two subsamples of sizes n_1 and n_2. The ith respondent in the first subsample is given a list of g innocuous items and asked to report the number, say X_i of items that are applicable to them ($X_i \leq g$). Similarly, the jth respondent in the second subsample is provided another list of $(g+1)$ items including the sensitive item and asked to report a number, say Y_j of the items that are applicable to them ($Y_j \leq g+1$). The g innocuous items may or may not be the same in both subsamples. An unbiased estimator of the proportion of sensitive item in the population say π is given by:

$$\widehat{\pi}_I = \overline{Y} - \overline{X} \tag{1}$$

where $\overline{Y}$ and $\overline{X}$ represent the sample mean from the second and first subsamples, respectively.

To our knowledge, no author has given the variance expression of the estimator given in (1). We have derived the variance of the estimator in (1), and it is given by:

$$V\left(\widehat{\pi}_I\right) = \frac{\pi\left(1-\pi\right)}{n_2} + \frac{n\sum\limits_{j=1}^{g}\theta_j\left(1-\sum\limits_{j=1}^{g}\theta_j\right)}{n_1 n_2} + \frac{n\sum\limits_{\substack{j,k=1\\ j\neq k}}^{g}\theta_j\theta_k}{n_1 n_2} \tag{2}$$

where θ_j is the known proportion of the item j in the population. More details about ICT can be found in Droitcour et al. (1991) and Droitcour & Larson (2002). Dalton, Wimbush & Daily (1994) named ICT as the unmatched count technique and applied it to study the illicit behaviors of the auctioneers, and as compared to the direct questioning they obtained higher estimates of six stigmatized items. Wimbush & Dalton (1997) applied this technique in estimating the employee theft rate in high-theft-exposure business and found higher theft rates. Tsuchiya (2005) extended the ICT to domain estimators by the stratified method, the cross-based method, and the double cross-based method. More recently, Tsuchiya, Hirai & Ono (2007) studied the properties of the ICT through an experimental web survey and found that ICT yielded higher estimates of the proportions of the shoplifters by nearly 10% as that of yielded by direct questioning. They also found that the cross-based method was the most appropriate one.

Besides its fruitful applications ICT has not been found fruitful in many studies; for example, Droitcour et al. (1991), Biemer & Wright (2004) and Ahart & Sackett (2004) failed to get higher estimates in their studies of different stigmatized traits. We have focused on the issue of the need of two subsamples in the usual application of ICT and have proposed an alternative ICT which does not

need two subsamples. Avoiding the need of two subsamples for our proposed ICT makes it more attractive in terms of cost and statistical efficiency. The following section provides a description of the proposed methodology.

2.1. Proposed Item Count Technique

Each respondent in a sample of size n is provided a questionnaire (list of questions) consisting of $g\,(\geq 2)$ questions. The jth question consists of queries about an unrelated item (F_j), and a sensitive characteristic (S). The respondent is requested to count 1 if he/she possesses at least one of the characteristics F_j and S, otherwise, count 0, as a response to the jth question, and to report the total count based on entire questionnaire.

The list of items is given to the respondents and they are sent to another room so that they are unseen to the interviewer. To illustrate, suppose the sensitive study item (S) be the cheating in exams and the unrelated items $(F_j, j = 1, 2.)$ are: (i) "Do you live in the hostel?" and (ii) "Is the last digit of your registration number odd?" It is obvious that there are almost (if not exactly) 50% (known) of the students having an odd registration number and proportion of the students living in hostel is easily available from the warden office. Let Z_i denote the total count of ith respondent, and then mathematically we can write it as:

$$Z_i = \sum_{j=1}^{g} \alpha_j \tag{3}$$

where α_j can assume values "1" and "0" with probabilities $(\pi + \theta_j - \pi\theta_j)$ and $(1 - \pi - \theta_j + \pi\theta_j)$, respectively.

Taking expectation on (3) we have:

$$\begin{aligned} E\,(Z_i) &= \sum_{j=1}^{g} E\,(\alpha_j) = g\pi + \sum_{j=1}^{g} \theta_j - \pi \sum_{j=1}^{g} \theta_j \\ &= \left(g - \sum_{j=1}^{g} \theta_j \right) \pi + \sum_{j=1}^{g} \theta_j \end{aligned}$$

This suggests defining an unbiased estimator of π as:

$$\widehat{\pi}_P = \frac{\overline{Z} - \sum_{j=1}^{g} \theta_j}{g - \sum_{j=1}^{g} \theta_j} \tag{4}$$

The estimator given in (4) serves the purpose of estimating π as is done by $\widehat{\pi}_I$ in (1). The estimator $\widehat{\pi}_P$ obtained through our proposed ICT does not demand two subsamples which are needed by $\widehat{\pi}_I$ based on the usual ICT. This property (avoiding the need of two subsamples) makes our proposal more attractive and practicable.

The variance of the estimator $\widehat{\pi}_P$ is given by (see Appendix)

$$V\left(\widehat{\pi}_P\right) = \frac{\pi\left(1-\pi\right)}{n} + \frac{\left(1-\pi\right)}{n\left(g-\sum_{j=1}^{g}\theta_j\right)^2}\left\{\left(\sum_{j=1}^{g}\theta_j\right)\left(1-\sum_{j=1}^{g}\theta_j\right)+\sum_{\substack{j,k=1\\ j\neq k}}^{g}\theta_j\theta_k\right\} \quad (5)$$

Some comments are in order. It is to be noted that in some surveys it may be possible to have unrelated traits $(F_j, j = 1, 2, \ldots, g)$ with equal proportions $(\theta_j, j = 1, 2, \ldots, g)$. In these situations we have $\theta_j = \frac{1}{g}$ for all j and consequently the variance of the proposed estimator $\widehat{\pi}_P$ reduces to

$$V\left(\widehat{\pi}_P\right) = \frac{\pi\left(1-\pi\right)}{n} + \frac{\left(1-\pi\right)}{ng\left(g-1\right)} \quad (6)$$

As pointed by the two referees, it is just possible that the actual status of the respondents about one (or all) the unrelated item(s) may be known to the interviewer by any means, then the response of 0 or g would disclose his/her status about the sensitive item. In this case privacy protection provided to the respondents will be limited. Thus, the unrelated items should be chosen in such a way that the actual status of the respondents about at least one of the unrelated items must be impossible to know by any means. To fix the idea, suppose the unrelated items are (i) and (ii) as we discussed above, then knowing the residential status of a particular student is difficult while actually conducting the survey but the proportion of students living in hostel may be readily available from the warden office. Similar is the case with the unrelated item of registration number. If it is possible to exactly guess or know about the particular item(s) for a given individual then such item(s) must not be included in the group of items. In this way, respondents would feel more protected and be motivated to answer truthfully. And, of course, the interviewer's ethical responsibility of being honest is more apparent, in the sense that he would be asking about those items about which he knows nothing of a particular respondent. The item count technique surveys are conducted in the hope that the respondents will be motivated more to reveal truthful answers rather than trapping them in mathematical tricks to trace their actual responses on the sensitive items. It will essentially be a direct questioning situation if surveyor is able to know the status of each respondent on each unrelated item. So, respondents must be assured that it is impossible to know the status of individual about an item but, of course, its population proportion is known somehow. It is easy to understand now that knowing the population proportion of an unrelated item is not harmful but knowing the individuals' status is. Moreover, another characteristic of such indirect survey methods is the anonymity. The identity (in terms of name or registration number, etc.) of the respondent is not required. The respondents may just write their answers on a sheet of paper and drop them in a box making it impossible to know the response of a particular respondent even the interviewer is able to know the status of a particular respondent on a given item. For example,

in our situation, if the surveyor is able enough to guess or know the residential status (*hostelite* or *non-hostelite*) of a student, due to anonymity, he/she is not able to know reported response of a given respondent. Thus, any unrelated item whose population proportion is known may be used in this technique.

The acceptance of the unrelated question by the respondents, as pointed by the two learned referees, is another key issue of concern. In some cases, it would be needed to explain the working of whole the technique to the respondents. But it depends on the nature and composition of the population. In such cases survey must be conducted under the supervision of a trained statistician. More specifically, if the studied population is composed of illiterate individuals the technique must be explained to them prior to actually conducting the survey. The explanation of the technique would possibly decrease the suspicion among the respondents of being tricked. Further, the suspicion depends upon the anonymity provided by the survey method. If the respondents are explained about the working of the survey in such a way that their anonymity is assured and they are giving meaningful answers in the sense that only population proportion of study item is estimated and individual's status can not be known through their reported response. With this explanation and provision of anonymity it is anticipated that any unrelated item with known population proportion of prevalence may be fairly used. One more thing about the acceptance of unrelated items by the respondents is the simplicity of the question. The unrelated question must not be an open ended or having multiple answers, that is, it must be a binary item.

3. Performance Evaluations and Comparison

In this section, we provide efficiency comparisons of the estimator $\widehat{\pi}_P$ of the proposed ICT with the $\widehat{\pi}_I$ of the usual ICT and another obtained through RRT of Warner (1965). As we have discussed, that ICT has been developed as an alternative to RRT, so we have also compared our technique with RR technique proposed by Warner (1965).

3.1. Proposed versus Usual ICT

We compare the proposed estimator $\widehat{\pi}_P$ with the usual ICT estimator $\widehat{\pi}_I$ in both the situations of having and not having unequal $\theta_j = \frac{1}{g}$. In case of having unequal $\theta_j's$ the proposed estimator $\widehat{\pi}_P$ would be more efficient than the estimator $\widehat{\pi}_I$ if

$$V\left(\widehat{\pi}_I\right) - V\left(\widehat{\pi}_P\right) \geq 0,$$

$$\frac{\pi\left(1-\pi\right)}{n_2} + \frac{n\sum\limits_{j=1}^{g}\theta_j\left(1-\sum\limits_{j=1}^{g}\theta_j\right)}{n_1 n_2} + \frac{n\sum\limits_{\substack{j,k=1\\ j\neq k}}^{g}\theta_j\theta_k}{n_1 n_2}$$

$$-\frac{\pi\left(1-\pi\right)}{n}-\frac{\left(1-\pi\right)}{n\left(g-\sum\limits_{j=1}^{g}\theta_j\right)^2}\left\{\left(\sum_{j=1}^{g}\theta_j\right)\left(1-\sum_{j=1}^{g}\theta_j\right)+\sum_{\substack{j,k=1\\ j\neq k}}^{g}\theta_j\theta_k\right\}\geq 0$$

$$\frac{\pi\left(1-\pi\right)n_1}{nn_2}\quad+\quad\left\{\sum_{j=1}^{g}\theta_j\left(1-\sum_{j=1}^{g}\theta_j\right)+\sum_{\substack{j,k=1\\ j\neq k}}^{g}\theta_j\theta_k\right\}\times\left[\frac{n^2\left(g-\sum\limits_{j=1}^{g}\theta_j\right)^2-\left(1-\pi\right)n_1n_2}{nn_1n_2\left(g-\sum\limits_{j=1}^{g}\theta_j\right)^2}\right]\geq 0$$

Moreover, in case of having $\theta_j=\frac{1}{g}\forall\, j$, such that $\sum_{j=1}^{g}\theta_j=1$, the proposed estimator $\widehat{\pi}_P$would be more efficient than the estimator $\widehat{\pi}_I$ if

$$\left[\frac{\pi\left(1-\pi\right)n_1}{nn_2}+\frac{n^2\left(g-1\right)^2-\left(1-\pi\right)n_1n_2}{nn_1n_2g\left(g-1\right)}\right]\geq 0 \tag{7}$$

which is always true for every value of g (≥ 2) (i.e., the number of innocuous items).

3.2. Proposed versus Warner's RRT

To have an efficiency comparison, we first give a short description of Warner (1965) RRT. Warner (1965) introduced this method to decrease the biasedness in the estimators and to increase the response rate. Warner's technique consists of two complimentary questions A (Do you belong to the sensitive group?) and A^c (Do you not belong to the sensitive group?) to be answered on a probability basis. Assuming a simple random sampling with replacement (SRSWR), the ith selected respondent is asked to select a question (A or A^{c}) and report "yes" if his/her actual status matches with selected question, and "no" otherwise. Assuming that p is the probability of selecting question A, and π is the population proportion of individuals with sensitive group, the probability of "yes" for a particular respondent, denoted by θ, is given by:

$$P\left(\text{yes}\right)=\theta=p\pi+\left(1-p\right)\left(1-\pi\right) \tag{8}$$

From (8), we have

$$\pi=\frac{\theta-\left(1-p\right)}{2p-1} \tag{9}$$

An unbiased estimator of π, by the methods of moment and maximum likelihood estimation, is given as:

$$\widehat{\pi}_W = \frac{\widehat{\theta} - (1-p)}{2p-1} \tag{10}$$

where $\widehat{\theta} = \frac{n'}{n}$ and n' is the number of "yes" responses in the sample of size n.

The variance of the estimator $\widehat{\pi}_W$ is given by:

$$Var\left(\widehat{\pi}_W\right) = \frac{\pi\left(1-\pi\right)}{n} + \frac{p\left(1-p\right)}{n\left(2p-1\right)^2} \tag{11}$$

Comparing (5) and (11) we can see that the proposed estimator $\widehat{\pi}_P$ will be more precise than $\widehat{\pi}_W$if

$$Var\left(\widehat{\pi}_W\right) - Var\left(\widehat{\pi}_P\right) \geq 0$$

$$\frac{p\left(1-p\right)}{n\left(2p-1\right)^2} - \frac{\left(1-\pi\right)}{n\left(g - \sum_{j=1}^{g}\theta_j\right)^2}\left\{\left(\sum_{j=1}^{g}\theta_j\right)\left(1-\sum_{j=1}^{g}\theta_j\right) + \sum_{\substack{j,k=1 \\ j\neq k}}^{g}\theta_j\theta_k\right\} \geq 0$$

Further comparing (6) and (11) we can see that the proposed estimator $\widehat{\pi}_P$ will be more precise than $\widehat{\pi}_W$ if

$$\frac{p\left(1-p\right)}{n\left(2p-1\right)^2} - \frac{\left(1-\pi\right)}{ng\left(g-1\right)} \geq 0$$

We have calculated the Relative Efficiency (RE) of the proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ when it is difficult/impossible to have $\theta_j = \frac{1}{g}$, and results are provided in Tables 1–9. The RE of the proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_W$ for $\theta_j \neq \frac{1}{g}$ is presented in Tables 10–12. For $\theta_j = \frac{1}{g}$ the RE of $\widehat{\pi}_P$ relative to $\widehat{\pi}_W$ is arranged in Table 13.

Table 1: RE of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 20$, $n_1 = 10$, $n_2 = 10$.

π	$g=2$		$g=3$		$g=4$		$g=5$	
	$\sum_{j=1}^{g}\theta_j = 0.3$	$\sum_{j=1}^{g}\theta_j = 1.7$	$\sum_{j=1}^{g}\theta_j = 0.6$	$\sum_{j=1}^{g}\theta_j = 2.4$	$\sum_{j=1}^{g}\theta_j = 1$	$\sum_{j=1}^{g}\theta_j = 3$	$\sum_{j=1}^{g}\theta_j = 1.5$	$\sum_{j=1}^{g}\theta_j = 3.5$
0.1	7.0298	0.4556	12.4787	1.6290	18.6250	4.1389	24.9068	8.4680
0.2	5.7590	0.5541	9.6476	1.8271	14.0400	4.3333	18.5551	8.2768
0.3	5.2484	0.6591	8.4993	2.0463	12.1764	4.6000	15.9676	8.3472
0.4	5.0701	0.7762	8.0579	2.3046	11.4419	4.9697	14.9373	8.6756
0.5	5.1150	0.9152	8.0701	2.6325	11.4231	5.500	14.8905	9.3253
0.6	5.3896	1.0953	8.5311	3.0887	12.0984	6.3076	15.7921	10.4674
0.7	6.0181	1.3610	9.6598	3.8089	13.8000	7.6667	18.0910	12.5347
0.8	7.4450	1.8447	12.2746	5.1979	17.7721	10.400	23.4744	16.8545
0.9	11.9614	3.2084	20.6151	9.2755	30.4773	18.6250	40.7140	30.100

From the above tables 1–13 it is advocated that

1. For larger values of $\sum_{j=1}^{g}\theta_j$ the proposed estimator $\widehat{\pi}_P$ is less efficient than $\widehat{\pi}_I$ when g and $\pi$are smaller, but when g increases it becomes more efficient even for smaller values of π.

2. For smaller values of $\sum_{j=1}^{g}\theta_j$ the proposed estimator $\widehat{\pi}_P$ is more efficient than $\widehat{\pi}_I$ even when g and $\pi$are smaller.

3. n, n_1 and n_2 do not have a significant effect on the RE of the proposed estimator relative to $\widehat{\pi}_I$ except the case when n and $\sum_{j=1}^{g}\theta_j$ are larger and $g = 2$.

4. When $\sum_{j=1}^{g}\theta_j = 1$ the proposed estimator is always more efficient.

5. For smaller p the proposed estimator is less efficient than $\widehat{\pi}_W$ but as g and π are increased the RE of the proposed estimator is increased.

6. When $\sum_{j=1}^{g}\theta_j$is smaller the proposed estimator is more efficient than $\widehat{\pi}_W$ when $\pi > 0.1$ and $g > 2$.

7. Compared to $\widehat{\pi}_W$ proposed estimator $\widehat{\pi}_P$ is more efficient than $\widehat{\pi}_W$ for $g > 3$ under the given condition of $\theta_j = \frac{1}{g}$.

8. The RE of the proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_W$ increases with an increase in p for a given value of g and $\pi$and it increases, for a given value of p, if g increases.

In the application scenario all the disciplines which are of sensitive nature and need extreme care in taking responses may take benefit out of the proposal, e.g., having more concern on time sensitivity (cf. Bonetti, Waeckerlin, Schuepfer & Frutiger 2000).

TABLE 2: RE of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 20$, $n_1 = 12$, $n_2 = 8$.

π	$g = 2$		$g = 3$		$g = 4$		$g = 5$	
	$\sum_{j=1}^{g}\theta_j = 0.3$	$\sum_{j=1}^{g}\theta_j = 1.7$	$\sum_{j=1}^{g}\theta_j = 0.6$	$\sum_{j=1}^{g}\theta_j = 2.4$	$\sum_{j=1}^{g}\theta_j = 1$	$\sum_{j=1}^{g}\theta_j = 3$	$\sum_{j=1}^{g}\theta_j = 1.5$	$\sum_{j=1}^{g}\theta_j = 3.5$
0.1	7.5462	0.4891	13.2303	1.7271	19.6354	4.3634	26.1792	8.9007
0.2	6.2910	0.6051	10.3474	1.9596	14.9250	4.6064	19.6285	8.7555
0.3	5.7906	0.7271	9.1825	2.2107	13.0147	4.9167	16.9640	8.8681
0.4	5.6240	0.8610	8.7410	2.5000	12.2674	5.3282	15.9087	9.2398
0.5	5.6834	1.0169	8.7665	2.8594	12.2596	5.9028	15.8716	9.9397
0.6	5.9783	1.2150	9.2843	3.3505	12.9713	6.7628	16.8191	11.1481
0.7	6.6398	1.5015	10.4363	4.1152	14.7500	8.1944	19.2198	13.3168
0.8	8.1312	2.0147	13.1650	5.5748	18.8924	11.0556	24.8323	17.8295
0.9	12.8399	3.4441	21.8568	9.8341	32.1307	19.6354	42.7940	31.6386

TABLE 3: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 20$, $n_1 = 8$, $n_2 = 12$.

π	$g = 2$		$g = 3$		$g = 4$		$g = 5$	
	$\sum_{j=1}^{g} \theta_j = 0.3$	$\sum_{j=1}^{g} \theta_j = 1.7$	$\sum_{j=1}^{g} \theta_j = 0.6$	$\sum_{j=1}^{g} \theta_j = 2.4$	$\sum_{j=1}^{g} \theta_j = 1$	$\sum_{j=1}^{g} \theta_j = 3$	$\sum_{j=1}^{g} \theta_j = 1.5$	$\sum_{j=1}^{g} \theta_j = 3.5$
0.1	7.0994	0.4601	12.7671	1.6667	19.1667	4.2592	25.7098	8.7411
0.2	5.7082	0.5492	9.7519	1.8468	14.3250	4.4213	19.0280	8.4877
0.3	5.1438	0.6459	8.5244	2.0523	12.3530	4.6667	16.3018	8.5219
0.4	4.9388	0.7561	8.0463	2.3013	11.5698	5.0252	15.2107	8.8344
0.5	4.9730	0.8898	8.0479	2.6250	11.5348	5.5556	15.1502	9.4879
0.6	5.2500	1.0670	8.5189	3.0843	12.2336	6.3782	16.0812	10.6589
0.7	5.8981	1.3340	9.6888	3.8202	14.0000	7.7778	18.4696	12.7970
0.8	7.3791	1.8284	12.4072	5.2540	18.1329	10.6111	24.0726	17.2841
0.9	12.0796	3.2401	21.0914	9.4897	31.3636	19.1667	42.0268	31.0714

TABLE 4: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 50$, $n_1 = 25$, $n_2 = 25$.

π	$g = 2$		$g = 3$		$g = 4$		$g = 5$	
	$\sum_{j=1}^{g} \theta_j = 0.3$	$\sum_{j=1}^{g} \theta_j = 1.7$	$\sum_{j=1}^{g} \theta_j = 0.6$	$\sum_{j=1}^{g} \theta_j = 2.4$	$\sum_{j=1}^{g} \theta_j = 1$	$\sum_{j=1}^{g} \theta_j = 3$	$\sum_{j=1}^{g} \theta_j = 1.5$	$\sum_{j=1}^{g} \theta_j = 3.5$
0.1	7.0299	0.4556	12.4788	1.6290	18.6250	4.1389	24.9067	8.4680
0.2	5.7590	0.5541	9.6477	1.8271	14.0400	4.3333	18.5551	8.2768
0.3	5.2484	0.6591	8.4993	2.0463	12.1765	4.6000	15.9675	8.3472
0.4	5.0701	0.7762	8.0579	2.3046	11.4419	4.9697	14.9373	8.6756
0.5	5.1150	0.9152	8.0709	2.6325	11.4231	5.5000	14.8904	9.3253
0.6	5.3896	1.0953	8.5311	3.0887	12.0984	6.3076	15.7912	10.4674
0.7	6.0182	1.3610	9.6598	3.8089	13.8000	7.6667	18.0910	12.5347
0.8	7.4450	1.8447	12.2747	5.1979	17.7722	10.40	23.4744	16.8545
0.9	11.9614	3.2084	20.6152	9.2755	30.4773	18.6250	40.7140	30.1008

TABLE 5: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 50$, $n_1 = 30$, $n_2 = 20$.

π	$g = 2$		$g = 3$		$g = 4$		$g = 5$	
	$\sum_{j=1}^{g} \theta_j = 0.3$	$\sum_{j=1}^{g} \theta_j = 1.7$	$\sum_{j=1}^{g} \theta_j = 0.6$	$\sum_{j=1}^{g} \theta_j = 2.4$	$\sum_{j=1}^{g} \theta_j = 1$	$\sum_{j=1}^{g} \theta_j = 3$	$\sum_{j=1}^{g} \theta_j = 1.5$	$\sum_{j=1}^{g} \theta_j = 3.5$
0.1	7.5462	0.4891	13.2303	1.7271	19.6354	4.3634	26.1792	8.9007
0.2	6.2910	0.6051	10.3474	1.9596	14.9250	4.6064	19.6285	8.7555
0.3	5.7906	0.7271	9.1825	2.2107	13.0147	4.9167	16.9640	8.8681
0.4	5.6240	0.8610	8.7410	2.5000	12.2674	5.3282	15.9087	9.2398
0.5	5.6834	1.0169	8.7665	2.8594	12.2596	5.9028	15.8716	9.9397
0.6	5.9783	1.2150	9.2843	3.3505	12.9713	6.7628	16.8191	11.1481
0.7	6.6398	1.5015	10.4363	4.1152	14.7500	8.1944	19.2198	13.3168
0.8	8.1312	2.0147	13.1650	5.5748	18.8924	11.0556	24.8323	17.8295
0.9	12.8399	3.4441	21.8568	9.8341	32.1307	19.6354	42.7940	31.6386

TABLE 6: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 50$, $n_1 = 20$, $n_2 = 30$.

π	$g = 2$		$g = 3$		$g = 4$		$g = 5$	
	$\sum_{j=1}^{g} \theta_j = 0.3$	$\sum_{j=1}^{g} \theta_j = 1.7$	$\sum_{j=1}^{g} \theta_j = 0.6$	$\sum_{j=1}^{g} \theta_j = 2.4$	$\sum_{j=1}^{g} \theta_j = 1$	$\sum_{j=1}^{g} \theta_j = 3$	$\sum_{j=1}^{g} \theta_j = 1.5$	$\sum_{j=1}^{g} \theta_j = 3.5$
0.1	7.0994	0.4601	12.7671	1.6667	19.1667	4.2592	25.7098	8.7411
0.2	5.7082	0.5492	9.7519	1.8468	14.3250	4.4213	19.0280	8.4877
0.3	5.1438	0.6459	8.5244	2.0523	12.3530	4.6667	16.3018	8.5219
0.4	4.9388	0.7561	8.0463	2.3013	11.5698	5.0252	15.2107	8.8344
0.5	4.9730	0.8898	8.0479	2.6250	11.5348	5.5556	15.1502	9.4879
0.6	5.2500	1.0670	8.5189	3.0843	12.2336	6.3782	16.0812	10.6589
0.7	5.8981	1.3340	9.6888	3.8202	14.0000	7.7778	18.4696	12.7970
0.8	7.3791	1.8284	12.4072	5.2540	18.1329	10.6111	24.0726	17.2841
0.9	12.0796	3.2401	21.0914	9.4897	31.3636	19.1667	42.0268	31.0714

TABLE 7: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 100$, $n_1 = 50$, $n_2 = 50$.

π	$g = 2$		$g = 3$		$g = 4$		$g = 5$	
	$\sum_{j=1}^{g} \theta_j = 0.3$	$\sum_{j=1}^{g} \theta_j = 1.7$	$\sum_{j=1}^{g} \theta_j = 0.6$	$\sum_{j=1}^{g} \theta_j = 2.4$	$\sum_{j=1}^{g} \theta_j = 1$	$\sum_{j=1}^{g} \theta_j = 3$	$\sum_{j=1}^{g} \theta_j = 1.5$	$\sum_{j=1}^{g} \theta_j = 3.5$
0.1	7.0298	0.4556	12.4787	1.6290	18.6250	4.1389	24.9068	8.4680
0.2	5.7590	0.5441	9.6476	1.8271	14.0400	4.3333	18.5555	8.2768
0.3	5.2484	0.6591	8.4993	2.0463	12.1764	4.6000	15.9675	8.3472
0.4	5.0701	0.7762	8.0579	2.3046	11.4419	4.9697	14.9373	8.6756
0.5	5.1150	0.9152	8.0701	2.6325	11.4231	5.5000	14.8904	9.3253
0.6	5.3896	1.0954	8.5311	3.0887	12.0936	6.3076	15.7922	10.4674
0.7	6.0181	1.3610	9.6598	3.8089	13.8000	7.6667	18.0910	12.5347
0.8	7.4450	1.8447	12.2746	5.1979	17.7721	10.40	23.4744	16.8545
0.9	11.9614	3.2084	20.6151	9.2755	30.4773	18.6250	40.7140	30.1008

TABLE 8: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 100$, $n_1 = 80$, $n_2 = 20$.

π	$g = 2$		$g = 3$		$g = 4$		$g = 5$	
	$\sum_{j=1}^{g} \theta_j = 0.3$	$\sum_{j=1}^{g} \theta_j = 1.7$	$\sum_{j=1}^{g} \theta_j = 0.6$	$\sum_{j=1}^{g} \theta_j = 2.4$	$\sum_{j=1}^{g} \theta_j = 1$	$\sum_{j=1}^{g} \theta_j = 3$	$\sum_{j=1}^{g} \theta_j = 1.5$	$\sum_{j=1}^{g} \theta_j = 3.5$
0.1	11.9895	0.7770	20.5405	2.6814	30.1562	6.7013	39.9729	13.5904
0.2	10.3074	0.9912	16.4144	3.1085	23.2875	7.1875	30.3435	13.5351
0.3	9.6561	1.2126	14.7610	3.5538	20.5147	7.7500	26.4391	13.8214
0.4	9.4637	1.4488	14.1534	4.0480	19.4477	8.4470	24.9101	14.4679
0.5	9.5908	1.7161	14.2275	4.6406	19.4711	9.3750	24.8896	15.5873
0.6	10.0600	2.0446	14.9845	5.4253	20.5635	10.7211	26.3356	17.4558
0.7	11.0721	2.5039	16.7765	6.6152	23.2500	12.9167	29.9551	20.7549
0.8	13.3247	3.3016	20.8840	8.8436	29.4778	17.2500	38.3881	27.5625
0.9	20.4003	5.4722	33.9334	15.2679	49.3466	30.1562	65.3419	48.3088

TABLE 9: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_I$ for $n = 100$, $n_1 = 20$, $n_2 = 80$.

π	$g = 2$		$g = 3$		$g = 4$		$g = 5$	
	$\sum_{j=1}^{g} \theta_j = 0.3$	$\sum_{j=1}^{g} \theta_j = 1.7$	$\sum_{j=1}^{g} \theta_j = 0.6$	$\sum_{j=1}^{g} \theta_j = 2.4$	$\sum_{j=1}^{g} \theta_j = 1$	$\sum_{j=1}^{g} \theta_j = 3$	$\sum_{j=1}^{g} \theta_j = 1.5$	$\sum_{j=1}^{g} \theta_j = 3.5$
0.1	9.9789	0.6467	18.4556	2.4092	28.04688	6.2326	37.8608	12.8723
0.2	7.6896	0.7398	13.7345	2.6010	20.5875	6.3542	27.6413	12.3289
0.3	6.7454	0.8471	11.7994	2.8407	17.5368	6.6250	23.4595	12.2637
0.4	6.3805	0.9768	11.0275	3.1539	16.3081	7.0833	21.7691	12.6435
0.5	6.3938	1.1441	10.9940	3.5859	16.2260	7.8125	21.6431	13.5542
0.6	6.7825	1.3784	11.6752	4.2271	17.2438	8.9904	23.0149	15.2548
0.7	7.7353	1.7492	13.4105	5.2880	19.8750	11.0417	26.5792	18.4158
0.8	9.9407	2.4631	17.4743	7.3997	26.0601	15.2500	34.9695	25.1079
0.9	16.9791	4.5544	30.4890	13.7181	45.8949	28.0469	61.8893	45.7563

TABLE 10: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_W$ for $n = 20$ and larger $\sum_{j=1}^{g} \theta_j$.

		π								
p	g, $\sum_{j=1}^{g} \theta_j$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	2, 1.7	0.09	0.13	0.16	0.20	0.24	0.28	0.34	0.42	0.63
	3, 2.4	0.19	0.25	0.32	0.38	0.44	0.51	0.59	0.72	1.06
	4, 3	0.32	0.42	0.50	0.58	0.65	0.73	0.83	1.00	1.44
	5, 3.5	0.49	0.60	0.69	0.77	0.85	0.93	1.04	1.23	1.74
0.2	2, 1.7	0.21	0.25	0.31	0.36	0.42	0.51	0.63	0.84	1.45
	3, 2.4	0.43	0.51	0.59	0.68	0.78	0.91	1.10	1.45	2.45
	4, 3	0.74	0.84	0.93	1.04	1.16	1.32	1.56	2.01	3.34
	5, 3.5	1.14	1.21	1.29	1.39	1.51	1.67	1.94	2.47	4.04
0.3	2, 1.7	0.54	0.62	0.71	0.81	0.95	1.15	1.46	2.06	3.81
	3, 2.4	1.13	1.25	1.38	1.54	1.76	2.07	2.57	3.54	6.44
	4, 3	1.95	2.05	2.18	2.35	2.60	2.99	3.62	4.91	8.77
	5, 3.5	2.98	2.96	3.01	3.15	3.39	3.80	4.52	6.02	10.61
0.4	2, 1.7	2.35	2.59	2.88	3.27	3.81	4.62	5.95	8.61	16.56
	3, 2.4	4.91	5.21	5.62	6.20	7.03	8.31	10.47	14.82	27.96
	4, 3	8.45	8.56	8.87	9.45	10.42	12.00	14.79	20.53	38.06
	5, 3.5	12.96	12.38	12.28	12.65	13.55	15.26	18.46	25.20	46.06

TABLE 11: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_W$ $n = 50$ and larger $\sum_{j=1}^{g} \theta_j$.

		π								
p	g, $\sum_{j=1}^{g} \theta_j$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	2, 1.7	0.09	0.13	0.16	0.20	0.24	0.28	0.34	0.42	0.63
	3, 2.4	0.19	0.25	0.32	0.38	0.44	0.51	0.59	0.72	1.06
	4, 3	0.32	0.42	0.50	0.58	0.65	0.73	0.83	1.00	1.44
	5, 3.5	0.49	0.60	0.69	0.77	0.85	0.93	1.04	1.23	1.74
0.2	2, 1.7	0.21	0.25	0.31	0.36	0.42	0.51	0.63	0.84	1.45
	3, 2.4	0.43	0.51	0.59	0.68	0.78	0.91	1.10	1.45	2.45
	4, 3	0.74	0.84	0.93	1.04	1.16	1.32	1.56	2.01	3.34
	5, 3.5	1.14	1.21	1.29	1.39	1.51	1.67	1.94	2.47	4.04
0.3	2, 1.7	0.54	0.62	0.71	0.81	0.95	1.15	1.46	2.06	3.81
	3, 2.4	1.13	1.25	1.38	1.54	1.76	2.07	2.57	3.54	6.44
	4, 3	1.95	2.05	2.18	2.35	2.60	2.99	3.62	4.91	8.77
	5, 3.5	2.98	2.96	3.01	3.15	3.39	3.80	4.52	6.02	10.61
0.4	2, 1.7	2.35	2.59	2.88	3.27	3.81	4.62	5.95	8.61	16.56
	3, 2.4	4.91	5.21	5.62	6.20	7.03	8.31	10.47	14.82	27.96
	4, 3	8.45	8.56	8.87	9.45	10.42	12.00	14.79	20.53	38.06
	5, 3.5	12.96	12.38	12.28	12.65	13.55	15.26	18.46	25.20	46.06

4. Concluding Remarks

An alternative item count technique has been presented in this article. One of the main features of this technique is that it does not require the selection of two subsamples of sizes n_1 and n_2. Therefore, we do not need to worry about the optimum values of n_1 and n_2 (as is the case with usual ICT estimator $\widehat{\pi}_I$). Furthermore, the response from a respondent is bounded to lie between 0 and g, which helps to provide the privacy to the respondent because the response can not be traced back to respondent's actual status about the possession of sensitive item (provided that the actual status of a particular respondent about at least one unrelated characteristic is unknown to the interviewer or anonymity is provided to respondents). To avoid this situation, we recommend conducting the survey in the absence of the interviewer or the whole process must be administered unseen to the interviewer.

TABLE 12: *RE* of proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_W$ for $n = 20$ and smaller $\sum_{j=1}^{g} \theta_j$.

		π								
p	g, $\sum_{j=1}^{g} \theta_j$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	2, 0.5	0.97	1.03	1.08	1.12	1.18	1.24	1.35	1.56	2.17
	3, 0.6	1.42	1.34	1.32	1.32	1.34	1.39	1.49	1.71	2.35
	4, 1	1.44	1.35	1.33	1.33	1.35	1.40	1.50	1.71	2.36
	5, 1.5	1.44	1.35	1.33	1.33	1.35	1.40	1.50	1.71	2.36
0.2	2, 0.5	2.25	2.07	2.01	2.02	2.09	2.24	2.52	3.13	5.02
	3, 0.6	3.30	2.70	2.46	2.37	2.39	2.51	2.79	3.43	5.45
	4, 1	3.34	2.72	2.47	2.38	2.40	2.52	2.80	3.44	5.47
	5, 1.5	3.35	2.72	2.48	2.39	2.40	2.52	2.81	3.44	5.47
0.3	2, 0.5	5.90	5.05	4.68	4.58	4.70	5.08	5.87	7.63	13.17
	3, 0.6	8.67	6.58	5.72	5.39	5.38	5.71	6.51	8.37	14.33
	4, 1	8.77	6.63	5.76	5.41	5.406	5.73	6.52	8.40	14.34
	5, 1.5	8.78	6.32	5.76	5.42	5.41	5.73	6.53	8.39	14.35
0.4	2, 0.5	25.59	21.13	19.10	18.42	18.81	20.40	23.95	31.93	57.21
	3, 0.6	37.62	27.51	23.35	21.67	21.56	22.94	26.54	35.01	62.15
	4, 1	38.06	27.72	23.48	21.78	21.63	23.01	26.61	35.09	62.28
	5, 1.5	38.11	27.74	23.50	21.78	21.64	23.02	26.62	35.10	62.30

TABLE 13: Relative efficiency of the proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_W$ for $0.1 \leq \pi \leq 0.9$ and $0.1 \leq p \leq 0.4$.

π \ p	0.1	0.2	0.3	0.4	π \ p	0.1	0.2	0.3	0.4
	$g = 4$					$g = 5$			
0.1	1.397	3.239	8.500	36.909	0.1	1.708	3.958	10.388	45.111
0.3	1.306	2.438	5.673	23.142	0.3	1.4311	2.671	6.214	25.346
0.5	1.339	2.380	5.357	21.428	0.5	1.420	2.525	5.681	22.727
0.7	1.492	2.784	6.478	26.425	0.7	1.558	2.908	6.766	27.600
0.9	2.345	5.435	14.262	61.932	0.9	2.427	5.625	14.763	64.105
	$g = 6$					$g = 7$			
0.1	1.921	4.453	11.687	50.750	0.1	2.069	4.796	12.586	54.653
0.3	1.502	2.804	6.525	26.614	0.3	1.546	2.887	6.716	27.397
0.5	1.464	2.604	5.859	23.437	0.5	1.491	2.651	5.965	23.863
0.7	1.593	2.974	6.920	28.227	0.7	1.614	3.013	7.011	28.598
0.9	2.470	5.726	15.026	65.250	0.9	2.496	5.785	15.181	65.922

It has been observed that the proposed item count technique estimator performs better than the usual item count technique under the conditions that $\theta_j = \frac{1}{g}$ and $\sum_{j=1}^{g} \theta_j = 1$. It may be difficult to select the items in such a way that their proportions in the population are the same and sum to one, but this would be the case if the number of items is large. Thus, in practice, one or two innocuous items with same proportions can be found and included in the item list (e.g., item 1: Were you born in the months from January to June?, and Item 2: Is your gender male?) If the condition to satisfy the inequality (7) is hard to meet we would suggest to look for a large number of innocuous items (4, 5, 6, etc.) such that their prevalence in the population is rare and consequently we have smaller $\sum_{j=1}^{g} \theta_j$, so that inequality (7) is easily satisfied.

In brief, based on the findings of the Section 4 and the concluding discussion above we recommend the use of the proposed ICT in surveys about sensitive items instead of the usual ICT and the Warner's RRT. Preferably, the data collecting phase must be administered unseen to the surveyor.

Acknowledgements

The authors are deeply thankful to the editor and the two learned referees for guiding towards the improvement of the earlier draft of this article.

[Recibido: enero de 2011 — Aceptado: septiembre de 2011]

References

Ahart, A. M. & Sackett, P. R. (2004), 'A new method for examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique', *Organizational Research Methods* **7**(1), 101–114.

Biemer, P. P. & Wright, D. (2004), Estimating cocaine use using the item count methodology, *in* 'Preliminary results form the national Survey on drug use and Health', Annual meeting of the American Association for Public Opinion research, Phoenix, Arizona.

Bjorner, J. B., Kosinski, M. & Ware, J. E. (2003), 'Using item response theory to calibrate the headache impact test (hit^{TM}) to the metric of traditional headache scales', *Quality of Life Research* **12**, 981–1002.

Bonetti, P. O., Waeckerlin, A., Schuepfer, G. & Frutiger, A. (2000), 'Improving time-sensitive processes in the intensive care unit: The example of 'door-to-needle time' in acute myocardial infarction', *International Journal for Quality in Health Care* **12**(4), 311–317.

Chaudhuri, A. & Christofides, T. C. (2007), 'Item count technique in estimating the proportion of people with a sensitive feature', *Journal of Statistical Planning and Inference* **137**(2), 589–593.

Chaudhuri, A. & Mukherjee, R. (1988), *Randomized Response: Theory and Methods*, Marcel-Decker, New York.

Dalton, D. R. & Metzger, M. (1992), 'Integrity testing for personal selection: An unsparing perspective', *Journal of Business Ethics* **12**, 147–156.

Dalton, D. R., Wimbush, J. C. & Daily, C. M. (1994), 'Using the unmatched count technique (uct) to estimate the base rates for sensitive behavior', *Personnel Psychology* **47**, 817–828.

Droitcour, J. A., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W. & Ezzati, T. M. (1991), The item count technique as a method of indirect questioning: A review of its development and a case study application, *in* P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz & S. Sudman, eds, 'Measurement Errors in Serveys', Wiley, New York.

Droitcour, J. A. & Larson, E. M. (2002), 'An innovative technique for asking sensitive questions: The three card method', *Sociological Methodological Bulletin* **75**, 5–23.

Droitcour, J. A., Larson, E. M. & Scheuren, F. J. (2001), The three card method: Estimating sensitive survey items with permanent anonymity of response, *in* 'Proceedings of the Social Statistics Section', American Statistical Association, Alexandria, Virginia.

Geurts, M. D. (1980), 'Using a randomized response design to eliminate nonresponse and response biases in business research', *Journal of the Academy of Marketing Science* **8**, 83–91.

Hubbard, M. L., Casper, R. A. & Lessler, J. T. (1989), Respondents' reactions to item count lists and randomized response, *in* 'Proceedings of the Survey Research Section', American Statistical Association, Washington, D. C., pp. 544–448.

Larkins, E. R., Hume, E. C. & Garcha, B. S. (1997), 'Validity of randomized response method in tax ethics research', *Journal of the Applied Business Research* **13**, 25–32.

Liu, P. T. & Chow, L. P. (1976), 'A new discrete quantitative randomized response model', *Journal of the American Statistical Association* **71**, 72–73.

Martin, M., Kosinski, M., Bjorner, J. B., Ware, J. E. & MacLean, R. (2007), 'Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale', *Quality of Life Research* **16**, 647–660.

Miller, J. D. (1985), 'The nominative technique: A new method of estimating heroin prevalence', *NIDA Research Monograph* **54**, 104–124.

Reinmuth, J. E. & Geurts, M. D. (1975), 'The collection of sensitive information using a two stage randomized response model', *Journal of Marketing Research* **12**, 402–407.

Tracy, D. & Mangat, N. (1996), 'Some development in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee', *Journal of Applied Statistical Science* **4**, 533–544.

Tsuchiya, T. (2005), 'Domain estimators for the item count technique', *Survey Methodology* **31**, 41–51.

Tsuchiya, T., Hirai, Y. & Ono, S. (2007), 'A study of the properties of the item count technique', *Public Opinion Quarterly* **71**, 253–272.

Warner, S. L. (1965), 'Randomized response: A survey technique for eliminating evasive answer bias', *Journal of the American Statistical Association* **60**, 63–69.

Wimbush, J. C. & Dalton, D. R. (1997), 'Base rate for employee theft: Convergence of multiple methods', *Journal of Applied Psychology* **82**, 756–763.

Appendix

To find the variance of the estimator $\widehat{\pi}_P$, consider

$$Z_i^2 = \sum_{j=1}^{g} \alpha_j^2 + \sum_{\substack{j,k=1 \\ j \neq k}}^{g} \alpha_j \alpha_k \tag{12}$$

After applying expectation operator on (12), we get:

$$\begin{aligned} E\left(Z_i^2\right) &= \sum_{j=1}^{g} E\left(\alpha_j^2\right) + \sum_{\substack{j,k=1 \\ j \neq k}}^{g} E\left(\alpha_j \alpha_k\right) \\ &= \sum_{j=1}^{g} \left(\pi + \theta_j - \pi\theta_j\right) + \sum_{\substack{j,k=1 \\ j \neq k}}^{g} \left\{\pi + \theta_j\theta_k\left(1-\pi\right)\right\} \\ &= \left(1-\pi\right)\sum_{j=1}^{g}\theta_j + g\pi + g\left(g-1\right)\pi + \left(1-\pi\right)\sum_{\substack{j,k=1 \\ j \neq k}}^{g}\theta_j\theta_k \end{aligned} \tag{13}$$

Now by definition of the variance of Z_i we have:

$$V\left(Z_i\right) = E\left(Z_i^2\right) - \left(E\left(Z_i\right)\right)^2 \tag{14}$$

Substituting (13) and $E\left(Z_i\right) = \left(g - \sum_{j=1}^{g}\theta_j\right)\pi + \sum_{j=1}^{g}\theta_j$ in (14), we get

$$\begin{aligned} V(Z_i t) &= \left(1-\pi\right)\sum_{j=1}^{g}\theta_j + g\pi + g(g-1)\pi + \left(1-\pi\right)\sum_{\substack{j,k=1 \\ j \neq k}}^{g}\theta_j\theta_k \\ &\quad - \left\{\left(g - \sum_{j=1}^{g}\theta_j\right)\pi + \sum_{j=1}^{g}\theta_j\right\}^2 \\ &= \left(1-\pi\right)\left[\left(g - \sum_{j=1}^{g}\theta_j\right)^2\pi + \sum_{j=1}^{g}\theta_j\left(1 - \sum_{j=1}^{g}\theta_j\right) + \sum_{\substack{j,k=1 \\ j \neq k}}^{g}\theta_j\theta_k\right] \end{aligned} \tag{15}$$

Now from (4) we have

$$V\left(\widehat{\pi}_P\right) = \frac{n^{-1}V\left(Z_i\right)}{\left(g - \sum_{j=1}^{g}\theta_j\right)^2} \tag{16}$$

Finally, using (15) in (16), we get the result in (5).

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 55 a 65

Measuring Degree of Departure from Extended Quasi-Symmetry for Square Contingency Tables

Medición del grado alejamiento del modelo extendido cuasi simétrico para tablas de contingencia cuadradas

Kouji Tahata[a], Keigo Kozai[b]

Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Chiba, Japan

Abstract

For square contingency tables with ordered categories, the present paper proposes a measure to represent the degree of departure from the extended quasi-symmetry (EQS) model. It is expressed by using the Cressie-Read power-divergence or Patil-Taillie diversity index. The present paper also defines the maximum departure from EQS which indicates the maximum departure from the uniformity of ratios of symmetric odds-ratios. The measure lies between 0 and 1, and it is useful for not only seeing the degree of departure from EQS in a table but also comparing it in several tables.

***Key words*:** Contingency table, Kullback-Leibler information, Quasi-symmetry, Shannon entropy.

Resumen

El presente artículo propone una medida para representar el grado de alejamiento del modelo extendido cuasisimétrico (EQS, por su sigla en inglés) para tablas de contingencia con categorías ordenadas. Esta medida se expresa mediante el uso de la divergencia de potencia de Cressie-Read o el índice de diversidad Patil-Taillie. Nuestro trabajo también define el máximo alejamiento de EQS, el cual indica el alejamiento máximo de la uniformidad de razones de odds-ratios simétricos. La medida cae entre 0 y 1 y es útil no solo para determinar el grado de alejamiento de EQS en una tabla, sino también para comparar este grado de alejamiento en varias tablas.

***Palabras clave*:** cuasi-simetría, entropía de Shannon, información de Kullback-Leibler, tablas de contingencia.

[a]Assistant professor. E-mail: kouji_tahata@is.noda.tus.ac.jp
[b]Graduate student. E-mail: keigo14@hotmail.co.jp

1. Introduction

Consider an $R \times R$ square contingency table with same row and column classifications. Let p_{ij} denote the probability that an observation will fall in the ith row and the jth column of the table ($i = 1, \ldots, R; j = 1, \ldots, R$). Bowker (1948) considered the symmetry (S) model defined by

$$p_{ij} = \phi_{ij} \quad \text{for} \quad i = 1, \ldots, R; j = 1, \ldots, R$$

where $\phi_{ij} = \phi_{ji}$ (Bishop, Fienberg & Holland 1975, p. 282). Caussinus (1965) considered the quasi-symmetry (QS) model defined by

$$p_{ij} = \alpha_i \beta_j \psi_{ij} \quad \text{for} \quad i = 1, \ldots, R; j = 1, \ldots, R$$

where $\psi_{ij} = \psi_{ji}$. A special case of this model obtained by putting $\{\alpha_i = \beta_i\}$ is the S model. For square tables with ordered categories, Tomizawa (1984) proposed the extended quasi-symmetry (EQS) model defined by

$$p_{ij} = \alpha_i \beta_j \psi_{ij} \quad \text{for} \quad i = 1, \ldots, R; j = 1, \ldots, R$$

where $\psi_{ij} = \gamma \psi_{ji}$ $(i < j)$. A special case of this model obtained by putting $\gamma = 1$ is the QS model. This is also expressed as, using the odds-ratios including the cell probabilities on the main diagonal,

$$\theta_{(i<j;j<k)} = \gamma \theta_{(j<k;i<j)} \quad \text{for} \quad i < j < k$$

where

$$\theta_{(i<j;j<k)} = \frac{p_{ij} p_{jk}}{p_{jj} p_{ik}}, \quad \theta_{(j<k;i<j)} = \frac{p_{ji} p_{kj}}{p_{ki} p_{jj}}$$

This indicates that the ratios of odds-ratios with respect to the main diagonal of the table are uniform for all $i < j < k$. The EQS model may be expressed as

$$D_{ijk} = \gamma D_{kji} \quad \text{for} \quad i < j < k,$$

where

$$D_{ijk} = p_{ij} p_{jk} p_{ki}, \quad D_{kji} = p_{kj} p_{ji} p_{ik}$$

For the analysis of square contingency tables, when a model does not hold, one may be interested in measuring how far the degree of departure from the model is. Thus some measures of various symmetry have been proposed. For example, Tomizawa (1994) and Tomizawa, Seo & Yamamoto (1998) proposed the measures to represent the degree of departure from the S model for square tables with *nominal* categories. Tomizawa, Miyamoto & Hatanaka (2001) proposed the measure for the S model for square tables with *ordered* categories. Tahata, Miyamoto & Tomizawa (2004) proposed the measure to represent the degree of departure from the QS model for square tables with *nominal* categories.

