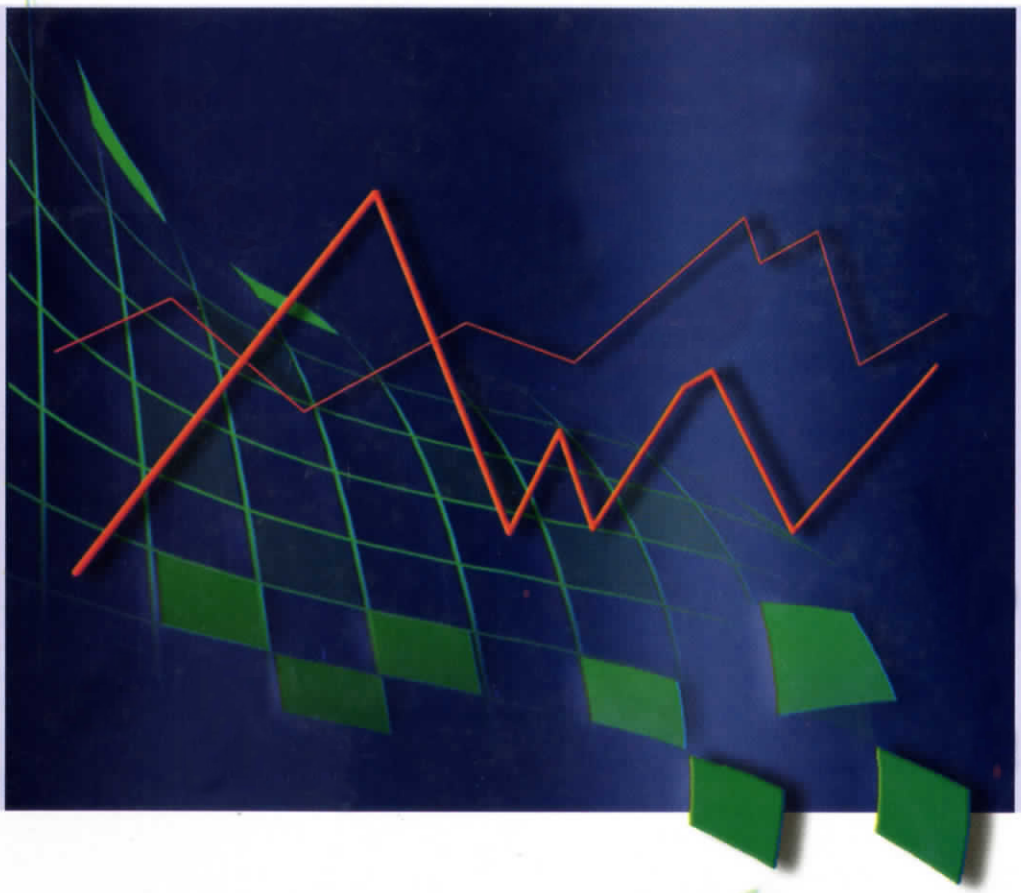
Memorias

del XXI Foro Nacional de Estadística



INEGI

INSTITUTO NACIONAL DE ESTADÍSTICA GEOGRAFÍA E INFORMÁTICA

AME

Memorias

del XXI Foro Nacional de Estadística

INEGI

INSTITUTO NACIONAL DE ESTADÍSTICA GEOGRAFÍA E INFORMÁTICA

AME

Edificio Sede
Av. Héroe de Nacozari Sur Núm. 2301
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.

www.inegi.gob.mx
atencion.usuarios@inegi.gob.mx

Memorias del XXI Foro Nacional de Estadística

Impreso en México

Presentación

El XXI Foro Nacional de Estadística se llevó a cabo en Acapulco Guerrero, siendo sede la Universidad Autónoma de Guerrero, del 11 al 13 de octubre de 2006.

En estas memorias se presentan resúmenes de algunas contribuciones libres presentadas en este evento. Los resúmenes incluidos en estas memorias, se revisaron con detalle, pero sin considerarse que fueron sometidos a un proceso de arbitraje.

La Asociación Mexicana de Estadística agradece a la Universidad Autónoma de Guerrero por el apoyo para la realización de este foro y al Instituto Nacional de Estadística, Geografía e Informática el apoyo para la edición de estas memorias.

El Comité Editorial:

J. Armando Domínguez Molina

Antonio V. González Fragoso

Jorge H. Sierra Cavazos

Contenido

Examining distributional characteristics of clusters

Alexander von Eye
Michigan State University
Patrick Mair
Wirtschaftuniversität Wien

1. Clustering and Data Generation Processes

Standard methods of cluster analysis, for example, Ward's method or complete linkage, create clusters without reference to the characteristics of the distribution the data were drawn from. Instead, the methods form clusters using criteria such as the one that minimizes the distance within a cluster while maximizing the distance between clusters. Based on this and other criteria, clusters result that reflect density centers in the data space. This strategy practically always yields interpretable clusters. However, this strategy cannot answer the question whether the thus identified density centers still qualify as such when the *Data Generation Process* (DGP) is taken into account that underlies the distribution of the data. In this contribution, we propose examining clusters from standard cluster analysis from a statistical perspective. Specifically, we propose estimating the probability of belonging to a particular cluster and comparing the resulting expected frequency with the observed number of cluster members.

There has been a number of attempts to evaluate cluster solutions from the perspective of distributional assumptions. Three data generation processes have been discussed in the literature (for an overview see Everitt, Landau, & Leese, 2001), the *random dissimilarity model*, the *Poisson model*, and the *unimodal model*.

The *random dissimilarity model* (cf. the random graph hypothesis; Jain & Dubes, 1988) states in its null hypothesis that all permutations of the ranks of the (dis)similarities of all pairs of cases are equally likely. Departures from this assumption are compatible with the hypothesis of clustering. This null hypothesis has been criticized because it creates an unrealistic distribution of test statistics.

The *Poisson model* assumes that the p-variate observations of the n cases in a sample are part of a uniform distribution over some region A of the p-space. If this assumption applies,

1. The underlying distribution has no mode;
2. The number of cases in each subregion, A_s, is a random number;
3. This number follows a Poisson distribution;
4. The numbers of non-overlapping subregions are independent; and
5. The number of cases within A_s is $\lambda\,|A_s|$ where λ is the constant intensity given by the mean of the Poisson distribution and $|A_s|$ is the volume of the subregion A_s (area in 2D).

For the evaluation of existing clusters, one assumes that λ is constant across all subregions of the p-space. That is, one assumes a *homogeneous Poisson process*. The subregions are defined by the clusters.

The *unimodal model* is based on a DGP that yields a frequency distribution with one mode, for example, the binomial or the normal distributions. The null hypothesis under this model states that the subregions (clusters) do not contain different numbers of cases than expected based on the assumption that the underlying distribution has one mode. In the present work, we assume that this is the mode of the multinormal distribution.

1.1. The Shape of Clusters

To be able to estimate the probability of belonging to a particular cluster, we first circumscribe the cluster by a convex hull. Specifically, we use spheroids and ellipsoids to circumscribe the subregion that is defined by a cluster. The hull of an ellipsoid or spheroid is, in p-space,

$$\boldsymbol{x}_d^T \boldsymbol{R}^T \boldsymbol{V} \boldsymbol{R} \boldsymbol{x}_d = 1,$$

where $\boldsymbol{x}_d$ is the $p \times 1$ vector of the differences of a point on the hull from the centroid of the hull, $\boldsymbol{R}$ is the $p \times p$ matrix of the orientation of the ellipse, and $\boldsymbol{V}$ is the inverse of the $p \times p$

matrix that contains the squared lengths of the semi-axes of the ellipsoid in its diagonal. If the semi-axes are equal in length, the hull circumscribes a spheroid, otherwise an ellipsoid.

To create the spheroids, we calculate the distance between the two farthest-apart points of a cluster. The midpoint of this distance is the centroid of the spheroid. The distance is the radius of the spheroid.

To determine the corresponding ellipsoid, we shrink radii as long as data points are still within or on the convex hull that circumscribes the subregion (cf. Löwner ellipsoids; Kumar & Yildirim, 2005). A data point X is located

- inside the convex hull, if $\left|\boldsymbol{x}_d^T \boldsymbol{R}^T \boldsymbol{V} \boldsymbol{R} \boldsymbol{x}_d\right| < 1$,
- on the hull of the convex hull, if $\left|\boldsymbol{x}_d^T \boldsymbol{R}^T \boldsymbol{V} \boldsymbol{R} \boldsymbol{x}_d\right| = 1$, and
- outside the convex hull, if $\left|\boldsymbol{x}_d^T \boldsymbol{R}^T \boldsymbol{V} \boldsymbol{R} \boldsymbol{x}_d\right| > 1$.

1.2. Estimating the Probability of Belonging to a Cluster

Let the volume of a convex hull in p-space be ν_A and the volume of the total data body ν_T. Then, under a homogeneous Poisson process, the probability of ν_A is estimated as ν_A/ν_T. To estimate the probability of ν_A under the multinormal model, we use the method proposed by Sommerville (1998a, 1998b, 2001). This method estimates the point probability for a prespecified number of random points inside ν_A. This number typically is selected to be as large as 10,000. The probability of ν_A is then the average of these point probabilities.

1.3. The Four Steps of Testing for Absence of Structure

1. *Clustering cases*: Clustering methods are selected based on the decisions discussed by von Eye and Mun (2004). In addition, clusters must be compact (convex).
2. *Circumscribing clusters*: Löwner ellipsoids (1) minimize the volume of the subregion that is constituted by a cluster, (2) minimize overlap between circumscribing hulls, and

(3) reflect correlations among variables.

3. *Determining the expected number of cases*: The determination of the expected number of cases is specific to the DGP and the shape of a cluster.

4. *Testing against lack of cluster structure*: If a test such as the binomial test suggests significant deviations from expectancy, a cluster structure may exist.

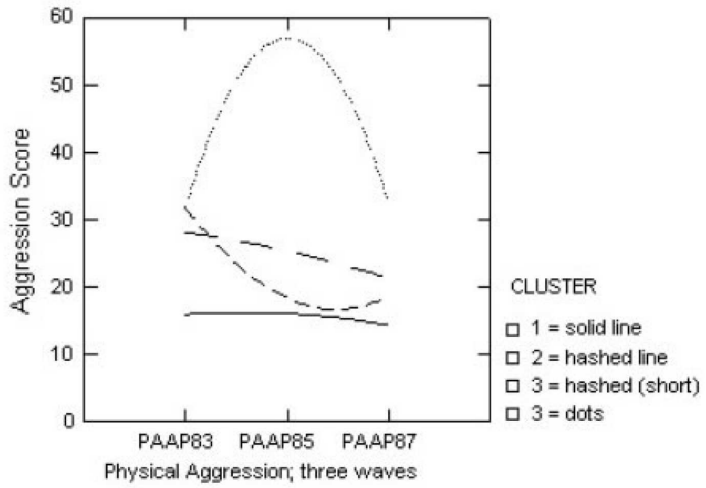


Figure 1. Developmental trajectories of physical aggression against peers (PAAP) in four clusters.

2. Data Example

The data analyzed in the following example were collected in 1985, in a study by Finkelstein, von Eye, and Preece (1994) on the development of aggression in adolescence. 1985 was the second of three data waves (the other data were collected in 1983 and 1987). In 1985, the adolescents were, on average, 13 years of age. 114 participants responded to the questionnaire, 46 of whom were boys. For the following example, we use the variable Physical Aggression against Peers which was observed at all three observation points (PAAP83, PAAP85, and PAAP87).

In Step 1, trajectory clusters were created using complete linkage. The intercluster distance diagram suggested that 4 clusters may exist, one of them (Cluster 4) being an isolate. Figure 1 displays the trajectories, by cluster.

Table 1 displays the results of the tests against lack of cluster structure.

The results in Table 1 show that Clusters 1 and 3 contain more cases than expected under either DGP and cluster shape. In contrast, Cluster 2 contains more cases than expected under the Poisson DGP when either shape of hull is used to circumscribe the subregion that is defined by this cluster, and it contains *fewer* cases than expected when the multinormality DGP is used. For the isolate, the test is not applicable.

Cluster				
	1	**2**	**3**	**4**
Size(N)	68	32	13	1
Poisson Model - spheroids				
Area	6406.79	7561.03	4902.87	
P	< ,000001	< ,000001	,002891	
$e > N$?	m	m	m	
Poisson Model - ellipsoids				
Area	4949.57	7561.03	4902.87	
P	< ,000001	< ,000001	,002891	
$e > N$?	m	m	m	
Multinormality Model - spheroids				
P	< ,000001	.023235	.000081	
$e > N$?	m	f	m	
Multinormality Model - ellipsoids				
P	< ,000001	.023235	.000081	
$e > N$?	m	f	m	

Table 1. Testing Hypotheses of Lack of Structure for the Cluster Solution in Figure 1.

3. Discussion

The method proposed here is neither a hybrid clustering method (e.g., Kwon & Han, 2002) nor a probabilistic clustering method (e.g., Raftery & Dean, 2006). Instead it is a method

for the evaluation of existing cluster solutions. For proper application of this method, it is of no importance how the clusters were created. The only condition is that they be compact (convex).

4. References

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4 ed.). London: Arnold.

Finkelstein, J. W., von Eye, A., & Preece, M. A. (1994). The relationship between aggressive behavior and puberty in normal adolescents: A longitudinal study. *Journal of Adolescent Health,* 15, 319-326.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data.* Englewood Cliffs, NJ: Prentice Hall.

Kumar, P., & Yildirim, E. A. (2005). Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and Applications*, 126, 1-12.

Kwon, S., & Han, C. (2002). Hybrid clustering method for DNA microarray data analysis. *Genome Informatics*, 13, 258-259.

Raftery, A. E., & Dean, N. (2006). Variable selection for model based clustering. *Journal of the American Statistical Association*, 101, 168-178.

Sommerville, P. N. (1998a). Numerical computation of multivariate normal and multivariate-t over convex regions. *Journal of Computational and Graphical Statistics*, 7, 529-544.

Sommerville, P. N. (1998b). A FORTRAN 90 program to evaluate multivariate normal and multivariate-t integrals over convex regions. *Journal of Statistical Software*, 3(4).

Sommerville, P. N. (2001). Numerical computation of multivariate normal and mutivariate-t probabilities over ellipsoidal regions. *Journal of Statistical Software*, 6(8).

von Eye, A., & Mun, E. Y. (2004). Classifying developmental trajectories -a decision making perspective. *Psychology Science*, 46, 65-98.

Prueba de bondad de ajuste para la distribución Gumbel basada en la divergencia de Kullback-Leibler

Paulino Pérez Rodríguez[1]
Colegio de Postgraduados
Humberto Vaquera Huerta[2]
Colegio de Postgraduados
José A. Villaseñor Alva[3]
Colegio de Postgraduados

1. Introducción

En el presente trabajo se desarrolla una prueba de bondad de ajuste para la distribución de valores extremos tipo Gumbel, utilizando la metodología propuesta por Song (2002) la cual se basa en estimaciones de la divergencia de Kullback-Leibler (1951). También se generan las tablas de valores críticos para la prueba para diferentes tamaños de muestra y diferentes niveles de significancia. La potencia de la prueba propuesta es comparada con la de otras pruebas conocidas, mediante un experimento de simulación Monte Carlo.

2. Estadística de prueba

Una variable aleatoria X tiene distribución Gumbel, si su función de densidad es de la forma:

$$f_0(x,\xi,\theta) = \frac{1}{\theta}\exp\left\{-\frac{x-\xi}{\theta} - \exp\left\{-\frac{x-\xi}{\theta}\right\}\right\} I_{(-\infty,\infty)}(x), \quad \xi \in \mathbb{R}, \theta > 0 \qquad (1)$$

[1] perpdgo@colpos.mx
[2] hvaquera@colpos.mx
[3] jvillasr@colpos.mx

Sea $\{X_i\}_{i=1}^n$ una muestra aleatoria de una distribución F, con función de densidad $f(x)$ con soporte en $\mathbb{R}$ y media finita. Se tiene interés en probar el siguiente juego de hipótesis:

$$H_0 : f(x;\cdot) = f_0(x;\xi,\theta) \text{ vs } H_1 : f(x;) \neq f_0(x;\xi,\theta) \tag{2}$$

Para discriminar entre H_0 y H_1 se propone utilizar la divergencia de Kullback-Leibler, para dos distribuciones:

$$KL(F,F_0) = \int_{-\infty}^{\infty} f(x)\log\left(f(x)/f_0(x)\right)dx = \int_{-\infty}^{\infty} f(x)\log f(x)dx - \int_{-\infty}^{\infty} f(x)\log f_0(x)dx \tag{3}$$

Para estimar $\int_{-\infty}^{\infty} f(x)\log f(x)dx = -H(F)$, se utiliza el estimador propuesto por Vasicek (1976), dado por:

$$H_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{2m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\} \tag{4}$$

Donde $m < [\![n/2]\!]$, $X_{(j)} = X_{(1)}\, si\ j < 1$, $X_{(j)} = X_{(n)}\ si\ j > n$ y $X_{(1)} \leq ... \leq X_{(n)}$ son las correspondientes estadísticas de orden, basadas en una muestra aleatoria de tamaño n.

Para estimar $\int_{-\infty}^{\infty} f(x)\log f_0(x)dx$ se utiliza la expresión propuesta por Song (2002), dada por:

$$\frac{1}{n}\sum_{i=1}^{n}\log f_0(X_i,\hat{\xi},\hat{\theta}) \tag{5}$$

Donde $\hat{\xi}$ y $\hat{\theta}$ son los estimadores máximo verosímiles de ξ y θ respectivamente. Si ξ y θ son parcial o completamente especificados, simplemente se sustituyen sus correspondientes valores en (5). Al sustituir (1) en (5) se obtiene:

$$\frac{1}{n}\sum_{i=1}^{n}\log f_0(X_i,\hat{\xi},\hat{\theta}) = -\log\hat{\theta} - \frac{\bar{X}}{\hat{\theta}} + \frac{\hat{\xi}}{\hat{\theta}} - \frac{1}{n}\sum_{i=1}^{n}\exp\left\{-\frac{X_i-\hat{\xi}}{\hat{\theta}}\right\} \tag{6}$$

Por lo tanto un estimador KL_{mn} de $KL(F,F_0)$ se obtiene al sustituir (4) y (6) en (3):

$$KL_{mn} = -H_{mn} + \log\hat{\theta} + \frac{\bar{X}}{\hat{\theta}} - \frac{\hat{\xi}}{\hat{\theta}} + \frac{1}{n}\sum_{i=1}^{n}\exp\left\{-\frac{X_i-\hat{\xi}}{\hat{\theta}}\right\} \tag{7}$$

Se rechaza H_0 si KL_{mn} es grande. Es decir, se rechaza H_0 en favor de H_1 al nivel de significancia α si $KL_{mn} \geq C_{mn}(\alpha)$, donde el valor de la constante crítica $C_{mn}(\alpha)$ queda determinado por el cuantil $(1-\alpha)\times 100$ de la distribución de KL_{mn} bajo la hipótesis nula.

Una vez que se tiene el tamaño de muestra n, se tiene que especificar el parámetro m. Dadas las observaciones $\{x_i\}_{i=1}^{n}$ se estima $KL(F,F_0)$ con KL_{mn}, la idea básica es tomar el valor de m que minimiza KL_{mn}:

$$\hat{m} = \min\left\{m^* : m^* = \arg\max_{m}\left\{H_{mn} : H_{mn} \leq -\frac{1}{n}\sum_{i=1}^{n}\log f_0(X_i,\hat{\xi},\hat{\theta})\right\}\right\}$$

El cálculo de KL_{mn} es relativamente fácil de hacer, pero el problema de obtención en forma analítica de su función de distribución es intratable. Para n grande se puede probar que su distribución no depende de θ ni de ξ, es decir:

$$KL_{mn} \approx -\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{2m}(Y_{(i+m)} - Y_{(i-m)})\right\} + \bar{Y} + \frac{1}{n}\sum_{i=1}^{n}\exp\{-Y_i\}$$

Donde $Y_i,\ i = 1,...,n$ son *v.a.i.i.d.* Gumbel(0,1)

3. Valores críticos

Tabla 1. Valores críticos $C_{m,n}(\alpha)$ de la estadística KL_{mn} obtenida mediante simulación

n	Nivel de significancia α							
	0.01		0.025		0.05		0.10	
	C_{mn}	m	C_{mn}	m	C_{mn}	m	C_{mn}	m
10	0.7434	4	0.6776	3	0.6245	3	0.5678	3
20	0.4812	4	0.4343	4	0.3970	3	0.3557	3
30	0.3555	5	0.3218	4	0.2940	4	0.2653	4
40	0.2890	5	0.2605	5	0.2399	5	0.2177	5
50	0.2430	6	0.2222	6	0.2051	5	0.1857	5
60	0.2125	6	0.1939	6	0.1793	6	0.1631	6
70	0.1910	7	0.1736	7	0.1604	6	0.1458	6
80	0.1718	7	0.1574	7	0.1451	7	0.1326	7
90	0.1578	7	0.1436	7	0.1329	7	0.1212	7
100	0.1464	7	0.1338	8	0.1232	8	0.1122	8
120	0.1276	9	0.1166	9	0.1077	9	0.0982	9
140	0.1132	9	0.1039	11	0.0961	11	0.0873	11
160	0.1028	10	0.0941	11	0.0869	10	0.0789	12
180	0.0933	12	0.0857	12	0.0790	12	0.0710	12
200	0.0865	12	0.0791	12	0.0731	13	0.0662	13

4. Potencia de la prueba

Tabla 2. Potencias estimadas para $\alpha = 0.05$ para algunas alternativas con $n = 20$

Alternativa	D	A^2	C. corr.	KL_{mn}
Normal estándar	0.1663	0.2297	0.1045	0.1782
Logística(0,0.7)	0.2359	0.3184	0.1731	0.2092
t(12)	0.2136	0.2843	0.1471	0.1980
t(4)	0.3157	0.4023	0.2782	0.2776
Cauchy estándar	0.8507	0.8857	0.8450	0.7478
Gamma(1,1)	0.2380	0.3954	0.1939	0.3907
Weibull(Γ(1+1/2), 2)	0.0593	0.0551	0.0144	0.0813
Weibull(Γ(1+1/0.5),0.5)	0.9211	0.9825	0.8348	0.9921
Log-Normal(-0.2,$\sqrt{0}$,4)	0.1415	0.2015	0.1784	0.1246
Fréchet estándar	0.8683	0.9278	0.8731	0.9280

La potencia de la prueba se compara con la de las pruebas desarrolladas por Stephens (1977), las desarrolladas por Chandra *et. al.* (1981) y la propuesta por Kinnison (1989).

5. Ejemplo de aplicación

Tabla 3. Lluvias máximas consecutivas(mm) para 1 día/año en Álamo, Ver.

Año	PP	Año	PP	Año	PP	Año	PP	Año	PP
67	86.8	75	161.6	82	188.3	89	100.0	96	39.7
68	78.5	76	187.6	83	113.9	90	64.3	97	80.3
69	93.1	77	89.9	84	42.5	91	98.0	98	116.4
70	95.5	78	73.4	85	80.0	92	30.7	99	120.0
71	78.1	79	78.1	86	142.6	93	37.9	00	160.0
73	89.9	80	73.3	87	42.9	94	60.7	01	129.0
74	109.5	81	130.1	88	60.2	95	48.7	02	80.0

El tamaño de muestra $n = 35$, para un nivel de significancia $\alpha = 0.05$, de la tabla 1 se toma $m = 4$, y el valor de la constante crítica $C_{4,35}(0{,}05) = 0{,}2639$, solo resta calcular el valor de KL_{mn}, para lo cual se utilizan los estimadores de máxima verosimilitud de los parámetros de localidad y escala, $\hat{\xi} = 74{,}5432$, $\hat{\theta} = 32{,}4328$, obteniéndose KL_{mn} =0.1956, como 0.1956<0.2639 no se rechaza H_0

6. Referencias

Chandra, M., Singpurwalla, N.D. y Stephens, M.A. (1981). Kolmogorov Statistics for Tests of fit for the Extreme Value and Weibull Distributions. *Journal of the American Statistical Association.* **74**, 729-735.

Kinnson, R. (1989). Correlation Coefficient Goodness of Fit Test for the Extreme Value Distribution. *American Statistician*, **43**, 98-100.

Kullback, S. y Leibler, R. A. (1951). On Information and Sufficiency, *Annals of Mathematical*

Statistics, **4**, 49-70.

Song, S. K. (2002). Goodness-of-Fit-Tests Based on Kullback-Leibler Discrimination Information, *IEEE Transactions On Information Theory*, **48**, 1103-1117.

Stephens, M. A. (1977). Goodness-of -Fit-Tests for the Extreme Value Distribution. *Biometrika*, **65**, 730–737.

Vasicek, O. (1976). A Test for Normality Based on Sample Entropy, *Journal of the Royal Statistical Society*, **38**, 54-59.

Prueba de bondad de ajuste para un proceso de Poisson no homogéneo

Francisco J. Ariza Hdez.[1]
Colegio de Postgraduados
Humberto Vaquera Huerta [2]
Colegio de Postgraduados
José A. Villaseñor A. [3]
Colegio de Postgraduados

1. Introducción

El Proceso Poisson No Homogéneo (PPNH) es frecuente y extensivamente utilizado para modelar las fallas en sistemas reparables y en pruebas de confiabilidad de software; uno de los modelos más utilizado para tales situaciones es el Proceso de Goel-Okumoto (1979), que puede ser considerado con diferentes distribuciones, tales como la exponencial, la Pareto, la Weibull, de valores extremos, etc.

Cox y Lewis (1966) mencionan que una de las primeras pruebas para contrastar que los datos siguen un Proceso Poisson Homogéneo (PPH), en la hipótesis nula, contra un PPNH con función de intensidad monótona creciente en la alternativa, es atribuida a Laplace y muestran que esta prueba es óptima para probar un PPNH con función de intensidad log-lineal. Crow (1974) realiza una prueba con la ji-cuadrada para el Proceso Poisson Weibull (PPW). Boswell (1966) desarrolla la Prueba de Razón de Verosimilitudes suponiendo un PPNH arbitrario. Park y Kim (1992) usan la estadística de Kolmogorov-Smirnov, la de Cramer-von Mises y la de Anderson-Darling para una prueba de bondad de ajuste para un proceso Ley Potencia, ellos presentan tablas de valores críticos para esas estadísticas; por su parte López (2002), realiza una prueba para el mismo proceso, utilizando el estimador de momentos del coeficiente de correlación.

[1] arizahfj@colpos.mx
[2] hvaquera@colpos.mx
[3] jvillasr@colpos.mx

El propósito de este trabajo es proponer una prueba de bondad de ajuste para un PPNH basada en el Coeficiente de Correlación, específicamente para el Proceso Goel-Okumoto (1979) tomando en cuenta la distribución Weibull. Esta prueba se aplica a un conjunto de datos reales que representan los tiempos de ocurrencia de fallas en un sistema de control de tácticas navales presentados por Kuo y Young (1996). Se obtienen los valores críticos para diferentes tamaños de muestra y niveles de significancia. También se realiza un estudio para estimar la potencia usando simulación Monte-Carlo.

2. Estadística de prueba

Partimos del supuesto que se observa un PPNH en un período de tiempo [0,T] y que el número de fallas, la cual es una variable aleatoria, denotada por N tiene una distribución Poisson con media θ. De modo que $\{N(t); t > 0\}$ es un PPNH con función de valor medio $m(t) = \theta F(t)$, donde F es la función de distribución acumulada de f. En particular cuando $F(t) = (1 - e^{-\beta t^{\alpha}})$, se tiene que $N(t)$ es un PPNH con función de valor medio:

$$m(t) = \theta(1 - e^{\beta t^{\alpha}}); \quad t \in [0, T]; \alpha > 0; \beta > 0; \theta > 0. \tag{1}$$

El cual es llamado Proceso Goel-Okumoto (1979), con función de distribución Weibull. Así, la prueba que se presenta se realiza para la función de valor medio de este proceso condicionando $N = n$, por lo que usando el modelo (1) se desea probar:

$$H_0 : m(t) = \theta(1 - e^{\beta t^{\alpha}}) \quad \text{vs} \quad H_1 : m(t) \neq \theta(1 - e^{\beta t^{\alpha}}) \tag{2}$$

La prueba se desarrolla linealizando la función de valor medio del proceso, la cual se obtiene mediante una transformación doble logarítmica para el modelo expresado en (1), que nos conduce a una forma lineal en $log(t)$, es decir:

$$\log\left[-\log(1 - m(t)/\theta)\right] = \log\beta + \alpha\log(t) \tag{3}$$

Bajo H_0 en (2) y dado que $N = n$, con tiempos de ocurrencia de eventos $t_1, t_2, ..., t_n$, resulta de (3) que:

$$\log\left[-\log(1 - m(t_i)/\theta)\right] = \log\beta + \alpha\log(t_i); \quad i = 1, ..., n \tag{4}$$

Ya que $m(t_i)$ es una cantidad no observable durante el proceso, entonces un buen representante de su valor desconocido, es su valor medio, por lo que podemos sustituir $m(t_i)$ por $E\left[(m(t_i)\right]$. Para calcular $E\left[m(t_i)\right]$, los tiempos en los cuales los eventos ocurren son distribuidos como las n estadísticas de orden de una muestra aleatoria de n observaciones de la distribución:

$$F(t) = \frac{\int_0^t \lambda(s)dx}{\int_0^T \lambda(s)ds} = \frac{m(t)}{m(T)} \tag{5}$$

Note que n es una realización de la variable aleatoria Poisson con parámetro $m(T)$. Dados los t_i y haciendo $\phi_T = m(T)$, de (5) se tiene:

$$m(t_i) = \phi_T F(t_i) \tag{6}$$

Así, la variable aleatoria en (6) se distribuye como la i-ésima estadística de orden, de una muestra de tamaño n de la distribucińn $U(0, \phi_T)$; ya que $F(\cdot)$ converge en probabilidad a una distribución uniforme estándar. Por lo tanto $E\left[m(t_i)\right] = \phi_T \frac{i}{n+1}$.

Entonces sustituyendo $m(t_i)$ por $E\left[m(t_i)\right]$ en la ecuación (4), resulta:

$$\log\left[-\log(1 - \frac{\phi_T}{\theta}\frac{i}{n+1})\right] = \log\beta + \alpha\log(t_i); \quad i = 1, ..., n \tag{7}$$

Las consideraciones para obtener (7) se establecen de condiciones reales; es decir, observando el número total de eventos en el intervalo de tiempo $[0, T]$, y los tiempos de ocurrencia de eventos t_i. En este contexto, se puede ver que $\phi_T = \theta$ las cuales representan el número de eventos promedio ocurridos hasta el tiempo T, y pueden ser estimados por $N = n$, ya que en este tipo de modelos, denotan el valor medio de la variable aleatoria N la cual tiene distribución Poisson, Kou y Yang (1966). De esta forma se hará referencia sobre ϕ_T, θ y n indistintamente suponiendo que el proceso es observado en un intervalo de tiempo fijo $[0, T]$, por lo que la ecuación (7), resulta ser:

$$\log\left[-\log(1 - \frac{i}{n+1})\right] = \log\beta + \alpha\log(t_i); \quad i = 1, ..., n \tag{8}$$

Haciendo $Y_i = \log\left[-\log(1-\frac{i}{n+1})\right]$ y $X_i = \log(t_i)$ para $i = 1, ..., n$ se puede escribir (8) en la forma lineal:

$$Y_i = \beta' + \alpha X_i \tag{9}$$

Entonces el modelo $m(t)$ será adecuado si cumple la relación (8) y la veracidad de H_0 en (2) estará sustentada por el grado de asociación lineal entre las variables X_i y Y_i de la ecuación (9). Esta dependencia lineal es medida mediante el estimador de momentos del coeficiente de correlación r, definido como:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{10}$$

Bajo H_0, la estadística r estará cercana a la unidad ya que se espera una asociación lineal perfecta entre X y Y. Por lo que se rechaza H_0 si $r \leq C_{\alpha,n}$; donde $C_{\alpha,n}$ es el valor de la constante crítica que queda determinada por el cuantil $(1-\alpha) \times 100$ de la distribución de r bajo la hipótesis nula. Dicha distribución se obtuvo vía simulación Monte-Carlo con 50000 repeticiones.

3. Valores críticos

Tabla 1. Valores críticos $C_{\alpha,n}$ de la estadística r

	Nivel de significancia α					Nivel de significancia α			
n	0,01	0,025	0,05	0,10	n	0,01	0,025	0,05	0,10
15	0.8566	0.8857	0.9081	0.9302	90	0.9388	0.9547	0.9655	0.9753
20	0.8712	0.8989	0.9204	0.9402	100	0.9442	0.9581	0.9678	0.9766
25	0.8774	0.9058	0.9271	0.9463	150	0.9575	0.9689	0.9763	0.9827
30	0.8892	0.9147	0.9345	0.9516	200	0.9649	0.9744	0.9805	0.9859
40	0.9018	0.9261	0.9438	0.9588	300	0.9745	0.9809	0.9855	0.9893
50	0.9151	0.9356	0.9507	0.9642	400	0.9798	0.9853	0.9886	0.9916
60	0.9224	0.9418	0.9558	0.9683	500	0.9828	0.9873	0.9902	0.9928
70	0.9284	0.9469	0.9596	0.9708	700	0.9872	0.9904	0.9926	0.9945
80	0.9352	0.9520	0.9632	0.9735	1000	0.9908	0.9929	0.9945	0.9958

4. Potencia de la prueba

Tabla 2. Potencias estimadas para un nivel de significancia $\alpha^* = 0.05$

F. de Intensidad	$n = 30$		$n = 50$		$n = 100$	
Alternativa	L_{aprox}	CC	L_{aprox}	CC	L_{aprox}	CC
$\lambda_{WL}(t) = \alpha\beta t^{\alpha-1} exp\left[\beta t^{\alpha}\right]$	0.3156	0.5534	0.3626	0.7414	0.4240	0.9376
$\lambda_{CL}(t) = e^{\alpha+\beta t}$	0.3160	0.5466	0.3468	0.7070	0.4042	0.9036
$\lambda_{MO}(t) = \frac{\alpha}{\beta+t}$	0.1248	0.1136	0.1212	0.3412	0.1120	0.9322
$\lambda_{LP}(t) = \alpha\beta t^{\alpha-1}$	0.2038	0.1840	0.2076	0.2486	0.2224	0.4042

5. Ejemplo de aplicación

Se tiene el siguiente conjunto de datos obtenidos de Kuo y Young Yang (1996), los cuales representan los tiempos entre fallas: 9, 12, 11, 4, 7, 2, 5, 8, 5, 7, 1, 6, 1, 9, 4, 1, 3, 3, 6, 1, 11, 33, 7, 91, 2, 1, 87, 47, 12, 9, 135.

Note que se tiene interés en probar la hipótesis en (2). Así, para $n = 31$, se calcula el valor de la estadística r a partir de (9) y (10) obteniendo $r = 0{,}9753$. Considerando un tamaño de prueba $\alpha^* = 0{,}05$ obtenemos el valor crítico $C_{0{,}05,30} = 0{,}9345$, de la tabla 1, por lo que se decide no rechazar H_0 en (2) ya que $r = 0{,}9753 > 0{,}9345$.

6. Referencias

Arnold, B. C., Balakrishnan, N., y Nagajara, H. N. (1992). *A First Course in Order Statistics*. John Wiley & Sons, Inc.

Basawa, I., y Prakasa R. (1980). *Statistical Inference for Stochastics Processes*. ACADEMY PRESS.

Boswell, M. T. (1966). Estimating and Testing Trend in a Stochastic Process of the Poisson Type, *Annals Mathematical Statistics*, **37**, 1564-1573.

Cox, D. R., y Lewis, P. A. (1966). *The Statistical Analysis of Series of Events*, METHUEN, London.

Crow, L. H. (1974). Reliability Analysis For Complex, Repairable System, *In Reliability and Biometry Statistical Analysis of Lifelength*, Philadelphia, 379-410.

Goel, A. L. y Okumoto, K. (1979). Time-Dependence Error Detection Rate Models for Software Reliability and Other Preformance Measures, *IEEE Transactions on Reliability*, **38**, 206-211.

López, S. L., Villaseñor, A. J. y Vaquera H. H. (2002). Dos Pruebas de Bondad de Ajuste Para Procesos de Poisson No Homogéneos, *Agrociencia*, **36**, 703-712.

Kuo, L. y Young Yang, T. (1996). Bayesian Computation for Non-Homogeneous Poisson Processes in Software Reliability, *Journal of the American Statistical Asssociation. Theory and Methods*, **91**, 763-773.

Park, W. J. y Kim, Y. G. (1992). Goodness of Fit Test For the Power-Law Process, *IEEE Transaction of Reliability*, *43*, 107-111.

Cartas multivariadas usando análisis de componentes principales

Arely Elizabeth Espinosa Jiménez[1]
Facultad de Ciencias Físico Matemáticas,Universidad Autónoma de Coahuila
Félix de Jesús Sánchez Pérez[2]
Centro de Investigación en Matemáticas Aplicadas,Universidad Autónoma de Coahuila
Emilio Padrón Corral[3]
Centro de Investigación en Matemáticas Aplicadas, Universidad Autónoma de Coahuila

1. Introducción

La productividad de la industria en la actualidad tiene retos a resolver ante las nuevas exigencias de los clientes de un mundo globalizado, lo cual lleva a un mejoramiento continuo en la calidad. Generando cambios importantes en la rentabilidad, producción, calidad y otras cuestiones del producto. Siendo la automatización una herramienta empresarial que ha crecido a través del tiempo y, generando una infinidad de información referente al artículo producido con las cuales se toman decisiones importantes para minimizar la variabilidad del proceso.

La industria se ha fortalecido en la prevención de los errores en la producción o por medio de ésta corregir los mismos con las herramientas estadísticas. Es el análisis multivariado, el cual consta de técnicas y métodos que ayudan a estudiar e interpretar un conjunto de variables. A través de las cartas de control multivariado se detectan errores o estabilidad en el proceso.

2. Análisis De Componentes Principales

El origen del análisis de componentes principales (ACP) data de 1901 con Karl Pearson que publicó un trabajo sobre el ajuste de un sistema de puntos en un multiespacio a una línea o

[1]arely1327@gmail.com
[2]fel1925@yahoo.com
[3]epadron@cima.uadec.mx

un plano. Siendo retomado en 1933 por Hotelling, quien fue el primero en formular ACP tal como se ha difundido hasta nuestros días. ACP deberá ser aplicado cuando se desee conocer la relación entre los elementos de una población y se sospeche que en dicha relación influye de manera desconocida un conjunto de variables o propiedades de los elementos y genera nuevas variables las cuales expresan la información más importante y relevante de los datos originales.

Al reducir la dimensión de los datos y formarse nuevas variables que no sean correlacionadas, por medio de la combinación lineal de las variables originales donde se describe la mayor tendencia de los datos. Los nuevos valores encontrados $\tau = \{x_k, k = 1, 2, \cdots, k\}$ contienen la mayor parte de información estadística, siendo presentada en los datos originales. Los ACP pueden ser hechos en base a los eigenvalores y los eigenvectores de una matriz de varianza-covarianza donde Sx_1^2, Sx_2^2 representan las varianzas de x_1 y x_2 respectivamente y la covarianza entre x_1 y x_2 es:

$$C(x_1, x_2) = \begin{pmatrix} Sx_1^2 & cov \\ cov & Sx_2^2 \end{pmatrix}$$

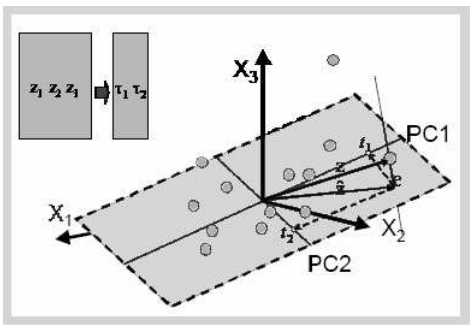


Figura 1. Ejemplo gráfico de ACP con dos variables

3. Modelos de Control de Calidad

El control estadístico de procesos (Statistical Process Control) consiste en monitorear el comportamiento de un proceso a lo largo del tiempo para detectar la ocurrencia de eventos especiales. Una vez detectada la ocurrencia de un evento especial se trata de diagnosticar el problema, encontrar las causas asignadas para la desviación y corregir el proceso, implementando medidas correctoras. El modelo más tradicional de control de calidad es el modelo de detección. Este modelo depende de un equipo de inspectores para verificar el producto en varias etapas de su producción y eliminar los defectos. El método resulta inadecuado e ineficiente. Se invierte tiempo, dinero y materiales en productos o servicios que no siempre son satisfactorios.

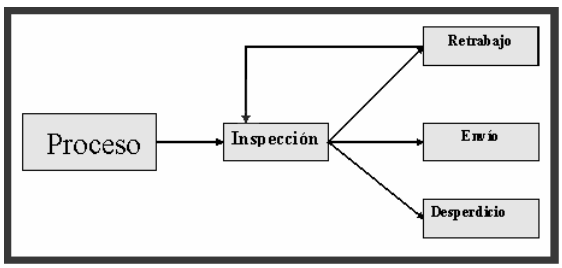


Figura 2. Modelo de dependencia

El modelo de prevención , utiliza la información de producción y provee un método eficiente para analizar el proceso e indicar el lugar y el momento en el que las mejoras pueden prevenir la producción de artículos defectuosos, es decir, monitorea el proceso de tal forma que los ajustes necesarios se realizarse antes de que la calidad sea afectada.

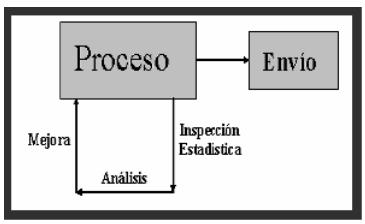


Figura 3. Modelo de Prevención

Los gráficos de control son la herramienta para revelar las causas asignables y el diseño de experimentos es la técnica que indican la forma de ajustar los parámetros del proceso. Detectan la presencia de causas asignables tan pronto como sea posible para permitir una acción correctiva adecuada que las elimine y regrese el proceso a un estado de control estadístico. Si una observación cae fuera de límites de control de un gráfico o se distingue algún patrón no aleatorio en la gráfica, se supone la existencia de causas asignables o especiales de variación y se dice que el proceso se encuentra fuera de control. La variabilidad se hace presente en el proceso de fabricación del producto y representa un gran obstáculo en su calidad, puede ser debida a una multitud de causas pequeñas que actúan en conjunto y son contables, denominada variabilidad inherente.

4. Cartas de Control Multivariadas con ACP

Las cartas de control es un proceso sujeto a la variable normal donde éste permanecerá bajo control hasta que se verifique un evento especial; de modo que los gráficos de control constituyen diferentes contrastes de hipótesis cuyo objetivo es detectar la ocurrencia de un evento especial lo más rápido posible. La aplicación de componentes principales supone la construcción de un modelo ACP a partir de un conjunto de referencia el cual determina la variaciones que forman parte de la operación normal del proceso, donde se deben incluir todas las variaciones que proporcionen resultados aceptables. Si el conjunto de variables deja fuera variables aceptables esto ocasionará falsas alarmas; en el caso contrario si se toman variables en exceso, se pierde la sensibilidad para detectar variables con resultados no aceptables. Cuando se tienen grandes cantidades de datos correlacionados es que los ACP son muy útiles por que permiten reducir la dimensión del problema, tomando en cuenta información acerca de la variación relativa existente entre las variables y reduciendo el nivel de ruido.

5. Estadísticos de las Cartas de Control con ACP

La T^2 de Hotelling: es un estadístico basado en la distancia de Mahalanobis que se emplea en la monitorización multivariada para medir la distancia de cada observación al centro del modelo ponderado según la estructura de covarianza. La expresión empleada cuando cumple

con lo anterior de una nueva observación z al origen en el espacio original k-dimensional de las variables del proceso y el estadístico T^2 de Hotelling es: $x^2 = z^T \sum^{-1} z$, $T^2 = z^T S^{-1} z \;\; con \;\; S = \frac{x^t x}{(N-1)}$ El limite de control superior de la T^2 de Hotelling se calcula a partir de la expresión:

$$T_\alpha^2 = \frac{(N^2-1)}{N(N-A)} F_\alpha(A, N-A)$$

donde $F_\alpha(A, N-A)$ es el percentil 100 $\times(1-\alpha)$ de la distribución F de Snedecor con (A,N-A) grados de libertad.

La T^2 de ACP: Obteniéndose las A (A=dimensión real del proceso) variables latentes o componentes principales que son combinaciones lineales de las k- variables del proceso (usualmente $A \ll k$), lo que incrementa la efectividad de la monitorización.

El estadístico de T^2_{ACP} es: $P^T_{ACP} = \sum_{a=1}^{2} \frac{t_\alpha^2}{\lambda_a} \quad t_\alpha = \sum_{j=1}^{J} P_{aj} \chi_j$

Su límite superior $T_\alpha^2 = \frac{(N^2-1)A}{N(N-A)} F_{(A,N-A)}$ donde $F_{(A,N-A)}$ es la distribución F de Snedecor con $(A, N-A)$ grados de libertad. Para el uso de este estadístico es necesario contar con las variaciones de los scores y residual.

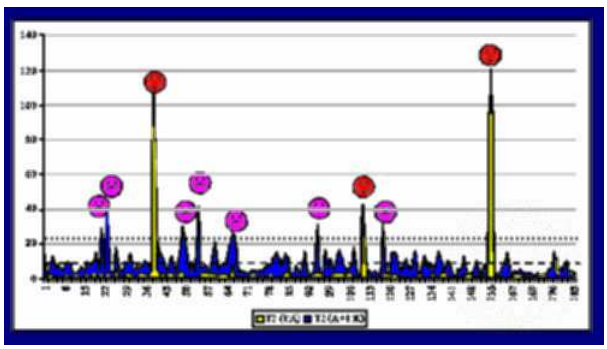


Figura 4. Comparación de las Cartas de Control entre la $T^2_{Hotelling}$ y T^2_{ACP}

6. Conclusiones

Con esta metodología se espera lograr tener estimaciones más eficientes en los análisis realizados, en empresas de servicios y manufactureras.

7. Referencias

De la Garza González, Mauricio(1996). *Desarrollo de Diagrama de Control Estadístico para proceso de alto volumen y corto tiempo de ciclo*. Tesis de Instituto Tecnológico y de Estudios superiores de Monterrey. Pág:2-18

Dallas E. Johnson(2000). *Métodos Multivariados aplicados al análisis de datos.* International Thomson Editores. Pág:1-13

Fuchs Camila, S. Kenett Ron (1998). *Multivariate Quality Control.* Marcel Dekker. Pág:9-13,115-120.

Ferrer Riquelme Alberto J. (2005). *Curso de Verano Técnicas Estadísticas Multivariantes para el Control Estadístico de Procesos Altamente Automatizados.* III Verano Estadística Industrial CIMAT, Guanajuato.

Análisis de componentes aplicado a la evaluación del rendimiento de hule

Emilio Padrón Corral[1]
Centro de Investigación en Matemáticas Aplicadas, Universidad Autónoma de Coahuila
Ignacio Méndez Ramírez[2]
Instituto de Investigación en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México
Armando Muñoz Urbina[3]
Asesoría Privada
Félix de Jesús Sánchez Pérez[4]
Centro de Investigación en Matemáticas Aplicadas,Universidad Autónoma de Coahuila

1. Introducción

Las investigaciones en fitomejoramiento hoy en día, en particular en genética molecular, están más interesadas en carácteres cualitativos monogénicos, debido a que presentan un menor problema en el desarrollo y aplicación de las técnicas modernas. Sin embargo, los carácteres complejos son demasiado importantes para permanecer de lado por largo tiempo. Para hacerlos más accesibles al mejoramiento convencional y quizás también para los métodos modernos de mejoramiento biotecnológico, un análisis de sus componentes es requerido. La identificación de los principales componentes y la determinación de su contribución relativa a la variación del carácter complejo es el primer objetivo de este análisis.

El objetivo general de este trabajo es efectuar un análisis secuencial de componentes para determinar qué variables contribuyen más al rendimiento de hule. Los resultados obtenidos nos indican que las dos componentes más importantes son: acumulación de contenido de hule por altura de planta $x_2 = \frac{\%H}{APL}$; y acumulación de peso seco por contenido de hule $x_3 = PS\,\%H$; explicando 22 % y 58 % respectivamente, de la variación del rendimiento de hule.

[1] epadron@cima.uadec.mx
[2] imendez@servidor.unam.mx
[3] epadron@cima.uadec.mx
[4] fel1925@yahoo.com

2. Materiales y Métodos

Las plantas de guayule utilizadas en la presente investigación provienen de una población silvestre del ejido Gómez Farías ubicado a 56 km de Saltillo, Coahuila, México. Este ejido presenta coordenadas geográficas de longitud Oeste 101ž 03' y 24ž 97' latitud Norte y una altura de 1900 msnm, en la provincia de la Sierra Madre Oriental, subprovincia de las Sierras Transversales. En este experimento se trabajó con una muestra de 35 plantas completas colectadas en el Otoño de 1997, plantas de aproximadamente dos años de edad determinada de acuerdo a Curtis (1947). De las plantas muestreadas se tomó la altura de planta (APL), posteriormente las plantas se secaron en una estufa para obtener el peso seco (PS). Una muestra de 5 g de tejido de la planta fue molida en un molino Wiley y fue utilizada para determinar el contenido de hule (% H) y de resina (% R) por el método de extracción de Soxhlet. Para el análisis de los datos de componentes del rendimiento de hule se utilizó el método propuesto por Sparnaaij & Bos (1993) y se consideraron las variables: altura de planta (APL), contenido de hule (% H), peso seco (PS), rendimiento de hule por planta (PH/PL).