Generally, when the EQS model does not hold, we may apply a model which is extension of EQS model. Such models have been discussed by, e.g., Yamaguchi (1990), Tomizawa (1990) and Lawal (2004). On the other hand, we are also interested in measuring the degree of departure from the EQS model as described above. However a measure, which represents the degree of departure from the EQS model, does not exist. Therefore, we are interested in proposing a measure to represent the degree of departure from the EQS model, for square tables with *ordered* categories.

TABLE 1: Cross-classification of father and son social classes; taken from Hashimoto (2003, p. 142).

(a) Examined in 1955

	Son's class					
Father's class	(1)	(2)	(3)	(4)	(5)	Total
(1)	39	39	39	57	23	197
(2)	12	78	23	23	37	173
(3)	6	16	78	23	20	143
(4)	18	80	79	126	31	334
(5)	28	106	136	122	628	1020
Total	103	319	355	351	739	1867

(b) Examined in 1975

	Son's class					
Father's class	(1)	(2)	(3)	(4)	(5)	Total
(1)	29	43	25	31	4	132
(2)	23	159	89	38	14	323
(3)	11	69	184	34	10	308
(4)	42	147	148	184	17	538
(5)	42	176	377	114	298	1007
Total	147	594	823	401	343	2308

(c) Examined in 1995

	Son's class					
Father's class	(1)	(2)	(3)	(4)	(5)	Total
(1)	68	48	36	23	1	176
(2)	33	191	102	33	3	362
(3)	25	147	229	34	2	437
(4)	48	119	146	129	5	447
(5)	40	126	192	82	88	528
Total	214	631	705	301	99	1950

Consider the data in Table 1, taken from Hashimoto (2003, p. 142). These data describe the cross-classification of father and son social classes in Japan, which were examined in 1955, 1975, and 1995. Note that status (1) is Capitalist; (2) New-middle; (3) Working; (4) Self-employed; and (5) Farming. For social mobility data, one may be interested in considering the structure of symmetry instead of independence between row and column variables. Thus, for example the S, QS and EQS models would be useful for analyzing the data. For these data in Table 1, "$i \to j$" denotes the move to the son's class j from his father's class i. Thus $\{p_{ij}\}$ could be interpreted as transition probabilities. The EQS model indicates

that for a given order $i < j < k$, the product of transition probabilities that connects a cyclic sequence of paths $i \to j \to k \to i$ (we shall call the probability for *right cyclic sequence of paths* $i \to j \to k \to i$ for convenience), which includes two upward moves $i \to j$ and $j \to k$ and one downward move $k \to i$, is γ times higher than the product of transition probabilities that represents a reverse cyclic sequence of paths $i \to k \to j \to i$ (we shall call the probability for *left cyclic sequence of paths* $i \to k \to j \to i$), which includes one upward move $i \to k$ and two downward moves $k \to j$ and $j \to i$.

The EQS model can also be expressed as

$$D_{ijk}^{(1)} = D_{ijk}^{(2)} \quad \text{for} \quad i < j < k, \tag{1}$$

where

$$D_{ijk}^{(1)} = \frac{D_{ijk}}{\sum_{s<t<u} D_{stu}}, \quad D_{ijk}^{(2)} = \frac{D_{kji}}{\sum_{s<t<u} D_{uts}}$$

For the data in Tables 1a, 1b and 1c, $D_{ijk}^{(1)}$ is conditional probability that for any three father-son pairs father's class and his son's class are (i, j), (j, k) and (k, i), on condition that there is right cyclic sequence of paths. Similarly, $D_{ijk}^{(2)}$ is conditional probability that for any three father-son pairs father's class and his son's class are (j, i), (k, j) and (i, k), on condition that there is left cyclic sequence of paths. In a similar manner to Tomizawa et al. (1998), we shall consider a measure which represents the degree of departure from EQS because the equation (1) states that there is a structure of symmetry between $\{D_{ijk}^{(1)}\}$ and $\{D_{ijk}^{(2)}\}$ for $i < j < k$.

Section 2 proposes the measure to represent the degree of departure from the EQS model. Section 3 gives the approximate confidence interval for the measure. Section 4 shows an example.

2. Measure of Extended Quasi-Symmetry

Assume that $\sum_{s<t<u} D_{stu} \neq 0$, $\sum_{s<t<u} D_{uts} \neq 0$ and $D_{ijk} + D_{kji} > 0$ for $i < j < k$. Let

$$E_{ijk}^{(1)} = \frac{D_{ijk}^{(1)}}{D_{ijk}^{(1)} + D_{ijk}^{(2)}}, \quad E_{ijk}^{(2)} = \frac{D_{ijk}^{(2)}}{D_{ijk}^{(1)} + D_{ijk}^{(2)}} \quad \text{for} \quad i < j < k$$

For the data in Tables 1a, 1b and 1c, $E_{ijk}^{(1)}$ is the proportion of the conditional probability $D_{ijk}^{(1)}$ to the sum of the conditional probabilities $D_{ijk}^{(1)} + D_{ijk}^{(2)}$. Similarly, $E_{ijk}^{(2)}$ is the proportion of $D_{ijk}^{(2)}$ to $D_{ijk}^{(1)} + D_{ijk}^{(2)}$. The EQS model can be expressed as

$$E_{ijk}^{(1)} = E_{ijk}^{(2)} = \frac{1}{2} \quad \text{for} \quad i < j < k$$

Consider the measure defined by

$$\Phi^{(\lambda)} = \frac{\lambda(\lambda+1)}{2(2^{\lambda}-1)} \sum_{i<j<k} \left(D_{ijk}^{(1)} + D_{ijk}^{(2)} \right) I_{ijk}^{(\lambda)} \quad \text{for} \quad \lambda > -1$$

where

$$I_{ijk}^{(\lambda)} = \frac{1}{\lambda(\lambda+1)} \left[E_{ijk}^{(1)} \left\{ \left(\frac{E_{ijk}^{(1)}}{1/2} \right)^{\lambda} - 1 \right\} + E_{ijk}^{(2)} \left\{ \left(\frac{E_{ijk}^{(2)}}{1/2} \right)^{\lambda} - 1 \right\} \right]$$

and the value at $\lambda = 0$ is taken to be the limit as $\lambda \to 0$. Thus,

$$\Phi^{(0)} = \frac{1}{2(\log 2)} \sum_{i<j<k} \left(D_{ijk}^{(1)} + D_{ijk}^{(2)} \right) I_{ijk}^{(0)}$$

where

$$I_{ijk}^{(0)} = E_{ijk}^{(1)} \log \left(\frac{E_{ijk}^{(1)}}{1/2} \right) + E_{ijk}^{(2)} \log \left(\frac{E_{ijk}^{(2)}}{1/2} \right)$$

Note that a real value λ is chosen by the user. The $I_{ijk}^{(\lambda)}$ is the modified power-divergence and especially $I_{ijk}^{(0)}$ is the Kullback-Leibler information. For more details of the power-divergence, see Cressie & Read (1984). The measure $\Phi^{(\lambda)}$ would represent, essentially, the weighted sum of the power-divergence $I_{ijk}^{(\lambda)}$.

The measure may be expressed as

$$\Phi^{(\lambda)} = 1 - \frac{\lambda 2^{\lambda-1}}{2^{\lambda}-1} \sum_{i<j<k} \left(D_{ijk}^{(1)} + D_{ijk}^{(2)} \right) H_{ijk}^{(\lambda)} \quad \text{for} \quad \lambda > -1$$

where

$$H_{ijk}^{(\lambda)} = \frac{1}{\lambda} \left[1 - \left(E_{ijk}^{(1)} \right)^{\lambda+1} - \left(E_{ijk}^{(2)} \right)^{\lambda+1} \right]$$

with

$$\Phi^{(0)} = 1 - \frac{1}{2(\log 2)} \sum_{i<j<k} \left(D_{ijk}^{(1)} + D_{ijk}^{(2)} \right) H_{ijk}^{(0)}$$

where

$$H_{ijk}^{(0)} = -E_{ijk}^{(1)} \log E_{ijk}^{(1)} - E_{ijk}^{(2)} \log E_{ijk}^{(2)}$$

Note that $H_{ijk}^{(\lambda)}$ is the Patil & Taillie (1982) diversity index, which includes the Shannon entropy when $\lambda = 0$. Therefore, $\Phi^{(\lambda)}$ would represent one minus the weighted sum of the diversity index $H_{ijk}^{(\lambda)}$.

For each λ, the minimum value of $H_{ijk}^{(\lambda)}$ is 0 when $E_{ijk}^{(1)} = 0$ (then $E_{ijk}^{(2)} = 1$) or $E_{ijk}^{(2)} = 0$ (then $E_{ijk}^{(1)} = 1$), and the maximum value is $(2^{\lambda}-1)/\lambda 2^{\lambda}$ (if $\lambda \neq 0$), $\log 2$ (if $\lambda = 0$), when $E_{ijk}^{(1)} = E_{ijk}^{(2)}$. Thus we see that $\Phi^{(\lambda)}$ lies between 0 and 1. Also

for each λ, (i) there is a structure of EQS in the table (i.e., $E_{ijk}^{(1)} = E_{ijk}^{(2)} = 1/2$, (thus $D_{ijk}^{(1)} = D_{ijk}^{(2)}$) for any $i < j < k$) if and only if $\Phi^{(\lambda)} = 0$; and (ii) the degree of departure from EQS is the largest, in the sense that $E_{ijk}^{(1)} = 0$ (then $E_{ijk}^{(2)} = 1$) or $E_{ijk}^{(2)} = 0$ (then $E_{ijk}^{(1)} = 1$) (i.e., $D_{ijk}^{(1)} = 0$ (then $D_{ijk}^{(2)} > 0$) or $D_{ijk}^{(2)} = 0$ (then $D_{ijk}^{(1)} > 0$)) for any $i < j < k$, if and only if $\Phi^{(\lambda)} = 1$. Note that $\Phi^{(\lambda)} = 1$ indicates that $D_{ijk}^{(1)}/D_{ijk}^{(2)} = \infty$ for some $i < j < k$ and $D_{ijk}^{(1)}/D_{ijk}^{(2)} = 0$ for the other $i < j < k$, and therefore it seems appropriate to consider that then the degree of departure from EQS (i.e., from $D_{ijk}^{(1)}/D_{ijk}^{(2)} = 1$ for $i < j < k$) is largest.

According to the weighted sum of power-divergence or the weighted sum of Patil-Taillie diversity index, $\Phi^{(\lambda)}$ represents the degree of departure from EQS, and the degree increases as the value of $\Phi^{(\lambda)}$ increases.

3. Approximate Confidence Interval for Measure

Let n_{ij} denote the observed frequency in the ith row and jth column of the table ($i = 1, \ldots, R; j = 1, \ldots, R$) with $n = \sum\sum n_{ij}$. Assume that $\{n_{ij}\}$ have a multinomial distribution. We shall consider an approximate standard error and large-sample confidence interval for the measure $\Phi^{(\lambda)}$ using the delta method as described by Bishop et al. (1975, Section 14.6). The sample version of $\Phi^{(\lambda)}$, i.e., $\widehat{\Phi}^{(\lambda)}$, is given by $\Phi^{(\lambda)}$ with $\{p_{ij}\}$ replaced by $\{\widehat{p}_{ij}\}$, where $\widehat{p}_{ij} = n_{ij}/n$. Using the delta method, $\sqrt{n}(\widehat{\Phi}^{(\lambda)} - \Phi^{(\lambda)})$ has asymptotically (as $n \to \infty$) a normal distribution with mean zero and variance

$$\sigma^2 = \sum_{a=1}^{R-1} \sum_{b=a+1}^{R} \left\{ \frac{1}{p_{ab}} \left(A_{ab}^{(\lambda)} \right)^2 + \frac{1}{p_{ba}} \left(B_{ab}^{(\lambda)} \right)^2 \right\} - \left\{ \sum_{a=1}^{R-1} \sum_{b=a+1}^{R} \left(A_{ab}^{(\lambda)} + B_{ab}^{(\lambda)} \right) \right\}^2$$

where for $\lambda > -1$ and $\lambda \neq 0$,

$$\begin{aligned} A_{ab}^{(\lambda)} = \frac{2^{\lambda-1}}{2^\lambda - 1} \sum_{i<j<k} \Big[& (E_{ijk}^{(1)})^\lambda D_{ijk}^{(1)} \Big\{ I_{(a=i,b=j)} + I_{(a=j,b=k)} \\ & - \sum_{s<t<u} D_{stu}^{(1)} (I_{(a=s,b=t)} + I_{(a=t,b=u)}) \Big\} \\ & + (E_{ijk}^{(2)})^\lambda D_{ijk}^{(2)} \Big\{ I_{(a=i,b=k)} - \sum_{s<t<u} D_{stu}^{(2)} I_{(a=s,b=u)} \Big\} \\ & + \lambda \Big(D_{ijk}^{(2)} (E_{ijk}^{(1)})^{\lambda+1} - D_{ijk}^{(1)} (E_{ijk}^{(2)})^{\lambda+1} \Big) \Big\{ \Big(I_{(a=i,b=j)} + I_{(a=j,b=k)} - I_{(a=i,b=k)} \\ & - \sum_{s<t<u} (D_{stu}^{(1)} I_{(a=s,b=t)} + D_{stu}^{(1)} I_{(a=t,b=u)} - D_{stu}^{(2)} I_{(a=s,b=u)}) \Big) \Big\} \Big] \end{aligned}$$

and

$$A_{ab}^{(0)} = \frac{1}{2\log 2} \sum_{i<j<k} \Big[D_{ijk}^{(1)} (\log E_{ijk}^{(1)}) \Big\{ I_{(a=i,b=j)} + I_{(a=j,b=k)} - \sum_{s<t<u} D_{stu}^{(1)} (I_{(a=s,b=t)} + I_{(a=t,b=u)}) \Big\} + D_{ijk}^{(2)} (\log E_{ijk}^{(2)}) \Big\{ I_{(a=i,b=k)} - \sum_{s<t<u} D_{stu}^{(2)} I_{(a=s,b=u)} \Big\} \Big]$$

with

$$I_{(a=i,b=j)} = \begin{cases} 1 & (a = i \quad \text{and} \quad b = j) \\ 0 & (\text{otherwise}) \end{cases}$$

and where $B_{ab}^{(\lambda)}$ for $\lambda > -1$ is defined as $A_{ab}^{(\lambda)}$ obtained by interchanging $D_{ijk}^{(1)}$ and $D_{ijk}^{(2)}$ and by interchanging $E_{ijk}^{(1)}$ and $E_{ijk}^{(2)}$.

Although the detail is omitted, (i) when $\Phi^{(\lambda)} = 0$, we can get $\sigma^2 = 0$ by noting $D_{ijk}^{(1)} = D_{ijk}^{(2)}$ and $E_{ijk}^{(1)} = E_{ijk}^{(2)} = 1/2$ for $i < j < k$, and (ii) when $\Phi^{(\lambda)} = 1$, we can get $\sigma^2 = 0$ by noting $D_{ijk}^{(1)} = 0$ (then $E_{ijk}^{(1)} = 0$ and $E_{ijk}^{(2)} = 1$) for some $i < j < k$ and $D_{ijk}^{(2)} = 0$ (then $E_{ijk}^{(1)} = 1$ and $E_{ijk}^{(2)} = 0$) for the other $i < j < k$. Thus we note that the asymptotic distribution of $\widehat{\Phi}^{(\lambda)}$ is not applicable when $\Phi^{(\lambda)} = 0$ and $\Phi^{(\lambda)} = 1$. Let $\widehat{\sigma}^2$ denote σ^2 with $\{p_{ij}\}$ replaced by $\{\widehat{p}_{ij}\}$. Then $\widehat{\sigma}/\sqrt{n}$ is an estimated approximate standard error for $\widehat{\Phi}^{(\lambda)}$.

4. An Example

Consider the data in Table 1 again. Then, the *maximum* departure from the EQS model indicates that for some $i < j < k$, the product of transition probabilities that connects $i \to j \to k \to i$ is zero, (and then the product of transition probabilities that represents $i \to k \to j \to i$ is not zero) and for the others the product of transition probabilities that connects $i \to j \to k \to i$ is not zero (and then the product of transition probabilities that represents $i \to k \to j \to i$ is zero); namely, the *stochastic circular* social mobility arises among any three father-son pairs.

Now we consider comparing the degree of departure from the EQS model for the data in Tables 1a, 1b and 1c. We choose $\lambda = 0$ because $\Phi^{(0)}$ is expressed as well known Kullback-Leibler information. Thus we apply the measure $\Phi^{(0)}$ for these data. Table 2 shows the estimated measure $\widehat{\Phi}^{(0)}$, estimated approximate standard error for $\widehat{\Phi}^{(0)}$, and approximate 95% confidence interval for $\Phi^{(0)}$. When the degrees of departure from the EQS model in Tables 1a, 1b and 1c are compared using the estimated measure $\widehat{\Phi}^{(0)}$, (i) the value of $\widehat{\Phi}^{(0)}$ is greater for Table 1a than for Tables 1b and 1c, and (ii) the value of $\widehat{\Phi}^{(0)}$ is greater for Table 1b than for Table 1c. Namely, the degree of departure from the EQS model for Table 1a is the largest, that for Table 1b is the second largest, and that for Table 1c is the

smallest. Thus, the data in Table 1a rather than in Tables 1b and 1c are estimated to be close to the *maximum* departure from the EQS model.

TABLE 2: Estimated measure $\widehat{\Phi}^{(0)}$, estimated approximate standard error for $\widehat{\Phi}^{(0)}$, and approximate 95% confidence interval for $\Phi^{(0)}$, applied to Tables 1a, 1b, and 1c.

Table	Estimated measure	Standard error	Confidence interval
1a	0.076	0.039	(−0.001, 0.153)
1b	0.036	0.034	(−0.031, 0.102)
1c	0.011	0.018	(−0.024, 0.046)

5. Discussions and Conclusion

The measure $\Phi^{(\lambda)}$ always ranges between 0 and 1 independently of the dimension R and sample size n. But the likelihood-ratio statistic for testing goodness-of-fit of the EQS model depends on sample size n. For example, consider two $R \times R$ contingency tables, say, A and B, where the observed frequency in each cell for Table A has ten times that in the corresponding cell for table B. Then the value of likelihood-ratio statistic for testing goodness-of-fit of the EQS model for table A is ten times that for table B. However, when the ratios of odds-ratios, $\widehat{\theta}_{(i<j;j<k)}/\widehat{\theta}_{(j<k;i<j)}$, $i < j < k$, for table A is equal to that for table B, the value of measure $\widehat{\Phi}^{(\lambda)}$ for table A is equal to that for table B. Therefore, $\widehat{\Phi}^{(\lambda)}$ would be useful for comparing the degree of departure from EQS in several tables, even if several tables have different sample sizes.

As described in Section 2, the proposed measure would be useful when we want to see with single summary measure how degree the departure from EQS is toward the maximum degree of departure from EQS. We have defined the maximum degree of departure from EQS, namely, $D^{(1)}_{ijk}/D^{(2)}_{ijk} = \infty$ for some $i < j < k$ and $D^{(1)}_{ijk}/D^{(2)}_{ijk} = 0$ for the other $i < j < k$. This seems natural as the definition of the maximum departure from EQS that indicates $D^{(1)}_{ijk}/D^{(2)}_{ijk} = 1$ for $i < j < k$.

TABLE 3: Values of power-divergence test statistic $W^{(\lambda)}$ (with 5 degrees of freedom), applied to Tables 1a, 1b, and 1c.

λ	For Table 1a	For Table 1b	For Table 1c
−0.4	13.70	4.63	1.62
0.0	13.59	4.66	1.60
0.6	13.48	4.73	1.56
1.0	13.43	4.79	1.55
1.4	13.40	4.86	1.53

TABLE 4: Artificial data (n is sample size).

(a) $n = 700$

30	81	79	120
10	39	83	16
13	20	38	31
7	35	77	21

(b) $n = 668$

30	29	60	10
110	39	33	36
21	42	38	61
15	61	62	21

TABLE 5: Values of $\widehat{\Phi}^{(\lambda)}$, the test statistic $W^{(\lambda)}$ and $W^{(\lambda)}/n$ applied to Tables 4a and 4b.

(a) Values of $\widehat{\Phi}^{(\lambda)}$

λ	For Table 4a	For Table 4b
-0.4	0.268	0.225
0.0	0.363	0.304
0.6	0.436	0.364
1.0	0.456	0.381
1.4	0.463	0.387

(b) Values of $W^{(\lambda)}$

λ	For Table 4a	For Table 4b
-0.4	27.76	52.90
0.0	28.33	51.95
0.6	30.13	51.03
1.0	32.12	50.72
1.4	34.92	50.64

(c) Values of $W^{(\lambda)}/n$

λ	For Table 4a	For Table 4b
-0.4	0.040	0.079
0.0	0.040	0.078
0.6	0.043	0.076
1.0	0.046	0.076
1.4	0.050	0.076

Consider the data in Table 1, again. Cressie & Read (1984) proposed the power-divergence test statistic for testing goodness-of-fit of a model. Denote the power-divergence statistic for testing goodness-of-fit of the EQS model with $R(R-3)/2$

degrees of freedom by $W^{(\lambda)}$. Table 3 gives the values of $W^{(\lambda)}$ applied to the data in Tables 1a, 1b and 1c. The EQS model fits the data in Table 1a poorly; however, fits the data in Tables 1b and 1c well. This is similar to the results described in Section 4. Then, it may seem to many readers that $W^{(\lambda)}/n$ (for a given λ) is also a reasonable measure for representing the degree of departure from EQS. However, we point out that $W^{(\lambda)}$ can not measure the degree of departure from EQS toward the maximum degree of departure from EQS that is defined in Section 2, although $W^{(\lambda)}$ can test the goodness-of-fit of the EQS model. For example, consider the artificial data in Tables 4a and 4b. From Table 5, the value of $W^{(\lambda)}/n$ ($W^{(\lambda)}$) is less for Table 4a than for Table 4b; however, the value of $\widehat{\Phi}^{(\lambda)}$ is greater for Table 4a than for Table 4b. When we want to measure the degree of departure from EQS toward the maximum departure from the uniformity of ratios of symmetric odds-ratios (i.e., the maximum departure from EQS), the measure $\Phi^{(\lambda)}$ rather than $W^{(\lambda)}$ may be appropriate. Also, $W^{(\lambda)}$ rather than $\Phi^{(\lambda)}$ would be appropriate to test the goodness-of-fit of the EQS model.

As described in Section 1, Lawal (2004), Tomizawa (1990) and Yamaguchi (1990) considered the extension of EQS model. For testing goodness-of-fit of the EQS model under the assumption that the extension of EQS model holds true, the difference between the likelihood ratio statistic for the EQS and extension of EQS models has an asymptotic chi-squared distribution with degrees of freedom equal to the difference between degrees of freedom for two models. This statistic, which is useful for comparing pairs of models, is well known. So, the readers may consider that this statistic is also a reasonable measure for representing the degree of departure from EQS. However, since this statistic can not measure the degree of departure from EQS toward the maximum departure from EQS, $\Phi^{(\lambda)}$ rather than it would be preferable when we want to measure the degree of departure from EQS toward the maximum degree of departure from EQS.

We observe that the EQS model and the measure $\Phi^{(\lambda)}$ should be applied to square tables with *ordered* categories because it is not invariant under the arbitrary similar permutations of row and column categories.

Acknowledgments

We would like to thank the anonymous referees for their helpful comments and suggestions.

[Recibido: marzo de 2011 — Aceptado: septiembre de 2011]

References

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge, Massachusetts.

Bowker, A. H. (1948), 'A test for symmetry in contingency tables', *Journal of the American Statistical Association* **43**, 572–574.

Caussinus, H. (1965), 'Contribution à l'analyse statistique des tableaux de corrélation', *Annales de la Faculté des Sciences de l'Université de Toulouse* **29**, 77–182.

Cressie, N. A. C. & Read, T. R. C. (1984), 'Multinomial goodness-of-fit tests', *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

Hashimoto, K. (2003), *Class Structure in Contemporary Japan*, Trans Pacific Press, Melbourne.

Lawal, H. B. (2004), 'Using a GLM to decompose the symmetry model in square contingency tables with ordered categories', *Journal of Applied Statistics* **31**, 279–303.

Patil, G. P. & Taillie, C. (1982), 'Diversity as a concept and its measurement', *Journal of the American Statistical Association* **77**, 548–561.

Tahata, K., Miyamoto, N. & Tomizawa, S. (2004), 'Measure of departure from quasi-symmetry and Bradley-Terry models for square contingency tables with nominal categories', *Journal of the Korean Statistical Society* **33**, 129–147.

Tomizawa, S. (1984), 'Three kinds of decompositions for the conditional symmetry model in a square contingency table', *Journal of the Japan Statistical Society* **14**, 35–42.

Tomizawa, S. (1990), 'Quasi-diagonals-parameter symmetry model for square contingency tables with ordered categories', *Calcutta Statistical Association Bulletin* **39**, 53–61.

Tomizawa, S. (1994), 'Two kinds of measures of departure from symmetry in square contingency tables having nominal categories', *Statistica Sinica* **4**, 325–334.

Tomizawa, S., Miyamoto, N. & Hatanaka, Y. (2001), 'Measure of asymmetry for square contingency tables having ordered categories', *Australian and New Zealand Journal of Statistics* **43**, 335–349.

Tomizawa, S., Seo, T. & Yamamoto, H. (1998), 'Power-divergence-type measure of departure from symmetry for square contingency tables that have nominal categories', *Journal of Applied Statistics* **25**, 387–398.

Yamaguchi, K. (1990), 'Some models for the analysis of asymmetric association in square contingency tables with ordered categories', *Sociological Methodology* **20**, 181–212.

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 67 a 76

Estimation of Reliability in Multicomponent Stress-strength Based on Generalized Exponential Distribution

Estimación de confiabilidad en la resistencia al estrés de multicomponentes basado en la distribución exponencial generalizada

Gadde Srinivasa Rao[a]

Department of Statistics, School of Mathematical and Computer Sciences, Dilla University, Dilla, Ethiopia

Abstract

A multicomponent system of k components having strengths following k-independently and identically distributed random variables $X_1, X_2, \ldots, X_k$ and each component experiencing a random stress Y is considered. The system is regarded as alive only if at least s out of k ($s < k$) strengths exceed the stress. The reliability of such a system is obtained when strength and stress variates are given by generalized exponential distribution with different shape parameters. The reliability is estimated using ML method of estimation in samples drawn from strength and stress distributions. The reliability estimators are compared asymptotically. The small sample comparison of the reliability estimates is made through Monte Carlo simulation. Using real data sets we illustrate the procedure.

***Key words*:** Asymptotic confidence interval, Maximum likelihood estimation, Reliability, Stress-strength model.

Resumen

Se considera un sistema de k multicomponentes que tiene resistencias que se distribuyen como k variables aleatorias independientes e idénticamente distribuidas $X_1, X_2, \ldots, X_k$ y cada componente experimenta un estrés aleatorio Y. El sistema se considera como vivo si y solo si por lo menos s de $k(s < k)$ resistencias exceden el estrés. La confiabilidad de este sistema se obtiene cuando las resistencias y el estrés se distribuyen como una distribución exponencial generalizada con diferentes parámetros de forma. La confiabilidad es estimada usando el método ML de estimación en muestras extraídas tanto para distribuciones de resistencia como de estrés. Los estimadores de confiabilidad son comparados asintóticamente. La comparación

[a]Professor. E-mail: gaddesrao@yahoo.com

para muestras pequeñas de los estimadores de confiabilidad se hace a través de simulaciones Monte Carlo. El procedimiento también se ilustra mediante una aplicación con datos reales.

***Palabras clave*:** confiabilidad, estimación máximo verosímil, intervalos de confianza asintóticos, modelo de resistencia-estrés.

1. Introduction

The two-parameter generalized exponential distribution (GE) has been introduced and studied quite extensively by Gupta & Kundu (1999, 2001, 2002). The two-parameter GE distribution is an alternative to the well known two-parameter gamma, two-parameter Weibull or two parameter log-normal distributions. The two-parameter GE distribution has the following density function and the distribution function, respectively

$$f(x;\alpha,\lambda) = \alpha\lambda e^{-x\lambda}(1-e^{-x\lambda})^{\alpha-1}; \quad \text{for } x > 0 \tag{1}$$

$$F(x;\alpha,\lambda) = (1-e^{-x\lambda})^{\alpha-1}; \quad \text{for } x > 0 \tag{2}$$

Here α and λ are the shape and scale parameters, respectively. Now onwards GE distribution with the shape parameter α and scale parameter λ will be denoted by GE(α, λ).

The purpose of this paper is to study the reliability in a multicomponent stress-strength based on X, Y being two independent random variables, where $X \sim$ GE(α, λ) and $Y \sim$ GE(β, λ).

Let the random samples $Y, X_1, X_2, \ldots, X_k$ being independent, $G(y)$ be the continuous distribution function of Y and $F(x)$ be the common continuous distribution function of $X_1, X_2, \ldots, X_k$. The reliability in a multicomponent stress-strength model developed by Bhattacharyya & Johnson (1974) is given by

$$\begin{aligned} R_{s,k} &= P[\text{at least } s \text{ of the } X_1, X_2, \ldots, X_k \quad \text{exceed} \quad Y] \\ &= \sum_{i=s}^{k} \binom{k}{i} \int_{-\infty}^{\infty} [1-F(y)]^i \, [F(y)]^{(k-i)} \, dG(y) \end{aligned} \tag{3}$$

Where $X_1, X_2, \ldots, X_k$ are independently identically distributed (iid) with common distribution function $F(x)$, this system is subjected to common random stress Y. The probability in (3) is called reliability in a multicomponent stress-strength model (Bhattacharyya & Johnson 1974). The survival probability of a single component stress-strength version has been considered by several authors assuming various lifetime distributions for the stress-strength random variates, e.g. Enis & Geisser (1971), Downtown (1973), Awad & Gharraf (1986), McCool (1991), Nandi & Aich (1994), Surles & Padgett (1998), Raqab & Kundu (2005), Kundu & Gupta (2005), Kundu & Gupta (2006), Raqab, Modi & Kundu (2008), Kundu & Raqab (2009). The reliability in a multicomponent stress-strength was developed

by Bhattacharyya & Johnson (1974), Pandey & Uddin (1985), and the references therein cover the study of estimating in many standard distributions assigned to one or both stress, strength variates. Recently, Rao & Kantam (2010) studied estimation of reliability in multicomponent stress-strength for the log-logistic distribution.

Suppose that a system, with k identical components, functions if $s(1 \leq s \leq k)$ or more of the components simultaneously operate. In this operating environment, the system is subjected to a stress Y which is a random variable with distribution function $G(.)$. The strengths of the components, that is the minimum stress to cause failure, are independent and identically distributed random variables with distribution function $F(.)$. Then, the system reliability, which is the probability that the system does not fail, is the function $R_{s,k}$ given in (3). The estimation of the survival probability in a multicomponent stress-strength system when the stress follows a two-parameter GE distribution has not received much attention in the literature. Therefore, an attempt is made here to study the estimation of reliability in multicomponent stress-strength model with reference to the two-parameter GE probability distribution. In Section 2, we derive the expression for $R_{s,k}$ and develop a procedure for estimating it. More specifically, we obtain the maximum likelihood estimates of the parameters. The Maximum Likelihood Estimators (MLEs) are employed to obtain the asymptotic distribution and confidence intervals for $R_{s,k}$. The small sample comparisons are made through Monte Carlo simulations in Section 3. Also, using real data, we illustrate the estimation process. Finally, some conclusion and comments are provided in Section 4.

2. Maximum Likelihood Estimator of $R_{s,k}$

Let $X \sim \text{GE}(\alpha,\ \lambda)$ and $Y \sim \text{GE}(\beta,\ \lambda)$ with unknown shape parameters α and β and common scale parameter λ, where X and Y are independently distributed. The reliability in multicomponent stress-strength for two-parameter GE distribution using (3) is

$$\begin{aligned}
R_{s,k} &= \sum_{i=s}^{k} \binom{k}{i} \int_0^\infty [1-(1-e^{-y\lambda})^\alpha]^i [(1-e^{-y\lambda})^\alpha]^{(k-i)}\ \beta\lambda e^{-y\lambda}(1-e^{-y\lambda})^{\beta-1}\ dy \\
&= \sum_{i=s}^{k} \binom{k}{i} \int_0^1 [1-t^\nu]^i\ [t^\nu]^{(k-i)}\ dt, \quad \text{where} \quad t = (1-e^{-y\lambda})^\beta \quad \text{and} \quad \nu = \frac{\alpha}{\beta} \\
&= \frac{1}{\nu} \sum_{i=s}^{k} \binom{k}{i} \int_0^1 [1-z]^i\ [z]^{(k-i+\frac{1}{\nu}-1)}\ dz \quad \text{if} \quad z = t^\nu \\
&= \frac{1}{\nu} \sum_{i=s}^{k} \beta(k-i+\frac{1}{\nu}, i+1)
\end{aligned}$$

After the simplification we get

$$R_{s,k} = \frac{1}{\nu}\sum_{i=s}^{k}\frac{k!}{(k-i)!}\left[\prod_{j=0}^{i}(k+\frac{1}{\nu}-j)\right]^{-1}, \quad \text{since } k \text{ and } i \text{ are integers} \tag{4}$$

The probability in (4) is called reliability in a multicomponent stress-strength model. If α and β are not known, it is necessary to estimate α and β to estimate $R_{s,k}$. In this paper we estimate α and β by the ML method. Once MLEs are obtained then $R_{s,k}$ can be computed using equation (4).

Let $X_1, X_2, \ldots, X_n$; $Y_1, Y_2, \ldots, Y_m$ be two ordered random samples of size n, m, respectively, on strength, stress variates following a GE distribution with shape parameters α and β and a common scale parameter λ. The log-likelihood function of the observed sample is

$$\begin{aligned} L(\alpha,\beta,\lambda) = (m+n)\ln\lambda + n\ln\alpha + m\ln\beta - \lambda\left[\sum_{i=1}^{n}x_i - \sum_{j=1}^{m}y_j\right] + \\ (\alpha-1)\sum_{i=1}^{n}\ln(1-e^{-x_i\lambda}) + (\beta-1)\sum_{j=1}^{m}\ln(1-e^{-y_j\lambda}) \end{aligned} \tag{5}$$

The MLEs of α, β and λ, say $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\lambda}$, respectively, can be obtained as the solution of

$$\widehat{\alpha} = \frac{-n}{\sum_{i=1}^{n} ln(1-e^{-x_i\lambda})} \tag{6}$$

$$\widehat{\beta} = \frac{-m}{\sum_{j=1}^{m} ln(1-e^{-y_j\lambda})} \tag{7}$$

$$\begin{aligned} g(\lambda) = 0 \Rightarrow \frac{m+n}{\lambda} - \frac{n\sum_{i=1}^{n}\frac{x_i\, e^{-x_i\lambda}}{1-e^{-x_i\lambda}}}{\sum_{k=1}^{n} ln(1-e^{-x_k\lambda})} - \frac{m\sum_{j=1}^{m}\frac{y_j\, e^{-y_j\lambda}}{1-e^{-y_j\lambda}}}{\sum_{k=1}^{m} ln(1-e^{-y_k\lambda})} \\ - \sum_{i=1}^{n}\frac{x_i}{1-e^{-x_i\lambda}} - \sum_{j=1}^{m}\frac{y_j}{1-e^{-y_j\lambda}} \end{aligned} \tag{8}$$

Therefore, $\widehat{\lambda}$ is a simple iterative solution of the non-linear equation $g(\lambda) = 0$. Once we obtain $\widehat{\lambda}$; $\widehat{\alpha}$ and $\widehat{\beta}$ can be obtained from (6) and (7), respectively. Therefore, the MLE of $R_{s,k}$ becomes

$$\widehat{R}_{s,k} = \frac{1}{\widehat{\nu}}\sum_{i=s}^{k}\frac{k!}{(k-i)!}\left[\prod_{j=0}^{i}(k+\frac{1}{\widehat{\nu}}-j)\right]^{-1}, \quad \text{where} \quad \widehat{\nu} = \frac{\widehat{\alpha}}{\widehat{\beta}} \tag{9}$$

To obtain the asymptotic confidence interval for $R_{s,k}$, we proceed as below: The asymptotic variance of the MLE is given by

$$V(\widehat{\alpha}) = \left[E(-\frac{\partial^2 L}{\partial \alpha^2})\right] = \frac{\alpha^2}{n} \quad \text{and} \quad V(\widehat{\beta}) = \left[E(-\frac{\partial^2 L}{\partial \beta^2})\right] = \frac{\beta^2}{n} \tag{10}$$

The asymptotic variance (AV) of an estimate of $R_{s,k}$ which is a function of two independent statistics $\widehat{\alpha}$ and $\widehat{\beta}$ is given by Rao (1973).

$$AV(\widehat{R}_{s,k}) = V(\widehat{\alpha}) \left[\frac{\partial R_{s,k}}{\partial \alpha}\right]^2 + V(\widehat{\beta}) \left[\frac{\partial R_{s,k}}{\partial \beta}\right]^2 \tag{11}$$

From the asymptotic optimum properties of MLEs (Kendall & Stuart 1979) and of linear unbiased estimators (David 1981), we know that MLEs are asymptotically equally efficient having the Cramer-Rao lower bound as their asymptotic variance, as given in (10). Thus, from equation (11), the asymptotic variance of $\widehat{R}_{s,k}$ can be obtained. To avoid the difficulty of the derivation of the $R_{s,k}$, we obtain the derivatives of $R_{s,k}$ for (s,k)=(1,3) and (2,4) separately and they are given by

$$\frac{\partial R_{1,3}}{\partial \alpha} = \frac{3}{\beta\left(3\widehat{\nu}+1\right)^2} \quad \text{and} \quad \frac{\partial R_{1,3}}{\partial \beta} = \frac{-3\widehat{\nu}}{\beta\left(3\widehat{\nu}+1\right)^2}$$

$$\frac{\partial R_{2,4}}{\partial \alpha} = \frac{12\widehat{\nu}(7\widehat{\nu}+2)}{\beta\left[(3\widehat{\nu}+1)(4\widehat{\nu}+1)\right]^2} \quad \text{and} \quad \frac{\partial R_{2,4}}{\partial \beta} = \frac{-12\widehat{\nu}^2(7\widehat{\nu}+2)}{\beta\left[(3\widehat{\nu}+1)(4\widehat{\nu}+1)\right]^2}$$

Thus $AV(\widehat{R}_{1,3}) = \frac{9\widehat{\nu}^2}{(3\widehat{\nu}+1)^4}\left(\frac{1}{n}+\frac{1}{m}\right)$

$$AV(\widehat{R}_{2,4}) = \frac{144\widehat{\nu}^4(7\widehat{\nu}+2)^2}{\left[(3\widehat{\nu}+1)(4\widehat{\nu}+1)\right]^4}\left(\frac{1}{n}+\frac{1}{m}\right)$$

as $n \to \infty, m \to \infty, \frac{\widehat{R}_{s,k}-R_{s,k}}{AV(\widehat{R}_{s,k})} \xrightarrow{d} N(0,1)$ and the asymptotic confidence 95% confidence interval for $R_{s,k}$ is given by

$$\widehat{R}_{s,k} \pm 1.96\sqrt{AV(\widehat{R}_{s,k})}$$

The asymptotic confidence 95% confidence interval for $R_{1,3}$ is given by

$$\widehat{R}_{1,3} \pm 1.96\frac{3\widehat{\nu}}{\left(3\widehat{\nu}+1\right)^2}\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}, \quad \text{where} \quad \widehat{\nu} = \frac{\widehat{\alpha}}{\widehat{\beta}}$$

The asymptotic confidence 95% confidence interval for $R_{2,4}$ is given by

$$\widehat{R}_{2,4} \pm 1.96\frac{12\widehat{\nu}^2(7\widehat{\nu}+2)}{\left[(3\widehat{\nu}+1)(4\widehat{\nu}+1)\right]^2}\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}, \quad \text{where} \quad \widehat{\nu} = \frac{\widehat{\alpha}}{\widehat{\beta}}$$

3. Simulation Study and Data Analysis

3.1. Simulation Study

In this subsection we present some results based on Monte Carlo simulations to compare the performance of the $R_{s,k}$ using different sample sizes. 3,000 random samples of size 10(5)35 each from stress population, strength population are generated for $(\alpha, \beta) = (3.0, 1.5)$, (2.5,1.5), (2.0,1.5), (1.5,1.5), (1.5,2.0),(1.5,2.5) and (1.5,3.0) in line with Bhattacharyya & Johnson (1974). The MLE of scale parameter λ is estimated by the iterative method, and the using λ the shape parameters α and β are estimated from (6) and (7). These ML estimators of α and β are then substituted in ν to get the reliability in a multicomponent reliability for $(s, k) = (1, 3), (2, 4)$. The average bias and average mean square error (MSE) of the reliability estimates over the 3000 replications are given in Tables 1 and 2. Average confidence length and coverage probability of the simulated 95% confidence intervals of $R_{s,k}$ are given in Tables 3 and 4. The true values of reliability in multicomponent stress-strength with the given combinations for $(s, k) = (1, 3)$ are 0.857, 0.833, 0.800, 0.750, 0.692, 0.643, 0.600, and for $(s, k) = (2, 4)$ are 0.762, 0.725, 0.674, 0.600, 0.519, 0.454, and 0.400. Thus, the true value of reliability in multicomponent stress-strength model decreases as β increases for a fixed α whereas reliability in multicomponent stress-strength increases as increases for a fixed β in both the cases (s, k). Therefore, the true value of reliability decreases as ν decreases, and *vice versa*. The average bias and average MSE decrease as sample size increases for both methods of estimation in reliability. Also the bias is negative in both situations of (s, k). It verifies the consistency property of the MLE of $R_{s,k}$. Whereas, among the parameters the absolute bias and MSE decrease as α increases for a fixed β in both cases of (s, k) and the absolute bias and MSE increase as β increases for a fixed α in both the cases of (s, k). The length of the confidence interval also decreases as the sample size increases. The coverage probability is close to the nominal value in all cases but slightly less than 0.95. Overall, the performance of the confidence interval is quite good for all combinations of parameters. Whereas, among the parameters we observed the same phenomenon for average length and average coverage probability that we observed in the case of average bias and MSE.

3.2. Data Analysis

In this subsection we analyze two real data sets and demonstrate how the proposed methods can be used in practice. The first data set is reported by Lawless (1982) and the second one is given by Linhardt & Zucchini (1986). Both are analyzed and fitted for various lifetime distributions. We fit the generalized exponential distribution to the two data sets separately. The first data set (Lawless 1982, p. 228) presented here arose in tests on endurance of deep groove ball bearings. The data presented are the number of million revolutions before failure for each of the 23 ball bearings in the life test, and they are: 17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12,

98.64, 105.12, 105.84, 127.92, 128.04, and 173.40. Gupta & Kundu (2001) studied the validity of the model and they compute the Kolmogorov-Smirnov (KS) distance between the empirical distribution function and the fitted distribution functions of generalized exponential distribution which is 0.1058 with a corresponding p-value of 0.9592.

Table 1: Average bias of the simulated estimates of $R_{s,k}$.

		(α,β)						
(s,k)	(n,m)	(3.0,1.5)	(2.5,1.5)	(2.0,1.5)	(1.5,1.5)	(1.5,2.0)	(1.5,2.5)	(1.5,3.0)
	(10,10)	−0.0029	−0.0047	−0.0072	−0.0109	−0.0150	−0.0183	−0.0207
	(15,15)	−0.0021	−0.0042	−0.0058	−0.0081	−0.0105	−0.0123	−0.0137
(1,3)	(20,20)	−0.0018	−0.0027	−0.0039	−0.0058	−0.0079	−0.0096	−0.0109
	(25,25)	−0.0012	−0.0020	−0.0030	−0.0046	−0.0064	−0.0078	−0.0089
	(30,30)	−0.0011	−0.0019	−0.0028	−0.0041	−0.0055	−0.0066	−0.0075
	(35,35)	−0.0002	−0.0006	−0.0012	−0.0021	−0.0031	−0.0040	−0.0047
	(10,10)	−0.0029	−0.0039	−0.0063	−0.0092	−0.0116	−0.0128	−0.0131
	(15,15)	−0.0022	−0.0034	−0.0059	−0.0075	−0.0087	−0.0092	−0.0091
(2,4)	(20,20)	−0.0017	−0.0027	−0.0040	−0.0056	−0.0070	−0.0077	−0.0080
	(25,25)	−0.0010	−0.0019	−0.0030	−0.0044	−0.0056	−0.0063	−0.0065
	(30,30)	−0.0009	−0.0011	−0.0030	−0.0041	−0.0051	−0.0057	−0.0059
	(35,35)	−0.0003	−0.0002	−0.0008	−0.0016	−0.0023	−0.0027	−0.0029

Table 2: Average MSE of the simulated estimates of $R_{s,k}$.

		(α,β)						
(s,k)	(n,m)	(3.0,1.5)	(2.5,1.5)	(2.0,1.5)	(1.5,1.5)	(1.5,2.0)	(1.5,2.5)	(1.5,3.0)
	(10,10)	0.0041	0.0052	0.0068	0.0092	0.0119	0.0139	0.0153
	(15,15)	0.0026	0.0033	0.0043	0.0058	0.0075	0.0087	0.0096
(1,3)	(20,20)	0.0017	0.0022	0.0029	0.0040	0.0052	0.0061	0.0068
	(25,25)	0.0014	0.0018	0.0024	0.0032	0.0042	0.0050	0.0055
	(30,30)	0.0011	0.0014	0.0018	0.0025	0.0032	0.0038	0.0043
	(35,35)	0.0009	0.0011	0.0015	0.0021	0.0027	0.0032	0.0036
	(10,10)	0.0096	0.0115	0.0141	0.0171	0.0193	0.0199	0.0196
	(15,15)	0.0062	0.0075	0.0091	0.0111	0.0125	0.0130	0.0128
(2,4)	(20,20)	0.0042	0.0051	0.0063	0.0078	0.0090	0.0094	0.0094
	(25,25)	0.0035	0.0043	0.0052	0.0065	0.0074	0.0078	0.0078
	(30,30)	0.0028	0.0033	0.0041	0.0050	0.0058	0.0060	0.0060
	(35,35)	0.0022	0.0027	0.0034	0.0042	0.0049	0.0052	0.0052

The second data set (from Linhardt & Zucchini 1986, p. 69) represents the failure times of the air conditioning system of an airplane (in hours): 23, 261, 87, 7, 120, 14, 62, 47, 225, 71, 246, 21, 42, 20, 5, 12, 120, 11, 3, 14, 71, 11, 14, 11, 16, 90, 1, 16, 52, 95. Gupta & Kundu (2003) studied the validity of the generalized exponential distribution and they compute the Kolmogorov-Smirnov (KS) distance between the empirical distribution function and the fitted distribution functions which is 0.1744 with a corresponding p-value 0.2926. Therefore, it is clear that the generalized exponential model fits quite well to both the data sets.

We use the iterative procedure to obtain the MLE of λ using (8), and MLEs of α and β are obtained by substituting MLE of λ in (6) and (7). The final estimates are $\widehat{\lambda} = 2.80609$, $\widehat{\alpha} = 1.00667$ and $\widehat{\beta} = 0.02098$. Based on the estimates of α and

β, the MLE of $R_{s,k}$ becomes $\widehat{R}_{1,3} = 0.893191$ and $\widehat{R}_{2,4} = 0.819677$. The 95% confidence intervals for $R_{1,3}$ become (0.841368, 0.945014) and for $R_{2,4}$ become (0.735472, 0.903882).

Table 3: Average confidence length of the simulated 95% confidence intervals of $R_{s,k}$.

		(α,β)						
(s,k)	(n,m)	(3.0,1.5)	(2.5,1.5)	(2.0,1.5)	(1.5,1.5)	(1.5,2.0)	(1.5,2.5)	(1.5,3.0)
	(10,10)	0.2112	0.2399	0.2762	0.3221	0.3627	0.3873	0.4012
	(15,15)	0.1747	0.1981	0.2279	0.2659	0.3000	0.3212	0.3337
(1,3)	(20,20)	0.1512	0.1716	0.1977	0.2311	0.2614	0.2804	0.2918
	(25,25)	0.1351	0.1534	0.1768	0.2069	0.2342	0.2515	0.2619
	(30,30)	0.1238	0.1404	0.1618	0.1893	0.2145	0.2304	0.2401
	(35,35)	0.1140	0.1295	0.1492	0.1748	0.1982	0.2132	0.2224
	(10,10)	0.3267	0.3628	0.4045	0.4485	0.4744	0.4782	0.4697
	(15,15)	0.2721	0.3020	0.3368	0.3742	0.3973	0.4020	0.3962
(2,4)	(20,20)	0.2366	0.2630	0.2939	0.3274	0.3486	0.3533	0.3486
	(25,25)	0.2119	0.2356	0.2635	0.2939	0.3134	0.3180	0.3141
	(30,30)	0.1943	0.2161	0.2416	0.2697	0.2878	0.2923	0.2890
	(35,35)	0.1794	0.1996	0.2234	0.2497	0.2669	0.2716	0.2688

Table 4: Average coverage probability of the simulated 95% confidence intervals of $R_{s,k}$.