La definición de componente corresponde con la dada por Thomas & Grafius (1976) y por Sparnaaij & Bus (1993): estrictamente aquellos carácteres los cuales cuando se multiplican conjuntamente dan exactamente el rendimiento (carácter complejo). En fitomejoramiento el análisis de componentes es utilizado generalmente para encontrar un criterio de selección para rendimiento. Cuando éste es el objetivo, no hay necesidad de prestar atención a la naturaleza y a la secuencia de los componentes. Altura de planta, peso seco y contenido de huele son características componentes del rendimiento de hule lo cual ha sido determinado por varios investigadores.

Para rendimiento de hule:

La primera componente $x_1 = a$, donde a = altura de planta en cm.

La segunda componente $x_2 = \frac{b}{a}$, donde b = contenido de hule en por ciento.

La tercera componente $x_3 = \frac{c}{b}$, donde c = peso seco en g.

La cuarta componente $x_4 = \frac{y}{c}$, donde y = rendimiento de hule en g.

En resumen: $x_1 * x_2 * x_3 * x_4 = y$

3. Resultados y Discusión

Cuadro 1. Coeficientes de correlación (r) entre los componentes (x_1, x_2, x_3, x_4) del carácter complejo y (rendimiento de hule) y los caracteres primarios (a, b, c, y). La determinación complementaria (cd), derivada de los valores de $r^2(y, a, y)$.

	a	b	c	y
$x_1 = a$	1,00	0.01	0,37*	0,33*
$x_2 = \frac{b}{a}$	−0,77**	0,55**	-0.16	0.02
$x_3 = \frac{c}{b}$	0,37*	**-0.16**	0,86**	0,68**
$x_4 = \frac{y}{c}$	0.01	1,00**	**0,33***	0,58**
y	0,33**	0,58**	0,95**	1.00
$r^2(y, a, \cdots, y)$	0.11	0.33	0.91	1.00
$cd(y, x_1, \cdots, x_4)$	0.11	0.22	0.58	0.09

*Significativo al 5 %, **Significativa al 1 %.

En dicho cuadro, se observa que la correlación entre cada componente y su carácter primario precedente (en negritas) ilustra cómo las componentes (y el producto) de las componentes precedentes están relacionados. La componente altura de planta $x_1 = a$ se correlaciona positiva y significativamente con peso seco c (r = 0.37*) y con rendimiento y (r = 0.33*) lo que nos indica que plantas con gran altura acumularon alto peso seco influyendo así positivamente en el rendimiento de hule.

La componente $x_2 = \frac{b}{a}$, se correlacionó negativa alta y significativamente con altura de planta **a** (r = -0.77**) y positiva alta y significativamente con contenido de hule **b** (r = 0.55**), lo que significa que hubo plantas con alto contenido de hule que presentaron baja altura de planta (plantas: 43, 24, 37, 14, 17, 20) y confirma que los altos valores de la componente $x_2 = \frac{b}{a}$ fueron dados por plantas que presentaron altos valores de contenido de hule o valores

muy reducidos de altura de planta (plantas: 16, 4). Las plantas 31, 45 y 21 presentaron alto contenido de hule pero no una reducida altura de planta.

La componente $x_3 = \frac{c}{b}$ se correlacionó positiva y significativamente con altura de planta **a** (r = 0.37*) y alta y significativamente con peso seco **c** (r = 0.86**) y con rendimiento de hule **y** (r = 0.68**) por lo tanto, algunas plantas con gran altura presentaron alto peso seco y alto rendimiento de hule, pero no muy altos contenidos de hule como las plantas: 35, 41, 19, 42. La planta 31 presentó alto peso seco pero también alto contenido de hule, por otro lado la planta 38 presentó la característica de acumular alto peso seco con una reducida altura de planta. La correlación de la componente $x_3 = \frac{c}{b}$ con contenido de hule **b** (r = -0.16) fue negativa pero no significativa.

Las plantas con mayor índice de cosecha $x_4 = \frac{y}{c}$ también presentaron los más altos valores de contenido de hule **b** (r = 1.00**) y alto rendimiento de hule **y** (r = 0.58**), plantas: 31, 45, 21, 43, 28, 24, 37, por lo tanto, una manera de mejorar el índice de cosecha y el rendimiento de hule es seleccionar plantas con altos contenidos de hule. Las plantas 31, 45 y 21 también fueron favorecidas por su alto peso seco, lo que explica la correlación positiva y significativa de índice de cosecha $x_4 = \frac{y}{c}$ con peso seco **c**(r = 0.33*).

Los valores de **cd** que indican incrementos en la determinación de **y** (rendimiento de hule), atribuible a la intervención de los componentes x_1, x_2, x_3, x_4. Las determinaciones complementarias indican que las dos componentes más importantes son x_2 y x_3, explicando 22 % y 58 %, respectivamente de la variación de **y**.

Las componentes x_1 y x_4 tienen menor influencia explicando el 11 % y 9 %, respectivamente. Las plantas mostraron alto contenido de hule con respecto a altura de planta como las plantas: 45, 21, 43 en la componente $x_2 = \frac{b}{a}$, y plantas que mostraron alto peso seco con respecto al contenido de hule como las plantas: 31, 35, 41, 19 en la componente $x_3 = \frac{c}{b}$, presentaron los más altos rendimientos de hule. Los genes que actúan en la componente $x_2 = \frac{b}{a}$ están principalmente activos durante el otoño e invierno, cuando las bajas temperaturas nocturnas estimulan la transcripción de genes que codifican para las enzimas incluídas en la síntesis de hule.

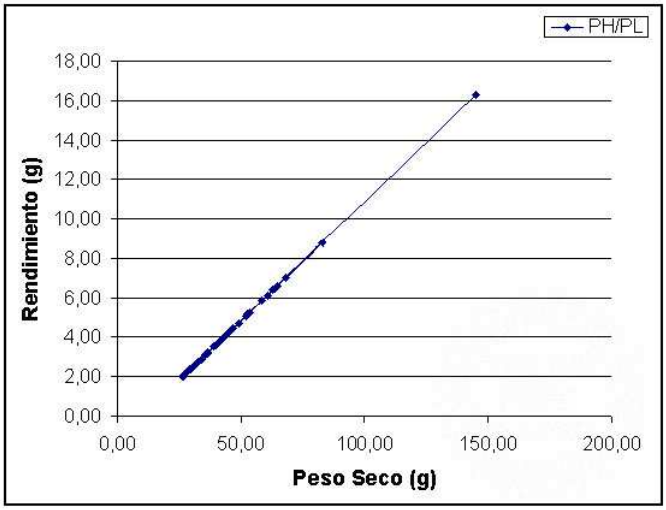


Figura 1. Relación entre rendimiento de hule $\frac{PH}{PL}$ con el peso seco. Valores ajustados para $\hat{y}_{\frac{PH}{PL}}$=-1.1681756+0.12039817(x), R^2=0.91

En la Figura 1. Se observa una tendencia lineal entre los rendimientos de hule (PH/PL) con el peso seco por planta (PS), por lo que plantas con mayor biomasa incrementan sus rendimientos de hule.

5 Conclusiones

El análisis de Componentes del rendimiento nos permitió examinar la amplia variabilidad que presentan las plantas de guayule de la población silvestre de Gómez Farías, Coah., México, con respecto a las componentes que pueden ser importantes para obtener progenitores que produzcan altos rendimientos de hule. De los resultados obtenidos se observa que la componente: $x_2 = \frac{b}{a}$, para rendimiento de hule nos permitió detectar plantas que con una reducida altura de planta y muy alto contenido de hule produjeron altos rendimientos de hule. La componente $x_3 = \frac{c}{b}$, nos permitió detectar plantas que a través de un alto peso seco y no muy bajos contenidos de hule produjeron el más alto rendimiento de hule.

El índice de cosecha indica que el ajuste común que las plantas hacen para soportar una situación de estrés es reducir el crecimiento y el tamaño. Pero cuando el estrés es severo o

no está bien distribuido durante el período de crecimiento, el índice de cosecha puede ser reducido. Por otra parte, las plantas bien adaptadas al estrés pueden dar un alto índice de cosecha con bajo rendimiento de materia seca como las platas: 43, 24, 37, para rendimiento de hule. Por lo tanto, con este tipo de plantas se podría aumentar el rendimiento de hule incrementando la densidad de plantas por hectárea.

Referencias

Curtis, O.F. (1947). *"Distribution of rubber and resins in guayule".* Plant Physiology. 22:333-459.

Sparnaaij, L.D. & I. Bos. (1993). *"Component analysis of complex characters in plant breeding. I. Proposed method for quantifying the relative contribution of individual components to variation of the complex character".* Euphytica 70: 225-235.

Thomas, R.L. & J.E. Grafius. (1976). *"Prediction of heterosis levels from parental information".* Proc. Seventh Congress of Eucarpia: 173-180.

Propiedades estadísticas del muestreo por línea intercepto y cuadros cargados en la estimación de la cobertura

Félix de Jesús Sánchez Pérez[1]
Centro de Investigación en Matemáticas Aplicadas
Emilio Padrón Corral[2]
Centro de Investigación en Matemáticas Aplicadas, Universidad Autónoma de Coahuila
Dino Ulises González Uribe
Depto. de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro

1. Introducción

En los inventarios de vegetación, frecuentemente se utiliza el muestreo estadístico para obtener información rápida, veraz y económica para la toma de decisiones. El objetivo del muestreo, en este caso, es la obtención de un estimación descriptiva de algunas características de la población vegetal en estudio, como lo son la cobertura y la densidad (Burguete y Carrillo (1972); Lyon (1968)).

Esta estimación debe representar suficientemente el parámetro en estudio y permitir detectar con precisión las diferencias entre poblaciones vegetales (Lyon(1968)).

Al muestreo que utiliza líneas rectas para conocer cobertura y densidad se le llama muestreo por **línea intercepto**; se le denomina así por considerar en la evaluación a aquellos individuos que se cortan por la línea en su parte aérea. Se utiliza porque es de fácil aplicación.

Si son cuadros, se cuentan sólo aquellas unidades de muestreo con los individuos de interés para obtener así la densidad vegetal; a este procedimiento se le denomina muestreo por **cuadros cargados**. El rango de aplicación de ambos procedimientos de muestreo es muy amplio (Cochran (1950); Kaiser(1983); Swindel(1983)).

[1] fel1925@yahoo.com
[2] epadron@cima.uadec.mx

Dado que el interés es la estimación de un parámetro de una característica, como la media poblacional de la cobertura y/o la media poblacional de la densidad a partir de una muestra, la estimación está sujeta a riesgo, entre otras razones, debido a la estructura del estimador con la que infiere el valor del parámetro. Por tal motivo, en un diseño de muestreo se propone un estimador y se analizan sus propiedades como sesgo, eficiencia, consistencia y otras de relevancia que son señaladas en teoría estadística, como la suficiencia (Burguete y Carrillo(1972)).

Si el estimador utilizado para calcular la media poblacional de la cobertura y la media poblacional de la densidad vegetal posee el mayor número de estas propiedades deseables, entonces se considera de buena calidad, por lo tanto la estimación de ambas variables es satisfactoria y, sin duda alguna, se puede utilizar en la estimación de ambos parámetros (Burguete y Carrillo (1972); Kisinger et al.(1960)). Dada la utilización de los procedimientos de muestreo mencionados, en este estudio se propone como objetivo demostrar las propiedades estadísticas básicas de los estimadores de la línea intercepto y muestreo por cuadros cargados, los cuales son: insesgamiento, suficiencia, consistencia y eficiencia.

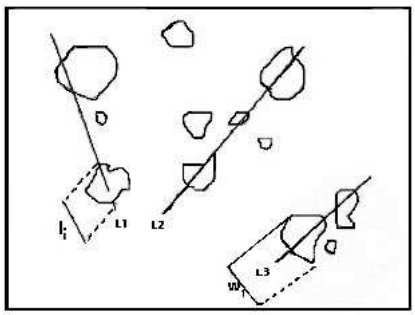


Figura 1. Área de estudio con $M = 12$ individuos de interés y $n = 3$ unidades de muestreo

2. Materiales y Métodos

Descripción del Muestreo por Línea Intercepto

El uso de la línea intercepto puede definirse como un procedimiento de muestreo de vegetación basado en la medición de todas las plantas interceptadas por un plano vertical de líneas,

localizadas aleatoriamente y de igual longitud (Canfield (1941)). Aunque también puede hacerse la estimación con líneas de diferente longitud (McDonald (1980)). Con el muestreo por línea intercepto pueden determinarse la cobertura de corona y la densidad vegetal (Fig.1).

Descripción del Muestreo por Cuadros Cargados.

Si tenemos un área A que se subdivide en n unidades de muestreo en forma de cuadro, cada uno de ellos de área a, a los cuadros con la presencia de individuos de interés serán los cuadros cargados (Figura 2). Si se denota por y el número de cuadrados no cargados en una muestra de unidades de muestreo de tamaño n, se puede obtener el número de individuos en el área A.

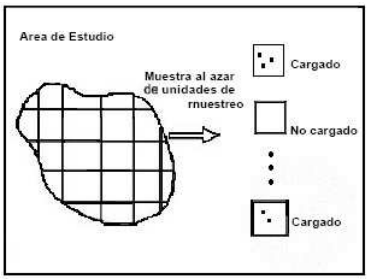


Figura 2. Cuadros cargado y no cargado tomados de un área A

Estimación y Estimador

Un estimador es una fórmula, la cual establece cómo calcular un valor dado contenido en una muestra aleatoria que se obtiene en campo; un estimador se designa como $\hat{\theta}$ y se toma como si fuera el valor verdadero de una población al cual se llama parámetro; el parámetro θ sólo se conocerá si se realiza un censo de población; por esta razón, el estimador es de gran importancia en el muestreo. La acción de utilizar al estimador y conocer las consecuencias de utilizarlo como una función de decisión al tomar el valor del estimador como si fuera el parámetro, es la estimación.

Los estimadores de la cobertura y densidad vegetal en el muestreo por línea intercepto y muestreo por cuadros cargados, estiman a la media poblacional del parámetro, y si el estimador usado posee la propiedad de que su valor esperado, o esperanza matemática sea

igual al parámetro se dice que el estimador es insesgado, teóricamente, si

$$E(\hat{\theta}) = \theta$$

Como la media poblacional se estima, el valor del estimador varía de acuerdo con la muestra aleatoria que se tome en campo; si se conoce su varianza se puede estimar su variación con respecto a su media. La consistencia de un estimador se prueba en la varianza del estimador de la media poblacional cuando el tamaño de muestra crece; si la varianza se aproxima a cero cuando se aumenta el tamaño de la muestra, se dice que el estimador es consistente y ha alcanzado su máxima eficiencia.

También es de interés saber si el estimador contiene la información necesaria para estimar el parámetro; cuando esto sucede, se dice que el estimador posee la propiedad de suficiencia, para lo cual se necesita saber la función de distribución de la variable en un estudio de población, aunque se puede suponer, no obstante que la distribución normal es la de uso más frecuente (Burguete y Carrillo(1972)).

Cuando dos o más estimadores insesgados estiman a la misma media poblacional, puede escogerse para su uso aquél estimador que tenga la menor varianza, a lo cual se le llama eficiencia relativa.

3. Resultados y Discusión

El estimador de la media poblacional de la cobertura vegetal que se usa en el muestreo por línea intercepto es insesgado. Si se quiere estimar la cobertura de corona de alguna especie vegetal, a partir de una muestra aleatoria de n unidades de muestreo en una población, el estimador a usar junto con su varianza es:

$$\hat{c} = 25\pi\left(\frac{\sum_{i=1}^{m}\ell_i}{\sum_{j=1}^{n}L_j}\right), \quad \hat{V}(\hat{c}) = \frac{S_\ell^2}{nL^2}, \quad S_\ell^2 = \frac{1}{n-1}\left(\sum_{i=1}^{m}\ell_i - \hat{c}L\right)$$

Si es una cobertura cuadrada o rectangular, se utiliza la expresión dentro del paréntesis (para el caso de la media poblacional); si la varianza estimada de la media de la cobertura es consistente, el estimador de la media también es eficiente.

Si el estimador de la densidad vegetal en el muestreo por línea intercepto estima insesgamiento a la media poblacional, su varianza es consistente y, además, el estimador posee la propiedad de suficiencia.

Los estimadores encontrados son los siguientes:

$$\hat{D} = \frac{\sum_{i=1}^{m} w_i^{-1}}{\sum_{j=1}^{n} L_j} \quad donde \quad \hat{V}(\hat{D}) = \frac{S_w^2}{nL^2} \quad y \quad S_w^2 = \frac{1}{n-1}\left(\sum_{i=1}^{m} w_i^{-2} - \frac{n\hat{D}}{A}\right)$$

En el muestreo por línea intercepto L la suma total de las unidades de muestreo l_i y w_i son el intercepto y ancho máximo, respectivamente, de las coberturas interceptadas (ver Fig.1).

El estimador de la media poblacional de la densidad vegetal y varianza del muestreo por cuadros cargados, posee las propiedades deseables de estimación; en este caso, la función de distribución Poisson es la adecuada para el muestreo. Así entonces, se contarán en aquella muestra de n unidades como 0, a los que no tienen presencia de individuos de intéres o no cargados, y como 1, a los que sí tienen individuos de interés o cargados (ver Fig. 2):

$$\hat{D} = -\left(\frac{1}{a}\right) In\left(\frac{y}{n}\right), \quad \hat{\tau} = N\hat{D} \quad y \quad \hat{V}(\hat{D}) = \frac{e^{\hat{D}a} - 1}{na^2}$$

Para obtener estimaciones de la media poblaciónal de la densidad vegetal con dos estimadores insesgados, es mejor utilizar el estimador del muestreo por línea intercepto, ya que su varianza es muy pequeña comparada con la del muestreo por cuadros cargados.

4. Conclusiones

En el muestreo por línea intercepto, es necesario conocer la forma promedio de la corona del individuo de intéres, para saber qué estimador utilizar; el estimador de la media poblacional de la cobertura junto con el de densidad, hacen que este procedimiento de muestreo proporcione más información sobre una población, por lo que su uso es recomendable. El muestreo por cuadros cargados proporciona, con rapidez, la estimación de la media poblacional de la densidad vegetal; sin embargo, es muy alta con respecto a la de la línea intercepto, por lo

que es recomendable un estudio más profundo sobre este procedimiento de muestreo para conocer con amplitud sus aplicaciones.

5. Referencias

Burguete, H. J. F. y A. C. Liz. (1972). Algunas propiedades de los estimadores en muestreo por áreas. *Agrociencia.* **10.9** 1 - 104.

Canfield, H. R. (1941). Application of the line interception method in sampling range vegetation. *J. of Forestry*388-394.

Cochran, G. W. (1950). *Estimation of bacterial densities by means of the most probable number*. Biometrics. 105-115.

Kaiser, L.(1983). *Unbiased estimation in line-intercept sampling.* Biometrics. 39: 965-976.

Kisinger, E. F.; R. E. Eckert and P. O. Currie.(1960). A comparison of the line-interception, variable plot and loop methods as used to measure shrub-crown cover. *J. Range Management* **13**. 17-21.

Lyon, J. (1968). An evaluation of density sampling methods in a shrub community. *J. Range Management* **21**16-20.

McDonald.(1980). Line-intercept sampling for attributes other than converage an density. *J. Wildlife Management.* **44(2)**.530-533.

Swindel, F. B. (1983). *Choice of size and number of quadrats to estimate density from frequency in poisson and binomially dispersed populations.* Biometrics. 39: 455-464.

R: Un ambiente y lenguaje para el cálculo y la graficación estadística

Gabriel Nuñez Antonio
Ernesto Barrios Zamudio

Instituto Tecnológico Autónomo de México

1 Antecedentes

R es un lenguaje y ambiente gratuito para el cálculo y la graficación estadística. Fue desarrollo en 1996, como una implementación del lenguaje S (Bell Labs, 1984–85), explotando el manejo de memoria y la versatilidad de *Scheme* (MIT, 1975-1978). R es actualmente la herramienta de cómputo más usada en la investigación de la estadística a nivel mundial. El grupo principal de desarrollo está formado por estadísticos de primer nivel, incluyendo al autor original del lenguaje S, John Chambers, y los creadores de R, Ross Ihaka y Robert Gentleman.

R es "código abierto" bajo licencia GNU GPL, por lo que su código fuente está disponible. Se tienen además versiones ya compiladas para las plataformas más comunes: MS Windows, Mac OS X, y varias versiones de Linux y Unix, lo que hace su instalación inmediata. Contrario a lo que se pueda pensar, es relativamente fácil de usar. Al ser un lenguaje orientado a objetos ofrece una gran flexibilidad para el análisis y graficación estadística y el desarrollo de nuevas técnicas aún no implementadas. El *Proyecto R* está abierto a contribuciones. Producto de éstas actualmente hay poco más de 1000 paquetes disponibles. Entre ellas varias aplicaciones bayesianas, financieras, genómicas, microarreglos, graficación de mapas, wavelets, etc.

Este trabajo tiene como objetivo mostrar y difundir R como una herramienta de investigación y una buena alternativa para realizar análisis gráficos y estadísticos.

2 Una breve introducción a R

R es un intérprete no un compilador. Esto significa que todos los comandos escritos sobre

la interface se ejecutan inmediatamente sin que se requiera la compilación de un programa como en C, Fortran, Pascal, etc. Una vez que se abre R aparece el prompt de default `">"`, lo que indica que se espera algún comando. En general, éste será una asignación, la evaluación de una función o ambos simultáneamente.

El nombre de un objeto debe comenzar con una letra (`A-Z` ó `a-z`) y además puede incluir dígitos y puntos. R es sensitivo a letras mayúsculas y minúsculas, por lo que `x` y `X` refieren distintos objetos.

Para ejecutar una función se deben incluir los argumentos de ésta entre paréntesis. Si se omiten, R toma los argumentos definidos por omisión. Si la función se invoca sin paréntesis, entonces el código de la función misma será desplegado, lo que permite su personalización. Los argumentos de una función pueden ser en sí objetos (datos, listas, fórmulas, matrices, tablas, etc.)

La forma de asignar objetos en R es a través del símbolo `<-`. Por ejemplo:

```
> x<- 56; n<- sqrt(x); m.aux<-10*n
[1] 56 [1] 7.483315 [1] 74.83315
```

Ayudas

En R se tienen distintos niveles de obtener ayuda mediante las funciones `help`, `help.search`, y `RSiteSearch`. Estas últimas son búsquedas inteligentes. Por ejemplo:

- `help(mean)`. Muestra en línea, entre otras cosas, una descripción de lo que hace la función `mean`; todos los argumentos que acepta ésta; el resultado de haber llamado la función; referencias bibliográficas; funciones relacionadas y ejemplos de su uso.
- `help.search("boxcox")`. Localiza todas las funciones en los paquetes cargados donde se incluya la palabra "boxcox" o "box cox".
- `RSiteSearch("rose diagram")`. Estando en línea, esta función buscará información sobre "rose diagram" (representación gráfica de datos circulares) en todos los paquetes

disponibles en el sitio de R en internet, y también en los mensajes en las listas de discusión de R donde el tema sea mencionado.

Creando sus propias funciones

En R uno puede crear funciones. La sintaxis general para la definición de una función es:

```
function(arguments){expression}
```

donde `arguments` son los argumentos de la función separados por comas y `expression` es cualquier estructura permitida en R, sea un cálculo o graficación. Por ejemplo, la función `grid.calc` calcula la suma de las coordenadas en cada punto de una malla.

```
grid.calc <- function(x,y){
    grid<-matrix(0,length(x),length(y))
    # Define la matriz para almacenar los resultados.
    for(i in 1:length(x)){
        for(j in 1:length(y)) {
            grid[i,j] <- x[i]+y[j]
        }
    }
    grid
}
```

Se incluye el código anterior para efectos de ilustración. Sin embargo, hay que señalar que R permite una programación más eficiente de la función.

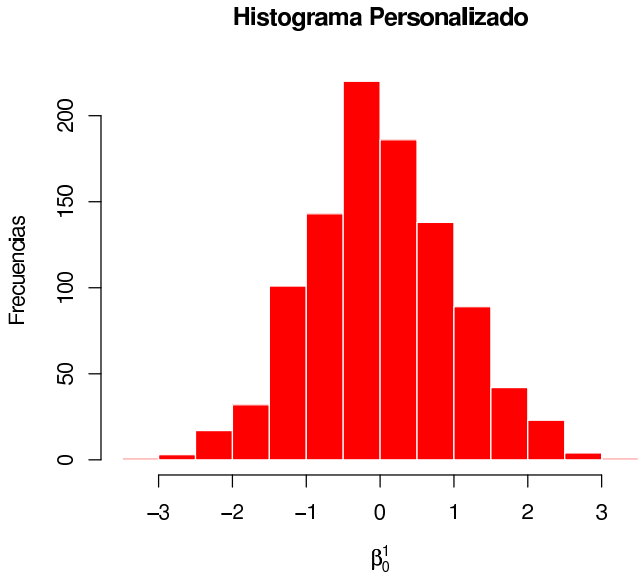
Generación de variables aleatorias

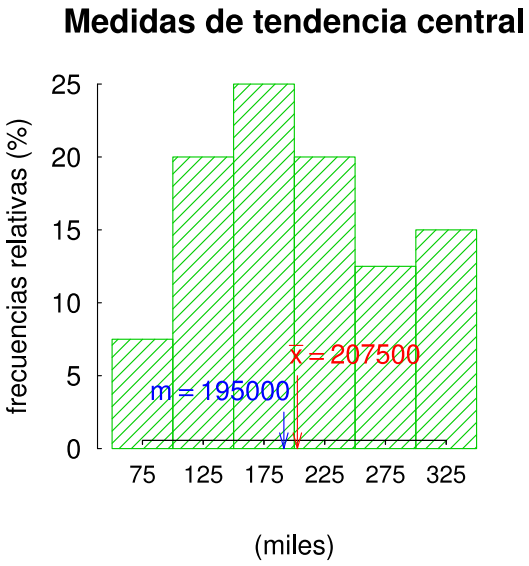
En R es posible generar realizaciones de variable aleatorias para una gran variedad de distribuciones tanto discretas como continuas. También se pueden obtener las correspondientes funciones de densidad, de probabilidad acumulada y cuantiles asociados. Por ejemplo,

`rnorm(n)`, `dnorm(x)`, `pnorm(x)` y `qnorm(p)`, respectivamente, para el caso de la normal estándar.

4 Análisis Gráfico

R ofrece una gran variedad de gráficos además de la posibilidad y flexibilidad de crearlos y personalizarlos. Para darse una idea del potencial gráfico se puede ejecutar el comando `demo(graphics)`. Resulta difícil exponer en este espacio las opciones y posibilidades disponibles para graficación. Como ilustración se presentan un par de gráficas personalizadas.





5 Análisis Estadístico

R ofrece también amplias posibilidades para realizar análisis estadísticos tanto descriptivos como inferenciales. Por ejemplo, ajuste de modelos lineales, lineales generalizados, modelos de supervivencia, de series de tiempo, análisis de datos multivariados, pruebas de hipótesis tanto paramétricas como no-paramétricas, etc. El comando `example(glm)` muestra el ajuste y análisis de varios modelos lineales generalizados.

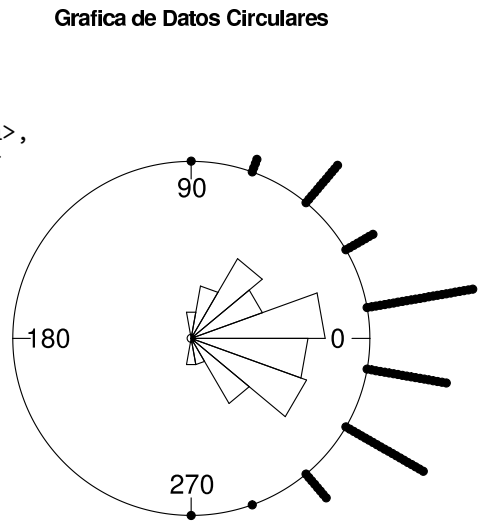
6 Extensiones

Contribución de paquetes

R se beneficia de la contribución de estadísticos de todo el mundo. Al momento de escribir esta nota, el sitio del *Proyecto R* muestra la participación de científicos de más de 50 países y contribuciones de alrededor de 1000 paquetes que enriquecen el software. Por ejemplo, para el análisis de *datos direccionales*, existen los paquetes CircStat y circular. Una ilustración del uso del primero es:

```
Title: Circular Statistics Package:
CircStats Author:
    S-plus original by Ulric Lund <ulund@calpoly.edu>,
    R port by Claudio Agostinelli <claudio@unive.it>
```

```
> install.packages(CircStats)
> library(CircStats)
> data.vm <- rvm(100, 0, 3)
> rose.diag(data.vm, bins = 18,
+ pts = TRUE, shrink=1.5,prop=1.5)
> title("Grafica de Datos Circulares")
```



Comunicación con otros lenguajes

Algunos procedimientos pueden realizarse de manera más eficiente fuera de R, usando Fortran y C. Por un lado, en simulaciones intensivas, es más rápida la ejecución en lenguajes de bajo nivel. Por el otro, se puede aprovechar el uso de programas y paqueterías existentes en estos lenguajes, e. g., IMSL, NAG, etc. Como R puede comunicarse con Fortran y C, resulta aún más flexible y consecuentemente más atractivo.

6 Consideraciones Finales

El equipo de desarrollo de R es de primer nivel estadístico y computacional. Siendo de código abierto se beneficia además de la colaboración de usuarios de todo el mundo. Desde nuestro punto de vista, consideramos que R es no solamente una opción, sino una *buena opción* para la graficación y el análisis estadístico, y una excelente herramienta en el desarrollo de nuevos métodos.

Lecturas Recomendadas

1. Dalgaard, P. (2002). *Introductory Statistics with R*. Springer-Verlag. New York.

2. Ihaka, R. y Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 3, 299-314.

3. *The R Project for Statistical Computing*. URL: `http://www.r-project.org`.

Elasticidades de la demanda por servicio telefónico de larga distancia

Dionicio Morales Ramírez[1]
Universidad Autónoma de Tamaulipas
Daniel Flores Curiel[2]
Universidad Autónoma de Nuevo León
Carmen Zenia Nava Vera
Universidad Autónoma de Tamaulipas

1. Introducción

Las telecomunicaciones son de vital importancia para los individuos y las naciones. Estos servicios permiten a los individuos mantenerse en contacto a pesar de que no exista presencia física, generando importantes ahorros de recursos y facilitando el proceso de las actividades personales, económicas y comerciales, según Kellerman (1992). Por ello, se puede esperar que exista una fuerte relación entre los flujos telefónicos de larga distancia y los movimientos comerciales internacionales.

2. Objetivo

El objetivo principal del presente trabajo es estimar las demandas por servicios internacionales y mundiales de larga distancia en México[3]. En particular, se buscan estimaciones de las elasticidades precio de estas demandas. Además, se trata de establecer el efecto que tienen diversas variables como el PIB, el comercio internacional y las remesas en los minutos de llamadas salientes de México hacia otros países. Para ello, se utilizó series de tiempo con datos trimestrales que abarcan el periodo de 1997 a 2004.

[1] dmorales@uat.edu.mx
[2] danflore@faeco.uanl.mx
[3] Internacionales se refiere a Estados Unidos y Canadá. Mundiales se refiere al resto del mundo.

3. Variables

Se emplea en el estudio un índice de precios real del servicio telefónico de larga distancia (PLD), el Producto Interno Bruto (PIB), exportaciones internacionales (XI), exportaciones mundiales (XM), importaciones internacionales (MI), importaciones mundiales (MM), remesas (R) y turismo (T). Las variables comerciales se encuentran desagregadas por país y, por lo tanto, pueden ligarse con el tráfico telefónico correspondiente. No se pudo hacer lo propio con las variables precio y remesas. Por lo tanto, se incluyeron las variables precio y PIB en ambas ecuaciones y la variable remesas solamente en la ecuación de demanda internacional. Finalmente, el subíndice t de las ecuaciones denota que la observación corresponde al trimestre t.

4. Modelo

Siguiendo el trabajo de Fiebig y Bewley (1987)[4], en el presente estudio se emplea un modelo logarítmico para estimar las funciones de demanda por servicios de larga distancia internacional y mundial. Para realizar el estudio sobre la demanda se emplearon datos de series de tiempo, así como técnicas de regresión basadas en Mínimos Cuadrados Ordinarios (MCO). En particular, las ecuaciones de demanda por minutos internacionales (I) y mundiales (M) que sirven como punto de partida para realizar las estimaciones son las siguientes:

$$lnI_t = \beta_0 + \beta_1 lnPLD_t + \beta_2 lnPIB_t + \beta_3 lnXI_t + \beta_4 lnMI_t + \beta_5 lnR_t + \beta_6 lnT_t + \varepsilon_t \qquad (1)$$

$$lnM_t = \beta_0 + \beta_1 lnPLD_t + \beta_2 lnPIB_t + \beta_3 lnXM_t + \beta_4 lnMM_t + \beta_5 T_t + \varepsilon_t \qquad (2)$$

[4]Estos autores emplean la transformación de Box Cox para estimar la forma funcional del modelo, encontrando que la forma óptima para estimar la función de demanda es una doble logarítmica.

5. Resultados

En el Cuadro 1 se presentan los resultados obtenidos empleando diversos modelos econométricos para estimar la demanda por servicio telefónico internacional. El modelo 1 es prácticamente idéntico al propuesto en la ecuación (1), solamente se agregó un rezago para corregir problemas de autocorrelación y se estimó mediante la opción covarianza consistente de white para corregir la heteroscedasticidad. Sin embargo, este modelo tiene problemas de multicolinealidad. Por ello, se construyeron los modelos 2, 3, 4, 5 y 6. La totalidad de los modelos, excepto el 6, fueron estimados empleando la opción de heterocedaticidad de covarianza consistente de white por la razón antes mencionada.

Considerando que el modelo 6 ofrece mejor ajuste que los otros, se empleó la estimación correspondiente para corroborar que la demanda es elástica mediante las pruebas t y Wald. En ambas pruebas se encontró que la elasticidad precio es significativamente diferente de 1.

En el Cuadro 2 se presentan los resultados obtenidos empleando diversos modelos econométricos para estimar la demanda por servicio telefónico mundial. Los modelos estimados presentan problemas estadísticos similares a los anteriores. Por ello, nuevamente se tuvieron que aplicar pruebas y corregir en caso necesario.

Una vez más se estimó las pruebas de Wald y t, considerando el modelo con mejor ajuste (i.e., el modelo 8), en donde el resultado indica que la elasticidad precio no es diferente de 1.

6. Conclusiones

Los resultados indican que las elasticidades precio estimadas para la demanda por servicio internacional (hacia EU y Canadá) se encuentran entre 1.29 y 1.45, mientras que las elasticidades estimadas para el servicio mundial (hacia el resto del mundo) se encuentran entre 1.35 y 1.63. Sin embargo, a pesar de que aparentemente la elasticidad de la demanda por servicio mundial es mayor que la elasticidad de la demanda por servicio internacional, una vez hechas las pruebas estadísticas correspondientes se pudo establecer que la demanda

Cuadro 1: Modelo minutos internacionales (MI)

Modelos	1	2	3	4	5	6
Variable						
Constante	13.549	12.891	13.605	13.56	12.725	12.999
	(8.353)	(8.420)	(8.563)	(8.576)	(8.044)	(766.443)
PLD	-1.444	-1.336	-1.457	-1.451	-1.297	-1.293
	(-5.273)	(-5.94)	(-5.516)	(5.627)	(-7.007)	(-13.644)
PIB	-0.602		-0.549	-0.58		
	(-1.373)		(-1.277)	(-1.31)		
XI	0.126	-0.171				
	(-0.246)	(-0.337)				
MI	0.016	0.203	0.117			
	(-0.032)	(0.404)	(0.497)			
R	0.033	0.032	0.048	0.04	0.0004	
	(0.298)	(0.293)	(0.645)	(0.541)	(0.006)	
T	0.110	-0.038	0.088	0.1	-0.018	
	(0.586)	(-0.251)	(0.495)	(0.544)	(-0.125)	
CI				0.067	0.028	
				(0.563)	(0.281)	
AR(1)	0.38	0.328	0.389	0.384	0.308	.339
	(1.856)	(1.617)	(1.943)	(1.912)	(1.498)	(1.943)
R ajustada	0.925	0.924	0.928	0.928	0.926	0.937
AIC	-2.676	-2.682	-2.740	-2.742	-2.741	-2.971
SC	-2.302	-2.355	-2.413	-2.415	-2.461	-2.833
F	52.52	60.04	63.9	64.02	74.59	227.08

Nota: el estadístico t se reporta entre paréntesis.

Cuadro 2: Modelo minutos mundiales (MM)

Modelos	1	2	3	4	5	6	7	8
Variable								
Constante	13.306	13.170	10.312	7.059	7.160	9.991	8.604	11.154
	(5.889)	(5.632)	(4.234)	(3.341)	(2.911)	(4.418)	(3.624)	(154.474)
PLD	-1.351	-1.795	-1.415					-1.634
	(-3.878)	(-3.785)	(-2.080)					(-3.568)
PIB	1.010			3.685	0.863			
	(5.048)			(7.738)	(1.458)			
XM	-1.079	-0.276		-1.050		-0.459		
	(-5.805)	(-0.867)		(-3.577)		(-1.258)		
MM	0.121					0.586		
	(0.871)					(2.372)		
T	-0.257	-0.093	-0.028	-0.895	-0.123	-0.031	0.067	
	(1.132)	(-0.449)	(-0.136)	(-3.611)	(-0.554)	(-0.169)	(0.370)	
CM			0.092		0.108		0.189	
			(0.569)		(0.651)		1.171)	
AR(1)	1.395	0.689	0.745	0.539	0.887	0.852	0.899	0.711
	(6.939)	(5.176)	(5.709)	(2.480)	(10.102)	(8.036)	(11.661)	(5.553)
AR(2)	-1.388			0.033				
	(-4.415)			(0.143)				
AR(3)	0.962			-0.271				
	(3.039)			(-1.461)				
AR(4)	-0.448							
	(-2.401)							
R ajustada	0.933	0.869	0.867	0.835	0.874	0.880	0.869	0.875
AIC	-2.106	-1.474	-1.456	-1.271	-1.512	-1.564	-1.499	-1.573
SC	-1.630	-1.243	-1.225	-0.941	-1.281	-1.333	-1.314	-1.434
F	42.83	50.872	49.827	24.688	53.085	56.280	67.315	105.843

Nota: el estadístico t se reporta entre paréntesis.

por servicio internacional es elástica mientras que la demanda por servicio mundial no es significativamente diferente de 1.

El PIB, que se puede interpretar como una medida de ingreso, solamente tuvo un efecto positivo y significativo en algunos modelos de demanda por servicio telefónico mundial. Además, esta variable no resultó importante para explicar los cambios en la demanda por servicio telefónico internacional. Por lo anterior, no se puede aseverar que los servicios telefónicos de larga distancia internacional o mundial sean bienes normales.

El resto de las variables incluidas en el estudio no fueron relevantes para explicar la demanda por servicio telefónico de larga distancia internacional o mundial una vez que se incluye el precio como variable explicativa.

7. Bibliografía

Fiebig, D. y R. Bewley (1987). International telecommunications forecasting: an investigation of alternative functional forms, In *Applied Economics* **19**, 949-60.

Gujarati, D. (2003). *Econometría*. México: McGraw Hill.

Kellerman, A. (1992). US international telecommunications, 1961-88: an international movement model, In *Telecommunications Policy* **16**, 401-414.

Muestreo por seguimiento de nominaciones: estimación de medias y totales de poblaciones de difícil detección[1]

Martín H. Félix Medina [2]
Escuela de Ciencias Físico-Matemáticas de la Universidad Autónoma de Sinaloa
Pedro E. Monjardin[3]
Escuela de Ciencias Físico-Matemáticas de la Universidad Autónoma de Sinaloa

1. Introducción

El Muestreo por Seguimiento de Nominaciones (denominado en Inglés como Link-tracing sampling o Snowball sampling) es un método que se ha propuesto para muestrear poblaciones de difícil detección, tales como poblaciones de drogadictos, niños de la calle, trabajadoras sexuales, etc. En este método se selecciona una muestra inicial de miembros de la población de interés, y se les pide a las personas que fueron seleccionadas que nominen a otros miembros de la población objetivo. A las personas que fueron nominadas se les puede pedir que nominen a otras personas, y el proceso de nominación puede continuar de esta manera hasta que se satisfaga alguna regla de terminación del muestreo. Para una revisón y discusión sobre este método ver Thompson and Frank (2000).

Félix Medina y Thompson (2004) desarrollaron una variante de este tipo de muestreo en la cual la muestra inicial es una muestra aleatoria simple de sitios, tales como parques, hospitales y cruceros de calles, que se selecciona de un marco muestral que sólo cubre una parte de la población de interés. Ellos propusieron estimadores máximo verosímiles del tamaño poblacional, y posteriormente, Félix Medina y Monjardin (2006) propusieron estimadores

[1]Trabajo realizado con apoyos parciales de los proyectos PIFI-2003-25-28 de la SEP y PAFI-UAS-2002-I-MHFM-06 de la UAS

[2]mhfelix@uas.uasnet.mx
[3] pemo@uas.uasnet.mx

del tamaño poblacional derivados bajo el enfoque Bayesiano, pero realizaron inferencias bajo un enfoque frecuentista basado en el diseño muestral.

En este trabajo consideramos el problema de estimar la media y/o el total poblacional de una variable respuesta, tal como gasto en drogas, gasto en medicamentos y edad. Proponemos estimar estos parámetros mediante estimadores tipo Horvitz-Thompson cuyo desempeño analizamos mediante un estudio de simulación.

2. Diseño muestral y notación

El diseño muestral que consideramos en este trabajo es el propuesto por Félix Medina y Thompson (2004). Así, supondremos que una parte U_1 de la población de interés U está cubierta por un marco muestral de N sitios $A_1, \ldots, A_N$, tales como parques, hospitales o cruceros de calles. De este marco se selecciona una muestra aleatoria simple sin reemplazo $S_0 = \{A_1, \ldots, A_n\}$ de n sitios, y a las personas de la población de interés que pertenecen al sitio seleccionado se les pide que nominen a otros miembros de la población. Como convención, diremos que una persona es nominada por un sitio si cualquiera de los miembros de ese sitio la nomina.

Denotaremos por τ el tamaño de U, por τ_1 el de U_1, por $\tau_2 = \tau - \tau_1$ el de $U_2 = U - U_1$, y por m_i el número de personas en A_i. Los conjuntos de variables $\{X_{ij}^{(1)}\}$ y $\{X_{ij}^{(2)}\}$ indicarán el proceso de nominación. Así, $X_{ij}^{(1)} = 1$ si la persona $j \in U_1 - A_i$ es nominada por el sitio A_i, y $X_{ij}^{(1)} = 0$ en otro caso. Similarmente, $X_{ij}^{(2)} = 1$ si la persona $j \in U_2$ es nominada por el sitio A_i, y $X_{ij}^{(2)} = 0$ en otro caso. La probabilidad de que la persona j en $U_1 - A_i$ sea nominada por el sitio A_i (llamada probabilidad de nominación) está dada por $p_i^{(1)} = \Pr(X_{ij}^{(1)} = 1)$, $j \in U_1 - A_i$. Similarmente $p_i^{(2)} = \Pr(X_{ij}^{(2)} = 1)$, $j \in U_2$. Denotaremos por $y_j^{(k)}$ el valor de la variable respuesta y asociado con la j-ésima persona en U_k, $k = 1, 2$. Finalmente, Y_k y $\bar{Y}_k$ denotarán el total y la media de los valores $y_j^{(k)}$, $j = 1, \ldots, \tau_k$, y Y y $\bar{Y}$ el total y la media de todos los valores y_j, $j = 1, \ldots, \tau$.

3. Estimadores del total y la media poblacional

El primer paso en la estimación de totales y medias poblacionales es la estimación de los tamaños poblacionales τ_1, τ_2 y τ, y de las probabilidades de nominación $p_i^{(1)}$ y $p_i^{(2)}$, $i = 1, \ldots, n$. Así, denotaremos por $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\tau}$, $\hat{p}_i^{(1)}$ y $\hat{p}_i^{(2)}$, $i = 1, \ldots, n$, ya sea los estimadores máximo verosímiles propuestos por Félix Medina y Thompson (2004) o los estimadores derivados bajo el enfoque Bayesiano y propuestos por Félix Medina y Monjardin (2006).

Un estimador tipo Horvitz-Thompson de Y_k es

$$\hat{Y}_k = \sum_{j \in S_k} \frac{y_j^{(k)}}{\hat{\pi}^{(k)}} = \frac{1}{\hat{\pi}^{(k)}} \sum_{j \in S_k} y_j^{(k)}, \qquad k = 1, 2,$$

donde S_k denota los elementos de U_k, $k = 1, 2$, contenidos en la muestra, y

$$\hat{\pi}^{(k)} = \begin{cases} 1 - (1 - n/N) \prod_{i=1}^{n} \left(1 - \hat{p}_i^{(1)}\right), & k = 1 \\ 1 - \prod_{i=1}^{n} \left(1 - \hat{p}_i^{(2)}\right), & k = 2 \end{cases}$$

Claramente, un estimador de Y es $\hat{Y} = \hat{Y}_1 + \hat{Y}_2$.

Las varianzas de estos estimadores se pueden estimar mediante estimadores tipo Horvitz-Thompson. Aunque ya contamos con expresiones para estos estimadores de varianza, por limitaciones de espacio no las presentamos.

Estimadores de las medias poblacionales $\bar{Y}_k$ y $\bar{Y}$ son $\hat{\bar{Y}}_k = \hat{Y}_k/\hat{\tau}_k$, $k = 1, 2$, y $\hat{\bar{Y}} = \hat{Y}/\hat{\tau}$. Obsérvese que estos estimadores son estimadores de razón, y al igual que las varianzas de los estimadores de totales, sus varianzas se pueden estimar mediante estimadores de varianza tipo Horvitz-Thompson.

4. Estudio Monte Carlo

Para realizar este estudio generamos dos poblaciones de $N = 250$ valores m_i. En la Población I, los m_i los generamos con la distribución Poisson truncada en cero con media 7.2 y varianza 7.17, mientras que en la Población II, con la distribución Binomial negativa truncada en cero con media 7.2 y varianza 2.4. En la Población I obtuvimos $\tau_1 = \sum_1^N m_i = 1897$, y en la Población II, $\tau_1 = 1764$. En ambos casos τ_2 lo fijamos en 700. Los valores $y_j^{(k)}$, $j = 1, \ldots, \tau_k$, $k = 1, 2$, los generamos con la distribución exponencial con media 1. Así, en la Población I obtuvimos Y_1 =1933.8, Y_2 =730.3, mientras que en la Población II, Y_1 =1802.9, Y_2 =679.9. Las probabilidades de nominación las generamos con el modelo $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, donde los valores de β_k los fijamos de tal manera que obtuvimos dos casos. Caso 1: $\left(\bar{p}^{(1)}), \bar{p}^{(2)}\right)$ ≈(0.05,0.03) y Caso 2: $\left(\bar{p}^{(1)}), \bar{p}^{(2)}\right)$ ≈(0.01,0.006). Consideramos dos conjuntos de estimadores de medias y totales poblacionales. Los estimadores $\tilde{Y}_1$, $\tilde{Y}_2$ y $\tilde{Y}$, los cuales obtuvimos a partir de los estimadores máximo verosímiles $\tilde{\tau}_1$, $\tilde{\tau}_2$ y $\tilde{\tau}$ propuestos por Félix Medina y Thompson (2004) y los estimadores $\hat{Y}_1$, $\hat{Y}_2$ y $\hat{Y}$, los cuales obtuvimos a partir de los estimadores $\hat{\tau}_1$, $\hat{\tau}_2$ y $\hat{\tau}$ derivados por Félix Medina y Monjardin (2006) bajo el enfoque Bayesiano y con distribuciones iniciales Poisson- Gamma de los τ_k. Los valores de los parámetros de las distribuciones iniciales los fijamos al igual que en la referencia anterior.

En la Tabla 1 presentamos los resultados de los estimadores de las medias y totales poblacionales. Los resultados muestran que los estimadores de los totales tienen buenos desempeños en el Caso 1, pero en el Caso 2 los estimadores $\tilde{Y}_2$ y $\tilde{Y}$ tienen pésimos desempeños. Esto se debe a los malos desempeños de los estimadores $\tilde{\tau}_2$ y $\tilde{\tau}$. Sin embargo, los estimadores $\hat{Y}_k$ y $\hat{Y}$ tienen desempeños aceptables. En el caso de los estimadores de las medias, ambos tipos de estimadores tienen buenos desempeños en el Caso 1, y desempeños aceptables en el Caso 2.

También observamos los desempeños de los estimadores de las varianzas de $\tilde{Y}_k$ y $\tilde{Y}$, así como de los intervalos de confianza tipo Wald del 95 % para los totales poblacionales basados en estos estimadores, esto es, intervalos de la forma $\tilde{Y}_k \pm \sqrt{\hat{\mathbf{V}}(\tilde{Y}_k)}$. Por restricciones de espacio no presentamos los resultados, pero los desempeños, tanto de los estimadores de varianza

como de los intervalos de confianza, fueron consistentes con el desempeño del correspondiente estimador del total poblacional. Esto es, en las situaciones en las que un estimador del total poblacional mostró buen (mal) desempeño también mostraron buenos (malos) desempeños el correspondiente estimador de varianza y el correspondiente intervalo de confianza.

Tabla 1. Sesgos relativos y raíces cuadradas de errores cuadráticos medios relativos de estimadores de totales y medias poblacionales. Resultados basados en 1000 iteraciones.