		(α,β)						
(s,k)	(n,m)	(3.0,1.5)	(2.5,1.5)	(2.0,1.5)	(1.5,1.5)	(1.5,2.0)	(1.5,2.5)	(1.5,3.0)
	(10,10)	0.9230	0.9247	0.9277	0.9220	0.9140	0.9070	0.9053
	(15,15)	0.9327	0.9330	0.9357	0.9323	0.9303	0.9280	0.9243
(1,3)	(20,20)	0.9373	0.9387	0.9397	0.9400	0.9360	0.9293	0.9243
	(25,25)	0.9287	0.9323	0.9347	0.9360	0.9340	0.9293	0.9247
	(30,30)	0.9347	0.9360	0.9393	0.9403	0.9420	0.9427	0.9363
	(35,35)	0.9453	0.9480	0.9497	0.9477	0.9450	0.9417	0.9347
	(10,10)	0.9197	0.9213	0.9230	0.9177	0.9133	0.9133	0.9097
	(15,15)	0.9320	0.9323	0.9340	0.9333	0.9307	0.9277	0.9237
(2,4)	(20,20)	0.9353	0.9373	0.9390	0.9387	0.9327	0.9310	0.9260
	(25,25)	0.9287	0.9320	0.9333	0.9383	0.9333	0.9300	0.9263
	(30,30)	0.9353	0.9380	0.9410	0.9397	0.9390	0.9393	0.9363
	(35,35)	0.9453	0.9490	0.9490	0.9453	0.9433	0.9380	0.9360

4. Conclusions

In this paper, we have studied the multicomponent stress-strength reliability for generalized exponential distribution when both stress, strength variates follow the same population. Also, we have estimated asymptotic confidence interval for the multicomponent stress-strength reliability. The simulation results indicate that the average bias and average the MSE decrease as sample size increases for both situations of (s,k). Among the parameters the absolute bias and MSE decrease (increase) as α increases (β increases) in both the cases of (s,k). The length of the confidence interval also decreases as the sample size increases and the coverage

probability is close to the nominal value in all sets of parameters considered here. Using real data, we illustrate the estimation process.

[Recibido: abril de 2011 — Aceptado: diciembre de 2011]

References

Awad, M. & Gharraf, K. (1986), 'Estimation of $P(Y < X)$ in Burr case: A comparative study', *Communications in Statistics-Simulation and Computation* **15**, 389–403.

Bhattacharyya, G. K. & Johnson, R. A. (1974), 'Estimation of reliability in multicomponent stress-strength model', *Journal of the American Statistical Association* **69**, 966–970.

David, H. A. (1981), *Order Statistics*, John Wiley and Sons, New York.

Downtown, F. (1973), 'The estimation of $P(X > Y)$ in the normal case', *Technometrics* **15**, 551–558.

Enis, P. & Geisser, S. (1971), 'Estimation of the probability that $Y < X$', *Journal of the American Statistical Association* **66**, 162–168.

Gupta, R. D. & Kundu, D. (1999), 'Generalized exponential distributions', *Australian and New Zealand Journal of Statistics* **41**, 173–188.

Gupta, R. D. & Kundu, D. (2001), 'Generalized exponential distributions; different method of estimations', *Journal of Statistical Computation and Simulation* **69**, 315–338.

Gupta, R. D. & Kundu, D. (2002), 'Generalized exponential distributions; statistical inferences', *Journal of Statistical Theory and Applications* **1**, 101–118.

Gupta, R. D. & Kundu, D. (2003), 'Discriminating between the Weibull and generalized exponential distributions', *Computational Statistics and Data Analysis* **43**, 179–196.

Kendall, M. G. & Stuart, A. (1979), *The Advanced Theory of Statistics*, Vol. 2, Charles Griffin and Company Limited, London.

Kundu, D. & Gupta, R. (2006), 'Estimation of $P(Y < X)$ for Weibull distribution', *IEEE Transactions on Reliability* **55**(2), 270–280.

Kundu, D. & Gupta, R. D. (2005), 'Estimation of $P(Y < X)$ for the generalized exponential distribution', *Metrika* **61**(3), 291–308.

Kundu, D. & Raqab, M. Z. (2009), 'Estimation of $R = P(Y < X)$ for three-parameter Weibull distribution', *Statistics and Probability Letters* **79**, 1839–1846.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.

Linhardt, H. & Zucchini, W. (1986), *Model Selection*, Wiley Eastern Limited, New York.

McCool, J. I. (1991), 'Inference on $P(Y < X)$ in the Weibull case', *Communications in Statistics-Simulation and Computation* **20**, 129–148.

Nandi, S. B. & Aich, A. B. (1994), 'A note on estimation of $P(X > Y)$ for some distributions useful in life-testing', *IAPQR Transactions* **19**(1), 35–44.

Pandey, M. & Uddin, B. M. (1985), Estimation of reliability in multicomponent stress-strength model following Burr distribution, *in* 'Proceedings of the First Asian congress on Quality and Reliability', New Delhi, India, pp. 307–312.

Rao, C. R. (1973), *Linear Statistical Inference and its Applications*, Wiley Eastern Limited, India.

Rao, G. S. & Kantam, R. R. L. (2010), 'Estimation of reliability in multicomponent stress-strength model: Log-logistic distribution', *Electronic Journal of Applied Statistical Analysis* **3**(2), 75–84.

Raqab, M. Z. & Kundu, D. (2005), 'Comparison of different estimators of $P(Y < X)$ for a scaled Burr type X distribution', *Communications in Statistics-Simulation and Computation* **34**(2), 465–483.

Raqab, M. Z., Modi, M. T. & Kundu, D. (2008), 'Estimation of $P(Y < X)$ for the 3-parameter generalized exponential distribution', *Communications in Statistics-Theory and Methods* **37**(18), 2854–2864.

Surles, J. G. & Padgett, W. J. (1998), 'Inference for $P(Y < X)$ in the Burr type X model', *Communications in Statistics-Theory and Methods* **7**, 225–238.

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 77 a 108

Quantification of Ordinal Surveys and Rational Testing: An Application to the Colombian Monthly Survey of Economic Expectations

Cuantificación de encuestas ordinales y pruebas de racionalidad: una aplicación con la encuesta mensual de expectativas económicas

Héctor Manuel Zárate[1,a], Katherine Sánchez[1,b], Margarita Marín[1,c]

[1]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

Expectations and perceptions obtained in surveys play an important role in designing the monetary policy. In this paper we construct continuous variables from the qualitative responses of the Colombian Economic Expectation Survey (EES). This survey examines the perceptions and expectations on different economic variables. We use the methods of quantification known as balance statistics, the Carlson-Parkin method, and a proposal developed by the Analysis Quantitative Regional (AQR) group of the University of Barcelona. Then, we later prove the predictive ability of these methods and reveal that the best method to use is the AQR. Once the quantification is made, we confirm the rationality of the expectations by testing four key hypotheses: unbiasedness, no autocorrelation, efficiency and orthogonality.

***Key words*:** Rational Expectations, Survey, Quantification.

Resumen

En este artículo se cuantifican las respuestas cualitativas de la "Encuesta Mensual de Expectativas Económicas (EMEE)" a través de métodos de conversión tradicionales como la estadística del balance de Batchelor, el método probabilístico propuesto por Carlson-Parkin (CP) y la propuesta del grupo de Análisis Cuantitativo Regional (ACR) de la Universidad de Barcelona. Para las respuestas analizadas de esta encuesta se encontró que el método ACR registra el mejor desempeño teniendo en cuenta su mejor capacidad predictiva. Estas cuantificaciones son posteriormente utilizadas en pruebas de racionalidad de expectativas que requieren la verificación de cuatro hipótesis fundamentales: insesgamiento, correlación serial, eficiencia y ortogonalidad.

***Palabras clave*:** cuantificación, encuestas, expectativas racionales.

[a]Lecturer. E-mail: hmzarates@unal.edu.co
[b]MSc student. E-mail: ksanchezc@unal.edu.co
[c]MSc student. E-mail: mmarinj@unal.edu.co

1. Introduction

Economic decisions are usually made under a scenario of uncertainty about economic conditions. Thus, expectations on key variables and how private agents form their expectations play a crucial role in macroeconomic analysis. The direct way in the measurement of expectations comes from the application of qualitative surveys[1] of firms, which try to gauge respondent's perceptions regarding current economic conditions and expected future activity. According to Pesaran (1997), the "Business Surveys" provide the only opportunity to explore one of the big black boxes in the economy that inquire about the expectations and which allows to obtain leading indicators of current changes in economic variables over the business cycle.

The main characteristic of this kind of surveys is that questions provide ordinal answers that reveal the direction of change for the variable under consideration[2]. In other words it increases, remains constant or declines. The information extracted with ordinal data is used to anticipate the behavior of economic variables of continuous type and to build indicators of economic activity[3]. However, the analysis requires a cardinal unit of measurement and therefore a conversion method from nominal to quantitative figures is a topic in business analysis.

In this paper we study the properties of several methodologies to quantify the qualitative answers and present an application from the monthly Economic Expectation Survey (EES) realized by the central bank of Colombia during the period October 2005 to January 2010. The article is organized into six sections including this introduction. In Section 2, briefly we describe traditional methods to convert variables from qualitative to continuous type. Later, in Section 3 we present the application of these methods with some of the questions contained in the EES. The models for expectations and the econometric strategy for testing are summarized in the Section 4. Section 5 shows the empirical results. Finally, in Section 6 we summarize the conclusions.

2. Quantification Methods of Expectations

In order to measure the attitudes of the respondents for variables such as prices, the central bank distributes monthly a questionnaire that can be classified into four broad categories: past business conditions, outlook of the business activity, pressures on firm's production capacity, outlook of wages and prices.

The EES survey answers contains three options classified as follows: "increases", "decreases" or "remains the same". In Table 1 is described the notation of the answers of the public-opinion poll in terms of judgments (perception in the period t

[1]The impact of the expectations of the agents on the economic variables is difficult to observe due to the fact that these are evaluated by quantitative measurements that present problems of sensitivity: Sampling errors, sampling plan and measurement errors.

[2]Berk (1999), Visco (1984) among others, analyzed opinion surveys with more than three categories of response.

[3]The evolution of cyclical movements is called the Business Climate indicator.

of the evolution of variable respect of the period $t-1$) and expectations (perception in t of the evolution expected from the variable for $t+1$)[4]. In this case, $JUP_t + JDO_t + JEQ_t = 1$ if they are judgments, or $EUP_t + EDO_t + EEQ_t = 1$ if they are expectations.

Table 1: Clasification of answers.

Notation	Description
JUP_t	Proportion of enterprises that at time t perceive that the observed variable is going 'Up' between period $t-1$ and period t.
JDO_t	Proportion of enterprises that at time t perceive that the observed variable is going 'Down' between period $t-1$ and period t.
JEQ_t	Proportion of enterprises that at time t perceive that the observed variable has 'A normal level' between period $t-1$ and period t.
EUP_t	Proportion of enterprises that at time t expect an 'Increase' of the variable from period t to period $t+1$.
EDO_t	Proportion of enterprises that at time t expect a 'Decrease' of the variable from period t to period $t+1$.
EEQ_t	Proportion of enterprises that at time t 'Don't expect any change' in the variable from period t to period $t+1$.

In this article, the expectations for growth in sales volume, the variation of the total raw material prices (national and imported) and the variation in price of products that will be sold; are quantified. The quantification techniques are based on two concepts. The first concerns with the distribution of expectations in which it is assumed that in the period t every individual i forms a distribution of subjective probability distribution $f_{it}(\mu_{it}, \tau_{it}^2)$ with mean μ_{it} and variance τ_{it}^2. The mean of this can be distributed through individuals as: $\mu_{it} g_t(\mu_t, \sigma_t^2)$ (where the expected value μ_t measures the average expectations in the survey population at time t and σ_t measures the dispersion of average expectations in that population); the second assumes that an individual with probability distribution f_{it} answers "increases" or "decreases" to the questions of the survey, according to whether the average subjective μ_{it} exceeds some rate limit δ_{it} or it is less to another rate limit $-\epsilon_{it}$ respectively, so that $\delta_{it} > 0$ and $\epsilon_{it} > 0$.

2.1. The Balance Statistics

Originally, this kind of statistics was introduced by Anderson (1952) in his work for the IFO survey. This statistic is obtained by:

$$S_t^{t+1} = EUP_t - EDO_t \tag{1}$$

The advantage of this statistic is that it can be used both for questions that investigate on judgments (S_t^{t-1}), and for making reference on expectations (S_t^{t+1}). Batchelor (1986) takes into account the key concepts of the general theory of quantification based on the following assumptions:

[4]See `www.banrep.gov.co/economia/encuesta_expeco/Cuestionario_CNC.pdf`

•The distribution of expectations follows a sign function (Pfanzagl 1952, Theil 1958), with a time-invariant parameter θ. It is to say $g_t(\mu_t, \sigma_t^2) = g(\mu_t, \sigma_t^2)$, where:

$$EDO_t \quad \text{si} \quad \mu_{it} = -\theta; \qquad EEQ_t \quad \text{si} \quad \mu_{it} = 0; \qquad EUP_t \quad \text{si} \quad \mu_{it} = \theta \tag{2}$$

•The distribution of the expectation is characterized by long terms unbiased, which means that in a period of time with T surveys, the average expectation μ_t is equal to the current average rate variable:

$$\sum_{t=1}^{T} \mu_t = \sum_{t=1}^{T} x_t \tag{3}$$

•The function of the response limits δ_{it} and ϵ_{it}; may be asymmetric and vary over the individuals and time, but must be strictly less than θ; it is to say:

$$\delta_{it} < \theta, \quad \epsilon_{it} < \theta \tag{4}$$

Therefore, the expected value and the variance of the distribution are:

$$\mu_{it} = \theta(EUP_t - EDO_t), \quad \sigma_t^2 = \theta^2[(EUP_t + EDO_t) - (EUP_t - EDO_t)^2] \tag{5}$$

By assuming the response function, the proportions of the sample:EUP_t, EDO_t and EEQ_t behave like maximum likelihood estimators, making it possible to estimate the parameter θ. With this estimate, it is obtained that

$$\begin{aligned} \sum_{t=1}^{T} \theta(EUP_t - EDO_t) &= \sum_{t=1}^{T} x_t, \\ \theta \sum_{t=1}^{T} (EUP_t - EDO_t) &= \sum_{t=1}^{T} x_t, \end{aligned} \qquad \widehat{\theta} = \frac{\sum_{t=1}^{T} x_t}{\sum_{t=1}^{T} (EUP_t - EDO_t} \tag{6}$$

Fluri & Spoerndli (1987) estimate the expectation of the variable as:

$$(E(X))_t = \widehat{\theta}(EUP_t - EDO_t) \tag{7}$$

Where $E(X)$ denotes the expectation of the random studied variable, x_t is the realization of the variable under study and $(\widehat{\theta})$ is the scaling factor determined by the unbiasedness of the equation . Thus, the Modified Balance Statistical (MBS) provides a measure of the expected average in the variable, taking into account the trend and the points of inflection.

2.1.1. Recent Proposals

Loffler (1999) estimates the measurement error introduced by the probabilistic and proposes a linear correction method[5]. On his part, Mitchell (2002) finds

[5]Claveria & Suriñach (2006).

evidence that the normal distribution, as well as any other stable distribution, provides accurate expectations[6]. Claveria & Suriñach (2006) posed different statistical expectations for the quantifications, including a method that proposes the use of random walks and another one that use Markov processes of first order.

Claveria (2010) proposes a statistical balance with nonlinear variation, called Weighted Balance, such that $WB_t = \frac{R_t - F_t}{R_t + F_t} = \frac{B_t}{1 - C_t}$. This statistic takes into account the percentage of respondents expecting no change in the evolution of an economic variable.

2.2. Probabilistic Method

This method was proposed originally for Theil (1952), initially applied by Knobl (1974), and identified by Carlson & Parkin (1975) as CP "Probabilistic Method". For these authors, x_{it} represents the percentage of change of a random variable X_i of period $t-1$ for the period t (with $t = 2, 3, \ldots, T$); the respondent is indexed by i and x^e_{it} symbolizes the expectation having i on the change in X_i from the period t to the period $t+1$ (with $t = 1, 2, \ldots, T-1$). Also, they assume intuitively that respondents have a range of indifference $(a_{it},\ b_{it})$, with $a_{it} < 0$ and $b_{it} > 0$, so that each one of the respondent answers "Decrease" if $x^e_{it} < a_{it}$ or "Increase" if $x^e_{it} > b_{it}$. If there is not change, $x^e_{it} \in (a_{it}, b_{it})$.

Thus, in the period t each respondent based his answers on a subjective probability distribution $f_i(x_{it}/I_{t-1})$ defined as from future change in X_i conditioned by information available at the time $t-1$ (represented by I_{t-1}). These subjective probability distributions $f_i(\cdot)$ are such that they can be used to obtain a probability distribution of added $g(x_i/\Omega_{t-1})$, where $\Omega_{t-1} = \bigcup_{i=1}^{N-1} I_{t-1}$ is the union of individual information groups (where N_t is the total number of respondents in the period t)[7]. For the estimation of x^e_t, ("Average expectation of respondents"), the equation $x^e_t = \sum_{i=1}^{N} w_i x^e_{it}$, is used where w_i represents the weight of the respondent i and x^e_{it} represents the individual expectations.

Carlson and Parkin make two additional assumptions: First, that the indifference interval is equal for all respondents ($a_{it} = a_t$ y $b_{it} = b_t$). Second $f_i(x_{it}/I_{t-1})$ has the same form for all players and the first and second moment are finite. Thus, x^e_{it} may be considered as independent samples of an aggregate distribution $g(\cdot)$ with mean $E(x_t/\Omega_{t-1}) = x^e_t$ and variance σ^2_t, that can be written as[8]:

$$EDO_t = prob\{x_t \leq a_t/\Omega_{t-1}\}, \quad EUP_t = prob\{x_t \geq b_t/\Omega_{t-1}\} \tag{8}$$

where each agent has the same subjective distribution of expectations based on the information available. In most applications the use of the normal distribution that is statistically appropriate, is completely specified by two parameters. Thus, if G is defined as the cumulative distribution of the aggregate distribution $g(\cdot)$;

[6] Ibíd.

[7] Which is constant for each period.

[8] Note that if individual distributions are independent through respondents, they have a common and finite first and second time, then by the Central Limit Theorem $g(\cdot)$, they have normal distribution.

it is obtained by standardizing f_t and r_t as the abscissa of the inverse of the G corresponding to EDO_t and $(1 - EUP_t)$. That is:

$$f_t = G^{-1}(EDO_t) = (a_t - x_t^e)/\sigma_t, \quad r_t = G^{-1}(1 - EUP_t) = (b_t - x_t^e)/\sigma_t \tag{9}$$

Solving the system of Equation 9 to find the average expectations x_t^e and the dispersion σ_t, we obtain:

$$x_t^e = \frac{b_t f_t - a_t r_t}{f_t - r_t}, \quad \sigma_t = -\frac{b_t - a_t}{f_t - r_t} \tag{10}$$

Carlson and Parkin assume that the indifference interval does not vary over time, remaining fixed between business, and is symmetric around zero; that is, $-a_t = b_t = c$. Given this, we obtain an expression for calculating operational x_t^e by the method of Carlson and Parkin (CP), defined as:

$$x_{t,cp}^e = c\frac{f_t + r_t}{f_t - r_t} \tag{11}$$

with $c = \frac{\sum_t x_t}{\sum_t d_t}$ and $d_t = \frac{f_t + r_t}{f_t - r_t}$, where x_t includes the annual variation of the observed variable. In this case, the role of c is scaled x_t^e, so that the average value of x_t equals x_t^e, which means that expectations are assumed to be average unbiased. Assuming that the random variable observed X has normal distribution, then f_t and r_t are found using the inverse of the cumulative distribution standard normal distribution, in the Equation 9. It is important to note that the imposition of expectations makes them unsuitable to apply rationality contrasts a posteriori. Moreover, it is assumed that $f_i(\cdot)$ has normal distribution. However, the uniform distribution also can be used. Assuming that X is distributed uniformly over the interval [0, 1], then f_t and r_t are calculated as:

$$f_t = \sqrt{12}(EDO_t - \frac{1}{2}), \quad r_t = \sqrt{12}(\frac{1}{2} - EUP_t) \tag{12}$$

2.2.1. Disadvantages and Extensions of the Carlson-Parkin Method

There are several shortcomings related to the Carlson-Parkin method. The same answers for all the respondents cause that the statistic goes to infinity, which, in turn, impedes the computation of expectations. Moreover, the assumption of constant and symmetric limits through time means that respondents are equally sensitive to an expected rise or an expected fall, of the variable under study. Seitz (1988) relaxes the assumptions of the Carlson-Parkin method allowing time variant boundaries of the indifference interval[9].

2.3. Regional Quantitative Analysis (RQA) Method

This method was implemented by Pons and Claveria at the Regional Quantitative Analysis Group (RQA); Department of Econometrics; Statistics and Economics at the University of Barcelona (Claveria, Pons & Suriñach 2003). The

[9]See Nardo (2003).

estimation is performed in two stages. The first stage gives a first set of expectations of the variation of the variable referred to as input, which can be defined as:

$$x^e_{input,t} = \widehat{c} * d_t \tag{13}$$

where $\widehat{c} = |x_{t-1}|$, $d_t = \frac{f_t + r_t}{f_t - r_t}$ and x_{t-1} shows the growth rate of the reference quantitative indicator of the previous period. The parameter estimation of indifference has a dual function: Firstly, it avoids the imposition of unbiasedness that occurs when estimating the range of indifference by the CP method, thus, the estimation allows movement in the indifference interval boundaries to incorporate changes in response time, and secondly, it relaxes the assumption of constancy over time of the scaling parameter because the parameter c will correspond to the rate of variation of quantitative indicator in the reference period $t-1$.

The re-scaling of the series Input obtained from Equation 13 is necessary, because the function of c is the scalar statistic d_t and, therefore, would be distorting the interpretation given by the over-dimension of the class EEQ_t, that requires less commitment from the respondent, and just distorting the interpretation that is the parameter c as the limit of visibility. This justifies the need for scaling in two stages.

In the second stage the model is re-scaled with parameters changing over time. This regression equation estimated by ordinary least squares (OLS) and the parameters obtained are used to estimate the new set of expectations, where the series Input acts as an exogenous variable:

$$x_t = \alpha + \beta x^e_{input,t} + u_t \tag{14}$$

where α y β are the parameters of the estimation and u_t is the error. On the OLS, estimation of the regression parameters is constructed following conversion equation:

$$x^e_t = \widehat{\alpha} + \widehat{\beta} x^e_{input,t} \quad \text{donde} \quad x^e_{input,t} = \widehat{c} * d_t \quad \text{y} \quad \widehat{c} = |x_{t-1}| \tag{15}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ parameters are estimated and x^e_t represents the number of estimated expectations of the rate of variation of the observed variable. Obtaining these set of directly observed expectations allows us to contrast some of the hypotheses usually assumed in economic models, such as the rationality of the agents.

3. Application to the EES

In this section we apply the methods of quantification submitted to the observed variables (EES); therefore expectations obtained are evaluated in terms of their predictive ability. This is evaluated under four statistics known as Mean Absolute Error (MAE), Median Absolute of the Percentage Error (MAPE), Root

Error Square Mean (RESM) and the coefficient U of Theil (TU1):

$$\begin{aligned} MAE &= \sum_{t=1}^{T} \frac{|x_t - x_t^e|}{T} \\ MAPE &= \frac{\sum_{t=1}^{T} \frac{|x_t - x_t^e|}{x_t}}{T} * 100 \\ RESM &= \sqrt{\sum_{t=1}^{T} \frac{(x_t - x_t^e)^2}{T}} \\ TU1 &= [\frac{\sum_{t=1}^{T}(x_t - x_t^e)^2}{\sum_{t=1}^{T}(x_t)^2}]^{\frac{1}{2}} \end{aligned} \tag{16}$$

3.1. Quantification of Question 2 for EES

The growth of sales volume (quantity) in the next 12 months, compared with growth in sales volume (quantity) in the past 12 months, is expected to be: a) Increased, b) Decreased, c) The same (See Figure 1).

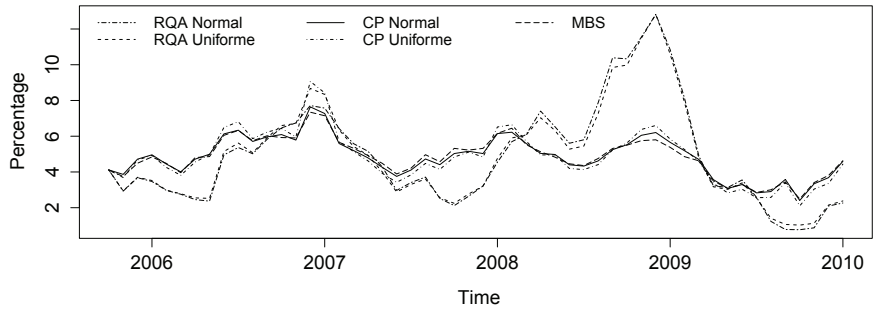

FIGURE 1: Expectations question 2.

For the quantification of this question the indicator of annual variation Total Index Sales[10], obtained from DANE is used as a reference. The methods applied were: RQA with normal and uniform distribution, method of CP with normal and uniform distribution and MBS.

It is noted that the expectations generated by normal RQA and the uniform method have very similar behaviors, and the patterns tend to have more movement when compared with other methods. Similarly, one can see that the series of expectations with the CP method with standard normal and uniform distribution, have similar behavior.

The results of the evaluation of the predictive power are presented in Table 2, and they suggest that the most appropriate method to carry out this quantification is the RQA with normal distribution, followed by the uniform distribution. In third

[10]In this case, the variable is nominal.

place is the CP method with uniform distribution, statistically below the MBS and finally by the normal CP method.

Table 2: Predictability Evaluation Question 2.

	MBS	Normal CP	Uniform CP	Normal RQA	Uniform RQA
MAE	0.046	0.047	0.042	0.029	0.032
MAPE	1.826	1.947	1.579	0.731	0.866
RESM	0.055	0.057	0.051	0.036	0.039
TU1	0.454	0.463	0.416	0.295	0.319

3.2. Quantification of Question 9 for EES

The increase in total prices of raw materials (domestic or imported) to buy in the next 12 months, compared with the total prices of raw materials purchased in the past 12 months is expected to be: a) Higher, b) Lower, c) The same (See Figure 2).

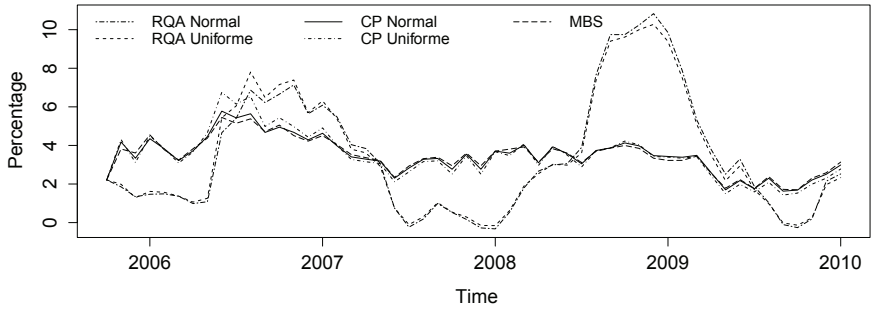

Figure 2: Expectations question 9.

The indicator used as reference is the annual variation Producer Price Index, obtained from the natinal statistical office in Colombia DANE. The series of expectations are estimated with the method of RQA with normal and uniform distributions and they exhibit similar behaviors on oscillations recorded over time. Moreover, the estimated normal uniform and CP and MBS fluctuate less than the other series.

The evaluation of the predictive ability (Table 3) indicates that the most appropriate method is RQA with normal distribution, followed by the uniform distribution. The third and fourth place corresponds to the CP method with uniform and normal distribution, respectively. The least predictive method presented is the MBS.

Table 3: Predictability Evaluation Question 9.

	MBS	Normal CP	Uniform CP	Normal RQA	Uniform RQA
MAE	2.648	2.623	2.616	1.648	1.704
MAPE	1.324	1.295	1.247	0.689	0.678
RESM	3.359	3.317	3.289	2.123	2.158
TU1	0.667	0.657	0.652	0.421	0.428

3.3. Quantification of Question 11 for EES

The increase in prices of products that will sell in the next 12 months, compared with the increase of prices of products sold in the past 12 months,are expected to be: a) Higher, b) Lower, c) The same (See Figure 3).

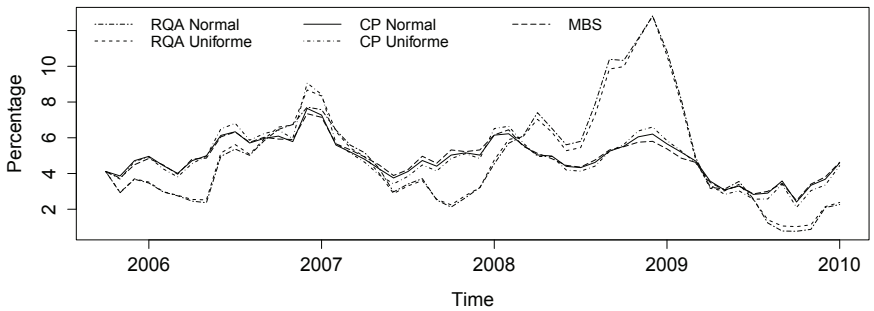

Figure 3: Expectations question 11.

The quantification is used as a reference indicator of annual variation rate of the Producer Price Index Produced and Consumed (PPIP&C).

It is noted that the expectations generated by the application of the method of MBS have a pattern that turns smoothly around the mean. The expectations series obtained with the CP method with normal and uniform distribution are similar but with a greater degree of variability. The expectations series obtained with the CP method with normal and uniform distribution are similar but with a greater degree of variability.

According to the statistics for the evaluation of the predictive ability (Table 4), the method with the best performance is the RQA with normal distribution, followed by the uniform distribution. The third and fourth place corresponds to the CP method uniform and the normal distributions respectively. Finally, the MBS method is the least predictive.

In general, there is evidence that the RQA methodology with standard normal distribution, followed by the uniform distribution; they present the best results in terms of evaluation of the predictive and their methods are attractive because the indifference parameter is asymmetric, changing over time and staying unbiased (which makes it optimal for the contrast of hypothesis about formation of expectations).

Nevertheless, due to the restriction of information on this method (both judgments and expectations), it is suggested to consider the CP method and the method of MBS in the quantification of the variables if you do not have all the information available.

Table 4: Predictability evaluation question 11.

	MBS	Normal CP	Uniform CP	Normal RQA	Uniform RQA
MAE	2.034	2.026	2.035	1.484	1.549
MAPE	0.697	0.691	0.660	0.446	0.461
RESM	2.792	2.772	2.753	1.980	2.058
TU1	0.477	0.474	0.470	0.339	0.351

4. Modeling the Expectations

4.1. Extrapolative and Adaptative Expectations

The pure model of extrapolative expectations is based on the assumption that the expectations depend only on the observed values of the variable that wil be predicted[11], of the variable to predict (Ece 2001), so this model can be represented as (Pesaran 1985):

$$ {}_t x^e_{t+1} = \alpha + \sum_{i=1}^{\infty} w_j x_{t-j} + u_{t+1} \tag{17} $$

where ${}_t x^e_{t+1}$ is the expectation of the variable formed in the period t, for the period $t+1$; x_{t-j}(with $j = 0, 1, 2, \ldots$) are the known data of the variable in the period t; w_j are the weights (fixed) given to each of the known values of the variable, and u_{t+1} is the random error term that attempts to capture the unobserved effects on the expectation.

Expectations of the adaptive model imply that if the variable value and expectations differ from the period of studies, then a correction to the expectation for the next period is made. However, if there is not difference, the expectation for the next period will stay unchanged (Ece 2001). On the imposition of certain restrictions to w_j in equation 17 it is possible to find the models used to testing adaptative expectations (this would support the hypothesis that such expectations are a special case of extrapolative expectations; (Pesaran 1985)). Thus, the four models used to represent the adaptive expectations are (Pesaran 1985, Ece 2001):

$$ x^e_{t+1} - x^e_t = w(x_t - x^e_t) + u_{t+1} \tag{18} $$

$$ x^e_{t+1} - x^e_t = \alpha_0(x_t - x^e_t) + \alpha_0(x_{t-1} - x^e_{t-1}) + u_{t+1} \tag{19} $$

$$ x^e_{t+1} - x^e_t = \beta_0(x_t - x^e_t) + \beta_1(x_{t-1} - x_{t-1}) + \beta_2(x_{t-1} - x^e_{t-1}) + u_{t+1} \tag{20} $$

$$ x^e_{t+1} = \lambda_0 + \lambda_1 x^e_t + \lambda_2 x^e_{t-1} + \lambda_3 x_t + \lambda_4 x_{t-1} + u_{t+1} \tag{21} $$

[11] See sections 2 and 3 of this paper.

Finally, to see if expectations are adaptive or extrapolative, it is necessary to perform an analysis on the coefficient of determination and the individual and joint significance level of the parameters. If all these indicators are significant, then it confirms the presence of these expectations. These models may have problems of serial correlation of errors and endogeneity, so it is necessary to apply appropriate econometric corrections to obtain estimators on which statistical inference can be made.

4.2. Rational Expectations

The rational expectations model was originally proposed by Muth (1961) and is based on the assumption that individuals (at least on average) use all available and relevant information when they make their predictions on the future behavior of the variable studied (Ece 2001). This can be expressed by:

$$x_t^e = E(x_t/I_{t-1}) \tag{22}$$

where x_t represents the value of the variable in the period t; x_t^e stands for the expected value of the variable for the period t reported in $(t-1)$ and I_{t-1} symbolizes the available and relevant information in $(t-1)$. The rational expectations must satisfy four tests (Ece 2001) and (Da Silva 1998):

1. *Unbiasedness*: For the regression $x_t = \alpha + \beta x_t^e + u_t$ the hypothesis $H_0 : \alpha = 0; \beta = 1$ cannot be rejected.

2. *Lack of serial correlations*: $E(u_t u_{t-i}) = 0, \forall_i \neq 0$

3. *Efficiency*: In the equation $u_t = \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_i x_{t-i}$, $i > 0$; the coefficients should not be significant.

4. *Orthogonality*: For the regression $x_t = \alpha + \beta x_t^e + \gamma I_{t-1} + u_t$ where, γ represent the effect of the information on the variable, the hypothesis $H_0 : \alpha = 0; \beta = 1, \gamma = 0$ cannot be rejected.

Some authors argue that orthogonality hypothesis contains the rest. Therefore, is sufficient to prove the existence of this to demonstrate the rationality of expectations (Da Silva 1998).

4.3. Endogeneity Problem and a Correction

Quantitative data for the expectations were calculated from the variable observed, which was also used for the tests of rationality. This may generate endogeneity problems that lead to inconsistent estimators. Then, to the covariance matrix, Hansen & Hodrick (1980) propose, that, given an equation:

$$y_{t+k} = \beta x_t + u_{t,k} \tag{23}$$

where y_{t+k} is a variable k steps-ahead; x_t is a row vector of $T \times p$ dimension (where p is the number of parameters that may or may not include the intercept[12] and T is the number of observations) containing all the relevant information in the period t and at least one of the variables is endogenous; β is a column vector of $p \times 1$ dimension and $u_{t,k}$ is the vector of residues, calculated by Ordinary Least Squares (OLS). It is possible to make a correction to the covariance matrix Θ such that:

$$\widehat{\boldsymbol{\Theta}}_T = T(\boldsymbol{X}'_T\boldsymbol{X}_T)^{-1}\boldsymbol{X}'_T\widehat{\boldsymbol{\Omega}}_T\boldsymbol{X}_T(\boldsymbol{X}'_T\boldsymbol{X}_T)^{-1} \tag{24}$$

with

$$\boldsymbol{X}_T = \begin{bmatrix} x_1 \\ \vdots \\ x_T \end{bmatrix}$$

And the symmetric $\widehat{\boldsymbol{\Omega}}_T$ matrix of $T \times T$ dimension, whose lower triangular representation is:

$$\begin{bmatrix} \overline{R}_u^T(0) & & & & & & \\ \overline{R}_u^T(1) & \overline{R}_u^T(0) & & & & & \\ \vdots & & \ddots & & & & \\ \overline{R}_u^T(k-1) & & & \ddots & & & \\ 0 & & & & \ddots & & \\ \vdots & & & & & \ddots & \\ 0 & \cdots & 0 & \overline{R}_u^T(k-1) & \cdots & \overline{R}_u^T(1) & \overline{R}_u^T(0) \end{bmatrix}$$

where

$$\overline{R}_u^T(j) = \frac{1}{T}\sum_{t=j+1}^{T} \widehat{u}_{t,k}\widehat{u}_{t-j,k}$$

for

$$j \geq 0, R_u(j) = R_u(-j)^{13}$$

5. Empirical Results

We checked the four fundamental hypotheses of the rational expectations model using estimates by OLS and the correction of the covariance matrix. The results of these tests are found in the tables at the end.

The variable in the question 2, corresponde to the year-on-year variation rate of the total sales index (denoted by S_t). In questions 9 and 11, we employ the year-on-year variation of the Producer Price Index (PPI) and the year-on-year variation

[12] As was shown in the previous section, the unbiasedness and orthogonality tests include the intercept. However, the efficiency test does not.

[13] See Hansen (1979)

of the Producer Price Index – Producer and Consumer (PPI_P&C), nominated in both cases as P_t. We denoted the lags of this variable as S_{t-i} (question 2) and P_{t-i} (questions 9 and 11). The variable x_t^e represents in the question 2 the sales expectations, S_t^e, and in the questions 9 and 11 ask for the inflation expectations in raw materials and in products to be sold (in both cases P_t^e). For the efficiency test we use as dependent variable the error term u_t, which is equal to $S_t - S_t^e$ (question 2) and $P_t - P_t^e$ (questions 9 and 11). We generated these errors from the regression used in the unbiasedness test.

In the orthogonality test we use the one period lagged dependent variable[14] in all the questions. For the question 2, we use as information variables the monthly variation of two periods lagged Market Exchange Rate (MER_{t-2}), the year-on-year variation of one period lagged PPI (PPI_{t-1})[15], and year-on-year variation of the two periods lagged Manufacturing Industry Real Production Index (IPI_{t-2}). In the questions 9 and 11 we employ as information variables the MER_{t-2} and the one period lagged Aggregated Monetary ($M3_{t-1}$)[16].

In the Hansen and Hodrick correction, we use as the y_{t+k} variables P_t, S_t and u_t. As x_t we use: for the unbiasedness test, S_t^e (question 2) and P_t^e (questions 9 y 11); for the efficiency test, S_{t-i} (question 2) and P_{t-i} (questions 9 and 11) and for the orthogonality tests S_{t-1}, PPI_{t-1}, IPI_{t-2}, MER_{t-2} (question 2) and P_{t-1}, MER_{t-2} and $M3_{t-1}$ (question 9 y 11). As $u_{t,k}$ variable we use the errors generated for each of the OLS regressions of the rational test. Finally, k is equal to 12, because in all the questions of the survey we ask about the behavior of the variables in 12 months[17].

5.1. Results of the Rational Test for the question 2

5.1.1. Results by OLS

Table 5 presents the results of the unbiasedness and serial correlation tests. Only by methods MBS and uniform and normal CP we can reject the null hypothesis of unbiasedness. In the hypothesis of serial correlation, the LM[18] statistic reveals that only in MBS there is evidence of serial correlation. Table 6 shows the

[14]For example see Ece (2001), Gramlich (1983), Keane & Runkle (1990), Mankiw & Wolfers (2003), Pesaran (1985).

[15]This variables were used because they are indicators of domestic and foreign prices of the products, which can affect sales expectations

[16]As reported by the Central Bank in its Inflation Report of September 2010 (Banco de la República de Colombia 2010), these variables have shown a greater influence on the country's inflation level.

[17]To view the full survey format see
`http://www.banrep.gov.co/economia/encuesta_expeco/Cuestionario_CNC.pdf`

[18]Which tests the null hypothesis of existence of correlation between the errors of the regression using a regression between the errors, as the dependent variable, and the variables of the equation and the p times lagged errors, as independent variables. From this, the statistic LM = nR2 is calculated, where n is the number of data in the regression of errors and R2 is the coefficient of determination. This statistic approximates the Chi-square distribution with p degrees of freedom. If this statistic is greater than the critical Chi-square, then it is possible to reject the null hypothesis of no autocorrelation among the errors.

results of the efficiency test. In all cases there is a relationship between the error term and S_{t-3}. Additionally the errors in the uniform RQA show relations with S_{t-1} and the errors in normal RQA present relation with S_{t-1} and S_{t-2}.

The results of the orthogonality tests using S_{t-1} (Table 7), MER_{t-2} (Table 8), PPI_{t-1} (Table 9), IPI_{t-2} (Table 10), and all the variables (Table 11), indicates that in the case of S_{t-1}, for all of the data set is possible reject the null hypothesis. For MER_{t-2} is possible reject the null hypothesis by MBS and uniform and normal CP. In the case of PPI_{t-1} we cannot reject the orthogonality for uniform and normal RQA. For IPI_{t-2} is possible reject the null hypothesis by MBS and uniform and normal CP. Finally, with all the variables we can reject the orthogonality for all the data sets.

5.1.2. Results by OLS with the Hansen and Hodrick Correction

Table 12 presents the results of the unbiasedness test with the correction of Hansen and Hodrick. It is not possible to reject the existence of unbiasedness for any of the data sets. The results of the efficiency tests (Table 13) show that there is no evidence to reject this hypothesis in either case. The orthogonality test using S_{t-1} (Table 14), MER_{t-2} (Table 15), PPI_{t-1} (Table 16), IPI_{t-2} (Table 17) and all the variables (Table 18) shows that we cannot reject the null hypothesis, for any of the variables and data sets.

We did not test for serial correlation, since this cannot be corrected by the Hansen and Hodrick method. However, we can say that this test is also satisfied, because it is a corollary of the orthogonality, which is fulfilled for all methods. Therefore, by extension, the serial correlation must be satisfied[19].

5.2. Results of the Rational Test for the question 9

5.2.1. Results by OLS

Table 19 presents the results of the unbiasedness test and serial correlation. For none of the cases it is possible to reject the null hypothesis of unbiasedness. The LM statistic shows that there is serial correlation for all data sets. Table 20 reports the results of the efficiency test. In all the cases there is a relation between the errors and P_{t-1}. For uniform and normal RQA there are also relation with P_{t-2}Finally, MB and uniform and normal CP present relation with P_{t-8}.

The results of the orthogonality test using P_{t-1} (Table 21), MER_{t-2} (Table 22), $M3_{t-1}$ (Table 23), and all the variables (Table 24) show that for P_{t-1} we cannot accept the hypothesis of orthogonality, for any of the data sets. For MER_{t-2}, it is possible to reject the null hypothesis for MBS and normal and uniform CP. In the case of $M3_{t-1}$ we can not reject the null hypothesis, for all the data sets. Finally, with all the variables, it is possible to reject the orthogonality for all the methods.

[19]This reason is used to justify the non-existence of serial correlation for the other two questions.

5.2.2. Results by OLS with the Hansen and Hodrick Correction

In Table 25, we present the results of the unbiasedness test with the Hansen and Hodrick correction. There is not evidence to reject this null hypothesis for any model. The efficiency test (Table 26) shows that we can not reject this hypothesis. The results of the orthogonality test with P_{t-1} (Table 27), MER_{t-2} (Table 28), $M3_{t-1}$ (Table 29), and all the variables (Table 30) show that we cannot reject the null hypothesis, for any of the data sets and variables.

5.3. Results of the Rational Test for the question 11

5.3.1. Results by OLS

The Table 31 shows the results of the unbiasedness and serial correlation test. Only for the case of MBS, we can reject the null hypothesis of unbiasedness. The LM statistic shows that there is serial correlation for all data sets. Table 32 presents the results of the efficiency test. For all methods there is a relationship between errors and P_{t-1}. For normal and uniform RQA there is also a relationship with P_{t-2}.

The results of the orthogonality test using P_{t-1} (Table 33), MER_{t-2} (Table 34), $M3_{t-1}$ (Table 35), and all the variables (Table 36) show that for the case of P_{t-1} we can reject the null hypothesis for all the data sets. In the case of MER_{t-2} is possible to reject the null hypothesis for MBS and normal and uniform CP. In the case of $M3_{t-1}$ we can not reject the null hypothesis for all the data sets. Finally, with all the variables, it is possible to reject the orthogonality for all the methods.

5.3.2. Results by OLS with the Hansen and Hodrick Correction

In Table 37 we present the results of the unbiasedness test with the Hansen and Hodrick correction. There is not evidence to reject this null hypothesis for any model. The efficiency test (Table 38) shows that we cannot reject this hypothesis. The results of the orthogonality test with P_{t-1} (Table 39), MER_{t-2} (Table 40), $M3_{t-1}$ (Table 41), and all the variables (Table 42) show that we cannot reject the null hypothesis, for any of the variables and data sets.

6. Conclusions and Recommendations

In order to identify the employers expectation formation process, we quantified the qualitative responses to questions on economic activity and prices in the Economic Expectation Survey (EES), applied by the division of Economic Studies of the central bank of Colombia, from October 2005 to January 2010. We used the conversion methods of Modified Balance Statistical, Carlson-Parkin with standard normal distribution and uniform distribution [0, 1] and the method proposed by

the Regional Quantitative Analysis Group (RQA) at the University of Barcelona with standard normal distribution and uniform distribution [0, 1].

The evaluation of the quantification methods was performed using four statistics to analyze their predictability: mean absolute error (MAE), absolute percentage error of the median (MAPE), Root Mean Square Error (RESM) and Theil U coefficient (TU1). According to the criteria above, for the four analyzed variables, it was found that the method with the best predictability was the one proposed by the RQA group with standard normal distribution, followed by the uniform distribution [0, 1]. However, due to the restriction of information on this method, it is suggested to take into account the methods of the MBS and CP, in the quantification of the variables that do not have all available information.

Subsequently, we confirmed the existence of rational expectations for three questions of the EES. By applying the correction proposed by Hansen and Hodrick for the endogeneity problem, it was found that the unbiasedness, efficiency, orthogonality and serial correlation tests were fulfilled for the three questions, considering the five methods of quantification. With these results we can conclude that the business expectations of the variation in sales, prices of raw materials and prices of domestic production in Colombia are compatible with the hypothesis of rational expectations.

However, this document was an initial approach to the quantification and verification of the rational expectations. Further studies on the topic should explore other methodologies Kalman filter or considering parameters that change over time. Additionally, other papers can implement other econometric methods for testing rationality hypotheses, such as maximum likelihood estimators or restricted cointegration tests.

Table 5: Unbiasedness and Serial Correlation tests by OLS question 2.

	$S_t = \alpha + \beta S_t^e + u_t$				
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	-0.3752	-0.3016	-13.3096***‡	-6.7867***	-1.3211***
	(0.9048)†	(0.9985)	(2.4780)	(1.8139)	(2.3076)
β	1.0168	1.0101***	2.359***	1.6852***	2.3574***
	(0.0766)	(0.0851)	(0.2464)	(0.1747)	(0.2299)
R^2	0.7789	0.738	0.647	0.6506	0.6777
adjusted R^2	0.7745	0.7328	0.6399	0.6436	0.6712
F-statistic	176.2***	140.8***	91.64***	93.09***	105.1***
Wald test‖					
χ^2	0.2238	0.1513	30.439***	15.422***	34.864***
F	0.1119	0.0756	15.219***	7.711***	17.432***
LM.OSC 12 ††	18.4087	17.2794	17.7599	16.1119	21.5569**
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

†† OSC = Order ... Serial Correlation; testing the H_0 : no correlation among the errors. If H_0 is rejected then the rational hypothesis is rejected.

TABLE 6: Efficiency tests by OLS question 2.

$u_t = \beta_1 S_{t-1} + \beta_2 S_{t-2} + \beta_3 S_{t-3} + \beta_4 S_{t-4} + \beta_5 S_{t-5} + \beta_6 S_{t-6} + \beta_7 S_{t-7} + \beta_8 S_{t-8} + v_t$

Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
β_1	0.2939*‡	0.3057*	0.0419	0.0705	-0.0521
	(0.1720)†	(0.1891)	(0.2091)	(0.2075)	(0.2053)
β_2	-0.3094*	-0.2471	0.3332	0.3166	0.2743
	(0.1730)	(0.1903)	(0.2104)	(0.2088)	(0.2065)
β_3	0.3450*	0.3838*	0.3542*	0.4126*	0.4494*
	(0.1874)	(0.2061)	(0.2279)	(0.2261)	(0.2237)
β_4	-0.0397	-0.0413	0.1501	0.1015	0.1359
	(0.1937)	(0.2130)	(0.2355)	(0.2337)	(0.2312)
β_5	-0.0384	-0.1114	-0.1049	-0.2137	-0.1689
	(0.2022)	(0.2224)	(0.2459)	(0.2440)	(0.2414)
β_6	0.0103	-0.0398	-0.3004	-0.2462	-0.1979
	(0.1686)	(0.1854)	(0.2050)	(0.2034)	(0.2012)
β_7	-0.2527	-0.2531	-0.2627	-0.2658	-0.2633
	(0.1588)	(0.1746)	(0.1931)	(0.1916)	(0.1895)
β_8	0.0243	0.0526	-0.1140	-0.0833	-0.0990
	(0.1554)	(0.1709)	(0.1889)	(0.1875)	(0.1855)
R^2	0.2466	0.2311	0.3024	0.306	0.266
adjusted R^2	0.1096	0.09124	0.1756	0.1799	0.1326
F -statistic	1.8	1.653	2.384**	2.425**	1.994
N	52	52	52	52	52

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 7: Orthogonality test with S_{t-1} as information variable, for question 2.

$S_t = \alpha + \beta S_t^e + \gamma S_{t-1} + u_t$

Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	-0.1833	-0.09118	-4.4334*‡	-2.0916	-4.1608*
	(0.8064)†	(0.84366)	(2.3257)	(1.5738)	(2.4478)
β	0.5371***	0.44677**	0.7877**	0.5415**	0.7722**
	(0.1444)	(0.14174)	(0.3092)	(0.2283)	(0.3385)
γ	0.4652***	0.54731***	0.6577***	0.6621***	0.6476***
	(0.1235)	(0.11872)	(0.1038)	(0.1076)	(0.1166)
R^2	0.8286	0.8173	0.8059	0.8028	0.8022
adjustedR^2	0.8216	0.8098	0.798	0.7948	0.7941
F-statistic	118.4***	109.6***	101.7***	99.76***	99.34***
Wald test‖					
χ^2	14.479***	21.466***	94.375***	64.633***	86.501***
F	4.8265***	7.1554**	31.458***	21.544***	28.834***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 8: Orthogonality test with MER_{t-2} as information variable, for question 2.