	Población I				Población II			
	Caso 1 $\bar{p}^{(1)} \approx ,05$ $\bar{p}^{(2)} \approx ,03$		Caso 2 $\bar{p}^{(1)} \approx ,05$ $\bar{p}^{(2)} \approx ,006$		Caso 1 $\bar{p}^{(1)} \approx ,05$ $\bar{p}^{(2)} \approx ,03$		Caso 2 $\bar{p}^{(1)} \approx ,05$ $\bar{p}^{(2)} \approx ,006$	
	sesgo-rel	$\sqrt{\text{ecm-rel}}$	sesgo-rel	$\sqrt{\text{ecm-rel}}$	sesgo-rel	$\sqrt{\text{ecm-rel}}$	sesgo-rel	$\sqrt{\text{ecm-rel}}$
$\tilde{Y}_1$	-0.00	0.02	-0.00	0.06	-0.00	0.03	-0.01	0.09
$\tilde{Y}_2$	0.00	0.07	L	L	0.00	0.08	L	L
$\tilde{Y}$	-0.00	0.03	L	L	-0.00	0.03	L	L
$\tilde{\tilde{Y}}_1$	-0.00	0.01	-0.00	0.03	-0.00	0.02	0.00	0.04
$\tilde{\tilde{Y}}_2$	0.00	0.04	-0.00	0.10	-0.00	0.03	-0.00	0.09
$\tilde{\tilde{Y}}$	-0.00	0.01	0.00	0.04	-0.00	0.01	-0.00	0.04
$\hat{Y}_1$	-0.00	0.02	-0.00	0.06	-0.00	0.03	-0.01	0.09
$\hat{Y}_2$	0.00	0.07	0.01	0.22	0.00	0.07	-0.00	0.22
$\hat{Y}$	-0.00	0.02	-0.00	0.08	-0.00	0.03	-0.01	0.09
$\hat{\hat{Y}}_1$	0.00	0.01	0.00	0.03	0.00	0.02	0.00	0.04
$\hat{\hat{Y}}_2$	0.00	0.04	-0.00	0.10	-0.00	0.03	-0.00	0.09
$\hat{\hat{Y}}$	0.00	0.01	0.00	0.04	-0.00	0.01	0.00	0.04

Notas: sesgo-rel=sesgo relativo; ecm-rel=error cuadrático medio relativo; $\tilde{Y}_k$ y $\tilde{\tilde{Y}}_k$, estimadores máximo verosímiles; $\hat{Y}_k$ y $\hat{\hat{Y}}_k$, estimadores bayesianos; L indica un valor mayor que 10^4.

5. Referencias

Félix-Medina, M.H., and Thompson, S.K. (2004). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics*,

20, 19-38.

Félix-Medina, M.H., and Monjardin, P.E. (2006). Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population: a Bayesian assisted approach. *Survey Methodology*, **32**, 187-195.

Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, **26**, 87-98.

Constrained linear regression models*

Gabriel Rodriguez-Yam**
Universidad Autónoma Chapingo
Richard A. Davis
Department of Statistics, Colorado State University, Fort Collins, Colorado
Louis L. Scharf
Departments of Electrical and Computer Engineering and Statistics
Colorado State University, Fort Collins, Colorado

1. Introduction

In this paper a linear regression model in which the regression parameters are subject to linear constraints of inequality and equality is considered. The motivation behind this line of research was an identification problem in hyperspectral imaging. In this problem, the spectrum $\mathbf{y}$ of a composite substance in a pixel can be represented as a linear combination of component spectra, i.e.,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where the columns of the full rank matrix $\mathbf{X}$ contain the spectra of the k materials in a pixel, $\boldsymbol{\beta}$ is a vector consisting of the "abundances" of the materials in the pixel, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is the noise of the model (see Manolakis and Shaw, 2002). Due to physical considerations, the abundance parameters are considered to be *non-negative*, i.e., $\beta_j \geq 0$, $j = 1, \ldots, k$ and satisfy the *sum-to-one* constraint $\beta_1 + \ldots + \beta_k = 1$. This model fits into a more general framework, where the vector of regression coefficients $\boldsymbol{\beta}$ from the linear regression in (1) is subject to a set of linear constraints given by $\mathbf{B}\boldsymbol{\beta} \leq \mathbf{b}$, and $\mathbf{C}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{B}$ and $\mathbf{C}$ are known matrices and $\mathbf{b}$ and $\mathbf{c}$ are known vectors. Judge and Takayama (1966) and Liew (1976) give the

*This work was supported in part by Colorado Advanced Software Institute (CASI) and Data Fusion Corporation (Scharf and Rodriguez-Yam) and NSF grant DMS-0308109 (Davis). It also forms part of the PhD dissertation of the first author, who received a scholarship from Consejo Nacional de Ciencia y Tecnologia (CONACYT).

**grodrigu@correo.chapingo.mx

inequality constrained least-squares (ICLS) estimate of $\boldsymbol{\beta}$ using the Dantzig-Cottle algorithm. The ICLS estimator reduces to the ordinary least squares estimator for a sufficiently large sample. Conditioning on knowledge of which constraints are binding and which are not, they compute an untruncated covariance matrix of the ICLS estimator. Geweke (1986) points out that this variance matrix is incorrect, since in practice it is not known ahead of time which constraints will be binding. The case when the vector of regression coefficients $\boldsymbol{\beta}$ from the linear regression in (1) is subject to a set of inequality linear constraints given by

$$\mathbf{B}\boldsymbol{\beta} \leq \mathbf{b}, \tag{2}$$

has been analyzed from the Bayesian perspective. Gelfand et al. (1992) suggest an approach based on a Monte Carlo Markov chain (MCMC) technique to routinely analyze problems with constrained parameters using the Gibbs sampler. Let $\mathcal{D}$ denote the data and $\boldsymbol{\theta}$ a parameter vector with some prior distribution. Suppose it is difficult or virtually impossible to draw samples from the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$. The Gibbs sampler, introduced by Geman and Geman (1984) in the context of image restoration, provides a method for generating samples from $p(\boldsymbol{\theta}|\mathcal{D})$. Suppose $\boldsymbol{\theta}$ can be partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$, where the $\boldsymbol{\theta}_i$'s are either uni- or multidimensional and that we can simulate from the conditional posterior densities $p(\boldsymbol{\theta}_i|\mathcal{D}, \boldsymbol{\theta}_j, j \neq i)$. The Gibbs sampler generates a Markov chain by cycling through $p(\boldsymbol{\theta}_i|\mathcal{D}, \boldsymbol{\theta}_j, j \neq i)$. In each cycle, the most recent information updates the posterior conditionals. Starting from some $\boldsymbol{\theta}^{(0)}$, after t cycles we have a realization $\boldsymbol{\theta}^{(t)}$ that under regularity conditions (Gelfand and Smith, 1990), approximates a drawing from $p(\boldsymbol{\theta}|\mathcal{D})$ for large t. Roberts (1996), Gilks and Roberts (1996) comment that the rate of convergence depends on the posterior correlation between the components in the vector $\boldsymbol{\theta}$. Geweke (1996) applies this procedure to the problem of linear regression when the inequality linear constraints in (2) are linearly independent. However, this implementation may suffer from poor mixing. Due to the requirement of independent constraints, the number of constraints can not exceed the number of parameters. Also, equality linear constraints are not considered. In Rodriguez-Yam et al. (2002), a Gibbs sampler implementation with good mixing is provided for the hyperspectral imaging problem when only the non-negativity constraints on the abundance parameters are considered. For this case, the constraints are linearly independent and the number of inequality linear constraints coincides with the number of regression coefficients. In this paper a new implementation of the Gibbs sampler for this constrained regression problem is proposed. The formulation of this implementation can cope with inequality lin-

ear constraints that are linearly dependent; constraints whose number exceeds the parameter dimension; and equality linear constraints. Furthermore, this implementation has faster mixing, requiring substantially fewer iterations of the Markov chain than previously published Gibbs sampler implementations. The organization of this paper is as follows. In Section 2 we provide a Bayesian framework for linear regression where the regression parameters are subject to the constraints in (2) and we present a new implementation of the Gibbs sampler to this model. In Section 3 this procedure is applied to a dataset consisting of aggregate data involving smokers preferences of three leading brands of cigarettes. For this example, equality linear constraints are needed in addition to inequality linear constraints and the number of inequality linear constraints exceeds the number of regression coefficients. Section 4 contains a summary of our findings.

2. Bayesian Constrained Regression

In this section we construct a Bayesian model for the linear regression given in (1) where the parameters satisfy the constraints in (2). The likelihood can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) &= (2\pi\sigma^2)^{-n/2} \exp\{-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/(2\sigma^2)\}, \quad (3) \\ &= (2\pi\sigma^2)^{-n/2} \exp\{-(n-k)\hat{\sigma}^2/(2\sigma^2) - (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/(2\sigma^2)\}, \quad (4) \end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the ordinary least squares estimates of (the unconstrained) $\boldsymbol{\beta}$ and σ^2 respectively. Now assume the "non-informative" prior for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, given by $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$, $\mathbf{B}\boldsymbol{\beta} \leq \mathbf{b}$. Thus,

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto \mathrm{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) p(\boldsymbol{\beta}, \sigma^2). \quad (5)$$

To sample from the posterior $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ we use the Gibbs sampler. To start, from (4) and (5), we obtain

$$\boldsymbol{\beta}|(\sigma^2; \mathbf{y}) \sim N(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}), \quad \mathbf{B}\boldsymbol{\beta} \leq \mathbf{b}, \quad (6)$$

while from (3) and (5), we have

$$S(\boldsymbol{\beta})\sigma^{-2}|(\boldsymbol{\beta}; \mathbf{y}) \sim \chi^2_n, \quad (7)$$

where $S(\boldsymbol{\beta}) := (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, and χ^2_n denotes a Chi-squared distribution with n degrees of freedom. Now, let $\mathbf{A}$ be a non-singular matrix for which $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A} = \mathbf{I}$, and

set $\boldsymbol{\eta} := \mathbf{A}\boldsymbol{\beta}$. Define $\mathbf{D} := \mathbf{B}\mathbf{A}^{-1}$ and $\boldsymbol{\alpha} := \mathbf{A}\hat{\boldsymbol{\beta}}$. Then, from (6), we obtain

$$\boldsymbol{\eta}|(\sigma^2, \mathbf{y}) \sim N(\boldsymbol{\alpha}, \sigma^2\mathbf{I}), \quad \mathbf{D}\boldsymbol{\eta} \leq \mathbf{b}. \tag{8}$$

Let $\boldsymbol{\eta}_{-j}$ denotes the vector $[\eta_1, \ldots, \eta_{j-1}, \eta_{j+1}, \ldots, \eta_k]^T$, and $\mathbf{D}_{-j}$ denotes the matrix obtained from $\mathbf{D} = [\mathbf{d}_1 \ldots \mathbf{d}_k]$ by removing the j-th column $\mathbf{d}_j$, then from (8), $\eta_j|(\boldsymbol{\eta}_{-j}, \sigma^2, \mathbf{y}) \sim N(\alpha_j,\ \sigma^2)$, where the random variable η_j is subject to the constraints

$$\mathbf{d}_j\eta_j \leq \mathbf{b} - \mathbf{D}_{-j}\boldsymbol{\eta}_{-j}. \tag{9}$$

Since the constraints on $\boldsymbol{\eta}$ form a convex subset of $\Re^k$, the solution of the inequalities in (9) can be written as one of the intervals $l_j \leq \eta_j \leq u_j$, $-\infty < \eta_j \leq u_j$ or $l_j \leq \eta_j < +\infty$. The values l_j and u_j can be easily obtained from the set of one-dimensional inequalities in (9).

Thus, the next component $\boldsymbol{\theta}^{(t+1)} = (\eta_1^{(t+1)}, \ldots, \eta_k^{(t+1)}, \sigma^{2(t+1)})$ based on the current path $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(1)}$, …, $\boldsymbol{\theta}^{(t)}$ of the Gibbs sampler is computed as follows • For $j = 1, \ldots, k$ generate $\eta_j^{(t+1)}$ from $p(\eta_j|\eta_1^{(t+1)}, \ldots, \eta_{j-1}^{(t+1)}, \eta_{j+1}^{(t)}, \ldots, \eta_k^{(t)}, \sigma^{2(t)}, \mathbf{y})$. • Generate $\sigma^{2(t+1)}$ from $p(\sigma^2|\boldsymbol{\eta}^{(t+1)}, \mathbf{y})$.

3. Example: Application to the cigarette-brand preference data

This example considers the estimation of the transition probability matrix of a finite Markov process when only the time series of the proportion of visits to each state is known. The numerical example given by Telser (1963) and Jugde and Takayama (1966) consists of the annual sales in billions of cigarettes for the three leading brands from 1925 to 1943. Given the time ordered market shares of these brands and assuming that the probability of a transition, p_{ij}, from brand i to brand j is constant over time, Telser gives the regression models

$$y_{jt} = \sum_{i=1}^{3} y_{i,t-1}p_{ij} + u_{jt}, \quad j = 1, 2, 3, \tag{10}$$

where y_{jt} is the proportion of individuals in state j at time t and u_{jt}, $t = 1, \ldots, T$ are independent errors. The probabilities p_{ij} are subject to the constraints

$$\sum_{j=1}^{3} p_{ij} = 1, \quad \text{for all } i, \tag{11}$$

$$p_{ij} \geq 0, \quad \text{for all } i \text{ and } j. \tag{12}$$

For this data set, the three models in (10) can be combined as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{W} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}, \tag{13}$$

where $\mathbf{y}_j := [y_{j2}, \ldots, y_{jT}]^T$, $\mathbf{W}$ is the common design matrix of dimension $3 \times T - 1$ from the models in (10), $\mathbf{p}_j$ is the j-th column of the probability transition matrix $\mathbf{P}$ of the finite Markov process, and $\mathbf{u}_j$ is the vector of errors from the model in (10). To handle the equality constraints in (11), denote by $\mathbf{y}$ the response vector of the full model in (13), by $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ the matrices having the columns 1 through 3, 4 through 6 and 7 through 9, respectively of the design matrix in (13). Substituting $p_{i3} = 1 - p_{i1} - p_{i2}$, $i = 1, 2, 3$, in this model, we obtain

$$\mathbf{y} - \mathbf{W}_3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = [\mathbf{W}_1 - \mathbf{W}_3 \quad \mathbf{W}_2 - \mathbf{W}_3] \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{bmatrix} + \mathbf{u}, \tag{14}$$

subject to the constraints

$$p_{i1} + p_{i2} \leq 1, \quad i = 1, 2, 3, \tag{15}$$

$$p_{ij} \geq 0, \quad i = 1, 2, 3, \ j = 1, 2, \tag{16}$$

where $\mathbf{u}$ is the vector of errors from the model in (13). In their method, Judge and Takayama (1966) assumed that $\text{var}(\mathbf{u}) = \sigma^2\mathbf{I}$. For simplicity we also assume that $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Notice that the number of constraints in (15) and (16) to the regression model in (14) exceeds the number of regression coefficients. A path of length 5000 for the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ was generated using the Gibbs sampler described in Section 2. Based on the last 2500 iterates of this sample, the estimate $\hat{\mathbf{P}}$ of the probability transition matrix and the matrix

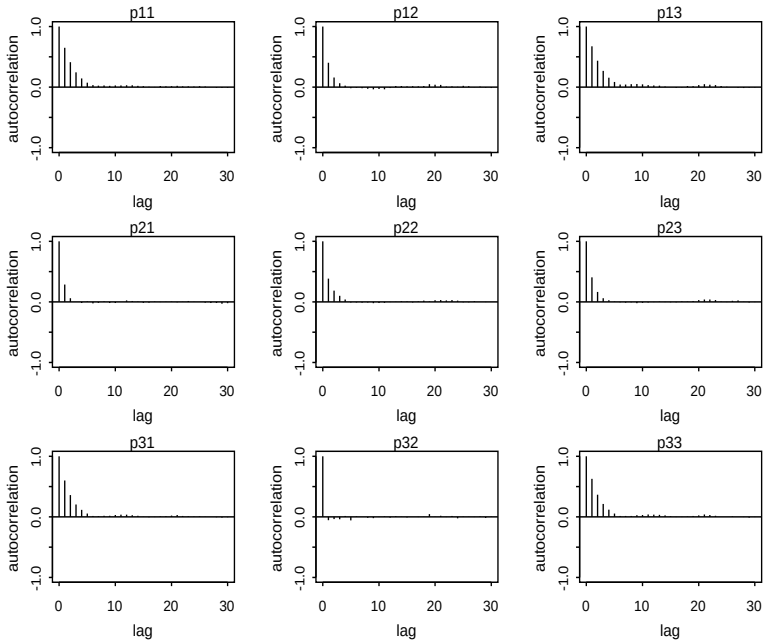


Figura 1: Autocorrelation plots of the components of the transition probability matrix $\mathbf{P}$ of the cigarettes data obtained with a Gibbs path of length 5000.

$\hat{\sigma}_{\hat{\mathbf{P}}}$ having in its entries the estimated standard error of each component of $\hat{\mathbf{P}}$ are

$$\hat{\mathbf{P}} = \begin{bmatrix} 0{,}692 & 0{,}116 & 0{,}193 \\ 0{,}033 & 0{,}848 & 0{,}119 \\ 0{,}334 & 0{,}058 & 0{,}608 \end{bmatrix}, \quad \hat{\sigma}_{\hat{\mathbf{P}}} = \begin{bmatrix} 0{,}0018 & 0{,}0009 & 0{,}0017 \\ 0{,}0005 & 0{,}0008 & 0{,}0009 \\ 0{,}0025 & 0{,}0010 & 0{,}0025 \end{bmatrix}. \tag{17}$$

The restricted least-squares estimates obtained by Judge and Takayama (1966) are given by

$$\hat{\mathbf{P}} = \begin{bmatrix} 0{,}6686 & 0{,}1423 & 0{,}1891 \\ 0 & 0{,}8683 & 0{,}1317 \\ 0{,}4019 & 0 & 0{,}5981 \end{bmatrix}. \tag{18}$$

The estimates in (17) differ slightly from the restricted least-squares in (18). Perhaps the most important difference is the fact that the estimates of p_{21} and p_{32} are non zero. The zero estimates of the elements of $\mathbf{P}$ can induce misleading interpretations. The autocorrelations of the components of the matrix $\mathbf{P}$ obtained with the Gibbs sample are shown in Figure 1. We observe a fast decay on these autocorrelations and following Chen, et al. (2000), we expect good mixing and fast convergence.

4. Conclusions

In this paper, a Bayesian analysis of a linear regression model where the parameters are subject to inequality linear constraints has been considered. Our method is based on a Gibbs sampler for an "orthogonal" transformation of the vector of regression coefficients. This sampler mixes fast, a property that is not always enjoyed by other implementations (see Rodriguez-Yam, 2003) and can cope with non-standard situations such as when the constraints are linearly dependent and when the number of constraints exceed the number of regression coefficients. We have shown with an example how to manage equality linear constraints in addition to inequality linear constraints; a case in which other implementations do not apply.

5. References

Chen, M-H. and Shao, Q-M. and Ibrahim, J. G. (2000). "Monte Carlo Methods in Bayesian Computation." Springer, New York, 2000.

Chen, M-H. and Deely, J. J. (1996) "Bayesian Analysis for a Constrained Linear Multiple Regression Problem for Predicting the New Crops of Apples," *J. Agric. Biol. Environ. Stat.*, **1**, 467-89.

Gelfand, A. E. and Smith, A. F. M. (1990) "Sampling-based Approaches to Calculating Marginal Densities," *J. Amer. Statist. Assoc.*, **85**, 398-409.

Gelfand, A. E., Smith, A. F. M. and Lee, T. M. (1992) "Bayesian Analysis of Constrained Parameters and Truncated Data Problems." *J. Amer. Statist. Assoc.*, **87**, 523-532.

Geman, S. and Geman, D. (1984) "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE trans. pattern anal. mach. intell*, **6**, 721-741.

Geweke, J. (1986) "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *J. Appl. Econ.*, **1**, 127-141.

Geweke, J. (1996) "Bayesian Inference for Linear Models Subject to Linear Inequality Constraints," In: Zellner, A., Lee, J. S. (Eds.), Modeling and Prediction: Honouring Seymour Geisser. Springer, New York.

Gilks, W. R. and Roberts, G. O. (1996) "Strategies for Improving MCMC," In Markov Chain Monte Carlo in Practice (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), Chapman & Hall, London, 89-114.

Jugdge, G. C. and Takayama, T. (1966) "Inequality Restrictions In Regression Analysis," *J. Amer. Statist. Assoc.*, **61**, 166-181.

Liew, C. K. (1976) "Inequality Constrained Least-Squares Estimation," *J. Amer. Statist. Assoc.*, **71**, 746-751.

Manolakis, D. and Shaw, G. (2002) " Detection Algorithms for Hyperspectral Imaging Applications," *IEEE Signal Processing Magazine*, **19**, 29-43.

Roberts, G. O. (1996) "Markov Chain Concepts Related to Sampling Algorithms." In Markov Chain Monte Carlo in Practice (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 45-57. London: Chapman & Hall.

Rodriguez-Yam, G. A., Davis, R. A. and Scharf, L. L. (2002) "A Bayesian Model and Gibbs Sampler for Hyperspectral Imaging," Proc. 2002 IEEE Sensor Array and Multichannel Signal Processing Workshop, Washington, D.C. 105-109.

Rodriguez-Yam, G. A. (2003) "Estimation for State-Space Models and Bayesian Regression Analysis with Parameter Constraints," Ph.D. Dissertation, Deparment of Statistics, Colorado State University, USA.

Telser, L. G. (1963) "Least Squares Estimates of Transition Probabilities," in Christ, C. F. and others (Eds.), Measurement in Economics: Studies in Mathematical Economics and Econometrics: In memory of Yehuda Grunfeld. Stanford University Press, Stanford.

Análisis de datos de suelos forestales en la caldera de Teziutlán, Puebla, por componentes principales y técnicas geoestadísticas

Gladys Linares Fleites[1], **Miguel Angel Valera Pérez**
Departamento de Investigaciones en Ciencias Agrícolas. Instituto de Ciencias de la Benemérita Universidad Autónoma de Puebla.
Maribel Castillo Morales
Estudiante Posgrado Ciencias Ambientales. Instituto de Ciencias de la Benemérita Universidad Autónoma de Puebla.

1. Introducción

Como resultado del aumento de concentraciones de gases de efecto de invernadero, existen evidencias científicas que sugieren que el clima global se verá alterado en este siglo. El mayor responsable del cambio climático global es el CO_2.

Los ecosistemas forestales pueden absorber cantidades significativas de CO_2, por lo que hay un gran interés por incrementar el contenido de carbono en estos ecosistemas, lo que se conoce como secuestro de carbono. A pesar de la importancia del secuestro de carbono, su evaluación se encuentra muy limitada en estos suelos.

El objetivo de este trabajo es estudiar el secuestro de carbono por suelos forestales en la Caldera de Teziutlán, Puebla, y establecer una metodología para la evaluación del secuestro de carbono en los suelos forestales.

A continuación se desarrolla la metodología empleada (Linares, 2004). Inicialmente se caracteriza la zona de estudio, posteriormente se realiza un estudio exploratorio de datos en sus aspectos univariado, bivariado y multivariado y finalmente se lleva a cabo el análisis geoestadístico.

[1]gladys.linares@icbuap.buap.mx

2. Características de la zona de estudio

El estudio se ha realizado en los suelos de la Caldera de Teziutlán situada en la porción nororiental del estado de Puebla, entre los paralelos $19^o43'30"$ y $20^o14'54"$ de latitud norte y los meridianos $97^o07'42"$ y $97^o43'30"$ de longitud occidental. Estos suelos, derivados de material piroclástico, se presentan cubriendo una superficie de 846 Km^2.

Fueron identificados como Andisoles y la vegetación corresponde a Bosques de Pino. El análisis fisicoquímico del suelo se efectuó de acuerdo a la Norma Oficial Mexicana NOM-021-RECNAT-2000.

Se determinaron las siguientes propiedades del suelo:

Materia Orgánica (MO),
% de Carbono Orgánico (Corg),
% de Nitrógeno Total (Ntotal), y
Relación C/N (C/N)

Estas propiedades se analizaron en muestras de suelo tomadas en 22 localizaciones no regulares, que eran representativas de la zona de estudio. Estas observaciones pudieron ser tratadas como datos geoestadísticos ya que son mediciones tomadas en localizaciones fijas y en escala continua. (Linares. et al. , 2006).

3. Análisis Exploratorio de los datos de la Caldera de Teziutlán

Las tablas 1 y 2 resumen las principales estadísticas univariadas y bivariadas. Puede apreciarse, en la tabla 1, que salvo Ntot, las restantes variables pueden considerarse que poseen distribución aproximadamente simétrica, dada la cercanía entre la media y la mediana de cada variable y presentar coeficientes de asimetría cercanos a cero. Se aplicó la transformación logaritmo a la variable Ntot, para continuar el análisis con una tabla de datos donde todas las variables tenían distribuciones no sesgadas.

La tabla 2 muestra el triángulo inferior de la matriz de correlaciones de Pearson entre las

Var	N	Xmedia	s	Mín	Medi	Máx	CAsi
Mo	38	6.48	4.5	0.4	6.19	18.1	0.51
COr	38	3.75	2.6	0.2	3.58	10.5	0.52
Ntot	38	0.36	0.68	0.01	0.24	4.27	5.43
CN	38	12.6	3.17	4.35	13.0	18.0	-0.9

Cuadro 1: Estadísticas Univariadas

cuatro variables consideradas, incluyendo debajo de cada coeficiente el valor de p empírico, lo que permite establecer las correspondientes pruebas de hipótesis de independencia. Puede apreciarse que únicamente la relación C/N no muestra alta correlación con las restantes variables.

Finalmente, la tabla 3 , muestra el Análisis de Componentes Principales (ACP), para sólo tres variables. (Linares, 1991). En el ACP con tres variables (se omite C/N que mostraba baja correlación con las restantes variables), se obtiene una sola componente que explica el 78 % de la variabilidad total.

	MO	**COrg**	**logNtotal**	**C/N**
MO	**1**			
	0.000			
COrg	**1.00**	**1**		
	0.000	0.000		
logNtotal	**0.488**	**0.488**	**1**	
	0.002	0.002	0.000	
C/N	**0.100**	**0.098**	**0.237**	**1**
	0.552	0.559	0.153	0.000

Cuadro 2: Estadísticas bivariadas: correlaciones.

Valor propio	2.3522
Proporción	**0.784**

Variable	**CP1**
MO	0.966
COrg	0.966
logNtotal	0.697

Cuadro 3: Análisis de Componentes Principales (3 variables)

Los cálculos se realizaron con MINITAB 14. Se decidió tomar los puntajes de la componente principal obtenida en la tabla 3 para realizar el Análisis Geoestadístico.

4. Análisis Geoestadístico de los Datos de la Caldera de Teziutlán.

Al llevar a cabo un análisis de datos geoestadístico deben estimarse las relaciones espaciales y las predicciones en los puntos no muestreados, así como, calcularse la estimación del error estándar de las predicciones. (Webster y Oliver, 2001).

Para estimar las relaciones espaciales debe contarse con el variograma, que da una medida de la correlación espacial describiendo cómo los datos muestrales están relacionados con la distancia y la dirección. De esta manera puede detectarse si el proceso es isotrópico (si no depende de la dirección) o es anisotrópico (si depende de la dirección). (Cressie, 1993).

Varias herramientas exploratorias como las nubes de variogramas y la matriz de anisotropía geométrica, señalaron que los datos analizados provenían de un proceso ligeramente anisotrópico. El variograma empírico, brindó la descripción de cómo los datos están correlacionados con la distancia y permitió estimar los parámetros de rango, sill y nugget, con los valores 9582.39, 0 y 2.298, respectivamente.

Dado que para desarrollar el método kriging es necesario especificar una función de variograma teórico, seleccionamos el modelo esférico. Se modeló el variograma esférico con los parámetros mencionados anteriormente y se obtuvieron las predicciones kriging a través de kriging ordinario. Previamente se comprobó que los puntajes de la componente principal tenían un coeficiente de asimétria de 0.21, lo que corroboraba empíricamente el supuesto de normalidad de la misma. El análisis fue realizado con S-PLUS: S+Spatial Stats, (2000).

La variabilidad espacial no fue particularmente significativa. Aparentemente, simples mediciones del carbono en el suelo, como la media, pudieran ser suficientes para estimar el carbono almacenado en el suelo. Los resultados coinciden con autores como Delise, et al (2001) al utilizar técnicas geoestadísticas para estimar la cantidad de carbono en suelos.

5. Conclusiones

En presencia de variables altamente correlacionadas, el análisis geoestadístico puede realizarse a través de la aplicación de técnicas factoriales como el ACP, que reducen la di-

mensión y evitan trabajar con modelos de corregionalización completa. Dado el carácter multidimensional de las propiedades de los suelos forestales, la metodología antes expuesta permite la evaluación geoestadística de estos suelos. Es necesario, en el estudio del secuestro de carbono integrar a las herramientas que brinda la Estadística Espacial otros enfoques y estrategias, que combinados con los anteriores, contribuirían a una mejor explicación de dicho fenómeno.

6. Referencias

Cressie, Noel A.C. (1993). *Statistics for Spatial Data.* New York : John Wiley.

Delise, *et al*(2001). *Modeling Soil Spatial Variability for C Stocks Estimation at the Field Level and Considerations for Scaling Up.*

Linares F., G. (1991). Análisis de Datos. *ENPES.* La Habana, Cuba.

Linares F., G. (2004). *Geoestadística en las ciencias del Suelo.* Puebla, México : Memorias de la XI Semana Nacional de Estadística. Facultad de Ciencias Fisico Matemáticas de la BUAP.

Linares F., G. *et al*(2006) *Análisis geoestadístico del secuestro de carbono en suelos forestales.* Oaxtepec, Morelos, México: Memorias del V Congreso Internacional y del XI Nacional de Ciencias Ambientales.

Webster, R. y Oliver, M. A. (2001). *Geostatistics for Environmental Scientist.* Chichester, England: John Wiley & Sons.

Diseño y análisis de un experimento fraccionado para determinar el tipo de arcilla óptima bajo diferentes condiciones de operación

H. Hervert Zamora [1]
Universidad Autónoma de Tamaulipas, Facultad de Ingeniería
M. Godínez Trejo
INEE, Instituto Nacional para la Evaluación de la Educación
D. Nieves Mendoza
Universidad Autónoma de Tamaulipas, Facultad de Ingeniería "Arturo Narro Siller"
C.Z. Nava Vera
Universidad Autónoma de Tamaulipas, Facultad de Ingeniería "Arturo Narro Siller"

1. Introducción

Las arcillas que tienen alto contenido de concentración de F, no son adecuadas para involucrarlas en procesos industriales referentes a la fabricación del cemento. Por ello, mientras menor sea el valor de F, ésta tendrá un mayor aprovechamiento dentro la industria. Una vez que se encontraron los métodos adecuados para el tratamiento de las arcillas, se requiere realizar un diseño de experimentos, en donde se establezcan los niveles de operación más convenientes y poder establecer así, las condiciones de operación optima. Para el presente estudio se tienen tres tipos de arcillas (distintas entre ellas) las cuales poseen un % inicial de F, por lo que se desea minimizar ese valor al final del tratamiento químico, para así obtener arcillas mejoradas con la menor concentración de F.

[1] hhervert@uat.edu.mx

2. Marco Teórico

Una réplica completa del diseño 3^4 requiere un total de 81 corridas, sin embargo un diseño factorial fraccionado a la un tercio, requiere solo 27 corridas. Cada efecto principal o componente de la interacción estimado a partir de este diseño tiene dos alias. Cada uno de los componentes AB y AB^2 tiene dos grados de libertad. Los niveles (0,1,2) de A y B se denotan por x_1 y x_2, respectivamente, las distintas combinaciones ocupan celdas de acuerdo con el patrón que se define en la tabla 1.

Cuadro 1: Patrón para el llenado de las celdas

Cuadrado a	Cuadrado b
$x_1 + x_2 = 0$	$x_1 + 2x_2 = 0$
$x_1 + x_2 = 1$	$x_1 + 2x_2 = 1$
$x_1 + x_2 = 2$	$x_1 + 2x_2 = 2$

Tomando en base el modelo: $x_1 + \alpha_2 x_2 + \alpha_3 x_3 = u$ y el patrón anterior; se tiene que $\alpha_1 = \alpha_3 = \alpha_4 = 1$ y $\alpha_2 = 2$ esto implica que $\beta_1 = (3-1)\alpha_1$ sea igual a $\beta_1 = (3-1)(1) = 2$; de la misma forma $\beta_2 = (3-1)\alpha_2$; $\beta_3 = (3-1)\alpha_2$ por lo tanto la ecuación del modelo nos queda de la siguiente forma: $2x_1 + x_2 + 2x_3 = x_4$. El diseño 3^{4-1}_{IV} resultante tiene 26 grados de libertad que pueden usarse para calcular las sumas de cuadrados de los 13 efectos principales y los componentes de las interacciones (y sus alias). La identidad es I = AB^2CD. Los componentes de las interacciones no tienen ninguna interpretación práctica, ya que se confunden con los bloques. Para un análisis factorial 2^{k-1} con K = 4 y resolución IV, tiene un total de 8 corridas, donde I= ABCD (cada una de las letras representa a un factor).

3. Metodología

Se tienen tres tipos de arcillas las cuales para fines del estudio se denominaron arcilla 0,1 y 2. El proceso parte de una arcilla inicial, de la cual se conoce su % F y estructura. Posteriormente se somete a una reacción química con distintos reactivos, siendo A, M y B los reactivos que corresponden a tres métodos previamente seleccionados. Una vez que se tiene el material arcilloso en contacto con el reactivo (A, M o B) se ajustan las variables de temperatura y

tiempo según el diseño de experimentos sugerido. El proceso experimental es el siguiente:

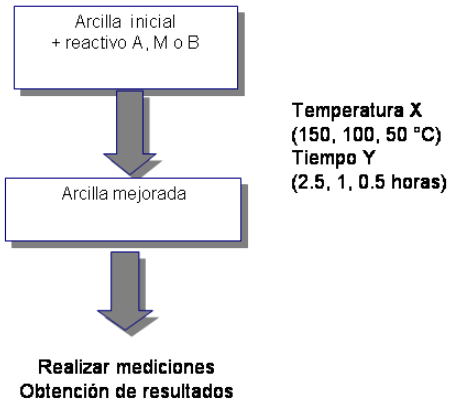


Figura 1. Fases del proceso experimental

Como se observa en la Figura 1, los factores de temperatura, tiempo y método (reactivo utilizado) son clave para la experimentación, partiendo de los distintos tipos de arcilla. Por lo tanto se tiene un análisis factorial, con cuatro variables cada una con tres niveles de experimentación los cuales se presentan en la tabla 2

Cuadro 2: Combinación de factores con sus distintos niveles

Nivel	Arcilla	Temp.	Tiempo	Método
0	tipo 1	100	2.5	A
1	tipo 2	75	1	M
2	tipo 3	50	0.5	B

Se realizó un análisis factorial fraccionado 3^{k-1} con K = 4, con un total de 27 corridas; para el segundo análisis se realizó un análisis factorial 2^{k-1} con K = 4 y resolución IV, con un total de 8 corridas

4. Resultados

Se probaron los supuestos de Normalidad de los residuales por medio de la prueba de bondad de ajuste de Kolmogorv con la cual se concluyó que los residuales sí se están comportando

normalmente con un p-valor de 0.150, además se probó la heteroscedasticidad de los errores; por tal motivo se corrió el análisis de varianza pertinente.

Cuadro 3: Análisis de Varianza ANOVA (usando valores ajustados SS).

Source	DF	Seq SS	Adj SS	Adj MS	F	P
arcilla	2	37.8034	37.8034	18.9017	51.33	0.000
temperat	2	8.3540	8.3540	4.1770	11.34	0.001
tiempo	2	1.6971	1.6971	0.8486	2.30	0.128
H	2	0.7601	0.7601	0.3801	1.03	0.376
Error	18	6.6278	6.6278	0.3682		
Total	26	55.2424				

En base al ANOVA, se concluye que tanto el efecto que causa el factor arcilla con un valor p de 0.000, como temperatura con un valor p de 0.001, son estadísticamente significativos; es decir afectan a la variable respuesta, sin embargo el efecto del tiempo y método[H] no son estadísticamente significativos. En la figura 1 se muestran las interacciones de los efectos, en el cual es posible observar que con la combinación del efecto arcilla tipo 2 (Valles) a una temperatura a nivel 0 (150), tiempo a nivel 0 (2.5) y utilizando el método 0 (A) se logra minimizar el % F. Un aspecto muy importante es, que si se utiliza la arcilla tipo 2 la variabilidad en el % F se mantiene en un rango constante. Sin embargo también cabe señalar que al variar los niveles de temperatura con los niveles de los efectos método y/o tiempo, el % F se altera considerablemente.

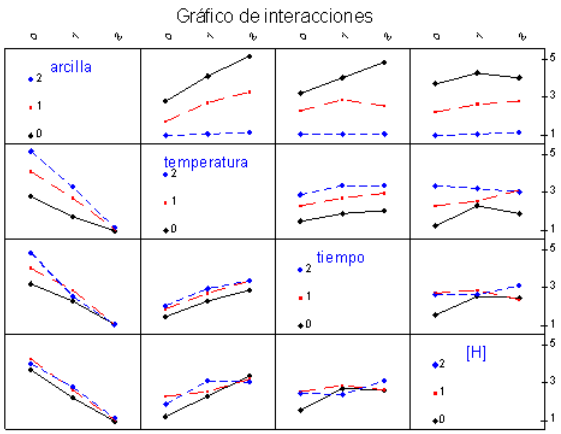


Figura 2. Interacciones (análisis general)

Para determinar las condiciones óptimas para cada tipo de arcilla se realizó un análisis factorial individual para cada una de ellas.

4.1. Análisis estadístico por tipo de arcilla

Arcilla tipo 0. En base al análisis realizado para esta variable, se concluyó que existe evidencia estadística suficiente con un valor p de 0.001 que el efecto de la temperatura sobre la variable respuesta (% F) afecta significativamente; así como también el tiempo y el [H] con un valor p de 0.003 y 0.018 respectivamente. En la figura 2 se observa que al tomar la temperatura 0 con el tiempo 0 y el [H] 0 se alcanza un mínimo, cabe señalar que con el tiempo a nivel 0 y se varía el [H] de nivel 2 a 0 se observa un cambio radical en la variable respuesta, lo mismo sucede con los demás niveles del [H].

Arcilla tipo 1. Existe evidencia estadística suficiente para concluir lo siguiente: la temperatura afecta significativamente %F, con un valor p de 0.017, sin embargo, el factor tiempo con un valor p de 0.114 no afecta significativamente a la variable respuesta es decir el tiempo que se aplique en el proceso no hace variar significativamente al % F; de manera similar el [H] con un valor p de 0.124, no afecta significativamente a la variable respuesta. En la figura 3 se observa que se alcanza un mínimo colocando la temperatura, y el [H] a nivel 0 y el tiempo a nivel 0. Además el cambio de temperatura de nivel 0 a nivel 2 ó 1 ocasiona variabilidad amplia en la variable respuesta.

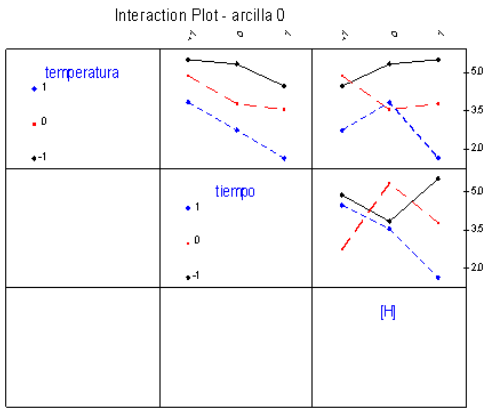


Figura 3. Interacciones referentes a la arcilla 0

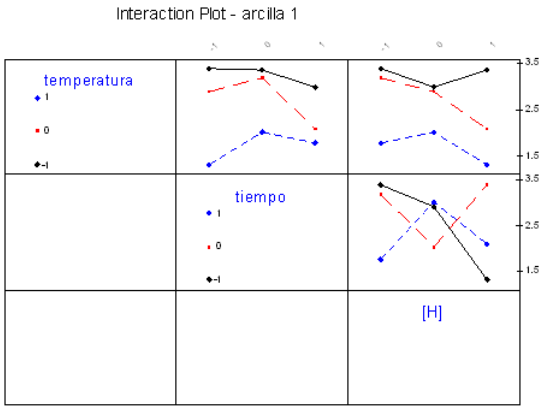


Figura 4. Interacciones de la arcilla tipo 1

Arcilla tipo 2 La anova nos muestra que existe evidencia estadística para concluir que todos los factores (tiempo, temperatura, y [H]) no son significativos, es decir con un valor p de 0.275, para la temperatura, 0.723, 0.369 para el tiempo y para el [H] respectivamente; los factores no están afectando considerablemente a la variable respuesta (porcentaje de F). En la figura 4 se observa que sí existe variación al cambiar de nivel en cada uno de los factores (temperatura, tiempo, [H]); además se obtiene un mínimo colocando la temperatura a nivel 0, el tiempo a nivel 1 y el [H] a nivel 0.

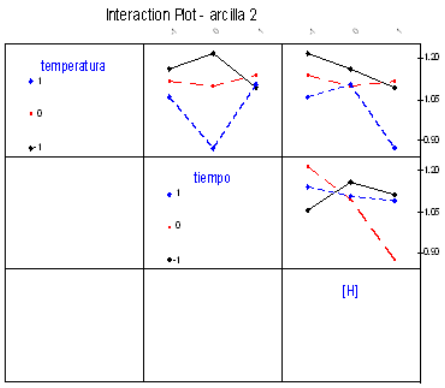


Figura 5. Interacciones de la arcilla tipo 2

5. Discusión

Se elaboraron los gráficos de las curvas de nivel para explicar mejor el comportamiento del modelo y poder definir dónde se encuentra el mínimo global. Las condiciones óptimas obtenidas por cada tipo de arcilla permitirán alcanzar una mejor calidad del producto final. Cabe señalar que las decisiones tomadas en base a los gráficos son solo sugerencias de apoyo, ya que no proporcionan información estadística, sino tan solo modelan el comportamiento de los datos, y queda a consideración del experto su interpretación.

6. Conclusiones

Los niveles con los cuales se logra minimizar el % F para las posibles combinaciones de factores clave del primer análisis son: temperatura a 150, a un tiempo de 2.5 horas, aplicando el método A, utilizando la arcilla tipo 2 (Valles); utilizando el tipo de arcilla 2 (Valles) se logra mantener el porcentaje de F casi estable, sin importar el método o tratamiento químico que se le aplique, la temperatura a la que se realice el proceso, ni el tiempo que dure el mismo.

Las combinaciones óptimas para cada tipo de arcilla, se muestran en la tabla 5 y 6 respectivamente.

Cuadro 4: Resumen de condiciones óptimas para cada tipo de arcilla

Arcilla	Temp.	Tiempo	Método
tipo 1	150	2.5	A
tipo 2	150	2.5	A
tipo 3	150	1	A

7. Referencias

Besoain, Eduardo (1985). *Mineralogía de Arcillas de Suelos*. Instituto Interamericano de Cooperación para la Agricultura. Costa Rica. Primera Edición. 8–13, 149, 158–163, 296.

Gieseking, John E. (1975). *Soil Components*. Editorial Springer-Verlag York Inc., E.U.A. Primera edición. 18–19, 458–466.

Montgomery (2001). *Diseño y Análisis de Experimentos*. Segunda Edición. Limusa Wiley, 379, 382.

Seoánez Calvo, Mariano (1999). *Contaminación del Suelo: Estudios, tratamiento y gestión*. Ediciones MundiPrensa, España , 27–28, 95.

Una clase flexible de modelos autorregresivos de primer orden utilizando cópulas[1]

Angélica Hernández Quintero[2]
Universidad Autónoma Metropolitana- Iztapalapa
Gabriel Escarela [3]
Universidad Autónoma Metropolitana- Iztapalapa

1. Introducción

La literatura de series de tiempo tiene un debate bien establecido para el modelado de respuestas Gaussianas. Sin embargo, cuando se trata de respuestas que no se distribuyen normalmente, no existe algún método que pueda generalizarse a cualquier tipo de respuesta. El propósito del presente trabajo es mostrar un modelo autorregresivo orientado a verosimilitud para respuestas que no necesariamente se ditrubuyen normalmente y que estén correlacionadas en forma adyacente.

El modelo que se presenta en este trabajo está definido por una clase de distribuciones condicionales las cuales son construídas a partir de un modelo cópula y de una distribución marginal dada, la cual pertenece a la familia exponencial de distribuciones. Esta clase representa una forma flexible y generalizada para el modelado de respuestas continuas que presentan correlaciones en forma adyacente; la dependencia es modelada a través de una función cópula y la distribución marginal puede darse en términos de un modelo lineal generalizado, permitiendo así un amplio rango de dependencia entre las respuestas, la elección de una marginal conveniente y la inclusión de información concomitante.

[1]Trabajo realizado con apoyos de SEP-CONACYT-ANUIES-ECOS
[2]cbi206280113@xanum.uam.mx
[3]ge@xanum.uam.mx

2. Cópulas

Una cópula bivariada es una función continua $C : \mathtt{I}^2 \to \mathtt{I} = [0, 1]$ con marginales uniformes en el intervalo unitario. La importacia de las cópulas en estadística es descrita por el siguiente teorema:

Teorema de Sklar. Sean Y_1 y Y_2 variables aleatorias con función de distribución conjunta H, y marginales F_1 y F_2 respectivamente. Entonces existe una cópula C que satisface

$$H(y_1, y_2) = C\left[F_1(y_1), F_2(y_2)\right] \tag{1}$$

para toda $y_1, y_2 \in \mathbb{R}$. Inversamente si C es una cópula y F_1 y F_2 son funciones de distribución, entonces H definida en (1) es una función de distribución conjunta con marginales F_1 y F_2.

Sea $c\left[F_1(y_1), F_2(y_2)\right] = \partial^2 C\left(F_2\left(y_1\right), F_2\left(y_2\right)\right)/\partial y_1 \partial y_2$ la función de densidad de la cópula. Si f_1 y f_2 denotan las funciones de densidad de F_1 y F_2 respectivamente, es posible demostrar que la función de densidad conjunta de (Y_1, Y_2) es $h(y_1, y_2) = f_1\left(y_1\right) f_2\left(y_2\right) c\left[F_1(y_1), F_2(y_2)\right]$. A partir de este resultado, es fácil de obtener la función de densidad condicional, que es

$$f_{2|1}(y_2 \mid y_1) = f_2(y_2) \times c[F_1(y_1), F_2(y_2)] \tag{2}$$

Las cópulas permiten un camino fácil en el estudio de medidas de asociación entre variables aleatorias. En particular se pueden destacar la tau de Kendall y la rho de Spearman, las cuales están dadas por $\tau = 4 \int \int_{\mathbf{I^2}} C(y_1, y_2) dC(y_1, y_2) - 1$ y $\rho = 12 \int \int_{\mathbf{I^2}} \left[C\left(y_1, y_2\right) - y_1 y_2\right] dy_1 dy_2$.

Entre las familias de cópulas encontradas en la literatura podemos destacar a dos: la cópula Gaussiana y la cópula Positiva Estable. La cópula *Gaussiana* tiene la representación

$$C_\alpha\left(u, v\right) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\alpha^2}} \exp\left\{-\frac{1}{2}\left(\frac{x^2 - 2\alpha xy + y^2}{1-\alpha^2}\right)\right\} \mathrm{d}x\mathrm{d}y$$

La tau de Kendall y la rho de Spearman para esta cópula están dadas por $\tau_\alpha = \frac{2}{\pi}\text{arcsen}(\alpha)$ y $\rho_\alpha = \frac{6}{\pi}\text{arcsen}(\frac{\alpha}{2})$ respectivamente. La función de densidad $Y_2|Y_1$ para respuestas con dis-

tribución gaussiana se puede expresar como

$$f_{2|1}(y_2|y_1) = \frac{f_2(y_2)}{\sqrt{1-\alpha^2}} \times \exp\left\{-\frac{1}{2}\left[\frac{(s_2-\alpha s_1)^2}{1-\alpha^2} - s_2^2\right]\right\},$$

donde $s_i = \Phi^{-1}((F_i(y_i))$, $i = 1, 2$, y Φ^{-1} es la inversa de la función de distribución normal estándar.

La cópula *Positiva Estable*, también conocida como Gumbel-Hougaard, es definida como $C_\alpha(u, v) = \exp\left\{-\left[(-\log u)^\alpha + (-\log v)^\alpha\right]^{1/\alpha}\right\}$, donde $\alpha \geq 1$. La tau de Kendall para esta cópula viene dada por $\tau_\alpha = 1 - \alpha^{-1}$.

3. El modelo autorregresivo

Considérese una serie de tiempo estacionaria $\{Y_t, t = 1, 2, ...\}$ con respuestas marginales $Y_t \sim \mathrm{N}(\boldsymbol{\beta}^T\mathbf{x}_t, \sigma^2)$, $t = 1, 2, ...$, donde $\mathbf{x}_t$ es un vector de variables explicativas en el tiempo t, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión, y σ^2 es la varianza marginal de las respuestas. Si la correlación entre las respuestas adyacentes Y_{t-1} y Y_t es r, el modelo de transición tiene la forma $Y_t \mid Y_{t-1} \sim \mathrm{N}\big(\boldsymbol{\beta}^T\mathbf{x}_t + r[Y_{t-1} - \boldsymbol{\beta}^T\mathbf{x}_{t-1}],\, \nu^2\big)$, donde $\nu^2 = \sigma^2(1-r^2)$, y $|r| < 1$. Usando la función de densidad condicional de la ecuación (2), $Y_t \sim f$, se puede construir un modelo autorregresivo de orden 1 para respuestas continuas en una forma similar al modelo Gaussiano al simplemente seleccionar una cópula y una marginal que por conveniencia puede ser una distribución de los modelos lineales generalizados.

La función de verosimilitud, para un modelo de transición de primer orden está dada por $L = f_1(y_1; x_1)\prod_{k=1}^n f(y_k|H_k)$; aquí, $H_k = (y_{k-1}; x_{k-1})$. El objetivo de presentar la función de máxima verosimilitud para un modelo de transición de orden 1, es que nos permitirá comparar dos modelos. Existen medidas de contraste entre modelos que penalizan en alguna medida que éstos tengan muchos parámetros, entre las cuales podemos destacar el criterio de información Akaike, el BIC y la devianza. El criterio de selección al mejor modelo será aquel que tenga valores de AIC y BIC más bajos.

4. Dos ejemplos

Se analizan los datos de un estudio realizado a un simio sobre las reacciones de su aparato neuronal (Zeger y Qaquish, 1988), en la figura 1 se muestra la serie de tiempo de las respuestas. Se observa que los datos presentan efectos de periodicidad. El modelo general de la regresión es expresado por, $tiempos$ = polinomio de grado n del $n.obs + \cos\left(\frac{2\pi n.obs}{13}\right) + \text{sen}\left(\frac{2\pi n.obs}{13}\right)$.

El objetivo es ajustar los datos utilizando la cópula positiva estable y marginales gamma, el cuadro 1 muestra el valor de la devianza, el AIC y el BIC para los diferentes modelos, donde $p(i)$,$i = 1, 2, 3, 4, 5$ denota el polinomio de grado i. Mientras que el cuadro 2 presenta los coeficientes de regresión con sus respectivos errores estándar del mejor modelo.

Cuadro 1: Valores del AIC y el BIC utilizando la Cópula Positiva y marginales Gamma

Modelo	-2xlogL	No. de Parámetros	AIC	BIC
Nulo	844.608	3	850.608	858.393
Periodo+p(1)	825.774	6	837.774	839.559
Periodo+p(2)	814.299	7	828.299	846.465
Periodo+p(3)	804.646	8	820.646	841.407
Periodo+p(4)	803.575	9	821.575	844.931
Periodo+p(5)	801.581	10	821.581	847.532

Cuadro 2: Coeficientes estimados y errores estándar para datos del simio

Parámtero	ordenada	$n.obs$	$n.obs^2$	$n.obs^3$	$\cos(2\pi t/13)$	$\sin(2\pi t/13)$	ϕ	θ
Estimación	3.501	-2.989	2.069	1.624	0.059	0.190	5.302	4.641
Error estándar	0.044	0.445	0.452	0.468	0.064	0.063	0.732	7.613

La figura 1 muestra el análisis de residuales, lo cual señala que el modelo propuesto es apropiado para ajustar las respuestas del simio.

Por otro lado se analizan las series de tiempo mensuales de casos de poliomielitis, reportados por U.S. Center Disease Control, para los años de 1970 a 1983, (Zeger, 1988). El objetivo es

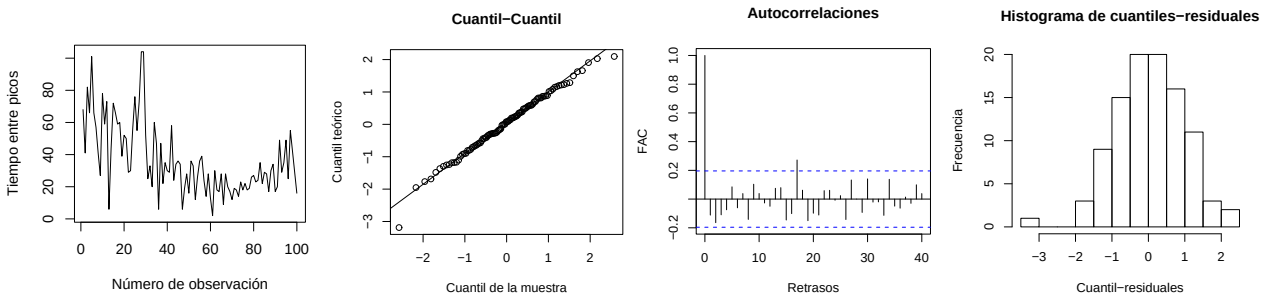


Figura 1: Serie de tiempo y gráficas de los cuantiles-residuales al mejor modelo

comparar el modelo independiente y el modelo cópula (marginales Poisson y cópula Positiva Estable). Se introducen efectos de tendencia y efectos de periodicidad en forma anual y semi-anual. El cuadro 3 muestra la devianza y el AIC para diferentes modelos. Es posible concluir que el utilizar la cópula Positiva Estable permite un ajuste mejorado de los datos en general.

Cuadro 3: Valores del AIC y el BIC utilizando la Cópula Positiva Estable, la Cópula independiente y marginales Poisson

Modelo	Cópula	-2xlogL	No. de Parámetros	AIC	BIC
Nulo	Positiva	584.978	2	588.978	594.168
	II	600.026	1	601.026	602.223
Sin efectos semi-anuales	Positiva	551.704	5	561.704	574.679
	II	566.643	4	574.643	585.023
Sin efectos anuales	Positiva	557.836	5	567.836	580.816
	II	570.643	4	578.643	589.023
Sin tendencia	Positiva	544.216	6	556.216	571.788
	II	556.535	5	566.535	579.511
Todas las variables explicativas	Positiva	532.205	7	546.205	564.37
	II	543.546	6	555.546	571.117

Referencias

Zeger SL, Qaquish B. (1988). Markov Regressions Models for Time Series A Quasi-Likelihood Approach. *Biometrics*; **44**: 1019-1031

Zeger SL. (1988). A regression model for time series of count. *Biometrika*; **75**: 621-629

Análisis de datos longitudinales en R

Miguel A. Polo Vuelvas[1]
Gabriel Escarela Pérez[2]
Universidad Autónoma Metropolitana - Iztapalapa

1. Introducción

Los datos longitudinales son arreglos en los cuales se consideran varias unidades experimentales (personas, empresas, ciudades, animales, etc.) de las cuales se registran repetidamente a lo largo de un periodo de estudio las observaciones de las respuestas de interés conjuntamente con sus variables explicativas.

La elección de un modelo particular para tratar un conjunto de datos longitudinales depende de los objetivos del estudio, pues la interpretación de los resultados varía de un modelo a otro. Esta elección también depende de la estructura de los datos, quiénes son las variables explicativas y qué tipo de relación tienen con la variable respuesta.

El presente trabajo tiene como propósito exponer los tres principales enfoques de modelación para datos longitudinales que se encuentran en la literatura mediante el análisis de ejemplos reales usando el paquete estadístico R.

2. Modelo marginal

Cuando se trata de modelar la respuesta media de una muestra longitudinal, se usan los modelos marginales, los cuales tienen una interpretación semejante a los estudios de corte transversal, con la diferencia que en los estudios de corte transversal se supone independencia entre las observaciones, y en los modelos marginales es necesario tomar en cuenta la correlación de las observaciones de un mismo sujeto.

En el siguiente ejemplo se analiza la efectividad del medicamento *progabide* en el tratamiento

[1] saygondragon@yahoo.com.mx
[2] ge@xanum.uam.mx

de ataques epilépticos, comparado con un placebo (Breslow y Clayton (1993), Thall y Vail (1990)). Para cada uno de 59 pacientes con epilepsia, se registró el número de ataques durante un periodo base de 8 semanas. Después fueron seleccionados al azar para formar uno de dos grupos: grupo "progabide" y grupo de control. El número de ataques fue registrado en cuatro intervalos bisemanales consecutivos.

Para ajustar el modelo de regresión usamos la siguiente función, que se incluye en el paquete `geepack` para R:

```
> m1 <- geeglm(y ~ offset(log(t)) + x + trt + x:trt, id = id,
               data=seiz.l, corstr="exch", family=poisson)
```

donde la variable `trt` nos da información de la diferencia entre los grupos de tratamiento. `x` es un indicador: si la observación es del periodo base, `x=0`, si es de las observaciones de tratamiento, `x=1`. El término `log(t)` es necesario para tomar en cuanta los diferentes periodos de observación. El coeficiente correspondiente a la variable `x:trt` representa la diferencia del logaritmo de la razón *tratamiento/base* entre el grupo de progabide y el de control. Un valor negativo de este coeficiente corresponde a una enorme reducción (o a un muy pequeño incremento) en el conteo de ataques para el grupo de progabide. Dado que son datos de conteo, se supone una distribución Poisson para las respuestas. La estructura de correlación asignada en este caso es "equicorrelación".

La siguiente tabla muestra los resultados del ajuste.

	estimate	san.se	wald	p
(Intercept)	1.35	0.16	73.34	0.00
x	0.11	0.12	0.93	0.33
trt	0.03	0.22	0.02	0.90
x:trt	−0.10	0.22	0.23	0.63

Estos valores sugieren que hay una muy pequeña diferencia entre los grupos de tratamiento "progabide" y de control en el cambio del conteo de ataques antes y después del tratamiento, pues `trt` no es significativa para el modelo de regresión.

3. Modelo de efectos aleatorios

Para exponer los modelos de efectos aleatorios se estudia un conjunto de datos longitudinales analizado en Frees et al. (2004). Tales datos corresponden a un estudio realizado para el programa de seguros médicos Medicare en 54 Estados de la Unión Americana.

La variable respuesta de interés es RCA, la cantidad de indemnizaciones cobradas por paciente en dólares. Las variables explicativas son Tiempo (en años), NA (número de pacientes dados de alta) y EP (estancia promedio en hospitalización).

En base al estudio previo de los datos, el modelo queda especificado por

$$\begin{aligned} \mathrm{RCA_i} &= \beta_0 + \beta_1 \mathrm{Tiempo_i} + \beta_2 \mathrm{NA_i} + \beta_3 \mathrm{EP_i} + \beta_4 \mathrm{Tiempo_i} : (\mathrm{Estado} = 31) \\ &\quad + \mathrm{b_{i1}} + \mathrm{b_{i2}} \mathrm{Tiempo_i} \end{aligned}$$

donde $Tiempo_i : (Estado = 31)$ es una interacción que considera el comportamiento inusual de RCA en el Estado 31 y β_j son los efectos fijos y b_{ij} son los efectos aleatorios. El ajuste se hace mediante la función `lme` del paquete `nlme` para R:

```
model<-lme(RCA~Tiempo:(Estado==31)+NA+EP,
           datos,random=~Tiempo|Estado,method="ML")
```

cuyos resultados son

	Value	Std.Error	DF	t-value	p-value
(Intercept)	4406.22	574.41	2650	7.67	0.00
Tiempo	738.12	34.57	265	21.35	0.00
NA	0.00	0.00	265	1.84	0.07
EP	340.30	45.65	265	7.45	0.00
Tiempo:Estado == 31TRUE	1530.42	184.04	265	8.32	0.00

Con estos valores se concluye que todas las variables son significativas en el modelo, excepto `NA`.

4. Modelo de transición

Los modelos de transición consideran la información de las variables explicativas, así como de las observaciones pasadas en la distribución condicional de la respuesta. A menudo, el estudio de datos longitudinales mediante modelos de transición se basa en el ajuste de un modelo de Markov multi-estados en tiempo continuo, cuyo algoritmo para la estimación de parámetros por el método de máxima verosimilitud fue propuesto por Kalbfleisch y Lawless (1985). En este caso, se hará uso de un conjunto de datos de monitoreo de transplante de corazón (Sharples et al. (2003)). Los datos son especificados como una serie de observaciones agrupadas por paciente (cada paciente tiene un número de identificación). Las variables que intervienen en el estudio son

- el tiempo de las observaciones (years),
- el estado observado del proceso (state).

Un modelo de este tipo es especificado por una matriz de intensidad de transición. Ajustar el modelo es un proceso de encontrar valores para las entradas de la matriz de intensidades en base a valores iniciales. Esto se hace usando la función `msm` del paquete del mismo nombre, como sigue:

```
ajuste <- msm(state ~ years, subject = PTNUM,
            data = heart, qmatrix = ini.q, death = 4)
```

donde `ini.q` es la matriz de valores iniciales.

Los resultados del ajuste, se presentan con la siguiente matriz de intensidades estimadas:

	State 1	State 2	State 3	State 4
State 1	-0.17	0.13	0	0.04
State 2	0.22	-0.61	0.34	0.04
State 3	0	0.13	-0.44	0.30
State 4	0	0	0	0

Se puede obtener la estimación de la matriz de probabilidad de transición P dentro de un tiempo dado. Por ejemplo, las probabilidades de transición dentro de 10 años están dadas por

	State 1	State 2	State 3	State 4
State 1	0.31	0.10	0.09	0.50
State 2	0.17	0.07	0.08	0.68
State 3	0.06	0.03	0.05	0.86
State 4	0.00	0.00	0.00	1.00

En estudios de enfermedades crónicas, un uso importante de los modelos multi-estados es predecir la probabilidad de sobrevivencia de los pacientes para algún tiempo en el futuro.

5. Referencias

Breslow, N. E. y Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9-25.

Diggle, P. J., Liang, K. Y., y Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Frees, E. W., Young, V. R., y Luo, Y. (2004). Case studies using panel data models, *North American Actuarial Journal*, **5**, 24-42.

Kalbfleisch, J. D. y Lawless, J. F. (1985). The analysis of panel data under a markov assumption, *Journal of the American Statistical Association*, **80**, 863-871.

Sharples, L. D., Jackson, C. H., et al. (2003). Diagnostic accuracy of coronary angiopathy and risk factors for post-heart-transplant cardiac allograft vasculopathy, *Transplantation*, **76**, 679-682.

Thall, P. F. y Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics*, **46**, 657-671.

Modelos de transición para analizar problemas de ecología[1]

Francisco Solano Tajonar Sanabria[2]
Benemérita Universidad Autónoma de Puebla
Gabriel Escarela Pérez [3]
Universidad Autónoma Metropolitana-Iztapalapa

1. Introducción

La dinámica de las poblaciones es importante para entender el desarrollo temporal y espacial de grupos de organismos de la misma especie que se desarrollan en distintos ambientes. En la práctica, el interés se centra en el manejo de plagas agrícolas, para entender la epidemiología de numerosas enfermedades, etc.

Una **población** desde el punto de vista ecológico es un grupo de organismos de la misma especie, que habitan un lugar determinado, en el cual utilizan recursos y se reproducen. El promedio de nacimientos que se den en el grupo constituye la **natalidad** del grupo.

La **densidad** es la representación de la abundancia de la población y se expresa como el número de individuos o biomasa en función del espacio o volumen ocupado. La densidad puede ser absoluta o ecológica y la abundancia determina algunos efectos a nivel de la población.

La densidad es producto del balance entre **natalidad** y **mortalidad poblacional**, y también del balance entre **inmigración y emigración**. Estos dos últimos factores suelen adscribirse por comodidad a la natalidad (b) y mortalidad (d) respectivamente. Estos parámetros poblacionales indican un cambio en el tamaño de la población en relación de los que nacen como los que mueren.

En este trabajo se presentan algunos modelos de crecimiento de una población, el **modelo matricial de Leslie** que se utiliza en ecología para determinar el crecimiento de una población y los porcentajes de distribución a lo largo del tiempo. También se presenta un

[1]Trabajo realizado con apoyo de la Fac. de Ciencias Físico Matemáticas
[2]ftajonar@fcfm.buap.mx
[3]ge@xanum.uam.mx

marco general de modelos de **estado y transición**, los cuales se utlizan en el estudio de la dinámica de la vegetación natural.

2. Modelos de Crecimiento

Iniciemos con el modelo más simple que permite determinar el crecimiento de una población, llamado modelo de crecimiento. Suponga que un organismo del cual se posee inicialmente N_0 individuos, tiene una capacidad de reproducción constante de λ especímenes. Así, tenemos que la reproducción para el periodo t es

$$N_t = N_{t-1}\lambda = N_0\lambda^t \tag{1}$$

En este caso, λ denota un crecimiento finito, es decir, un crecimiento por pulsos discreto. Por ello se denomina **tasa discreta de crecimiento poblacional** o **tasa finita de crecimiento**, este parámetro informa cómo crece o decrece una población entre periodos de tiempo. Esto se puede usar para definir tasas de extracción de especies silvestres. Tomando el crecimiento en función de d y b, se tiene que: $b - d = \lambda = \Delta N/N\Delta t$, si el crecimiento es por pulsos.

En ausencia de factores limitantes, esto es, con alimento suficiente y adecuado, con espacio suficiente y adecuado, una población crecerá exponencialmente, un modelo con estas características se denomina de **Crecimiento Exponencial**.

En muchas situaciones el crecimiento definido por periodos de tiempo no permite realizar comparaciones entre poblaciones que tienen diferentes periodos reproductivos, ni tampoco estimar con precisión las variaciones del desarrollo poblacional en cada instante, para resolver esto se utiliza la **tasa instantánea de crecimiento o tasa de crecimiento específico**, que es el parámetro de mayor importancia relativa en la dinámica de cualquier población. En este caso tenemos que

$$dN/Ndt = b - d = r$$

o

$$dN/dt = rN \tag{2}$$

donde r es constante para cada especie y se le denomina **tasa intrínseca de crecimiento**. Para obtener una expresión del crecimiento poblacional en función del tiempo se obtiene

integrando (2),

$$N_t = N_0 e^{rt} \tag{3}$$

Esto indica que se puede conocer el crecimiento o tamaño de una población en cualquier instante, si se conoce la población inicial N_0 y el valor de r.

3. Modelo Matricial de Leslie

El modelo matricial de Leslie es una herramienta usada para determinar el crecimiento de una población así como la distribución por edad a lo largo del tiempo. Esta descripción fue hecha por Leslie en (1945). Se ha usado para estudiar la dinámica de poblaciones de una amplia variedad de organismos, como truchas, conejos, escarabajos, piojos, orcas, humanos y para predecir distribuciones de clases de edad estable en especies leñosas. La matriz de Leslie es un caso especial de una matriz de transición. La primera fila contiene el número de descendientes de cada padre en esa clase (generalmente clases de edad/tamaño). La subdiagonal primaria (la línea inmediatamente debajo de la diagonal principal) contiene las probabilidades de moverse de una clase a la siguiente más alta. El modelo de Leslie está definido por la ecuación

$$X_k = L^k X_0, \tag{4}$$

donde X_0 es el vector inicial de distribución de la población, y X_k el vector de distribución de la población en el instante k. Si la matriz de Leslie L es diagonalizable, entonces $L = VDV^{-1}$, donde D es una matriz diagonal formada por los eigenvalores de la matriz L. Las columnas de V son los eigenvectores correspondientes. En este caso, el modelo de Leslie se puede escribir como

$$X_k = c_1\lambda_1^k v_1 + c_2\lambda_2^k v_2 + ... + c_n\lambda_n^k v_n, \tag{5}$$

donde λ_i, v_i son el eigenvalor y eigenvector asociados. Si λ_1 es el eigenvalor estrictamente dominante de L, entonces para valores grandes de k se tiene que

$$X_k \approx c_1\lambda_1^k v_1, \tag{6}$$

y la proporción de objetos en cada clase de edad tiende a una constante. Estas proporciones límites se pueden determinar a partir de las componentes de v_1. Por último, el eigenvalor

dominante λ_1 determina la tasa de cambio de un año para otro. Como

$$X_k \approx \lambda_1 X_{k-1}, \tag{7}$$

para valores grandes de k, el vector de población en el instante k es un múltiplo del vector de población en el instante $k-1$. Si $\lambda_1 > 1$ entonces la población tendrá un crecimiento indefinido. Si $\lambda_1 < 1$ entonces la población se extinguirá.
Observaciones:

1. La ecuación (4) nos indica que si conocemos el vector de distribución inicial X_0 y la matriz de Leslie L podemos determinar el vector de distribución de la población en cualquier instante. En general, la matriz de Leslie L es un caso especial de una matriz de transición y usualmente no tiene un vector de probabilidades estacionarias, sin embargo, una proporción estable límite de clases (edad/tamaño) es alcanzada y está dada por (5), al vector v_1 se le llama vector de probabilidades pseudo-estacionarias.

2. Se puede notar que la ecuación (4) tiene una expresión semejante a una ecuación en diferencias.

3. Si la matriz de Leslie L es diagonalizable, se pueden utlizar las de cadenas de Markov para calcular la distribución estacionaria de (4), utilizando la

$$X = PX. \tag{8}$$

4. Modelos de Estado y Transición

En la actualidad se sabe que los cambios climáticos pueden alterar el curso de los cambios en la vegetación (de pastos, hierbas o matorrales) y de esta manera alterar la producción agrícola. Estos cambios climáticos son ocasionados por fenómenos naturales (e.g., precipitación abundante y sequías). Se argumenta que los factores que ocasionan estos cambios son la contaminación y la erosión de las tierras (Rodríguez y Kothmann, 1997). El modelo de estado y transición (ET), así denominado por (Westoby et al. 1989), actualmente

es una herramienta popular que ayuda a productores y administradores de tierras a tomar decisiones sobre el manejo de éstas (Bellamy y Brown, 1994). Este puede ser usado para describir la dinámica y manejo de plantas leñosas (Grice y Macleod1994), la dinámica de pastos en praderas tropicales. En un modelo de (ET), los estados están caracterizados como entidades ecológicas y usualmente se describen como composición botánica de vegetación dominante. Las transiciones no siempre están claramente definidas y pueden ser clasificadas como simples o complejas. Transiciones simples involucran la acción de una sola posible causa (aunque se puede tener más de una componente; tratamiento químico de plantas leñosas y pastos, por ejemplo) que pueden involucrar uno o más factores (e.g., tratamiento químico de plantas leñosas, pastizales, precipitación, fertilización, estacionalidad). Transiciones complejas pueden ser provocadas por más de una causa (pastizales o precipitación y fuego de verano) cada una de las cuales involucra a uno o más factores. Los modelos de (ET) son similares en estructura a los procesos de Markov, ya que éstos poseen estados, transiciones y además, son procesos que están evolucionando con el tiempo. Además, la expresión analítica de un modelo de estado y transición tiene la siguiente forma

$$X_{t+1} = PX_t \tag{9}$$

donde X_{t+1} y X_t, son vectores cuyos elementos son proporciones del sistema en ese estado, y P es una matriz cuadrada de probabilidades de transición, i.e., probabilidades de moverse de cada uno de los estados a todos los otros estados en un periodo de tiempo. Así, que los procesos de Markov a tiempo discreto se pueden aplicar para modelar aspectos de la dinámica de cambios de vegetación y de cultivos. Dos modelos matemáticos relacionados son las matrices de Leslie y los procesos Semi-Markov (Howard, 1971).

5. Referencias

Bellamy, J.A. y Brown, J.R. (1994). Building a state and transition model for management and research on rangelands. *Tropical Grasslands*, **Volume 28**, pp. 247-255.

Grice, A.C. y Macleod, N.D. (1994). State and transition models as aids to communication between scientists and land managers. *Tropical Grasslands*, **Volume 28**, pp. 241-246.

Howard R.A. (1971). *Dynamic Probabilistic Systems. Volume II: Semi-Markov and Decision Process*. New York: John Wiley and Sons.

Leslie P.H. (1945). On the use of matrices in certain population mathematics. *Biometrika*, **Volume 33**, pp. 183-212.

Rodríguez, I.R., y Kothmann M.M. (1996). Structure of vegetation change in state and transition model applications. *Journal of range management*, **50**, 399-408.

Consideraciones para aplicar pruebas de equivalencia

Cecilia Ramírez Figueroa[4]

Colegio de Postgraduados. Montecillo, Texcoco Estado de México

David Sotres Ramos [5]

Colegio de Postgraduados. Montecillo, Texcoco Estado de México.

1. Introducción

Existen cuando menos dos motivaciones principales para recomendar pruebas de equivalencia. La primera es cuando el propósito del estudio no sea mostrar que los tratamientos son idénticos, (como en la hipótesis nula tradicional, H_0: $\mu_1 = \mu_2$) pero que las diferencias entre los tratamientos son demasiado pequeñas para ser consideradas significativas.

La segunda es que conforme el tamaño de muestra aumenta, la probabilidad de encontrar diferencias muy pequeñas de medias estadísticamente significativas se acerca a uno con la prueba tradicional de equivalencia de hipótesis nula, como la prueba t de Student para dos muestras independientes.

Algunos investigadores (Lin, 1995) han demostrado que cuando se prueba la equivalencia de dos tratamientos usando la metodología de la hipótesis nula de no equivalencia a menudo se llega a contradicciones. Es decir que las medias de los dos grupos experimentales se pueden declarar estadísticamente diferentes con la prueba t de Student, pero estadísticamente equivalentes usando la prueba de equivalencia de Schuirman.

Se plantea comparar el método para valorar la equivalencia de medias de dos grupos propuesto por Schuirmann (1987), con el método tradicional de la prueba t de Student.

[4] ceciliarf@colpos.mx
[5] sotres.davida@kendle.com

2. Metodología

Un estudio de simulación se usó para comparar la prueba de equivalencia de Schuirmann con la prueba t de Student para detectar la equivalencia de poblaciones. Las variables en este estudio fueron (a) tamaño de muestra, (b) configuración de la media poblacional, y (c) las varianzas de la población. La diferencia de medias crítica fue de $D = 1$ en todas las condiciones.

Método de Schuirman

La hipótesis alternativa de equivalencia

$$H_a : -D < \mu_1 - \mu_2 < D$$

es convenientemente expresada como la intersección de los dos conjuntos

$$\Theta_1^c = \{(\mu_1, \mu_2, \sigma_D^2) : \mu_1 - \mu_2 > -D\} \quad y \quad \Theta_2^c = \{(\mu_1, \mu_2, \sigma_D^2) : \mu_1 - \mu_2 < D\}$$

La prueba que rechaza $\mathrm{H}_{01} : \mu_1 - \mu_2 \leq -D$ si $t_1 \geq t_v^\alpha$ es de tamaño α donde

$$t_1 = \frac{(\overline{X}_1 - \overline{X}_2) - (-D)}{\sqrt{\left((n_1+n_2)[(n_1-1)s_1^2 + (n_2-1)s_2^2]\right)/(n_1 n_2 (n_1+n_2-2))}}$$

de igual manera, la prueba que rechaza $\mathrm{H}_{02} : \mu_1 - \mu_2 \geq D$ si $t_2 \leq -t_v^\alpha$ es de tamaño α. Así la prueba que rechaza H_0 solamente si ambas pruebas son rechazadas es una prueba de nivel α.

3. Resultados

Con 10 ó 25 observaciones por grupo, la prueba t de Student fue mejor que la prueba de Schuirmann en declarar dos medias de grupo equivalentes, sin considerar el tamaño de la diferencia de medias. Por otro lado, con 50 o 100 sujetos por grupo, la prueba de equivalencia de Schuirmann era más parecida a la t de Student. Cuando las varianzas de la población

se incrementaron en un grupo, la prueba de Schuirmann fue muy mala para detectar la equivalencia. Cuando las medias de grupo de población eran iguales, la prueba de t de Student encontró la equivalencia en aproximadamente el 95 % de las veces, sin que la afecte el tamaño de muestra, en cambio la prueba de Schuirmann se ve muy disminuida para muestras pequeñas. (Tabla 1 y Tabla 2).

Cuando se tiene a priori una diferencia de medias más grande que la diferencia crítica, la prueba de Schuirmann es muy exacta ($> 95\,\%$ en todos los casos) en detectar las diferencias. Sin embargo, esta habilidad superior es función del sesgo de la prueba para declarar las poblaciones no equivalentes incluso cuando las diferencias son más pequeñas que la diferencia crítica . El poder de la t de Student para detectar las diferencias de medias fue afectado por los tamaños de muestra y las varianzas, con el poder maximizado en los tamaños de muestra más grandes y las varianzas más pequeñas. (Tabla 3).

Los resultados muestran que el tamaño de muestra es un factor crucial en decidir entre la prueba de Schuirmann y la t de Student. Si el número de sujetos por tratamiento es grande (25 o más), la prueba de Schuirmann puede ser más apropiada para detectar la equivalencia de población que la t de Student, especialmente cuando las diferencias de media son menores que la diferencia crítica. Cuando el tamaño de muestra y la diferencia entre las medias aumentan, la prueba de t de Student es más poderosa para detectar diferencias. Cuando las varianzas por tratamiento son muy grandes, la habilidad del procedimiento de Schuirmann para detectar la equivalencia es reducida.

Tabla 1. Estudio de Simulación de la Probabilidad de Detectar a las Poblaciones Equivalentes Usando los Parametros ($\mu_1 = 65$, $\mu_2 = 65{,}7$, $\sigma_1 = 8$, $\sigma_2 = 9$, $D = 5$). De Seaman y Serlin (1997, p. 405)

Prueba estadística	Tamaño de muestra ($n_1 = n_2$)			
	25	50	75	100
Prueba de equivalencia de Schuirmann	0.311	0.768	0.913	0.974
Prueba t de Student	0.939	0.934	0.914	0.912

Tabla 2. Probabilidad de Detectar a las Poblaciones Equivalentes usando las Pruebas de Schuirmann [S] y la t de Student [t] ($D = 1$).

		$\sigma_1^2 = \sigma_2^2 = 1$		$\sigma_1^2 = \sigma_2^2 = 4$		$\sigma_1^2 = \sigma_2^2 = 8$	
n	$\mu_1 - \mu_2$	S	t	S	t	S	t
10	0	0.3930	0.9488	0.0248	0.9466	0.0026	0.9412
	0.4	0.2764	0.8576	0.0220	0.9076	0.0016	0.9234
	0.8	0.0966	0.5962	0.0114	0.7940	0.0008	0.8582
25	0	0.9384	0.9534	0.4426	0.9506	0.0698	0.9488
	0.4	0.6778	0.7174	0.3016	0.8314	0.0526	0.8922
	0.8	0.1698	0.2036	0.1056	0.5832	0.0308	0.7528
50	0	0.9996	0.9510	0.8660	0.9494	0.5230	0.9514
	0.4	0.9100	0.4962	0.5922	0.7676	0.3512	0.8530
	0.8	0.2562	0.0202	0.1544	0.2974	0.1174	0.5320
100	0	1.0000	0.9496	0.9946	0.9510	0.9083	0.9498
	0.4	0.9956	0.1922	0.8444	0.5688	0.6324	0.7330
	0.8	0.3998	0.0000	0.2388	0.0544	0.1714	0.2526

Tabla 3. Probabilidad de Detectar a las Poblaciones No Equivalentes usando las Pruebas de Schuirmann [S] y la t de Student [t] ($D = 1$).

		$\sigma_1^2 = \sigma_2^2 = 1$		$\sigma_1^2 = \sigma_2^2 = 4$		$\sigma_1^2 = \sigma_2^2 = 8$	
n	$\mu_1 - \mu_2$	S	t	S	t	S	t
10	1.2	0.9832	0.7254	0.9928	0.3598	0.9982	0.2290
	1.6	0.9980	0.9280	0.9974	0.5890	0.9992	0.3780
25	1.2	0.9900	0.9874	0.9834	0.7486	0.9886	0.5006
	1.6	0.9998	1.0000	0.9988	0.9414	0.9980	0.7446
50	1.2	0.9942	0.9998	0.9856	0.9584	0.9798	0.7928
	1.6	1.000	1.000	0.9998	0.9990	0.9990	0.9588
100	1.2	0.9984	1.000	0.9924	0.9994	0.9876	0.9764
	1.6						

4. Conclusiones

Si el número de sujetos por tratamiento es grande (25 o más), la prueba de Schuirmann puede ser más apropiada para detectar la equivalencia de población que la t de Student, especialmente cuando las diferencias de media son menores que la diferencia crítica y las varianzas son homogéneas.

Con los resultados mostrados se enfatiza la necesidad de reconocer que la prueba de equivalencia de Schuirmann y la prueba t de Student son diametralmente opuestas en su enfoque para poner a prueba hipótesis, por lo que se debe poner especial cuidado en la elección de un procedimiento de constraste de hipótesis.

Referencias

Lin S.C. (1995). Sample Size for Therapeutic Equivalence Based on Confidence Interval, *Drug Information Journal*, **29**, 45-50

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15, 657-680.

Seaman, M.A., & Serlin, R.C. (1998). Equivalence confidence intervals for two-group comparisons of means. Psychological Methods, 3, 403-411

Student (W.S. Gosset) (1908). "On the probable error of a mean". Biometrika 6, 1-25

Selección de modelos de supervivencia en la industria farmacéutica[1]

Rafael E. Borges[2]

Departamento de Estadística, Universidad de Los Andes, Mérida, Venezuela

1. Introducción

El propósito de este trabajo es alertar sobre posibles problemas que pudieran presentarse al utilizar el modelo de Cox y proponer modelos alternativos en casos de observarse violación de los supuestos u otros problemas más dificiles de identificar. Se presenta además un ejemplo aplicado a la industría farmacéutica.

2. Elementos del análisis de supervivencia

El análisis de supervivencia engloba un conjunto de técnicas para analizar el tiempo de seguimiento hasta la ocurrencia de un evento de interés, tiempo que puede observarse completa o parcialmente. La observación parcial puede ser debida a mecanismos de censura y truncamiento.

Sea T una variable aleatoria positiva (o no negativa) con función de distribución $F(t)$ y función de densidad de probabilidad $f(t)$. La función de supervivencia $S(t)$ se define como:

$$S(t) = 1 - F(t) = P[T > t]$$

Otra función importante es la función de razón de riesgos o tasa instantánea de fallas $\lambda(t)$, definida como la probabilidad de que a un individuo le ocurra el evento de interés en la siguiente unidad de tiempo Δt dado que ha sobrevivido hasta el tiempo t.

[1]Trabajo realizado con apoyos del Consejo de Desarrollo Científico, Humanístico y Tecnológico de la Universidad de Los Andes (CDCHT-ULA), Mérida, Venezuela, a través del proyecto E-199-02-09-C.

[2]borgesr@ula.ve

En análisis de supervivencia, el primer paso consiste en estimar la función de supervivencia, para lo cual existen varios estimadores, siendo el más utilizado el estimador de Kaplan y Meier (1958).

Otro aspecto importante es la utilización de modelos de regresión, que sean capaces de explicar la función de riesgo pudiendo incluirse variables explicativas, denominadas covariables. Algunos de estos modelos serán presentados en las próximas secciones.

2.1. El modelo de regresión de Cox

El modelo de regresión de Cox (1972) es el modelo de regresión más utilizado para datos de supervivencia en el área médica. En el modelo de regresión de Cox, el riesgo para el i-ésimo individuo se define mediante:

$$\lambda\left(t; Z_i\left(t\right)\right) = \lambda_0\left(t\right) e^{\beta' Z_i(t)}$$

donde $Z_i\left(t\right)$ es el vector de covariables para el i-ésimo individuo en el tiempo t.

El ajuste del modelo de Cox se hace a través del método de verisimilitud parcial. Una vez que se ha ajustado un modelo de Cox, existen tres contrastes de hipótesis para verificar la significación del modelo, estas pruebas son asintóticamente equivalentes pero no siempre sucede lo mismo en la práctica. Estos contrastes son el de razón de verosimilitud, el de Wald y el de los puntajes (score test).

2.2. Análisis de residuos, violación de supuestos, modelos alternativos y criterios para la selección de modelos

Una de las ventajas que han surgido del enfoque del análisis de supervivencia es la posibilidad de efectuar análisis de residuos (Therneau y Grambsch, 2000). Los residuos se pueden utilizar para verificar el supuesto de riesgo proporcional (supuesto más importante) (residuos de Schoenfeld), para identificar los individuos influyentes en la estimación del modelo (residuos

deviance), para identificar los individuos influyentes en la estimación del parámetro asociado a cada covariable (residuos score) y para descubrir la forma funcional correcta de un predictor continuo (residuos de martingala).

Existen otras situaciones en las que, a pesar de la no violación de los supuestos, pareciera ser necesario utilizar otros modelos. Por ejemplo, para los casos en que se crucen las funciones de supervivencia para dos o más grupos se pudiera utilizar un modelo de Cox Generalizado (Hseih, 2001, Bagdonavicius y Nikulin, 2001), un patrón quebradizo en el gráfico para verificar el supuesto de riesgo proporcional pudiera sugerir el uso de modelos locales de Cox y, algunos patrones cíclicos en este gráfico sugieren la posibilidad de utilizar modelos de Cox con covariables en el tiempo (Vaupel *et al*, 1979).

Otros modelos alternativos son el modelo aditivo de Aalen (1980), el modelo de Cox-Aalen y otros modelos dinámicos (Martinussen y Scheike, 2006), el modelo de Gray (1992), el modelo de regresión de Bukley y James (1979), los modelos acelerados (Bagdonavicius y Nikulin, 2001).

Exite otra clase de modelos basados en familias paramétricas (exponencial, Weibull, logística, normal, log-logística, log-normal, valor extremo, entre otras), algunas de las cuales pueden ser sugeridas mediante gráficos de probabilidad o mediante algunos gráficos específicos (Meeker y Escobar, 1998).

Un aspecto importante a considerar es que cada modelo tiene unos supuestos que deben ser verificados. Pudiera además presentarse el caso de tener varios modelos, y para la selección del mejor de ellos pueden utilizarse varios criterios, siendo el más utilizado el criterio de información de Akaike.

3. Ejemplo

En esta seccíon se presenta un ejemplo aplicado a la industría farmacéutica. Los datos han sido tomados de un informe de asesoría llevada a cabo por Borges (2004).

El análisis efectuado consistió en la estimación de la función de supervivencia mediante el

estimador de Kaplan y Meier, y el ajuste del modelo de Cox, incluyendo además la verificación de los residuos.

El modelo de Cox obtenido incluía a las covariables DOSIS y POSTDATISMO. Tanto el modelo como las variables resultaron significativas y no se encontró violación de los supuestos del modelo.

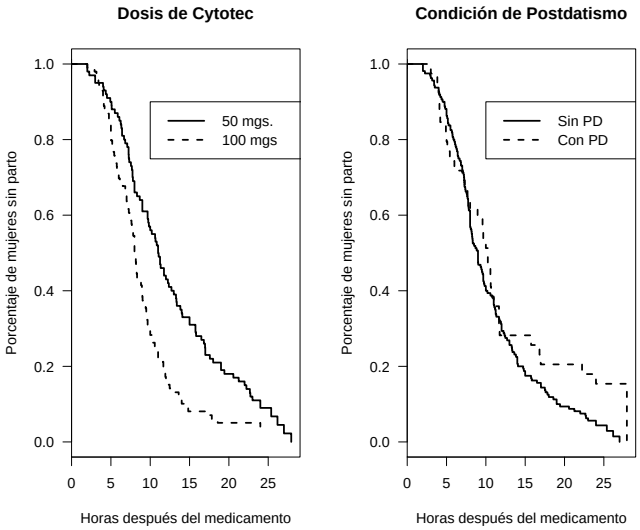


Gráfico 1. Función de Supervivencia estimada a través del estimador de Kaplan y Meier.

En el gráfico 1 se observa, por una parte que el medicamento reduce significativamente el tiempo de trabajo de parto (resultado similar al del modelo de Cox) y, el entrecruzamiento de las curvas para el postdatismo, indica que el modelo de Cox no es del todo adecuado, por lo que debería ajustarse otro modelo (modelo de Cox generalizado o modelo aditivo de Aalen).

4. Referencias

Aalen, O.O. (1980). A Model for Non-parametric Regression Analysis of Counting Processes. In *Lecture Notes in Statistics 2* (eds. W. Klonecki *et al*), pp. 1-25. New York: Springer-Verlag.

Bagdonavicius, V. y Nikulin, V. (2001). *Accelerated Life Models: Modeling and Statistical Analysis*. London: Chapman & Hall.

Borges, R. (2004). Estudio Comparativo de dos dosis del Medicamento Cytotec como Potencial Fármaco para la Inducción del Parto, *Informe Técnico de Asesoría*, Mérida, Venezuela: Grupo de Investigación en Bioestadística de la Universidad de Los Andes.

Buckley, J.J. y James, I.R. (1979). Linear regression with censored data, *Biometrika*, **66**, 429-436.

Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society: Series B*, **34**, 187-220.

Gray, R.J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis, *Journal of the American Statistical Association*, **87**, 942-951.

Hsieh, F. (2001). On heteroscedastic hazard regression models: theory and application, *Journal of the Royal Statistical Society: Series B*, **61**, 63-79.

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

Martinussen, T. y Scheike, T.H. (2006). *Dynamic Regression Models for Survival Data*. New York: Springer-Verlag.

Meeker, W.Q. y Escobar, L.A (1998). *Statistical Methods for Reliability Data*. New York: Wiley.

Therneau, T.M. y Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Vaupel, J., Manton, K. y Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16**, 439-454.

Uso de distribución de valores extremos para investigar tendencias en niveles muy altos de ozono

Hortensia J. Reyes Cervantes. [1]
Colegio de Postgraduados
Humberto Vaquera Huerta [2]
Colegio de Postgraduados
José A. Villaseñor A. [3]
Colegio de Postgraduados

1. Introducción

El ozono es un contaminante secundario producto de la reacción de óxidos de nitrógeno, hidrocarburos reactivos y la luz ultravioleta. También es altamente oxidante y forma parte de una mezcla compleja de contaminantes fotoquímicos asociado a radicales libres como lo menciona Seinfeld (1991).

La zona Metropolitana de la Ciudad de México (ZMCM) es considerada recientemente como la metrópoli con más problemas de contaminación en el país, y quizá en el mundo (Bravo et al, 1992). Se estima que poco más de 9.5 millones de habitantes están expuestos a diferentes grados de contaminación por ozono, ya que frecuentemente se encuentra por arriba de la norma oficial mexicana de 0.11 ppm promedio máximo por hora. En el caso particular de la zona metropolitana de la ciudad de México, los aspectos geográficos, las condiciones socioeconómicas de sus habitantes, así como la sobre explotación de recursos hídricos y forestales contribuyen de manera substantiva a explicar los altos niveles de ozono.

Un problema importante en muchos modelos actuales es la estimación de las tendencias en los niveles de ozono, ya que esto permite la evaluación de las medidas ambientales tomadas por las autoridades con el fin de disminuir los niveles de contaminación. Smith R.(1989)

[1] hreyes@colpos.mx
[2] hvaquera@colpos.mx
[3] jvillasr@colpos.mx

muestra un procedimiento basado en la teoría de valores extremos que maneja el análisis de la excedencia sobre umbrales altos con el fin de investigar tendencias en los niveles altos de ozono urbano.

En este trabajo se presenta una propuesta para investigar las tendencias en los niveles muy altos de ozono basada en la teoría de valores, se ilustra la metodología aplicándose al caso particular de la contaminación ambiental de la Ciudad de México.

2. La Teoría de Valores Extremos.

Una expresión usada por Reiss y Tomas (2001) de la distribución de valores extremos generalizada (GEV), involucra los parámetros: de localización μ, de escala σ y de forma ϵ, la expresión es:

con $(\mu, \sigma, \epsilon) \in \Re$ x $\Re_+$x $\Re$.

$$G_{\mu,\sigma,\epsilon}(z) = \begin{cases} exp\left\{-\left[1+\epsilon\frac{z-\mu}{\sigma}\right]^{-\frac{1}{\epsilon}}\right\} & \text{si } 1+\epsilon\frac{z-\mu}{\sigma} > 0, \\ exp\left[-exp^{-\frac{z-\mu}{\sigma}}\right] & \text{si } \epsilon = 0. \end{cases} \tag{1}$$

Fisher (1928) dice que una variable aleatoria X tiene una función de distribución $F(X)$ que pertenece al dominio de atracción de una distribución de Valor Extremo Generalizado $G(\theta)$, si existen constantes $c_n > 0$ y $d_n \in \Re$ tales que $M_n = max\{X_1, ..., X_n\}$ donde $X_1, ..., X_n$ es una muestra aleatoria tomada de $F(X)$, por lo cual

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} G_\theta$$

El interés es encontrar la distribución de M_n. Basándose en el teorema de Fisher-Tippet, se tiene que el límite de la función de distribución del valor máximo normalizado (si existe), es la función de distribución de la VEG,

$$G_\theta = G_{\mu,\sigma,\epsilon}.$$

Suponiendo que las distribuciones VEG son continuas y diferenciables, y además, existen las

constantes c_n y d_n. Se puede hacer la inferencia en términos de la función de verosimilitud (Coles,2001).

Smith (1985) señala que para la distribución VEG, cuando $\epsilon > -0{,}5$ los estimadores de máxima verosimilitud tienen propiedades asintóticas usuales; si $-1 < \epsilon < -0{,}5$ los estimadores son fáciles de obtener pero no tienen propiedades asintóticas y $\epsilon < -1$ los estimadores de máxima verosimilitud no existen.

El término cuantil se refiere a que dado un valor $F_X(x;\theta)$ es el valor q_α tal que $\alpha = P(X \leq q_\alpha) = F_X(q_\alpha;\theta) \leftrightarrow F_X^{-1}(\alpha;\theta) = q_\alpha$ es la función cuantil de $F_X(x;\theta)$.

En términos del tema en cuestión, el q_α, es un nivel de ozono, tal que para valores mayores de ese valor se considera que hay un $(1-\alpha)\,\%$ de observaciones más altas de ozono. Suponiendo que $Y_1, ...Y_n$ tienen una distribución VEG y tomando su Función Cuantil para un umbral alto en las observaciones, con α fijo (Coles, 2001).

$$Z_\alpha(p) = \begin{cases} \mu - \frac{\sigma}{\epsilon}(1 - y_p^{-\epsilon}) \text{ si} & \epsilon \neq 0, \\ \mu - \sigma log(y_p) \quad \text{si} & \epsilon = 0. \end{cases} \tag{2}$$

donde $G(z_\alpha(p)) = 1 - p$ con $y_p = -log(1-p) = -log((z_\alpha))$, para un p fijo y además se conoce a $Z_\alpha(p)$ como el nivel de retorno asociado al periodo $\frac{1}{p}$.

Si se supone que se tiene una muestra grande, la expresión anterior se puede simplificar para conocer la distribución de probabilidad de $Z_\alpha(p)$. Finalmente se estima por máxima verosimilitud la expresión final (Coles, 2001), quedando

$$\hat{Z}_\alpha(p) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\epsilon}}(1 - y_p^{-\hat{\epsilon}}) \quad \text{si} \quad \epsilon \neq 0. \tag{3}$$

Por lo anterior, se tienen las condiciones para el siguiente resultado que permite conocer la distribución asintótica de $\hat{Z}_\alpha(p)$.

Resultado: Sean $Y_1, ...Y_n$ observaciones independientes idénticamente distribuidas con función de distribución GEV. Para $y_p = -log(1-p) = -log(G_\theta(z_\alpha(p)))$ y $\alpha = 1 - p$ fijo, $0 < p < 1$,

sea $\hat{Z}_\alpha(p)$ un estimador en (3) de la Función Cuantil $Z_\alpha(p)$, donde μ es conocida y $\hat{\sigma}, \hat{\epsilon}$ son los estimadores de máxima verosimilitud.

Entonces

$$\sqrt{n}\{z(\hat{\sigma},\hat{\epsilon}) - [\mu - \sigma log(y_p) + \frac{(log(y_p))^2}{2}(cov(\hat{\sigma},\hat{\epsilon}) + \sigma\varepsilon)]\} \xrightarrow[n\to\infty]{d}$$

$$N(0, [log(y_p) - \frac{\epsilon(log(y_p))^2}{2}]^2 var(\hat{\sigma}) + [\frac{(log(y_p))^2\sigma}{2}]var(\hat{\epsilon}) -$$

$$2[log(y_p) - \frac{\epsilon(log(y_p))^2}{2}\frac{(log(y_p))^2\sigma}{2}cov(\hat{\sigma},\hat{\epsilon})])$$

Demostración. Se usa el método delta para la expresión (2)

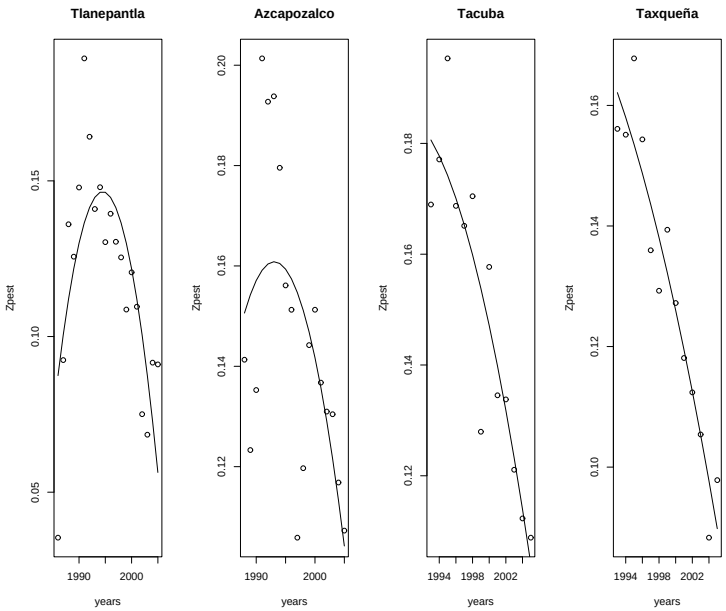
3. Ejemplo de aplicación

La información utilizada fue tomada de SIMAT. Los datos analizados se agruparon para formar bloques de observaciones durante 4 días y se determinó el máximo de cada bloque (Villaseñor, 1996). Se presentan estimaciones de los valores del ozono a través de los años en algunas Estaciones Meteorológicas de la Ciudad de México, usando Regresión Cuantil en $p = 0{,}05$.

4. Conclusiones

El uso de este procedimiento, sólo es para muestras grandes, se puede utilizar en el Modelo de Regresión Lineal para analizar tendencias en los niveles de ozono.