$S_t = \alpha + \beta S_t^e + \gamma MER_{t-2} + u_t$

Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	-0.3810	-0.3105	-13.3414***‡	-6.8124***	-13.3343***
	(0.9138)†	(1.0085)	(2.5070)	(1.8343)	(2.3334)
β	1.0180***	1.0119***	2.3631***	1.6888***	2.3722***
	(0.07754)	(0.08618)	(0.24963)	(0.1769)	(0.23296)
γ	0.02916	0.03622	0.03219	0.0379	0.08637
	(0.12527)	(0.13642)	(0.15844)	(0.1577)	(0.15167)
R^2	0.7792	0.7384	0.6473	0.651	0.6798
adjusted R^2	0.7702	0.7277	0.6329	0.6367	0.6667
F-statistic	86.45***	69.15***	44.96***	45.69***	52.01***
Wald test‖					
χ^2	0.2738	0.2189	29.896***	15.189***	34.717***
F	0.0913	0.073	9.9654***	5.0631***	11.572***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 9: Orthogonality test with PPI_{t-1} as information variable, for question 2.

$S_t = \alpha + \beta S_t^e + \gamma PPI_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	-0.4232	-0.3362	-14.6134***‡	-8.0029***	-15.7273***
	(1.0643)†	(1,1712)	(2.6991)	(2.0440)	(2.5210)
β	1.0173***	1.0105***	2.4171***	1.7301***	2.4879***
	(0.0775)	(0.0862)	(0.2502)	(0.1772)	(0.2304)
γ	0.0122	0.0088	0.2107	0.2221	0.3569**
	(0.1395)	(0.1519)	(0.1767)	(0.1758)	(0.1669)
R^2	0.779	0.738	0.6569	0.6616	0.7052
adjusted R^2	0.7699	0.7273	0.6429	0.6478	0.6931
F -statistic	86.34	69.02***	46.92***	47.9***	58.6***
Wald test‖					
χ^2	0.2271	0.1516	32.116***	17.202***	41.925***
F	0.0757	0.0505	10.705***	5.7341***	13.975***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 10: Orthogonality test with IPI_{t-2} as information variable, for question 2.

$S_t = \alpha + \beta S_t^e + \gamma IPI_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.7586	1.4121	-5.1226	-1.2549	-5.9224*‡
	(1.2175)†	(1.3030)	(3.0658)	(2.2896)	(3.3640)
β	0.8496***	0.7583***	1.3824***	0.9870***	1.4906***
	(0.1432)	(0.1522)	(0.3358)	(0.2560)	(0.3746)
γ	0.1434	0.2138*	0.3671***	0.3559***	0.3103***
	(0.1041)	(0.1084)	(0.0959)	(0.1027)	(0.1098)
R^2	0.7872	0.7573	0.7282	0.7194	0.7229
adjusted R^2	0.7785	0.7474	0.7171	0.7079	0.7116
F -statistic	90.61***	76.44***	65.65***	62.81***	63.91***
Wald test‖					
χ^2	2.1241	4.05	53.397***	30.839***	47.736***
F	0.708	1.35	17.799***	10.280***	15.912***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 11: Orthogonality test with S_{t-1}, MER_{t-2}, PPI_{t-1} and IPI_{t-2} as information variables, for question 2.

$S_t = \alpha + \beta S_t^e + \gamma_1 S_{t-1} + \gamma_2 MER_{t-2} + \gamma_3 PPI_{t-1} + \gamma_4 IPI_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.3742	0.6892	-2.8770	-0.7323	-2.7849
	(1.1816)†	(1.2228)	(2.8390)	(2.0988)	(3.3901)
β	0.49085***‡	0.38402**	0.66780*	0.43262*	0.65278
	(0.1693)	(0.1646)	(0.3378)	(0.2560)	(0.4120)
γ_1	0.4511***	0.5179***	0.5445***	0.5638***	0.5445***
	(0.1353)	(0.1327)	(0.1337)	(0.1350)	(0.1465)
γ_2	0.0326	0.0389	0.0241	0.0248	0.0259
	(0.1235)	(0.1269)	(0.1292)	(0.1306)	(0.1312)
γ_3	-0.0409	-0.0428	0.0488	0.0393	0.0776
	(0.1454)	(0.1497)	(0.1563)	(0.1581)	(0.1674)
γ_4	0.0517	0.0790	0.1529	0.1472	0.1447
	(0.1079)	(0.1105)	(0.1023)	(0.1051)	(0.1061)
R^2	0.8302	0.8205	0.8149	0.8109	0.8101
adjusted R^2	0.8118	0.8009	0.7948	0.7904	0.7894
F -statistic	44.99***	42.04**	40.51***	39.46***	39.24***
Wald test‖					
χ^2	14.168**	21.325***	95.162***	65.25***	86.506***
F	2.3613**	3.5542***	15.860***	10.875***	14.418***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 12: Unbiasedness tests with Hansen and Hodrick correction question 2.

$S_t = \alpha + \beta S_t^e + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	-0.3752	-0.3016	-13.3096	-6.7867	-13.2109
	(7.2319)†	(8.5205)	(31.1082)	(20.4776)	(29.2455)
β	1.0168*‡	1.0101	2.3591	1.6852	2.3574
	(0,5966)	(0.6957)	(2.9707)	(1.8725)	(2,7932)
R^2	0.7789	0.738	0.647	0.6506	0.6777
adjusted R^2	0.7745	0.7328	0.6399	0.6436	0.6712
Wald test‖					
χ^2	0.003486278	0.001466486	0.392368	0.2437437	0.4402235
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 13: Efficiency tests with Hansen and Hodrick correction question 2.

$u_t = \beta_1 S_{t-1} + \beta_2 S_{t-2} + \beta_3 S_{t-3} + \beta_4 S_{t-4} + \beta_5 S_{t-5} + \beta_6 S_{t-6} + \beta_7 S_{t-7} + \beta_8 S_{t-8} + v_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
β_1	0.2939	0.3057	0.0419	0.0705	-0.0521
	(1.1371)†	(1.2570)	(1.3959)	(1.3692)	(1.3897)
β_2	-0.3094	-0.2471	0.3332	0.3166	0.2743
	(1.1487)	(1.2589)	(1.3909)	(1.3777)	(1.3560)
β_3	0.3450	0.3838	0.3542	0.4125	0.4494
	(1.2488)	(1.3636)	(1.5038)	(1.4907)	(1.4568)
β_4	-0.0397	-0.0413	0.1501	0.1015	0.1359
	(1.2757)	(1.3972)	(1.5686)	(1.5418)	(1.5316)
β_5	-0.0384	-0.1114	-0.1049	-0.2137	-0.1689
	(1.3369)	(1.4661)	(1.6424)	(1.6283)	(1.5956)
β_6	0.0102	-0.0398	-0.3004	-0.2462	-0.1979
	(1.1377)	(1.2423)	(1.3559)	(1.3502)	(1.3021)
β_7	-0.2527	-0.2531	-0.2627	-0.2658	-0.2633
	(1.0567)	(1.1501)	(1.2481)	(1.2397)	(1.2079)
β_8	0.0242	0.0526	-0.1140	-0.0833	-0.0990
	(1.0405)	(1.1430)	(1,2523)	(1.2431)	(1.2147)
R^2	0.2466	0.2311	0.3024	0.306	0.266
adjusted R^2	0.1096	0.09124	0.1756	0.1799	0.1326
N	52	52	52	52	52

† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 14: Orthogonality test with S_{t-1} as information variable and Hansen Hodrick correction question 2.

$S_t = \alpha + \beta S_t^e + \gamma S_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	-0.1832	-0.0911	-4.4331	-2.0916	-4.1608
	(5.5455)†	(5.8321)	(16.2375)	(10.9572)	(17.0546)
β	0.5370	0.4467	0.7876	0.5414	0.7721
	(1.0447)	(1.0154)	(2.0544)	(1.5138)	(2.2452)
γ	0.4651	0.5473	0.6577	0.6620	0.6475
	(0.8946)	(0.8517)	(0.702)	(0.7315)	(0.7776)
R^2	0.8286	0.8173	0.8059	0.8028	0.8022
adjusted R^2	0.8216	0.8098	0.798	0.7948	0.7941
Wald test‖					
χ^2	0.4678	0.7101	0.9642	0.9474	0.7634
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 15: Orthogonality test with MER_{t-2} as information variable and Hansen Hodrick correction question 2.

$S_t = \alpha + \beta S_t^e + \gamma MER_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	-0.3810	-0.3105	-13.3414	-6.8123	-13.3343
	(7.2155)†	(8.5095)	(31.4805)	(20.6374)	(29.6241)
β	1.0180**‡	1.0119*	2.3631	1.6887	2.3722
	(0.5987)	(0.7001)	(3.0209)	(1.8999)	(2.8478)
γ	0.0291	0.0362	0.0321	0.0379	0.0863
	(0,8810)	(0.9671)	(1.1653)	(1.1535)	(1.1263)
R^2	0.7792	0.7384	0.6473	0.651	0.6798
adjusted R^2	0.7702	0.7277	0.6329	0.6367	0.6667
Wald test‖					
χ^2	0.0047	0.0030	0.3839	0.2414	0.4406
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 16: Orthogonality test with PPI_{t-1} as information variable and Hansen Hodrick correction question 2.

$S_t = \alpha + \beta S_t^e + \gamma PPI_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	-0.4232	-0.3362	-14.6134	-8.0028	-15.7273
	(8.4651)†	(10.0479)	(35.1934)	(23.6066)	(33.1838)
β	1.0173**‡	1.0105*	2.4171	1.7300	2.4878
	(0.5996)	(0.7006)	(3.0898)	(1.9161)	(2.8738)
γ	0.0122	0.0088	0.2107	0.2220	0.3569
	(1.0047)	(1.1431)	(1.5247)	(1.4676)	(1.4876)
R^2	0.779	0.738	0.6569	0.6616	0.7052
adjusted R^2	0.7699	0.7273	0.6429	0.6478	0.6931
Wald test‖					
χ^2	0.0034	0.0014	0.4018	0.2830	0.5502
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 17: Orthogonality test with IPI_{t-2} as information variable and Hansen Hodrick correction question 2.

$S_t = \alpha + \beta S_t^e + \gamma IPI_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.7585	1.4121	-5.1226	-1.2548	-5.9224
	(9.6215)†	(10.4423)	(31.4487)	(21.7297)	(35.5760)
β	0.8495	0.7582	1.3824	0.9870	1.4906
	(1.1024)	(1.1700)	(3.2767)	(2.2757)	(3.8010)
γ	0.1433	0.2138	0.3671	0.3558	0.3103
	(0.7542)	(0.7857)	(0.6786)	(0.7180)	(0.8198)
R^2	0.7872	0.7573	0.7282	0.7194	0.7229
adjusted R^2	0.7785	0.7474	0.7171	0.7079	0.7116
Wald test‖					
χ^2	0.0609	0.1350	0.3329	0.2490	0.1876
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 18: Orthogonality test with S_{t-1}, MER_{t-2}, PPI_{t-1}, IPI_{t-2} as information variable and Hansen Hodrick correction question 2.

$S_t = \alpha + \beta S_t^e + \gamma_1 S_{t-1} + \gamma_2 MER_{t-2} + \gamma_3 PPI_{t-1} + \gamma_4 IPI_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.3742	0.6892	-2.8770	-0.7323	-2.7849
	(8.1915)†	(8.5302)	(20.4118)	(14.6932)	(24.8905)
β	0.4908	0.3840	0.6677	0.4326	0.6527
	(1.1798)	(1.1340)	(2.2492)	(1.6495)	(2.8157)
γ_1	0.4511	0.5179	0.5445	0.5638	0.5445
	(0.9669)	(0.9530)	(0.9798)	(0.9846)	(1.0352)
γ_2	0.0326	0.0389	0.0241	0.0248	0.0259
	(0.8280)	(0.8467)	(0.8421)	(0.8521)	(0.8569)
γ_3	-0.0409	-0.0428	0.0488	0.0393	0.0776
	(0.9562)	(0.9895)	(1.0622)	(1.0583)	(1.1453)
γ_4	0.0517	0.0790	0.1529	0.1471	0.1447
	(0.7599)	(0.7846)	(0.7518)	(0.7635)	(0.7811)
R^2	0.8302	0.8205	0.8149	0.8109	0.8101
adjusted R^2	0.8118	0.8009	0.7948	0.7904	0.7894
Wald test‖					
χ^2	0.4140	0.6111	0.3948	0.4882	0.3443
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 19: Unbiasedness and Serial Correlation tests by OLS question 9.

$P_t = \alpha + \beta P_t^e + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.1300	0.1069	-2.7941	-1.4007	-2.8873
	(0.4499)†	(0.4619)	(1.6719)	(1.3892)	(1.8107)
β	0.9566***‡	0.9632***	1.8114***	1.4075***	1.8393***
	(0.09604)	(0.09929)	(0.4649)	(0.3785)	(0.5061)
R^2	0.665	0.6531	0.2329	0.2166	0.2089
adjusted R^2	0.6583	0.6461	0.2176	0.201	0.1931
F-statistic	99.24***	94.12***	15.18***	13.83***	13.21***
Wald test‖					
χ^2	0.2085	0.1421	3.0477	1.1596	2.7509
F	0.1043	0.071	1.5238	0.5798	1.3755
LM OSC 12††	38.8449***	37.7988***	43.0731***	43.241***	44.9366***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
†† OSC = Order ... Serial Correlation; testing the H_0 : no correlation among the errors. If H_0 is rejected then the rational hypothesis is rejected.

TABLE 20: Efficiency tests by OLS question 9.

$u_t = \beta_1 P_{t-1} + \beta_2 P_{t-2} + \beta_3 P_{t-3} + \beta_4 P_{t-4} + \beta_5 P_{t-5} + \beta_6 P_{t-6} + \beta_7 P_{t-7} + \beta_8 P_{t-8} + v_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
β_1	1.0853***‡	1.0675***	0.6612*	0.7173*	0.6594*
	(0.2527)†	(0.2676)	(0.3649)	(0.3622)	(0.3598)
β_2	-1.1349**	-1.13470**	-0.4884	-0.5771	-0.4770
	(0.4847)	(0.5133)	(0.6998)	(0.6945)	(0.6901)
β_3	0.5139	0.5418	0.4260	0.4874	0.4125
	(0.5198)	(0.5505)	(0.7504)	(0.7448)	(0.7400)
β_4	-0.5755	-0.6271	-0.3916	-0.4130	-0.3199
	(0.5345)	(0.5659)	(0.7715)	(0.7657)	(0.7608)
β_5	0.6086	0.6961	0.8154	0.8519	0.7431
	(0.5357)	(0.5673)	(0.7734)	(0.7676)	(0.7627)
β_6	-0.5728	-0.6346	-0.8417	-0.8800	-0.8253
	(0.5378)	(0.5695)	(0.7763)	(0.7705)	(0.7656)
β_7	0.1490	0.2260	0.7968	0.7824	0.8490
	(0.5166)	(0.5470)	(0.7457)	(0.7401)	(0.7353)
β_8	-0.0091	-0.0600	-0.7338*	-0.7278*	-0.7923*
	(0.2804)	(0.2969)	(0.4048)	(0.4018)	(0.3992)
R^2	0.5088	0.4681	0.553	0.5689	0.5785
adjusted R^2	0.4195	0.3714	0.4718	0.4905	0.5019
F -statistic	5.698***	4.841***	6.805***	7.257***	7.549***
N	52	52	52	52	52

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 21: Orthogonality test with P_{t-1} as information variable question 9.

$P_t = \alpha + \beta P_t^e + \gamma P_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.1104	0.0875	-1.679***†	-1.1334**	-1.9572***
	(0.2730)‡	(0.2748)	(0.613)	(0.5122)	(0.6422)
β	0.0638	0.0814	0.5937***	0.4288***	0.6706***
	(0.1121)	(0.1091)	(0.1825)	(0.1498)	(0.1895)
γ	0.8954***	0.8844***	0.8835***	0.8901***	0.8870***
	(0.0961)	(0.0920)	(0.0489)	(0.0498)	(0.0473)
R^2	0.8791	0.8797	0.8999	0.8957	0.9031
adjusted R^2	0.8742	0.8747	0.8958	0.8915	0.8991
F -statistic	178.1***	179.1***	220.3***	210.4***	228.2***
Wald test‖					
χ^2	87.34***	92.66***	349.38***	327.59***	372.84***
F	29.113***	30.886***	116.46***	109.20***	124.28***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 22: Orthogonality test with MER_{t-2} as information variable question 9.

$P_t = \alpha + \beta P_t^e + \gamma MER_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.2502	0.2348	-1.7089	-0.5015	-1.7879
	(0.4872)†	(0.5006)	(1.6635)	(1.3797)	(1.7769)
β	0.9260***‡	0.9305***	1.5157***	1.1661***	1.5404***
	(0.1070)	(0.1107)	(0.4618)	(0.3755)	(0.4959)
γ	0.0547	0.05736	0.2638**	0.2687**	0.2790**
	(0.0824)	(0.0838)	(0.1111)	(0.1122)	(0.1113)
R^2	0.668	0.6563	0.312	0.2988	0.2989
adjusted R^2	0.6544	0.6423	0.2839	0.2702	0.2702
F -statistic	49.29***	46.79***	11.11***	10.44***	10.44***
Wald test‖					
χ^2	0.6483	0.6082	8.9621**	7.0104*	9.3267**
F	0.2161	0.2027	2.9874**	2.3368*	3.1089**
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 23: Orthogonality test with $M3_{t-1}$ as information variable question 9.

$P_t = \alpha + \beta P_t^e + \gamma M3_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.2493	0.2256	-2.8228	-1.4501	-2.9258
	(0.4828) †	(0.4947)	(1.6926)	(1.4140)	(1.8322)
β	0.9695*** ‡	0.9765***	1.7966***	1.3939***	1.8203***
	(0.0982)	(0.1016)	(0.4738)	(0.3852)	(0.5149)
γ	-0.1315	-0.1321	0.0638	0.0772	0.0832
	(0.1865)	(0.1899)	(0.2812)	(0.2838)	(0.2851)
R^2	0.6683	0.6565	0.2337	0.2178	0.2103
adjusted R^2	0.6548	0.6424	0.2025	0.1859	0.1781
F -statistic	49.37***	46.82***	7.473***	6.823***	6.525***
Wald test‖					
χ^2	0.7038	0.6247	3.0414	1.2123	2.7859
F	0.2346	0.2082	1.0138	0.4041	0.9286
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:α= 0,β= 1, γ= 0 . If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 24: Orthogonality test with P_{t-1}, MER_{t-2}, $M3_{t-1}$ as information variable question 9.

$P_t = \alpha + \beta P_t^e + \gamma_1 P_{t-1} + \gamma_2 MER_{t-2} + \gamma_3 M3_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.1691	0.1473	-1.6066***‡	-1.0665*	-1.8822***
	(0.3251)†	(0.3270)	(0.6523)	(0.5528)	(0.6785)
β	0.0422	0.0617	0.5815***	0.4173***	0.6586***
	(0.1186)	(0.1155)	(0.1887)	(0.154822)	(0.1953)
γ_1	0.8955***	0.8835***	0.8760***	0.8819***	0.8789***
	(0.0990)	(0.0948)	(0.0526)	(0.0536)	(0.0512)
γ_2	0.0350	0.0333	0.0200	0.0217	0.0212
	(0.0514)	(0.0514)	(0.0467)	(0.0476)	(0.0458)
γ_3	0.0199	0.0162	-0.0001	0.0053	0.0001
	(0.1179)	(0.1176)	(0.1061)	(0.1082)	(0.1043)
R^2	0.8805	0.8809	0.9003	0.8962	0.9035
adjusted R^2	0.8703	0.8708	0.8918	0.8874	0.8953
F -statistic	86.58***	86.91***	106.1***	101.5***	110***
Wald test‖					
χ^2	85.322***	90.3***	336.69***	316***	359.56***
F	17.064***	18.06***	67.337***	63.2***	71.912***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:α= 0,β= 1, γ= 0 . If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 25: Unbiasedness tests with Hansen and Hodrick correction question 9.

$P_t = \alpha + \beta P_t^e + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.1300	0.1069	-2.7941	-1.4006	-2.8872
	(4.6061)†	(4.7575)	(23.8860)	(19.3813)	(26.4556)
β	0.9566	0.9632	1.8114	1.4074	1.8392
	(0.8196)	(0.8544)	(6.4540)	(5,1573)	(7,1860)
R^2	0.665	0.6531	0.2329	0.2166	0.2089
adjusted R^2	0.6583	0.6461	0.2176	0.201	0.1931
Wald test‖					
χ^2	0.0035	0.0023	0.0294	0.0114	0.0255
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:α= 0,β= 1 . If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 26: Efficiency tests with Hansen and Hodrick correction question 9.

$u_t = \beta_1 P_{t-1} + \beta_2 P_{t-2} + \beta_3 P_{t-3} + \beta_4 P_{t-4} + \beta_5 P_{t-5} + \beta_6 P_{t-6} + \beta_7 P_{t-7} + \beta_8 P_{t-8} + v_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
β_1	1.0853	1.0674	0.6611	0.7173	0.6593
	(1.6476)†	(1.7389)	(2.2721)	(2,2610)	(2,2282)
β_2	-1.1349	-1.1347	-0.4884	-0.5770	-0.4770
	(2.8641)	(3.0105)	(3.8336)	(3,8159)	(3,7332)
β_3	0.5139	0.5418	0.4260	0.4874	0.4125
	(2.7332)	(2.8858)	(3.9358)	(3,9090)	(3,8289)
β_4	-0.5755	-0.6271	-0.3916	-0.4129	-0.3199
	(2.7601)	(2.9172)	(3.9647)	(3,9462)	(3,8422)
β_5	0.6086	0.6961	0.8153	0.8518	0.7430
	(2.8105)	(2.9667)	(3.9029)	(3,8920)	(3,7725)
β_6	-0.5728	-0.6346	-0.8417	-0.8799	-0.8252
	(3.0761)	(3.2309)	(4.0643)	(4,0547)	(3,9467)
β_7	0.1490	0.2260	0.7967	0.7824	0.8490
	(3.1474)	(3.2958)	(3.9544)	(3,9540)	(3,8434)
β_8	-0.0091	-0.0600	-0.7338	-0.7278	-0.7923
	(1.8502)	(1.9480)	(2.4722)	(2,4648)	(2,4242)
R^2	0.5088	0.4681	0.553	0.5689	0.5785
adjusted R^2	0.4195	0.3714	0.4718	0.4905	0.5019
N	52	52	52	52	52

† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 27: Orthogonality test with P_{t-1} as information variable and Hansen Hodrick correction question 9.

$P_t = \alpha + \beta P_t^e + \gamma P_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.1104	0.0875	-1.6796	-1.1334	-1.9572
	(2.5251)†	(2.5428)	(5.8454)	(4.8321)	(6.2624)
β	0.0638	0.0814	0.5937	0.4288	0.6706
	(0.9506)	(0.9352)	(1.7178)	(1.4046)	(1.8168)
γ	0.8954	0.8844	0.8834**‡	0.8900**	0.8870**
	(0.8640)	(0.8284)	(0.3876)	(0.3987)	(0,3726)
R^2	0.8791	0.8797	0.8999	0.8957	0.9031
adjusted R^2	0.8742	0.8747	0.8958	0.8915	0.8991
Wald test‖					
χ^2	2.0460	2.1057	5.3352	5.2044	5.7971
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 28: Orthogonality test with MER_{t-2} as information variable and Hansen Hodrick correction question 9.

$P_t = \alpha + \beta P_t^e + \gamma MER_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.2502	0.2348	-1.7088	-0.5014	-1.7879
	(4.7485)†	(4.8909)	(21.6063)	(17.4986)	(23.5648)
β	0.9260	0.9305	1.5157	1.1661	1.5404
	(0.8620)	(0.8935)	(5.8503)	(4.6708)	(6,4124)
γ	0.0547	0.0573	0.2637	0.2687	0.2790
	(0.5724)	(0.5757)	(0.6934)	(0.7026)	(0.6979)
R^2	0.668	0.6563	0.312	0.2988	0.2989
adjusted R^2	0.6544	0.6423	0.2839	0.2702	0.2702
Wald test‖					
χ^2	0.0193	0.0182	0.1587	0.1483	0.1727
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness byH_0:$\alpha = 0, \beta = 1\gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 29: Orthogonality test with $M3_{t-2}$ as information variable and Hansen Hodrick correction question 9.

$P_t = \alpha + \beta P_t^e + \gamma M3_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.2493	0.2255	-2.8228	-1.4501	-2.9258
	(4.7699)†	(4.9244)	(23.9604)	(19.5093)	(26.5471)
β	0.9695	0.9764	1.7966	1.3939	1.8203
	(0.8256)	(0.8582)	(6.4563)	(5.1604)	(7.1712)
γ	-0.1315	-0.1321	0.0638	0.0772	0.0832
	(1.2956)	(1.3116)	(1.8766)	(1.9078)	(1,8920)
R^2	0.6683	0.6565	0.2337	0.2178	0.2103
adjusted R^2	0.6548	0.6424	0.2025	0.1859	0.1781
Wald test‖					
χ^2	0.0144	0.0129	0.0302	0.0129	0.0271
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 30: Orthogonality test with P_{t-1}, MER_{t-2} and $M3_{t-2}$ as information variable and Hansen Hodrick correction question 9.

$P_t = \alpha + \beta P_t^e + \gamma_1 P_{t-1} + \gamma_2 MER_{t-2} + \gamma_3 M3_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.1691	0.1473	-1.6066	-1.0665	-1.8822
	(2.8078)†	(2.8263)	(5.9434)	(4.9616)	(6.3311)
β	0.0422	0.0617	0.5815	0.4173	0.6586
	(0.9755)	(0.9569)	(1.7114)	(1.3978)	(1.8047)
γ_1	0.8955	0.8835	0.8760**‡	0.8819**	0.8789***
	(0.8647)	(0.8306)	(0.4103)	(0.4215)	(0.3968)
γ_2	0.03502	0.0333	0.0200	0.0217	0.0212
	(0.3667)	(0.3645)	(0.3289)	(0.3358)	(0.3232)
γ_3	0.0199	0.0162	-0.0001	0.0053	0.0001
	(0.8168)	(0.8152)	(0.7400)	(0.7560)	(0.7276)
R^2	0.8805	0.8809	0.9003	0.8962	0.9035
adjusted R^2	0.8703	0.8708	0.8918	0.8874	0.8953
Wald test‖					
χ^2	2.0500	2.1045	4.6949	4.6013	5.0347
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 31: Unbiasedness and Serial Correlation tests by OLS question 11.

$P_t = \alpha + \beta P_t^e + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.3335	0.3467	-3.4767**‡	-1.3048	-3.673**
	(0.561)†	(0.5935)	(1.6759)	(1.3689)	(1.747)
β	0.921***	0.9192***	1.7199***	1.2712***	1.760***
	(0.0992)	(0.1056)	(0.3372)	(0.2711)	(0.352)
R^2	0.6329	0.6024	0.3422	0.3054	0.3333
adjusted R^2	0.6255	0.5944	0.3291	0.2915	0.32
F -statistic	86.2***	75.75***	26.01***	21.98***	25***
Wald test‖					
χ^2	0.6721	0.6147	4.5589	1.0014	4.6611*
F	0.336	0.3074	2.2795	0.5007	2.3305
LM OSC 12††	41,1284***	40.5504***	42.492***	42.7165***	44.1127***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
†† OSC = Order ... Serial Correlation; testing the H_0 : no correlation among the errors. If H_0 is rejected then the rational hypothesis is rejected.

TABLE 32: Efficiency tests by OLS question 11.

$u_t = \beta_1 P_{t-1} + \beta_2 P_{t-2} + \beta_3 P_{t-3} + \beta_4 P_{t-4} + \beta_5 P_{t-5} + \beta_6 P_{t-6} + \beta_7 P_{t-7} + \beta_8 P_{t-8} + v_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
β_1	1.3276***‡	1.3014***	0.7293**	0.7698**	0.6903**
	(0.2202)†	(0.2395)	(0.3389)	(0.3379)	(0.3378)
β_2	-1.4680***	-1.4017***	-0.5923	-0.6023	-0.5849
	(0.3962)	(0.4308)	(0.6097)	(0.6079)	(0.6078)
β_3	0.4327	0.4003	0.2593	0.2663	0.2716
	(0.4230)	(0.4599)	(0.6509)	(0.6489)	(0.6488)
β_4	-0.2804	-0.2568	-0.0817	-0.0723	-0.0543
	(0.4381)	(0.4763)	(0.6741)	(0.6721)	(0.6720)
β_5	0.2264	0.2022	0.2568	0.2457	0.2982
	(0.4396)	(0.4780)	(0.6765)	(0.6745)	(0.6744)
β_6	-0.3410	-0.3364	-0.3445	-0.3552	-0.3332
	(0.4347)	(0.4726)	(0.6689)	(0.6669)	(0.6668)
β_7	0.2913	0.3381	0.4574	0.4536	0.4148
	(0.4127)	(0.4487)	(0.6351)	(0.6332)	(0.6331)
β_8	-0.1389	-0.1898	-0.5509	-0.5687	-0.5644
	(0.2320)	(0.2522)	(0.3569)	(0.3559)	(0.3558)
R^2	0.5354	0.4929	0.3858	0.422	0.3978
adjustedR^2	0.4509	0.4007	0.2742	0.3169	0.2883
F -statistic	6.337***	5.345***	3.455***	4.016***	3.633***
N	52	52	52	52	52

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 33: Orthogonality test with P_{t-1} as information variable question 11.

$P_t = \alpha + \beta P_t^e + \gamma P_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.3853	0.3823	-1.2100	-0.6584	-1,5631**‡
	(0.3421)†	(0.3482)	(0.7500)	(0.5919)	(0.7577)
β	-0.0918	-0.0781	0.3756**	0.2486*	0.4558**
	(0.1251)	(0.1190)	(0.1741)	(0.1357)	(0.1732)
γ	1.0060***	0.9928***	0.8679***	0.8808***	0.8606***
	(0.1088)	(0.1011)	(0.0596)	(0.0594)	(0.0572)
R^2	0.8662	0.8659	0.8765	0.8734	0.8813
adjusted R^2	0.8607	0.8604	0.8714	0.8682	0.8765
F -statistic	158.6***	158.2***	173.8***	169***	181.9***
Wald test‖					
χ^2	87.243***	98.079***	235.67***	225.21***	251.88***
F	29.081***	32.693***	78.556***	75.07***	83.96***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 34: Orthogonality test with MER_{t-2} as information variable question 11.

$P_t = \alpha + \beta P_t^e + \gamma MER_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.4864	0.5076	-2.6471	-0.6182	-2.8272
	(0.5995)†	(0.6368)	(1.7360)	(1.4056)	(1.7783)
β	0.8925***‡	0.8890***	1.5559***	1.1374***	1.5934***
	(0.1067)	(0.1140)	(0.3487)	(0.2780)	(0.3577)
γ	0.0551	0.0556	0.1481	0.1624*	0.1648*
	(0.0738)	(0.0771)	(0.0950)	(0.0968)	(0.0939)
R^2	0.637	0.6066	0.3733	0.3431	0.3727
adjusted R^2	0.6222	0.5905	0.3477	0.3163	0.3471
F -statistic	43***	37.77***	14.59***	12.8***	14.56***
Wald test‖					
χ^2	1.2237	1.1289	7.1196*	3.8525	7.9321**
F	0.4079	0.3763	2.3732*	0.2902	2.644*
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.

† Standard errors in parentheses

‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 35: Orthogonality test with $M3_{t-1}$ as information variable question 11.

$P_t = \alpha + \beta P_t^e + \gamma M3_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.3540	0.3542	-3.5557***‡	-1.4424	-3.7411**
	(0.5830)†	(0.6168)	(1.6941)	(1.3927)	(1.7649)
β	0.9233***	0.9200***	1.7042***	1.2600***	1.7433***
	(0.1014)	(0.1079)	(0.3408)	(0.2732)	(0.3561)
γ	-0.0257	-0.0093	0.1243	0.1538	0.1194
	(0.17312)	(0.1800)	(0.2290)	(0.2347)	(0.2307)
R^2	0.633	0.6024	0.3462	0.3114	0.3369
adjusted R^2	0.6181	0.5862	0.3195	0.2833	0.3099
F -statistic	42.27***	37.12***	12.97***	11.08***	12.45***
Wald test‖					
χ^2	0.681	0.6052	4.7894	1.4198	4.8607
F	0.227	0.2017	1.5965	0.4733	1.6202
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 36: Orthogonality test with P_{t-1}, MER_{t-2} and $M3_{t-1}$ as information variable question 11.

$P_t = \alpha + \beta P_t^e + \gamma_1 P_{t-1} + \gamma_2 MER_{t-2} + \gamma_3 M3_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.5093	0.5158	-1.0763	-0.5428	-1.4357**‡
	(0.3833)†	(0.3916)	(0.8014)	(0.6417)	(0.7995)
β	-0.1179	-0.1051	0.3561*	0.2335	0.4400**
	(0.1286)	(0.1225)	(0.1794)	(0.1395)	(0.1775)
γ_1	1.0055***	0.9920***	0.8598***	0.8713***	0.8514***
	(0.1099)	(0.1020)	(0.0618)	(0.0617)	(0.0596)
γ_2	0.0488	0.0494	0.0282	0.0308	0.0293
	(0.0461)	(0.0463)	(0.0448)	(0.0452)	(0.0436)
γ_3	0.0144	0.0117	0.0051	0.0089	0.0007
	(0.1078)	(0.1077)	(0.1039)	(0.1051)	(0.1018)
R^2	0.8697	0.8694	0.8776	0.8748	0.8825
adjusted R^2	0.8586	0.8583	0.8672	0.8641	0.8725
F -statistic	78.39***	78.2***	84.23***	82.09***	88.24***
Wald test‖					
χ^2	87.152***	97.82***	228.57***	218.95***	244.52***
F	17.430***	19.564***	45.714***	43.791***	48.904***
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)

TABLE 37: Unbiasedness tests with Hansen and Hodrick correction question 11.

$P_t = \alpha + \beta P_t^e + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.3335	0.3466	-3.4766	-1.3047	-3.6727
	(5.7962)†	(6.2000)	(21.3691)	(17.0799)	(23.3012)
β	0.9210	0.9192	1.7199	1.2712	1.7599
	(0.8650)	(0.926)	(3.9649)	(3.0456)	(4.331)
R^2	0.6329	0.6024	0.3422	0.3054	0.3333
adjusted R^2	0.6255	0.5944	0.3291	0.2915	0.32
Wald test‖					
χ^2	0.0116	0.0107	0.0594	0.0137	0.0556
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 38: Efficiency tests with Hansen and Hodrick correction question 11.

$u_t = \beta_1 P_{t-1} + \beta_2 P_{t-2} + \beta_3 P_{t-3} + \beta_4 P_{t-4} + \beta_5 P_{t-5} + \beta_6 P_{t-6} + \beta_7 P_{t-7} + \beta_8 P_{t-8} + v_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
β_1	1.3276	1.3013	0.7293	0.7697	0.6903
	(1.4474)†	(1.5688)	(2.1654)	(2.1639)	(2.1503)
β_2	-1.4679	-1.4017	-0.5923	-0.6022	-0.5849
	(2.5093)	(2.7035)	(3.5178)	(3.5227)	(3.4562)
β_3	0.4327	0.4002	0.2593	0.2662	0.2716
	(2,5421)	(2.7276)	(3.6347)	(3.6334)	(3.5887)
β_4	-0.2803	-0.2567	-0.0817	-0.0723	-0.0543
	(2.6373)	(2.8332)	(3.7724)	(3.7728)	(3.7207)
β_5	0.2264	0.2021	0.2568	0.2456	0.2982
	(2.7005)	(2.9065)	(3.8625)	(3.8619)	(3.8077)
β_6	-0.3409	-0.3364	-0.3445	-0.3551	-0.3332
	(2.7524)	(2.9764)	(4.0017)	(3.9987)	(3.9545)
β_7	0.2913	0.3381	0.4574	0.4536	0.4148
	(2.7136)	(2.9419)	(3.9136)	(3.9129)	(3.8567)
β_8	-0.1388	-0.1897	-0.5509	-0.5687	-0.5644
	(1.6054)	(1.7531)	(2.4628)	(2.4591)	(2,4512)
R^2	0.5354	0.4929	0.3858	0.422	0.3978
adjusted R^2	0.4509	0.4007	0.2742	0.3169	0.2883
N	52	52	52	52	52

† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 39: Orthogonality test with P_{t-1} as information variable and Hansen Hodrick correction question 11.

$P_t = \alpha + \beta P_t^e + \gamma P_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.3853	0.3823	-1.2100	-0.6584	-1.5631
	(2.9489)†	(2.9989)	(6.1028)	(4.8656)	(6.1776)
β	-0.0918	-0.0781	0.3756	0.2486	0.4558
	(0.9689)	(0,9102)	(1.3596)	(1.0453)	(1.3582)
γ	1.0060	0.9928	0.8679**‡	0.8808**	0.8606**
	(0.8816)	(0.8142)	(0.4627)	(0.4627)	(0.4425)
R^2	0.8662	0.8659	0.8765	0.8734	0.8813
adjusted R^2	0.8607	0.8604	0.8714	0.8682	0.8765
Wald test‖					
χ^2	2.5890	2.9060	3.7690	4.1586	4.0065
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 40: Orthogonality test with MER_{t-2} as information variable and Hansen Hodrick correction question 11.

$P_t = \alpha + \beta P_t^e + \gamma MER_{t-2} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.4864	0.5076	-2.6471	-0.6181	-2.8272
	(6.1306)†	(6.5513)	(20.7389)	(16.4281)	(22.1554)
β	0.8925	0.8890	1.5558	1.1374	1.5934
	(0.9111)	(0.9745)	(3.8422)	(2.9227)	(4,1180)
γ	0.0551	0.0556	0.1481	0.1624	0.1648
	(0.5348)	(0.5554)	(0.6179)	(0.6320)	(0.6143)
R^2	0.637	0.6066	0.3733	0.3431	0.3727
adjusted R^2	0.6222	0.5905	0.3477	0.3163	0.3471
Wald test‖					
χ^2	0.0308	0.0289	0.0946	0.0696	0.1090
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 41: Orthogonality test with $M3_{t-2}$ as information variable and Hansen Hodrick correction question 11.

$P_t = \alpha + \beta P_t^e + \gamma M3_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.3540	0.3542	-3.5557	-1.4424	-3.7411
	(5.9232)†	(6.3404)	(21.5437)	(17.3899)	(23.4422)
β	0.9233	0.9200	1.7042	1.2600	1.7432
	(0.8674)	(0.9277)	(3.9278)	(3.0139)	(4.2931)
γ	-0.0257	-0.0093	0.1243	0.1538	0.1194
	(1.2101)	(1.2516)	(1.5350)	(1.5888)	(1.5401)
R^2	0.633	0.6024	0.3462	0.3114	0.3369
adjusted R^2	0.6181	0.5862	0.3195	0.2833	0.3099
Wald test‖					
χ^2	0.0118	0.0106	0.0659	0.0236	0.0614
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1, \gamma = 0$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

TABLE 42: Orthogonality test with P_{t-1}, MER_{t-2} and $M3_{t-2}$ as information variable and Hansen Hodrick correction question 11.

$P_t = \alpha + \beta P_t^e + \gamma_1 P_{t-1} + \gamma_2 MER_{t-2} + \gamma_3 M3_{t-1} + u_t$					
Method	RQA Normal	RQA Uniform	CP Normal	CP Uniform	MBS
α	0.5093	0.5158	-1.0763	-0.5428	-1.4357
	(3.2973)†	(3.3678)	(6.4512)	(5.2262)	(6.4411)
β	-0.1179	-0.1051	0.3561	0.2335	0.4400
	(0.9772)	(0.9209)	(1.3809)	(1.0581)	(1.3750)
γ_1	1.0055	0.9920	0.8598**	0.8713**‡	0.8514**
	(0.8690)	(0.8037)	(0.4772)	(0.4784)	(0.4593)
γ_2	0.0488	0.0494	0.0282	0.0308	0.0293
	(0.3336)	(0.3352)	(0.3220)	(0.3248)	(0.3147)
γ_3	0.0144	0.0117	0.0051	0.0089	0.0007
	(0.7570)	(0.7575)	(0.7349)	(0.7435)	(0.7205)
R^2	0.8697	0.8694	0.8776	0.8748	0.8825
adjusted R^2	0.8586	0.8583	0.8672	0.8641	0.8725
Wald test‖					
χ^2	2.6931	3.0092	3.4991	3.8624	3.6605
N	52	52	52	52	52

‖ Wald Test verifies the unbiasedness by H_0:$\alpha = 0, \beta = 1$. If H_0 it is rejected (statistically significant) then the rational hypothesis is rejected.
† Standard errors in parentheses
‡ The * denotes if the the estimator is significant at 10% (*), 5% (**) or 1% (***)
The correction of Hansen & Hodrick (1980) was applied to the covariance matrix

[Recibido: abril de 2011 — Aceptado: marzo de 2012]

References

Anderson, O. (1952), 'Business Test of the IFO-Institute for Economic Research, Munich, and its theoretical model', *Review of the International Statistical Institute* **20**(1), 1–17.

Banco de la República de Colombia (2010), Informe sobre inflación, Technical report, Banco de la República de Colombia, Departamento de Programación e Inflación, Bogotá.

Batchelor, R. (1986), 'Quantitative vs. qualitative measures of inflation expectations', *Oxford Bulletin of Economics and Statistics* (48), 99–120.

Berk, J. (1999), 'Measuring inflation expectations: A survey data approach', *Applied Economics* (31), 1467–1480.

Carlson, J. & Parkin, M. (1975), 'Inflation expectations', *Economica, New Series* (42), 123–138.

Claveria, O. (2010), Qualitative survey data on expectations: Is there an alternative to the balance statistic?, *in* A. T. Molnar, ed., 'Economic Forecasting', Nova Science Publishers, Hauppauge, New York, pp. 181–189.

Claveria, O., Pons, E. & Suriñach, J. (2003), 'Las encuestas de opinión empresarial como instrumento de control y predicción de los precios industriales', *Cuadernos Aragoneses de Economía* (13), 515–541.

Claveria, O. Pons, E. & Suriñach, J. (2006), 'Quantification of expectations. Are they useful for forecasting inflation?', *Economic Issues* **11**(2), 19–38.

Da Silva, A. (1998), 'On the restricted cointegration test as a test of the rational expectations hypothesis', *Applied Economics* **30**(2), 269–278.

Ece, O. (2001), Inflation expectations derived from business tendency survey of the Central Bank, Technical report, Statistics Department of the Central Bank of the Republic of Turkey.

Fluri, R. & Spoerndli, E. (1987), Rationality of consumers: Price expectations-empirical tests using Swiss qualitative survey data, Technical report, paper presented to 18th CIRET Conference.

Gramlich, E. (1983), 'Models of inflation expectations formation: A comparison of household and economist forecasts', *Journal of Money, Credit and Banking* **15**(2), 155–173.

Hansen, L. (1979), The asymptotic distribution of least squares estimators with endogenous regressors and dependent residuals, Technical report, Carnegie-Mellon University.

Hansen, L. & Hodrick, R. (1980), 'Forward exchange rates as optimal predictors of future spot rates: An econometric analysis', *The Journal of Political Economy* **88**(5), 829–853.

Keane, M. & Runkle, D. (1990), 'Testing the rationality of price forecasts: New evidence from panel data', *The American Economic Review* **80**(4).

Knobl, A. (1974), 'Price expectations and actual price behavior in Germany', *International Monetary Staff Papers* **21**, 83–100.

Loffler, G. (1999), 'Refining the Carlson-Parkin method', *Economics Letters* (64), 167–171.

Mankiw, G. Reis, R. & Wolfers, J. (2003), Disagreements about inflation expectations, Technical Report 9796, National Bureau of Economic Research.

Mitchell, J. (2002), 'The use of non-normal distributions in quantifying qualitative survey data on expectations', *Economics Letters* (76), 101–107.

Muth, J. (1961), 'Rational expectations and the theory of price movements', *Econometrica* **29**(3), 315–335.

Nardo, M. (2003), 'The quantification of qualitative survey data', *Journal of Economic Surveys* (17), 645–664.

Pesaran, H. y Schmidt, P. (1997), *Handbook of Applied Econometrics*, Vol. 2, Blackwell Publishing, Oxford.

Pesaran, M, H. (1985), 'Formation of inflation expectations in British manufacturing industries', *The Economic Journal* (380), 948–975.

Pfanzagl, J. (1952), 'Zur methodik des konjunkturtest-verfahrens', *Statistische Vierteljahresschrift* (5), 161–173.

Seitz, H. (1988), 'The estimation of inflation forecasts from business survey data', *Applied Economics* (20), 427–438.

Theil, H. (1952), 'On the time shape of economics microvariables and the Munich business test', *Review of the International Statistical Institute* (20), 105–120.

Theil, H. (1958), *Economic Forecast and Policy*, 1 edn, North-Holland, Amsterdam.

Visco, I. (1984), *Price Expectations in Rising Inflation*, 1 edn, North-Holland, Amsterdam.

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 109 a 129

Intraday-patterns in the Colombian Exchange Market Index and VaR: Evaluation of Different Approaches

Patrones del IGBC y valor en riesgo: evaluación del desempeño de diferentes metodologías para datos intra-día

Julio César Alonso-Cifuentes[a], Manuel Serna-Cortés[b]

Cienfi - Departamento de Economía, Facultad de Ciencias Administrativas y Económicas, Universidad Icesi, Cali, Colombia

Abstract

This paper evaluates the performance of 16 different parametric, non-parametric and one semi-parametric specifications to calculate the Value at Risk (VaR) for the Colombian Exchange Market Index (IGBC). Using high frequency data (10-minute returns), we model the variance of the returns using GARCH and TGARCH models, that take in account the leverage effect, the day-of-the-week effect, and the hour-of-the-day effect. We estimate those models under two assumptions regarding returns' behavior: Normal distribution and t distribution. This exercise is performed using two different ten-minute intraday samples: 2006-2007 and 2008-2009. For the first sample, we found that the best model is a TGARCH(1,1) without day-of the week or hour-of-the-day effects. For the 2008-2009 sample, we found that the model with the correct conditional VaR coverage would be the GARCH(1,1) with the day-of-the-week effect, and the hour-of-the-day effect. Both methods perform better under the t distribution assumption.

***Key words*:** Leverage, Finance, GARCH model, Risk estimation, Stock returns.

Resumen

El documento evalúa el desempeño de 16 métodos paramétricos, uno no paramétrico y uno semiparamétrico, para estimar el VaR (Valor en Riesgo) de un portafolio conformado por el Índice General de la Bolsa de Valores de Colombia (IGBC). El ejercicio se realiza analizando dos muestras de datos intra-día con una periodicidad de 10 minutos para los períodos 2006-2007 y 2008-2009. Los modelos paramétricos evaluados consideran la presencia o no

[a]Professor. E-mail: jcalonso@icesi.edu.co
[b]Research Assistant. E-mail: mserna.cortes@gmail.com

de patrones de comportamiento, tales como: el efecto "Leverage", el efecto día de la semana, el efecto hora y el efecto día-hora. Nuestros resultados muestran que para la primera muestra el mejor modelo es un TGARCH(1,1) sin el efecto día de la semana ni la hora del día y bajo el supuesto de una distribución t. Para la segunda muestra, 2008-2009, el método que presenta el mejor comportamiento corresponde al modelo GARCH(1,1), que tiene en cuenta el efecto del día y la hora. Estos dos modelos presentan una correcta cobertura condicional y menor función de pérdida.

***Palabras clave*:** apalancamiento, estimación de riesgo, finanzas, GARCH, rendimientos financieros,.

1. Introduction

On January 5, 2007 the Colombian Stock Market Index (IGBC, from the Spanish acronym) dropped 3.1% within the first ten minutes after the stock market opened. Such a drop in the stock market had never occurred before, and it would only happen again in the year 2009. At the time of closing that day, the IGBC had bounced back to such an extent that the overall index loss was 2.1% for that day. This means that the IGBC was down 334 points for the first ten minutes of trade, but then it recouped during the course of the day with a cumulative overall loss of 280.8 points at the end of the day. A financial analyst who only keeps track of information about the index at closing will probably come to the conclusion that it was a relatively ordinary day. That, however, was not just an ordinary day for traders. The kind of risk that materialized during the first ten minutes of trade that day would have gone unnoticed if an analyst had focused on a daily time horizon.

In fact, the behavior of the IGBC during the course of a day seems to follow a relatively stable pattern which can be taken into consideration to improve risk measures. This paper is aimed to illustrate how the involvement of previously documented behavior patterns can be used for improving the performance of risk measures.

It is a hard fact that making decisions in financial markets is exposed to different sources of risk. Hence, there is a need to acknowledge the importance of measuring risk and developing techniques that allow to make improved decisions considering the market circumstances. After the unstability episodes and financial crises in the 1980s[1] and 1990s[2], measuring financial risk has become a routine everyday task in the "back office" of financial institutions (see Alonso & Berggrun 2008). It appears, moreover, that the financial crisis in 2008 led risk management to become the center of discussion again. The reliability of methods such as the Value-at-Risk (VaR) approach is part of the discussion where na reun-

[1]For example, the external debt crisis in most Latin American countries in the 1980s or the collapse of the New York Stock Exchange in 1987.

[2]For example, the burst of the financial and real state bubbles in Japan in the 1990s and of the dot com companies in the late 1990s, the Mexican "Tequila" crisis in 1994, the financial crisis in Southeast Asia in 1997, the Russian crisis in 1997, and the Argentinean crisis in 1998.

derstanding of the limitations of this approach to detect risk in the latest financial crisis has raised great interest among academics, regulators, and the mass media.

VaR is a measure (an estimate) of the largest potential loss for a given time horizon and a given significance level under circumstances which are considered "normal" in the market. VaR is, without question, the most popular measure of financial risk among regulators, financial market stakeholders, and academics. Yet, despite its popularity, the recent financial crisis in 2008 made evident the limitations of this risk measurement tool. In fact, at the onset of the crisis, the collective conscience in the financial community seemed to concur that one of the major culprits of the financial crisis was, undoubtedly, excessive reliance on the VaR (Nocera 2009)[3]. Such reliance actually meant that the appropriation of the meaning and interpretation of the VaR disappeared at some point in time along with the bubble. The market players believed that mathematical and statistical models would be sufficient for managing risk and forgot that VaR was only one of the components of the analysis. Although it was a good way to measure risk, it still had limitations as to estimating it and incorporating other kinds of risks such as liquidity and systemic risk.