Se revisaron algunos métodos no paramétricos para estimar tendencias a los parámetros desconocidos (μ, σ, ϵ). Para el caso de tener ϵ se utiliza el estimador de Pickand o de Hill (Panaretos, 2003). Para muestras grandes se observó que la distribución resultante es una distribución normal y que los parámetros de escala y de forma están muy ligados, de donde no es sencillo encontrar los valores para los que ocurre la convergencia.



5. Referencias

Bravo et al . (1992). *La Contaminación Atmosférica por ozono en la Ciudad de México. La Contaminación Atmosférica en México. Sus causas y sus efectos*. Comisión Nacional de Derechos Humanos, México.

Coles, S. (2001). *An Introduction to statistical Modeling of Extreme Values*. Springer Verlag.

Lou, T., Reynolds J., Cox L., Guttorp P. y Sampson P. (2001). A review of Statistical Methods for the Meteorological adjustment of tropospheric ozone. *Environment*, **35**, 617-630.

Molina M. y Molina L. (2002). *Air Quality in the Mexico: An Integrated Assessment*. Kluwer Academia Publishers.

Paranetos y Zoi (2003). Extreme Values Index Estimators and Smoothing alternatives: A critical review, Stochastic Musings Perspectives from the Pionners of the late 20 .

Seinfeld J. (1991), *Comittee on troposfheric ozone formation and measurement; Board on*

Environment Studies and Taxicology; Board on Atmospheric Siences and Climate; Commission on Geosciences, Environment, and Resources; National Research Council, Rethinking the ozone problem in urban and regional air Pollution National, Academic Press.

Smith R. (1995). Model Selection in Environmental Statistic, *National Institute of Statistical Sciences*, Technical Report 32.

Villaseñor,A.(1996). A model for cluster maxima of excedances over a threshold for ozone data, *INEGI*, **8-10**, 7-16.

Muestreo de respuestas aleatorizadas en poblaciones finitas: un enfoque unificador[1]

Víctor Soberanis Cruz[2]
Universidad de Quintana Roo
Gustavo Ramírez Valverde[3]
Colegio de Postgraduados
Sergio Pérez Elizalde[4]
Colegio de Postgraduados
Félix González Cossio[5]
Colegio de Postgraduados

1. Introducción

La técnica de respuesta aleatorizada (RA) introducida por Warner (1965) es una propuesta de solución para la protección de la confidencialidad del entrevistado y consiste en la utilización de un mecanismo aleatorio (MA) por medio del cual se selecciona una de dos preguntas: ¿Pertenece al grupo con la característica A? o ¿Pertenece al grupo que no tiene la característica A?, donde A es la característica sensible de interés. El entrevistado contestará si o no y el entrevistador no tiene la posibilidad de saber qué pregunta contestó el entrevistado; protegiendo así la confidencialidad del mismo. Las principales técnicas de respuesta aleatorizada que se han propuesto son: a) el modelo W (Warner, 1965), b) el modelo U con pregunta inocua W no relacionada (Greenberg et al., 1969), c) el modelo C, d) el modelo H (Horvitz et al., 1976), e) el modelo D (Devore, 1977) y, f) el modelo M (Mangat y Singh, 1990). En este trabajo se propone un nuevo enfoque que utiliza la información contenida en la correlación entre la variable de interés y y una variable inocua W, enfoque que denominaremos modelo C. Así mismo, bajo un muestreo de poblaciones finitas y en el marco de la

[1]Trabajo realizado con apoyos del Colegio de Postgraduados y la Universidad de Quintana Roo
[2]vsobera@uqroo.mx
[3]gramirez@colpos.mx
[4]sergiop@colpos.mx
[5]felixgc@colpos.mx

teoría de los estimadores π (Särndal et al., 1992; Wretman et al., 1977), se generalizan varios de los modelos propuestos de RA por medio del modelo G. Además, se obtienen las varianzas de los estimadores en los distintos modelos y se observa que, bajo ciertas restricciones, el estimador para el modelo C es más eficiente que los correspondientes a los otros modelos.

2. Modelo general

2.1. Población bajo estudio

Consideremos una población finita $U = \{1, 2, ..., N\}$. En este trabajo, el tamaño de la población N se supondrá conocido. El tamaño de la muestra lo denotaremos por n, el cual no necesariamente es fijo.

Sea y la variable dicotómica que denota la pertenencia de un individuo al grupo con la característica sensible de interés, con y_k el valor de y para el k-ésimo elemento de la población. Así y_k es desconocida pero no aleatoria. Además, $y_k = 1$ si el k-ésimo individuo de la población tiene la característica sensitiva A y $y_k = 0$ si el k-ésimo individuo no tiene la característica sensitiva A. Lo que se desea estimar es $t_A = \sum_U y_k$, el total de los individuos en la población con la característica sensible A.

2.2. Procedimiento de muestreo

Para el modelo general (modelo G) el procedimiento de muestreo es como sigue:

Etapa 1 (selección de la muestra). Se extrae una muestra de tamaño n de acuerdo al diseño de muestreo $p(s)$ con probabilidades positivas de inclusión π_k y π_{kl} donde

$$\pi_k = \Pr\{S \ni k\} = \sum_{s \ni k} p(s) \quad \text{y} \quad \pi_{kl} = \Pr(S \ni k \& l) = \sum_{s \ni k \& l} p(s).$$

Para cada elemento k en la muestra S se tiene $I_k = 1$ si $k \in S$, $I_k = 0$ de otra forma; nótese que $I_k(S)$ es función de la variable aleatoria S. Además, $\pi_{kk} = \pi_k$, $\breve{y}_k = \frac{y_k}{\pi_k}$, $\breve{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$ y

$$\begin{aligned} \Delta_{kl} &= Cov(I_k, I_l) = E_p(I_k I_l) - E_p(I_k) E_p(I_l) \\ &= \pi_{kl} - \pi_k \pi_l \leq 0, k \neq l; \Delta_{kk} = \pi_k(1 - \pi_k); \end{aligned} \tag{1}$$

Etapa 2 (recopilación de la información). Las entrevistas se realizan a los individuos en la muestra de acuerdo al MA definido por el modelo de respuesta aleatorizada empleado. El MA induce para cada $k \in S$ una variable aleatoria Z_k tal que la combinación lineal $\hat{Z}_k = aZ_k + b_k$ es una estimación insesgada de y_k, donde a y b_k son constantes conocidas que dependen del MA; por tanto, $E_{MA}\left(\hat{Z}_k\right) = y_k$, $V_{MA}\left(\hat{Z}_k\right) = a^2 V_{MA}(Z_k)$, $E_{MA}(b_k) = b_k \forall k \in S$ y el cálculo de $V_{MA}(Z_k)$ también depende de MA.

2.3. Enfoque unificador (modelo G)

El estimador general que se propone para la estimación del total $t_A = \sum_U y_k$ es:

$$\hat{t}_{AG,\pi} = \sum\nolimits_S \frac{\hat{Z}_k}{\pi_k} \quad . \tag{2}$$

Se tiene que $E\left(\hat{t}_{AG,\pi}\right) = t_A$, por lo que el estimador es insesgado bajo el diseño de muestreo $p(s)$. También,

$$V\left(\hat{t}_{AG,\pi}\right) = \sum\sum\nolimits_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + a^2 E_p\left(\sum\nolimits_S \frac{V_{MA}(Z_k)}{\pi_k^2}\right) \tag{3}$$

Nótese que la varianza para cualquier modelo que cumpla con las condiciones del modelo G está formada por dos términos, el primero depende del diseño de muestreo $p(s)$ y los valores y_k, esta parte es común a todos los modelos y será denotada por V_G; el segundo término depende del mecanismo aleatorio empleado. Por tanto, para comparar la varianza de los distintos modelos es suficiente la comparación del segundo término de la varianza. Se

puede verificar que todos los modelos considerados son casos particulares del modelo G. En el Cuadro 1 se muestran los coeficientes asociados a cada modelo y en el Cuadro 2 se dan las varianzas de los estimadores de t_A.

Cuadro 1: Resumen de las constantes asociadas a los distintos modelos considerados

Modelo	a	b
W	$\frac{1}{2p-1}$	$-\frac{1-p}{2p-1}$
H	$\frac{1}{p_1}$	$-\frac{p_2}{p_1}$
U, C	$\frac{1}{p}$	$-\frac{(1-p)w_k}{p}$
D	$\frac{1}{p}$	$-\frac{(1-p)}{p}$
M	$\frac{1}{t+(1-t)(2p-1)}$	$-\frac{(1-t)(1-p)}{t+(1-t)(2p-1)}$

Cuadro 2: Ecuaciones para los estimadores y sus respectivas varianzas

Modelo	Estimador	Varianza del estimador
W	$\frac{1}{2p-1}\left[\sum_S \frac{Z_k}{\pi_k} - (1-p)\sum_S \frac{1}{\pi_k}\right]$	$V_G + V_0\left(\sum_U \frac{1}{\pi_k}\right)$
U, C	$\frac{1}{p}\left[\sum_S \frac{Z_k}{\pi_k} - (1-p)\sum_S \frac{w_k}{\pi_k}\right]$	$V_G + \frac{1-p}{p}\sum_U \frac{(w_k-y_k)^2}{\pi_k}$
H	$\frac{1}{p_1}\left[\sum_S \frac{Z_k}{\pi_k} - p_2\sum_S \frac{1}{\pi_k}\right]$	$V_G + \frac{1}{4}\frac{1-p_1^2}{p_1^2}\left(\sum_U \frac{1}{\pi_k}\right)$
D	$\frac{1}{p}\sum_S \frac{Z_k}{\pi_k} - \frac{1-p}{p}\sum_S \frac{1}{\pi_k}$	$V_G + \frac{1-p}{p}\left(\sum_U \frac{1-y_k}{\pi_k}\right)$
M	$\frac{1}{t+(1-t)(2p-1)}\sum_S \frac{Z_k}{\pi_k} - \frac{(1-t)(1-p)}{t+(1-t)(2p-1)}\sum_S \frac{1}{\pi_k}$	$V_G + \frac{\beta(1-\beta)}{(2\beta-1)^2}\sum_U \frac{1}{\pi_k}$, $\beta = (1-p)(1-t)$

3. Resultados y conclusiones

Los modelos considerados en este trabajo son generalizados mediante el modelo G (Cuadro 1), lo que permite comparar las varianzas de los estimadores de los mismos. La expresión de la varianza del estimador del total en el modelo C indica la existencia de una relación inversa entre la varianza del estimador y la correlación entre las variables sensitiva e inocua (Cuadro 2). De aquí que estudios de simulación muestran que, para un conjunto amplio de valores de los parámetros de los distintos MA, el estimador $\hat{t}_{A6,\pi}$ para el modelo C tiene menor varianza que en los demás modelos. Esto es, se logra una reducción muy importante en la varianza del estimador $\hat{t}_{A2,\pi}$ del modelo U al considerar una variable W inocua pero altamente correlacionada con la variable sensible y.

4. Referencias

Cassel, C. M., Wretman, J. K. y Särndal, C. E. (1977). *Foundations of Inference in Survey Sampling*. Wiley. New York.

Devore, J. L. (1977). A note on the randomized response technique. *Communications in Statistics Theory and Methods* **6**, 1525-1529.

Greenberg, B. G., Abul-ela, A. A., Simmons, W. R. y Horvitz, D. C. (1969). The unrelated question RR model: theoretical framework. *Journal of the American Statistical Association* **64**, 520-539.

Horvitz, D. C., B. G. Greenberg y J. R. Abernathy (1976). Randomized response. A data gathering device for sensitive questions. *International Statistical Review* **44**, 181-196.

Mangat, N. S. y Singh R. (1990). An alternative randomized response procedure. *Biometrika* **77**, 439-442.

Särndal, C. E., B. Swensson y J. Wretman (1992). *Model Assisted Survey Sampling*. Springer Verlag. New York. 694 p.

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

Utilización de un paquete de cómputo matemático en apoyo a la enseñanza de la estadística y la probabilidad[1]

Agustín Jaime García Banda[2]
Facultad de Ciencias Administrativas y Sociales, U.V.
Luis Cruz-Kuri
Instituto de Ciencias Básicas, U.V.
Ismael Sosa Galindo
Facultad de Ciencias Administrativas y Sociales, U.V.

1. Introducción

En algunos cursos de estadística y probabilidad básica se utilizan tablas previamente preparadas tales como las de la distribución binomial, la distribución normal, la distribución "t" de Student, etc. Con dicho enfoque se tiene un esquema rígido que limita los alcances de las aplicaciones, particularmente en los cursos más avanzados. En el presente trabajo se hace ver cómo mediante la utilización de un paquete de cómputo matemático se pueden realizar fácilmente los procesamientos relevantes a una variedad más amplia de modelos probabilistas. En nuestro caso hemos decidido hacer las ilustraciones mediante la utilización del paquete *Mathematica*, aunque, por supuesto, se tienen otras opciones tales como *Maple V*, *Matlab*, tan solo por mencionar algunos. Los procesamientos que se pueden hacer son de tipo gráfico, algebraico, así como numérico, permitiendo en este último caso la elaboración de tablas personalizadas.

[1]Trabajo realizado con apoyos de la Universidad Veracruzana
[2]jaimegarciabanda@yahoo.com

2. Estudio de Algunos Modelos Probabilistas Clásicos Vía Mathematica

En esta sección se presentan ejemplos de modelos probabilistas clásicos, tanto de tipo continuo como de tipo discreto. Se proporcionan las fórmulas matemáticas para las correspondientes funciones de densidad así como para las probabilidades discretas. Lo anterior se ilustra con la familia de la distribución normal, la del modelo "t" de Student, la familia de la distribución binomial y la familia de la distribución hipergeométrica. Por supuesto, sin dificultades adicionales, se puede hacer el estudio de algunos otros modelos clásicos, pero nuestro propósito no es el de hacer una presentación exhaustiva sino únicamente indicar cómo mediante la utilización de un programa de computo matemático se puede hacer un estudio gráfico y aritmético de algunas propiedades de los modelos probabilistas que se desee considerar. El programa de cómputo que se utilizó es *Mathematica*, pero fácilmente se pueden hacer las conversiones con otros programas de propósitos similares.

Estudio de la Distribución Normal para distintos valores de sus parámetros

Con instrucciones del programa *Mathematica* se ingresa la familia de funciones de densidad de la distribución normal, la cual dependerá de dos parámetros, a saber, b $>$ 0 parámetro de escala y a $\in$ R. Se tiene,

$$f(x) = \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-a}{b})^2} \tag{1}$$

Las instrucciones en Mathematica para la declaración de esta familia están dadas por

$$f[x\_, a\_, b\_] := Exp[-((x-a)/b)^2/2]/(Sqrt(2*Pi)*b) \tag{2}$$

Si se ejecuta la instrucción de *Mathematica* siguiente:

f[x_,a_,b_] (3)

se obtiene la fórmula simbólica en (1) de la familia de la distribución normal.

A continuación se presentan las instrucciones en *Mathematica* que permiten generar las gráficas de funciones de densidad normal, con igual medida de ubicación y valores varios de la desviación estándar; es decir, con las instrucciones que siguen, al ser ejecutadas, el programa Mathematica produce como salida las gráficas que aparecen en las figuras 1 al 3.

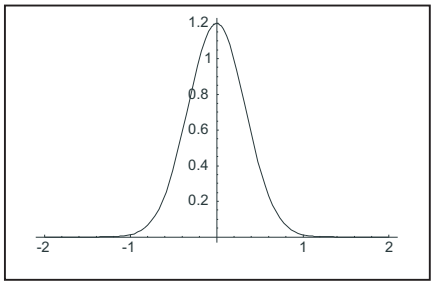
Instrucciones

Plot[f[x,0,1],{x,-2,2}] (Out[5]) (4)

Plot[f[x,0,1/3],{x,-2,2}] (Out[6]) (5)

Plot[f[x,0,1/2],{x,-2,2}] (Out[7]) (6)

Show[Out[5],Out[6],Out[7]](Out[8]) (7)



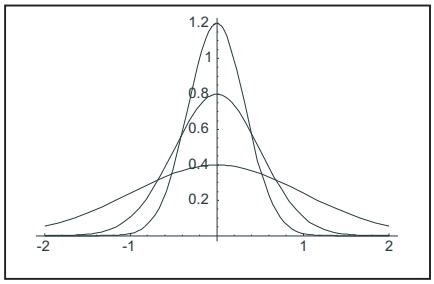


Fig. 1. Salida gráfica de la función de densidad normal con parámetros a = 0, b = 1/3 para valores de la variable en el intervalo [-2, 2]. Ver instrucción (3). (Izquierda)

Fig. 2. Salida gráfica de las funciones de densidad, presentadas simultáneamente, para valores de los parámetros a = 0 y b = 1, 1/2, 1/3. Ver instrucción (5). (Derecha)

Con las instrucciones que siguen, se generan conjuntamente varias gráficas de funciones de densidad con el mismo parámetro de escala cada una y con distintos valores para el parámetro de ubicación.

Show[Plot[f[x,0,1],{x,-2,2}],Plot[f[x,1,1],{x,-2,2}],Plot[f[x,2,1],{x,-2,2}]] (8)

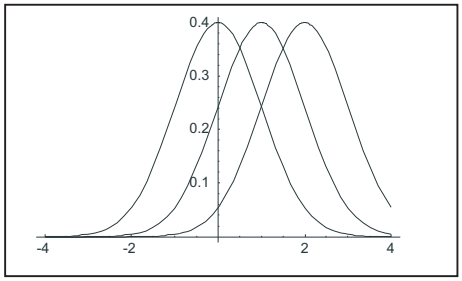
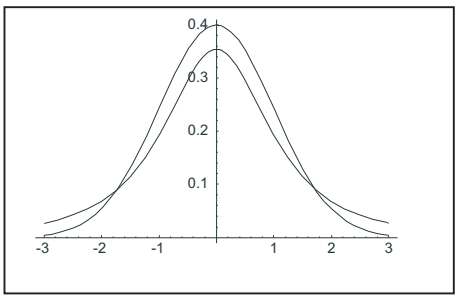


Fig. 3. Tres curvas normales con parámetro de escala b = 1, cada una, y con parámetros de ubicación a=0, 1, 2, respectivamente.

La Distribución "t" de Student

Con la función gama, la cual se encuentra en el repertorio interno del programa *Mathematica*, se pueden generar números factoriales tanto para cantidades enteras como fraccionarias; con la fórmula matemática de la familia de funciones de densidad que obedecen el modelo "t" de Student se puede declarar de manera inmediata con instrucciones de Mathematica dicha familia. Esto nos permite presentar, si así se desea, gráficas de densidades varias de dicho modelo para valores selectos del parámetro que se conoce como "grados de libertad". Las instrucciones son las que se presentan en la figura 4. Dicha gráfica está generada con instrucciones del programa y permite visualizar conjuntamente tanto la densidad de la distribución normal estándar como la del modelo "t" de Student con un número seleccionado de grados de libertad (2 para nuestra ilustración).



Instrucciones

```
c[n_] := Gamma[(n + 1) / 2] / (Sqrt[n * Pi] * Gamma[n / 2])

student[t_, grados_] :=
 c[grados] * (1 + (t^2 / grados))^-(grados+1)/2

Plot[student[t, 2], {t, -2, 2}]
```

Fig. 4. Comparación de las distribuciones de probabilidad Normal (ver fig. 1) y "t" de Student con 2 grados de libertad. Ver instrucciones (2) y las que aparecen a la derecha de la presente gráfica.

Construcción de Tablas de Probabilidad

Las instrucciones en *Mathematica* que se presentan a continuación hacen un manejo aritmético del programa utilizando en particular su capacidad de integración numérica y la iteración de dicho algoritmo para valores varios del límite superior de integración con incrementos especificados. Lo anterior permite la construcción de una tabla original de probabilidades acumuladas para cualquier modelo probabilista seleccionado. En nuestro caso, para propósitos de ilustración, se escogió el modelo de la distribución normal estándar con probabilidades acumuladas desde -∞ hasta z, para valores selectos de z. Es pertinente mencionar que la precisión aritmética se puede variar.

Instrucciones

```
Table[NIntegrate[f[x,0,1],{x,0,0.1i+0.1j}], {i,0,10},{j,0,10}]//MatrixForm
```
(9)

Tabla 1. Generada con instrucciones del paquete Mathematica para probabilidades acumuladas del modelo normal estándar.

0.000000	0.039828	0.079260	0.117911	0.155422	0.191462	0.225747	0.258036	0.288145	0.315940
0.039828	0.079260	0.117911	0.155422	0.191462	0.225747	0.258036	0.288145	0.315940	0.341345
0.079260	0.117911	0.155422	0.191462	0.225747	0.258036	0.288145	0.315940	0.341345	0.364334
0.117911	0.155422	0.191462	0.225747	0.258036	0.288145	0.315940	0.341345	0.364334	0.384930
0.155422	0.191462	0.225747	0.258036	0.288145	0.315940	0.341345	0.364334	0.384930	0.403200
0.191462	0.225747	0.258036	0.288145	0.315940	0.341345	0.364334	0.384930	0.403200	0.419243
0.225747	0.258036	0.288145	0.315940	0.341345	0.364334	0.384930	0.403200	0.419243	0.433193
0.258036	0.288145	0.315940	0.341345	0.364334	0.384930	0.403200	0.419243	0.433193	0.445201

Binomial

Las instrucciones que aparecen a continuación, permiten declarar la función de probabilidad de la familia de la distribución binomial con parámetros n=1, 2, 3, ... y $\pi \in [0,1]$. Se hace énfasis en que se utilizan funciones primitivas del paquete *Mathematica*, tales como las de la función gama, etc; con esto es posible definir funciones personalizadas tales como las de coeficiente combinatorio, que aunque existen en el repertorio, siempre es instructivo el generarlas a partir de las funciones primitivas.

```
factorial[n_]:=Gamma[n+1]
```
(10)

combinatorio[n_,j_]:=(factorial[n]/(factorial[j]*factorial[n-j]) (11)

binomial[n_,p_,k_]:=combinatorio[n,k]*p^k*(1-p)^n^-^k (12)

Table[binomial[10,1/2,k],{k,0,10}]//MatrixForm (13)

Generación de Probabilidades Hipergeométricas.

Utilizando funciones primitivas, se puede especificar un modelo de tipo discreto con valores selectos de los parámetros que intervienen. Para nuestros propósitos de ilustración, las instrucciones que siguen permiten generar probabilidades de un modelo hipergeométrico, así como la gráfica correspondiente en el formato de histograma de probabilidad.

Instrucciones

combina[n_,k_]:=Factorial[n]/(Factorial[k]*Factorial[n-k]) (14)

hipergeo[npob_,ncar_,nmuestra_,k_]:=combina[ncar,k]*combina[npob-ncar,

nmuestra-k]/combina[npob,nmuestra] (15)

Table[hipergeo[20,8,6,k],{k,0,6}] (16)

Tabla 2. Probabilidades del Modelo Hipergeométrico con Parámetros N=20, N_1=8, n=6

0.02383900929,0.1634674923,0.3575851393,0.3178534572,
0.1191950464,0.01733746130,0.0007223942208

Objetos Tridimensionales

El programa *Mathematica* permite la graficación de objetos tridimensionales. Las instrucciones que aparecen a continuación se usan para generar la gráfica de una superficie inmersa en un espacio de tres dimensiones. Para nuestros propósitos se ilustra con la función de densidad de una distribución normal bivariada, donde el parámetro de correlación es nulo.

Instrucciones

Plot3D[Exp[$-1{,}5x^2 - y^2$],{x,-2,2},{y,-2,2},PlotPoints →100,ViewPoint→ {1,1,0},

AxesLabel→{'Estatura','Peso','Densidad'},LightSources→{{{1.‘,0.‘,1.‘},RGBColor[0,1,0]},

{{1.‘,1.‘,1.‘},RGBColor[1,0,0]},{{0.‘,1.‘,1.‘},RGBColor[1,0,0]}}] (17)

La gráfica que aparece abajo es monocromática; aunque las instrucciones utilizadas permiten selección de colores, perspectiva desde donde la gráfica se puede observar y otros detalles.

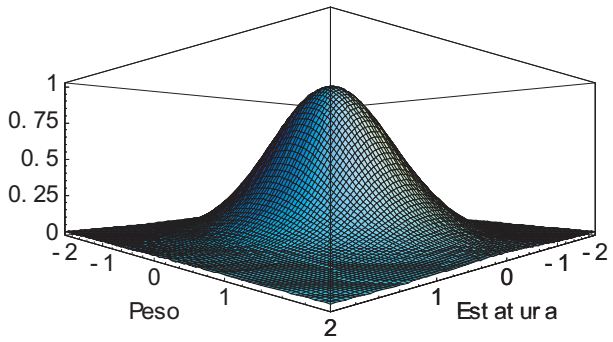
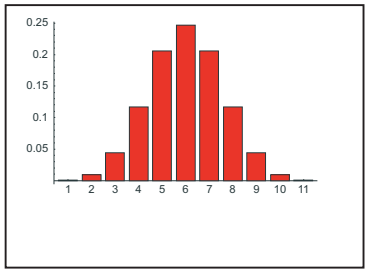
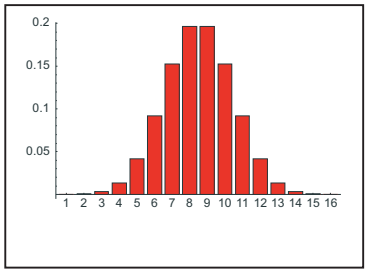


Fig. 5. Distribución normal bivariada. Superficie generada por el programa Mathematica.

3. Ilustración Gráfica del Teorema Central de Límite

Utilizando la capacidad gráfica del programa *Mathematica*, a continuación se presenta una serie de histogramas de la familia de la distribución binomial para valores crecientes de n, así como para distintos valores del parámetro. Ver figuras 6 a 8 que aparecen a continuación. De manera semejante, podría utilizarse otra familia de modelos probabilistas discretos, tales como la familia hipergeométrica, la familia de la distribución binomial negativa, etc. Por supuesto, algo análogo puede realizarse para modelos de tipo continuo, tales como la familia de distribuciones uniformes, de distribuciones exponenciales, etc. Por consideraciones de espacio se evitan los detalles.





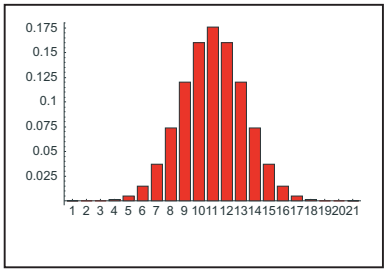


Fig. 6. Histograma Binomial generado por el programa *Mathematica* (n=10, p=0.5, gráfica izquierda). **Fig. 7.**Histograma Binomial (n=15, p=0.5, gráfica central). **Fig. 8.** Histograma Binomial (n=20, p=0.5, gráfica derecha).

4. Comentarios Finales

La utilización de un programa de cómputo matemático, con capacidad aritmética y simbólica además de graficación, permite una gran flexibilidad para la exploración de conceptos importantes de la probabilidad y también de la estadística. En nuestro trabajo, el énfasis ha sido en los aspectos probabilistas, los cuales, aunque elementales en la presente exposición son susceptibles de extender y de profundizar; algo que se requiere es la presentación de las fórmulas matemáticas para los modelos probabilistas utilizados.

5. Referencias

Blachman, N. (1992). *Mathematica: A Practical Approach.* Prentice Hall, Inc. London, UK.

Derman, C., Gleser, L. y Olkin I. (1973). *A Guide to Probability Theory and Application.* Holt, Rinehart and Winston, Inc. New York, USA.

Shawn, W. y Tigg, J. (1994). *Applied Mathematica - Getting Started, Getting It Done.* Addison Wesley. New York, USA.

El método de coordenadas principales y algunas de sus aplicaciones [1]

Ismael Sosa Galindo[2]
Facultad de Ciencias Administrativas y Sociales, U.V.
Luis Cruz-Kuri
Instituto de Ciencias Básicas, U.V.
Agustín Jaime García Banda
Facultad de Ciencias Administrativas y Sociales, U.V.

1. Introducción

En situaciones donde se cuenta con los datos cuantitativos y muchas variables, ya sea individualmente o por grupos se pueden aplicar algunas técnicas estadísticas multivariadas como por ejemplo: *Componentes Principales, Análisis de Correspondencia* o *Promediado Recíproco*, etc., pero en el caso en el cual no se cuenta con ellos directamente, se pueden proponer estos de acuerdo a algún tipo de criterio técnico (medida de *semejanza* o de *disimilitud*) y trabajar con los datos propuestos, también hay casos en los cuales se pierde la información o no es posible contar con ella y en estos casos es de mucha utilidad contar con una técnica que permita trabajar con datos alternativos y éste es el caso del método de Coordenadas Principales. En todas ellas intervienen la geometría del espacio s - dimensional para analizar los «enjambres» de puntos y los patrones que presentan. En términos geométricos, conviene proyectar a tales enjambres sobre espacios apropiados de dimensiones "visualizables", es decir, de dimensiones 3, 2 ó 1. Las técnicas anteriores remiten en determinado momento a la descomposición espectral de ciertas matrices, que pueden ser simétricas o no simétricas.

En una gran variedad de situaciones en las cuales se estudian poblaciones, surge la necesidad de clasificar o de ordenar a los individuos que la constituyen de acuerdo a ciertos criterios. Estos criterios pueden traducirse a condiciones cuantificables que remiten a problemas matemáticos cuyas soluciones están lejos de ser triviales. El propósito del Método de Coordenadas Principales es detectar patrones presentes entre los individuos de una población

[1]Trabajo realizado con apoyos de la Universidad Veracruzana
[2]isoga77@yahoo.com.mx

los cuales pueden o están ocultos en la multitud de circunstancias que rodean a éstos. En algunas situaciones puede ser obvio cuáles son los patrones que intervienen en la ordenación de los individuos de una población y en otras situaciones se hace patente la necesidad de desarrollar técnicas estadísticas y matemáticas que permitan la ordenación de una colección de objetos pertenecientes a una colección dada de interés y que dicha ordenación se traduzca en la detección de ciertos patrones que pueden ser reveladores de la estrucura oculta que existe en dicha población.

Se aborda y resuelve el problema de ordenar una colección finita de objetos mediante la utilización de criterios geométricos derivados de algunas características cuantificables y se describe la metodología que permite la ordenación, igualmente se realizan algunas aplicaciones en el contexto: Objetos de 13 botellas, además de los procesamientos correspondientes a la ordenación de dos colecciones de objetos en las que el tamaño y la forma constituyen las características esenciales para la formación de diversos patrones de semejanza. El procedimiento que constituye la parte central del presente trabajo es el llamado *Método de Coordenadas Principales* (este método describe a detalle una técnica estadística importante de ordenación de datos, dentro del marco de la estadística multivariada cuando los objetos que se quieren ordenar no tienen asignadas coordenadas naturales inicialmente y lo único con que se cuenta es con una medida de disimilitud entre ellos). A diferencia de la técnica de Componentes Principales, en la cual los objetos individuales son los que se miden de acuerdo a cierto número de variables, y se encajan de manera inmediata en algún espacio euclidiano, en el procedimiento que aquí nos ocupa, el proceso de medición corresponde a pares de objetos a través de una medida de semejanza, o si se quiere de disimilitud, la cual puede satisfacer o no la desigualdad del triángulo; en este supuesto, la solución al problema planteado puede hacerse en términos geométricos. Más explícitamente, el problema a resolver es el de la inmersión de la colección de objetos que se quiere ordenar en algún espacio euclidiano que permita tener una asignación de coordenadas para aquellas situaciones en que la medida de disimilitud sea métrica en el sentido técnico de satisfacer la desigualdad del triángulo, y lograr la inmersión de tal manera que la distancia entre dos objetos corresponda lo más aproximadamente posible a la medida de semejana o de disimilitud.

2. Aspectos Generales

Dada una colección finita de objetos, $\boldsymbol{U}$, el problema que se debe resolver es el de la ordenación de los miembros de $\boldsymbol{U}$ de acuerdo a ciertos criterios. Más específicamente, se busca una «estructura» o patrón sistemático que indique, *e.g.*, que ciertos grupos de objetos (o especies si se trata de un contexto ecológico) tendieron a ocurrir juntos. Para detectar tal estructura en una matriz de datos en bruto, se puede buscar reordenar columnas y renglones de ésta de una manera apropiada (lo cual no es una tarea trivial). Si se opta por una matriz de datos para la realización de la ordenación requerida, se tienen dos procedimientos indicados anteriormente, dentro de los que destaca la técnica de Componentes Principales. Por otra parte, se puede utilizar como criterio una medida de disimilitud como el punto de arranque. En cualquier caso, para la ordenación se requiere o bien de una matriz de datos o una matriz de disimilitudes. Nosotros describremos el procedimiento que permite una ordenación utilizando la segunda opción.

Para efectuar un análisis de Coordenadas Principales, debe contarse primero con una medida de disimilitud que satisfaga la condición de ser métrica, es decir, la desigualdad:

$\delta(x,y) \leq \delta(x,z) + \delta(z,y),\ \forall\ x,y,z$ en $\boldsymbol{U}$. Si el número de objetos es n y si $\delta(j,j')$ es la medida de disimilitud entre los objetos j y j', se requiere una asignación de coordenadas a cada uno de ellos de tal manera que la distancia euclidiana que lo separa $d\,(j,j')$, sea igual a $\delta(j,j')$, lo más cercano que sea posible. Si se logra este propósito, se puede decir en términos geométricos que se ha «sumergido» a la colección de objetos en cierto espacio métrico. Con frecuencia no es posible sumergir a todos los objetos de una colección dada en algún espacio euclidiano por más dimensiones que se permitan, y solo puede uno contentarse con hacerlo de manera aproximada.

3. Algoritmo Numérico

A continuación se presenta en forma detallada la descripción del procedimiento para la inmersión de una colección dada de objetos en algún espacio euclidiano, y por consiguiente, la asignación de coordenadas a cada uno de los objetos correspondiente a su ubicación en

tal espacio. Dichas coordenadas constituyen las así llamadas Coordenadas Principales con respecto a la medida de disimilitud δ.

i) Asígnese una disimilitud métrica entre los n objetos considerados. Denótese como antes con $\delta(j, j')$ $(j, j' = 1, ..., n)$; de esta manera se obtiene una matriz Δ de dimensiones $n \times n$ cuyas entradas son $\delta(j, j')$.

ii) Para cada $j, j' = 1, ..., n$, se calcula el número $\alpha[j, j']$ mediante la fórmula:

$$A = (I_n - H_n) \cdot \Delta_2 \cdot (I_n - H_n), \tag{1}$$

donde I_n es la matriz identidad de dimensiones $n \times n$, H_n está dada por

$$H_n = (1/n) \cdot \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}. \tag{2}$$

$$\Delta_2 = [(-1/2) \cdot \delta^2(j, j')];\ j, j' = 1, ..., n. \tag{3}$$

$$A = [\alpha_{j,j'};\ j, j' = 1, 2, 3, ..., n]\,.$$

iii) Efectúese la descomposición espectral para la matriz simétrica A; es decir, obténganse matrices O y D, donde O: $n \times n$ es ortogonal constituida por los eigenvectores normalizados de A y D: $n \times m$ es diagonal con sus entradas en la diagonal principal siendo los eigenvalores de A.

iv) Defínase ahora una matríz C: $n \times n$ mediante la relación

$$C = D^{1/2} \cdot O' \tag{4}$$

Las columnas de C constituyen las Coordenadas Principales de los n objetos. La primera coordenada está asociada al valor característico (*eigenvalor*) máximo de A; la segunda coordenada corresponde al segundo valor característico, y así sucesivamente. Pueden ocurrir eigenvalores negativos lo cual sería una indicación de que es posible encajar a los n objetos en un espacio euclidiano real, no importa de cuantas dimensiones fuera éste.

4. Ordenación de una Colección Específica de Objetos

Como una ilustración concreta se presenta el caso de una colección de trece botellas a cada una de las cuales se le miden cinco variables, tales como longitud total de la botella, circunferencia en la base, circunferencia en el cuello, etc. los datos son los que aparecen en la tabla 1 adelante. Aquí, la parte importante para ordenarlas o agruparlas es definir una medida adecuada de disimilitud. Resulta más natural definir una medida de disimilitud en términos de las diferencias en valor absoluto de las mediciones de las botellas (longitud total, circunferencia de la base, circunferencia del cuello, largo del cuello y circunferencia de la base del cuello), que hacerlo con las diferencias elevadas al cuadrado; en el primer caso, la disimilitud corresponde a una métrica L_1 o *distancia de Manhattan*, en tanto que en la segunda instancia remite a una métrica o distancia euclidiana. En este sentido, Componentes Principales producirá una solución basada en la métrica euclidiana, mientras que la técnica de coordenadas principales dará la posibilidad de hacer la ordenación de objetos para cualquier métrica elegida. Aquí usaremos la métrica L_1 o de Manhattan.

Tabla 1. Matriz de Datos. Cinco Variables Medidas a un Conjunto de Botellas para la Obtención de una Matriz de Disimilitudes (Todas las mediciones son en centímetros).

Botella	Longitud Total	Circunferencia Base	Circunferencia Cuello	Largo Cuello	Circunferencia Base Cuello
1	29.5	27.0	9.0	23.5	23.0
2	28.5	21.0	8.0	11.5	21.0
3	24.5	20.0	8.0	19.5	18.5
4	28.5	20.5	7.5	20.5	19.0
5	29.5	26.5	8.0	16.0	24.5
6	24.5	20.0	8.0	11.5	18.5
7	24.5	20.5	8.0	14.0	20.5
8	24.5	20.0	8.0	10.0	18.0
9	24.0	19.5	7.5	12.5	19.5
10	17.0	21.0	8.0	4.5	19.0
11	25.5	22.0	8.5	10.0	11.0
12	17.5	31.0	21.0	3.0	23.0
13	18.0	27.5	26.5	4.0	24.0

Con las 13 botellas se pudo constatar que la medida de disimilitud que se propuso para

la ilustración discutida, como una aplicación del Método de Coordenadas Principales, es bastante satisfactoria; es decir la métrica L_1, o métrica de Manhattan , corresponde a una medida de disimilitud la cual, al compararla con la métrica L_2 o *métrica de Pitágoras*, determina un espacio euclidiano donde quedan inmersos los objetos que se desean ordenar, obteniéndose un encaje en que las distancias métricas guardan una relación estrecha con las correspondientes disimilitudes. Se propuso la métrica L_1, como punto de partida ya que, si se hubiera utilizado la métrica L_2, el encaje sería perfecto puesto que es precisamente la distancia euclidiana la que se utiliza en un análisis de componentes principales; en ese caso, el Análisis de Componentes Principales produce los mismos resultados que el Análisis de Coordenadas Principales.

Tabla 2. Matriz de Disimilitudes con Métrica L_1 para un Conjunto de 13 Botellas. Las cantidades fueron generadas con el programa *Mathematica*. Ver instrucciones al pie de la tabla.

0.	22.	21.5	16.	10.5	29.5	24.5	31.5	29.	42.5	35.	48.5	50.
22.	0.	15.5	12.	14.5	7.5	7.5	9.5	9.	20.5	16.	44.5	46.
21.5	15.5	0.	6.5	21.	8.	8.	10.	9.5	24.	20.5	52.	53.5
16.	12.	6.5	0.	17.5	14.5	12.5	16.5	14.	28.5	24.	56.5	58.
10.5	14.5	21.	17.5	0.	22.	17.	24.	21.5	35.	28.5	44.	43.5
29.5	7.5	8.	14.5	22.	0.	5.	2.	3.5	16.	12.5	44.	45.5
24.5	7.5	8.	12.5	17.	5.	0.	7.	4.5	19.	16.5	44.	45.5
31.5	9.5	10.	16.5	24.	2.	7.	0.	5.5	15.	10.5	43.	44.5
29.	9.	9.5	14.	21.5	3.5	4.5	5.5	0.	17.5	16.	44.5	46.
42.5	20.5	24.	28.5	35.	16.	19.	15.	17.5	0.	23.5	29.	31.5
35.	16.	20.5	24.	28.5	12.5	16.5	10.5	16.	23.5	0.	48.5	50.
48.5	44.5	52.	56.5	44.	44.	44.	43.	44.5	29.	48.5	0.	11.5
50.	46.	53.5	58.	43.5	45.5	45.5	44.5	46.	31.5	50.	11.5	0.

Table [$\sum_{j=1}^{5}$Abs[datosbotellas[[m,j]] -datosbotellas[[n,j]]], {m,1,13},{n,1,13}]//MatrixForm

5. Comentarios Finales

Aunque en el presente trabajo se ha ilustrado la técnica para una colección de objetos específicos (botellas), el procedimiento es aplicable a cualquier colección finita de objetos en los cuales sea más accesible una asignación de medidas de disimilitud que su contraparte de ubicación geométrica en algún espacio euclidiano. Asimismo, la medida de disimilitud debe

satisfacer la condiciones usuales de distancia de un espacio métrico.

Tabla 3. Matriz de Coordenadas Principales en un Espacio de Dimensión 7 Asignadas a una Colección de 13 Objetos. Ver ecuación (4).

	Botellas												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	8.7	6.9	14.4	19.5	4.5	6.0	6.6	4.7	6.4	-10.4	7.4	-36.8	-38.2
2	23.0	0.3	-1.2	3.4	15.7	-6.7	-1.8	-8.5	-5.6	-14.7	-10.8	2.7	4.2
3	0.8	-2.8	5.5	4.9	-4.0	-0.1	0.9	-1.1	1.6	6.1	-11.9	1.3	-1.2
4	-4.0	1.9	2.3	0.3	4.4	0.6	1.4	-0.4	2.6	0.1	-3.2	-5.0	4.0
5	0.3	-5.9	4.7	-2.7	0.5	0.7	1.1	1.5	0.4	-1.9	0.8	-2.2	2.8
6	-0.9	-0.7	-0.4	-1.7	1.0	0.01	2.5	-0.7	3.1	-1.5	-0.8	2.4	-2.4
7	0.2	0.01	0.1	0.02	-0.1	-0.1	-0.2	0.1	1.1	-0.2	0.1	-0.03	0.1

El número de dimensiones relevantes para una visualización geométrica debe tomarse ≤ 3.

6. Referencias

Pielou, E.C. (1984). *The Interpretation of Ecological Data - A Primer on Classification and Ordination.* Editorial: John Wiley & Sons. New York.

Wolfram, S. (1999). *MATHEMATICA -A System for Doing Mathematics by Computer.* Fourth Edition. Editorial: Addison-Wesley Publishing Company. U.S.A.

Ordenación discriminante y algunas aplicaciones[1]

Luis Cruz-Kuri [2]
Instituto de Ciencias Básicas, U.V.
Agustín Jaime García Banda
Facultad de Ciencias Administrativas y Sociales, U.V.
Ismael Sosa Galindo
Facultad de Ciencias Administrativas y Sociales, U.V.

1. Introducción

Cuando se tiene una colección de objetos que se quieren ordenar o clasificar, se dispone, dentro de las técnicas multivariadas clásicas, aquellas que utilizan, por ejemplo, el análisis de componentes principales o el de coordenadas principales. Para la realización de tales ordenaciones o clasificaciones, se requiere un análisis espectral de matrices que da lugar a cambios de escala y rotaciones rígidas en un espacio euclidiano. Por otra parte, si se tienen varias poblaciones o colecciones de objetos que se requieren ordenar conjuntamente, un procedimiento útil es el de una ordenación de tipo discriminante, la cual resulta de un análisis espectral de la matriz que representa a todas las poblaciones; dicho análisis remite a rotaciones desarticuladas y por supuesto también a cambios de escala.
En el presente trabajo, además de hacer una descripción de la técnica, se hace mención de una serie de ilustraciones en varios contextos, dentro de los que se incluyen el biológico y el ecológico. El procesamiento computacional de los datos correspondientes, para el análisis espectral, se ejecutó mediante el programa Mathematica.

2. Matriz de Datos

Es una representación simbólica de una comunidad ecológica (o de otro tipo). Todos los análisis y representaciones subsecuentes se basan en una de tales matrices. En lo que sigue,

[1]Trabajo realizado con apoyos de la Universidad Veracruzana
[2]kruz1111@yahoo.com.mx

$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, etc. denotan matrices de datos. Por ejemplo, para una comunidad ecológica, se tiene

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{s1} & x_{s2} & \cdots & x_{sn} \end{pmatrix}, \tag{1}$$

donde x_{ij} = "abundancia" de la i-ésima especie en el j-ésimo cuadrat ($i = 1, \ldots, s; j = 1, \ldots n$).

A partir de $\mathbf{X}$, se puede obtener una matriz simétrica $X'X$, donde el apóstrofo denota transpuesto matricial. Análogamente para otras matrices de datos, $\mathbf{Y}, \mathbf{Z}$, etc. Cualquiera que sea el caso, se sabe que, si $\mathbf{M}$ es una matriz simétrica (tal como lo es $X'X$), su análisis espectral da la factorización

$$\mathbf{M} = \mathbf{U}'\mathbf{L}\mathbf{U} \tag{2}$$

donde $\mathbf{U}$ es una matriz ortogonal y $\mathbf{L}$ es una matriz diagonal. La ordenación, una a la vez, se obtiene con, e.g.,

$$\mathbf{X}_o = \mathbf{U}\mathbf{X} \tag{3}$$

$$\mathbf{Z}_o = \mathbf{U}\mathbf{Z}$$

En términos geométricos, el efecto de $\mathbf{U}$ corresponde a una rotación rígida del "enjambre" de puntos; esto es, siempre que $\mathbf{U}$ sea una matriz ortogonal.

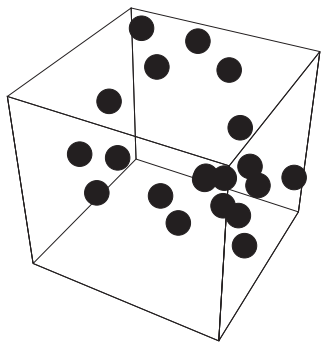


Fig. 1. *"Enjambre" de puntos en un espacio euclidiano.*

Si los datos fueron generados, por ejemplo, por una distribución normal multivariada, el enjambre presenta una forma elipsoidal, tal como se sugiere en la Fig. 1. Para otros modelos, el patrón del enjambre puede ser más complejo, e.g., helicoidal.

2.1. Objetivos

Se busca una "estructura" o patrón sistemático que indique, e.g., que ciertos tipos de especies tendieron a ocurrir juntos. Para detectar tal estructura en una matriz de datos en bruto, se puede intentar reordenar columnas y renglones de una manera apropiada (una tarea no trivial)

3. Ordenaciones Simultáneas

Las técnicas anteriores se refieren a ordenaciones una-a-la-vez. Cuando se tienen varias matrices de datos, correspondientes cada una de ellas a una población perteneciente a una colección de poblaciones, un análisis multivariado de varianza podría servir para establecer contrastes pero, si se quieren realizar ordenaciones simultáneas de los datos de todas las comunidades, se necesita entonces otra clase de técnica.

Conviene tener la capacidad de "visualizar" en una ordenación bidimensional cómo están relacionados los varios conjuntos de datos. En el diagrama que sigue, se presenta un esquema general.

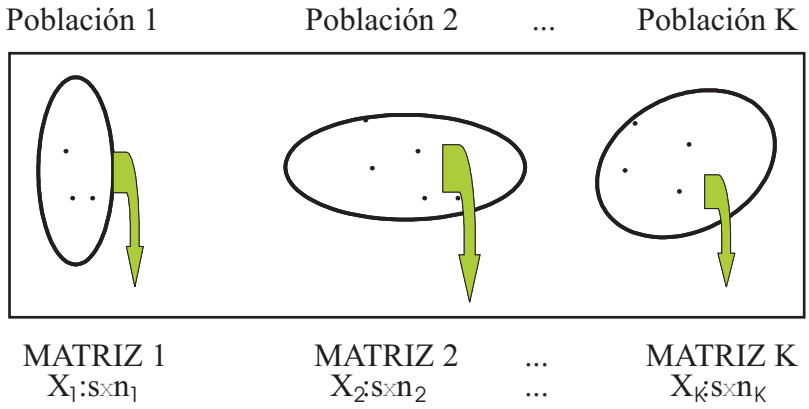


Fig. 2. *Esquema general para la obtención de varias matrices de datos.*

"ORDENACION DISCRIMINANTE"

Para el método de ordenación discriminante, se requiere un análisis espectral de cierta matriz no-simétrica. El procedimiento que se describe a continuación se discute, por ejemplo, en Pielou (1984) y también está relacionado con una técnica debida a Hotelling, que se conoce como el método de correlación canónica de variables (ver e.g. Morrison, 1990).

Si $\mathbf{B} : s \times s$ es no-simétrica, entonces su factorización está dada por

$$\mathbf{B} = \mathbf{WLW}^{-1} \tag{4}$$

donde $\mathbf{W}$ es una matriz de vectores característicos normalizados y $\mathbf{L}$ es una matriz diagonal de valores característicos de $\mathbf{B}$.

Ordenación Conjunta -El Método

Para L = 1, 2,. . . ,K, sea $\mathbf{X}_L$: s×n_L matriz de datos de "abundancias" para s "especies" en n_L cuadrats. $\mathbf{X}_L$ se aumenta para obtener $\mathbf{X}^*{}_L$ con vector columna "codificador" de dimensión K-1 tal y como se indica a continuación.

Codificación

$$\begin{aligned} \text{MATRIZ 1} \quad &\leftrightarrow [1, 0, \ldots, 0]' \\ \text{MATRIZ 2} \quad &\leftrightarrow [0, 1, \ldots, 0]' \\ &\ldots \\ \text{MATRIZ K-1} &\leftrightarrow [0, \ldots, 0, 1]' \\ \text{MATRIZ K} \quad &\leftrightarrow [0, \ldots, 0, 0]' \end{aligned} \tag{5}$$

Sea $n = n_1 + n_2 + \cdots + n_K$ y defínase la matriz mancomunada de datos, $\mathbf{X} : (s + K - 1) \times n$, mediante

$$X = [X *_1 | X *_2 | \ldots | X *_K]. \tag{6}$$

Procedimiento

El procedimiento se describe en 6 pasos, como sigue.

1. Dada $\mathbf{X}$, transformar $\mathbf{X} \Diamond \mathbf{Z}$ donde $\mathbf{Z}$ es la matriz de datos estandarizados.

2. Calcular $\mathbf{ZZ}'$, la cual casi corresponde a la matriz de correlaciones.

3. Particionar $\mathbf{ZZ}'$ en la forma

$$\mathbf{ZZ}' = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix}, \tag{7}$$

donde, $\mathbf{R}_{11} : (K - 1) \times (K - 1), R_{22} : s \times s$, etc.

4. Calcúlese la matriz $\mathbf{D}$: s×s mediante

$$\mathbf{D} = \mathbf{R}_{22}^{-1} \mathbf{R}_{21} \mathbf{R}_{11}^{-1} \mathbf{R}_{12}, \tag{8}$$

la cual establece correlaciones parciales.

5. $\mathbf{D}$ es no-simétrica. Analizarla espectralmente. El número de valores característicos no-nulos es igual a mín$\{s, k - 1\}$.

6. Las puntuaciones asignadas a cada cuadrat, digamos l con vector $\mathbf{x} : s \times 1$ de abundancias, corresponden a una combinación lineal de las coordenadas de $\mathbf{x}$, con ponderaciones dadas por los vectores característicos.

(Si la ordenación es bidimensional, solo se toman los correspondientes a los dos valores característicos más grandes).

Aspectos computacionales

La factorización de la matriz no-simétrica D que se describe en el paso 4 del procedimiento es más difícil que un caso simétrico, pero existen programas de cómputo fácilmente accesibles. Debe cubrirse la posibilidad de valores-no reales (i.e. números complejos). En nuestros estudios hemos analizado matrices 20×20. Fácilmente es posible de extender esto a dimensiones mayores. Algunos programas de cómputo matemático, tales como Mathematica, Maple, etc. permiten resolver los problemas de análisis espectral.

4. Ejemplos de aplicabilidad de ordenaciones conjuntas

Supóngase que se tienen varias matrices de datos, las cuales se quieren ordenar en un marco coordenado común. A continuación se mencionan algunas de las muchas situaciones en las que una técnica de ordenación conjunta es aplicable.

1. Vegetación (o flora) de varios lagos; para cada lago, se tiene una matriz de datos.

2. Muestreos de fauna de insectos en campos de trigo cada mes de Julio; los datos de abundancias de especies, para cada año, constituyen los elementos de tales matrices.

3. Comparación de las condiciones ambientales en varias regiones geográficas separadas; dentro de cada región, se miden cierto número de variables ambientales para las estaciones de muestreo (e.g., datos climáticos, para los que se formado un banco de datos, ver detalles adelante)

Una ilustración de aplicabilidad –caso del Pico de Orizaba

A partir de marzo del año de 1999, dentro de un proyecto en el que uno de nosotros es participante (Cruz-Kuri) se dio inicio a la instalación de aparatos de medición de parámetros climáticos en distintas ubicaciones de la cara sur del Pico de Orizaba, por debajo, por encima y dentro de lo que técnicamente se conoce como "línea de árboles" de la montaña. Es pertinente mencionar que esta montaña se distingue por tener la línea a mayor elevación de todas las líneas de árboles de nuestro planeta; motivo por el cual es de gran interés el estudio de los factores climáticos y de otra naturaleza que determinan la ubicación de tales líneas de árboles, especialmente en ambientes tropicales alpinos.

El registro de las variables meteorológicas consideradas continúa en proceso a la fecha y se tiene una extensa base de datos en alrededor de 20 ubicaciones, cada una de las cuales se denomina "estación". Esto ha dado lugar a un "Banco de datos climáticos" el cual abarca los años 1999-2006 y 20 estaciones en las cuales se miden cada hora variables tales como las que se indican en la lista a continuación.

Tabla 1. *Algunas Variables Medidas en el Pico de Orizaba*

Temperatura Aire Precipitación Total
Temperatura Suelo a 0 cm. de Profundidad
Temperatura Suelo a 10 cm. de Profundidad
Temperatura Suelo a 20 cm. de Profundidad
Temperatura Suelo a 40 cm. de Profundidad
Humedad Relativa
Humedad Absoluta
Punto de Rocío
Temperatura Interna de Arbol

Para tener una idea del tamaño de los archivos, los datos de este banco ocupan un espacio equivalente a casi 300 hojas de papel de impresora de computadora. Las técnicas de ordenaciones conjuntas de las matrices resultantes de datos son susceptibles de ser ejecutadas tanto para la búsqueda de patrones espaciales como temporales. Para las ordenaciones espaciales, se tomaron en consideración las varias ubicaciones geográficas de las estaciones; para las ordenaciones temporales, se utilizaron los distintos años. En cualquiera de estas situaciones, los datos fueron generados por la medición de los parámetros climáticos arriba menciona-

dos. Por consideraciones de espacio, los resultados de tales análisis no se presentan en este trabajo.

5. Comentarios Finales

Se ha visto que, si se tienen varias poblaciones o colecciones de objetos que se requieren ordenar conjuntamente, un procedimiento útil es el de una ordenación de tipo discriminante, el cual resulta de un análisis espectral de la matriz que representa a todas las poblaciones; dicho análisis remite a rotaciones desarticuladas y por supuesto también a cambios de escala.

En el presente trabajo, además de hacer una descripción de la técnica, se han mencionado una serie de posibilidades de aplicación en varios contextos, dentro de los que se han incluido el biológico y el ecológico. Algunos tipos de procesamientos de los datos (cuyos resultados no se presentan aquí por consideraciones de espacio) se ejecutaron mediante el programa Mathematica, en tanto que otros, particularmente para las series de tiempo multivariadas, requirieron programas de cómputo estadístico, tales como SPSS y Statistica.

6. Referencias

Blachman, N. (1992). *Mathematica: A Practical Approach.* Prentice Hall, Inc. CityplaceLondon, country-regionUK.

Morrison, D.F. (1990). *Multivariate Statistical Methods.* Third Edition. McGraw-Hill. Stateplace, New York. country-regionplace USA.

Pielou, E.C. (1984). *The Interpretation of Ecological Data* – A Primer on Classification and Ordination. Editorial: John Wiley & Sons. Stateplace, New York.

Shawn, W. y Tigg, J. (1994). *Applied Mathematica – Getting Started, Getting It Done.* Addison Wesley. Cityplace, New York, country-region USA.

Una propuesta de mejora en un proceso de servicio de salud bajo un contexto seis sigma

Samantha Lucill Silva Chavelas[1]
Universidad de las Américas, Puebla
Jorge Domínguez Domínguez[2]
Centro de Investigación en Matemáticas, Guanajuato
Antonio González Fragoso[3]
Universidad de las Américas, Puebla
Gladys Linares Fleites[4]
Universidad de las Américas, Puebla

1. Introducción

El objetivo de este trabajo es la aplicación de las tres primeras etapas de cinco que corresponden a DMAMC: Definir, Medir, Analizar, Mejorar y Controlar, bajo la metodología Seis Sigma a un proceso de servicio del sector salud: atención a los pacientes que solicitan y se les otorga una consulta médica en la clínica de la Universidad de las Américas, Puebla (UDLAP), con la finalidad de presentar una propuesta de mejora en este proceso.

2. Etapa de Definición

Un primer paso fue identificar al cliente (paciente) del proceso, posteriormente usando los datos históricos de la clínica, se realizó un diagrama de Pareto de las principales inconformidades expresadas por los pacientes hasta este momento. De esta manera se identificaron las oportunidades de mejora, por lo que el problema se determina como:

[1]samantha_lucielle@hotmail.com
[2]jorge@cimat.mx
[3]antonio.gonzalez@udlap.mx
[4]gladys.linares@udlap.mx

La Falta de un instrumento para medir la calidad en la atención que se brinda a los pacientes que asisten a las ***consultas médicas*** *de los servicios médicos de la Universidad de las Américas Puebla. Y poder así tener un monitoreo continuo para verificar una mejora en el servicio.*

Se trabajó conjuntamente con el personal de la clínica, médicos, enfermeras y paramédicos, lográndose el mapa del proceso. Con esto último, junto con el diagrama de Pareto y una insistente retroalimentación, se lograron establecer las variables críticas.

3. Etapa de Medición

En esta etapa se construye el instrumento de medición, considerando las variables definidas en la etapa anterior, siendo este instrumento un cuestionario formulado en base al proyecto del Cualitómetro Akao (1990), el cual se aplica en dos etapas: antes y después de recibir el servicio. En la primera etapa, se le pregunta al encuestado por una parte, la importancia que tiene el servicio para él y, por otra se le cuestiona sobre la calidad que espera del servicio, esto para cada una de las variables críticas. En la segunda aplicación del cuestionario, el paciente califica la Calidad Percibida.

4. Etapa de Análisis

Primero se analizan los datos utilizando las gráficas de barra para después continuar con un análisis cuantitativo propuesto por el Proyecto del Cualitómetro en donde: **Calidad del Servicio = Calidad Percibida − Calidad Esperada**

Esta medida de la calidad del servicio se basa en el grado de satisfacción o insatisfacción del paciente (cliente). Señalando como puntos de mejora la *capacidad de diagnóstico (variable de mayor peso negativo), limpieza, comunicación de la enfermera, rapidez de la atención y confiabilidad*, variables en donde la Calidad Percibida es menor a la Esperada.

Con el fin de encontrar todas las posibles relaciones existentes entre las variables que determinen la satisfacción del paciente se aplica un análisis de correspondencia simple, Greenacre et al (1987).

En el análisis de correspondencias simple se definieron dos ejes factoriales con el fin de explicar el 33 % del problema, definiendo el primer eje o dimensión como la Calidad Esperada y la segunda dimensión como la Calidad Percibida, basándonos en la fuerza de correlación existente con las variables de medición. Utilizando los gráficos obtenidos del análisis de correspondencia se analiza la distancia entre las variables y respecto a los ejes de los gráficos identificando así el grado de satisfacción de la muestra, Benzécri (1992).

Posteriormente se aplican dos de las cartas del Despliegue de la Función de Calidad llamado QFD por sus siglas en ingles, ReVelle et al (1998).

Primero se construye la carta de *Despliegue de la Calidad Demandada*. Matriz que determina el nivel de calidad demandada relacionando los factores a evaluar del servicio con las variables definidas dentro del plan de calidad.

De esta tabla se obtiene QUÉ es lo que hay que mejorar, el siguiente paso es encontrar CÓMO lograr dicha mejora y para ello se implementa la *Carta de Despliegue de los Elementos de la Calidad*. Se muestra un fragmento de esta carta en la siguiente página.

Cabe señalar que en la construcción de la Carta de la Calidad Demandada se utilizó el análisis de correspondencia para determinar el argumento de atención dando dicho valor de ponderación en base a la correlación existente entre el factor del servicio (variable de medición) y el eje representante de la Calidad Esperada. Siendo una aportación de este trabajo de investigación, ya que anteriormente el argumento de atención había sido otorgado de acuerdo al criterio de los directivos.

Con el fin de encontrar las posibles causas del problema bajo estudio se utiliza el diagrama de Causa-Efecto (Ishikawa), indicando 6 factores causales generales; en el caso de la capacidad de diagnóstico estos factores son: Equipo, paciente, médico, método, análisis clínicos e infraestructura. De estas causas generales se amplía el diagrama con todas las causas posibles de dispersión.

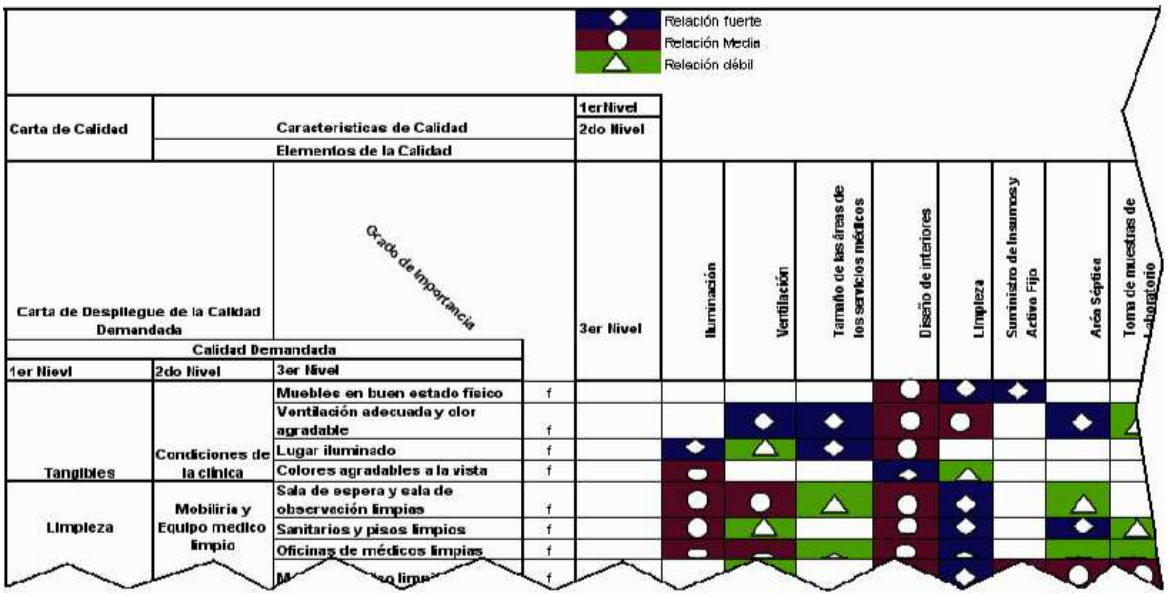


Figura 1: Fragmento de la Carta de Despliegue de los Elementos de Calidad.

5. Recomendaciones

Una vez terminadas las tres primeras fases de la metodología DMAIC de Seis Sigma, se formula una propuesta como mecanismo de mejora entre algunas recomendaciones presentamos las siguientes.

- A través de la carta de despliegue de la calidad demandada determinamos que la *capacidad de diagnóstico* es la variable más importante del paciente, por lo que en base al diagrama de Ishikawa recomendamos que el médico tenga un diálogo mayor con el paciente para que no exista confusión en los síntomas.

- También encontramos por medio del análisis de correspondencia que el paciente considera que la capacidad de diagnostico es representativa de la calidad global del servicio por lo que recomendamos se implemente un seguimiento sobre el paciente y las posibles causas de fallo de no existir recuperación en el tiempo estipulado.

- Otra variable a la cual hay que prestar atención es el tiempo de espera; es recomendable tomar el tiempo de espera de cada paciente y analizar aquellos casos en los que el

tiempo de espera fue mayor al promedio y en base al Diagrama de Ishikawa relacionar los puntos de falla.

- En general se recomienda, basándose en la última Carta del QFD, darle mayor fuerza e importancia a las características de calidad en donde se encontró una relación con las variables que requieren una mejora. Tomar en cuenta las relaciones que se obtuvieron entre las variables con el análisis de correspondencia al momento de brindar el servicio, e identificar la interacción de estas relaciones en el diagrama de procesos.

6. Referencias

Akao Yoji (1990) *Quality Function Deployment: Integrating Customer Requirements into Product Design*, Cambrige, Mass.

Benzécri J.-P. (1992) *Correspondence Analysis Handbook*, pp. 18, Marcel Deckker Inc, New York.

Greenacre, M. and Hastie, T. (1987) *The Geometric Interpretation of Correspondence Analysis* pp. 437-447 Journal of the American Statistical Association, New York.

ReVelle, Jack B., John W. Moran, Charles A. Cox, *The QFD. Handbook*, New York: Wiley and Sons, 1998.

Diseños experimentales óptimos en modelos de compartimientos [1]

Víctor Ignacio López Ríos [2]

Estudiante de Doctorado en Probabilidad y Estadística- Centro de Investigación en Matemáticas- CIMAT. Profesor Escuela de Estadística, Universidad Nacional de Colombia

Rogelio Ramos Quiroga [3]

Investigador Titular A, Centro de Investigación en Matemáticas (CIMAT)

1. Introducción

El propósito de este trabajo es la construcción de diseños óptimos para la estimación de los parámetros en un modelo de cuatro compartimientos con tasas de transferencia reversibles; además de diferentes funciones no lineales de éstos. Los modelos de compartimientos son usados para el análisis de un sistema dividido en un número finito de componentes, llamados compartimientos. Estos modelos son de gran utilidad en farmacocinética (rama de la farmacología que se ocupa de la liberación, absorción, distribución, metabolismo y excreción de los medicamentos desde el organismo), y su interés está en el estudio de las curvas concentración-tiempo de un medicamento, a partir de estos modelos.

Estamos interesados en hallar tiempos de muestreo "óptimos" con el fin de estimar, entre otras cantidades, el área bajo la curva concentración-tiempo (AUC) (importante en la deducción de parámetros farmacocinéticos como el aclaramiento -volumen de sangre o plasma depurado de medicamento en unidad de tiempo-, el volumen de distribución - forma en la que un medicamento podría llegar a penetrar en tejidos y líquidos orgánicos- Clark y Smith 1989), el tiempo para la concentración máxima ($t_{\text{máx}}$) y la concentración máxima ($C_{\text{máx}}$).

El modelo de estudio en este trabajo, se representa mediante el siguiente esquema:

$$A \overset{\theta_5}{\to} D \overset{\theta_1}{\to} B \underset{\theta_3}{\overset{\theta_2}{\rightleftharpoons}} C \overset{\theta_4}{\to} . \tag{1}$$

[1] Trabajo realizado con apoyos de Centro de Investigación en Matemáticas (CIMAT), México, Secretaría de Relaciones Exteriores de México, y Universidad Nacional de Colombia, Sede Medellín.

[2] vilopez@cimat.mx

[3] rramosq@cimat.mx

siendo los θ_i las tasas de transferencia del medicamento de un compartimiento a otro, se supone que la tasa del compartimiento receptor es proporcional a la concentración en el compartimiento fuente. En B y C se tienen tasas reversibles. Varios modelos similares, con tasas reversibles, fueron considerados en Allen, D. M (1983), Waterhouse, T. H. (2005), y sin tasa reversible está el trabajo de Atkinson et. al (1993).
El esquema 1 se expresa por medio del siguiente sistema de ecuaciones diferenciales:

$$\frac{d\eta_A(t,\Theta)}{dt} = -\theta_5\eta_A(t,\Theta) \qquad \frac{d\eta_B(t,\Theta)}{dt} = \theta_1\eta_D(t,\Theta) - \theta_2\eta_B(t,\Theta) + \theta_3\eta_C(t,\Theta)$$

$$\frac{d\eta_D(t,\Theta)}{dt} = \theta_5\eta_A(t,\Theta) - \theta_1\eta_D(t,\Theta) \quad \frac{d\eta_C(t,\Theta)}{dt} = \theta_2\eta_B(t,\Theta) - (\theta_3+\theta_4)\eta_C(t,\Theta),$$

siendo $\Theta^T = (\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$ y η_R: concentración del medicamento en el compartimiento R. Como el objetivo es estimar las cantidades de interés en el compartimiento C, se presenta sólo la solución para este compartimiento:

$$\eta_C(t,\Theta) = g_1(\Theta)\left[g_2(\Theta)e^{-\theta_1 t} + g_3(\Theta)e^{-\theta_5 t} + g_4(\Theta)e^{k_2^{(1)}t} + g_5(\Theta)e^{k_2^{(2)}t}\right] \tag{2}$$

$\gamma = \frac{\theta_1 C_0 \theta_5}{\theta_5-\theta_1}$ y $\theta_2+\theta_3+\theta_4)^2 - 4\theta_2\theta_4 > 0$ y $k_2^{(1)}$, $k_2^{(2)}$ soluciones a: $k_2^2 + k_2\left(\theta_2+\theta_3+\theta_4\right) + \theta_2\theta_4 = 0$.
Y las funciones $g_i(\Theta)$ están dadas por:

$$g_1(\Theta) = \frac{\theta_1 C_0 \theta_5 \left(k_2^{(2)}+\theta_2\right)\left(k_2^{(1)}+\theta_2\right)}{\theta_3(\theta_5-\theta_1)\left(k_2^{(1)}-k_2^{(2)}\right)}, \quad g_2(\Theta) = \frac{k_2^{(2)}-k_2^{(1)}}{\left(k_2^{(1)}+\theta_1\right)\left(k_2^{(2)}+\theta_1\right)}, \quad g_3(\Theta) = \frac{k_2^{(1)}-k_2^{(2)}}{\left(k_2^{(1)}+\theta_5\right)\left(k_2^{(2)}+\theta_5\right)},$$

$$g_4(\Theta) = \frac{\theta_1-\theta_5}{\left(k_2^{(1)}+\theta_1\right)\left(k_2^{(1)}+\theta_5\right)} \quad \text{y} \quad g_5(\Theta) = \frac{\theta_5-\theta_1}{\left(k_2^{(2)}+\theta_1\right)\left(k_2^{(2)}+\theta_5\right)}.$$

Se denotarán por $H_i(\Theta)$, $i = 1,2,3,4,5$, las funciones asociadas al área bajo la curva de $\eta_C(t,\Theta)$, (AUC), tiempo para la concentración máxima, $t_{\text{máx}}$, concentración máxima $(\eta(t_{\text{máx}},\Theta))$, y además, la diferencia entre las tasas reversibles, $(\theta_2-\theta_3)$ y la suma de las todas las tasas, $(\sum\theta_j)$.

2. Preliminares y Definiciones

Considere un experimento con N observaciones independientes Y_i, $i = 1,2,\ldots,N$, en las condiciones experimentales t_i, modelado por

$$Y_i = \eta(t_i,\Theta) + \epsilon_i, \;\; i = 1,2,\ldots,N, \;\; t_i > 0, \Theta \in \mathbb{R}^p, \tag{3}$$

donde los errores ϵ_i se suponen independientes e idénticamente distribuídos con media cero y varianza común, σ^2. Un diseño aproximado

$$\xi = \begin{pmatrix} t_1 & \cdots & t_n \\ w_1 & \cdots & w_n \end{pmatrix}$$

es una medida de probabilidad con soporte finito, es decir, las observaciones son tomadas en los puntos de soporte t_i proporcional al peso w_i. Para el modelo descrito en 3, y bajo algunas suposiciones de regularidad, el estimador de mínimos cuadrados es asintóticamente insesgado con matriz de covarianza asintótica proporcional a la inversa de:

$$M(\xi, \Theta) = \sum_{i=1}^{n} w_i f(t_i, \Theta) f^T(t_i, \Theta) = \int f(t, \Theta) f^T(t, \Theta) d\xi(t),$$

donde $f(t, \Theta) = \frac{\eta(t,\Theta)}{\partial \Theta}$. En la literatura de diseños, la matriz $M(\xi, \Theta)$ se conoce como la matriz de información del diseño ξ y un criterio de optimalidad, ϕ, maximiza algún funcional de valor real (con algún significado estadístico) de la matriz de informacion sobre la clase de todos los diseños aproximados (ver Pukelsheim 1993). Ejemplos de tales criterios son: $\phi(\xi) = \ln |M(\xi, \Theta)|^{1/p}$ (D-optimalidad), y $\phi(\xi) = \left\{ \mathrm{Tr}\left[AM(\xi, \Theta)^{-1} \right] \right\}^{-1}$, con A una matriz definida positiva (L-optimalidad). El último criterio es apropiado para la estimación de combinaciones lineales de los parámetros, $K^T\Theta$, ya que equivale a minimizar el promedio de las varianzas de las combinaciones lineales estimadas, con $A = K^T K$. Un diseño ξ que maximiza ϕ se dice que es ϕ-óptimo para estimar Θ. En este trabajo, para la estimación conjunta de las cantidades de interés, $H_i(\Theta)$, funciones no lineales de Θ ($i = 1, 2, 3$), y lineales ($i = 4, 5$); se va a considerar una versión linealizada, tomando $A = K^T(\Theta) K(\Theta)$, donde la columna i de K es el gradiente de $H_i(\Theta)$. La matriz de información y también, en general, K dependen de Θ, una solución es considerar un valor apriori para Θ, Θ_0, y los diseños así obtenidos se denominan locales (Chernoff 1953), además si se tiene una distribución apriori π para Θ, se define el criterio ϕ optimal promediado por la apriori π, abreviado por ϕ_π, como: $\phi_\pi(\xi) = E[\phi(\xi, \Theta)] = \int \phi(\xi, \Theta) d\pi(\Theta)$. Si no hay lugar a confusión, en lo que sigue, se omite la dependencia de Θ en f, K y M. Para la construcción y verificación de que un diseño dado, ξ, es ϕ-óptimo, en la literatura se tienen varios teoremas de equivalencia (Fedorov y Hackl 1997):

ξ es $\mathbb{D}-$ óptimo local sií $f^T(t)^T M^{-1}(\xi) f(t) - p \leq 0, \forall\ t \in [0, \infty)$,

ξ es $\mathbb{L} -$ óptimo local sií $f^T(t) M^{-1}(\xi) K K^T M^{-1}(\xi) f(t) - \mathrm{Tr}(K^T M^{-1}(\xi) K) \leq 0, \forall t \in [0, \infty)$,

ξ es $\mathbb{D}_\pi$-óptimo sií $\int \left[f^T(t, \Theta) M^{-1}(\xi, \Theta) f(t, \Theta) \right] d\pi(\Theta) - p \leq 0\ \ \forall t \in [0, \infty)$.

ξ es $\mathbb{L}_\pi$-óptimo sií $E_\Theta \left[f^T(t, \Theta) M^{-1} K(\Theta) K^T(\Theta) M^{-1} f(t, \Theta) - Tr(K^T(\Theta) M^{-1} K(\Theta)) \right] \leq 0$,

$\forall t \in [0, \infty)$, en todos los casos anteriores, con igualdad en los puntos de soporte del diseño óptimo.

Θ_0	D-óptimo local	L-óptimo local
Θ_0^1	$\left\{\begin{array}{ccccc} 0{,}73 & 1{,}60 & 3{,}05 & 5{,}45 & 10{,}10 \\ 0{,}20 & 0{,}20 & 0{,}20 & 0{,}20 & 0{,}20 \end{array}\right\}$	$\left\{\begin{array}{ccccc} 0{,}544 & 1{,}471 & 2{,}988 & 5{,}719 & 12{,}046 \\ 0{,}233 & 0{,}152 & 0{,}171 & 0{,}199 & 0{,}245 \end{array}\right\}$
Θ_0^2	$\left\{\begin{array}{ccccc} 0{,}80 & 1{,}92 & 3{,}85 & 7{,}15 & 13{,}30 \\ 0{,}20 & 0{,}20 & 0{,}20 & 0{,}20 & 0{,}20 \end{array}\right\}$	$\left\{\begin{array}{ccccc} 0{,}639 & 1{,}797 & 3{,}775 & 7{,}549 & 15{,}797 \\ 0{,}215 & 0{,}124 & 0{,}141 & 0{,}185 & 0{,}335 \end{array}\right\}$
Apriori	D_π-óptimo	L_π-óptimo
π	$\left\{\begin{array}{ccccc} 0{,}730 & 1{,}600 & 3{,}050 & 5{,}450 & 10{,}196 \\ 0{,}20 & 0{,}20 & 0{,}20 & 0{,}20 & 0{,}20 \end{array}\right\}$	$\left\{\begin{array}{ccccc} 0{,}531 & 1{,}493 & 3{,}077 & 5{,}754 & 11{,}579 \\ 0{,}473 & 0{,}194 & 0{,}121 & 0{,}096 & 0{,}116 \end{array}\right\}$

Cuadro 1: Diseños óptimos locales y promediados por una apriori π con valores locales para Θ de $\Theta_0^{1T} = (1{,}1\ 2{,}82\ 2{,}17\ 0{,}54\ 1{,}57)$, $\Theta_0^{2T} = (0{,}5\ 2{,}82\ 2{,}17\ 0{,}54\ 1{,}57)$

3. Resultados

En esta sección se presentan los diseños óptimos, locales y promediados por una apriori, obtenidos por los diferentes criterios, para la estimación de las funciones $H_i(\Theta)$ simultáneamente y también para la estimación de todas las tasas de transferencia. Se hizo un programa en R para la construcción de los diseños óptimos utilizando los algoritmos propuestos por Fedorov y Hackl (1997). Como ilustración se consideraron dos valores apriori para Θ y una distribución apriori uniforme discreta π, para Θ_0^1 y las 32 combinaciones de los valores de éste, obtenidas al perturbar cada componente en un 10 %, es decir, cada θ_i tomó dos valores $\theta_{i0}^1 \pm 0{,}10\theta_{i0}^1$, donde el soporte de π es: $\Omega = \left\{\Theta :\ \Theta = \Theta_0^1 \pm 0{,}10\Theta_0^1,\ \text{o}\ \Theta = \Theta_0^1\right\}$. En el cuadro 1, se muestran los diferentes diseños obtenidos, donde los pesos de los diseños D-óptimos locales y D_π-óptimos son iguales a 1/5, lo cual está acorde con la teoría de diseños óptimos. Cada uno de los diseños proporcionan los tiempos de muestreo óptimos y la frecuencia de la toma de muestras en cada uno de los tiempos. Se observan diferencias de los diferentes diseños obtenidos por cada uno de los criterios. Note que los diseños L-óptimos locales dan un peso mayor a los puntos extremos del diseño hallado, mientras que el diseño L_π-óptimo da un peso del 47 % al primer punto. De la Figura 1 se verifica que todos los diseños son óptimos, uso directo de los teoremas de equivalencia de la sección 2.

4. Conclusiones

Se construyeron diseños óptimos para la estimación de funciones no lineales de Θ, tanto locales como promediados por una apriori, en modelos de compartimientos útiles en farmacocinética, usando L-optimalidad. Como trabajo futuro se construirán diseños óptimos que permitan estimar las cantidades de interés y también discriminen entre dos modelos de compartimientos.

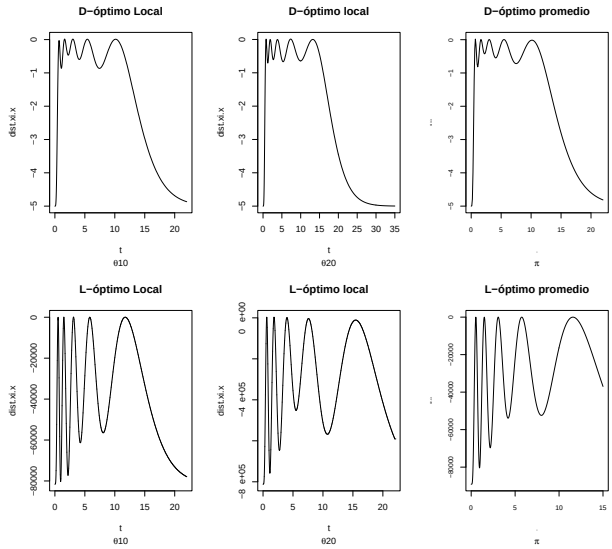


Figura 1: Gráficos para verificar ϕ-optimalidad

5. Referencias

Allen, D. M. (1983). Parameter Estimation for Nonlinear Models with Emphasis on Compartmental Models. *Biometrics*, 39, 629–637.

Atkinson, A. C., Chaloner, K., Herzberg, A. M. y Juritz, J. (1993). Optimum Experimental Designs for Properties of a Compartmental Model. *Biometrics*, 49 (2), 325–337.

Chernoff, H. (1953). Locally Optimal Designs for Estimating Parameters. *The Annals of Mathematical Statistics*, 24 (24), 586–602.

Clark, B. y Smith, D. A. (1989). *An Introduction to Pharmacokinetics*, Oxford.

Fedorov, V. V. y Hackl (1997). *Model-Oriented Design of Experiments*, New York.

Pukelsheim, F. (1993). *Optimal Design of Experiments*, New York: Wiley.

R Development Core Team (2005). R Foundation for Statistical Computing, http://www.R-project.org

Waterhouse, T. H (2005). Optimal Experimental Design for Nonlinear and Generalised Linear Models. *Ph. D. Thesis*. University of Queensland.

Pronósticos en modelos autorregresivos con umbral[1]

María Guadalupe Russell Noriega[2]
Facultad de Matemáticas, Universidad de Guanajuato
Graciela González Farías[3]
Centro de Investigación en Matemáticas, A.C.
Jesús Gonzalo[4]
Universidad Carlos III, Mardrid, España

1. Introducción

Los modelos de series de tiempo no lineal son una herramienta útil para describir y pronosticar una gran variedad de fenómenos naturales. Uno de los modelos más utilizados y estudiados en la literatura es el modelo autorregresivo con umbral, identificado como modelo TAR. La creciente aplicabilidad de dichos modelos radica en que describen adecuadamente una gran variedad de fenómenos naturales, y a que son relativamente fáciles de especificar y estimar. Sin embargo, en relación a la obtención y cuantificación de los desempeños de los pronósticos se ha encontrado que los modelos TAR en general no presentan mejores desempeños que los pronósticos generados a partir de un modelo autorregresivo lineal (AR), cuantificados sus desempeños mediante sus respectivos errores cuadráticos medios de pronóstico (ECMP). El objetivo de este trabajo es explorar nuevas alternativas de obtención de pronósticos en modelos TAR, lo cual nos lleva a proponer métodos de estimación y obtención de pronósticos que tomen en cuenta la estructura no lineal de los modelos TAR y evaluar los desempeños de los pronósticos bajo cada alternativa considerada. Ante la problemática ¿Por qué, dado un proceso generador de datos de un modelo no lineal, los pronósticos a través de un modelo lineal resultan más satisfactorios en términos de sus ECM? proponemos dos alternativas de solución conocidas como Verosimilitud Predictiva (VP) y Regresión por Cuantiles (RC), mostramos las propiedades asintóticas de los estimadores bajo cada alternativa y discutimos los desempeños de las mismas bajo un amplio ejercicio de simulación.

[1]Trabajo apoyado por CONACyT, beca 86663 y proyecto 39017E
[2]mgrussell@quijote.ugto.mx
[3]farias@cimat.mx
[4]jesus.gonzalo@uc3m.es

2. El modelo autorregresivo con umbral (modelo TAR)

El proceso Y_t sigue un modelo TAR de dos regímenes, con $p = 1$, condición inicial Y_0, y variable umbral Z_{t-d}, si Y_t puede expresarse de la forma siguiente:

$$Y_t = [\phi_1 I(Z_{t-d} \leq \gamma) + \phi_2 I(Z_{t-d} > \gamma)]Y_{t-1} + \varepsilon_t = \delta_t Y_{t-1} + \varepsilon_t, \tag{1}$$

con $\delta_t = \phi_1 I(Z_{t-d} \leq \gamma) + \phi_2 I(Z_{t-d} > \gamma)$, Z_{t-d} y $\varepsilon_t \sim iidD(0, \sigma^2)$ procesos estacionarios, ergódicos y mutuamente independientes, γ el parámetro de umbral y d el valor del rezago de la variable umbral. El modelo bajo (1) se denota como TAR$(2; 1, 1)$. González y Gonzalo (1998) muestran las propiedades de estacionariedad y ergodicidad del modelo TAR, se tiene por ejemplo que Y_t es estacionario en covarianzas si $E(\delta_t^2) < 1$. Cuando $Z_{t-d} = Y_{t-d}$ o una función de rezagos de Y_t el modelo en (1) se conoce como SETAR. Las propiedades de ergodicidad en los modelos SETAR se deben principalmente a Chan *et al* (1985), entre otros.

3. Estimación en modelos TAR

Mínimos Cuadrados (MC): Sea $\boldsymbol{\psi} = (\phi_1, \phi_2, \sigma)'$ el vector de parámetros en el modelo TAR de la ecuación (1), los estimadores de MC para $\boldsymbol{\psi}$**,** considerando γ fijo, son aquellos valores que minimizan $\sum_{t=2}^{T}[Y_t - E(Y_t|\mathfrak{F}_{t-1}; \boldsymbol{\psi})]^2$, donde T es el tamaño de la muestra observada y $E(Y_t|\mathfrak{F}_{t-1}; \boldsymbol{\psi}) = [\phi_1 I(Z_{t-1} \leq \gamma) + \phi_2 I(Z_{t-1} > \gamma)]Y_{t-1}$ es la esperanza condicional de Y_t dado $\mathfrak{F}_{t-1}$. Russell-Noriega (2006) demuestra que los estimadores de MC $\hat{\phi}_1(\gamma)$, $\hat{\phi}_2(\gamma)$ y $\hat{\sigma}^2(\gamma)$ para γ fijo son estimadores consistentes y asintóticamente normales. La estimación para γ de se da por medio de un procedimiento de búsqueda directa, de tal forma que dicho valor minimice el valor de la varianza $\hat{\sigma}^2(\gamma)$.

Máxima verosimilitud (MV): Sea el modelo TAR de la ecuación (1) y Z_{t-1} ∼AR(1) estacionario con ruido $u_t \sim iidN\left(0, \sigma_u^2\right)$ y coeficiente ρ. La función de verosimilitud condicional del vector de parámetros $\boldsymbol{\theta} = \left(\phi_1, \phi_2, \sigma^2, \gamma, \rho, \sigma_u^2\right)'$, para una serie observada $Y_1, ..., Y_T, Z_0, ..., Z_{T-1}$ con $\varepsilon_t \sim iidN\left(0, \sigma^2\right)$, está dada por: $L(\boldsymbol{\theta}; Y_t, Z_{t-1}) = (\sigma^{T-1})^{-1} \exp[-(2\sigma^2)^{-1} \sum_{t=2}^{T}(Y_t - \delta_t Y_{t-1})^2](\sigma_u^{T-1})^{-1} \exp[-(2\sigma_u^2)\sum_{t=2}^{T}(Z_{t-1} - \rho Z_{t-2})^2]$. En Russell-Noriega (2006) se demuestra que los EMV para γ conocido son estimadores consistentes y asintóticamente normales. Bajo normalidad en los errores los estimadores de MC y MV son iguales. Para γ desconocido y bajo ciertas condiciones de regularidad se demuestra también que los estimadores de MV son consis-

tentes. La estimación del parámetro γ es posible, i.e. calculados los EMV para γ fijo, se evalúa la función de log verosimilitud en estos valores y se obtiene la función de log verosimilitud perfil, $\ell_p^*(\gamma) = -\frac{T-1}{2}\log[\frac{1}{T-1}\sum_{t=2}^{T}(Y_t - \hat{\delta}_t(\gamma)\, Y_{t-1})^2] - \frac{T-1}{2}$.

4. Alternativas de pronósticos en modelos TAR

Monte Carlo Recursivo (MCR): Este método es el más utilizado en la literatura. Su uso no está ligado a un método particular de estimación. Al utilizarse para la obtención y evaluación de pronósticos en modelos TAR la conclusión en general en la literatura es que: los pronósticos de un modelo TAR no son mejores que los pronósticos bajo un modelo AR, en el sentido de que sus ECM son mayores que los ECM bajo el modelo lineal AR.

El pronóstico h-pasos adelante por la alternativa MCR consiste en aproximar la esperanza condicional por medio de un procedimiento recursivo: $\hat{Y}_{T+h} = E(Y_{T+h}|\mathfrak{F}_T) = E(\delta_{T+h}Y_{T+(h-1)}+\varepsilon_{T+h}|\mathfrak{F}_T)$ $= E(\delta_{T+h}Y_{T+(h-1)}|\mathfrak{F}_T)$, ó $\hat{Y}_{T+h} = \phi_1 E[I(Z_{T+h-1} \leq \gamma)Y_{T+h-1}|\mathfrak{F}_T]+\phi_2 E[I(Z_{T+h-1} > \gamma)Y_{T+h-1}|\mathfrak{F}_T]$. A diferencia de los modelos SETAR en los modelos TAR tenemos que, para obtener el pronóstico h-pasos adelante para Y_t, es necesario conocer el valor de la variable umbral Z_t al tiempo $T+h-1$ o equivalentemente conocer el valor de la variable indicadora $I(Z_{T+h-1} \leq \gamma)$. De esta forma aproximamos el pronóstico para Y_t por medio de las siguientes ecuaciones, según utilicemos el pronóstico de Z_{T+h-1} o de $I(Z_{T+h-1} \leq \gamma)$ respectivamente, $\hat{Y}_{T+h} \simeq \phi_1 I(\hat{Z}_{T+h-1} \leq \gamma)\hat{Y}_{T+h-1} + \phi_2 I(\hat{Z}_{T+h-1} > \gamma)\hat{Y}_{T+h-1}$, ó $\hat{Y}_{T+h} \simeq \phi_1 \hat{I}(Z_{T+h-1} \leq \gamma)\hat{Y}_{T+h-1} + \phi_2 \hat{I}(Z_{T+h-1} > \gamma)\hat{Y}_{T+h-1}$. Para γ conocido, pronosticar Z_t implica conocer o seleccionar un modelo que describa el comportamiento de dicha variable, mientras que pronosticar $I(Z_{T+h-1} \leq \gamma)$ implica considerar la teoría relacionada a series binarias, estudiada en Kedem y Fokianos (2002).

Verosimilitud Predictiva (VP): La propuesta de verosimilitud predictiva para ordenar la plausibilidad de valores futuros, consiste fundamentalmente en lo siguiente: Sean $\mathbf{x}^*$ y $\boldsymbol{\theta}$ dos cantidades desconocidas, donde el interés primario es obtener información sobre $\mathbf{x}^*$ con $\boldsymbol{\theta}$ jugando el papel de parámetro de estorbo. El principio de verosimilitud para predicción, parte de que toda la evidencia sobre $(\mathbf{x}^*, \boldsymbol{\theta})$ está contenida en la función de verosimilitud conjunta, $L(\mathbf{x}^*,\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x},\mathbf{x}^*;\boldsymbol{\theta})$ y el objetivo es desarrollar una función de verosimilitud para $\mathbf{x}^*$ conocida como verosimilitud predictiva, digamos $L(\mathbf{x}^*|\mathbf{x})$ eliminando $\boldsymbol{\theta}$ de la ecuación anterior. La propuesta que consideramos consiste en eliminar $\boldsymbol{\theta}$ maximizando $L_p(\mathbf{x}^*|\mathbf{x}) = \sup_{\boldsymbol{\theta}} f(\mathbf{x},\mathbf{x}^*;\boldsymbol{\theta}) = L(\mathbf{x}^*, \hat{\boldsymbol{\theta}}_{\mathbf{w}};\mathbf{x})$, con $\mathbf{w} = (\mathbf{x},\mathbf{x}^*)'$, Bjørnstad

(1990). Intuitivamente, la motivación de la función de verosimilitud predictiva L_p es como sigue: para $\mathbf{x}^*$ el vector de parámetros de interés y $\boldsymbol{\theta}$ el vector de parámetros de estorbo, se obtienen los valores más probables para $\boldsymbol{\theta}$ dado $\mathbf{w}$ dando como resultado la función de verosimilitud. En inferencia paramétrica esta verosimilitud corresponde a la verosimilitud perfil y de aquí el nombre de verosimilitud predictiva perfil. Ver Russell-Noriega (2006) para más detalles y referencias.

Regresión por Cuantiles (RC): La alternativa de regresión por cuantiles permite estudiar dinámicas asimétricas presentes en los modelos TAR, introduce además una medida de cuantificación distinta al ECMP la cual reproduce de forma natural la asimetría del modelo dentro y fuera de la muestra observada. Implementamos sobre dicho modelo un proceso de predicción. Los estimadores de RC son consistentes y asintóticamente normales para γ fijo, Russell-Noriega (2006). Consideramos el modelo TAR dado en la ecuación (1) el cual reescribimos como: $Y_t = f_t(\boldsymbol{\beta}, Y_{t-1}, Z_{t-1}) + \varepsilon_{t\theta},\ t = 1, \ldots T$, con $f_t(\boldsymbol{\beta}, Y_{t-1}, Z_{t-1}) = \phi_1 I(Z_{t-1} \leq \gamma) Y_{t-1} + \phi_2 I(Z_{t-1} > \gamma) Y_{t-1}$ y $\boldsymbol{\beta}= (\phi_1, \phi_2)'$ el vector de parámetros en el modelo, considerando el parámetro de umbral γ conocido, con $\varepsilon_{t\theta} \sim iidN(0, \sigma^2)$. El θ-ésimo cuantil no lineal con $0 < \theta < 1$ es cualquier vector $\widehat{\boldsymbol{\beta}}(\theta)$ que minimice la función $\text{mín}_{\boldsymbol{\beta}} \sum_{t=1}^{T} \{\theta - I[Y_t < f_t(\boldsymbol{\beta}, X_{t-1}, Z_t)]\}[Y_t - f_t(\boldsymbol{\beta}, X_{t-1}, Z_t)]$ con respecto a β.

Identificamos que el problema del mal desempeño de los pronósticos en modelos TAR, se debe principalmente a problemas de especificaciones incorrectas, como son: Clasificación incorrecta de las observaciones en cada uno de los regímenes (Dacco y Satchell (1999), Russell-Noriega (2006)). Probabilidades grandes de mala clasificación en cada uno de los regímenes implicarán ECM de pronósticos grandes. Así como estimaciones incorrectas del parámetro γ o malos pronósticos Z_t implicarán desempeños pobres de los pronósticos bajo un TAR. Se añade también el efecto de modelos mal especificados en la obtención de los pronósticos; esto es, al suponer un modelo AR mostramos que el proceso de los errores se identifica como un proceso de volatilidad. A pesar de que los errores estimados bajo un AR se identifican como un proceso de volatilidad, si se utiliza Monte Carlo para aproximar los pronósticos y el ECMP notamos que el aumento en la variabilidad de error bajo el AR no compensa el problema de la mala especificación del régimen y de la medida utilizada.

5. Resultados y descripción del ejercicio de simulación

Se generan 500 series de tiempo de tamaño 255, provenientes de un modelo TAR(2;1,1) dado en la ecuación (1). Los escenarios simulados contemplan una gran variedad de situaciones de interés, sin

embargo por cuestiones de espacio presentamos los resultados para un sólo caso, sin embargo las conclusiones para el resto de los casos son equivantentes al aquí presentado, Russell- Noriega (2006). Para cada una de las series simuladas, con γ conocido, consideramos las primeras 250 observaciones simuladas para la fase de estimación, mientras que las últimas 5 observaciones de cada serie las usamos para la comparación de los pronósticos por medio de las pérdidas ECMP y de cuantiles. Sean $\phi_1 = -0{,}1$, $\phi_2 = 1$, $\Pr(Z_{t-1} \leq \gamma) = 0{,}5$, $E(\delta_t) = 0{,}45$ y $E(\delta_t^2) = 0{,}505$, este caso corresponde al modelo TUR estudiado en González y Gonzalo (1998). El comportamiento gráfico de los 500 valores estimados para cada uno de los parámetros bajo los modelos considerados se muestra en la Figura 1. Las líneas verticales, en el caso de los histogramas asociados al modelo TAR$(2;1,1)$ (primeros tres histogramas en cada renglón de la Figura 1) indican los verdaderos valores de los parámetros bajo dicho modelo. Los últimos dos histogramas de cada renglón están asociados al modelo lineal AR(1), donde las líneas verticales indican la $E(\delta_t)$, y varianza del error $\sigma^2 = 1$, respectivamente.

La Figura 2 representa los comportamientos promedios de los ECMP y de la pérdida de cuantiles para cada una de las alternativas de pronóstico bajo los modelos TAR$(2;1,1)$ y AR(1). La línea punteada identificada con la etiqueta TAR_VP está asociada al modelo TAR$(2;1,1)$ con pronósticos generados por verosimilitud predictiva. Observamos que para ambas funciones de pérdida, la ganancia del método de VP es clara.

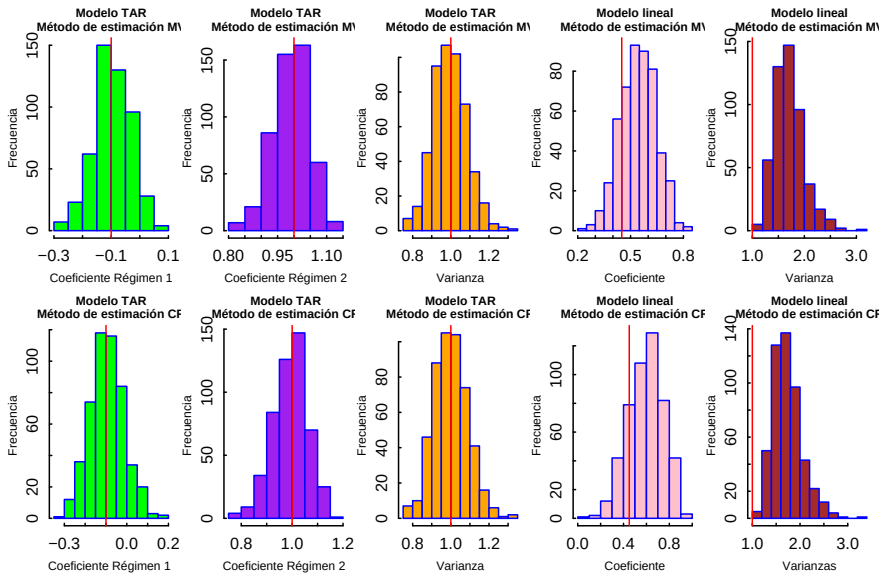


Figura 1. Histogramas de los parámetros estimados, bajo el modelo TAR$(2;1,1)$ y modelo AR(1) respectivamente.