During the mortgage bubble, the easy earnings derived at a risk that had been transformed into mathematical certainty, which made people overlook the true meaning of the VaR. Agents forgot that the VaR was only intended to describe what occurred 99% of the times. A VaR of USD 25,000, for example, implied that this amount was not only the most one could lose 99% of the times, but also the least one could lose 1% of the times. It was precisely this 1% where analysts had to link other analyses to incorporate the quantification of liquidity risk or scenarios where the economy would go into a recession and portfolio diversification (systemic risk) had little importance. It was maybe losing sight of this 1% of the times what allowed the bubble to go on for such a long time. Financial entities focused on minimum-risk low-yield investments (VaR), but when they lost that 1% of the times, they did so in a disproportional matter.

After the storm associated with the financial crisis, it now seems clear to both academics and financial analysts that risk measures were not responsible for the crisis; what failed was judgment on the part of the individuals who interpreted these numbers (for a documented discussion of this issue, see Nocera (2009)). It is also evident that these measures, especially VaR, should not fall into oblivion, but should, on the contrary, be polished. Although VaR has some major limitations, incorporating these limitations to the analyses allows for more effective risk management 99% of the times-disregarding them would be going to the extreme of absolute risk aversion.

Consequently, the latest financial crisis is not the end of risk measurement. It is, on the contrary, a wake-up call to encourage reaching a deeper understanding of it, particularly of how it is calculated and interpreted. It is precisely at this point where our work is geared tow and illustrating how the incorporation of previously

[3]Nocera (2009) examines the issue of excessive reliance on the VaR as a result of a lack of an understanding of this measure as a supporting tool for analyzing risk.

documented behavior patterns can be used for improving the performance of risk measures.

It is important to acknowledge that the calculation of the VaR typically follows a simple and intuitive concept, but estimating it poses some practical difficulties. This kind of measure is difficult to estimate because it requires knowledge of the future value distribution of the asset or portfolio being reviewed. In most approaches, the distribution function is not directly estimated. A distribution is assumed for which parameters are calculated for the first moment (mean) and second moment around the mean (variance). Therefore, in practice, various approaches to its calculation take into account considerations that go from assuming normal distribution with constant variance or yields to assume other kinds of distribution and to allow the variance to be updated on a period-after-period basis.

Regardless of the kind of approach used for calculating VaR, a daily time horizon is the most common way to estimate it. This customary calculation of the VaR on a daily basis is partly due to the need to report this number to the regulatory agencies. Nevertheless, in recent years the calculation of the VaR for shorter periods of time has become increasingly popular because of two factors. Firstly, there is a need to have information regarding the risk associated with their business[4] on the part of the stakeholders involved in the market on minute-to-minute basis. On the other hand, there is an increasing availability of intraday information and a widespread use of statistical methods and computer capabilities for processing this information.

The purpose of this work is to evaluate the performance of various approaches to estimate VaR for the following ten minutes. As far as the authors are aware, these kinds of exercises for the Colombian case have not been published in the past. To accomplish this objective, VaR is calculated using different approaches, including parametric, non-parametric, and semi-parametric approaches, and a portfolio that replicates the Colombian Stock Market General Index (IGBC, from its Spanish acronym). To this end, it is essential to acknowledge that the calculus of the VaR entails predicting the performance of the conditional distribution of the portfolio for the following period. A conditional distribution can be different from one day of the week to another (see Alonso & Romero 2009) and even within the same day (see Alonso & García 2009). Therefore, we use VaR models that capture both the weekday effect and the hour-effect as well as the most commonly discussed (Alonso & Arcos 2006) stylized facts such as, the volatility clustering and the heavy tails of the distribution of returns.

This paper is organized as follows: The first section provides a brief introduction and the second provides a brief discussion of the calculations and the evaluation of the Value at Risk in this exercise. The third section addresses the kind of data to be used and the necessary considerations for estimating models using intraday data. The fourth section summarizes the results obtained, and finally, the last section presents the final comments.

[4]See Giot (2000) for an extensive discussion of the reasons for the usefulness of calculating VaR for short time periods.

2. Calculation and Evaluation of Value at Risk

As mentioned above, the concept behind Value at Risk (VaR) is very straightforward and intuitive; these characteristics make this technique very popular. Nevertheless, despite its conceptual simplicity, its calculation poses a relatively sophisticated statistical problem. VaR is intuitively defined as the maximum loss expected from a portfolio with a certain confidence level in a given period of time (see, for example, Alonso & Berggrun 2008).

Formally speaking, the VaR for the following trading period $t+1$ given the information available in the current period t ($VaR_{t+1|t}$) is defined as:

$$P\left(Z_{t+1} < VaR_{t+1|t}\right) = \alpha \tag{1}$$

whereas Z_{t+1} stands for future yield (in Colombian pesos) of the portfolio value for the following period and $(1-\alpha)$ is the level of confidence of the VaR. Therefore, the calculation of the VaR depends on the assumptions regarding the function of distribution of potential losses or gains (absolute yield) from the portfolio Z_{t+1}.

It can be easily proved that if Z_{t+1} follows a distribution with its first two finite moments (such as a normal or a t-distribution), then the value at risk will be as follows:

$$VaR_{t+1|t} = F(\alpha) \cdot \sigma_{t+1} \tag{2}$$

whereas σ_{t+1} stands for the standard deviation of the distribution of Z_{t+1}, and $F(\alpha)$ represents the α percentile of the corresponding (standardized) distribution. Thus, the calculation of the VaR critically depends on two assumptions regarding the behavior of the distribution of Z_{t+1}: (i) its volatility (standard deviation σ) and (ii) its distribution $F(\cdot)$.

As noted earlier, there are several methodological approaches to estimate VaR. These approaches can be classified in three large groups: (i) historical simulation or the non-parametric approach, which does not assume a distribution and does not require estimating parameters; (ii) a parametric approach which involves assuming a distribution and estimating a set of parameters; and (iii) a semi-parametric approach which involves techniques that combine estimating parameters and using the non-parametric approach, such as, for example, filtered historical simulation.

In general, the results obtained after using the various kind of approaches are different and, in each case, the adequacy of these models must be evaluated on an individual case basis. On the other hand, backtesting the performance of an approach to the calculation of the VaR is not an easy task either. A description of the various types of approaches used here for estimating VaR is provided below, including a discussion of the methods to be used for evaluating the performance of the various approaches.

2.1. Some Approaches for the Estimation of VaR

The most common non-parametric approach is the historical simulation. This kind of approach involves determining the α percentile based on historical data.

In other words, this method assumes that past realizations of yield values from the portfolio represent the best approximation of the portfolio yield distribution for the following period. Therefore, $VaR_{t+1|t}$ will be equal to the α percentile of historical portfolio yield values. This approximation will be named as the specification 1.

On the other hand, any parametric approach involves assuming a given distribution function $F(\cdot)$ and the behavior of the parameter that characterizes it, e.g., σ_{t+1}. A well-documented fact of yields on assets is the presence of clustered variance (volatility clustering) (for a discussion of this stylized fact for the Colombian case, see, for example, Alonso & Arcos (2006)). This means, in other words, that volatility is not constant and, therefore, σ will be a function of time. Taking into account this stylized fact, the VaR of a portfolio can then be estimated using the following expression:

$$VaR_{t+1|t} = F(\alpha) \cdot \sigma_{t+1|t} \tag{3}$$

where $\sigma_{t+1|t}$ stands for the standard deviation for period $t+1$ subject to the information available in period t. Thus, it will be necessary to model the conditional variance in order to obtain a one-step-ahead forecast and to assume a distribution for calculating VaR.

Following Alonso & García (2009), we will use ten different approximations in this exercise to estimate the behavior of the variance[5], and for each parametric approach, a model is estimated assuming a normal distribution and a t-distribution[6].

In our case, we will consider eight parametric specifications of the GARCH model. Specification 2 reflects the GARCH(p,q) model proposed by Bollerslev (1986). We will particularly use the GARCH(1,1) model, which, as suggested by Brooks (2008), is usually sufficient to capture the clustered volatility phenomenon. Hence, specification 2 can be represented as follows:

$$\sigma^2_{t+1} = \alpha_0 + \alpha_1 \sigma^2_t + \alpha_2 z^2_t \tag{4}$$

where z_t stands for the error in the mean equation, and α_0, α_1 and α_2 are nonnegative. Additionally, a necessary and sufficient condition for the variance generating process to be stationary is that $\alpha_1 + \alpha_2 < 1$.

Specification 3, which was proposed by Berument & Kiymaz (2003), among others, incorporates dummy variables in the GARCH(1,1) model, capturing the effect of each day on the volatility of returns[7]. In this case, the variance has the following behavior:

$$\sigma^2_{t+1} = \alpha_0 + \alpha_1 \sigma^2_t + \alpha_2 z^2_t + \sum_{i=1}^{4} \beta_i D_{it} \tag{5}$$

[5]The mean is modeled using an autoregressive moving average (ARMA) process, particularly an ARMA(1,1) model that was selected based on the Akaike, Schwarz and Hannan-Quinn information criteria.

[6]A Student's t distribution is assumed in order to take into account a possible heavy tail behavior in the yield distribution, which is another stylized fact of yields from a portfolio or an asset (Alonso & Arcos 2006).

[7]These dummy variables are also incorporated into the mean equation.

where D_{1t} equals one, if day t is a Monday or, otherwise, zero. $D_{2t} = 1$ if t is a Tuesday or zero otherwise, and so forth, respectively. Summarizing, D_{it} are the dummy variables for the first four days of the week.

The day of the week effect (DOW) has been documented in several countries. The findings of recent studies such as Mittal & Jain (2009), for the case of India, and, Kamath & Chinpiao (2010) for the case of Turkey, have shown that this effect exists in emerging markets. For the Colombian case, Alonso & Romero (2009) and Rivera (2009) have shown that this effect is present in the volatility of the IGBC and the Colombian peso-US dollar exchange rate.

Following Giot (2000), specification 4 considers the hour of the day effect (HOD) in our GARCH(1,1) model for variance. There are a fairly good number of studies available in financial literature documenting a U-shaped behavior of volatility and returns within a day. Panas (2005) presented an extensive bibliographic review that documents the presence of this effect in various financial markets worldwide. This specification is aimed at capturing intraday behavior by means of dummy variables. Taking into account that the trading hours at the Colombian Stock Exchange run from 9:00 am to 1:00 pm in the periods being reviewed, this means that there are, in total, four hours for trading. Hence, the hour of the day effect is incorporated into the model using three dummy variables for time,

$$\sigma_{t+1}^2 = \alpha_0 + \alpha_1\sigma_t^2 + \alpha_2 z_t^2 + \sum_{i=1}^{3} \beta_i H_{it} \tag{6}$$

where H_{it} takes the value of one, if t equals trading hour i, or zero otherwise.

The fifth specification to be considered incorporates both the day of the week effect and the hour of the day effect into the GARCH(1,1) model. Dummy variables are included, which take the value of one taking into account time and day. All together, $(4\times5)-1 = 19$ dichotomous variables are used. In this case, the variance is modeled as follows:

$$\sigma_{t+1}^2 = \alpha_0 + \alpha_1\sigma_t^2 + \alpha_2 z_t^2 + \sum_{i=1}^{5}\sum_{j=1}^{4} \varphi_{ij} D_{it} H_{jt} - \varphi_{54} D_{5t} H_{4t} \tag{7}$$

The GARCH(1,1) specifications considered so far do not capture one of the common stylized facts observed for yields: The leverage effect. The leverage effect reflects the trend in volatility of having a greater increase when the price drops than when the price rises to the same extent. The TGARCH model (GARCH model with a threshold)[8] is customarily used in financial literature for capturing the asymmetric behavior of volatility. Thus, specification 6 corresponds to a TGARCH model. This model matches the model proposed by Glosten, Jagannathan & Runkle (1993), which can be expressed as follows:

$$\sigma_{t+1}^2 = \alpha_0 + \alpha_1\sigma_t^2 + \alpha_2 z_t^2 + \alpha_3 d_t z_t^2 \tag{8}$$

where $d_t = 1$ if $z_t < 0$ and $d_t = 0$ if $z_t > 0$. It can also be expected that if the leverage effect is present, then $\alpha_3 > 0$. The condition for obtaining non-negative

[8]This model is also known as the GARCH GJR model.

variances will continue to be that α_0, α_1 and α_2 must be non-negative values. Another condition is that $\alpha_1 + \alpha_3 \geqslant 0$. On the other hand, the model continues to be admissible if $\alpha_3 < 0$, provided that $\alpha_1 + \alpha_3 \geqslant 0$. A necessary and sufficient condition for the variance generating process to be stationary is $\alpha_1 + \alpha_2 < 1$.

Specifications 7, 8, and 9 incorporate the day of the week effect, the hour of the day effect, and the day of the week and hour of the day effect, respectively, into the TGARCH model.

Table 1: Summary of specifications of the models used in this exercise.

Specification	Acronym	Model
1	HS	Historical simulation
2	GARCH	$\sigma^2_{t+1} = \alpha_0 + \alpha_1\sigma^2_t + \alpha_2 z^2_t$
3	GARCH + DOW	$\sigma^2_{t+1} = \alpha_0 + \alpha_1\sigma^2_t + \alpha_2 z^2_t + \sum_{i=1}^{4} \beta_i D_{it}$
4	GARCH + HOD	$\sigma^2_{t+1} = \alpha_0 + \alpha_1\sigma^2_t + \alpha_2 z^2_t + \sum_{i=1}^{3} \beta_i H_{it}$
5	GARCH + DOW + HOD	$\sigma^2_{t+1} = \alpha_0 + \alpha_1\sigma^2_t + \alpha_2 z^2_t + \sum_{i=1}^{5}\sum_{j=1}^{4} \varphi_{ij} D_{it}H_{jt} - \varphi_{54} D_{5t}H_{4t}$
6	TGARCH	$\sigma^2_{t+1} = \alpha_0 + \alpha_1\sigma^2_t + \alpha_2 z^2_t + \alpha_3 d_t z^2_t$
7	TGARCH + DOW	$\sigma^2_{t+1} = \alpha_0 + \alpha_1\sigma^2_t + \alpha_2 z^2_t + \alpha_3 d_t z^2_t + \sum_{i=1}^{4} \beta_i D_{it}$
8	TGARCH + HOD	$\sigma^2_{t+1} = \alpha_0 + \alpha_1\sigma^2_t + \alpha_2 z^2_t + \alpha_3 d_t z^2_t + \sum_{i=1}^{3} \beta_i H_{it}$
9	TGARCH + DOW + HOD	$\sigma^2_{t+1} = \alpha_0 + \alpha_1\sigma^2_t + \alpha_2 z^2_t + \alpha_3 d_t z^2_t + \sum_{i=1}^{5}\sum_{j=1}^{4} \varphi_{ij} D_{it}H_{jt} - \varphi_{54} D_{5t}H_{4t}$
10	FHS	Filtered historical simulation

Note: DOW: Day of the week effect, HOD: Hour of the day effect.
d_t is defined as follows: $d_t = 1$ if $z_t < 0$ and $d_t = 0$ if $z_t > 0$.
H_{it} are the dummy variables for the first three hours of trading at the stock exchange.
D_{it} are the dummy variables for the first four days of the week.

Lastly, a semi-parametric approach, an average-filtered historical simulation is considered. This approach makes it possible to filter an autocorrelation of yields. Thus, yields are filtered by using an ARMA (p,q)[9] process. The estimated residual will be used to perform a historical simulation as described above. This approach has an advantage over historical simulation since it provides an empirical function of density that is more "realistic" in capturing autocorrelation (see, for example,

[9] As described above, in this case an ARMA(1,1) model will be used.

Dowd 2005). A summary of the specifications described above is presented in Table 1.

2.2. Approaches to Assess the Estimated VaR Models

The fit of our models is assessed based on two backtesting or calibration tests and one loss-function criterion. The difficulty in assessing a calculation of VaR lies in the fact that the VaR cannot be observed directly. If one wants to conduct an assessment outside of the sample, in practice there is only information about the realization of yield for the following period, but information about the realization of VaR for that period will not be available.

The most commonly used test available in the literature is Kupiec's (1995) proportion of failures. The purpose of this test is to determine whether the observed proportion of losses that exceed the VaR (also known as proportion of failures) is consistent with the theoretical proportion of failures which provided the basis for constructing the VaR. In other words, the model must provide (non-conditional) coverage in constructing the VaR. In particular, according to the null hypothesis that our model has a "good fit", the n number of failures follows a binomial distribution[10]. In general, considering a total number of observations N and a theoretical level of proportion of failures equal to α (significance level), the probability of observing n losses is calculated as follows:

$$P(n|\, N, \alpha) = \left(\begin{array}{c} N \\ n \end{array} \right) \alpha^n (1-\alpha)^{N-n} \tag{9}$$

To test the null hypothesis that the proportion of failures (ρ) is the same as the theoretically expected value (α) ($H_0 : \rho = \alpha$), Kupiec (1995) suggested the following statistic:

$$t_U = \frac{\widehat{\rho} - \alpha}{\sqrt{\widehat{\rho}(1-\hat{\rho})/N}} \tag{10}$$

where $\widehat{\rho}$ is the observed proportion of failures. Kupiec (1995) demonstrated that t_U follows a t distribution with $N-1$ degrees of freedom.

Christoffersen (1998) suggested a test which considers that the calculation of VaR for $t+1$ represents a forecast subject to the information available in period t. Then, VaR provides coverage subject to the information available at t and, therefore, the backtest should take this into consideration.

The idea underlying this test is that if the best model of VaR is being used, then using all information available at the time of predicting VaR, one should not be able to predict whether the VaR value was exceeded or not. This means that the observed number of failures must be random over time. Thus, a risk model will be said to have suitable non-conditional coverage if the probability of failure equals ρ, i.e. $P(\ PL_{t+1} > \ \ VaR^p{}_{t+1}) = \rho$.[11] A risk model will be said to have correct conditional coverage if $P_t(\ PL_{t+1} > VaR^p{}_{t+1}) = \rho$.

[10] A random variable is defined which takes the value of one if a loss is greater than the VaR or zero otherwise.

[11] Where PL_{t+1} stands for the portfolio loss in period $t+1$.

Therefore, having correct non-conditional coverage means that a model has failures with a probability of ρ on average as the days go by. Having correct conditional coverage, on the other hand, means that the model has failures with a probability of ρ every day, given all the information available on the previous day. It must be noted that correct non-conditional coverage is a necessary, yet not sufficient, condition for correct conditional coverage.

Christoffersen's (1998) idea entails separating specific forecasts being tested and then testing each individual forecast separately. The first is the equivalent of examining whether the model generates a correct proportion of failures, i.e., whether it provides correct non-conditional coverage. The latter implies to test that observed failures are statistically independent from each other. This means that failures should not cluster over time. Evidence of such clustering would mean that the model specification is not correct, even if the model meets the non-conditional coverage requirement.

Given that the theoretical probability of failures is α, Christoffersen (1998) suggests a test that can be expressed in terms of a likelihood ratio (LR) test. Under the null hypothesis of correct non-conditional coverage, the test statistic will be as follows:

$$LR_{uc} = -2\ln[(1-\alpha)^{N-n}\alpha^{n}] + 2\ln[(1-\rho)^{N-n}\rho^{n}] \tag{11}$$

This statistic follows an χ^2_1 distribution. Coming back to the independence test, let n_{kl} be the number of days on which status l occurs at t after status k occurred a $t-1$, where the status refer to failures or non-failures. Besides, let π_{kl} be the probability that status l occurs for any t, given that the status at $t-1$ was k. Under the null hypothesis of independence, the test statistic is as follows:

$$\begin{aligned} LR_{ind} = &-2\ln[(1-\widehat{\pi}_2)^{n_{00}+n_{11}}\widehat{\pi}_2^{n_{01}+n_{11}}] + \\ &2\ln[(1-\widehat{\pi}_{01})^{n_{00}}\widehat{\pi}_{01}^{n_{01}}(1-\widehat{\pi}_{11})^{n_{10}}\widehat{\pi}_{11}^{n_{11}}] \end{aligned} \tag{12}$$

This statistic also follows an χ^2_1. Additionally, the estimated probabilities are defined as follows:

$$\widehat{\pi}_{01} = \frac{n_{11}}{n_{00}+n_{01}}, \quad \widehat{\pi}_{11} = \frac{n_{11}}{n_{10}+n_{11}}, \quad \widehat{\pi}_2 = \frac{n_{01}+n_{11}}{n_{00}+n_{10}+n_{01}+n_{11}} \tag{13}$$

Overall, under the combined hypothesis of correct coverage and independence –i.e., the hypothesis of correct conditional coverage– the test statistic is as follows:

$$LR_{cc} = LR_{uc} + LR_{ind} \tag{14}$$

which follows an χ^2_2 distribution. Thus, Christoffersen's (1998) test allows testing the hypothesis of coverage and independence concurrently. It also tests these hypotheses separately, making possible to identify where the model is failing.

Meanwhile, López (1998) proposes a different approach to evaluate the behavior of VaR using a utility function for selecting the best model based on a set of models that meet the correct conditional coverage requirement. López's (1998)

loss function considers the number of failures and the magnitude of each failure in the following manner:

$$\Psi_{t+1} = \begin{cases} 1 + (PL_{t+1} - VaR_{t+1|t})^2 & \text{if} \quad PL_{t+1} < VaR_{t+1|t} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where Ψ_{t+1} represents the loss function.

Thus, by penalizing the method with the largest failures, the intent is to find a model that minimizes:

$$\Psi = \sum_{t=1}^{N} \Psi_{t+1} \tag{16}$$

3. Description of Our Exercise

In order to evaluate the behavior of the ten approaches[12] above with $\alpha = 0.05$ a recursive window is used. The evaluation exercise involves the following steps:

1. Calculate the VaR for period $T + 1$ (next 10 minutes) using the first T observations;
2. Save the estimated $VaR_{T+1|T}$ and compare it against the observed loss or gain;
3. Update the sample by incorporating an additional observation;
4. Repeat 1,000 times steps 1 to 3 using the last 1,000 observations; and
5. Perform the tests described above.

Below a description is provided of the data used and some special considerations for using intraday data.

3.1. Data

We used ten-minute observations of the returns from the IGBC (General Colombian Stock Exchange Index). In order to achieve our objective of determining the behavior of the various VaR's specifications for a very short time horizon. This exercise was carried out with two samples that represent different environments in the international and macroeconomic markets.

The two samples were used to perform different comparisons of the effectiveness of the VaR's specifications in a relatively steady scenario (2006-2007) versus a scenario of increased uncertainty and volatility (2008-2009). The first sample began at 9:00 am on December 27, 2006, and ended at 1:00 pm on November 9, 2007, with a total of 5,088 observations. The second sample (2008-2009), which

[12]The exercise is carried out under the assumptions of either a normal distribution or a t-distribution.

corresponds to the financial crisis period, started at 9:00 am on June 3, 2008, and ended at 1:00 pm on March 17, 2009, with a total of 4,655 observations.

The IGBC series for the first period was obtained from the Bloomberg information system, while the series for the second period of analysis was obtained from Reuters[13] financial information platform. Figures 1 and 2 show the IGBC series, both for 2006-2007 and 2008-2009, including the returns, the corresponding histograms, and probability charts for the normal, t with 3 and 4 degrees of freedom theoretical distributions.

Based on the charts it is possible to infer that the distribution of yields has relatively heavy tails in comparison to the normal distribution. The descriptive statistics for both samples are reported in Table 2. Jarque-Bera's normality test leads to the conclusion that there is no evidence in favor of the normal distribution of yields for either of the samples. This result is consistent with the stylized facts of yields as discussed by Alonso & Arcos (2006) for this same series.

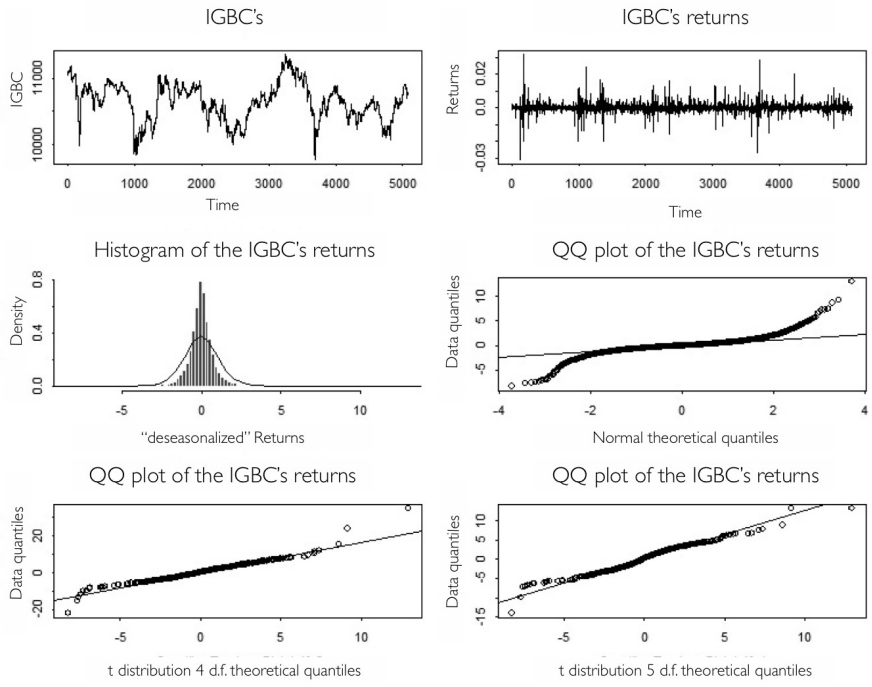

FIGURE 1: IGBC series and returns 2006-2007.

This means that the probability of obtaining extreme values is much greater in the empirical distribution of yields than the expected from a normal distribution. Consequently, in addition to the parametric estimation of VaR under an

[13]The use of different sources of information does not pose any issues to this exercise. Both sources obtain information from the registry system at the Colombian Stock Exchange, so data reported from both sources are identical. There was a change in the source of information because one of the information service providers charged a more convenient fee.

assumption of normality, the parametric VaR estimation was carried out using a Student's t distribution, which relatively adjusts better to the reality of data used[14], as shown by the qq-plots of the t distribution with 3 and 4 degrees of freedom for the 2006-2007 sample and 4 and 5 degrees of freedom for the 2008-2009 sample. As can be seen, the theoretical percentiles from the aforementioned distribution not only adjust more closely to those observed, but also incorporate the stylized fact of heavy tails.

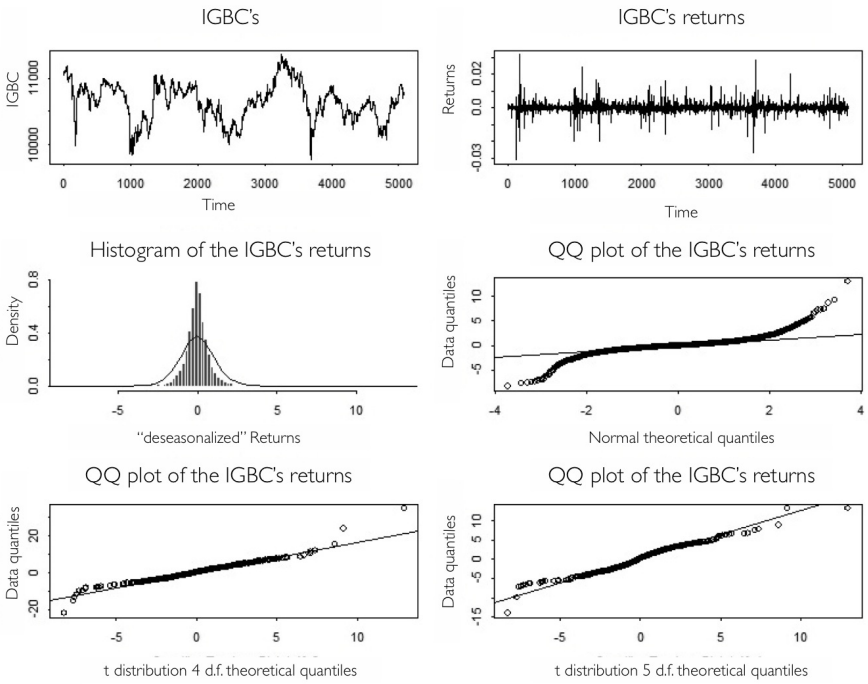

FIGURE 2: IGBC series and returns 2008-2009.

Table 2 shows the apparent symmetry that can be observed in each of the histograms. The kurtosis of both samples is relatively high, especially for the sample from the period that coincides with the financial crisis. On the other hand, the variance of the sample from the financial crisis period is 2.23 greater than that of the other sample. These two results confirm that the financial crisis period is more volatile. These characteristics of the samples, but particularly of the second sample, represent a challenge to the modeling of variance based on GARCH models.

[14]The degrees of freedom were estimated for each iteration based on the conditional variance of the returns assumed under GARCH models.

Table 2: Descriptive statistics of returns for every 10 minutes of the IGBC for both samples.

Sample	2006-2007	2008-2009
Mean	−4.60E-006	−0.5083E-004
Variance	4.89E-06	1.092E-05
Asymmetry coefficient	0.22	0.30
Kurtosis	44.56	196.23
Jarque-Bera	77671.30***	6764.49***

(***) The null hypothesis of normality is rejected with a 99 % level of confidence.

3.2. Special Considerations for Intraday Data

The nature of the data used for this research brings some methodological problems as mentioned by Andersen & Bollerslev (1997) and Giot (2005). By modeling the volatility of returns at high frequencies (i.e., every 5, 10 or 20 minutes), Andersen & Bollerslev (1997) show that the existence of bias is more likely to occur in GARCH and ARCH parameters with high-frequency data when a GARCH model is estimated. Particularly, the probability that the sum of the coefficients equals one increases[15], and thus the probability of estimating non-stationary models for the variance process will also increase. This means that using higher frequency samples involves the risk of capturing the "noise" associated with intraday seasonality and, ultimately, the existence of bias in the estimation of parameters for the GARCH model.

As suggested in literature, there are several alternatives for preventing bias or noise associated with intraday seasonality. Andersen & Bollerslev (1997) propose the use of "deseasonalized" returns (r_t^*). Deseasonalization can be assumed to be deterministic, and when intraday observations are available at regular intervals (e.g., every 10 or 30 minutes), deseasonalized returns can be calculated using the formula:

$$r_t^* = \frac{r_t}{\sqrt{\phi(i_t)}} \tag{17}$$

where r_t stands for observed returns and $\phi(i_t)$ represents the deterministic component of intraday seasonality. To calculate this component, Giot (2005) proposes an average of all square returns that correspond to the same time and day of the week of the observed return r_t. Hence, for 10-minute periods, for each of the five days of the week, the same number of $\phi(i_t)$ is obtained as the number of 10-minute periods in a trading day[16]. Therefore, the specifications will be estimated using the "deseasonalized" series. Later, intraday seasonality will be incorporated in order to calculate the VaR.

[15]In this case, the probability that $\alpha_1 + \alpha_2 = 1$ increases, which implies that the variance process will explode.

[16]In the case of the IGBC index, there are four hours of trading and six 10-minute periods per hour. This means that there are 24 different $\phi(\ i_t)$.

4. Results

Table 3 shows the proportion of failures for each of the samples as well as for the non-parametric approach (specification 1), the semi-parametric approach (specification 10), and GARCH and TGARCH specifications under the assumption of a normal distribution. For the ten approaches, the hypothesis of correct unconditional coverage is rejected if the Kupiec's (1995) test is used. In other words, the forecast proportion of failures using our models is different from the observed proportion of failures.

In fact, it can be observed in Table 3 that the proportions of failure are lower than the 5%, expected proportion of failures. This could be an indication that these specifications are fairly conservative in estimating the VaR. Table 4 on the other hand, reports the same results for GARCH and TGARCH specifications that use a t-distribution. The results are different. In the case of the first sample, the hypothesis of correct unconditional coverage cannot be rejected for specifications 2, 3, 6, and 7. This means that the observed proportion of failures for these specifications is the same as the theoretical proportion used for designing the VaR. For other specifications, the coverage is relatively lower than theoretically expected ($\alpha = 0.05$). For the second sample, specification 8 is the only specification using does not provide correct unconditional coverage.

Thus, if only nunconditional coverage is considered, those specifications using a t-distribution exhibit a better behavior than those where a normal distribution is assumed.

Table 3: Proportion of failures and Kupiec's (1995) test. Normal distribution.

	Sample 2006-2007		Sample 2008-2009	
Spec.	$\widehat{\rho}$	t_U	$\widehat{\rho}$	t_U
1	0.038	−1.985**	0.016	−8.569**
2	0.030	−3.708**	0.028	−4.217**
3	0.027	−4.487**	0.032	−3.234**
4	0.033	−3.009**	0.032	−2.234**
5	0.031	−3.467**	0.035	−2.581**
6	0.030	−3.708**	0.028	−4.217**
7	0.027	−4.487**	0.03	−3.708**
8	0.033	−3.009**	0.033	−3.009**
9	0.031	−3.467**	0.032	−3.234**
10	0.035	−2.367**	0.020	−7.234**

t_U = Kupiec's t-statistic.

(**) Rejects the null hypothesis of non-conditional coverage ($\rho = 0.05$) with a 5% significance level.

Let us now consider Lopez's magnitude loss function (see Table 5). Under the assumption of normality for parametric approaches, it is found that, for the first sample (2006-2007), the third specification is the one that minimizes the loss

Table 4: Proportion of failures and Kupiec's (1995) test. t-distribution.

Spec.	Sample 2006-2007		Sample 2008-2009	
t-distribution	$\widehat{\rho}$	t_U	$\widehat{\rho}$	t_U
2	0.04	−1.614	0.041	−1.435
3	0.042	−1.261	0.041	−1.435
4	0.038	−1.985**	0.042	−1.261
5	0.037	−2.178**	0.043	−1.091
6	0.039	−1.797	0.041	−1.435
7	0.040	−1.435	0.042	−1.261
8	0.037	−2.178**	0.033	−3.009**
9	0.037	−2.178**	0.044	−0.925

t_U = Kupiec's t-statistic.

(**) Rejects the null hypothesis of non-conditional coverage ($\rho = 0.05$) with a 5% significance level.

Table 5: Results of Lopez's (1998) loss function. Normal distribution.

	Sample 2006-2007	Sample 2008-2009
Spec.	Ψ	Ψ
1	2333046836358.57	2647046846407.870
2	1653110917303.23	451650358867.146
3	1578224677463.77*	453951659137.534
4	1684453701968.83	431233167551.233
5	1603126383687.97	435309131396.309
6	1657300659233.63	453207153145.048
7	1581276624589.99	452283235911.192
8	1689813156174.26	430361362409.036*
9	1613824157927.02	433582401277.377
10	2033048131561.07	2036136846407.654

(*) Lowest loss from Lopez's magnitude loss function.

The units of measure for this test are square Colombian pesos.

The initial portfolio value for each period equals 100 million Colombian pesos.

function. On the other hand, for the 2008-2009 sample, specification 8 is the one that exhibits the best behavior with regard to this criterion. None of these specifications, however, provides correct unconditional coverage.

In the case of parametric specifications where a t-distribution is assumed, we find that specification 6 is the one that minimizes Lopez's loss for the first sample, and specification 5 does the same for the second sample. Both specifications provide correct coverage for their corresponding samples.

This would mean that, based on these two criteria, the VaR calculated from a TGARCH model, without considering week day effect or day time effect and a t-distribution, and a GARCH model considering week day and day time effects

Table 6: Results of Lopez's (1998) loss function. t-distribution.

Spec.	Sample 2006-2007	Sample 2008-2009
t-distribution	Ψ	Ψ
2	1850889691317.980	512915110922.186
3	1872451233890.160	504166038902.471
4	1858706617825.380	511117944321.063
5	1865178096796.600	502818217742.356*
6	1846764259665.510*	541384048077.450
7	1868859264549.290	550978138480.490
8	1852149851803.150	3540151295984.520
9	1858768144055.600	529705197041.650

(*) Lowest loss from Lopez's magnitude loss function.

The units of measure for this test are square Colombian pesos. The initial portfolio value for each period equals 100 million Colombian pesos.

and a t-distribution would be the best specifications for estimating VaR for the first and second samples, respectively.

Finally, Table 7 shows the results for Christoffersen's (1998) correct conditional coverage test for the estimated models under the assumption of normality. It can be observed that, for the 2006-2007 sample, there are four specifications that stand out, namely, 1, 4, 8, and 9, because there is not sufficient evidence to reject the null hypothesis of correct conditional coverage. Thus, for this sample, there are a GARCH model considering the day time effect (specification 4), a GARCH model with leverage and day time effects (specification 8), a historical simulation (specification 1), and a filtered historical simulation (specification 10). None of these approaches, however, provides the lowest Lopez's loss function for that sample. The results differ from those of the 2008-2009 sample. In fact, based on Christoffersen's (1998) test, none of the specifications provides correct conditional coverage.

If we consider the parametric models estimated under the assumption of a t-distribution (see Table 8), we find that, for the first sample, all models with the exception of models 5, 8, and 9, the hypothesis of correct conditional coverage cannot be rejected. Out of these specifications, specification 6 (TGARCH without considering the day and time effect) is the one that exhibits the lowest loss function. For the second sample, model 8 is the only one that rejects the hypothesis of correct conditional coverage. And in this case the specification 5 (GARCH model considering day and time effect) provides both correct conditional coverage and the lowest Lopez's loss function.

López's (1998) loss function allows to make a comparison of models that have correct conditional coverage estimated both under the assumption of normal distribution and the assumption of a t-distribution for each of the samples. For the 2006-2007 sample, specification 6 (TGARCH model without considering day and time effect), which was estimated under the assumption of a t-distribution, mini-

mizes Lopez's loss function and, therefore, has a better behavior than a historical simulation or a filtered historical simulation. For the second sample, the best model is model 5, which represents a GARCH(1,1) model considering week day and day time effect, estimated under the assumption of a t-distribution[17].

Table 7: Christoffersen's (1998) coverage and independence test. Normal distribution.

	Sample 2006-2007			Sample 2008-2009		
Spec.	LR_{uc}	LR_{ind}	LR_{cc}	LR_{uc}	LR_{ind}	LR_{cc}
1	3.29	2.43	0.864	32.74**	0.01	32.74^{++}
2	9.77**	−0.51	9.26^{++}	12.04**	0.02	12.06^{++}
3	13.28**	0.02	13.29^{++}	7.78**	0.03	7.81^{++}
4	6.88**	−1.28	5.59	7.78**	0.035	7.81^{++}
5	8.74**	−0.77	7.96^{++}	5.27**	0.04	5.31
6	9.77**	−0.51	9.26^{++}	12.04**	0.02	12.06^{++}
7	13.28**	0.0206	13.299^{++}	9.769**	0.0285	9.797^{++}
8	6.88**	−1.2873	5.591	6.878 ∗ ∗	0.0381	6.916^{++}
9	8.74**	−0.7769	7.962^{++}	7.777**	0.0347	7.811^{++}
10	4.93	2.93	1.99	35.84**	0.05	35.79^{++}

(**) Rejects the null hypothesis of non-conditional coverage at a 5% significance level.
(++) Rejects the null hypothesis of correct conditional coverage at a 5% significance level.

Table 8: Christoffersen's (1998) coverage and independence test. t-distribution.

	Sample 2006-2007			Sample 2008-2009		
Spec.	LR_{uc}	LR_{ind}	LR_{cc}	LR_{uc}	LR_{ind}	LR_{cc}
2	2.25	−2.84	−0.59	1.812	0.07	1.89
3	1.42	−3.23	−1.81	1.812	0.07	1.89
4	3.29	$3.86^{\circ\circ}$	7.15^{++}	1.42	0.079	1.50
5	3.89**	−0.74	3.15	1.08	0.0858	1.17
6	2.75	−2.639	0.11	1.81	0.074	1.89
7	1.81	−3.039	−1.23	1.42	0.079	1.50
8	3.89**	−0.74	3.15	6.88**	0.67	7.56^{++}
9	3.99**	−0.74	3.15	0.79	0.09	0.88

(**) Rejects the null hypothesis of non-conditional coverage at a 5% significance level.
(∘∘) Rejects the null hypothesis of independence at a 5% significance level.
(++) Rejects the null hypothesis of correct conditional coverage at a 5% significance level.

5. Final Remarks and Conclusions

In order to test our hypothesis that intraday behavior patterns could provide relevant information that could be used for improving risk measures such as the

[17]After completing all of the calculations reported above, an exercise was carried out in order to guarantee the robustness of our results both at the beginning and at the end of the two samples being reviewed. For this purposed, the same exercises were replicated, starting with one month, two months, and three months less of data. The results and conclusions remained unchanged. This exercise was also carried out omitting one month, two months, and three months of data at the end of both samples. The conclusions did not change substantially. For the purpose of saving space, these results are not reported here.

VaR, we evaluated the behavior for the next ten minutes of trading at the Colombian Stock Exchange using 18 different ways to estimate VaR for a portfolio with the same behavior as that of the Colombian Stock Exchange index. We considered a non-parametric approach, eight parametric models under the assumption of a normal distribution, and 8 models under the assumption of a t-distribution and one semi-parametric approach. These methods were applied to two different samples, one for a relatively steady period (2006-2007)[18] and another sample for a scenario of increased uncertainty and volatility (2008-2009)[19].

The parametric specifications include the day of the week effect and the hour of the day effect as well as different ways to forecast volatility for the following ten minutes (for a summary of specifications used, see Table 1).

In all cases, prior to the estimation of the models, data is deseasonalized following Giot's (2005) recommendations. The results obtained can be summarized as described below. Firstly, in the case of parametric VaRs with the assumption of normality, we find that there is no model that provides correct non-conditional coverage for the two samples being considered.

Secondly, for both samples, the estimated VaR models under the assumption of a t-distribution have, overall, a better performance than those under the assumption of a normal distribution. Thirdly, we found that, using Christoffersen's (1998) test and López's (1998) loss function to compare models that have correct conditional coverage, we found that the TGARCH(1,1), model without considering week-day and day-time effect and a t-distribution, is the best model for the 2006-2007 sample[20]. For the second sample, the best model is GARCH(1,1), which considers week-day effect and day-time effect estimated under the assumption of a t-distribution. This result validates our hypothesis that intraday behavior patterns can provide relevant information for improving risk measures such as the VaR.

The normal probability charts, Jarque-Bera's normality test, and conditional coverage tests are useful for inferring that, in general, using the assumption of a t-distribution seems to be a better approach than using a normal distribution assumption, which supports our results.

Lastly, our results suggest that there is a need to study intraday behavior of stock portfolios in more detail and encourage a review of approaches that incorporate the dynamic of each of the assets that comprise the portfolio. In other words, it will be necessary to investigate the effect of modeling the multivariate conditional distribution of all assets involved. In order to be able to achieve this, the conditional matrix of variance and covariance will have to be estimated.

[18]This sample, consisting of 5,088 observations in total, runs from 9:00 am on December 27, 2006, to 1:00 pm on November 9, 2007.

[19]This sample, consisting of 4,655 observations in total, begins at 9:00 am on June 3, 2008 and ends at 1:00 pm on March 17, 2009.

[20]It is worth mentioning that Alonso & García (2009) found that, using the first sample, the best model for forecasting the IGBC average for the next ten minutes is a model that did not consider the day or time effect. In other words, these authors showed that day and time are not important when it comes to forecasting the behavior (average) of the IGBC for the next ten minutes. Our results for this sample allow drawing a similar conclusion with regard to the behavior of the variance.

[Recibido: agosto de 2010 — Aceptado: enero de 2011]

References

Alonso, J. C. & Arcos, M. A. (2006), 'Cuatro hechos estilizados de las series de rendimientos: una ilustración para Colombia', *Estudios Gerenciales* **22**(110).

Alonso, J. C. & Berggrun, L. (2008), *Introducción al Análisis de Riesgo Financiero*, Colección Discernir. Serie Ciencias Administrativas y Económicas, Universidad ICESI, Cali, Colombia.

Alonso, J. C. & García, J. C. (2009), '¿qué tan buenos son los patrones del IGBC para predecir su comportamiento?: una aplicación con datos de alta frecuencia', *Estudios Gerenciales* **25**(112), 1–50.

Alonso, J. C. & Romero, F. (2009), The day-of-the-week effect: The Colombian exchange rate and stock market case, *in* 'Selected Abstracts and Papers. Latin American Research Consortium 2009', pp. 112–120.

Andersen, T. G. & Bollerslev, T. (1997), 'Intraday periodicity and volatility persistence in financial markets', *The Journal of Empirical Finance* **4**(2-3), 115–158.

Berument, H. & Kiymaz, H. (2003), 'The day of the week effect on stock market volatility and volume: International evidence', *Review of Financial Economics* **12**(3), 363–380.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **31**(3), 307–327.

Brooks, C. (2008), *Introductory Econometrics for Finance*, Cambridge University Press, London.

Christoffersen, P. (1998), 'Evaluating interval forecasts', *International Economic Review* **39**(4), 841–862.

Dowd, K. (2005), *Measuring Market Risk*, 2 edn, John Wiley & Sons Ltd, England.

Giot, P. (2000), Intraday value-at-risk, CORE Discussion Papers 2000045, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Giot, P. (2005), 'Market risk models for intraday data', *European Journal of Finance* **11**, 309–324.

Glosten, L., Jagannathan, R. & Runkle, D. E. (1993), 'On the relation between the expected value and the volatility of the nominal excess return on stocks', *Journal of Finance* **48**(5), 1779–1801.

Kamath, R. & Chinpiao, L. (2010), 'An investigation of the day-of-the-week effect on the Istanbul stock exchange of Turkey', *Journal of International Business Research* **9**(1), 15–27.

Kupiec, P. H. (1995), 'Techniques for verifying the accuracy of risk measurement models', *Journal of Derivatives* **3**(2), 73–84.

López, J. A. (1998), 'Methods for evaluating value at risk estimates', *Economic Policy Review* **4**(3).

Mittal, S. K. & Jain, S. (2009), 'Stock market behaviour: evidences from Indian market', *Vision* **13**(3), 19–29.

Nocera, J. (2009), 'Risk mismanagement', *The New York Times* .

Panas, E. (2005), 'Generalized Beta distributions for describing and analysing intraday stock market data: Testing the U-shape pattern', *Applied Economics* **37**(2), 191–199.

Rivera, D. M. (2009), 'Modelación del efecto del día de la semana para los índices accionarios de Colombia mediante un modelo STAR GARCH', *Revista de Economía del Rosario* **12**(1), 1–24.

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 131 a 138

The Multinomial Logistic Model for the Case in which the Response Variable Can Assume One of Three Levels and Related Models

El modelo logístico multinomial para el caso en que la variable de respuesta puede asumir uno de tres niveles y modelos relacionados

Humberto Llinás[1,a], Carlos Carreño[2,b]

[1]Departamento de Matemática y Estadística, División de Ciencias básicas, Universidad del Norte, Barranquilla, Colombia

[2]Facultad de Ingeniería de Sistemas, Corporación Universitaria Americana, Barranquilla, Colombia

Abstract

The aim of this work is to examine multinomial logistic models when the response variable can assume three levels, generalizing a previous work of logistic models with binary response variables. We also describe some related models: The null, complete, and saturated models. For each model, we present and prove some theorems concerning to the estimation of the corresponding parameters with details that we could not find in the current literature.

***Key words*:** Binomial distribution, Logistic model, Multinomial logit.

Resumen

El objetivo de este trabajo es examinar los modelos de regresión logística multinomial cuando la variable de respuesta puede asumir tres niveles, generalizando un trabajo anterior con variables respuesta binarias. También describimos algunos modelos relacionados: los modelos nulo, completo y saturado. Para cada modelo, presentamos y demostramos teoremas relacionados con la estimación de los parámetros correspondientes con detalles que no fueron posibles encontrar en la literatura.

***Palabras clave*:** distribución binomial, logit multinomial, modelo logístico.

[a]Associate professor. E-mail: hllinas@uninorte.edu.co
[b]Lecturer. E-mail: ccarreno@coruniamericana.edu.co

1. Introduction

Llinás (2006) studied logistic models with a dichotomous response variable. A theorem was proved on the existence and uniqueness of maximum likelihood (ML) estimations for the logistic model and also about its calculations. Additionally, based on asymptotic theory for these ML-estimations and the score vector, approximations were found for different deviations $-2\log L$, where L is the likelihood function. Based on these approximations, statistics were obtained for several hypothesis tests, each with an asymptotic chi-squared distribution. The asymptotic theory was developed for the case of independent, non-identically distributed variables; thus, modifications are required to apply this theory to the case of identically distributed variables. In this article, a distinction is always made between grouped data and ungrouped data.

Applications of the multinomial logistic model in various fields of engineering and health sciences have made this technique as a fundamental tool for data analysis and subsequent decision making. For this reason firstly, it is important to clarify the theoretical foundations of these models so that they can be applied to specific situations within the data analysis process, which requires more than the use of a statistical program.

We will present to the reader the theoretical background of this model in an effort to describe the continuity of its construction and the elements that are used to perform different analyses with respect to hypothesis tests, relative risks, odds, odds ratios, etc.

For this reason, and following the methodology proposed by Llinás (2006), this article studies multinomial logistic models only for the case in which the variable of interest can assume one of three levels. We describe related models, such as the null, full, and saturated models. For each model, the estimation theorems for the corresponding parameters are presented, providing details that are not found in the current literature (e.g., Agresti 1990, Hosmer & Lemeshow 2000, Kleinbaum & Klein 2002).

The article is organized into six sections. The first section consists of a introduction motivating this reason. The second section explains the basic Bernoulli model. The third section explains the full model. The fourth section explains the null model. The fifth section studies the saturated model and the basic assumptions, and the sixth section develops the theory corresponding to the multinomial logistic model.

2. The Bernoulli Model

Let us suppose that the variable of interest Y can assume one of three values or levels: 0, 1 or 2. For each $r = 0, 1, 2$, we let $p_r := P(Y = r)$ denote the probability that Y assumes the value r.

With n independent observations of Y, a sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is obtained with data $y_i \in \{0, 1, 2\}$ for $i = 1, \ldots, n$, in which y_i is a possible value of Y_i, which are independent of one another.