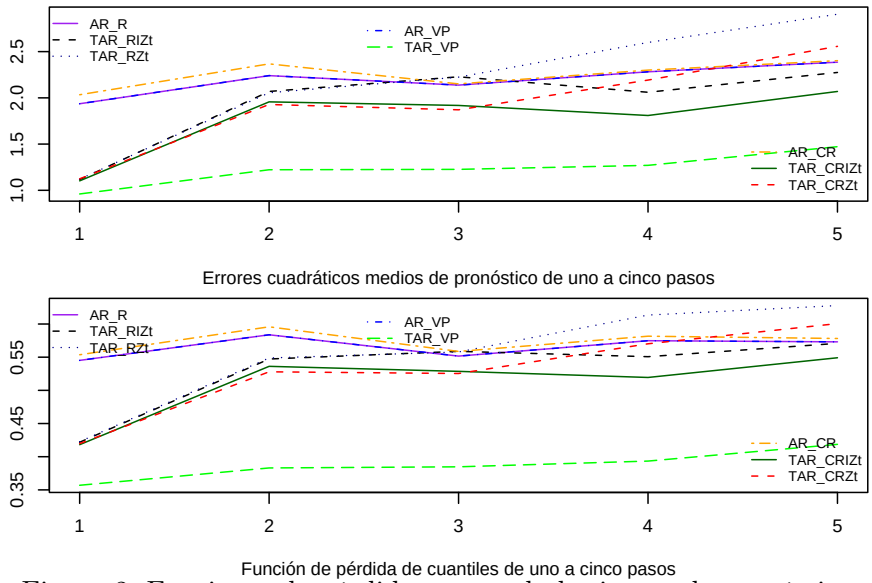


Figura 2. Funciones de pérdida para cada horizonte de pronóstico.

6. Conclusiones

Ante especificaciones erróneas, la asimetría natural del modelo TAR se "hereda" a los residuales en el modelo AR. Bajo normalidad, las estimaciones vía MC y MV son iguales, sin embargo, los métodos de pronósticos asociados a MC y MV para modelos TAR no son iguales; mientras que bajo el modelo AR sí lo son.

VP considera toda la información de manera conjunta, utilizando así las propiedades probabilísticas de la variable umbral, lo cual hace más eficiente el proceso de optimización. Los métodos de RC y MCR presuponen conocida la variable umbral, ya que el proceso de obtención de los pronósticos de dicha variable se realiza como un paso previo. La recomendación en base al estudio de simulación, es generar los pronósticos TAR utilizando los pronósticos de la variable indicadora. Una ventaja de regresión por cuantiles es que no se requiere de supuestos distribucionales para los errores.

7. Referencias

Bjœrnstad, J. F. (1990). Predictive Likelihood: A Review. *Stat. Science.* **5**, No. 1, 242-265.

Chan, K. S., Petruccelli ,J. D, Tong, H. and Woolford, S. W. (1985). A Multiple Threshold AR(1) Model. *Journal App. Probability.* **22**, 267-279

Dacco, R. and Satchell, S. (1999). Why do Regime-Switching Models Forecast so Badly. *Journal of Forecasting.* **18**, 1-16.

González, M. and Gonzalo, J. (1998). Threshold Unit Root Models. *Working paper, Universidad Carlos III.*

Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis.* Wiley Series in Probability and Statistics.

Russell Noriega (2006). Pronósticos en modelos autorregresivos con umbral. Tesis de doctorado del Centro de Investigación en Matemáticas, A.C.

Inferencia sobre el punto de cambio estructural en modelos lineales

Blanca Rosa Pérez Salvador[1] **y Alberto Castillo Morales**[2]

Universidad Autónoma Metropolitana, Iztapalapa

Departamento de Matemáticas

1. Introducción

Considerere un conjunto de observaciones Y_t indexados en el tiempo, tales que

$$Y_t = \begin{cases} \mu + \varepsilon_i & \text{si } t \leq m \\ \mu^* + \varepsilon_i & \text{si } t > m \end{cases}$$

donde $\mu \neq \mu^*$ son parámetros desconocidos, y $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2 < \infty$. El número m se conoce como el punto de cambio.

Este modelo se utiliza entre otras cosas para detectar si hubo impacto social de un programa implementado por el gobierno; o para detectar si hubo cambios en la filtración del suelo de una presa, después de ocurrido un sismo; o para detectar posibles cambios en los índices de pobreza, después de aplicar algunas políticas de estado.

En este contexto, se puede efectuar la prueba

$$H_0 : \ m \geq n \quad \text{contra} \quad H_a : \ m < n.$$

Observe que la hipótesis nula indica que en el periodo de observación no se encuentra el punto de cambio, mientras que la hipótesis alternativa indica que en el periodo de observación se presentó el punto de cambio.

Este es un problema ampliamente estudiado y a lo largo del tiempo se han propuesto varias estadísticas de prueba, entre las cuales se tienen las siguientes.

Sen y Srivastava (1975) propusieron las siguientes tres estadísticas de prueba y encontraron su

[1] psbr@xanum.uam.mx
[2] aacm@xanum.uam.mx

potencia para la hipótesis alternativa mediante el método de Monte Carlo,

$$S = \max_{1\le r\le N-1} \frac{\overline{Y}_{n-r} - \overline{Y}_r}{\sqrt{\left(\frac{1}{r}+\frac{1}{n-r}\right)\left(\sum_{i=1}^{r}(Y_i-\overline{Y}_r)^2+\sum_{i=r+1}^{n}(Y_i-\overline{Y}_{n-r})^2\right)}}$$

$$P = \frac{\sum_{i=1}^{n-1} i(Y_{i+1}-\overline{Y})}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{N-1}}} \qquad P_1 = \frac{\sum_{i=1}^{n-1} i(Y_{i+1}-\overline{Y})}{\sqrt{\frac{\sum_{i=1}^{n-1}(Y_{i+1}-Y_i)^2}{2(N-1)}}}$$

James, Ling y Siegmund (1987) presentaron las siguientes estadísticas y también calcularon la potencia aproximada.

$$\max_{1\le k\le n}[|kS_n/n - S_k|/\{k(1-k/m)\}^{1/2}], \quad \max_{1\le k\le n}[kS_n/n - S_k], \quad \max_{1\le k\le n}[kS_n/n - S_k - \frac{1}{2}k - k/n)\delta_0],$$

para un valor arbitrario de δ_0 y $S_n = Y_1 + Y_2 + \ldots + Y_n$ y con el cambio de variable, $z_i = \frac{n}{\sqrt{n+1}}(Y_{i+1} - \overline{Y}_i)$ y $\tilde{S}_i = z_1 + \ldots + z_i$ presentaron las estadísticas

$$\max_{x_0\le k\le n} \tilde{S}_i/n^{1/2}, \quad \max_{x_0\le k\le n}(\tilde{S}_{n-1} - \tilde{S}_{n-k})/(k-1)^{1/2} \quad \text{y} \quad C = \sum_{i=1}^{n-1}\sqrt{i(i+1)}z_i$$

$$\max_{1\le k<n}\left[-\frac{1}{2}n\,\log\left\{1 - \frac{(S_k - kS_n/n)^2}{k(1-k/n)\sum(x_i-x_n)^2}\right\}\right]$$

Gombay y Horvath (1990) consideraron las siguientes estadísticas,

$$Z(i,j) = \max_{i\le k\le j}(Z_k/g''(\mu))$$

para una función g y $Z_k = 2(k\,g(\overline{X}_k) + (n-k)g(\overline{X}_{n-k}) - ng(\overline{X}_n)$

Antoch y Hušcová (2001), proponen como estadística de prueba la expresión

$$T_{nl}(R) = \max_{1<k<n}\left\{\sqrt{\frac{n}{k(n-k)s^2}}\left|\sum_{i=1}^{k}(X_{Ri} - \overline{X}_n)\right|\right\}$$

donde R indica que el cálculo se hace sobre todas las posibles permutaciones de los datos muestrales, y encuentran que

$$P(\sqrt{2\log\log nT_{n1}(R)} \leq y + 2\log\log n - \frac{1}{2}\log\log\log n - \frac{1}{2}\log\pi \mid X_1, \ldots, X_n)$$

$$\longrightarrow exp\{-2exp\{-y\}\}, \quad \text{casi seguramente}$$

En este trabajo se presenta una región crítica diferente y la función de densidad de la estadística de prueba, lo que se ve en las secciones siguientes.

2. La propuesta

Proposición 1 Si $X_i,\ \ i = 1, 2, 3, \ldots, n$, es una muestra aleatoria tal que

$$X_i = \mu + \delta_n I_{\{i>m\}} + \varepsilon_i \quad \text{con } \delta_n \neq 0 \ \text{ y } \ \varepsilon \sim N(0, \sigma^2)$$

la región crítica obtenida por el cociente de verosimilitud para la prueba

$$H_0 :\ m \geq n \quad \text{contra} \quad H_a :\ m < n$$

es

$$\frac{\text{mín}_m(\sum_{i=1}^{m}(X_i - \overline{X}_1)^2 + \sum_{i=m+1}^{n}(X_i - \overline{X}_2)^2)}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \leq \lambda$$

Observe que en esta estadística se compara la variación de dos submuestras con la variación en toda la muestra. Si en la región de observación se presenta un cambio, es de esperarse que el numerador sea mucho más pequeño que el denominador.

3. La distribución de la estadística de prueba

La estadística de prueba se puede escribir como $\text{mín}_{1\leq m<n} \frac{Y_{m+1}}{Y_1} \leq \lambda$, donde

$Y_1 = \sum_{i=1}^{n}(X_i - \overline{X})^2$

$Y_2 = \sum_{i=2}^{n}(X_i - \frac{\sum_{j=2}^{n} X_j}{n-1})^2$

$Y_3 = \sum_{i=1}^{2}(X_i - \frac{X_1+X_2}{2})^2 + \sum_{i=3}^{n}(X_i - \frac{\sum_{j=2}^{n} X_j}{n-2})^2$

$\vdots$

$Y_m = \sum_{i=1}^{m-1}(X_i - \frac{\sum_{j=1}^{m-1} X_j}{m-1})^2 + \sum_{i=m}^{n}(X_i - \frac{\sum_{j=m}^{n} X_j}{n-m+1})^2$

$\vdots$

$Y_n = \sum_{i=1}^{n-1}(X_i - \frac{\sum_{j=1}^{n-1} X_j}{n-1})^2$

La suposición del modelo implica que $Y_1/\sigma^2 \sim \chi^2_{n-1}$ y para $i = 2, 3, \ldots, n$ $Y_i/\sigma^2 \sim \chi^2_{n-2}$, por lo tanto Y_1 se puede escribir como la suma de $n-1$ variables normales independientes al cuadrado con media 0 y varianza 1, y para $i = 2, \ldots, n$, Y_i se puede escribir en términos de las mismas variables. Es importante notar que Y_i y Y_j no son independientes para $i \neq j$.

Proposición 2 Si Y_1, Y_2, ..., Y_n son las variables definidas antes, entonces bajo H_0, se sigue que $Y_1 = W^T W$, and $Y_{m+1} = Y_1 - \frac{n(n-m)}{m}(\sum_{i=1}^{m} \frac{w_i}{\sqrt{(n-i)(n-i+1)}})^2$ con $W \sim N(0, \sigma^2 I_{(n-1)})$.

Proposición 3 Sea el vector $V^T = (V_1, \ldots, V_{n-2})$ con $V_i = W_i/\sqrt{\sum_{i=1}^{n-1} W_i^2}$, entonces su función de densidad conjunta es

$$f_V(v_1, v_2, \ldots, v_{n-2}) = \begin{cases} \frac{\Gamma((n-1)/2)}{(\pi)^{(n-1)/2}\sqrt{1-v_1^2-v_2^2-\ldots-v_{n-2}^2}}, & \text{si } v_1^2 + \cdots + v_{n-2}^2 < 1, \\ 0, & \text{en otro caso.} \end{cases}$$

Entonces la función $f_V(v_1, v_2, \ldots, v_{n-2})$ no depende de ningún parametro desconocido, y la región crítica obtenida usando la densidad anterior depende sólo de la muestra.

Proposición 4 Sea $V_i = W_i/W^TW$ y $U_i = \sqrt{\frac{n(n-i)}{i}} \sum_{j=1}^{i} \frac{V_j}{\sqrt{(n-j)(n-j+1)}}$, las variables aleatorias $W \sim N(0, \sigma^2 I)$ para $i = 1, 2, \ldots, n-1$; y sea $|U|_{(n-1)} = \text{máx}_m |U_{1 \leq m \leq n-1}|$. La función de distribución de $|U|_{(n-1)}$ es

$$F_{|U|_{(n-1)}}(u) = \int_{A_1} \ldots \int_{A_{n-3}} \int_{A_{n-2} \cap A_{n-1}} f_V(v_1, v_2, \ldots, v_{n-2}) dv_{n-2} \ldots dv_1$$

donde $A_{iu} = \{v \in R^{n-2} \mid \left| \sum_{j=1}^{i} \frac{v_j}{\sqrt{(n-j)(n-j+1)}} \right| \leq \sqrt{\frac{i}{n(n-i)}} u\}$.

Por su definición, se tiene que $Y_{m+1}/Y_1 = 1 - U_m^2$, entonces la región $\text{mín}_m Y_{m+1}/Y_1 < \lambda$ es equivalente a la región $\text{máx}_m |U_m| > u_0$.

4. Referencias

Antoch, J. y Huŭková, M. (2001). Permutation test in change point analysis; *Statistics & Probability letters*, **no. 53**; pp 37-46.

Gombay, E. y Horvath, L. (1990). Asymptotic distributions of maximum likelihood test for change in the mean; *Biometrika* ; **77** no. 2 pp 411-414

James B.; Ling James K. y Siegmund D. (1987). Tests for a change/point, *Biometrika*; **74**, no. 1, pp 71-83

Sen, A. y Srivastava, M. S. (Febreaury 1975). Some One-Side Tests for Change in Level; *Technometrics*, **vol. 17**, No. 1; ; pp 61-64.

Bayesian detection of active effects in factorial experiments with dichotomous response

Román de la Vara[1]
Centro de Investigación en Matemáticas
Víctor Aguirre-Torres[2]
Instituto Tecnológico Autónomo de México

1. Introduction

In many industrial experiments the response variable is discrete. Notably, when the data are counts or proportion of defectives (see Lewis, *et al.* 2001). The traditional approach for analyzing such data is to apply a variance-stabilizing transformation to the response variable, and then use ordinary least squares with the transformed data. Hopefully the transformation will also induce normality and constancy of variance on the response and simplify the empirical model. Another approach is to use a generalized linear model (GLM), in which the normality and constant variance is no longer required (Myers and Montgomery, 1997; Hamada and Nelder, 1997).

This paper proposes a Bayesian method for detecting active effects in an unreplicated factorial experiment analyzed by a GLM with dichotomous response. The idea is to generalize the Box and Meyer (1993) method for normal case.

2. Generalized Linear Model

We have $\mathbf{y}^t = [y_1, y_2, \ldots, y_n]$ a vector of independent observations, with vector of means $[\mu_1, \mu_2, \ldots, \mu_n]$. The observation y_i has a distribution that is a member of the exponential family. The systematic part of the model involves the factors of the experiment represented by the variables $x_1, ...x_k$. The model is built around the linear predictor $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. The model is found through the use of a link function $\eta_i = g(\mu_i)$, which relates the linear predictor with the mean (μ_i) of the

[1] delavara@cimat.mx
[2] aguirre@itam.mx

specified distribution for the response variable. The link function is monotonic and differentiable. The canonical link is such that $\eta_i = \zeta_i$ and $Var(y_i)$ is a function of μ_i (McCullagh, P. y Nelder, J. A. (1989)).

3. Bayesian Model Selection

If the distribution of $\mathbf{y}$ given the model is denoted by $f(\mathbf{y}\,|M_i, \boldsymbol{\theta}_i)$, the a priori probability of M_i by $p(M_i)$ and the *prior* probability density of $\boldsymbol{\theta}_i$ is $f(\boldsymbol{\theta}_i\,|M_i)$, then the prior predictive density (ppd) of $\mathbf{y}$ or integrated likelihood, given the model M_i, is

$$f(\mathbf{y}\,|M_i) = \int_{\Theta_i} f(\mathbf{y}\,|M_i, \boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i\,|M_i) d\boldsymbol{\theta}_i \tag{1}$$

where Θ_i is the space set of $\boldsymbol{\theta}_i$. The posterior probability of the model M_i, given the data $\mathbf{y}$, is

$$p(M_i\,|\mathbf{y}\,) = \frac{p(M_i) f(\mathbf{y}\,|M_i)}{\sum_{h=0}^{m} p(M_h) f(\mathbf{y}\,|M_h)}. \tag{2}$$

The a priori probability of observing a model with t_i significant effects is $p(M_i) \propto \alpha^{t_i}(1-\alpha)^{n-t_i}$ where α is the a priori probability that any one effect is active, ($0 < \alpha <$ 0,4) assuming the factor sparcity principle (Box and Meyer, 1993). The posterior probability that the effect T_j is active is

$$P_j = \sum_{M_i : T_j \; is \; active} p(M_i\,|\mathbf{y}) \; . \tag{3}$$

In this work we propose to approximate the ppd (1) by using two methods: Quasi-Monte Carlo simulation and Bayesian information criteria (BIC).

4. Computing the Prior Predictive Density

4.1. Quasi-Monte Carlo Approach

Basically this approach consists in the average of the N evaluations of the prior distribution $f(\boldsymbol{\theta}_i\,|M_i)$ for approximating the integrated likelihood, that is,

$$f(\mathbf{y}\,|M_i) \approx \widehat{E}[f(\mathbf{y}\,|M_i, \boldsymbol{\theta}_i)] = \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{y}\,|M_i, \boldsymbol{\theta}_{ij})\;. \tag{4}$$

Consider first the parameter β_0 which is linked to original scale by the relationship $\beta_0 = g(\mu_0)$. In order to obtain a prior distribution for β_0 notice that μ_0 could be considered the mean response when none of the effects are significant, or mean response in the center of the experimental region. Whatever interpretation applies, this approach assumes that the experimenter has some broad idea about the value of this mean response and it is stated in terms of a probability interval of the form

$$P(L_\mu < \mu_0 < U_\mu) = 1 - \delta, \tag{5}$$

where L_μ and U_μ are a lower and upper bound for μ_0, and δ is a small fraction, say 5 % or 1 %. Assuming a strictly increasing link function and (5) it follows that

$$P(g(L_\mu) < \beta_0 < g(U_\mu)) = 1 - \delta \tag{6}$$

Then, one way to fulfil (6) with a normal distribution is to take the parameters for the prior density of β_0 as

$$\mu_{\beta_0} = \frac{g(L_\mu) + g(U_\mu)}{2}; \qquad \sigma_{\beta_0} = \frac{g(U_\mu) - \mu_{\beta_0}}{z_{1-(\delta/2)}} \tag{7}$$

where z_ξ is the $\xi - th$ percentile of the standard normal distribution.

For the rest of the parameters β_i, $1 \le i \le k$, we will assume a $N(0, \sigma^2_{\beta_0})$ distribution, where the zero mean is introduced since it is supposed that in advance there is no information about the sign

of the effect. In the example that we discuss later, for the case of the experiment with 7 effects, we use this approximation for models with the 128 possible models using $N = 1000$ quasi-random repetitions for each model.

4.2. The BIC Approach

In this approach (see Raftery (1995)) the posterior probability of the model M_k is given by

$$p(M_k \mid \mathbf{y}) \approx \frac{p(M_k)\exp(-\frac{1}{2}BIC_k)}{\sum_{h=0}^{m} p(M_h)\exp(-\frac{1}{2}BIC_h)}. \tag{8}$$

with $BIC_k = D_k^2 - df_k \log n$, where D_k^2 the deviance for model M_k and df_k is the corresponding number of degrees of freedom. This method does not require the explicit specification of a prior density $f(\theta_i \,|M_i)$.

5. Binomial Response

Consider an experiment with binary response, let y_j be the number succeses observed out of n_j trials $(j = 1, 2, ..., n)$, processed at treatment j, with p_j the success probability in treatment $\mathbf{x}_j$. Can be shown that the likelihood function or the n observations, considering the logit link $\eta_j = \log(\frac{p_j}{1-p_j})$, is given by in the experiment is given by

$$f\left(\mathbf{y} \,|M_i\, , \boldsymbol{\theta}_i\right) \propto \prod_{j=1}^{n} \left(e^{y_j \mathbf{x}_j^t \boldsymbol{\theta}_i}\right) \left(\frac{1}{1+e^{\mathbf{x}_j^t \boldsymbol{\theta}_i}}\right)^{n_j} \tag{9}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ is the parameter vector in the linear predictor.Once again, we make the mild assumption that the experimenter has some vague idea on the proportion of defectives p_0. That is we expect an interval of the form $L_p < p_0 < U_p$ and some probability associated to this interval, say $1 - \delta$. In this case the link function is given in terms of $\eta_i = \log(\frac{p_i}{1-p_i})$, hence we have the relation $\beta_0 = \log(\frac{p_0}{1-p_0}) = g(p_0)$, and the formulas (7) can be applied directly to obtain the hyperparameters as follows, since the link function is increasing:

$$\mu_{\beta_0} = \frac{\log(\frac{L_p}{1\text{-}L_p}) + \log(\frac{U_p}{1-U_p})}{2}; \qquad \sigma_{\beta_0} = \frac{\log(\frac{U_p}{1-U_p}) - \mu_{\beta_0}}{z_{1-(\delta/2)}}, \tag{10}$$

notice that if the interval $L_p < p_0 < U_p$ is symmetric around 0,5 then $\mu_{\beta_0} = 0$. Considering again the prior normal distribution with mean zero and variance $\sigma^2_{\beta_0}$ for the parameters in the model M_i that multiply the independent variables, the following predictive prior density is obtained

$$f(\mathbf{y}\,|M_i) = \int_{\Theta_i} \prod_{j=1}^{n} \left(e^{\mathbf{x}_j^t \boldsymbol{\theta}_i}\right)^{y_j} \left(\frac{1}{1+e^{\mathbf{x}_j^t \boldsymbol{\theta}_i}}\right)^{n_j} \frac{1}{(2\pi)^{t_i/2}\left(\sigma_{\beta_0}\right)^{t_i}} \times \tag{11}$$

$$e^{-(1/2\sigma^2_{\beta_0})\sum_{k=1}^{t_i}\beta_k^2} \times \frac{1}{(2\pi)^{1/2}\sigma_{\beta_0}} e^{-(1/2\sigma^2_{\beta_0})(\beta_0-\mu_{\beta_0})^2} d\boldsymbol{\theta}_i.$$

5.1. Example: Survival of Sperm Experiment.

The 2^3 factorial experiment given in Table 1 study the number of sperm samples that survive, meaning that the sample has the ability to impregnate (Myers, *et al.* 2002, page 116), where fifty samples of material were used in each experimental point.

Table 1. Data from Survival of Sperm Experiment.

A (Sodium Citrate)	B (Glycerol)	C (Equilibrium Time)	Y (survived)
−1	−1	−1	34
1	−1	−1	20
−1	1	−1	8
1	1	−1	21
−1	−1	1	30
1	−1	1	20
−1	1	1	10
1	1	1	25

In the frequentist analysis Myers *et al.* (2002) find, using Wald test, that the active effects are B (glycerol) and AB (interaction of sodium citrate and glycerol). Assuming that the experimenter

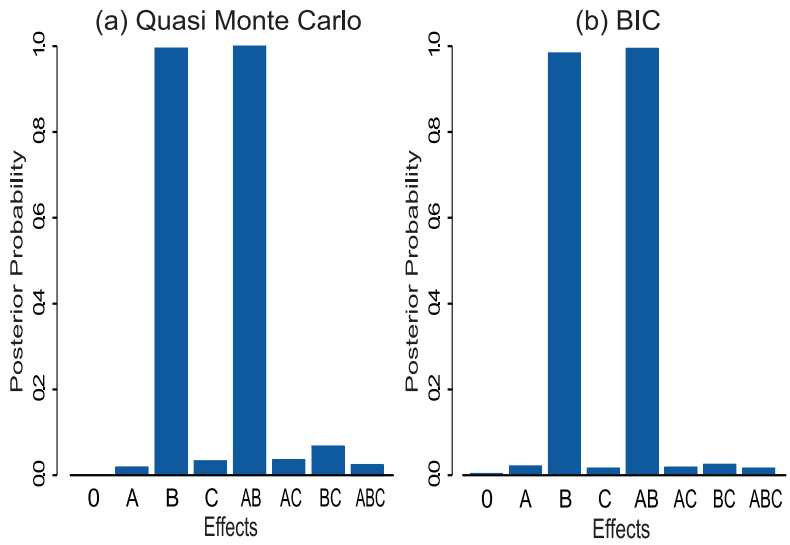


Figura 1: Post. probability of being active, binomial logit link, (a) Quasi Monte Carlo and (b) BIC approaches

expects a proportion survival of $0{,}1 < p_0 < 0{,}9$ with probability of 99 % then the following hyper parameters are obtained using (10)

$$\mu_{\beta_0} = 0; \quad \sigma_{\beta_0} = 0{,}85.$$

Figure 1(a) is obtained with a priori probability of active effect $\alpha = 0{,}2$. Clearly the effects B and AB are active, with posterior probabilities almost equal to one, while the other effects have smallish posterior probabilities. This result agrees with the frequentist analysis. The BIC approximation detects the same active effects A and AB (see Figure 1(b)).

5.2. Concluding Remarks

In this paper we propose a method based on Bayesian model selection for detecting the active effects in unreplicated factorial experiments where a dichotomous response variable is modeled by a GLM.

Two approximations of the integrated likelihood were considered, the Quasi-Monte Carlo simulation approach and the BIC approach. Both approximations to the integrated likelihood gave good results in the discussed binomial example.

6. References

Box, G. E. P. and Meyer, R. (1993). Finding the active factors in fractionated screenig experiments. *Journal of Quality Technology,* 25, pp. 94-104.

Hamada, M. and Nelder, J. A. (1997). Generalized linear models for quality-improvement experiments. *Journal of Quality Technology,* 29, pp. 292-303.

Lewis, S. L. and Montgomery, D. C. (2001). Examples of designed experiments with nonnormal responses. *Journal of Quality Technology,* 33, pp. 265-278.

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2nd ed., Chapman and Hall, New York.

Myers, R. H. and Montgomery, D. C. (1997). A tutorial on generalized linear models. *Journal of Quality Technology,* 29, pp. 274-291.

Myers, R. H., Montgomery, D. C. and G. Vining (2002). Generalized Linear Models. Wiley, New York.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion) *Sociological Methodology,* 25, pp. 111-196.

Optimización simultánea multi-respuesta aplicando técnicas de graficación[1]

Luz Vanessa Bacio Parra[2]
Centro de Investigación en Matemáticas
Jorge Domínguez Domínguez[3]
Centro de Investigación en Matemáticas

1. Introducción

Es común que en diferentes áreas de estudio se consideren problemas representados por muchas características de interés. El tipo de diseño que se utiliza involucra la selección de un conjunto de factores de interés que resultarán en un producto con la mejor combinación de estas características. A este proceso se le conoce como un diseño de optimización multi-respuesta en la que las características de interés se definen como múltiple respuesta. El objetivo es definir un conjunto de factores que proporcionen la mejor mediación simultáneamente de la múltiple respuesta.

Una aproximación usada comúnmente para resolver los problemas de diseño multi-respuesta es considerar una función objetivo unificada; esto es, que las respuestas individuales son matemáticamente combinadas para generar una función simple. Inicialmente las variables de respuesta individuales son modeladas para crear una superficie de respuesta de un diseño experimental. A cada variable de respuesta se le aplica una transformación de tal manera que todas las respuestas se puedan combinar en una sola función. A partir de ahí se varían los niveles de los factores tal que se puedan cumplir de la mejor manera los óptimos individuales hasta alcanzar un óptimo global.

En esta presentación se mostrarán cinco procedimientos para construir una función que represente a la combinación de los objetivos de las respuestas individuales. Estos procesos se caracterizan por ser modelos de optimización multiplicativos o aditivos, mediante un ejemplo se hará una comparación de estos métodos. Los resultados de los procedimientos de optimización se ilustrarán mediante técnicas de graficación.

[1]Trabajo apoyado por el Consejo de Ciencia y Tecnología de Guanajuato: convenio 06-02-K117-87.
[2]lbacio@cimat.mx
[3]jorge@cimat.mx

2. Descripción del problema Optimización multi-respuesta

La información se genera mediante un esquema experimental, la matriz $D(n \times k)$ representa alguno de estos esquemas, donde n es el número de combinaciones (tratamientos), de valores de los k factores $(x_1, ..., x_k)$ (variables de entrada al proceso). Varios esquemas experimentales se pueden plantear para este proyecto, tales como, diseños factoriales, diseños factoriales fraccionados, diseño Box - Benhken, diseño central compuesto y diseños óptimos entre otros, ver Box-Draper (1987) y Myers-Montgomery (2002). Se tienen r respuestas para cada respuesta en las n combinaciones, con la información generada por el experimento se pueden modelar de manera individual cada una de las r respuestas. Por lo general estos modelos son lineales y de forma cuadrática, estos están en función de los k factores. Así para r respuestas se tienen r modelos, el $j-$ésimo modelo, un polinomio de orden 2, para esa respuesta Y_j se escribe como: $Y_j = \beta_{j0} + X^t\beta_j + X^tB_jX + \varepsilon_j$, $j = 1, ..., r$, donde X^t $=(x_1, ..., x_k)$, β_0 la constante, $\beta^t = (\beta_1, ..., \beta_k)$ un vector de parámetros $B = (\beta_{11}, ..., \beta_{1k}, \beta_{k1}, ..., \beta_{kk})$ matriz simétrica de parámetros de segundo orden, en esta situación $q = k(k+3)/2+1$, q corresponde al número de términos: constante, lineales y de segundo orden, ε_j es una variable aleatoria que tiene una distribución de probabilidad. Para la teoría que se expondrá en esta presentación se hará bajo el supuesto de que ε_j sigue una distribución normal con media cero y varianza σ_j^2 es decir: $\varepsilon_j \sim N(0, \sigma_j^2)$. Es también relevante considerar esta temática de optimización multi respuesta para el caso de que la variable aleatoria ε_j siga una distribución de probabilidad no normal.

El problema consiste en determinar la mejor combinación de los factores tal que produzcan el óptimo global, es decir que todas las respuestas den su mejor valor. Un planteamiento más general del problema de optimización en presencia de j objetivos: O_j flexibles $j = 1, ..., r$ $(Y_1 = O_1, ..., Y_r = O_{r-1})$ cada una de las Y_j es un vector de observaciones en los n tratamientos, y l restricciones $g_l(x)$, en esta presentación se utilizá el caso $X\epsilon\mathfrak{R}$, $\mathfrak{R}$: región experimental, así:

$$\begin{array}{ll} \text{Optimizar} & [Y_1, Y_2, ..., Y_r], \\ \text{Sujeto a} & g_l(x),\ l = 1, 2, ..., m. \end{array} \tag{1}$$

Como resultado de la estrategia experimental los modelos Y_j se sustituyen por los mejores modelos que genera el método de mínimos cuadrados Box-Draper (1987) y Khuri -Cornell (1996), y el *j-ésimo* modelo de manera simplificada se expresa por:

$$\widehat{Y}_j = \widehat{\beta}_{j0} + X^t\widehat{\beta}_j + X^t\widehat{B}_jX. \tag{2}$$

La solución por el método de mínimos cuadrados es: $\widehat{\delta}_j = [(Z(x))^t Z(x)]^{-1}(Z(x))^t Y_j$, $Z(x)$ es una matriz de orden $(n \times q)$, donde n es el número de tratamientos y $q = k(k+3)/2 + 1$ es el número de términos, constante, lineales, cuadráticos e interacciones de segundo orden, $\widehat{\delta}_j = (\widehat{\beta}_{j0}, \widehat{\beta}_j, \widehat{B}_j)$, cada una de las $\widehat{Y}_j$ es un vector de valores predichos por el modelo en los n tratamientos, estos se sustituyen en los planteamientos (1).

3. Planteamientos de optimización multi-respuesta

A continuación se indican las referencias de algunos procedimientos de optimización desarrollados por diferentes autores. La expresión (1) describe el planteamiento típico de un problema de programación lineal, la solución es un valor x_o de X, que genera una respuesta óptima global bajo estas condiciones. Entre las ventajas de este procedimiento están su planteamiento matemático y que puede ser resuelto mediante una hoja de cálculo. Existen otros procedimientos analíticos eficientes de optimización, el principio de varios métodos consiste en transformar la variable de respuesta j *ésima* descrita en el modelo (2) a una escala de valores, denominada valor deseable u y sus valores están entre 0 y 1. Este valor crece conforme se requiera el mejor valor de la variable de respuesta correspondiente.

$$u_j(\widehat{Y}_j(x)) = \begin{cases} 0 & \text{si} \quad \widehat{Y}_j(x) < Y_j^{\text{mín}} \quad \text{o} \quad \widehat{Y}_j(x) > Y_j^{\text{máx}}, \\ 1 - \frac{M_j - \widehat{Y}_j(x)}{M_j - Y_j^{\text{mín}}} & \text{si} \quad Y_j^{\text{mín}} \leq \widehat{Y}_j(x) \leq M_j, \\ 1 - \frac{\widehat{Y}_j(x) - M_j}{Y_j^{\text{máx}} - M_j} & \text{si} \quad M_j < \widehat{Y}_j(x) < Y_j^{\text{máx}}, \end{cases} \tag{3}$$

donde M es un valor objetivo fijado de acuerdo al interés del investigador y $\left(Y_j^{\text{mín}}, Y_j^{\text{máx}}\right)$ son dos cotas en la j *ésima* respuesta; éstas se deben de especificar inicialmente. Existen varios criterios para determinar estas cotas, por ejemplo los límites de especificación de un producto, regulaciones o estandarizaciones de una empresa, o simplemente de manera subjetiva. Si es necesario determinar las cotas con base a un rango físico de las respuestas, es razonable considerar los mínimos y máximos de las respuestas individuales estimadas, es decir $\left(Y_j^{\text{mín}} = \underset{x \in R}{\text{mín}}[\widehat{Y}_j(x)],\ Y_j^{\text{máx}} = \underset{x \in R}{\text{máx}}[\widehat{Y}_j(x)]\right)$.

La función $u_j(\widehat{Y}_j(x))$ depende de las condiciones del proceso y por lo que también se puede desear minimizar la respuesta, en el primer caso $M_j = Y_j^-$ se sustituye en la tercera función en la expresión

(3). Si se desea maximizar $M_j = Y_j^+$ va en la segunda función en (3).

En este resumen sólo se indicarán el nombre de los cinco métodos de optimización utilizados, la descripción de los métodos aparece en Domínguez (2007). La idea general de varios de los métodos de optimización está en función de la expresión (3), estos son: la función de deseabilidad FD. Otro procedimiento consiste en optimizar de manera simultánea el modelo (3) y con ello obtener un grado de satisfacción λ, GS. El tercer procedimiento corresponde a una aplicación de la lógica difusa y se denotó por PD. El cuarto óptimo se obtiene utilizando la función distancia generalizada, DI. Finalmente se realizó la optimización para el planteamiento topsis, TO.

4. Caso de estudio

El caso de estudio trata de la elaboración de un queso, Schmidt et. al. (1979) y se desea conocer la combinación de los efectos de la cistina (cuajo): X_1 y el clorido de calcio: X_2 en la texturización y en las características de agua-caliente dializada en una concentración de proteína de suero en un gel. En este proceso experimental se aplicó un diseño central compuesto, en este diseño cada factor X tiene cinco valores. Las características de la textura son medidas por la dureza: Y_1, cohesividad: (coherencia) Y_2, elasticidad: Y_3, y el agua comprimible: Y_4. El objetivo es obtener máximos simultáneos para las cuatro variables.

En la Tabla 1 se describe en las columnas correspondientes a x_1 y x_2 el diseño central compuesto para dos factores y en este caso $n_o = 5$, en las últimas cuatro columas se muestran los valores de las cuatro respuestas para cada uno de los tratamientos.

La matriz $Z(x)$ para este ejemplo corresponde a los $n = 13$ renglones correspondientes a los tratamientos y a las columnas de la 2 a la 7, es decir $q = 6$, $(q = 2(2+3)/2+1)$. Los resultados de

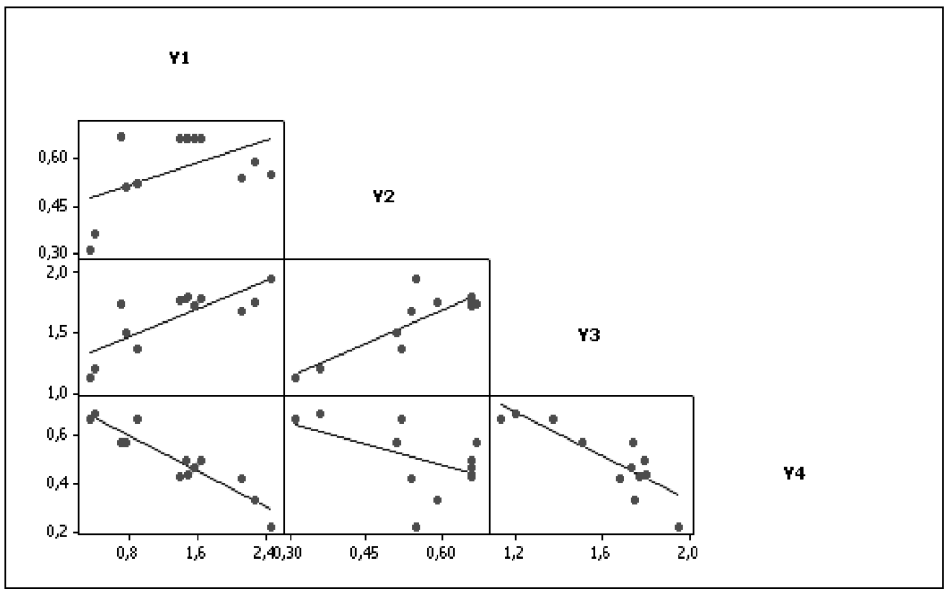


Figura 1: Matriz que describe la relación por pares entre las cuatro variables de respuesta.

mínimos cuadrados para cada uno de los 4 modelos se obtienen por: $\widehat{\delta}_j = [(Z(x))^t Z(x)]^{-1}(Z(x))^t Y_j$.

Tabla 1. Esquema experimental y las respuestas en cada tratamiento

Tratamiento	X_1	X_2	cte.	x_1	x_2	x_1x_2	x_1^2	x_2^2	Y_1	Y_2	Y_3	Y_4
1	8.0	6.5	1	-1	-1	1	1	1	2,48	0,55	1,95	0,22
2	34	6.5	1	1	-1	-1	1	1	0,91	0,52	1,37	0.67
3	8.0	25.9	1	-1	1	-1	1	1	,71	0,67	1,74	0,57
4	34	25.9	1	1	1	1	1	1	,41	0,36	1,20	0,69
5	2.6	16.2	1	$-\sqrt{2}$	0	0	2	0	2,28	0,59	1,75	0,33
6	39.4	16.2	1	$\sqrt{2}$	0	0	2	0	0,35	0,31	1,13	0,67
7	21	2.3	1	0	$-\sqrt{2}$	0	0	2	2,14	0,54	1,68	0,42
8	21	29.9	1	0	$\sqrt{2}$	0	0	2	0,78	0,51	1,51	0,57
9	21	16.2	1	0	0	0	0	0	1,50	0,66	1,80	0,44
10	21	16.2	1	0	0	0	0	0	1,66	0,66	1,79	0,50
11	21	16.2	1	0	0	0	0	0	1,48	0,66	1,79	0,50
12	21	16.2	1	0	0	0	0	0	1,41	0,66	1,77	0,43
13	21	16.2	1	0	0	0	0	0	1,58	0,66	1,73	0,47

Relación entre las respuestas

En esta parte se presenta la descripción gráfica de los modelos para evaluar si existe alguna relación entre ellos, a continuación se muestran los modelos ajustados por mínimos cuadrados y sus óptimos individuales. En la Figura 1 se muestra la relación entre las cuatro variables de respuesta.

Se puede notar en la Figura 2 que existe una consistente relación inversa entre las variables de

respuesta Y_1 y Y_4 y entre Y_3 y Y_4, y más leve esa tendencia entre Y_2 y Y_4, en ese sentido cuando se desea maximizar las respuestas Y_1, Y_3 y Y_2 entonces Y_4 disminuirá más en el caso de Y_1 y Y_3. Ante esta situación se tiene que valorar la importancia de cada respuesta en el proceso real para ver qué conviene más, si sacrificar la respuesta Y_4 o encontrar un equilibrio entre las cuatro respuestas. Este escenario se considerará en el procedimiento de optimización simultánea. En la fase de maximizar las respuestas Y_1, Y_2 y Y_3 no hay mayor complicación porque existe una relación directa entre ellas, sin embargo, es importante considerar la eficiencia de los métodos de optimización ante la correlación de las variables de respuesta. Los planteamientos de optimización consideran la importancia de las funciones mediante pesos w_j.

En la Figura 2, se sobreponen las cuatro variables de respuesta, en el punto se encuentra un punto común para las cuatro respuestas que se aproxima a ser un posible óptimo global, $\bullet$ $(x_1, x_2) = (0{,}33, -1{,}40)$. Como puede observarse en la gráfica se generan varias regiones de soluciones óptimas posibles.

Se obtuvieron los valores mínimos $\widehat{Y}_j^{\text{mín}}$ y máximos $\widehat{Y}_j^{\text{máx}}$ individuales de cada respuesta ajustada, estos son:

Respuesta	$\widehat{Y}_1$	$\widehat{Y}_2$	$\widehat{Y}_3$	$\widehat{Y}_4$
$\widehat{Y}_j^{\text{mín}}$	0,37	0,33	1,11	0,23
$\widehat{Y}_j^{\text{máx}}$	2,68	0,66	1,88	0,71

Estos valores se utilizaron como cota para cada respuesta y como se desea maximizar el valor $\widehat{Y}_j^{\text{máx}} (j = 1, 2, 3, 4)$ se utilizó como valor objetivo.

Los resultados de optimización que generó cada método se presentan en la Tabla 2, los j resultados alcanzados en cada una de estas cinco funciones se podrían usar para evaluar las otras cuatro, de tal manera de llevar a cabo una comparación de la eficiencia entre ellos. Sin embargo se empleó la función de pérdida estandarizada para hacer una evaluación de la eficiencia de los métodos, así el que tenga la menor función de pérdida será considerado como el mejor método. Se define la función de pérdida estandarizada como:

$$P = \sqrt{\textstyle\sum_{j=1}^{r} (Z_j - Z_j^{\text{máx}})^2}, \text{ donde } Z_j = \widehat{Y}_j(x_o)/\widehat{Y}_j^{\text{máx}}, Z_j^{\text{máx}} = 1.$$

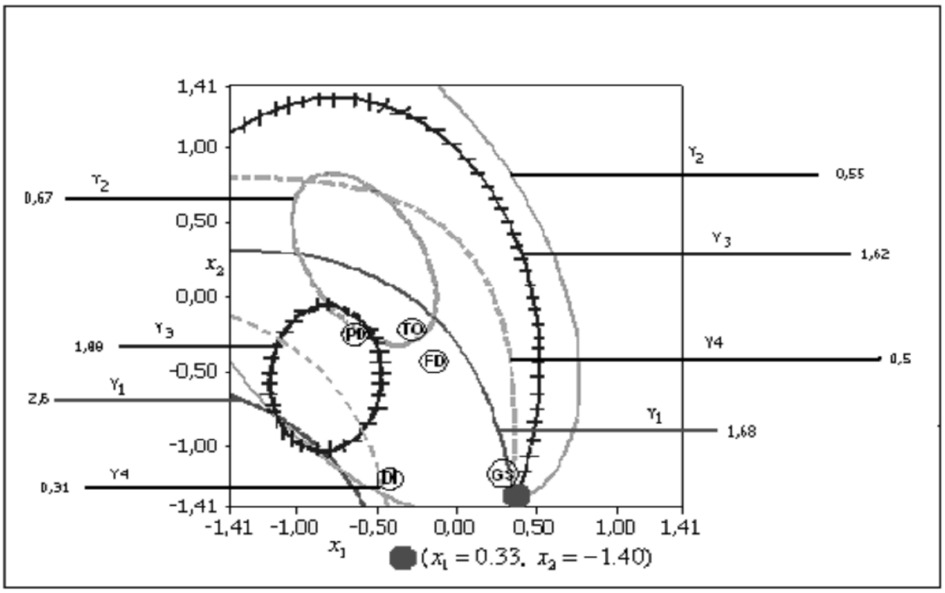


Figura 2: Ubicación de las funciones simples en la gráfica de superposición de las cuatro respuestas.

Tabla 2. Resultados de los diferentes procesos de optimización

	Óptimo		Valores				P
Función	x_{1o}	x_{2o}	$\widehat{Y}_1(x_o)$	$\widehat{Y}_2(x_o)$	$\widehat{Y}_3(x_o)$	$\widehat{Y}_4(x_o)$	
FD	-0.10	-0.41	1.79	0.66	1.81	0.43	0.517
GS	0.35	-1.37	1.68	0.55	1.62	0.50	0.523
PD	-0.53	-0.19	1.91	0.67	1.88	0.38	0.545
DI	-0.46	-1.38	2.47	0.54	1.83	0.31	0.598
TO	-0.24	-0.21	1.78	0.67	1.84	0.42	0.529

5. Discusión y Conclusiones

El planteamiento de optimización que obtuvo el mejor óptimo común para las cuatro respuestas corresponde a la funcion de deseabilidad, y como se esperaba dada la estrecha relación entre las variables (Figura 2), se nota que algunos procedimientos alcanzaron un mejor máximo en la respuesta $\widehat{Y}_1$, pero salieron afectados en la respuesta $\widehat{Y}_4$ como es, principalmente el caso de la función distancia DI.

Se observa que el planteamiento GS prácticamente coincide con el óptimo generado de manera gráfica y como se observa es una solución bastante adecuada. Será una actividad interesante aplicar estos planteamientos a varios ejemplos para evaluar su precisión en el proceso de optimización, desde luego

considerando situaciones donde no exista una fuerte relación entre las variables. También pueden llevarse a cabo estrategias de simulación.

6. Referencias

Box, G. E. P. and Draper, N. R. (1987). Empirical Model Building and Response Surfaces. John Wiley & Sons, New York, NY.

Domínguez, D.J. (2007). Optimización Estadística: Multi-respuesta, sometido a Comunicaciones Internas del CIMAT.

Khuri, A. I. and Cornell, J A. (1996). Response Surface, Designs and Analysis. Marcel Dekker, Inc, New York.

Myers, R. and Montgomery, D.C. (2002). Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Wiley Series in Probability and Statistics. New York.

Schmidt, R.H., Illingworth, B.L., Deng, J.C., and Cornell, J.A. (1979). Multiple Regression and Response Surface Analysis of the Effects of Calcium Chloride and Cystine on Heat-Induced Whey Protein Gelation, J Agric. Food Chem., 27, pp 529-532.

Simulación de un proceso de manufactura en un contexto seis sigma

Fernando Valenzuela Camacho[1]
Universidad de las Américas, Puebla
Jorge Domínguez Domínguez[2]
Centro de Investigación en Matemáticas
Antonio González Fragoso [3]
Universidad de las Américas, Puebla

1. Introducción

La finalidad de este trabajo es ilustrar, mediante un procedimiento de simulación, la aplicación de la metodología seis sigma a un proceso de manufactura. Este último es representado por un artefacto denominado helicóptero, tal que al soltarlo desde una altura determinada éste debe tardarse el mayor tiempo en caer y girar en su trayectoria de manera vistosa. Para alcanzar esta meta el artefacto depende de una serie de factores que influyen en su eficiencia. Esta situación permite considerar el desarrollo del helicóptero como un proceso similar al que se realiza en la industria al manufacturar un producto.

En resumen, el desarrollo de este prototipo, helicóptero, permitirá mostrar cómo se realizan proyectos en la metodología seis sigma. Además en este marco se hará énfasis en dos técnicas estadísticas: la función despliegue de la calidad (QFD, siglas en inglés) y del diseño de experimentos (DoE), Breyfogle III (2002).

Seis Sigma es un programa que adopta una empresa en el que se considera determinante la opinión del cliente Pande y Holpp (2001). En su desarrollo se invierte la mayor parte del tiempo en la planeación de la mejora de los procesos que presentan variación considerable; es decir, se planea y se mide de manera óptima a los causantes de variación en los procesos con el fin de mejorarlos y así reducir el tiempo invertido en el desarrollo/planeación, los desperdicios generados en el desempeño, y finalmente, aumentar las ganancias. Todo esto bajo la aplicación de varias herramientas estadísticas

[1]nandov@gmail.com
[2]jorge@cimat.mx
[3]antonio.gonzalez@udlap.mx

y la participación de todo el personal de la empresa. Las etapas de la metodología Seis Sigma son: Definir, Medir, Analizar, Mejorar y Controlar, Breyfogle III (2002).
QFD es una metodología que captura exactamente lo que el cliente requiere y transforma esos datos en requerimientos de producción para la empresa. Sus objetivos son: a) el mantener la voz explícita y no explícita del cliente de forma íntegra durante todo el proceso a fin de evitar que se realicen procesos que no vayan bien enfocados; y b) transformar requerimientos muy generales (e.g. mayor rapidez, mayor comodidad) en acciones fácilmente identificables y realizables por los ingenieros de la empresa. El elemento principal de la metodología QFD es la Matriz de la Calidad, la cual es un mapa conceptual que relaciona los requerimientos y necesidades que demanda el cliente, con las características técnicas que la empresa debe tomar en cuenta para satisfacerlos.
QFD y DoE trabajan juntos de la siguiente manera: QFD se encarga inicialmente de tomar las variables que se quieren medir, lo correspondiente a los "qué"de la casa de la calidad; sus respectivos "cómo"se encuentran divididos en dos: a)Factores de Control y b)Factores de Ruido. Una vez que queda hecha la casa de la calidad, DoE toma cada uno de los "qué"de la Casa de la Calidad como sus Factores de Control para desarrollar el experimento. Cuando DoE determina los factores e interacciones principales, éstos van a ser usados en un nuevo diagrama QFD y se eliminan los no trascendentes. Por último, se vuelve a hacer el análisis de DoE para comprobar que todos los factores e interacciones son importantes, en caso contrario se repite el proceso.