In order to construct the likelihood function, we create three independent binary variables with values of 0 and 1 as follows:

$$U_{ri} = \begin{cases} 1, & \text{if } Y_i = r \\ 0, & \text{otherwise} \end{cases}$$

where $r = 0, 1, 2$ and $i = 1, \ldots, n$. Observe that $U_{ri} \sim \mathcal{B}(1, p_{ri})$, where $p_{ri} = P(Y_i = r)$.

In terms of the U_{ri} variables, the sample variables are $Y_i = (U_{0i}, U_{1i}, U_{2i})$, with values of $y_i = (u_{0i}, u_{1i}, u_{2i})$, in which $\sum_{r=0}^{2} u_{ri} = 1$, for a fixed i. A statistical model is obtained in which

$$P(Y_i = y_i) = \prod_{r=0}^{2} p_{ri}^{u_{ri}}, \quad i = 1, \ldots, n$$

Setting $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, we obtain the logarithm of the likelihood function for the $2n$-dimensional parameter $\boldsymbol{p} = (p_{01}, p_{11}, \ldots, p_{0n}, p_{1n})^T$:

$$\mathcal{L}(\boldsymbol{p}) = \sum_{i=1}^{n} [u_{0i} \ln p_{0i} + u_{1i} \ln p_{1i} + (1 - u_{0i} - u_{1i}) \ln(1 - p_{0i} - p_{1i})] \qquad (1)$$

3. The Complete Model

The *complete model* is characterized by the assumption that all p_{ri} (with $r = 0, 1, 2$ and $i = 1, \ldots, n$) are considered parameters.

Theorem 1. *In the complete model, the ML-estimations of p_{ri} are $\widehat{P}_{ri} = U_{ri}$ with values $\widehat{p}_{ri} = u_{ri}$ for $r = 0, 1, 2$ and $i = 1, \ldots, n$. Additionally, $\mathcal{L}_c := \mathcal{L}(y) = 0$.*

Proof. Consider equation (1), in which

$$\mathcal{L}(\boldsymbol{p}) = \sum_{\substack{i \\ u_{0i}=1, u_{1i}=0}} \ln p_{0i} + \sum_{\substack{i \\ u_{0i}=0, u_{1i}=1}} \ln p_{1i} + \sum_{\substack{i \\ u_{0i}=0, u_{1i}=0}} \ln(1 - p_{0i} - p_{1i}).$$

Consider that $\mathcal{L}(\boldsymbol{p}) \stackrel{!}{=} 0$ if and only if $p_{0i} = u_{0i}$ and $p_{1i} = u_{1i}$ for each $i = 1, \ldots, n$. This condition proves the existence of the ML-estimations. If for some i it is true that $p_{ri} \neq u_{ri}$, $r = 0, 1$, then $\mathcal{L}(\boldsymbol{p}) < 0$. This condition demonstrates that the ML-estimations are unique because if $\widetilde{\boldsymbol{p}}$ is a vector that has at least one p_{ri} component that is different from u_{ri}, then $\mathcal{L}(\boldsymbol{p}) < \mathcal{L}_c$ (given that upon replacing $p_{ri} = u_{ri}$ in $\mathcal{L}(\boldsymbol{p})$, $\mathcal{L}_c = 0$). □

4. The Null Model

The *null model* is characterized by the assumption that for each $r = 0, 1, 2$, all the p_{ri} values $(i = 1, \ldots, n)$ are considered equal; that is, there are two parameters, p_0 and p_1. In this case, equation (1) becomes

$$\mathcal{L}(\boldsymbol{p}) \quad = \quad n[\overline{u}_0 \ln p_0 + \overline{u}_1 \ln p_1 + (1 - \overline{u}_0 - \overline{u}_1) \ln(1 - p_0 - p_1)] \tag{2}$$

in which $\overline{u}_r = \sum_{i=1}^{n} \frac{u_{ri}}{n}$.

Theorem 2. *In the null model, the ML-estimation of p_r is $\widehat{P}_r = \overline{U}_r$ with value $\widehat{p}_r = \overline{u}_r$. Additionally, $\mathcal{L}_o := \mathcal{L}(\widehat{p}) < 0$ if and only if $0 < \overline{u}_0 + \overline{u}_1 < 1$.*

Proof. It is clear that from equation (2) that

- If $\overline{u}_0 + \overline{u}_1 = 0$, then $\overline{u}_0 = \overline{u}_1 = 0$. Therefore, $\mathcal{L}(p) = 0$ if and only if $p_r = 0 = \overline{u}_r$.
- If $\overline{u}_0 + \overline{u}_1 = 1$, then $\overline{u}_0 = 0$ or $\overline{u}_1 = 0$. Therefore, for $\overline{u}_0 = 0$, $\mathcal{L}(p) = 0$ if and only if $p_1 = 1 = \overline{u}_1$ and $\overline{u}_1 = 1$, $\mathcal{L}(p) = 0$ if and only if $p_0 = 1 = \overline{u}_0$.
- Now let us assume that $0 < \overline{u}_0 + \overline{u}_1 < 1$. From equation (2) and for a given r, it can be proven that

  $$\frac{\partial \mathcal{L}(\boldsymbol{p})}{\partial p_r} \quad = \quad \frac{\overline{u}_r}{p_r} - \frac{1 - \overline{u}_0 - \overline{u}_1}{1 - p_0 - p_1} \quad = \quad 0$$

  if and only if $\widehat{p}_r = \overline{u}_r$. Given that

  $$\frac{\partial^2 \mathcal{L}(\widehat{\boldsymbol{p}})}{\partial p_r^2} \quad = \quad -\left[\frac{\overline{u}_r}{\widehat{p}_r^2} - \frac{1 - \overline{u}_0 - \overline{u}_1}{(1 - \widehat{p}_0 - \widehat{p}_1)^2}\right] \quad < \quad 0$$

  this solution is unique. Additionally, $\ln \overline{u}_r$ and $\ln(1 - \overline{u}_0 - \overline{u}_1)$ are both negative. Therefore, $\mathcal{L}_o < 0$. □

5. The Saturated Model and Assumptions

The saturated model is characterized by the following assumptions:

Assumption 1. It is assumed that:

1. There are K explanatory variables $X_1, \ldots, X_K$ (some may be numerical and other may be categorical) with values $x_{1i}, \ldots, x_{Ki}$ for $i = 1, \ldots, n$ (which are set or observed by the statistician depending on whether the variables are deterministic or random);
2. Among the n individual vectors $(x_{1i}, \ldots, x_{Ki})$ of the values of the explanatory variables Xs, there are J different possible combinations, defining J populations. Therefore, $J \leq n$. J is often referred to as the number of covariate patterns in the applied literature.

Notation. The notation for each population $j = 1, \ldots, J$ is denoted as follows:

- The number of Y_{ij} observations (or of U_{rij} observations in the r category) in each jth population is n_j, with $n_1 + \cdots + n_J = n$;
- For a fixed $r = 0, 1, 2$; the random variable corresponding to the sum of the n_j observations of U_{rij}, given by $Z_{rj} := \sum_{i=1}^{n_j} U_{rij}$ with value $z_{rj} = \sum_{i=1}^{n_j} u_{rij}$, in which $\sum_{j=1}^{J} z_{rj} = \sum_{i=1}^{n} u_{ri}$.

For simplicity, the jth population $(x_{1j}, \ldots, x_{Kj})$ will be abbreviated with the symbol $\star$.

Assumption 2. For each fixed $r = 0, 1, 2$, each population $j = 1, \ldots, J$ and each observation $i = 1, \ldots, n$ in population j, it is assumed that

- $(U_{rij} \mid \star) \sim \mathcal{B}(1, p_{rj})$
- The $(U_{rij} \mid \star)$ variables are independent of one another.

Below, the $\star$ symbol will be omitted. Assumption 2 implies the following:

1. For each $r = 0, 1, 2$ and each fixed $j = 1, \ldots, J$, all the p_{rij}, $i = 1, \ldots, n$, in each jth population are equal. In other words, the $2J$-dimensional $p = (p_{01}, p_{11}, \ldots, p_{0J}, p_{1J})^T$ vector is the parameter.
2. For each $r = 0, 1, 2$ and each population $j = 1, \ldots, J$:
   - $Z_{rj} \sim \mathcal{B}(n_j, p_{rj})$
   - The Z_{rj} variables are independent among populations.

In the saturated model, the logarithm of the maximum likelihood function will be

$$\mathcal{L}(\boldsymbol{p}) = \sum_{j=1}^{J} [z_{0j} \ln p_{0j} + z_{1j} \ln p_{1j} + (n_j - z_{0j} - z_{1j}) \ln(1 - p_{0j} - p_{1j})] \tag{3}$$

Theorem 3. *In the saturated model, the ML-estimations of p_{rj} are $\widetilde{P}_{rj} = \frac{Z_{rj}}{n_j}$, with the values $\widetilde{p}_{rj} = \frac{z_{rj}}{n_j}$, $j = 1, \ldots, J$. Furthermore,*

$$\mathcal{L}(\widetilde{\boldsymbol{p}}) = \sum_{j=1}^{J} n_j [\widetilde{p}_{0j} \ln \widetilde{p}_{0j} + \widetilde{p}_{1j} \ln \widetilde{p}_{1j} + (1 - \widetilde{p}_{0j} - \widetilde{p}_{1j}) \ln(1 - \widetilde{p}_{0j} - \widetilde{p}_{1j})] \tag{4}$$

It also holds that $\mathcal{L}_s := \mathcal{L}(\widetilde{\boldsymbol{p}}) < 0$ for $0 < \widetilde{p}_{rj} < 1$.

Proof. Let us hold r and j. If $0 < \widetilde{p}_{rj} < 1$, then we have

$$\frac{\partial \mathcal{L}}{\partial p_{rj}} = \frac{z_{rj}}{p_{rj}} - \frac{n_j - z_{0j} - z_{1j}}{1 - p_{0j} - p_{1j}} = 0$$

if and only if $\widetilde{p}_{rj} = \frac{z_{rj}}{n_j}$. Therefore, if $0 < z_{rj} < n_j$, for each r and j, then we have

$$\left.\frac{\partial^2 \mathcal{L}}{\partial p_{rj}^2}\right|_{p_{rj}=\widetilde{p}_{rj}} = -\left[\frac{n_j^2}{z_{rj}} + \frac{n_j^2}{n_j - z_{0j} - z_{1j}}\right] < 0$$

Two extreme cases must be analyzed:

- If $z_{rj} = 0$, then $\frac{\partial \mathcal{L}}{\partial p_{rj}} = -\frac{n_j}{1-p_{0j}-p_{1j}}$ decreases in p_j. In this case, $\mathcal{L}$ decreases in p_{rj}; that is, $\mathcal{L}(\boldsymbol{p})$ is maximized for $p_{rj} = 0$.
- If $z_{rj} = n_j$, then, $\frac{\partial \mathcal{L}}{\partial p_j} = \frac{n_j}{p_j}$ increases in p_{rj}. In this case, $\mathcal{L}$ increases in p_{rj}; that is, $\mathcal{L}(\boldsymbol{p})$ is maximized for $p_{rj} = 1$.

In the saturated model, the value of $\mathcal{L}$ can be obtained by replacing in equation (3), each p_{rj} with $\widetilde{p}_{rj}$, $j = 1, \ldots, J$. Thus, we obtain equation (4). Under the condition that $0 < \widetilde{p}_{rj} < 1$ it can be shown that $\ln \widetilde{p}_{rj}$ y $\ln(1 - \widetilde{p}_{rj})$ are both negative. Therefore, the sum of the right side of equation (4) is also negative. □

6. The Multinomial Logistic Model

6.1. Assumptions

Assumptions 1 and 2 from section 5 are preserved, with the additional assumption that a matrix

$$\boldsymbol{C} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{K1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1J} & \cdots & x_{KJ} \end{pmatrix}$$

has a complete range $Rg(C) = 1 + K \leq J$. To obtain a logistic model, one of the categories of the dependent variable Y, such as 0, is used as a reference. The following additional assumption is also made:

Assumption 3.

$$g_1(\boldsymbol{x}_j) = \ln\left(\frac{p_{1j}}{p_{0j}}\right) = \delta_1 + \beta_{11}x_{j1} + \cdots + \beta_{1K}x_{jK} \quad (5)$$

$$g_2(\boldsymbol{x}_j) = \ln\left(\frac{p_{2j}}{p_{0j}}\right) = \delta_2 + \beta_{21}x_{j1} + \cdots + \beta_{2K}x_{jK} \quad (6)$$

in which $\boldsymbol{x}_j := (1, x_{j1}, \ldots, x_{jK})^T$. Let

$$\boldsymbol{\alpha} = (\delta_1, \beta_{11}, \ldots, \beta_{1K}, \delta_2, \beta_{21} \ldots, \beta_{2K})^T$$

denote the vector of the $2(1+K)$ parameters in the model. Note that the assumption that $Rg(\boldsymbol{C}) = 1 + K$ allows the $\boldsymbol{\alpha}$ parameter to be identified.

For a given observation x_j in population j and for the so-called risk is calculated as follows:

$$p_{rj} = \frac{\exp\{g_r(x_j)\}}{\sum_{s=0}^{2} \exp\{g_s(x_j)\}} \tag{7}$$

for each $r = 0, 1, 2$ and with $g_0(x_j) = 0$. The logarithm of the likelihood function can be written as a function of α, as follows:

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{j=1}^{J} \left[z_1 g_1(x_j) + (n_j - z_{0j} - z_{1j}) g_2(x_j) - n_j \ln \left(\sum_{r=0}^{2} \exp\{g_r(x_j)\} \right) \right] \tag{8}$$

6.2. Relation between the Multinomial Logistic Model and the Saturated Model

The equations of assumption 3 in Section 6.1 can be written in a vector form, where $\boldsymbol{g}_r = \boldsymbol{C}\boldsymbol{\beta}_r$, $r = 1, 2$, in which $\boldsymbol{g}_r$ is a J-dimensional vector with elements $g(x_j)$, $j = 1, 2, \ldots, J$.

Given the above, the following cases can be highlighted:

Case 1. $J = 1 + K$

In this case, $\boldsymbol{C}$ is an invertible matrix. Therefore,

$$\beta_r = \boldsymbol{C}^{-1} g_r, \quad r = 1, 2$$

That is, there is a one-to-one relationship between the parameters of the saturated model and those of the logistic model. In other words, the two models express the same thing.

Particularly, the ML-estimations of the probabilities p_{rj} are equal in both models: $\widehat{p}_{rj} = \widetilde{p}_{rj}$ for each $j = 1, 2, \ldots, 1 + K$.

Case 2. $J > 1 + K$

In this case, $\widehat{\alpha}$ must first be calculated, and based on these values, the p_{rj} values can be calculated. In general, we observe that $\widehat{p}_{rj} \neq \widetilde{p}_{rj}$.

7. Likelihood Equations

The likelihood equations are found by calculating the first derivatives of $\mathcal{L}(\boldsymbol{\alpha})$ with respect to each one of the $2(1 + K)$ unknown parameters, as follows. For every $k = 0, 1, \ldots, K$, we have

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \beta_{1k}} = \sum_{j=1}^{J} \left[z_{1j} x_{jk} - \frac{n_j \, x_{jk} \, e^{\boldsymbol{g}_1(x_j)}}{1 + e^{\boldsymbol{g}_1(x_j)} + e^{\boldsymbol{g}_2(x_j)}} \right] = \sum_{j=1}^{J} x_{jk}(z_{1j} - n_j p_{1j})$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \beta_{2k}} &= \sum_{j=1}^{J}\left[(n_j - z_{0j} - nz_{1j})x_{jk} - \frac{n_j\, x_{jk}\, e^{\boldsymbol{g}_2(x_j)}}{1 + e^{\boldsymbol{g}_1(x_j)} + e^{\boldsymbol{g}_2(x_j)}}\right] \\ &= \sum_{j=1}^{J} x_{jk}[(n_j - z_{0j} - z_{1j}) - n_j p_{2j}] \\ &= \sum_{j=1}^{J} x_{jk}(z_{2j} - n_j p_{2j})\end{aligned}$$

Therefore, for every $k = 0, 1, \ldots, K$ and every $r = 0, 1, 2$, the likelihood equations are given by

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \beta_{rk}} = \sum_{j=1}^{J} x_{jk}(z_{rj} - n_j p_{rj})$$

The estimator of maximum likelihood is obtained by setting these equations equal to zero and solving for the logistic parameters. The solution requires the same type of iterations that were used to obtain the estimations in the binary case, as demonstrated in Llinás (2006).

8. Conclusions

We have studied the multinomial logistic models when the response variable can assume one of three values and also described some related models such as the null, complete, and saturated models. We have presented and proved the theorems 1, 2 and 3, which give us the estimation of the corresponding parameters.

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

References

Agresti, A. (1990), *Categorical Data Analysis*, 2 edn, John Wiley and Sons, Inc., New York.

Hosmer, D. & Lemeshow, S. (2000), *Applied Logistic Regresssion*, 2 edn, John Wiley and Sons, Inc., New York.

Kleinbaum, D. & Klein, M. (2002), *Logistic Regression: A Self-Learning Text*, 2 edn, Springer, New York.

Llinás, H. (2006), 'Precisiones en la teoría de los modelos logísticos', *Revista Colombiana de Estadística* **29**(2), 239–265.

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 139 a 166

Aggregation of Explanatory Factor Levels in a Binomial Logit Model: Generalization to the Multifactorial Unsaturated Case

La agregación de niveles en un factor explicativo del modelo logit binomial: generalización al caso multifactorial no saturado

Ernesto Ponsot-Balaguer[1,a], Surendra Sinha[2,b], Arnaldo Goitía[2,c]

[1]Departamento de Estadística, Facultad de Ciencias Económicas y Sociales (FACES), Universidad de Los Andes (ULA), Mérida, Venezuela

[2]Programa de Doctorado en Estadística, Instituto de Estadística Aplicada y Computación (IEAC), FACES-ULA, Mérida, Venezuela

Abstract

We discuss a situation in which, once a logit model is fitted to the data in a contingency table, some factor levels are grouped. Generally, researchers reapply a logit model on the pooled data, however, this approach leads to the violation of the original distributional assumption, when the probabilities of success of the random variables of aggregation differ. In this paper we suggest an alternative procedure that operates under the unsaturated, multifactorial, binomial, logit model. Based on asymptotic theory and taking advantage of the decrease in the variance when the correct distributional assumption is made, the suggested procedure significantly improves the estimates, reduces the standard error, produces lower residuals and is less likely to reject the goodness of fit test on the model. We present the necessary theory, the results of an extensive simulation designed for this purpose, and the suggested procedure contrasted with the usual approach, through a complete numerical example.

***Key words*:** Contingency tables, Generalized linear model, Levels sets, Logit model.

Resumen

Se discute la situación en la que, una vez ajustado un modelo logit a los datos contenidos en una tabla de contingencia, se selecciona un factor cualquiera de los participantes y se agregan algunos de sus niveles. Generalmente los investigadores proceden a postular nuevamente un modelo logit

[a]Associate Professor. E-mail: ernesto@ula.ve
[b]Professor. E-mail: sinha32@yahoo.com
[c]Professor. E-mail: goitia@ula.ve

sobre los datos agrupados, sin embargo, este proceder conduce a la violación del supuesto distribucional original, cuando las probabilidades de éxito de las variables aleatorias de la agregación, son disímiles. En este trabajo se sugiere un procedimiento alternativo que opera en el marco del modelo logit binomial no saturado, multifactorial. Con base en la teoría asintótica y aprovechando la disminución en la varianza cuando se postula el modelo distribucional correcto, el procedimiento sugerido mejora apreciablemente las estimaciones, reduce el error estándar, produce valores residuales más cercanos al cero y menores probabilidades de rechazo en la prueba de bondad del ajuste del modelo. Sustentan tales afirmaciones tanto los desarrollos teóricos necesarios, como los resultados de una extensa simulación diseñada al efecto. También se expone el procedimiento sugerido contrastado con el habitual, mediante un ejemplo numérico completo.

***Palabras clave*:** conjuntos de niveles, modelo lineal generalizado, modelo logit, tablas de contingencia.

1. Introduction

Assume a Bernoulli phenomenon, that is, an experiment whose outcome regarding an individual can only be a success or a failure (or equivalently, the presence or absence of a feature, membership to a particular group or other similar forms). Assume also that a researcher wants to test whether the outcome of the experiment is determined by certain characteristics, measurable in each individual and possibly the direction of the relationship if it exists. For this, the researcher collects data from a previous study or by sampling, for example, and builds a contingency table including the levels of the factors under study, the number of cases in which tests the response of interest (success or failure) and total individuals examined, for each combination of these levels.

A statistical model is related to a contingency table in order to capture the essence of the phenomenon of study in a manageable way and to draw valid conclusions for the population regarding about the causal relationships between the observed response and the measured characteristics.

Now, assuming that the responses are distributed as independent binomials, a model that postulates a certain function of the probability of success of the response and relates linearly with the measured characteristics in individuals looks suitable for analysis. Thus, taking the probit model as a precursor, are the logistic regression for continuous variables and its counterpart, the logit model for categorical explanatory variables or factors introduced by Joseph Berkson in 1944 (Hilbe 2009, p. 3).

In the case of a logit model, the link function considered is $\text{logit}(p) = \log(\frac{p}{1-p})$. Applied to the probability of success p of a Bernoulli random variable, $\text{logit}(p)$ represents the logarithm of the possibility. This, in turn, is defined as the ratio between the probability of success p and its complement, the probability of failure $1 - p$.

Moreover, suppose that the researcher, after fitting a logit model to the data, decides to add some levels of one factor, and repeat the analysis, i.e., fit a new logit model on a contingency table resulting from the aggregation.

It happens that in reiterating a logit model on a second contingency table, with grouped levels of the factors, generally the original binomial assumption is violated, with important implications on the estimated variances (Ponsot, Sinha & Goitía 2009)[1].

In seeking to address this problem and keep the situation under the generalized linear model frame (Nelder & Wedderburn 1972), this paper postulates the problem of aggregation of factor levels in a broad context, i.e., in multifactorial unsaturated logit model situation, and proposes and demonstrates some theorems needed to suggest a procedure, alternative to the usual, that takes advantage of the true variance of the random variables added. It is shown theoretically by asymptotic means, and by simulation, that the suggested procedure is appropriate and in many cases, better than the usual procedure.

This paper continues with the next section presenting a summary of the main background of the work. The third section presents the problem and its solution, including the theorems that support the suggested procedure and their proofs. The fourth section illustrates the suggested procedure with a numerical example. The fifth section summarizes the extensive simulation results comparing the two procedures (normal and suggested). The sixth section is devoted to conclusions, and the work ends with the acknowledgments, references and a brief appendix on the design matrix for the saturated and unsaturated models.

2. Backgrounds

Ponsot et al. (2009) present the problem of aggregation levels of an explanatory factor in the saturated logit model. The authors study the affectation of the binomial distributional assumption and show that, once factor levels are grouped, which involves adding independent binomial random variables (RV's), in the general case where the probabilities of success are different, the random variable (RA) resulting from the aggregation does not follow a binomial distribution. Proper distribution is as follows:

Let X_1 and X_2 be two independents RV's such that $X_1 \sim \text{Bin}(n_1, p_1)$ and $X_2 \sim \text{Bin}(n_2, p_2)$ with $n_1 \leq n_2$. Then, the RV $Z = X_1 + X_2$ is distributed as follows:

$$\text{P}[Z = k] = \left(\frac{p_1}{1 - p_1}\right)^k (1 - p_1)^{n_1} (1 - p_2)^{n_2} S(k) \tag{1}$$

[1]This is central in the doctoral thesis of first author (Ponsot 2011), one of whose results is this paper.

where

$$S(k) = \begin{cases} \sum_{i=0}^{k} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = 0, \ldots, n_1 \\ \sum_{i=k-n_1}^{k} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = n_1+1, \ldots, n_2 \\ \sum_{i=k-n_1}^{n_2} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = n_2+1, \ldots, n_1+n_2 \end{cases}$$

The authors also prove that as the difference between the probabilities of success of the RV's involved in the aggregation increases, the correct variance of the resulting RV [distributed as in (1)], becomes less than the variance calculated assuming that the RV resulting is binomially distributed.

In general, let $X_1, X_2, \ldots, X_a$ be independents RV's such that $X_i \sim \text{Bin}(n_i, p_i)$ for $i = 1, \ldots, a$. Let $X_{a-k+1}, X_{a-k+2}, \ldots, X_a$, the k last RV's being added $(1 < k < a)$ forming the RV $Z = X_{a-k+1} + X_{a-k+2} + \cdots + X_a$. Due to the independence of the originals RV's, $\text{V}[Z]$ is the simple sum of $\text{V}[X_i]$ for $i = a-k+1, \ldots, a$. However, if Z is assumed (incorrectly) binomial, the variance (V_{Bin}) should be calculated differently, making assumptions about the probability of success. By studying the difference $\Delta\text{V} = \text{V}_{\text{Bin}}[Z] - \text{V}[Z]$, it follows that:

$$\Delta\text{V} = \frac{\sum_{i=a-k+1}^{a-1} \sum_{j=i+1}^{a} n_i n_j (p_i - p_j)^2}{\sum_{i=a-k+1}^{a} n_i} \tag{2}$$

Clearly $\Delta\text{V} \geq 0$, then the correct variance is generally smaller than the binomial (equal if and only if $p_i = p_j, \forall i, j$).

Based on these facts and using arguments of asymptotic nature, these authors suggest an alternative procedure to the reiteration of the logit model fitting when factor levels are added. This procedure improves the precision of the estimates, using the true variance of the RV's involved.

Now, as mentioned, the entire development applies in the univariate situation and saturated model exclusively, leaving pending the study of unsaturated logit model in the multifactorial situation. Such an extension is the aim in this work.

Besides, it must be mentioned that there are different courses of action than the asymptotic approach to the problem. For example, we may include (1) as a factor in the likelihood function; however, clearly an analytically intractable expression is obtained, and therefore, very difficult to derivate.

Another possible course of action is to postulate the exact distribution for each given data set, from the contingency table. This way to avoid the assumption of

binomial populations, leading to the hypergeometric distribution and combinatorial analysis. This path has been explored successfully in the theory of generalized linear model; however, it is not of very frequent application because it imposes considerable computational challenges.

It should also be mentioned that the aggregation of factor levels and subsequent repetition of a logit setting is of common practice among statisticians. Hosmer & Lemeshow (2000, p. 136) suggested as a strategy to overcome the drawback of responses with very low or no representation in the contingency table. Examples abound in which the researcher adds factor levels, simply to reduce the complexity of the analysis or because wish to concentrate *posteriori* on some levels and try the other anonymously. An exercise that illustrates this approach can be seen in Hilbe (2009, pp. 74 y 88). In his text the author develops models from the Canada's National Cardiovascular Registry, using a first opportunity to age with four levels as an explanatory factor, and another time, this factor grouping up to only two levels. Another example of the latter type is shown in Menard (2010). In his text the author uses data from the National Center for Opinion Research (University of Chicago, USA), taken from the General Social Survey. In some instances, operates with three or even more levels for the factor "race" (Caucasian, African descent and others), while in alternative examples, it does so with only two levels (not Caucasian and other), grouping the original levels.

3. The Problem and Its Solution

Let T a contingency table for a binary response with s crossed factors $A_1, A_2, \ldots, A_s$, each with $t_1, t_2, \ldots, t_s$ levels, respectively. Each combination of factor levels has an observed response ($y_{i_1 i_2 \cdots i_s}$) as the number of successes, all assumed independently binomially distributed with a total number of observations ($n_{i_1 i_2 \cdots i_s}$), $i_j = 1, \ldots, t_j$ and $j = 1, \ldots, s$. On T, an unsaturated logit model is fitted with the reference parameterization [see for example Rodríguez (2008, cap. 2, p. 29) or SAS Institute Inc. (2004, p. 2297)], then let:

$$\eta_{i_1 i_2 \cdots i_s} = \text{logit}(p_{i_1 i_2 \cdots i_s}) = \boldsymbol{x}^T_{i_1 i_2 \cdots i_s} \boldsymbol{\beta}, \quad \begin{array}{l} i_j = 1, \ldots, t_j; \\ j = 1, \ldots, s \end{array} \tag{3}$$

be the univariate version of the logit model for crossed factors $A_1, \ldots, A_s$. To simplify the treatment of the subscripts of the model, assume that each combination of factors is reindexed orderly, making it correspond to a single value as:

$$1 \equiv (1, 1, \ldots, 1), \ldots, i \equiv (i_1, i_2, \ldots, i_s), \ldots, k \equiv (t_1, t_2, \ldots, t_s)$$

so as to produce $k = t_1 \times \cdots \times t_s$ sequenced indexes. In turn, the response is reindexed as $y_1, y_2, \ldots, y_k$ and so the totals as $n_1, n_2, \ldots, n_k$. Then (3) is expressed in the usual way as:

$$\eta_i = \text{logit}(p_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad i = 1, \ldots, k \tag{4}$$

In (4) $\boldsymbol{x}_i^T$ is the row vector corresponding to the combination of levels $i_1, i_2, \ldots, i_s$ of the design matrix $\boldsymbol{X}_{k\times m}$ and $\boldsymbol{\beta}_{m\times 1}$ is the vector of parameters to be estimated. Let $\boldsymbol{\eta} = [\eta_1 \ \cdots \ \eta_k]^T$ be the vector that groups the logit elements, then the multivariate version of the binomial logit model can be expressed as $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$.

Suppose that after fitting the model to the data, we decide to group some levels of a factor. In the multifactorial situation, the grouping of levels of a factor occurs in several separate clusters, whose number is directly related to the number of levels of other factors of the model. For example, let $s = 3$, A_1, A_2, A_3 be crossed ordered factors and $t_1 = t_2 = t_3 = 3$ its levels. This factor structure contains the tuples $(1,1,1), (1,1,2), \ldots, (3,3,3)$, resulting in $3 \times 3 \times 3 = 27$ tuples.

Let examine the following situation for illustrative purposes: Levels 2 and 3 of A_3 are grouped. In this situation, the new number of levels of A_3 is $t_3^* = 2$ and factor structure is reduced to $3 \times 3 \times 2 = 18$ tuples. For $i = 1, 2, 3$ and $j = 1, 2, 3$, original tuples $(i, j, 2)$ and $(i, j, 3)$ collapse in the new tuples $(i, j, 2^*)$ by adding the corresponding values of the response variables $y_{ij2} + y_{ij3}$ and the totals $n_{ij2} + n_{ij3}$. It is easy to notice that 9 aggregation sets are required, $c_k, k = 1, 2, \ldots, 9$, each one with two elements or levels $c_1 = \{(1,1,2), (1,1,3)\}, \ldots, c_9 = \{(3,3,2), (3,3,3)\}$.

If the proposed model is saturated ($k = m$), i.e. the number of available observations equals the number of model parameters, the $\boldsymbol{X}$ matrix is a square, full rank, and therefore invertible matrix. Moreover, when assuming an unsaturated logit model, generally $k > m$, the design matrix $\boldsymbol{X}$ is no longer square and it has no inverse.

It has been proved by McCullagh & Nelder (1989, p. 119) that

$$V[\widehat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$$

where $\boldsymbol{W} = \text{diag}[n_i p_i (1-p_i)]$. These authors also discuss that the problems of over or under dispersion, deserve detailed study and that they can be solved by simply scaling $V[\widehat{\boldsymbol{\beta}}]$ by a constant, obtained from the deviance or Pearson's statistics and residual degrees of freedom ratio.

Thus, assuming no over or under dispersion (which simply involves the appropriate scaling of the estimated variance-covariance matrix), an immediate consequence of the fact that $\boldsymbol{X}$ has no inverse is that, once parameters have been estimated by iterative reweighted least squares (Searle, Casella & McCulloch 2006, p. 295), $V[\boldsymbol{X}\widehat{\boldsymbol{\beta}}] = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T$, do not support further simplification.

Let be $\boldsymbol{\Sigma} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T$, with elements $[\sigma_{ij}]$, $i, j = 1, \ldots, k$. In general, though not necessarily, $\sigma_{ij} \neq 0$. Then, due to the central limit theorem (Lehmann 1999, p. 73) and asymptotic properties of maximum likelihood estimators:

$$\widehat{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \sim \text{AN}(\boldsymbol{X}\boldsymbol{\beta}; \boldsymbol{\Sigma}_{k\times k}) \tag{5}$$

In (5), "AN" is the abbreviation for "Asymptotically Normal", commonly used in the statistical literature. Moreover, it is necessary the asymptotic distribution of the $\widehat{p}_i$. It is developed in the following theorem:

Theorem 1. *If* $\widehat{\boldsymbol{\eta}} = [logit(\widehat{p}_1)\ logit(\widehat{p}_2)\ \cdots\ logit(\widehat{p}_k)]^T$ *is distributed as in* (5)*, then* $\widehat{\boldsymbol{p}} = [\widehat{p}_1\ \widehat{p}_2\ \cdots\ \widehat{p}_k]^T$*, such that:*

$$\widehat{p}_i = \frac{e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}}}, \quad i=1,\ldots,k$$

is asymptotically distributed as multivariate normal with $E[\widehat{p}_i] = p_i = e^{\boldsymbol{x_i^T\beta}}/(1+e^{\boldsymbol{x_i^T\beta}})$ *and variance covariance matrix* $\boldsymbol{\Psi} = [\psi_{ij}]$ *with elements* $\psi_{ij} = \sigma_{ij}p_i(1-p_i)p_j(1-p_j)$*,* $i,j=1,\ldots,k$*.*

Proof. Let g_i^{-1} for $i=1,\ldots,k$ be real-valued functions defined as

$$g_i^{-1}(\widehat{\eta}_1,\ldots,\widehat{\eta}_i,\ldots,\widehat{\eta}_k) = \frac{e^{\widehat{\eta}_i}}{1+e^{\widehat{\eta}_i}}$$

then,

$$\frac{\partial g_i^{-1}}{\partial\widehat{\eta}_j} = \begin{cases} 0 & ,i\neq j \\ e^{\widehat{\eta}_i}/(1+e^{\widehat{\eta}_i})^2 & ,i=j \end{cases}$$

$$\begin{aligned}\psi_{ij} &= \sum_{s=1}^{k}\sum_{t=1}^{k}\sigma_{st}\frac{\partial g_i^{-1}}{\partial\widehat{\eta}_s}\frac{\partial g_j^{-1}}{\partial\widehat{\eta}_t}\bigg|_{\widehat{\boldsymbol{\eta}}=\boldsymbol{\eta}} = \sum_{s=1}^{k}\sigma_{sj}\frac{\partial g_i^{-1}}{\partial\widehat{\eta}_s}\frac{\partial g_j^{-1}}{\partial\widehat{\eta}_j}\bigg|_{\widehat{\boldsymbol{\eta}}=\boldsymbol{\eta}} \\ &= \sigma_{ij}\frac{\partial g_i^{-1}}{\partial\widehat{\eta}_i}\frac{\partial g_j^{-1}}{\partial\widehat{\eta}_j}\bigg|_{\widehat{\boldsymbol{\eta}}=\boldsymbol{\eta}} = \sigma_{ij}\frac{e^{\eta_i}}{(1+e^{\eta_i})^2}\frac{e^{\eta_j}}{(1+e^{\eta_j})^2} \\ &= \sigma_{ij}p_i(1-p_i)p_j(1-p_j)\end{aligned}$$

Thus, given the existence of the partial derivatives around $\widehat{\boldsymbol{\eta}}$, multivariate version of the delta method (Lehmann 1999, p. 315) ensures that $\widehat{\boldsymbol{p}} = [\widehat{p}_1\ \widehat{p}_2\ \cdots\ \widehat{p}_k]^T$ is asymptotically normal with $\mathrm{E}[\widehat{p}_i] = p_i = e^{\boldsymbol{x_i^T\beta}}/(1+e^{\boldsymbol{x_i^T\beta}})$ and variance covariance matrix $\boldsymbol{\Psi} = [\psi_{ij}]$ with $\psi_{ij} = [\sigma_{ij}p_i(1-p_i)p_j(1-p_j)]$, $i,j=1,\ldots,k$. □

Suppose then that the researcher wants to add r levels $(1<r<t_i)$ of i-th factor A_i and, therefore, $a = t_1\times\cdots\times t_{i-1}\times t_{i+1}\times\cdots\times t_s$ sets are produced, whose elements are each r of the indexes $1,\ldots,k$ without repetition, affected by the aggregation. Let the sets (called "aggregation sets"), be defined by:

$$\begin{aligned} c_\nu = \{\xi_1^i,\xi_2^i,\ldots,\xi_r^i\}, \qquad & \xi_j^i\in\{1,\ldots,k\};\ j=1,\ldots,r; \\ & i=1,\ldots,a;\ \ \nu=\min\{\xi_1^i,\xi_2^i,\ldots,\xi_r^i\}\text{ for each }i; \\ & c_\nu\cap c_{\nu'}=\phi,\forall\,\nu,\nu'\end{aligned}$$

for each of which, in turn is defined:

$$n_\nu^* = \sum_{c_\nu} n_i, \qquad \widehat{p}_\nu^* = \frac{\sum_{c_\nu} n_i\widehat{p}_i}{n_\nu^*} \tag{6}$$

Since $\widehat{p}^*_\nu$ is the weighted sum of asymptotically normal RV's, $\widehat{p}^*_\nu$ is an asymptotically normal RV for all ν and covaries with the other probability estimators. It is easy to verify that $\mathrm{E}[\widehat{p}^*_\nu] = p^*_\nu = (\sum_{c_\nu} n_i p_i)/n^*_\nu$, however, the variance and covariance associated with $\widehat{p}^*_\nu$ are more complex, as is proved in the following theorem:

Theorem 2. *Given* $\widehat{\boldsymbol{p}} = [\widehat{p}_1\ \widehat{p}_2\ \cdots\ \widehat{p}_k]^T$ *distributed as in Theorem 1, if* $\widehat{p}^*_\nu = (\sum_{c_\nu} n_i \widehat{p}_i)/n^*_\nu$ *with* $n^*_\nu = \sum_{c_\nu} n_i$*, then:*

$$V[\widehat{p}^*_\nu] = \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2\psi_{ii} + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j \psi_{ij}\right\}$$

$$Cov[\widehat{p}^*_\nu, \widehat{p}_j] = \frac{\sum_{i\in c_\nu} n_i\psi_{ij}}{n^*_\nu},\ \textit{for all } j \notin \bigcup_{i=1}^{a} c_i$$

$$Cov[\widehat{p}^*_\nu, \widehat{p}^*_{\nu'}] = \frac{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \psi_{ij}}{n^*_\nu n^*_{\nu'}}$$

for any two aggregation sets $c_\nu, c_{\nu'}$*.*

Proof.

$$(\widehat{p}^*_\nu)^2 = \left\{\frac{\sum_{c_\nu} n_i\widehat{p}_i}{n^*_\nu}\right\}^2$$

$$= \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2\widehat{p}_i^2 + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j \widehat{p}_i\widehat{p}_j\right\}$$

$$(\mathrm{E}[\widehat{p}^*_\nu])^2 = \left\{\frac{\sum_{c_\nu} n_i\mathrm{E}[\widehat{p}_i]}{n^*_\nu}\right\}^2$$

$$= \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2(\mathrm{E}[\widehat{p}_i])^2 + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j \mathrm{E}[\widehat{p}_i]\mathrm{E}[\widehat{p}_j]\right\}$$

$$\mathrm{E}[(\widehat{p}^*_\nu)^2] = \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2\mathrm{E}[\widehat{p}_i^2] + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j \mathrm{E}[\widehat{p}_i\widehat{p}_j]\right\} \Rightarrow$$

$$\mathrm{V}[\widehat{p}^*_\nu] = \mathrm{E}[(\widehat{p}^*_\nu)^2] - (\mathrm{E}[\widehat{p}^*_\nu])^2$$

$$= \frac{1}{(n^*_\nu)^2}\left\{\sum_{c_\nu} n_i^2(\mathrm{E}[\widehat{p}_i^2] - \mathrm{E}[\widehat{p}_i]^2)\right.$$

$$\left. +2\sum_{i\in c_\nu - \max\{c_\nu\}}\ \sum_{j\in c_\nu > i} n_i n_j(\mathrm{E}[\widehat{p}_i\widehat{p}_j] - \mathrm{E}[\widehat{p}_i]\mathrm{E}[\widehat{p}_j])\right\}$$

$$= \frac{1}{(n_\nu^*)^2}\left\{\sum_{c_\nu} n_i^2 \mathrm{V}[\widehat{p}_i] + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\sum_{j\in c_\nu > i} n_i n_j \mathrm{Cov}[\widehat{p}_i,\widehat{p}_j]\right\}$$
$$= \frac{1}{(n_\nu^*)^2}\left\{\sum_{c_\nu} n_i^2 \psi_{ii} + 2\sum_{i\in c_\nu - \max\{c_\nu\}}\sum_{j\in c_\nu > i} n_i n_j \psi_{ij}\right\}$$

Furthermore, for $j \notin c_\nu$:

$$\begin{aligned}\mathrm{Cov}[\widehat{p}_\nu^*,\widehat{p}_j] &= \mathrm{E}[\widehat{p}_\nu^*\widehat{p}_j] - \mathrm{E}[\widehat{p}_\nu^*]\mathrm{E}[\widehat{p}_j] \\ &= \frac{\sum_{c_\nu} n_i \mathrm{E}[\widehat{p}_i\widehat{p}_j]}{n_\nu^*} - \frac{\sum_{c_\nu} n_i \mathrm{E}[\widehat{p}_i]\mathrm{E}[\widehat{p}_j]}{n_\nu^*} \\ &= \frac{\sum_{c_\nu} n_i \mathrm{Cov}[\widehat{p}_i,\widehat{p}_j]}{n_\nu^*} = \frac{\sum_{c_\nu} n_i \psi_{ij}}{n_\nu^*}\end{aligned}$$

Finally:

$$\widehat{p}_\nu^*\widehat{p}_{\nu'}^* = \left(\frac{\sum_{c_\nu} n_i\widehat{p}_i}{n_\nu^*}\right)\left(\frac{\sum_{c_{\nu'}} n_i\widehat{p}_i}{n_{\nu'}^*}\right) = \frac{1}{n_\nu^* n_{\nu'}^*}\left\{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \widehat{p}_i\widehat{p}_j\right\} \Rightarrow$$

$$\begin{aligned}\mathrm{Cov}[\widehat{p}_\nu^*,\widehat{p}_{\nu'}^*] &= \mathrm{E}[\widehat{p}_\nu^*\widehat{p}_{\nu'}^*] - \mathrm{E}[\widehat{p}_\nu^*]\mathrm{E}[\widehat{p}_{\nu'}^*] \\ &= \frac{1}{n_\nu^* n_{\nu'}^*}\left\{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \mathrm{E}[\widehat{p}_i\widehat{p}_j]\right\} \\ &\qquad - \frac{1}{n_\nu^* n_{\nu'}^*}\left\{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \mathrm{E}[\widehat{p}_i]\mathrm{E}[\widehat{p}_j]\right\} \\ &= \frac{1}{n_\nu^* n_{\nu'}^*}\left\{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \mathrm{Cov}[\widehat{p}_i,\widehat{p}_j]\right\} \frac{\sum_{i\in c_\nu}\sum_{j\in c_{\nu'}} n_i n_j \psi_{ij}}{n_\nu^* n_{\nu'}^*}\end{aligned}$$

□

Note that the cardinality of the index range of the model (originally k) has been reduced given the aggregation levels and is now $k^* = k - a(r-1)$. Each group of r originals RV's, for each of the a different combinations of the levels of the other factors $(A_1,\ldots,A_{i-1},A_{i+1},\ldots,A_k)$, gives way to a single random variable constructed from the sum, renamed in its index at the lower value of the aggregation set that corresponds. So, having both likelihood estimators not affected by aggregation, as those who effectively are, we can settle the new vector:

$$\widehat{\boldsymbol{p}}_{k^*\times 1}^* \sim \mathrm{AN}(\boldsymbol{p}_{k^*\times 1}^*; \boldsymbol{\Psi}_{k^*\times k^*}^*) \tag{7}$$

where:

$$\widehat{\boldsymbol{p}}^{*} = \begin{bmatrix} \widehat{p}_1^{*} \\ \vdots \\ \widehat{p}_k^{*} \end{bmatrix}, \ \boldsymbol{p}^{*} = \begin{bmatrix} p_1^{*} \\ \vdots \\ p_k^{*} \end{bmatrix}, \ \boldsymbol{\Psi}^{*} = \begin{bmatrix} \psi_{11}^{*} & \psi_{12}^{*} & \cdots & \psi_{1k}^{*} \\ \psi_{21}^{*} & \psi_{22}^{*} & \cdots & \psi_{2k}^{*} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{k1}^{*} & \psi_{k2}^{*} & \cdots & \psi_{kk}^{*} \end{bmatrix}$$

except that the range of the index $1, \ldots, k$ in $\boldsymbol{\Psi}^{*}$, although ordered, is not correlated with $\mathbb{N}$, that is, some of their values are no longer present.

Also, $\widehat{p}_i^{*} \equiv \widehat{p}_i$, $p_i^{*} \equiv p_i$, $\psi_{ij}^{*} \equiv \psi_{ij}$ for all $i, j \notin \cup c_\nu$ and $\widehat{p}_i^{*}$, p_i^{*}, ψ_{ij}^{*} are as in the definition and Theorem 2 for the remaining i, j.

Example 1. Let T be a contingency table with two factors A_1 and A_2, the first with 2 levels $(1, 2)$ and the second with three $(1, 2, 3)$. Reindexing the original subscripts properly, we have:

$$1 \equiv (1,1); 2 \equiv (1,2); 3 \equiv (1,3); 4 \equiv (2,1); 5 \equiv (2,2); 6 \equiv (2,3)$$

with the original logit model estimates $\widehat{\boldsymbol{p}} = [\widehat{p}_1 \quad \widehat{p}_2 \quad \widehat{p}_3 \quad \widehat{p}_4 \quad \widehat{p}_5 \quad \widehat{p}_6]^T$.

Now suppose we add levels 2 and 3 of factor A_2. Aggregation sets that arise are $c_2 = \{2, 3\}$ and $c_5 = \{5, 6\}$, and the new model estimates $\widehat{\boldsymbol{p}}^{*} = [\widehat{p}_1^{*} \quad \widehat{p}_2^{*} \quad \widehat{p}_4^{*} \quad \widehat{p}_5^{*}]^T$, where:

$$\begin{aligned} \widehat{p}_1^{*} &= \widehat{p}_1 \\ \widehat{p}_2^{*} &= \frac{n_2 \widehat{p}_2 + n_3 \widehat{p}_3}{(n_2 + n_3)} \\ \widehat{p}_4^{*} &= \widehat{p}_4 \\ \widehat{p}_5^{*} &= \frac{n_5 \widehat{p}_5 + n_6 \widehat{p}_6}{(n_5 + n_6)} \end{aligned}$$

In addition, the variance covariance matrix of $\widehat{\boldsymbol{p}}$ is

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} & \psi_{14} & \psi_{15} & \psi_{16} \\ \psi_{21} & \psi_{22} & \psi_{23} & \psi_{24} & \psi_{25} & \psi_{26} \\ \psi_{31} & \psi_{32} & \psi_{33} & \psi_{34} & \psi_{35} & \psi_{36} \\ \psi_{41} & \psi_{42} & \psi_{43} & \psi_{44} & \psi_{45} & \psi_{46} \\ \psi_{51} & \psi_{52} & \psi_{53} & \psi_{54} & \psi_{55} & \psi_{56} \\ \psi_{61} & \psi_{62} & \psi_{63} & \psi_{64} & \psi_{65} & \psi_{66} \end{bmatrix}$$

while the variance covariance matrix of $\widehat{\boldsymbol{p}}^{*}$ (symmetric) is

$$\boldsymbol{\Psi}^{*} = \begin{bmatrix} \psi_{11}^{*} & \psi_{12}^{*} & \psi_{14}^{*} & \psi_{15}^{*} \\ \psi_{21}^{*} & \psi_{22}^{*} & \psi_{24}^{*} & \psi_{25}^{*} \\ \psi_{41}^{*} & \psi_{42}^{*} & \psi_{44}^{*} & \psi_{45}^{*} \\ \psi_{51}^{*} & \psi_{52}^{*} & \psi_{54}^{*} & \psi_{55}^{*} \end{bmatrix}$$

where, following the Theorem 2:

$$\begin{aligned}
\psi^*_{11} &= \psi_{11}\\
\psi^*_{12} &= (n_2\psi_{12} + n_3\psi_{13})/(n_2 + n_3)\\
\psi^*_{14} &= \psi_{14}\\
\psi^*_{15} &= (n_5\psi_{15} + n_6\psi_{16})/(n_5 + n_6)\\
\psi^*_{22} &= (n_2^2\psi_{22} + n_3^2\psi_{33} + 2n_2n_3\psi_{23})/(n_2 + n_3)^2\\
\psi^*_{24} &= (n_2\psi_{24} + n_3\psi_{34})/(n_2 + n_3)\\
\psi^*_{25} &= (n_2n_5\psi_{25} + n_2n_6\psi_{26} + n_3n_5\psi_{35} + n_3n_6\psi_{36})/[(n_2 + n_3)(n_5 + n_6)]\\
\psi^*_{44} &= \psi_{44}\\
\psi^*_{45} &= (n_5\psi_{45} + n_6\psi_{46})/(n_5 + n_6)\\
\psi^*_{55} &= (n_5^2\psi_{55} + n_6^2\psi_{66} + 2n_5n_6\psi_{56})/(n_5 + n_6)^2
\end{aligned}$$

Now, returning to the theoretical development, the following theorem shows the required distribution of $\text{logit}(\widehat{p}^*_i), i = 1, \ldots, k$, prior to the estimation of the parameters associated with the factors.

Theorem 3. *If* $\widehat{\boldsymbol{p}}^*$ *is distributed as in* (7)*, then:*

$$\textit{logit}(\widehat{\boldsymbol{p}}^*) = \begin{bmatrix}\textit{logit}(\widehat{p}^*_1) & \cdots & \textit{logit}(\widehat{p}^*_k)\end{bmatrix}^T$$

is asymptotically distributed multivariate normal with $E[\textit{logit}(\widehat{p}^*_i)] = \textit{logit}(p^*_i)$ *and variance covariance matrix* $\boldsymbol{\Sigma}^* = [\sigma^*_{ij}] = [\psi^*_{ij}[p^*_i(1-p^*_i)p^*_j(1-p^*_j)]^{-1}]$.

Proof. Lets $g_i(i = 1, \ldots, k)$, real-valued functions defined as

$$g_i(\widehat{p}^*_1, \ldots, \widehat{p}^*_i, \ldots, \widehat{p}^*_k) = \text{logit}(\widehat{p}^*_i)$$

then,

$$\frac{\partial g_i}{\partial \widehat{p}^*_j} = \begin{cases} 0 & , i \neq j \\ [\widehat{p}^*_i(1-\widehat{p}^*_i)]^{-1}, & i = j \end{cases}$$

and

$$\sigma^*_{ij} = \sum_{s=1}^{k}\sum_{t=1}^{k}\psi^*_{st}\frac{\partial g_i}{\partial \widehat{p}^*_s}\frac{\partial g_j}{\partial \widehat{p}^*_t}\bigg|_{\widehat{\boldsymbol{p}}^* = \boldsymbol{p}^*} = \psi^*_{ij}[p^*_i(1-p^*_i)p^*_j(1-p^*_j)]^{-1}$$

And because in this case also there are the partial derivatives around $\widehat{\boldsymbol{p}}^*$, using again a multivariate version of the delta method, $\textbf{logit}(\widehat{\boldsymbol{p}}^*)$ is asymptotically distributed multivariate normal with $\text{E}[\text{logit}(\widehat{p}^*_i)] = \text{logit}(p^*_i)$ and variance covariance matrix $\boldsymbol{\Sigma}^* = [\sigma^*_{ij}] = [\psi^*_{ij}[p^*_i(1-p^*_i)p^*_j(1-p^*_j)]^{-1}]$. □

Finally, the following theorem shows the distribution of the new parameters $\widehat{\boldsymbol{\beta}}^*$, from the new design matrix $\boldsymbol{X}^*$. Its proof is omitted since it is easily obtained by appealing to the results included in the Appendix.

Theorem 4. *Given the model*

$$\boldsymbol{Y} = logit(\widehat{\boldsymbol{p}}^{*}) = \boldsymbol{X}^{*}\boldsymbol{\beta}^{*} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim AN(\boldsymbol{0}, \boldsymbol{\Sigma}^{*})$$

in which, $\boldsymbol{Y} = logit(\widehat{\boldsymbol{p}}^{*})$ *is a column vector whose elements are* $logit(\widehat{p}_i^{*}), i = 1, \ldots, k$ *and* $\boldsymbol{\Sigma}^{*}$ *is the variance covariance matrix, both constant and known, calculated according to the Theorem 3. Let* $\boldsymbol{X}^{*}$ *be the new design matrix*[2]*, using the reference parameterization, proposed after the process of aggregation of factor levels, constrained to include the same factors that included the original design matrix* $\boldsymbol{X}$*. And let* $\boldsymbol{\beta}^{*}$ *be the new vector of parameters to be estimated by maximum likelihood (*$\widehat{\boldsymbol{\beta}}^{*}$*). Then:*

- $\widehat{\boldsymbol{\beta}}^{*} = [(\boldsymbol{X}^{*})^{T}\boldsymbol{X}^{*}]^{-1}(\boldsymbol{X}^{*})^{T}\boldsymbol{Y}$.
- $V[\widehat{\boldsymbol{\beta}}^{*}] = [(\boldsymbol{X}^{*})^{T}\boldsymbol{X}^{*}]^{-1}(\boldsymbol{X}^{*})^{T}\boldsymbol{\Sigma}^{*}\boldsymbol{X}^{*}[(\boldsymbol{X}^{*})^{T}\boldsymbol{X}^{*}]^{-1}$.
- $\widehat{\boldsymbol{\beta}}^{*}$ *is distributed asymptotically normal.*

A modification in this asymptotical distribution has been induced for the original and transformed RV's, by applying aggregation some factor levels and some required sets of aggregation (c_ν). Thus, the suggested procedure is as follows:

1. Fit a logit model by preserving the calculation of the vector of estimates of p_i and the variance covariance matrix $\boldsymbol{\Sigma}$ estimated for $\widehat{\boldsymbol{\beta}}$.
2. Define the required aggregation sets, in order to calculate point estimates for the p_ν^{*} as in the Theorem 2 and the variance covariance matrix $\boldsymbol{\Psi}^{*}$ of $\widehat{\boldsymbol{p}}^{*}$, like in (7).
3. Compute $\text{logit}(\widehat{p}_i^{*})$ for the resulting range of values i and its variance covariance matrix $\boldsymbol{\Sigma}^{*}$, following Theorem 3.
4. Build the new design matrix $\boldsymbol{X}^{*}_{k^{*}\times m^{*}}$ according to the new desired parameters vector $\boldsymbol{\beta}^{*}_{m^{*}\times 1}$.
5. In general, setting a generalized least squares regression (Christensen 2002, pp. 33, 86) to estimate $\widehat{\boldsymbol{\beta}}^{*}_{m^{*}\times 1}$ with a new model formulated as follows:

   $$\boldsymbol{Y} = \text{logit}(\widehat{\boldsymbol{p}}^{*}) = \boldsymbol{X}^{*}\boldsymbol{\beta}^{*} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim \text{AN}(\boldsymbol{0}, \boldsymbol{\Sigma}^{*})$$

   in which the matrix $\boldsymbol{\Sigma}^{*}$ is the result of step 3. However, using the reference parameterization, the computation of both, the vector of parameters to be estimated and the variance covariance matrix, is greatly simplified by using the Theorem 4.

Finally, it is clear that the deductions have been made for the aggregate of r levels of a single factor. However, this approach does not diminish generality. If the researcher wants to group two or more factors, we just weed to iteratively apply the suggested procedure, one factor at a time. In other words, we simply applies the suggested procedure, repeatedly.

[2]Note that $\boldsymbol{X}^{*}$ has a smaller number of columns than $\boldsymbol{X}$ (hence $\boldsymbol{\beta}^{*}$ has fewer elements $\boldsymbol{\beta}$), because it models a smaller number of junctions in the levels of the factors.

4. Illustration of the Suggested Procedure

Table 1 presents a situation where the interest lies in studying the relationship between a response variable Y and two explanatory factors A_1 and A_2 with 2 and 3 levels, respectively. The observed frequencies or number of successes for each levels combination are shown in Table 1 above.

TABLE 1: Example $Y(0,1)$ vs. $A_1(1,2)$, $A_2(1,2,3)$.

i	A_1	A_2	No. of successes	Total
1	1	1	53	133
2	1	2	11	133
3	1	3	127	133
4	2	1	165	533
5	2	2	41	533
6	2	3	476	533
Total			873	1998

In this particular case, the proposed logit model omits interactions between the factors, and therefore is not saturated. Using the first level of each factor as a reference, the equations are as follows:

$$\begin{bmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \text{logit}(p_3) \\ \text{logit}(p_4) \\ \text{logit}(p_5) \\ \text{logit}(p_6) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \tag{8}$$

In (8), β_1 represents the intercept effect, β_2 represents the level 2 of A_1 effect, β_3 represents the level 2 of A_2 effect and β_4 the level 3 of A_2 effect. The levels 1 of A_1 and 1 of A_2 are not explicitly represented in the model (are the references levels).

Preliminary tests and goodness of fit for this model are shown in Table 2. The model fits the data appropriately, as it is deduced from the deviance and Pearson's statistics. Also, according to the Pearson's statistic, the overdispersion is negligible.

Table 3 contains the parameter estimates with their corresponding standard tests for $\text{H}_0 : \beta_i = 0, i = 1,2,3$ and 95% confidence intervals (CI) for β_i. The predicted probabilities and their CI's are also shown in Table 4.

Now suppose that we cluster levels 2 and 3 of the factor A_2 in Table 1, producing the Table 5 with aggregate data, and postulate the usual procedure: a new logit model.

Table 2: Original model. Preliminary tests and goodness of fit.

Test	Value
Residual deviance:	2.5065
Residual degrees of freedom (DF):	2
Deviance χ^2/DF:	0.2856
Deviance test:	No reject
Pearson's statistic:	2.3104
Pearson's χ^2/DF:	0.315
Pearson's test:	No reject
Deviance/DF:	1.2533
Pearson's/DF:	1.1552

Table 3: Original model. $\widehat{\beta}_i$ and normal tests ($H_0 : \beta_i = 0$).

	Estimation of β_i					95% CI	
i	$\widehat{\beta}_i$	SE	z-Value	p ($>\|z\|$)	Conclusion	Ll	Ul
1	−0.39857	0.14688	−2.71369	0.00666	Reject	−0.68644	−0.11070
2	−0.40725	0.15536	−2.62130	0.00876	Reject	−0.71176	−0.10275
3	−1.75603	0.16676	−10.53046	0.00000	Reject	−2.08286	−1.42919
4	2.99348	0.15665	19.10956	0.00000	Reject	2.68646	3.30051

SE: Standard Error. Ll: Lower limit. Ul: Upper limit.

Table 4: Original model. Predicted probabilities and 95% CI's.

i	$\widehat{p_i}$	Ll	Ul	i	$\widehat{p_i}$	Ll	Ul
1	0.4017	0.3325	0.4708	4	0.3088	0.2713	0.3463
2	0.1039	0.0702	0.1376	5	0.0716	0.0521	0.0912
3	0.9305	0.9068	0.9542	6	0.8991	0.8752	0.9231

Table 5: Example $Y(0,1)$ vs. $A_1(1,2)$, $A_2(1,2^*)$.

i	A_1	A_2	No. of success (y)	Total (t)	y/t
1	1	1	53	133	0.3984
2	1	2^*	138	266	0.5188
3	2	1	165	533	0.3096
4	2	2^*	517	1066	0.4850
Total			873	1998	

The new unsaturated model (ignoring interactions), using the reference parameterization (with level 1 of both factors by reference), unfolds as follows:

$$\begin{bmatrix} \text{logit}(p_1^*) \\ \text{logit}(p_2^*) \\ \text{logit}(p_3^*) \\ \text{logit}(p_4^*) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \beta_3^* \end{bmatrix} \tag{9}$$

Now in (9), β_1^* represents the intercept effect, β_2^* level 2 of A_1 effect and β_3^* the level 2^* of A_2 effect. Some measures of goodness of fit for this model are reproduced in Table 6.

The model fits the data with negligible overdispersion. Re-adjusting a logit model over the resulting contingency table, the new estimates are shown in Table

7. Without regard the parameters significance, the probabilities predicted by the model are reproduced in Table 8.

TABLE 6: Usual procedure. Testing goodness of fit of the aggregate data model.

Test	Value
Residual deviance:	1.0986
Residual degrees of freedom (DF):	1
Deviance χ^2/DF:	0.2946
Deviance test:	No reject
Pearson's statistic:	1.1051
Pearson's χ^2/DF:	0.2932
Pearson's test:	No reject
Deviance/DF:	1.0986
Pearson's/DF:	1.1051

TABLE 7: Usual procedure. $\widehat{\beta}_i^*$ and normal tests ($H_0 : \beta_i^* = 0$).

	β_i^* estimation					95% CI	
i	$\widehat{\beta}_i$	SE	z-Value	p ($>|z|$)	Conclusion	Ll	Ul
1	−0.5485	0.1220	−4.4965	0.0000	Reject	−0.7876	−0.3094
2	−0.2162	0.1137	−1.9014	0.0573	No reject	−0.4390	0.0067
3	0.6885	0.0992	6.9401	0.0000	Reject	0.4941	0.8830

TABLE 8: Usual procedure. Predicted probabilities and 95% CI without regard to model parameters statistical significance.

i	$\widehat{p}_i^*$	Ll	Ul
1	0.3662	0.3107	0.4217
2	0.5349	0.4831	0.5868
3	0.3176	0.2811	0.3542
4	0.4810	0.4519	0.5100

The new parameter vector $\boldsymbol{\beta^*}$ is estimated differently in both models (original and aggregated data). With $\alpha = 0.05$, Table 7 suggests the absence of sufficient evidence to reject the null hypothesis about β_2^*. This finding has important implications for the analysis: Since it is not possible to conclude that β_2^* is significantly different that 0, the predicted probabilities in Table 8, in strict statistical sense, should not be considered valid. Statistical valid predictions are as follows:

$$\widehat{p}_1^* = \frac{e^{\widehat{\beta}_1^*}}{1 + e^{\widehat{\beta}_1^*}} = 0.3662$$

$$\widehat{p}_2^* = \frac{e^{\widehat{\beta}_1^* + \widehat{\beta}_3^*}}{1 + e^{\widehat{\beta}_1^* + \widehat{\beta}_3^*}} = 0.5349$$

$$\widehat{p}_3^* = \frac{e^{\widehat{\beta}_1^* + \widehat{\beta}_2^*}}{1 + e^{\widehat{\beta}_1^* + \widehat{\beta}_2^*}} = \frac{e^{\widehat{\beta}_1^* + 0}}{1 + e^{\widehat{\beta}_1^* + 0}} = 0.3662 \tag{10}$$

$$\widehat{p}_4^* = \frac{e^{\widehat{\beta}_1^*+\widehat{\beta}_2^*+\widehat{\beta}_3^*}}{1+e^{\widehat{\beta}_1^*+\widehat{\beta}_2^*+\widehat{\beta}_3^*}} = \frac{e^{\widehat{\beta}_1^*+0+\widehat{\beta}_3^*}}{1+e^{\widehat{\beta}_1^*+0+\widehat{\beta}_3^*}} = 0.5349 \tag{11}$$

Finally, Table 9 contains the estimates and the 95% CIs for the logit model postulated in (9), obtained by the procedure suggested in this paper. Note that the point estimates of the standard procedure and the suggested procedure are slightly but significantly different. Using the suggested procedure, the Pearson's goodness of fit of the model produces a χ^2 of 0.0104 that leaves a probability of 0.9188 at right. Then, the model analyzed by the suggested procedure properly fits the data; in fact it fits in a better way than with the usual procedure, which produces a Pearson's statistic 1.1051 that leaves a probability of 0.2932 at right (see Table 6).

Table 9: Suggested procedure. $\widehat{\beta}_i^*$ and normal tests ($H_0 : \beta_i^* = 0$).

	β_i^* estimation					95% CIs	
i	$\widehat{\beta}_i$	SE	z-Value	p ($>\|z\|$)	Conclusion	Ll	Ul
1	−0.4685	0.1257	−3.7280	0.0002	Reject	−0.7149	−0.2222
2	−0.2673	0.1019	−2.6236	0.0087	Reject	−0.4670	−0.0676
3	0.6074	0.0931	6.5234	0.0000	Reject	0.4249	0.7899

Table 9 shows that the estimated standard errors for the parameters β_2^* and β_3^*, using the suggested procedure are lower than those found by conventional procedure (Table 7).

Table 10 presents the predicted probabilities, now fitting the data according to the procedure suggested in this paper. Note that the predicted probabilities are considerably closer to those expected for the new data set (see the column and y/t in the Table 5), than those predicted with the usual procedure.

Table 10: Suggested procedure. Predicted probabilities and 95% CIs.

i	$\widehat{p}_i^*$	Ll	Ul
1	0.4017	0.3325	0.4708
2	0.5172	0.4927	0.5417
3	0.3088	0.2713	0.3463
4	0.4854	0.4696	0.5012

Also, note in Table 9 that the conclusion about the significance of β_2^* is no lunger the same. The standard procedure statistically valid estimates (10) and (11) and look considerably different from using the suggested procedure. The predictions on Table 10 are statistically valid, approaching in a better way than would be expected from the available data.

Finally, Figure 1 presents the Pearson's standardized residuals, calculated using both methods. Clearly, the estimates produced by the suggested procedure are much closer to the expected value than those produced by the conventional procedure.

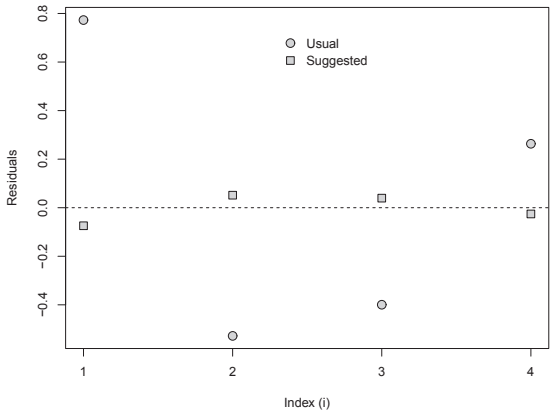

FIGURE 1: Pearson's standardized residuals calculated by both procedures.

5. Comparison of Alternative Procedures Through Simulation

In a situation with a binary response Y and two explanatory factors A_1, A_2, the first with 2 levels and the second with 3 leves, we propose a simulation in order to study the effect of the aggregation of levels 2 and 3 for the factor A_2, using pseudo-random generation of a large number of contingency tables of the type shown in Table 11.

TABLE 11: Original arrangement for simulation Y vs. $A_1(1,2)$, $A_2(1,2,3)$.

i	A_1	A_2	No. of successes (y_i)	Total (n_i)
1	1	1	y_1	n_1
2	1	2	y_2	n_2
3	1	3	y_3	n_3
4	2	1	y_4	n_4
5	2	2	y_5	n_5
6	2	3	y_6	n_6
Total			$y.$	$n.$

An unsaturated logit model is fitted to of the generated tables signoring the effect of interactions, using the first level of each factor as a reference, by

$$\begin{bmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \text{logit}(p_3) \\ \text{logit}(p_4) \\ \text{logit}(p_5) \\ \text{logit}(p_6) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \tag{12}$$

In (12), p_i represents the probability of success of the i-th combination of levels of the two explanatory factors identified in Table 11 ($i = 1, \ldots, 6$), while β_j are

the parameters to be fitted ($j = 1, \ldots, 4$). Specifically, β_1 represents the effect of the intercept, β_2 is the effect of level 2 of factor A_1, β_3 is the effect of level 2 of the factor A_2 and β_4 is the effect of the level 3 of the factor A_2.

The Table 12 is formed by grouping the last two levels of the second factor in the Table 11.

TABLE 12: Aggregated data for simulation Y vs. $A_1(1,2)$, $A_2(1,2^*)$.

i	A_1	A_2	No. of successes (y_i)	Total (n_i)
1	1	1	y_1	n_1
2	1	2^*	$y_2 + y_3$	$n_2 + n_3$
3	2	1	y_4	n_4
4	2	2^*	$y_5 + y_6$	$n_5 + n_6$
Total			$y_\cdot$	$n_\cdot$

Following the usual procedure, we set a new unsaturated logit model for the Table 12, than also ignores the effect of interactions and uses the first level of each factor as a reference:

$$\begin{bmatrix} \text{logit}(p_1^*) \\ \text{logit}(p_2^*) \\ \text{logit}(p_3^*) \\ \text{logit}(p_4^*) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \beta_3^* \end{bmatrix} \tag{13}$$

Now, in (13) p_i^* represents the probability of success of the i-th combination of levels of the two explanatory factors identified in the Table 12 ($i = 1, \ldots, 4$), with β_1^* representing the effect of the intercept, β_2^* the effect of level 2 of the factor A_1 and β_3^* the effect of level 2^* of the factor A_2.

Lastly, following the suggested procedure, from the original model results (12) we estimate the parameters of the new model with aggregate levels of the factor. The comparison between the two procedures (usual and suggested) is then used to analyze the resulting performance statistics in each case.

5.1. Design of the Experiment of Simulation

That total in the experiment set $n_\cdot = 2000$ is distributed in two randomized numbers to each level of A_1 and, within these, in three randomized numbers for each level of A_2. In this particular study, it is not of interest to compare the effect of both procedures on the levels of factor A_1, or on the first level of factor A_2, p_1 y p_4. Then, independent pseudo-randomly uniform $(0, 1)$ samples are generated. Using the generated values of n_1, n_4 (selected randomly from $n_\cdot$) and p_1 and p_4, the samples $Y_1 \sim \text{Bin}(n_1, p_1)$ and $Y_4 \sim \text{Bin}(n_4, p_4)$ are generated.

For the factor levels being compared in A_2, the samples $Y_2 \sim \text{Bin}(n_2, p_2)$, $Y_3 \sim \text{Bin}(n_3, p_3)$, $Y_5 \sim \text{Bin}(n_5, p_5)$ and $Y_6 \sim \text{Bin}(n_6, p_6)$, are generated n_j randomized as before and sequentially using combinations of $\Delta_p = |p_2 - p_3| = |p_5 - p_6| = 0.0, 0.2, 0.4, 0.6, 0.8$. Such combinations are obtained by maintaining

the values $p_2 = p_5 = 0.1$ as constant and varying by the values of $p_3 = p_6 = 0.1, 0.3, 0.5, 0.7, 0.9$.

For each combinations of Δ_p to experiment several, contingency tables are produced, regrading to by the binomials generated, which are independent within each table and between tables. We only incorporate samples that meet the following conditions:

1. Lead to acceptance of the original logit model, as assessed by the Pearson's goodness of fit.

2. Lead to an original logit model the does not present important problems on subdispertion. That is, that produces a statistical ratio of the Pearson's and residual degrees of freedom in the range $(0.75; 1.25)$.

3. Lead to acceptance of the logit model with levels 2 and 3 of the factor added A_2, also following the Pearson's test of goodness of fit.

4. Lead to a logit model with aggregate levels of the factor, which does not have important problems of subdispertion. This in order to produce a statistical ratio of the Pearson's and residual degrees of freedom in the range $(0.75; 1.25)$.

Finally, there are 10,000 valid samples, 2,000 for each combination of Δ_p, and significance level is set up with for testing $\alpha = 0.05$. The performance measures considere were:

a) Firstly, we examine descriptive statistics of the differences the Pearson's χ^2 goodness of fit test, obtained using standard procedures and suggested (in that order).

b) We compare the absolute differences in point estimates of β_1^*, β_2^* y β_3^*, obtained by the standard and suggested procedures, regardless to their statistical significance.

c) Compare the differences in the lengths of the calculated CIs using the usual and suggested procedure. It uses the average ratio between the lengths of the first and the second (in that order). These ratios are calculated for the CIs accompanying the parameter estimates β_1^*, β_2^* and β_3^*.

d) We study the absolute frequency of occurrence of the change in the conclusion of the analysis of variance (acceptance to rejection, or vice versa) for testing hypotheses about the parameters $H_0 : \beta_1^* = 0$, $H_0 : \beta_2^* = 0$ y $H_0 : \beta_3^* = 0$, when they are contrasted by the usual way, and when they are contrasted by the suggested procedure.

e) Finally, for each sample Pearson's standardized residuals produced by both methods were calculated. Also, analysis of each value of Δ_p, we construct boxplots their corresponding.

5.2. Results of the Simulation Experiment

a) Firstly, Table 13 shows means and standard deviations (SD) of the simple differences between the probabilities that leaves to the right Pearson's χ^2 test, in the examination of the goodness of fit of the model, obtained by the usual and suggested procedure.

Table 13: Mean and standard deviations of the differences to the Pearson's χ^2 probabilities (Usual - Suggested).

Δ_p	Mean	SE
0.0	−0.0002	0.0006
0.2	−0.0104	0.0214
0.4	−0.0219	0.0615
0.6	−0.0314	0.1148
0.8	−0.1173	0.1884

Since the average values in Table 13 are all negative, it is clear that the suggested procedure fits the data consistently better than the usual, with the increase in the differences Δ_p.

As evidence of goodness of fit of the model, the Pearson's statistic is particularly suitable in this case, since it is based on the accumulation of the standardized residuals. Although the variability is high, Table 13 that steadily as there are greater differences between the probabilities of the variables involved in the aggregation, the probability to the right of Pearson's χ^2 goodness of fit test increases in the suggested procedure compared with the usual.

In practice this means that, on average, the estimated parameters using the suggested procedure are closer to the expected for a given dataset in comparison to the estimates produced by the usual procedure. It also means that the model fitted using the suggested procedure is less likely to be rejected than the other model.

b) Without considering the significance of the estimated parameters, the Table 14 contains the ranges obtained by both methods (standard and suggested) for each estimate. It can be seen that these ranges are very similar in general and, as should be verified, the same when $\Delta_p = 0$, and slightly more dissimilar as Δ_p increase.

Table 15 contains the averages and standard deviations of the absolute differences between the parameters. As seen there, both the average and the standard deviation of the differences between the parameters estimated by the usual procedure (u) and suggested (s), $|\beta_i^*(\mathrm{u}) - \beta_i^*(\mathrm{s})|, i = 1, 2, 3$ behave similarly. This is, they grow as the probabilities of the variables involved in the aggregation are more dissimilar.

Nevertheless, given the ranges shown in Table 14, these differences do not seem important on average. The conclusion here is that both procedures

(usual and suggested) essentially estimate the same values of model parameters, in most situations.

TABLE 14: Ranges of $\widehat{\beta}_i^*$ according to the usual and suggested procedures.

	$\widehat{\beta}_1^*$				$\widehat{\beta}_2^*$				$\widehat{\beta}_3^*$			
	Usual		Suggested		Usual		Suggested		Usual		Suggested	
Δ_p	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
0.0	-2.43	2.51	-2.43	2.51	-1.31	1.20	-1.31	1.20	-4.69	0.43	-4.69	0.43
0.2	-2.57	2.50	-2.58	2.50	-0.70	0.92	-0.71	0.92	-4.04	1.22	-4.04	1.22
0.4	-2.38	2.65	-2.38	2.65	-0.53	0.54	-0.54	0.55	-3.51	1.63	-3.50	1.63
0.6	-2.51	2.70	-2.51	2.69	-0.57	0.45	-0.66	0.50	-3.13	2.09	-3.13	2.09
0.8	-2.44	2.52	-2.47	2.59	-0.38	0.32	-0.50	0.39	-2.54	2.52	-2.54	2.45

TABLE 15: Average of absolute differences and deviations between the parameters estimated by both methods.

	$\lvert\widehat{\beta}_1^*(u) - \widehat{\beta}_1^*(s)\rvert$		$\lvert\widehat{\beta}_2^*(u) - \widehat{\beta}_2^*(s)\rvert$		$\lvert\widehat{\beta}_3^*(u) - \widehat{\beta}_3^*(s)\rvert$	
Δ_p	Mean	SD	Mean	SD	Mean	SD
0.0	0.000	0.000	0.000	0.000	0.000	0.000
0.2	0.002	0.002	0.001	0.001	0.002	0.003
0.4	0.004	0.005	0.004	0.004	0.006	0.007
0.6	0.008	0.010	0.009	0.008	0.011	0.013
0.8	0.012	0.018	0.014	0.014	0.016	0.022

c) Regarding to the lengths of the CIs for each estimator, Table 16 presents the results of the average rates and standard deviations obtained. In general terms, the CI length for the intercept effect shows no appreciable variations in both procedures. However, for the other parameters, the higher Δ_p is the higher the average ratio of the CIs lengths estimated by both methods. Then, it consistently appears that the confidence intervals related to the suggested procedure are narrower and therefore preferable than those estimated by the usual procedure.

TABLE 16: Averages of the ratio between the lengths of confidence intervals (LCI) obtained by the usual method (u) and the suggested method (s).

	β_1^*: LCI(u)/LCI(s)		β_2^*: LCI(u)/LCI(s)		β_3^*: LCI(u)/LCI(s)	
Δ_p	Mean	SD	Mean	SD	Mean	SD
0.0	1.00	0.00	1.00	0.00	1.00	0.00
0.2	1.00	0.00	1.01	0.00	1.01	0.00
0.4	1.01	0.01	1.03	0.01	1.03	0.01
0.6	1.01	0.02	1.06	0.01	1.05	0.02
0.8	1.00	0.04	1.09	0.03	1.06	0.05

Another aspect to note is that while in average terms the conclusion is clear, the differences for the unsaturated case are not as significant as they were in the saturated case developed by Ponsot et al. (2009). The introduction of the covariance and the fact that it examines a larger number of factors have somewhat dampened these differences.

d) Table 17 shows the absolute frequencies of occurrence of the change in the conclusions on the significance of model parameters (H_0: $\beta_i^* = 0$ for $i = 1, 2, 3$), when they are examined with the usual procedure and when they are examined with the suggested procedure.

β_1^* changes occur in similar frequency and any direction. This indicates that is not possible to suggest preferences between the two procedures for intercept estimation. On the other hand, for the remaining two parameters, the conclusion about the statistical significance of not rejecting the null hypothesis and its rejection, greatly promotes the suggested procedure. Improvements in the results on β_2^* are remarkable. There was no change from rejection to acceptance of the null hypothesis, however, there were considerable changes to the contrary, i.e., acceptance to rejection of this hypothesis. The suggested procedure allows us to reject the null hypothesis of model parameters, in a higher proportion of cases, generally increasing with Δ_p.

TABLE 17: Change of the conclusions for H_0: $\beta_i^* = 0$ from the suggested procedure, compared to usual.

Δ_p	Rejection to acceptance	without changes	Acceptance to rejection
	For H_0: $\beta_1^* = 0$		
0.0	0	2000	0
0.2	1	1998	1
0.4	1	1997	2
0.6	3	1989	8
0.8	9	1983	8
	For H_0: $\beta_2^* = 0$		
0.0	0	2000	0
0.2	0	1982	18
0.4	0	1946	54
0.6	0	1951	49
0.8	0	1901	99
	For H_0: $\beta_3^* = 0$		
0.0	0	2000	0
0.2	0	2000	0
0.4	1	1992	7
0.6	1	1990	9
0.8	2	1973	25

e) Finally, the Figures from 2 to 6 contain Pearson's standardized residuals boxplots, grouped according to the procedure that gave rise to (usual and suggested), for each $y_i, i = 1, \ldots, 4$. Observe that for Δ_p from 0.0 to 0.4, boxplots not vary appreciably, indicating that the residuals produced by both procedures are very similar. However, when Δ_p is greater, although the variability of residuals becomes less stable, their averages are closer to 0 in those settings using the suggested procedure. This confirms that in trend terms, the suggested procedure produces better fits than the usual procedure[3].

[3]All programs, both for example, as for the simulation, were made with R statistical system (R Development Core Team 2007).

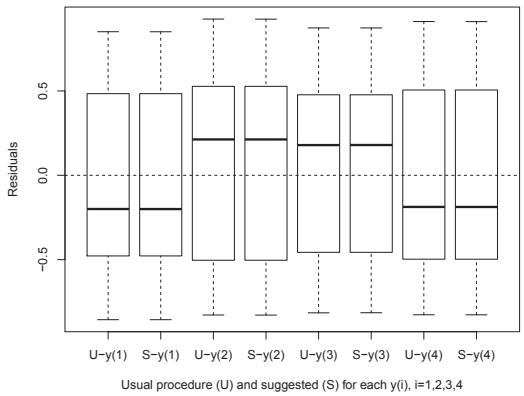

FIGURE 2: Pearson's standardized residuals boxplots for$\Delta_p = 0.0$.

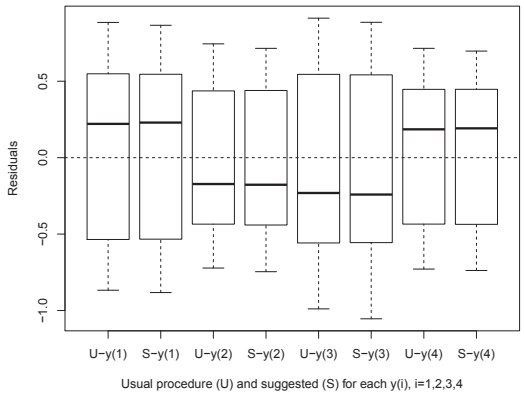

FIGURE 3: Pearson's standardized residuals boxplots for $\Delta_p = 0.2$.

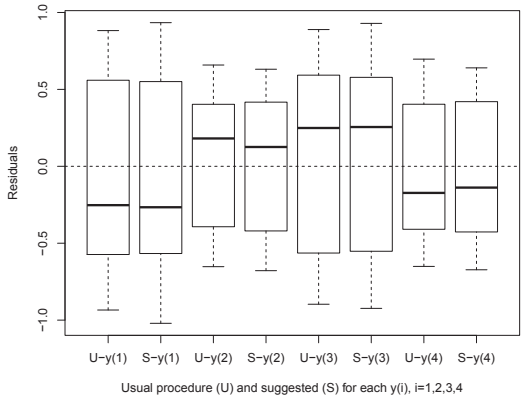

FIGURE 4: Pearson's standardized residuals boxplots for $\Delta_p = 0.4$.

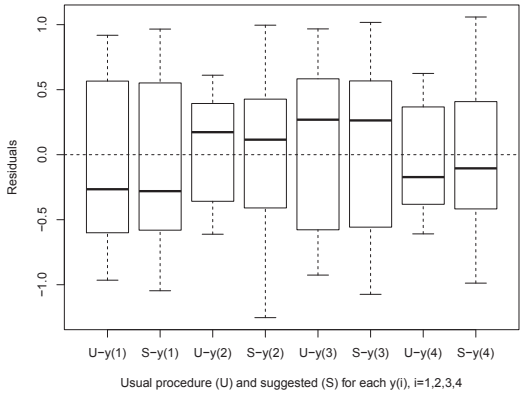

FIGURE 5: Pearson's standardized residuals boxplots for $\Delta_p = 0.6$.

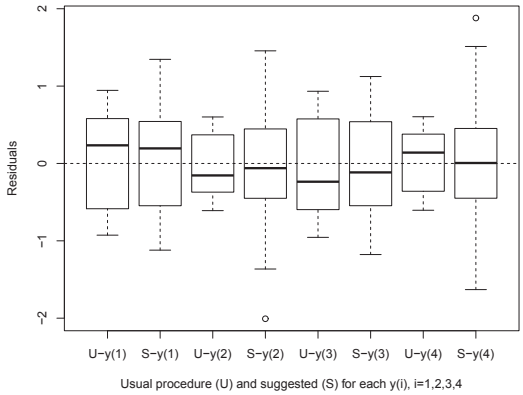

FIGURE 6: Pearson's standardized residuals boxplots for $\Delta_p = 0.8$.

6. Conclusions

This paper addresses and resolves a problem rarely studied, which arises from the practical application of the binomial logit model. We discuss the situation in which, once fitted a logit model to the data in a contingency table, a factor from any of the participants is selected and some levels are added as a new level, to reiterate a logit setting.

In general, there is a problem in the logit model fit with aggregate levels of the factor, particularly when the probabilities of success of RV's involved in aggregation are far from each other. Consequently, this paper suggests a procedure that operates in a broader context, i.e., under the binomial unsaturated multifactorial logit model, and with arguments of asymptotic nature, taking advantage of the reduction in variance when postulates proper distributional model instead of the binomial model, significantly improves the estimates, while lowering the standard error.

As the difference in the probabilities of success accentuates, it becomes better supported by the suggested procedure, instead of the usual. The model fitted by the suggested procedure, also produces closer to zero residuals and less chance of rejection in the goodness of fit test.

In summary, it is proposed to the researcher logit model user, an alternative procedure that can provide theoretical correctness, greater accuracy and less computational effort in the state of aggregation levels of a factor, especially when they involve sample proportions which are markedly dissimilar.

Acknowledgements

The authors thank the Council of Scientific, Humanistic, Technology and the Arts (CDCHTA) of the Los Andes University, the financial support to carry out this work, registered with the code E-303-09-09-ED.

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

References

Christensen, R. (2002), *Plain Answers to Complex Questions. The Theory of Linear Models*, 3 edn, Springer-Verlag, Nueva York, Estados Unidos.

Graybill, F. (1969), *Introduction to Matrices with Applications in Statistics*, 1 edn, Wadsworth Publishing, California, Estados Unidos.

Hilbe, J. M. (2009), *Logistic Regression Models*, 1 edn, Chapman & Hall, Florida, Estados Unidos.

Hosmer, D. W. & Lemeshow, S. (2000), *Applied Logistic Regression*, 2 edn, John Wiley & Sons, Nueva York, Estados Unidos.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, 1 edn, Springer-Verlag, Nueva York, Estados Unidos.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, 2 edn, Chapman & Hall, London, United Kingdom.

Menard, S. (2010), *Logistic Regression: From Introductory to Advanced Concepts and Applications*, 1 edn, SAGE Publications, Inc., California, Estados Unidos.

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized Linear Models', *Journal of the Royal Statistical Society. Serie A* (135), 370–384.

Ponsot, E. (2011), Estudio de la Agrupación de Niveles en el Modelo Logit, Unpublisehd PhD Thesis, Instituto de Estadística Aplicada y Computación, Facultad de Ciencias Económicas y Sociales, Universidad de Los Andes, Mérida, Venezuela.

Ponsot, E., Sinha, S. & Goitía, A. (2009), 'Sobre la agrupación de niveles del factor explicativo en el modelo logit binario', *Revista Colombiana de Estadística* **32**(2), 157–187.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org

Rodríguez, G. (2008), 'Lectures notes about generalized linear models'. *Http://data.princeton.edu/wws509/notes

SAS Institute Inc. (2004), *SAS/STAT(R) 9.1 User's Guide*, SAS Institute Inc., Carolina del Norte, Estados Unidos.

Searle, S., Casella, G. & McCulloch, C. (2006), *Variance Components*, 1 edn, John Wiley and Sons, Inc., Nueva Jersey, Estados Unidos.

Appendix. Study of the Design Matrix $\boldsymbol{X}$ for Saturated and Unsaturated Models

Theorem 5. *Using the reference parameterization, the design matrix of the saturated logit model is invertible.*

Proof. We prove the invertibility of the matrices of design, both in the univariate situation, as in the multifactorial situation, then:

1. Let the saturated univariate logit model design matrix be:

$$\boldsymbol{X}_{k\times k} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

   The calculation of the determinant of $\boldsymbol{X}$ by cofactors decomposition (X_{ij}) is $|\boldsymbol{X}| = (-1)^{k+1}|I| = (-1)^{k+1} \neq 0$, pivoting the last row of the matrix, since the only one nonzero element is x_{k1}. Then, since $|\boldsymbol{X}| \neq 0$, $\boldsymbol{X}^{-1}$ exists.

2. Suppose $A_1, A_2, \ldots, A_s$ factors, each one $t_1, t_2, \ldots, t_s$ levels, respectively. Using the reference parameterization, the followings are postulated:

   a) 1 parameter for the effect of the intercept.

   b) $t_1 - 1$ parameters for the main effects of A_1 factor levels, except the reference; $t_2 - 1$ parameters for the main effects of A_2 factor levels, except the reference; and so on until $t_s - 1$ parameters for the main effects of A_s factor levels, except the reference; in total, $\sum_{i=1}^{s}(t_i - 1)$ parameters for the main effects.

c) $(t_1-1)(t_2-1)$ parameters for the double interaction effects between levels of the factors A_1 and A_2; $(t_1-1)(t_3-1)$ parameters for the double interaction effects between levels of the factors A_1 and A_3; so on until $(t_{s-1}-1)(t_s-1)$ parameters for the double interaction effects between levels of the factors A_{s-1} and A_s; in total$\sum_{i=1}^{s-1}\sum_{j=i+1}^{s}(t_i-1)(t_j-1)$.

d) $(t_1-1)(t_2-1)(t_3-1)$ parameters for the triple effects of interaction between levels of the factors A_1, A_2 and A_3; $(t_1-1)(t_2-1)(t_4-1)$ parameters for the triple effects of interaction between levels of the factors A_1, A_2 and A_4; so on until $(t_{s-2}-1)(t_{s-1}-1)(t_s-1)$ parameters for the triple effects of interaction between levels of the factors A_{s-2}, A_{s-1} and A_s; in total $\sum_{i=1}^{s-2}\sum_{j=i+1}^{s-1}\sum_{k=j+1}^{s}(t_i-1)(t_j-1)(t_k-1)$.

In general, for order a interactions $(1 \leq a \leq s)$ the followings parameters are postulated

$$\sum_{i_1=1}^{s-a+1}\sum_{i_2=i_1+1}^{s-a+2}\cdots\sum_{i_a=i_{a-1}+1}^{s}\prod_{j=1}^{a}(t_{i_j}-1)$$

As the model is saturated, the k total number of postulated parameters equals the number of observations in the contingency table ($k = t_1 \times t_2 \times \cdots \times t_s$). Now, including the interaction of order a $(1 \leq a \leq s)$ in its $i_1, i_2, \ldots, i_a$ levels, requires a row of $\boldsymbol{X}$ like:

$$\begin{bmatrix} 1 & x_1 \cdots x_s & x_{12} \cdots x_{(s-1)s} & \cdots & x_{i_1 i_2 \cdots i_a} & 0 \cdots 0 \end{bmatrix}$$

where

$$x_i = \begin{cases} 1, & i \in \{i_1, i_2, \ldots, i_a\} \\ 0, & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1, & i \text{ y } j \in \{i_1, i_2, \ldots, i_a\} \\ 0, & \text{otherwise} \end{cases}$$

and so on until $x_{i_1 i_2 \cdots i_a} = 1$. In other words, the equation that introduces a new parameter representing the interaction of any kind involves only the parameter representing this interaction, and those representing the lower-order interactions contained in it.

Appropriately arranging the rows of $\boldsymbol{X}$ thus constructed, it is easy to verify that a triangular matrix is formed, whose diagonal consists of ones only. Then, using Theorems 1.5.3 and 8.6.5 of Graybill (1969, pp. 8, 191), $|\boldsymbol{X}| \equiv \pm 1 \neq 0$, and therefore $\boldsymbol{X}^{-1}$ exist.

□

Corollary 1. *With reference to parameterization, the design matrix $\boldsymbol{X}$ of the logit model is such that there is $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.*

Proof. Clearly in the saturated model situation, as there is $\boldsymbol{X}^{-1}$, $(\boldsymbol{X}^T)^{-1}$ exist and then $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ also exist.

In the unsaturated model situation, the design matrix $\boldsymbol{X}$ is no longer square and has no inverse. However, the unsaturated model starts from the saturate model ignoring parameters in reverse order of the interactions (high order to low order) as desired by the researcher, always following the construction rules described in item 2 of Theorem 5. Therefore, the construction of an unsaturated model is produced by simply removing

columns in the design matrix of the corresponding saturated model. However, as the columns of the saturated model design matrix are linearly independent, any subset of the columns in it (in this case $\boldsymbol{X}$) is also such that its columns are linearly independent, whereby the unsaturated model matrix is columns full range and following the corollary B.53 of Christensen (2002, p. 415), $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exist. □

Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 167 a 184

Random Time-Varying Coefficient Model Estimation through Radial Basis Functions

Estimación de los coeficientes de un modelo de coeficientes dinámicos y aleatorios a través de funciones radiales *kernel*

Juan Camilo Sosa[1,a], Luis Guillermo Díaz[2,b]

[1]Mathematics Department, Universidad Externado de Colombia, Bogotá, Colombia
[2]Statistics Department, Faculty of Science, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

A methodology to estimate a time-varying coefficient model through a linear combination of radial kernel functions which are centered around all the measuring times, or their quantiles is developed. The linear combination is weighted by a bandwidth that may change or not among coefficients. The proposed methodology is compared with the local polynomial kernel methods by means of a simulation study. The proposed methodology shows a better behavior in a high proportion of times in all cases, or at least it has a similar behavior in relation with the estimation through local polynomial kernel regression, that in a low rate of times has a better behavior in relation with the average mean square error. In order to illustrate the methodology the data set ACTG 315 related with an AIDS study is taken into account. The dynamic relationship between the viral load and the CD4+ cell counts is investigated.

***Key words*:** Cross validation, Kernel function, Longitudinal data analysis, Mixed model.

Resumen

Se propone una metodología para estimar los coeficientes de un modelo de coeficientes dinámicos y aleatorios a través de una combinación lineal de funciones radiales kernel centradas en los diferentes puntos de medición, o en cuantiles de éstos, escalada por un ancho de banda que puede cambiar de coeficiente a coeficiente. En un estudio de simulación se compara la metodología propuesta con la estimación mediante los métodos de polinomios locales *kernel*, obteniéndose que la nueva metodología propuesta es la

[a]Lecturer. E-mail: juan.sosa@uexternado.edu.co
[b]Associated professor. E-mail: lgdiazm@unal.edu.co

mejor opción en un alto porcentaje de veces para todos los escenarios simulados, o por lo menos se desempeña similarmente a la estimación a través de la regresión de polinomios locales *kernel*, que pocas veces se desempeña mejor que la estimación mediante funciones radiales *kernel*, en relación al error cuadrático medio promedio. Para ilustrar la estrategia de estimación propuesta se considera el conjunto de datos ACTG 315 asociado con un estudio del SIDA, en el que se modela dinámicamente la relación entre la carga viral y el conteo de células CD4+.

***Palabras clave*:** análisis de datos longitudinales, función *kernel*, modelo mixto, validación cruzada.

1. Introduction

Longitudinal Data Analysis (LDA) takes place when a set of subjects are observed repeatedly along time, measuring the response variable in accordance with the covariates that may or may not be time-dependent. Given the characteristics of this kind of data, an underlying property that must be thought fitting a statistical model, is the correlation between repeated measures of the response variable within each experimental unit, considering measures independent between subjects. That is, measurements are correlated inside experimental units and independent between subjects. This way, the main purpose is to identify and describe the evolution of the response variable and to determine how it is affected by the covariates. For instance, in clinic trials, it is of interest to evaluate the impact of a dose or other related factors, over the progress of a disease along time.

Parametric techniques for LDA have been exhaustively studied in the literature (Diggle, Liang & Zeger 1994, Davis 2000, Verbeke & Molenberghs 2005, Fitzmaurice, Davidian, Verbeke & Molenberghs 2009). While these tools are useful under some reasonable restrictions, always arise doubts and questions about the adequacy of the model assumptions and the potencial impact of model misspecifications on the analysis (Hoover, Rice, Wu & Yang 1998). Non parametric techniques recently introduced in LDA allow a functional dependence more flexible between the response variable and the covariates.

Hart & Wehrly (1986), Altman (1990), Hart (1991) propose methods for choosing smoothing parameter through cross-validation using kernel functions and considered kernel methods for estimating the expectation of the response variable without covariates, while Rice & Silverman (1991) did it by using a class of smoothing splines. Although the kernel and splines methods are successful in predicting the mean curve of the response variable, they only consider the time effect and do not take into account other important covariates (Hoover et al. 1998).

In order to quantify the influence of covariates, Zeger & Diggle (1994) studied a semi-parametric model as follows:

$$\begin{gathered} y_{ij} = \mu(t_{ij}) + \boldsymbol{x}_i(t_{ij})^T\boldsymbol{\beta} + e_{ij} \\ j = 1, \ldots, n_i,\ i = 1, \ldots, n \end{gathered} \tag{1}$$

where n is the number of subjects, n_i is the number of repeated measures associated with the i-th experimental unit, t_{ij}, $y_{ij} \equiv y_i(t_{ij})$,

$$\boldsymbol{x}_i(t_{ij}) = [x_{i0}(t_{ij}), x_{i1}(t_{ij}), \ldots, x_{id}(t_{ij})]^T$$

and $e_{ij} \equiv e_i(t_{ij})$ are respectively the measuring time, the response variable, the covariate vector in $\mathbb{R}^{d+1}$ and the error term, associated with the j-th measure of the i-th subject. Moreover, $\mu(\cdot)$ is an arbitrary smooth real function and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_d]^T$ is a parameter vector in $\mathbb{R}^{d+1}$. Working with longitudinal data, it is usually assumed that repeated measures are independent between experimental units and that $e_i(t)$ is a Gaussian Process (GP) with $\mathbb{E}[e_i(t)] = 0$, for each $t \in \mathcal{T}_i$, with covariance function $\gamma_{\boldsymbol{e}_i}(r, s)$, $r, s \in \mathcal{T}_i$, and $\mathcal{T}_i = \{t_{ij} : j = 1, \ldots, n_i\}$; this is written as

$$\boldsymbol{e}_i = [e_{i1}, \ldots, e_{in_i}]^T \sim PG(\mathbf{0}_{n_i}, \boldsymbol{\Gamma}_i)$$

where $\mathbf{0}_{n_i}$ is a column-vector with $n_i \times 1$ zeros and $\boldsymbol{\Gamma}_i = [\gamma_{\boldsymbol{e}_i}(t_{ik}, t_{il})]_{k,l=1,\ldots,n_i}$.

Hoover et al. (1998) considered a generalization of the model (1) that allows the parameters to vary over time. This extension is as follows:

$$\begin{aligned} y_{ij} &= \boldsymbol{x}_i(t_{ij})^T \boldsymbol{\beta}(t_{ij}) + e_{ij}, \\ j &= 1, \ldots, n_i,\ i = 1, \ldots, n \end{aligned} \tag{2}$$

where

$$\boldsymbol{\beta}(t_{ij}) = [\beta_0(t_{ij}), \beta_1(t_{ij}), \ldots, \beta_d(t_{ij})]^T$$

is a vector of arbitrary real smooth functions. Components in vector $\boldsymbol{\beta}(t)$ are called dynamic coefficients or dynamic parameters, and the statistical model (2) is referred as Time-Varying Coefficient Model (TVCM). This kind of model has been widely studied by Wu & Zhang (2006) who investigated various alternatives for estimating the model coefficients. Sosa & Díaz (2010) proposed a methodology to estimate true-varying coefficients models through generalized estimation equations.

A Random Time-Varying Coefficient Model (RVCM) is an extension of a TVCM, and it was firstly investigated by Guo (2002). As in a Linear Mixed Effects Model (LMEM), this extension decomposes the term error $e_i(t_{ij})$ of model (2) into two parts: one of them that describes the characteristics of each subject that differs of the mean population behavior, and other related with the pure random error; that is, it is done by the decomposition

$$\begin{aligned} e_i(t_{ij}) &= \boldsymbol{z}_i(t_{ij})^T \boldsymbol{v}_i(t_{ij}) + \epsilon_i(t_{ij}) \\ j &= 1, \ldots, n_i,\ i = 1, \ldots, n \end{aligned}$$

where $\boldsymbol{z}_i(t_{ij})^T \boldsymbol{v}_i(t_{ij})$ is the model component that describes the characteristics related with each subject (random effects component), with

$$\boldsymbol{z}_i(t_{ij}) = [z_{i0}(t_{ij}), z_{i1}(t_{ij}), \ldots, z_{id^*}(t_{ij})]^T$$

a covariate vector in $\mathbb{R}^{d^*+1}$, with components that vary along time, associated with the vector

$$\boldsymbol{v}_i(t_{ij}) = [v_{i0}(t_{ij}), v_{i1}(t_{ij}), \ldots, v_{id^*}(t_{ij})]^T$$

of random time-varying coefficients with size $(d^* + 1) \times 1$ and $\epsilon_{ij} \equiv \epsilon_i(t_{ij})$ is the random error term associated with the j-th measurement of the i-th experimental unit. Thus, a RVCM is a model with the following form:

$$\begin{gathered} y_{ij} = \boldsymbol{x}_i(t_{ij})^T \boldsymbol{\beta}(t_{ij}) + \boldsymbol{z}_i(t_{ij})^T \boldsymbol{v}_i(t_{ij}) + \epsilon_{ij} \\ j = 1, \ldots, n_i,\ i = 1, \ldots, n \end{gathered} \tag{3}$$

where

$$\boldsymbol{v}_i(t) \sim PG(\mathbf{0}_{d^*+1}, \boldsymbol{\Gamma})$$

and

$$\boldsymbol{\epsilon}_i(t) = [\epsilon_{i1}, \ldots, \epsilon_{in_i}]^T \sim PG(\mathbf{0}_{n_i}, \mathbf{R}_i)$$

with $\boldsymbol{\Gamma} = [\gamma(t_{ik}, t_{il})]_{k,l=1,\ldots,d^*+1}$ and $\mathbf{R}_i = [\gamma_{\boldsymbol{\epsilon}_i}(t_{ik}, t_{il})]_{k,l=1,\ldots,n_i}$. It is supposed that the repeated measurements are independent between subjects, and $\boldsymbol{v}_i(t)$ and $\boldsymbol{\epsilon}_{i^*}(t)$ are independent Gaussian processes.

This paper is structured as follows: In Section 2 and Section 3 the estimation through local polynomial kernel techniques is presented and an estimation methodology by means of radial kernel functions is proposed, respectively. In Section 4 some techniques to choose the bandwidth associated with the estimation methodologies is studied. In section 5 it is shown a simulation study where the estimation alternatives through the average mean square error are compared. In Section 6 the methodology is illustrated by analyzing the data set ACTG 315 (Liang, Wu & Carroll 2003), where the relationship between viral load and CD4+ cell counts is investigated dynamically in a AIDS clinical trial. Finally, results are discussed in 7.

2. Estimation Through Local Polynomial Kernel Regression

The basic idea behind the estimation through Local Polynomial Kernel (LPK) regression is to approximate the dynamic coefficients by means of a Taylor expansion. Thus, in a fix time point t_0, it is supposed that the dynamic parameters $\beta_r(t_0)$, $r = 0, 1, \ldots, d$, and $v_{is}(t_0)$, $s = 0, 1, \ldots, d^*$, have $(p+1)$ continuous derivatives for some non-negative integer p. Then, by means of an approximation in a Taylor expansion of order p around t_0, it follows that:

$$\beta_r(t_{ij}) \approx \boldsymbol{h}_{ij}^T \boldsymbol{\alpha}_r,\ r = 0, 1, \ldots, d \tag{4}$$

and

$$v_{si}(t_{ij}) \approx \boldsymbol{h}_{ij}^T \boldsymbol{b}_{si},\ s = 0, 1, \ldots, d^* \tag{5}$$

for $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, where

$$\boldsymbol{h}_{ij} = [1, t_{ij} - t_0, (t_{ij} - t_0)^2, \ldots, (t_{ij} - t_0)^p]^T$$

is the vector of $(p+1)\times 1$ components related with the polynomials in the approximation, $\boldsymbol{\alpha}_r = [\alpha_{r0}, \alpha_{r1}, \ldots, \alpha_{rp}]^T$ and $\boldsymbol{b}_{si} = [b_{si0}, b_{si1}, \ldots, b_{sip}]^T$, with

$$\alpha_{rk} = \frac{\beta_r^{(k)}(t_0)}{k!} \tag{6}$$

and

$$b_{sik} = \frac{v_{si}^{(k)}(t_0)}{k!} \tag{7}$$

for $k = 0, 1, \ldots, p$.

Let $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_d^T]^T$ and $\boldsymbol{b}_i = [\boldsymbol{b}_{0i}^T, \boldsymbol{b}_{1i}^T, \ldots, \boldsymbol{b}_{d^*i}^T]^T$ be the vectors associated with the approximation of the dynamic coefficients. Given that the repeated measurements are independent between subjects and that $\boldsymbol{v}_i(t) \sim PG(\mathbf{0}_{d^*+1}, \boldsymbol{\Gamma})$, it follows that the sequence of vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ constitutes a random sample from a population with a multivariate normal distribution with mean $\mathbf{0}_{(d^*+1)(p+1)}$ and covariance matrix $\mathbf{D} \equiv \mathbf{D}(t_0)$ with size $d^*(p+1)\times d^*(p+1)$. Thus, in a neighborhood of t_0, model (3) can be approximately expressed as

$$\begin{gathered} y_{ij} \approx \boldsymbol{x}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \epsilon_{ij} \\ j = 1, \ldots, n_i,\ i = 1, \ldots, n \end{gathered} \tag{8}$$

where $\boldsymbol{x}_{ij} = \boldsymbol{x}_i(t_{ij}) \otimes \boldsymbol{h}_{ij}$, $\boldsymbol{z}_{ij} = \boldsymbol{z}_i(t_{ij}) \otimes \boldsymbol{h}_{ij}$, with $\boldsymbol{b}_i \sim N(\mathbf{0}_{(d^*+1)(p+1)}, \mathbf{D})$ and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}_{n_i}, \mathbf{R}_i)$

Thus, in a neighborhood of t_0, model (8) is a standard LMEM where it is required to estimate $\boldsymbol{\alpha}$ and find the Best Linear Unbiased Predictor (BLUP) of $\boldsymbol{b}_i$, with the purpose of finding the estimations of $\boldsymbol{\beta}(t)$ and $\boldsymbol{v}_i(t)$. In order to incorporate the information given in the neighborhood, as in Wu & Zhang (2006, p. 297), it is constituted the following objective function:

$$(\boldsymbol{y} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{b})^T\mathbf{K}_h^{1/2}\mathbf{R}^{-1}\mathbf{K}_h^{1/2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{b}) + \boldsymbol{b}^T\widetilde{\mathbf{D}}^{-1}\boldsymbol{b} \tag{9}$$

where

$$\begin{aligned} \boldsymbol{b} &= [\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T]^T \\ \boldsymbol{y} &= [\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T]^T, \quad \boldsymbol{y}_i = [y_{i1}, \ldots, y_{in_i}]^T \\ \mathbf{X} &= [\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T]^T, \quad \mathbf{X}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}]^T \\ \mathbf{Z} &= diag[\mathbf{Z}_1, \ldots, \mathbf{Z}_n], \quad \mathbf{Z}_i = [\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i}]^T \\ \widetilde{\mathbf{D}} &= diag[\mathbf{D}, \ldots, \mathbf{D}], \quad \mathbf{R} = diag[\mathbf{R}_1, \ldots, \mathbf{R}_n] \\ \mathbf{K}_h &= diag[\mathbf{K}_{1h}, \ldots, \mathbf{K}_{nh}], \quad \mathbf{K}_{ih} = diag[K_h(t_{i1} - t_0), \ldots, K_h(t_{in_i} - t_0)] \end{aligned} \tag{10}$$

with $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ a kernel function and h a bandwidth.

The estimators can be found fitting the model

$$\begin{gathered} \widetilde{\boldsymbol{y}} = \widetilde{\mathbf{X}}\boldsymbol{\alpha} + \widetilde{\mathbf{Z}}\boldsymbol{b} + \boldsymbol{\epsilon} \\ \boldsymbol{b} \sim N(\mathbf{0}_N, \widetilde{\mathbf{D}}),\ \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \mathbf{R}) \end{gathered} \tag{11}$$

where $\widetilde{\boldsymbol{y}} = \mathbf{K}_h^{1/2}\boldsymbol{y}$, $\widetilde{\mathbf{X}} = \mathbf{K}_h^{1/2}\mathbf{X}$, $\widetilde{\mathbf{Z}} = \mathbf{K}_h^{1/2}\mathbf{Z}$ and $N = \sum_{i=1}^{n} n_i$.

Therefore, given the variance components $\widetilde{\mathbf{D}}$ and $\mathbf{R}$, the kernel function $K(\cdot)$ and the bandwidth h, to minimize (9) in relation with $\boldsymbol{\alpha}$ and $\boldsymbol{b}$ leads to

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{X}^T\mathbf{K}_h^{1/2}\mathbf{V}^{-1}\mathbf{K}_h^{1/2}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{K}_h^{1/2}\mathbf{V}^{-1}\mathbf{K}_h^{1/2}\boldsymbol{y} \tag{12}$$

$$\widehat{\boldsymbol{b}} = \widetilde{\mathbf{D}}\mathbf{Z}^T\mathbf{K}_h^{1/2}\mathbf{V}^{-1}\mathbf{K}_h^{1/2}(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\alpha}}) \tag{13}$$

and

$$\widehat{\boldsymbol{b}}_i = \mathbf{D}\mathbf{Z}_i\mathbf{K}_{ih}^{1/2}\mathbf{V}_i^{-1}\mathbf{K}_{ih}^{1/2}(\boldsymbol{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\alpha}})$$

where

$$\mathbf{V} = diag[\mathbf{V}_1, \ldots, \mathbf{V}_n] = \mathbf{K}_h^{1/2}\mathbf{Z}\widetilde{\mathbf{D}}\mathbf{Z}^T\mathbf{K}_h^{1/2} + \mathbf{R}$$

with

$$\mathbf{V}_i = \mathbf{K}_{ih}^{1/2}\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T\mathbf{K}_{ih}^{1/2} + \mathbf{R}_i$$

3. Estimation through Radial Kernel Functions

The idea behind the estimation through Radial Kernel Functions (RKF) is to approximate the dynamic coefficients by means of a linear combination of kernel functions treated as radial basis functions. Thus, it is possible to express the dynamic parameters by means of

$$\boldsymbol{\beta}(t) = \boldsymbol{\Xi}(t)^T\boldsymbol{\alpha} \tag{14}$$

and

$$\boldsymbol{v}_i(t) = \boldsymbol{\Theta}(t)^T\boldsymbol{b}_i,\ i = 1, \ldots, n \tag{15}$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_d^T]^T$, $\boldsymbol{\Xi}(t) = diag[\boldsymbol{\Xi}_0(t), \boldsymbol{\Xi}_1(t), \ldots, \boldsymbol{\Xi}_d(t)]$,

$$\boldsymbol{\alpha}_r = [\alpha_{r1}, \ldots, \alpha_{rM}]^T\ \boldsymbol{\Xi}_r(t) = \left[\xi_r\left(\frac{|t - t_1|}{h}\right), \ldots, \xi_r\left(\frac{|t - t_M|}{h}\right)\right]^T$$

for $r = 0, 1, \ldots, d$, $\boldsymbol{b}_i = [\boldsymbol{b}_{0i}^T, \boldsymbol{b}_{1i}^T, \ldots, \boldsymbol{b}_{d^*i}^T]^T$, $\boldsymbol{\Theta}(t) = diag[\boldsymbol{\Theta}_0(t), \boldsymbol{\Theta}_1(t), \ldots, \boldsymbol{\Theta}_{d^*}(t)]$

$$\boldsymbol{b}_{si} = [b_{si1}, \ldots, b_{siM}]^T\ \boldsymbol{\Theta}_s(t) = \left[\theta_s\left(\frac{|t - t_1|}{h}\right), \ldots, \theta_s\left(\frac{|t - t_M|}{h}\right)\right]^T$$

for $i = 1, \ldots, n$, $s = 0, 1, \ldots, d^*$, with $\xi_r : [0, \infty) \to \mathbb{R}$ and $\theta_s : [0, \infty) \to \mathbb{R}$ kernel functions, $t_1, \ldots, t_M$ are all the M measurements time points that are different (or quantils of these) and h is a bandwidth.

If $\xi_r \equiv \xi$ for each $r = 0, 1, \ldots, d$ and $\theta_s \equiv \theta$ for each $s = 0, 1, \ldots, d^*$, then

$$\boldsymbol{\beta}(t) = \left[\mathbf{I}_{d+1} \otimes \boldsymbol{\xi}(t)\right]^T\boldsymbol{\alpha} \tag{16}$$

and

$$\boldsymbol{v}_i(t) = \left[\mathbf{I}_{d^*+1} \otimes \boldsymbol{\theta}(t)\right]^T\boldsymbol{b}_i,\ i = 1, \ldots, n \tag{17}$$

where $\mathbf{I}_k$ denote the identity matrix of $k \times k$,

$$\boldsymbol{\xi}(t) = \left[\xi\left(\frac{|t-t_1|}{h}\right), \ldots, \xi\left(\frac{|t-t_M|}{h}\right)\right]^T \tag{18}$$

and

$$\boldsymbol{\theta}(t) = \left[\theta\left(\frac{|t-t_1|}{h}\right), \ldots, \theta\left(\frac{|t-t_M|}{h}\right)\right]^T \tag{19}$$

As above, given that $\boldsymbol{v}_i(t) \sim PG(\mathbf{0}_{d^*+1}, \boldsymbol{\Gamma})$ and that the repeated measurements are independent between subjects, it follows that the sequence of vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ constitutes a random sample from a population with a multivariate normal distribution with mean $\mathbf{0}_{(d^*+1)(p+1)}$ and covariance matrix $\mathbf{D} \equiv \mathbf{D}(t)$ with size $d^*(p+1) \times d^*(p+1)$. Due to (3) and (15), it follows that $\gamma(s,t) = \boldsymbol{\Theta}(s)^T \mathbf{D} \boldsymbol{\Theta}(t)$, so that an estimator of $\mathbf{D}$ leads directly to an estimator of $\boldsymbol{\Gamma}$.

Thus, model (3) can be approximately expressed as

$$\begin{gathered} y_{ij} \approx \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i + \epsilon_{ij} \\ j = 1, \ldots, n_i,\ i = 1, \ldots, n \end{gathered} \tag{20}$$

where $\boldsymbol{x}_{ij} = \boldsymbol{\Xi}(t_{ij})\boldsymbol{x}_i(t_{ij})$ and $\boldsymbol{z}_{ij} = \boldsymbol{\Theta}(t_{ij})\boldsymbol{z}_i(t_{ij})$, with $\boldsymbol{b}_i \sim N(\mathbf{0}_{(d^*+1)(p+1)}, \mathbf{D})$ and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}_{n_i}, \mathbf{R}_i)$.

If $\xi_r \equiv \xi$ and $\theta_s \equiv \theta$ then

$$\boldsymbol{x}_{ij} = (\mathbf{I}_{d+1} \otimes \boldsymbol{\xi}(t_{ij}))\boldsymbol{x}_i(t_{ij})$$

and

$$\boldsymbol{z}_{ij} = (\mathbf{I}_{d^*+1} \otimes \boldsymbol{\theta}(t_{ij}))\boldsymbol{z}_i(t_{ij})$$

where $\boldsymbol{\xi}(t)$ and $\boldsymbol{\theta}(t)$ are given in (18) and (19).

Given the vectors $\boldsymbol{\Xi}_r(t)$, $r = 0, 1, \ldots, d$, and $\boldsymbol{\Theta}_s(t)$, $s = 0, 1, \ldots, d^*$, and the bandwidth h, model (20) is a standard LMEM where it is required to estimate $\boldsymbol{\alpha}$ and find the BLUP of $\boldsymbol{b}_i$ in order to calculate the estimations of $\boldsymbol{\beta}(t)$ and $\boldsymbol{v}_i(t)$.

4. Bandwidth Selection

By estimating the dynamic components of model 3 through LPK or RKF, it is mandatory to choose the bandwidth h carefully. In this section are presented two selection criterions designed to choose smoothing parameters: Subject Cross-Validation (SCV) and Point Cross-Validation (PCV).

4.1. Subject Cross-Validation

This criterion was proposed by Rice & Silverman (1991), and has been studied by many authors, as Hoover et al. (1998) for instance. The idea behind this criteria

is to choose the smoothing parameter vector that minimize the expression

$$SCV(h) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_i \left[y_{ij} - \boldsymbol{x}_i(t_{ij})^T \widehat{\boldsymbol{\beta}}^{(-i)}(t_{ij}) \right]^2 \tag{21}$$

where y_{ij} and $x_i(t_{ij})$ are defined as in model (3), $\widehat{\boldsymbol{\beta}}^{(-i)}(t)$ denotes the estimation of $\boldsymbol{\beta}(t)$ excluding the data related with the i-th subject, and w_i for $i = 1 \ldots, n$, is a weight given by some of the following schemes:

Scheme 1 All weights are given by $w_i = 1/N$, $i = 1, \ldots, n$, where $N = \sum_{i=1}^{n} n_i$.

Scheme 2 All weights are given by $w_i = 1/(nn_i)$, $i = 1, \ldots, n$.

Scheme 1 uses the same weight for all experimental units and was proposed by Hoover et al. (1998). Scheme 2 is considered by Huang, Wu & Zhou (2002) and uses different weights for the subjects taken into account in the study. In Huang et al. (2002) it is shown that scheme 1 could lead to inconsistent estimators of $\boldsymbol{\alpha}$.

4.2. Point Cross Validation

Let $\{t_l : l = 1, \ldots, M\}$ be the set formed by all the measuring times that are different (or quantiles of these) in all the data set. For a given time point t_l, let $\{i_{l^*} : l^* = 1, \ldots, m_l\}$ be the set of all experimental units at time t_l.

The idea behind this criteria is to choose the smoothing parameter vector that minimize the expression

$$PCV(h) = \sum_{l=1}^{M} \sum_{l^*=1}^{m_l} w_l \left[y_{i_{l^*}}(t_{l^*}) - \widehat{s}_{i_{l^*}}^{(-l)}(t_l) \right]^2 \tag{22}$$

where $y_{i_{l^*}}(t_{l^*})$ is the value of the response variable for subject i_{l^*} at time t_{l^*}, $w_l = (Mm_l)^{-1}$ is the weight associated with the l-th measuring time and $\widehat{s}_{i_{l^*}}^{(-l)}(t_l)$ denotes the estimation of the response variable for experimental unit i_{l^*} at time t_l when all the observations related with the response variable at time t_l have been excluded.

5. Simulation

This section presents a simulation study to evaluate the performance of the estimation methods. The comparison is performed through the Average Mean Square Error (AMSE) given by

$$AMSE(\kappa) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\kappa(t_{ij}) - \widehat{\kappa}(t_{ij}) \right]^2 \tag{23}$$

with $\kappa(\cdot)$ a function that corresponds to any dynamic coefficient of model (3).

Simulation strategy is similar to that followed by Wu & Liang (2004). The simulation model is structured as follows:

$$\begin{aligned} y_i(t) &= \beta_0(t) + x_{i1}(t)\left[\beta_1(t) + v_{1i}(t)\right] + \epsilon_i(t),\ i = 1, \ldots, n \\ x_{i1}(t) &= 1 - \exp\left[-0.5t - (i/n)\right] \\ \beta_0(t) &= 3\exp(t),\ \beta_1(t) = 1 + \cos(2\pi t) + \sin(2\pi t) \\ v_{1i}(t) &= a_{i0} + a_{i1}\cos(2\pi t) + a_{i2}\sin(2\pi t) \\ a_i &= [a_{i0}, a_{i1}, a_{i2}]^T \sim N\left([0,0,0]^T, diag[\sigma_0^2, \sigma_1^2, \sigma_2^2]\right) \\ \epsilon_i(t) &\sim N(0, \sigma_\epsilon^2 x_{i1}^2(t)) \end{aligned} \tag{24}$$

where $\beta_0(t)$, $\beta_1(t)$ and $v_{1i}(t)$, are the dynamic parameters of the model, $x_i(t)$ is the covariate of the model associated with $\beta_1(t)$ and where $v_{1i}(t)$ and $\epsilon_i(t)$ are random errors. This model corresponds to the RVCM given in (3) where

$$\boldsymbol{\beta}(t) = [\beta_0(t), \beta_1(t)]^T,\ \boldsymbol{v}_i(t) = [v_{1i}(t)],\ \boldsymbol{x}_i(t) = [x_{i0}(t), x_{i1}(t)]^T,\ \boldsymbol{z}_i(t) = [z_{i1}(t)]$$

with $x_{i0}(t) \equiv 1$ and $z_{i1}(t) \equiv x_{i1}(t)$. Note that in the simulated model $\mathbf{R}_i$ is a diagonal matrix and $\mathbf{D}$ is an unstructured covariance matrix. The observations between subjects are simulated independent.

It is assumed that $\sigma_1^2 = \sigma_2^2 = \sigma_\epsilon^2 = \sigma^2$. Then, the correlation coefficient between repeated measurements within each experimental unit is

$$\rho = \mathbb{C}orr[y_i(t), y_i(s))] = \frac{\sigma_0^2 + \sigma^2\cos[2\pi(t-s)]}{\sigma_0^2 + 2\sigma^2},\ \text{for } s \neq t$$

therefore

$$\frac{\sigma_0^2 - \sigma^2}{\sigma_0^2 + 2\sigma^2} \leq \rho_y \leq \frac{\sigma_0^2 + \sigma^2}{\sigma_0^2 + 2\sigma^2}$$

To simulate different intensities of correlation are considered three cases:

Case 1 In which $\sigma_1^2 = \sigma_2^2 = \sigma_\epsilon^2 = \sigma^2 = 0.01$ and $\sigma_0^2 = 0.01$. This corresponds to $\rho_y \in (\ 0\ ,\ 0.67\ )$.

Case 2 In which $\sigma_1^2 = \sigma_2^2 = \sigma_\epsilon^2 = \sigma^2 = 0.01$ and $\sigma_0^2 = 0.04$. This corresponds to $\rho_y \in (\ 0.50\ ,\ 0.83\ )$.

Case 3 In which $\sigma_1^2 = \sigma_2^2 = \sigma_\epsilon^2 = \sigma^2 = 0.01$ and $\sigma_0^2 = 0.09$. This corresponds to $\rho_y \in (\ 0.73\ ,\ 0.91\ )$.

Design times are simulated in accordance with the expression

$$t_{ij} = j/(m+1),\ i = 1, \ldots, n,\ j = 1, \ldots, m$$

where m is a positive integer. To simulate unbalanced data sets, a main characteristic of the structure of longitudinal data, in each subject are removed randomly repeated measures with a rate $r_m = 30\%$. Thus, there is approximately $m(1 - r_m)$ repeated measurements per experimental unit and $nm(1 - r_m)$ measurements in

Table 1: Scenarios for the simulation study.

Scenario	n	m	σ_0^2	Scenario	n	m	σ_0^2	Scenario	n	m	σ_0^2
1	5	5	0.01	10	10	5	0.01	19	20	5	0.01
2	5	5	0.04	11	10	5	0.04	20	20	5	0.04
3	5	5	0.09	12	10	5	0.09	21	20	5	0.09
4	5	10	0.01	13	10	10	0.01	22	20	10	0.01
5	5	10	0.04	14	10	10	0.04	23	20	10	0.04
6	5	10	0.09	15	10	10	0.09	24	20	10	0.09
7	5	15	0.01	16	10	15	0.01	25	20	15	0.01
8	5	15	0.04	17	10	15	0.04	26	20	15	0.04
9	5	15	0.09	18	10	15	0.09	27	20	15	0.09

total. Smoothing parameters are chosen by using PCV. Table 1 contains all the scenarios considered in the simulation study.

Each scenario was repeated $N = 500$ times and each time was calculated $AMSE(\beta_0)$ and $AMSE(\beta_1)$, in order to compare the relative performance of the Local Polynomial Kernel Regression Estimation (LPKE) with Radial Kernel Functions Estimation (RKFE). For these estimations the next indicators are define

$$AMSER_{(RKFE/LPKE)} = \frac{1}{N}\sum_{k=1}^{N}\frac{AMSE_k(\kappa, LPKE)}{AMSE_k(\kappa, RKFE)} \times 100\% \tag{25}$$

and

$$AMSERKF_{(RKFE/LPKE)} = \frac{\sum_{k=1}^{N} I_{\{AMSE_k(\kappa,LPKE)>AMSE_k(\kappa,RKFE)\}}}{N} \times 100\% \tag{26}$$

where $AMSE_k(\kappa, LPKE)$ and $AMSE_k(\kappa, RKFE)$ denote the value of $AMSE(\kappa)$ obtained in the k-th simulation replicate, $k = 1, \ldots, N$, by using the RKFE and the LPKE respectively, and I_A denotes the indicator function of set A. $AMSER$ represents the average relative efficiency associated with the N replications and $AMSERKF$ is the percentage of estimations obtained through RKF that are better than those obtained through LPK in relation to the AMSE in AMSE in the N replications. If $AMSER \approx 100\%$ and $AMSERKF \approx 50\%$, LPKE and the RKFE perform similarly; if $AMSER > 100\%$ y $AMSERKF > 50\%$, RKFE has better performance than LPKE; and if $AMSER < 100\%$ and $AMSERKF < 50\%$, LPKE has better performance than RKFE.

Table 2 contains the results of the simulation. According to this table, the choice rules of an alternative estimation by using indicators (25) and (26), and Tables 3 and 4 which summarizes the results, it follows that at 48% of cases the best estimation strategy is the RKFE; by approximation to the rules given, that is, following the criteria $AMSER_0 \approx 100\%$ y $AMSERKF_0 \approx 50\%$, it has that in the 35.2% of the situations the two strategies behave similarly; furthermore, just 9.3% of cases the best strategy is LPKE and for 7.4% of the scenarios the criterion does not decide ($AMSER > 100\%$ and $AMSERKF < 50\%$, or, $AMSER < 100\%$

TABLE 2: Simulation results.

		$n=5$			$n=10$			$n=20$		
		$m=5$	$m=10$	$m=15$	$m=5$	$m=10$	$m=15$	$m=5$	$m=10$	$m=15$
$\sigma_0^2=0.01$	$AMSER_0$	100.0%	100.2%	101.1%	100.9%	100.0%	100.0%	101.0%	102.0%	100.8%
	$AMSERKF_0$	49.9%	50.2%	50.8%	50.8%	49.2%	48.4%	50.9%	46.0%	62.0%
	$AMSER_1$	100.5%	102.4%	101.4%	100.0%	101.5%	100.9%	99.6%	100.1%	100.4%
	$AMSERKF_1$	45.8%	51.2%	50.8%	48.1%	51.0%	50.4%	43.8%	61.9%	56.4%
$\sigma_0^2=0.04$	$AMSER_0$	100.1%	100.0%	100.0%	100.1%	101.0%	100.0%	100.0%	102.1%	100.5%
	$AMSERKF_0$	50.7%	49.4%	45.8%	47.8%	45.9%	49.9%	49.0%	63.1%	51.8%
	$AMSER_1$	99.6%	102.2%	100.7%	100.0%	100.8%	100.5%	100.7%	101.3%	100.4%
	$AMSERKF_1$	42.0%	52.7 %	48.1%	46.9%	50.1%	54.9%	50.2%	49.9%	52.8%
$\sigma_0^2=0.09$	$AMSER_0$	100.9%	100.5%	100.1%	100.0%	100.0%	100.3%	100.0%	100.7%	99.9%
	$AMSERKF_0$	50.5%	50.4%	51.1%	47.7%	49.7 %	51.0%	47.7%	48.6%	46.8%
	$AMSER_1$	99.3%	101.8%	101.6%	100.5%	101.3%	100.9%	99.5 %	102.0%	100.2 %
	$AMSERKF_1$	44.5%	49.0%	49.3%	49.6%	43.6%	51.6%	46.7%	52.0%	51.4%

and $AMSERKF > 50\%$). It is also noted that the strategy most appropriate for estimating, considering $\beta_0(t)$ and $\beta_1(t)$ simultaneously, is type $RKFE$ which corresponds to $n=5$, $m=10$ and $\sigma_0^2=0.01$, $n=5$, $m=15$ and $\sigma_0^2=0.01$, $n=10$, $m=15$ and $\sigma_0^2=0.09$, $n=10$, $m=15$ and $\sigma_0^2=0.01$, and $n=10$, $m=15$ and $\sigma_0^2=0.04$; there is no case where LPKE improved the outcomes for both dynamic components simultaneously. Furthermore, there are a variety of cases where the best strategy is RKFE for one of the dynamic parameters and for the other two strategies perform similarly.

According to Table 3, it is concluded that while the value of σ_0^2 decreases, and at the same time the correlation between repeated measurements, the proportion of times that the best strategy is RKFE increase. Moreover, in all degrees of correlation, the proportion of times that performs better RKFE is superior compared to the proportion for LPKE. In the same way, in all degrees of correlation, the proportion of times where the two strategies perform similarly is higher than the proportion where LPKE is the best option. Also, these relationships are maintained in each case for the dynamic intercept and the dynamic slope. Therefore, with any degree correlation and any dynamic parameter, in 83.3% of cases, RKFE performs better or similarly than LPKE. Thus, it is concluded that in such circumstances, to choose RKFE is the best alternative.

TABLE 3: Proportion of times that a strategy is better than another for σ_0^2 and $\beta_r(t)$.

		RKF	Equal	LPK	No
$\sigma_0^2=0.01$	$\beta_0(t)$	9.3%	5.6%	0.0%	1.9%
	$\beta_1(t)$	11.1%	1.9%	1.9%	1.9%
Total		20.4%	7.4%	1.9%	3.7%
$\sigma_0^2=0.04$	$\beta_0(t)$	5.6%	9.3%	0.0%	1.9%
	$\beta_1(t)$	9.3%	5.6%	1.9%	0.0%
Total		14.8%	14.8%	1.9%	1.9%
$\sigma_0^2=0.09$	$\beta_0(t)$	7.4%	7.4%	1.9%	0.0%
	$\beta_1(t)$	5.6%	5.6%	3.7%	1.9%
Total		13.0%	13.0%	5.6%	1.9%
Total general		48.1%	35.2%	9.3%	7.4%

Table 4: Proportion of times that a strategy is better than another for n and m.

		RKF	Equal	LPK	No
	$m = 5$	3.7%	1.9%	3.7%	1.9%
$n = 5$	$m = 10$	7.4%	3.7%	0.0%	0.0%
	$m = 15$	5.6%	5.6%	0.0%	0.0%
Total		16.7%	11.1%	3.7%	1.9%
	$m = 5$	1.9%	9.3%	0.0%	0.0%
$n = 10$	$m = 10$	3.7%	3.7%	0.0%	3.7%
	$m = 15$	7.4%	3.7%	0.0%	0.0%
Total		13.0%	16.7%	0.0%	3.7%
	$m = 5$	3.7%	3.7%	3.7%	0.0%
$n = 20$	$m = 10$	5.6%	3.7%	0.0%	1.9%
	$m = 15$	9.3%	0.0%	1.9%	0.0%
Total		18.5%	7.4%	5.6%	1.9%
Total		48.1%	35.2%	9.3%	7.4%

Furthermore, according to Table 4, for all sample sizes, when the number of repeated measurements of each individual increases, the proportion of scenarios where RKFE performs better RKFE increases as well. It must also be noted that this proportion is similar for all sample sizes, and is always significantly higher than the proportion where LPKE is the best option. Moreover, when $n = 10$ is notorious the proportion of times where the two strategies perform similarly. Finally, it is observed the fact that the proportion of times where LPKE is better is equal to 0.0% in most cases for any value of n y m. Thus, it is concluded that the proposed methodology is the best option a high percentage of times in all simulated scenarios, or at least performs similarly to the LPKE, which very rarely performs better than the RKFE.

6. Application

The viral load (*plasma VIH RNA copies/mL*) and cell count CD4+ are currently key indicators to assess AIDS treatments in clinical research. Initially it was considered the CD4+ cell count as a primary indicator of AIDS immunodeficiency, but it was newly found that viral load is more predictive for clinical outcomes. However, recently some researchers have suggested that a combination of these two indicators may be more appropriate to evaluate the treatment of HIV and AIDS. Therefore it is pertinent to study the relationship between viral load and CD4+ cell count during treatment (Liang et al. 2003).

Figure 2 presents some graphs of a linear regression of viral load (log(RNA)) against to CD4+ cell counts in some measuring times of a clinical study of AIDS (ACTG 315). In this investigation, there are 46 infected patients with an antiviral therapy consisting of *ritonavir*, 3TC and AZT. After starting treatment, viral load and CD4+ cell count were observed simultaneously at days 0, 2, 7, 10, 14, 28, 56, 84, 168, and 336. The number of repeated measurements for individual varies from 4 to 10 and in total 361 observations were obtained.

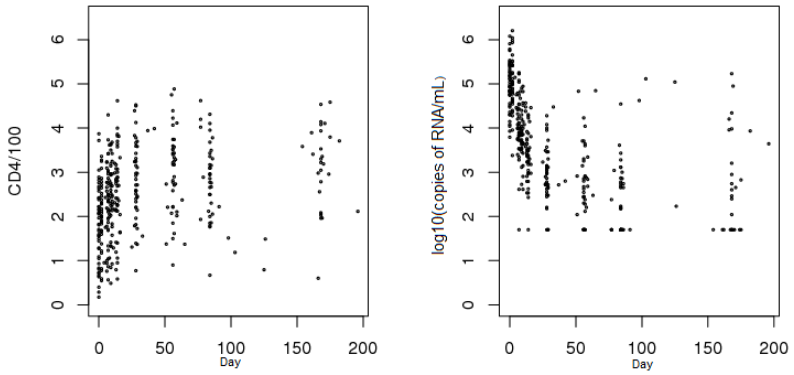

FIGURE 1: Scatter plot for CD4+ cell count and viral load.

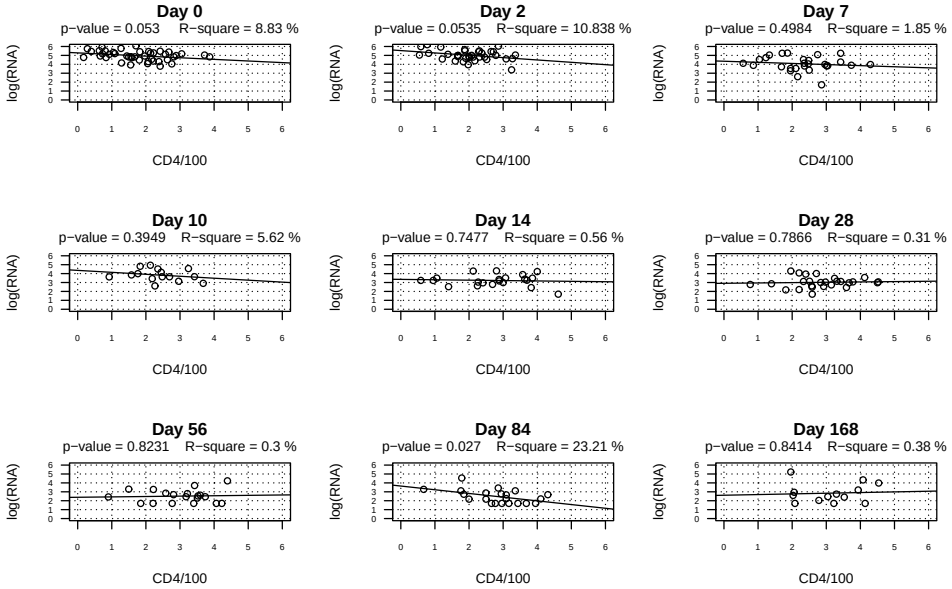

FIGURE 2: Graphs related with the linear regression of viral load ($\log_{10}$(RNA)) against CD4+ cell count in some measuring times. The model adjusted in each case has got the form $\log_{10}(RNA) = \beta_0 + \beta_1(\text{CD4}/100) + e$. The p-value corresponding to $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is also presented in each case.

In general, it appears that the virologic (measured by the viral load) and the immune response (measured by the CD4+ cell count) of the patient are negatively correlated, and that their relationship is approximately linear during antiviral therapy. Figure 1 shows the scatter plots associated with CD4+ cell count and viral load. The logarithm of viral load is used to stabilize the variance for the estimation procedures of the model fitted in the following.

Figure 2 shows that the slope of the linear regression of viral load versus CD4+ cell count changes over time because in a few days the slope is significantly different from zero and in others not. This motivates the fitting of a model with dynamic coefficients in order to describe and quantify the change in the relationship. However, because it may be of interest to investigate the relationship between viral load and CD4+ cell count in a particular patient, the fitting of a RVCM is needed.

The ACTG 315 data set has been studied extensively by Liang et al. (2003), who showed a strong inverse relationship between viral load and CD4+ cell count. In this section, a RVCM is fitted to investigate the dynamic relationship between viral load (in logarithmic scale) and CD4+ cell count, and also to describe this relationship particularly in any patient.

The RVCM fitted is

$$y_{ij} = \beta_0(t_{ij}) + \beta_{1i}(t_{ij})x_{i1}(t_{ij}) + e_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, 46 \qquad (27)$$

where y_{ij}, $x_{i1}(t_{ij})$, and e_{ij} are viral load (in logarithmic scale), the CD4+ cell count and the error associated with the j-th measurement of the i-th patient, respectively, $\beta_0(t)$ is the dynamic coefficient associated with the intercept and $\beta_{1i}(t)$ is the dynamic and random coefficient associated with the CD4+ cell count. This parameter is given by

$$\beta_{1i}(t) = \beta_1(t) + v_i(t),\ i = 1, \ldots, 46$$

with $\beta_1(t)$ the coefficient associated with the mean dynamic relationship between viral load and cell count CD4+ and $v_i(t)$ the coefficient related to the characteristics of the i-th patient that differ from the average behavior.

The dynamic components of the model are estimated through LPK and the proposed methodology by using RKF. The kernel functions used in the estimation are Gaussian, and for selecting the smoothing parameters (bandwidths) the PCV is implemented which gives the bandwidths $h_{RKF} = 0.999$ and $h_{RKF} = 0.401$ using RKF and LPK respectively (Figure 3). Furthermore, models (8) and (20) are fitted by using function `lme4` (Bates, Maechler & Bolker 2011) in `R` (R Development Core Team 2008).

Figure 4 shows the residuals of the RVCM fitted. It is observed that in both cases, the RVCM has a good fit to the data. The value of the residuals by using both estimation methods are similar prior 150-th day. From that day the value of the residuals is less by using LPK, suggesting that the relationship at the end of the treatment by using LPK is more accurate; however, both techniques indicate the same at the end of treatment as it is evidenced in Figure 5 where are illustrated the graphs associated with the estimation of $\beta_0(t)$ and of $\beta_1(t)$ by using LPK and RKF, respectively. In both cases, the graphics are very similar to those obtained by Liang et al. (2003).The right chart shows that the dynamic relationship between viral load and CD4+ cell count is approximately direct to day 50, point at which the association is weak; from this day the relationship between the indicators is inverse to the end of treatment. Moreover, between week 1 and 14, RKF estimate suggests that the relationship is apparently stronger. Also, major differences between the estimation methodologies from day 150 of treatment are noted, where the estimate

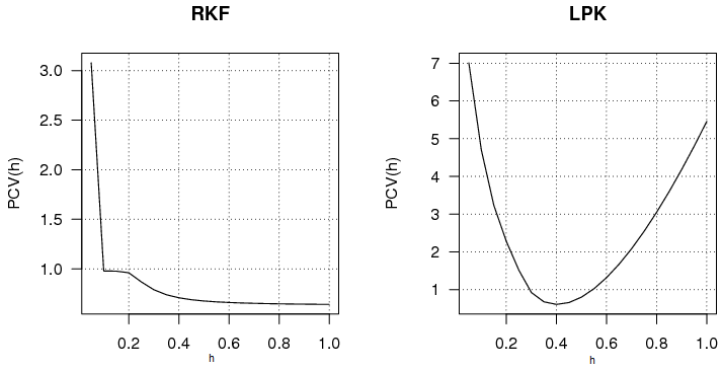

FIGURE 3: Graphs related with PCV against the bandwidth.

through LPK suggests that the relationship changes and it is strengthened –in an inverse way– to the end of treatment. Overall, the dynamic relationship between viral load and CD4+ cell count decreases gradually until the seventh week of the study where the relationship begins to strengthen gradually until the end of treatment.

One advantage of fitting a RVCM is that it is possible to characterize the performance of the dynamic relationship of interest for any particular subject. Figure 6 shows the estimates of the deviations typical of the population $v_i(t)$ for patients 1, 3 y 16 using RKF and LPK. Not only the magnitude but also the direction of changes can be see among individuals. Due to the high variation within each of the individuals, the estimation of the relationship between the indicators for each patient is very important because it allows to customize the treatment and care of each patient. Using LPK more variability between individuals in the dynamic relationship of viral load and CD4+ cell count is perceived. It is observed how the relationship may even be direct. While using RKF variability is lower and the pattern is very similar to the average dynamic relationship.

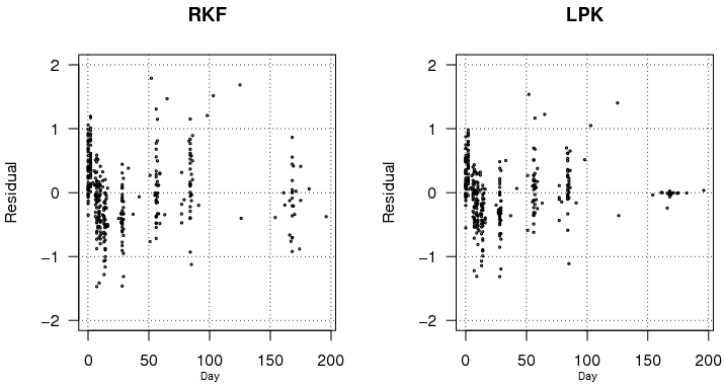

FIGURE 4: Residuals of RVCM fitted by using RKF and LPK.

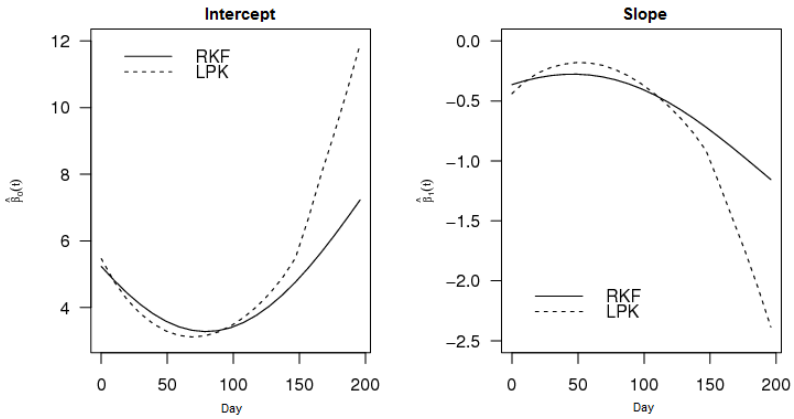

FIGURE 5: Graphs associated with the estimation of $\beta_0(t)$ and $\beta_1(t)$ for the RVCM fitted by using RKF and LPK.

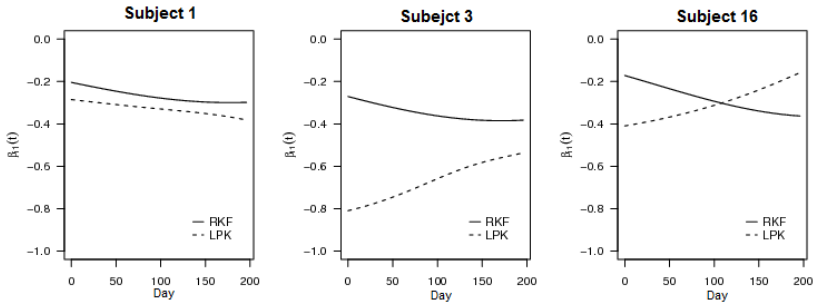

FIGURE 6: Graphs associated with the estimation of $v_i(t)$ for the RVCM fitted by using RKF and LPK for patients 1, 3 and 16.

7. Discussion and Conclusions

This paper proposes a methodology to estimate the coefficients of a random time-varying coefficient model through radial kernel functions, where model coefficients are approximated by a linear combination of kernel functions which centered around all the measuring points, or their quantiles, weighted by a bandwidth that may change or not among coefficients (Hastie, Tibshirani & Friedman 1990).

By means of a simulation study the estimation method is compared by using a local polynomial kernel regression with the use of radial kernel functions in relation with the average mean square error, resulting that the proposed methodology is the best one in a high percentage of times in all simulated scenarios, or at least performs similarly to the LPKE, who rarely performs better than the RKFE, in relation with the average mean square error.

Analyzing the ACTG 315 data set (Liang et al. 2003), it was found that the relationship between viral load and CD4+ cell count is inverse. Furthermore, as a

future alternative modeling, it can be thought a model in which the response variable is bivariate, consisting of viral load and CD4+ cell count, and the predicted correspond to some covariates related to the treatment of patients with AIDS.

Further studies may investigate the consistency and asymptotic properties of the estimators proposed, the impact of the functional form of the dynamic coefficients of the model and mechanisms for testing hypotheses related to both the dynamic and random coefficients model.

[Recibido: abril de 2010 — Aceptado: febrero de 2012]

References

Altman, N. S. (1990), 'Kernel smoothing of data with correlated errors', *Journal of the American Statistical Association* **85**(411), 749–759.

Bates, D., Maechler, M. & Bolker, B. (2011), *lme4: Linear Mixed-Effects Models Using S4 classes*. R package version 0.999375-42.
*http://CRAN.R-project.org/package=lme4

Davis, C. S. (2000), *Statistical Methods for the Analysis of Repeated Measurements*, Springer.

Diggle, P. J., Liang, K. Y. & Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford University Press.

Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (2009), *Longitudinal Data Analysis*, Champman & Hall.

Guo, W. (2002), 'Functional mixed effects models', *Biometrics* **58**(1), 121–128.

Hart, J. D. (1991), 'Kernel regression estimation with time series errors', *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(1), 173–187.

Hart, J. D. & Wehrly, T. E. (1986), 'Kernel regression estimation using repeated measurements data', *Journal of the American Statistical Association* **81**(396), 1080–1088.

Hastie, T., Tibshirani, R. & Friedman, J. (1990), *The elements of Statistical Learning*, Springer.

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L. P. (1998), 'Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data', *Biometrika* **85**(4), 809–822.

Huang, J. Z., Wu, C. O. & Zhou, L. (2002), 'Varying-coefficient models and basis function approximations for the analysis of repeated measurements', *Biometrika* **89**(1), 111–128.

Liang, H., Wu, H. & Carroll, R. J. (2003), 'The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error', *Biostatistics* **4**(2), 297–312.

R Development Core Team (2008), *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

Rice, J. A. & Silverman, B. W. (1991), 'Estimating the mean and covariance structure nonparametrically when the data are curves', *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(1), 233–243.

Sosa, J. C. & Díaz, L. G. (2010), 'Estimación de las componentes de un modelo de coeficientes dinámicos mediante las ecuaciones de estimación generalizadas', *Revista Colombiana de Estadística* **33**(1), 89–109.

Verbeke, G. & Molenberghs, G. (2005), *Models for Discrete Longitudinal Data*, Springer.

Wu, H. & Liang, H. (2004), 'Backing random varying-coefficient models with time-dependent smoothing covariates', *Scandinavian Journal of Statistics* **31**, 3–19.

Wu, H. & Zhang, J. T. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis*, Wiley.

Zeger, S. L. & Diggle, P. J. (1994), 'Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters', *Biometrics* **50**(3), 689–699.

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 personas, que no integran los Comités Editorial y Científico, por su colaboración en el volumen 34 (2011).

Aluisio de Souza Pinheiro, Ph.D.
Ahmad Parsian, Ph.D.
Aureli Alabert, Ph.D.
Andreas Wienke, Ph.D.
Anna Castañer Garriga, Ph.D.
Aroldo Perez, Ph.D.
Arjun Kumar Gupta, Ph.D.
Arturo Ruiz-Falco, Ph.D.
Boyan Dimitrov, Ph.D.
Carmen Elisa Flórez, Ph.D.
César Sánchez Sellero, Ph.D.
Cristian Coletti, Ph.D.
Chris Jones, Ph.D.
Daniela Dunkler, Ph.D.
Donald Richards, Ph.D.
Edgar Brunner, Ph.D.
Eliana González, M.Sc.
Ernesto Juan Darias, Ph.D.
Ernesto Ponsot Balaguer, Ph.D.
Francisco Lopez Herrera, Ph.D.
Gadde Srinivasa Rao, Ph.D.
Hans van Houwelingen, Ph.D.
Humberto Gutiérrez, Ph.D.
Humberto Jesus Llinas, Ph.D.
Ivan A. Carrillo, Ph.D.
Jon Kettenring, Ph.D.
Jaime Castillo, M.Sc.
Jaime Londoño, Ph.D.
James Wilson, Ph.D.
José Carmona, Ph.D.
José Bermúdez, Ph.D.
Jose Fabian Gonzalez Flores, Ph.D.
Joseph Kadane, Ph.D.
Jorge Achcar, Ph.D.
Jong-Min Kim, Ph.D.
Juvencio Santos, Ph.D.
Kumaraswamy Ponnambalam, Ph.D.
Lawrence M. Leemis, Ph.D.
Lelys Bravo, Ph.D.
Ma. Purificación Galindo, Ph.D.
Mariano Ruiz Espejo, Ph.D.
Mauricio Gutiérrez Urzua, Ph.D.
Manuel Salvador, Ph.D.
Martin Egozcue, Ph.D.
Michael Doherty, Ph.D.
Michael Schemper, Ph.D.
Michael Wolf, Ph.D.
Michelli Barros, Ph.D.
Nandini Dendukuri, Ph.D.
Nick Horton, Ph.D.
Paolo Rosso, Ph.D.
Pablo Olivares Rieumont, Ph.D.
Pranab Sen, Ph.D.
Raúl Fernández, Ph.D.
Reinaldo Arellano Valle, Ph.D.
Rezaul Karim, Ph.D.
Rocío Ribero, Ph.D.
Rohana Ambagaspitiya, Ph.D.
Ron S. Kenett, Ph.D.
Sarjinder Singh, Ph.D.
Sergio Perez Elizalde, Ph.D.
Shuo Jie Wu, Ph.D.
Tarciana Liberal Pereira, Ph.D.
Thierry Duchesne, Ph.D.
Valderio A. Reisen, Ph.D.
Víctor Leiva, Ph.D.
Vladimir Moreno, M.Sc.