2. Simulación de un proceso de manufactura

En una empresa ficticia se producen helicópteros de papel (HP), los cuales cuando se dejan caer de una cierta altura comienzan a rotar de una manera "vistosa", asemejándose a un helicóptero. El propósito del HP es ser un juguete para niños. Por medio de un proceso Seis Sigma, se mejorará el tiempo que los HP tardan en caer, y la forma en que cae el juguete. La única restricción es que no se cambió el diseño general de los HP. Se considerará como mejora el aumentar el tiempo promedio de caída de los HPs manteniendo lo vistoso de ésta, es por esta razón que este proyecto de mejora es un proyecto de desarrollo tecnológico. De manera muy resumida se describirán los aspectos relevantes de la metodología seis sigma Pyzdek (1999).

Definir

Esta etapa constituye una parte esencial en la elaboración de un proyecto seis sigma, como elementos básicos considera el caso del negocio ya que al final dirá el impacto económico que alcanzará la empresa al finalizar el proyecto y determinará el éxito del proyecto. También mide la métrica seis

sigma que se alcanza. En esta parte se elabora una carta conocida como Team Charter, la cual es una hoja que describe primordialmente el problema: caso de negocio, las metas a alcanzar, las tareas en cada una de las siguientes etapas, el cronograma, el personal que compone el equipo y el papel que estos desempeñarán a lo largo del proceso. Esta hoja la supervisa constantemente un responsable, que en la metodología seis sigma se denomina Champion, éste representa a la gerencia de la empresa. Por cuestiones de espacio en este trabajo no se presentará la carta, a continuación únicamente se definirán las características del helicóptero.

El producto con el que se trabaja en este proceso Seis Sigma es una hoja de papel que cortada y doblada de cierta manera, produce un objeto que se asemeja a un helicóptero, el cual cuando se deja caer de una cierta altura, lo hace rotando lentamente hasta tocar el suelo. Las especificaciones de las medidas antes de la mejora son las siguientes: el tamaño de la hoja es de $12 \times 10 cm^2$; dentro de la hoja se tienen delimitados dos espacios de $8 \times 5 cm^2$ que se le llaman "alas largas", dos espacios de $3 \times 3 cm^2$ que se llaman "alas cortas"y un espacio que divide los dos tipos de alas de $1 \times 10 cm^2$ llamado "cinturón".

Medir

Para este proceso de la elaboración de Helicópteros de Papel, como fue visto anteriormente, existen dos características que son críticas para la calidad y son: a) El HP debe tardar en promedio el mayor tiempo posible en el aire (se debe tardar en caer); y b) La caída del HP debe ser vistosa (debe dar vueltas al caer y no debe desviarse en su trayecto). Estas dos características se manejaron de forma separada. En la primera se recolectaron datos de tipo continuo ya que se midió el tiempo de caída de los HP. En la segunda característica crítica para la calidad, los datos recolectados fueron de forma nominal.
Se realizaron dos estudios Gage R&R, se seleccionaron a tres operadores en forma aleatoria, cada uno mediría el tiempo de caída de 10 diferentes Helicópteros de Papel elegidos al azar y también determinaría si la caida fue vistosa o no.
Con respecto al tiempo de caída de los HP (Helicópteros de Papel), el porcentaje de variación debido a repetibilidad es de $3.77\,\%$, mientras que el de reproducibilidad es de $2.12\,\%$; es por esto que la variabilidad debida a la diferencia de partes constituye el $94.11\,\%$ de la variabilidad total. Conforme a lo vistoso de la caída, los resultados de congruencia entre operadores muestran que los operadores tuvieron una congruencia de por lo menos $90\,\%$ consigo mismos y con respecto al estándar.
Se realizaron Cartas de Control de tipo $\overline{x}$ y S para tiempo de caída de los HP y una carta de tipo P para lo vistoso de la caída. Ninguna de estas cartas reportó que hubiera problemas mayores que impidan continuar con el proyecto.

Analizar

Se realizó un estudio de capacidad de proceso y se determinó que al obtener un índice C_{pk} de 0.54, el nivel sigma antes de la mejora es de 0.7. La Figura 1 muestra la situación antes del proceso.

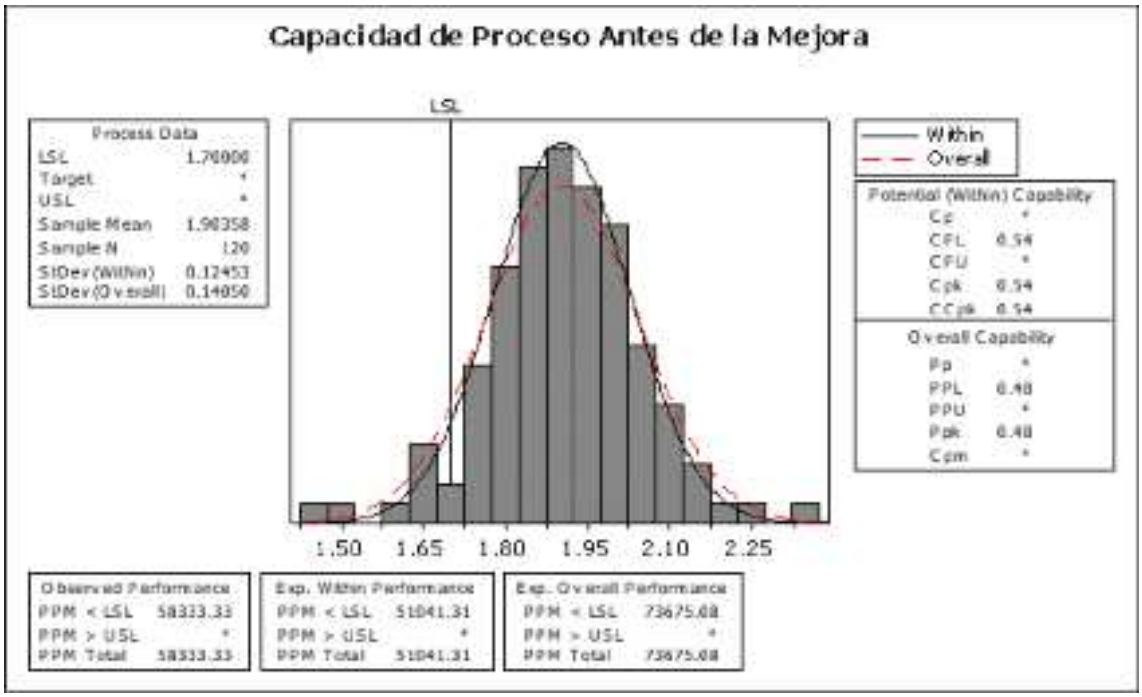


Figura 1: Capacidad del Proceso Antes de la Mejora.

Mejorar

Se comenzó determinando los factores de control y eligiendo sus niveles. Los factores de control y sus niveles son: Largo de alas (niveles: $8cm$, $4cm$); Ancho del HP (niveles: $5cm$, $11cm$); Doblez de la Base (niveles: $1.5cm$, $3cm$); Ancho del Cinturón (niveles: $0.3cm$, $2cm$); Tipo de Papel (Bond, Micro, Albanene Delgado, Mantequilla Delgado); Aditamento (niveles: Clip, Sin Clip). Después, se elaboró el primer diagrama QFD que muestra de forma ordenada los factores de control y ruido, así como los requerimientos del cliente.

El objetivo del primer experimento fue eliminar dos tipos de papel. Se realizaron Cuatro Diseños Factoriales Completos 2^5, con un tipo de papel y tres replicaciones cada uno. El resultado de los cuatro experimentos fue que se eliminan los tipos de papel Mantequilla Delgado y Albanene Delgado. En estos experimentos también se observó que el utilizar un clip en los HPs reduce considerablemente el tiempo de caída, aunque la hace más atractiva. Se realizaron varios estudios rápidos y se determinó que se fija el factor de Grapa como aditamento de ahora en adelante. Otras modificaciones

que se hicieron fueron el medir Ancho del Ala del HP en lugar de medir Ancho Total del HP, y en lugar de medir Doblez de la Base del HP, dividirlo en mediciones de Ancho de Cuerpo y Ala del Cuerpo. Al término de este primer experimento, se actualiza el diagrama QFD con la información obtenida.
El segundo experimento realizado fue un Escalamiento Ascendente para papel Micro y papel Bond; se realizaron Dos Diseños Factoriales Fraccionados 2^{5-2} de cinco replicaciones cada uno. Los resultados del experimento son que aunque el papel Micro presenta tiempos de caída menores a los presentados por el papel Bond, resulta ser muy inestable, por lo que se decide eliminar el papel Micro del experimento. El papel Bond utilizado hasta el momento es de $75gr/m^2$, se propuso probar un papel Bond más ligero, por lo que se hicieron experimentos rápidos con papel Bond de $58gr/m^2$ (al cual se le llamará papel Bond Ligero); el resultado es que el papel Bond Ligero presenta tiempos de caída superiores a los presentados por el papel Bond de $75gr/m^2$, sin perder lo vistoso. Se vuelve a actualizar el diagrama QFD con la información obtenida.
A continuación se realizó un Escalamiento Ascendente para Papel Bond Ligero. Los diseños utilizados fueron Factorial Fraccionado 2^{5-2} y Factorial Completo 2^3 entre otros experimentos menores. El resultado es que se obtuvo la primera propuesta a diseño final del HP. Finalmente, se llevó a cabo un Diseño Central Compuesto 2^5. Durante el análisis de los resultados de este DCC 2^5, se encontró un diseño de HP que presenta un mejor tiempo de caída, y mayor estabilidad; este HP encontrado es el diseño final, sus especificaciones son:

Largo de Ala: $12cm$	Ancho de Ala: $3cm$	Cinturón: $1cm$
Largo de Cuerpo: $2.5cm$	Ancho de Cuerpo: $4.5cm$	Papel: Bond de $58gr/m^2$
Aditamento: Grapa		

Controlar

Ahora que ya se logró la mejora, se realiza de nuevo un estudio de Capacidad de Proceso, para determinar la magnitud de la mejora. Antes de la mejora, el índice C_{pk} era de 0.54 y el tiempo promedio de caída era de 1.9 segundos; ahora, el índice C_{pk} es de 7.29 y el tiempo de caída promedio es de 3.68 segundos, lo que nos asegura que se logró la capacidad Seis Sigma. En la Figura 2 se muestra la Capacidad de Proceso después de la mejora.

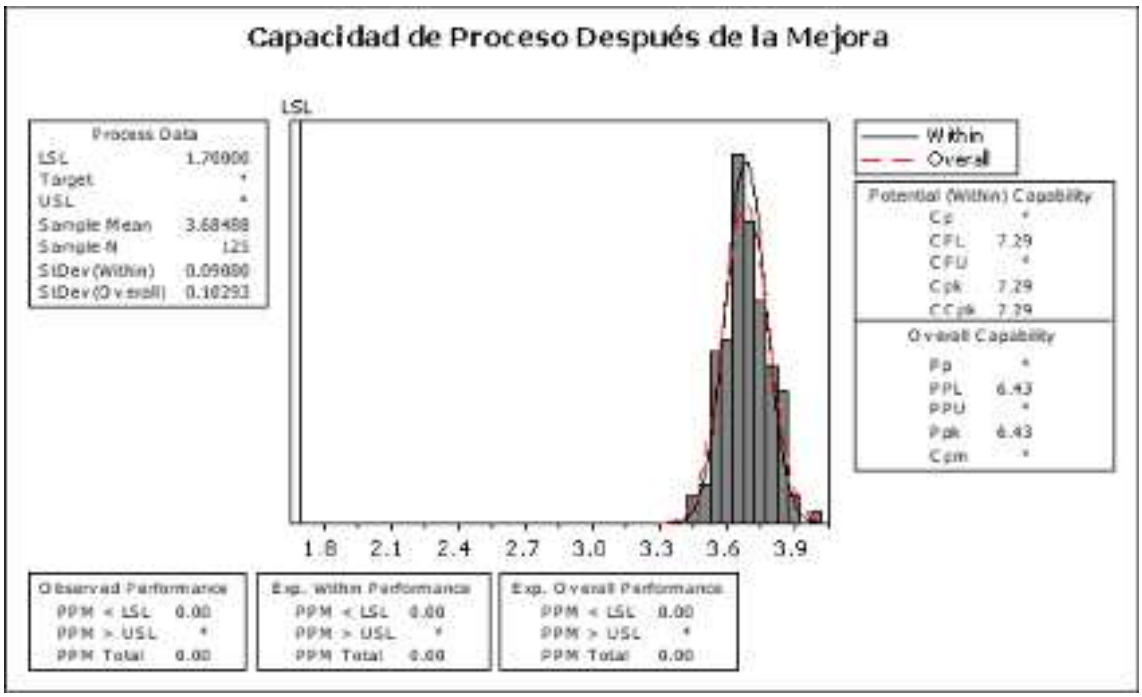


Figura 2: Capacidad del Proceso Después de la Mejora.

3. Conclusión

En esta breve exposición se puede observar que al aplicar de manera eficiente las etapas de la metodología seis sigma, se ha podido mejorar el desempeño del tiempo de caída del helicóptero y se ha aumentado su vistosidad. El desarrollo de este proyecto nos permite generar material educativo para explicar como se puede realizar un QFD y llevar a cabo un experimento. Los usuarios que reproduzcan este proyecto podrán adquirir experiencia en el manejo de procesos y lo más importante es que tendrán una guía en la aplicación de la metodología seis sigma.

Referencias

1. Breyfogle III FW. (2002). Implementing Six Sigma: Smarter Solutions Using Statistical Methods. Wiley: New York

2. Pande P. y Holpp I.. (2001). What is Six Sigma? McGaw-Hill: Maidenhead, Berkshire.

3. Pyzdek T. (1999). The Complete Guide to Six Sigma. Quality Press: Milwukee, WI.

Construcción de una escala clínica - ultrasonográfica para el diagnóstico de coledocolistiasis[1]

Ana Bertha Irineo Cabrales[2]
Universidad Autónoma de Sinaloa
Carlos Zambada-Sentíes
Universidad Autónoma de Sinaloa
Felipe Peraza
Universidad Autónoma de Sinaloa

1. Introducción

Para el diagnóstico de colestasis concurren hallazgos clínicos, paraclínicos y morfológicos definidos, entre los cuales la ictericia es el signo cardinal, y aunque diagnosticarla es muy simple, aclarar su causa puede ser un desafío. Machnik G. (1993) y Lucas WB, Chutan R. (1995) establecieron que es de particular importancia la identificación temprana de pacientes con ictericia secundaria a obstrucción extrahepática o quienes son referidos erróneamente con diagnostico de ictericias de resolución quirúrgica. Martin WB (1960) clasificó a la colestasis en Intrahepática y extrahepática. En donde la causa más común de colestasis extrahepática es la coledocolitiasis (Obstrucción aguda del conducto biliar generalmente por litos provenientes de la vesícula biliar). La ictericia es un signo alarmante que obliga al paciente, en la mayoría de los casos, a acudir con el médico y sirve como clave o signo guía para el diagnostico clínico. El abordaje del paciente con ictericia colestásica requiere una evaluación clínica cuidadosa que incluye interrogatorio, examen físico y pruebas de laboratorio que orientan la selección apropiada de investigaciones adicionales para diferenciar entre las numerosas causas de esta patología. Machnik G. (1993) y Lucas WB, Chutan R. (1995) concluyen que desafortunadamente, la clínica y la presentación bioquímica con frecuencia son inespecíficas e indistinguibles de otras alteraciones hepatocelulares, por ello, Stern RB, Knill-Jones RP, Williams R. (1974) comentan que basándose en tales datos, fácilmente disponibles, sólo los clínicos experimentados pueden diferenciar entre enfermedad hepática parenquimatosa y obstrucción extrahepática del conducto biliar en 80 a 90 % de los pacientes y Shea JA (1994), concluye que identifica su causa sólo en 45 a 65 % de ellos. Deitch EA (1981) en su trabajo comprobó que el ultrasonido es la técnica de elección para

[1]Trabajo realizado con apoyos de la CGIP-UAS
[2]drairineo@hotmail.com

detectar cálculos vesiculares, con una sensibilidad de 84 % (IC 95 % 76 a 98), una especificidad de 99 %(IC 95 % 97 a 100) y una exactitud diagnóstica que varía de 90 a 95 %. Sin embargo, Gross BH (1983) documentó que la sensibilidad para detectar cálculos en el conducto biliar común muestra una gran variabilidad, con un rango de 25 a 90 %. La enfermedad por cálculos es con frecuencia un proceso obstructivo intermitente y, por este fenómeno, en 24 a 36 % de los pacientes no hay dilatación biliar, lo que dificulta aún más el diagnostico. El rápido desarrollo de la tecnología médica, que induce a nuevas y casi siempre costosas opciones diagnósticas y terapéuticas, coexiste con recursos financieros cada vez más escasos de los sistemas de salud. La asignación óptima de estos recursos permite un uso racional, en beneficio del paciente. Este fenómeno fortalece el interés en pruebas diagnósticas no invasivas de bajo costo para enfermedades tales como la coledocolitiasis. El propósito de este trabajo es describir la construcción de una escala clínica-ultrasonográfica basada en datos universales y fácilmente disponibles, de uso fácil y practico, con capacidad para realizar un diagnóstico correcto en 90 % o más de los casos de colestasis extrahepática aguda.

2. Material y Métodos

Se realizó un diseño de pruebas diagnósticas. Los casos se obtuvieron de tres hospitales que atienden a población derechohabiente de la ciudad de Culiacán, Sinaloa, México. Se revisaron los registros de ingresos de octubre de 1998 a septiembre de 1999. Se incluyeron los datos de los expedientes de pacientes adultos con diagnósticos relacionados con la presencia de ictericia: Fueron 94 pacientes adultos que cumplían con los criterios de inclusión: edad $>$ 18 años, de uno u otro sexo, ictéricos con bilirrubina total sérica $>$ 2 mg/dl, con ultrasonido de vesícula, vías biliares y páncreas, con diagnóstico concluyente por colangiopancreatografía retrógrada endoscópica (CPRE) o cirugía. Se consideró coledocolitiasis la extracción de litos del colédoco por CPRE o cirugía. Inicialmente se realizó un análisis univariado para identificar variables que estuvieran asociadas con la enfermedad. En una segunda fase se hizo un análisis de componentes principales. Una vez construida la escala se calcularon sensibilidad y especificidad de cada punto de corte y se construyó una curva ROC estimando el área bajo la curva.

3. Resultados

El grupo I, con coledocolitiasis, estuvo constituido por 57 pacientes, con un promedio de edad de 56 años (rango 18-94); 45 fueron mujeres (79 %), y 12 hombres ($p< 0.05$). En el grupo II, sin coledocolitiasis, hubo un total de 37 pacientes, con 17 (46 %) mujeres y 20 hombres ($p< 0.05$), la edad promedio fue de 57 años (rango 19 a 84). La tabla 1 muestra la asociación de las variables independientes con la presencia de la enfermedad, coledocolitiasis, siendo algunas de ellas estadísticamente significativas ($p<0.05$). Del análisis de componentes principales se extrajeron dos componentes, según muestran sus valores de varianza total explicada. Donde el componente 1 explica el 69.10 % de la varianza total y el factor 2 de una importancia considerablemente menor en su capacidad de explicar la varianza observada (16.08 %). Usamos al componente 1 como una escala diagnostico de coledocolitiasis, que se caracteriza como dolor, ictericia, fiebre, tipo de dolor y ultrasonido. El componente 2 sólo es positivo a la fiebre. De acuerdo a las cargas del primer componente, la escala diagnóstica queda de la siguiente manera:

0.960* Inicia su padecimiento con dolor en epigastrio

y/o hipocondrio derecho postprandial hasta de 8 horas

+ 0.964*Instalación de la Ictericia posterior al dolor

+ 0.556*Afebril o en caso de fiebre aparece simultánea o posterior al dolor

+ 0.872*dolor tipo cólico

+ 0.732*Ultrasonido: cualquier opción positiva

a) Litiasis vesicular con colédoco > 5mm de diámetro

b) Litiasis vesicular con colédoco <5 mm

c) Litiasis vesicular con lito visible en colédoco con o sin dilatación de colédoco

d) Sin Litiasis vesicular con lito visible en colédoco con o sin dilatación de colédoco. Se estimaron sensibilidad y especificidad para cada punto de corte, siendo el valor de 3.1 el mejor punto de corte

con una sensibilidad del 98 % y especificidad del 91.8 % con una prevalencia de la enfermedad del 60 %. La exactitud predictiva del modelo fue de 98 % (IC 95 % 90,99 %) lo que corresponde al área bajo la curva ROC.

4. Referencias

Berkowitz D.(1964) Pitfall in the differential diagnosis of jaundice. Am J Gastroenterol 41:488-98.

Conn HO, Blei AT, Chojkeir M, Schade R, Taggart GJ, Atterbury CE (1979). The naked physician: The blind interpretation of liver function tests in the differential diagnosis of jaundice. En: Preising R, Bircher J, Ed. The liver. Quantitative aspects of structure and function. Aulendorf: Editio Cantor, 386-94.

Cronan JJ, Mueller PR, Simeone JF, O'Connell RS, vanSonnenberg E (1983). Prospective diagnosis of choledocholithiasis. Radiology 146:567-9.

Deitch EA (1981). In vivo measurements of the internal and external diameters of the common bile duct in man. Surg Gynecol Obstet. 152:642-5.

Dong B, Chen M. (1987) Improved sonographic visualization of choledocholithiasis. J Clin Ultrasound 15:185-90.

Gross BH, Harter LP, Gore RM, Callen PW, Filly RA, Shapiro HAet (1983). Ultrasonic evaluation of common bile duct stones: Prospective comparison with endoscopic retrograde cholangiopancrestography. Radiology;146:471-4.

Haubek A, Pedersen JH, Burcharth F, Gammel-Gard J, Hancke S, Willumsen L. (1981) Dynamic sonography in the evaluation of jaundice. Am J Radiol. 136:1071-4.

Laing FC, Jeffery RB (1983). Choledocholithiasis and cystic duct obstruction: Difficult ultrasonographic diagnosis. Radiology 146:475-9.

Laing FC, Jeffery RBJ, Wing VW (1984). Improved visualization of choledocholithiasis by sonography. ARJ Am J Roentgenol.143:949-52.

Lucas WB, Chuttani R. (1995) Pathophysiology and current concepts in the diagnosis obstructive jaundice. Gastroenterologist 3:105-118.

Machnik G. (1993) Histological changes in liver tissue in cholestasis. Z Gastroenterol. 31:7-10.

Martin WB, Apostolakos PC, Roazen H. (1960). Clinical versus actuarial prediction in the differential diagnosis of jaundice. Am J Med Sci 240:571-8.

Schenker S, Balint J, Schiff L. (1962) Differential diagnosis of jaundice: Report of a prospective study of 61 proved cases. Am J Dig Dis 7:449-63.

Shea JA, Berlín JA, Escarce JJ. (1994) Revised estimates of diagnostic test sensitivity and specificity in suspected biliary tract disease. Arch Intern Med ;154:2573-81.

Stern RB, Knill-Jones RP, Williams R (1974). Clinician versus computer in the choice of 11 differential diagnoses of jaundice based on formalised data. Methods Inf Med ;13:79-82.

Modelación no estocástica

José Elías Rodríguez Muñoz[1]

Facultad de Matemáticas, Universidad de Guanajuato

1. Introducción

En muestreo de poblaciones finitas cuando se desea estimar un parámetro poblacional con base a información auxiliar casi siempre recurrimos a un modelo que relaciona la variable de interés con las variables auxiliares. Además si se emplea un enfoque para la inferencia basado en modelos o asistido por modelos, Särndal et al. (1992), esto llevará, en ambos enfoques, a incluir un elemento estocástico en la relación de las variables.

Por ejemplo, si se considera para cada unidad poblacional una relación lineal:

$$y = \beta_0 + \beta_1 x + \sigma_k \varepsilon,$$

entonces se supone al menos que las variables aleatorias ε tienen valor esperado cero y una cierta estructura regular de correlación.

Este tipo de supuestos (sobre la parte estocástica del modelo) es después difícil de justificar y verificar en diseños de muestreo que, por ejemplo, se llevan en dos o más etapas y las últimas unidades de muestreo son conglomerados.

Una posible alternativa al problema anterior es plantear un modelo no estocástico y posteriormente la inferencia basarla en el diseño de muestreo. El presente trabajo pretende exponer, a través de ejemplos, dicha estrategia de modelación y análisis en muestreo de poblaciones finitas.

Cabe aclarar que en la presente exposición la población es representada por U, los valores de la variable de interés por $y_1, \ldots, y_N$ y el parámetro de interés es el total $t_y = \sum_{k \in U} y_k$.

[1] elias@cimat.mx

2. Ejemplos

Ejemplo 1: Estimador de Regresión Logística

Supóngase que el problema a resolver es estimar el total de individuos en una subpoblación (total de desempleados, discapacitados, asistentes al casino de la Feria de San Marcos, etc.). Entonces la variable de interés tendrá dos posibles valores: 1 si pertenece a la subpoblación ó 0 si no pertenece. Además supongamos que tenemos variables auxiliares $x_k = (1\ x_{k2} \cdots x_{kr})$, cuyos valores se conocen para cada elemento de la población y se pueden utilizar para resolver dicho problema.

En este ejemplo se utilizará el siguiente modelo operacional (*working model*):

$$y_k = G\left(x_k\beta\right) + \varepsilon_k, \quad \text{(modelo no estocástico)} \tag{1}$$

donde G es una función de distribución, β es un vector de parámetros y los residuos ε_k (valores desconocidos y no aleatorios) absorben lo no explicado de y_k por parte de las variables auxiliares; por $G\left(x_k\beta\right)$.

Un problema aparente en primera instancia con el presente modelo es que existen muchos vectores de parámetros β y respectivos vectores ε con los cuales se puede describir a la variable de interés. Esta situación se ilustra en la gráfica 1.

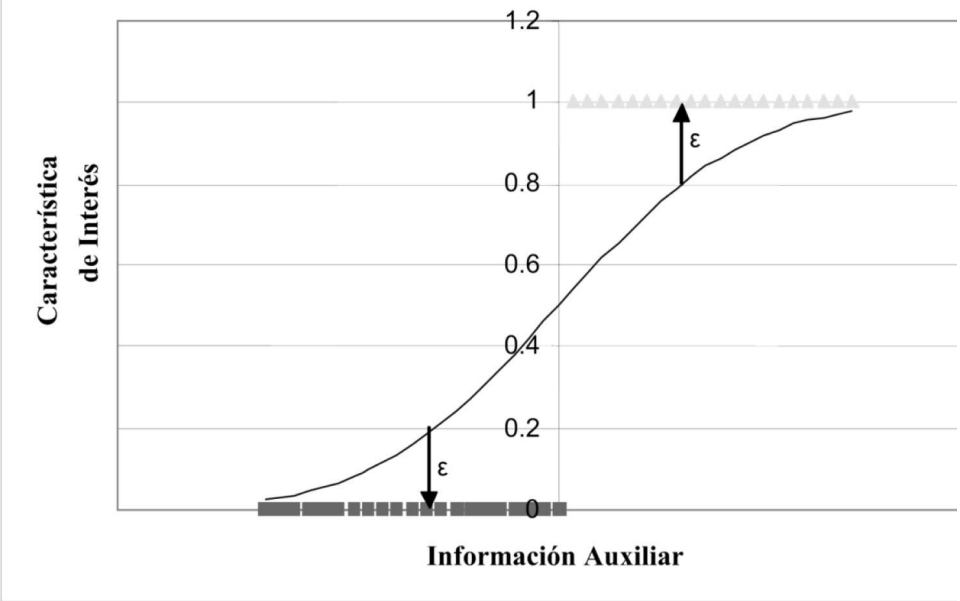
Para este problema aparente se propone utilizar los parámetros que minimicen la media geométrica de los valores absolutos de los residuos. En otros términos, se utilizarán los parámetros tales que

$$\sqrt[N]{\prod_{k\in U} |\varepsilon_k|} = \sqrt[N]{\prod_{k\in U} G\left(x_k\beta\right)^{1-y_k} \left(1 - G\left(x_k\beta\right)\right)^{y_k}}$$

sea mínimo.

Lo anterior es equivalente a maximizar:

$$MG\left(|\varepsilon|\right) = \sqrt[N]{\prod_{k\in U} G\left(x_k\beta\right)^{y_k} \left(1 - G\left(x_k\beta\right)\right)^{1-y_k}}$$



Grafica 1. Una posible distribución G para el modelo (1)

o bien maximizar

$$\ln\left[MG\left(|\varepsilon|\right)\right] = \frac{1}{N}\sum_{k\in U}\left\{y_k \ln\left[G\left(x_k\beta\right)\right] + \left(1-y_k\right)\ln\left[1-G\left(x_k\beta\right)\right]\right\}. \tag{2}$$

Si G es la función de distribución logística, entonces maximizar (2) es equivalente a resolver las siguientes ecuaciones: $\sum\limits_{k\in U}\left\{y_k - G\left(x_k\beta^{MG}\right)\right\}x_k^T = 0$.

Observemos que de la forma de $x_k = (1 \quad x_{k2}\cdots x_{kr})$ y de la primera de las ecuaciones a resolver anteriores: $\sum\limits_{k\in U}\left\{y_k - G\left(x_k\beta^{MG}\right)\right\} = 0$, obtenemos que $t_y = \sum\limits_{k\in U} y_k = \sum\limits_{k\in U} G\left(x_k\beta^{MG}\right)$.

En otras palabras, sin importar las propiedades de β^{MG}, la suma de los valores de la distribución $G\left(x_k\beta^{MG}\right)$ sobre todos los elementos de la población es igual a t_y.

También la anterior relación sugiere una forma de estimar t_y, cuando se tiene información de una muestra:

$$\hat{t}_{yRL} = \sum_{k\in U} G\left(x_k\hat{\beta}^{MG}\right), \tag{3}$$

donde $\hat{\beta}^{GM}$ es el vector de parámetros que resuelve la ecuaciones:

$$\sum_{k\in U}\left\{y_k - G\left(x_k\hat{\beta}^{MG}\right)\right\}x_k^T\frac{I_k}{\pi_k} = 0,$$

donde I_k es la variable aleatoria que indica si la unidad k de la población está en la muestra y π_k es la probabilidad de que dicha unidad sea selecciona en la muestra.

Entonces el estimador en (3) se convierte en $\hat{t}_{yRL} = \sum_{k\in U} G\left(x_k\hat{\beta}^{MG}\right)$. A este estimador se le conoce como el *estimador de regresión logística* del total.

Lo anterior se puede reproducir si la característica de interés se modela con variables aleatorias independientes $Bernoulli\left(G\left(x_k\beta\right)\right)$ y se utilizan los estimadores de máxima verosimilitud para β. Precisamente, la ventaja de no suponer lo anterior es que no se tiene que justificar ni verificar la distribución que se le pueda asociar a la característica de interés.

Ejemplo 2: Estimador de regresión localmente polinomial

Ahora el problema es estimar el total cuando la variable de interés (continua) y la variable auxiliar (escalar) siguen la siguiente relación:

$$y_k = m\left(x_k\right) + \varepsilon_k, \tag{4}$$

donde los residuos son como en el ejemplo 1. Además la función m es desconocida pero se elige tal que alrededor de x_0 se puede aproximar por una serie de Taylor de orden q:

$$m\left(x\right) \approx m\left(x_0\right) + m^{(1)}\left(x_0\right)\times\left(x - x_0\right) + \cdots + \frac{m^{(q)}\left(x_0\right)}{q!}\times\left(x - x_0\right)^q.$$

Otra forma de expresar lo anterior es como sigue

$$m\left(x\right) \approx \beta_0 + \beta_1\times\left(x - x_0\right) + \cdots + \beta_q\times\left(x - x_0\right)^q,$$

donde $\beta_j = \frac{m^{(j)}(x_0)}{j!}$. Observemos que $m\left(x_0\right)$ está representado por β_0 en la anterior ecuación.

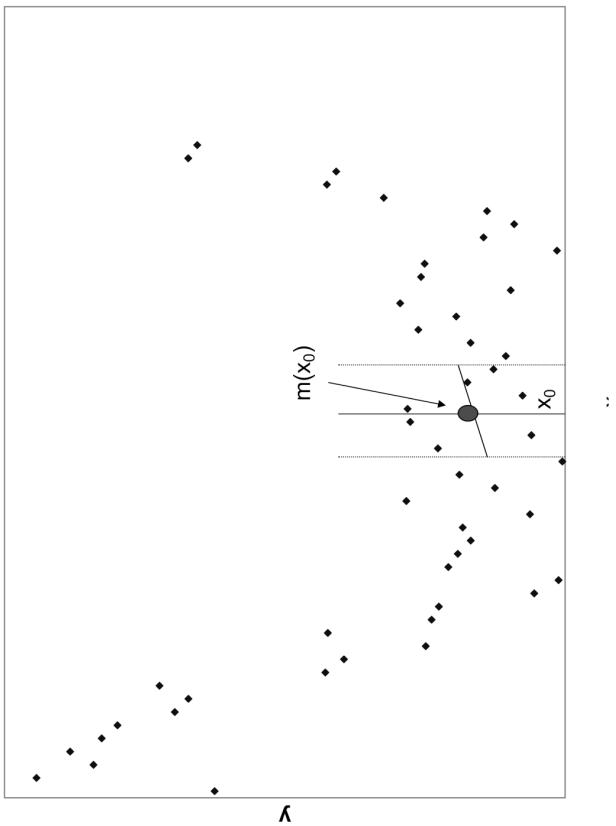
Entonces si se tuvieran varias observaciones alrededor de x_0, se podría estimar $m\left(x_0\right)$ a través de un estimador de β_0. Obsérvese que también se podrían estimar las primeras q derivadas de m. Dada la forma de la última ecuación se podría estimar β_0 minimizando

$$\sum_{k\in U}\left(y_k - \beta_0 - \beta_1\times\left(x_k - x_0\right) + \cdots + \beta_q\times\left(x_k - x_0\right)^q\right)^2 K_h\left(x_k - x_0\right),$$

donde $K_h(x - x_0) = \frac{1}{h} g\left(\frac{x-x_0}{h}\right)$, g es una función de densidad simétrica alrededor de cero y h se le denomina ancho de banda. El papel de la función K_h es definir la vecindad de x_0 y medir la influencia de las observaciones alrededor de éste en la estimación de β_0. El valor estimado para β_0 y por tanto de $m(x_0)$ es

$$\widehat{m(x_0)} = e_1 \left(\sum_{k \in U} K_h(x_k - x_0) \mathbf{x}_k^T(x_0) \mathbf{x}_k(x_0) \right)^{-1} \times \sum_{k \in U} K_h(x_k - x_0) \mathbf{x}_k^T(x_0) y_k.$$

donde $\mathbf{x}_k(x_0) = (1 \quad (x_k - x_0) \quad \cdots \quad (x_k - x_0)^q)$ y e_1 es un vector de orden $q+1$ con un uno en la primer entrada y ceros en las demás. Esto es simplemente hacer $\widehat{m(x_0)} = \hat{\beta}_0$. La forma de estimar la función m se ilustra con la gráfica 2.



Gráfica 2. Estimación a través de un polinomio localmente lineal.

Cuando se tiene información de una muestra, la anterior cantidad se puede estimar por

$$\widehat{m(x_0)}_\pi = e_1 \left(\sum_{k \in U} K_h (x_k - x_0) \mathbf{x}_k^T (x_0) \mathbf{x}_k (x_0) \frac{I_k}{\pi_k} \right)^{-1} \times \sum_{k \in U} K_h (x_k - x_0) \mathbf{x}_k^T (x_0) y_k \frac{I_k}{\pi_k}.$$

Ahora obsérvese que bajo el modelo (4), el total queda como $t_y = \sum_{k \in U} m(x_k) + \sum_{k \in U} \varepsilon_k$.

Una forma de estimar lo anterior es como sigue:

$$\hat{t}_{yLP} = \sum_{k \in U} \widehat{m(x_k)} + \left(\widehat{\sum_{k \in U} \varepsilon_k} \right) = \sum_{k \in U} \widehat{m(x_k)}_\pi + \left(\sum_{k \in U} \left(y_k - \widehat{m(x_k)}_\pi \right) \frac{I_k}{\pi_k} \right).$$

Este último estimador es el *estimador de regresión localmente polinomial* del total, Breidt y Opsomer (2000).

Nuevamente, como no se hizo ningún supuesto para los residuos ε_k, no será necesario justificar ni verificar supuesto alguno antes de utilizar el anterior estimador del total. Esto refuerza también la ventaja de conceptualizar el modelo operacional como un modelo no estocástico.

3. Comentarios Finales

El presente enfoque de modelación y análisis está más apegado al enfoque asistido por modelos, donde el modelo es sólo un modelo operacional y la inferencia se basa en el diseño de muestreo.

A parte de los modelos aquí presentados, se está trabajando en la deducción de un estimador de total cuando utilizamos regresión con coeficientes variables y en la estimación de la probabilidad de pasar de la subpoblación de desempleados a empleados en dos períodos de muestreo consecutivos.

4. Referencias

Breidt, F.J. y Opsomer, J.D. (2000). Local Polynomial Regression Estimators in Survey Sampling. *The Annals of Statistics*, 28, 1026–1053.

Särndal, C.E., Swensson, B. y Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

El método del cubo: Un algoritmo eficiente para la selección de muestras balanceadas

Abel Alejandro Coronado Iruegas[1]
Instituto Nacional de Estadística Geografía e Informática
José de Jesús Suárez Hernández[2]
Instituto Nacional de Estadística Geografía e Informática

1. Introducción

La finalidad de este trabajo es aportar elementos que faciliten la comprensión del algoritmo denominado *método del cubo*, Deville y Tillé (2004), a través de una explicación general para una posterior implementación en el lenguaje R.

2. Muestreo Balanceado

En el muestreo balanceado el objetivo es estimar el total:

$$Y = \sum_{k \in U} y_k$$

de una población U de tamaño N. Suponga un vector de p variables auxiliares para cada uno de los elementos de la población $\mathbf{x}_k = (x_{k1}, ..., x_{kj}, ..., x_{kp})'$

Es decir una matriz de variables auxiliares de la siguiente forma:

$$\mathbf{x} = \begin{pmatrix} x_{11} & & x_{k1} & & x_{N1} \\ \vdots & & \vdots & & \vdots \\ x_{1j} & ... & x_{kj} & .. & x_{Nj} \\ \vdots & & \vdots & & \vdots \\ x_{1p} & & x_{kp} & & x_{Np} \end{pmatrix}_{p \times N}$$

[1] abel.coronado@inegi.gob.mx
[2] jesus.suarez@inegi.gob.mx

Se pueden calcular los totales poblacionales de las p variables auxiliares

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$$

los cuales pueden ser estimados mediante el estadístico de Horvitz Thompson

$$\widehat{\mathbf{X}}_{HT} = \sum_{k \in U} \frac{\mathbf{x}_k S_k}{\pi_k}$$

Entonces un diseño *p(s)* se dice que es balanceado respecto a las variables auxiliares $x_1, ..., x_p$ si y solo si satisface las ecuaciones de balanceo dadas por:

$$\widehat{\mathbf{X}}_{HT} = \mathbf{X}$$

Lo cual puede ser escrito como:

$$\sum_{k \in U} \frac{x_{kj} S_k}{\pi_k} = \sum_{k \in U} x_{kj}$$

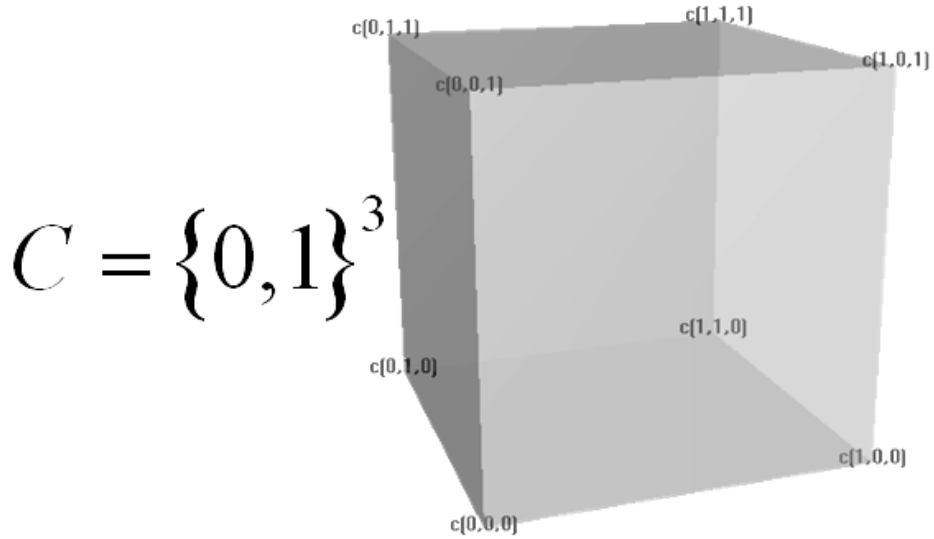
Para toda $s \in S$ tal que $p(s) > 0$ y para toda $j = 1, ..., p$ o en otras palabras var $\left(\widehat{\mathbf{X}}_{HT}\right) = 0$.

3. Representación Geométrica

El método del cubo se basa en una representación geométrica del diseño de muestreo. Las 2^N muestras posibles corresponden a 2^N vectores en $\mathbb{R}^N$ de la siguiente forma. Cada vector s es un vértice de un N-cubo y el número de posibles muestras es el número de vértices de un N-cubo, donde

$$C = \{0, 1\}^N$$

denota a tales vértices. Supongamos una población de tamaño N=3, entonces la representación geométrica correspondiente es la siguiente



Sea la matriz

$$\mathbf{A} = \begin{pmatrix} x_{11}/\pi_1 & & x_{k1}/\pi_k & & x_{N1}/\pi_N \\ \vdots & & \vdots & & \vdots \\ x_{1j}/\pi_1 & \ldots & x_{kj}/\pi_k & \ldots & x_{Nj}/\pi_N \\ \vdots & & \vdots & & \vdots \\ x_{1p}/\pi_1 & & x_{kp}/\pi_k & & x_{Np}/\pi_N \end{pmatrix}_{p \times N}$$

Y los vectores

$$\mathbf{s} = \begin{pmatrix} s_1 \\ \vdots \\ s_k \\ \vdots \\ s_N \end{pmatrix}_{N \times 1}, \boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_k \\ \vdots \\ \pi_N \end{pmatrix}_{N \times 1}$$

Luego

$$\mathbf{As} = \begin{pmatrix} x_{11}/\pi_1 & x_{k1}/\pi_k & x_{N1}/\pi_N \\ \vdots & \vdots & \vdots \\ x_{1j}/\pi_1 \ \ldots & x_{kj}/\pi_k \ \ldots & x_{Nj}/\pi_N \\ \vdots & \vdots & \vdots \\ x_{1p}/\pi_1 & x_{kp}/\pi_k & x_{Np}/\pi_N \end{pmatrix}_{p\times N} \begin{pmatrix} s_1 \\ \vdots \\ s_k \\ \vdots \\ s_N \end{pmatrix}_{N\times 1} = \begin{pmatrix} \sum_{i=1}^{N} \frac{\mathbf{x}_{i1}}{\pi_i} s_i \\ \vdots \\ \sum_{i=1}^{N} \frac{\mathbf{x}_{ij}}{\pi_i} s_i \\ \vdots \\ \sum_{i=1}^{N} \frac{\mathbf{x}_{ip}}{\pi_i} s_i \end{pmatrix}_{p\times 1}$$

y

$$\mathbf{A}\boldsymbol{\pi} = \begin{pmatrix} x_{11}/\pi_1 & x_{k1}/\pi_k & x_{N1}/\pi_N \\ \vdots & \vdots & \vdots \\ x_{1j}/\pi_1 \ \ldots & x_{kj}/\pi_k \ \ldots & x_{Nj}/\pi_N \\ \vdots & \vdots & \vdots \\ x_{1p}/\pi_1 & x_{kp}/\pi_k & x_{Np}/\pi_N \end{pmatrix}_{p\times N} \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_k \\ \vdots \\ \pi_N \end{pmatrix}_{N\times 1} = \begin{pmatrix} \sum_{i=1}^{N} \frac{\mathbf{x}_{i1}}{\pi_i} \pi_i \\ \vdots \\ \sum_{i=1}^{N} \frac{\mathbf{x}_{ij}}{\pi_i} \pi_i \\ \vdots \\ \sum_{i=1}^{N} \frac{\mathbf{x}_{ip}}{\pi_i} \pi_i \end{pmatrix}_{p\times 1} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_j \\ \vdots \\ \mathbf{X}_p \end{pmatrix}_{p\times 1}$$

Entonces las ecuaciones de balanceo pueden ser expresadas de la siguiente forma

$$\mathbf{As} = \mathbf{A}\boldsymbol{\pi}$$

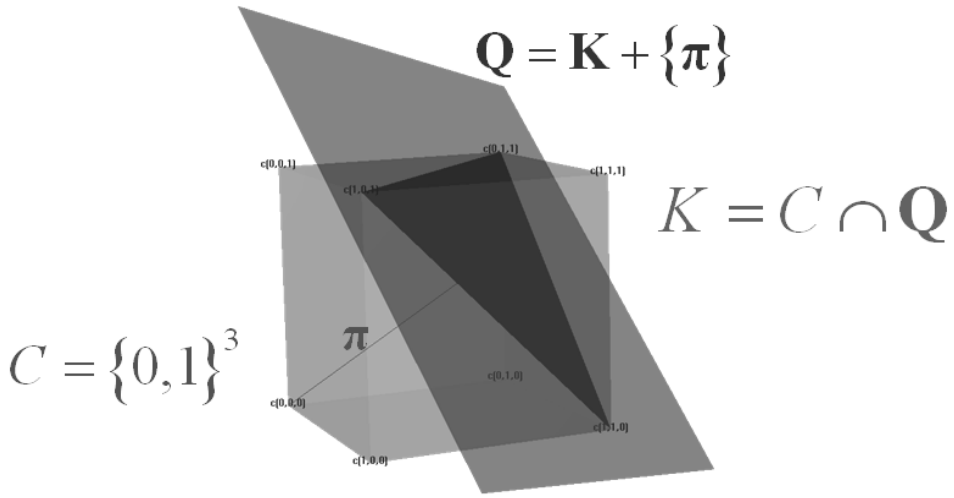
Así pues

$$\mathbf{Q} = \left\{ \mathbf{s} \in \mathbb{R}^N : \mathbf{As} = \mathbf{A}\boldsymbol{\pi} \right\}$$

Se puede demostrar que

$$\mathbf{Q} = \mathrm{K} + \{\boldsymbol{\pi}\}$$

Donde **K** es el kernel de la matriz A. Geométricamente el hipercubo C es intersectado por el plano de restricciones Q cuya solución se encuentra en dicha intersección K que es no vacía y de dimension (N-p) Deville y Tillé (2004).



4. El Método del Cubo

El método del cubo consiste en dos fases

1.-**Fase de Vuelo**

Consiste en seleccionar aleatoriamente un vértice de K = C ∩ Q

Algoritmo

- $\pi(0) = \pi$
- Mientras sea posible realizar el paso 1

Paso 1

- Generar cualquier vector $u(t)$= $[u_k(t)]\neq 0$, aleatorio o no, tal que $u(t)$ esté en el kernel de **A**, y $u_k(t)=0$ si $\pi_k(t)$ es un número entero.

Paso 2

- Calcular $\lambda_1^*(t)$ y $\lambda_2^*(t)$ lo más grande posibles, tal que

  $0 \leqslant \pi(t) + \lambda_1(t)\, u(t) \leqslant 1,$

  $0 \leqslant \pi(t) - \lambda_2(t)\, u(t) \leqslant 1$

Paso 3

- Seleccionar

  $$\pi(t+1) = \begin{cases} \pi(t) + \lambda_1^*(t)\, u(t), & \text{con probabilidad } q_1(t) \\ \pi(t) - \lambda_2^*(t)\, u(t), & \text{con probabilidad } q_2(t), \end{cases}$$

  donde $q_1 = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ y $q_2 = \lambda_1^*(t)\,[\lambda_1^*(t) + \lambda_2^*(t)]$.

2.-**Fase de Aterrizaje**

Enfrentar lo mejor posible el hecho de que las ecuaciones de equilibrio no serán siempre satisfechas de manera exacta. La fase de aterrizaje se puede completar por medio de un algoritmo enumerativo (MPL) sobre la subpoblación.

$$C(\pi*)$$

$$\min_{p*(.)} \sum_{s \in C(\pi*)} \text{Costo}(s)\, p*(s)$$

Sujeto a

$$\sum_{s \in C(\pi*)} p*(s) = 1$$

$$\sum_{s \in C(\pi*)} sp*(s) = \pi*$$

$$0 \leqslant p*(s) \leqslant 1, \text{ para toda } s \in C(\pi*)$$

Donde

$$C(\pi*)$$

Denota al conjunto de los elementos para los cuales ha sido satisfecho el balanceo

5. Referencias

Deville, J.-C., Tillé, Y., (2004). Efficient balanced sampling: The Cube Method. *Biometrika; 91 (4)*, pp. 893-912.

Un modelo para datos longitudinales con dependencia espacial-temporal[1]

Felipe Peraza[2]
Universidad Autónoma de Sinaloa
Graciela González-Farías [3]
Centro de Investigación en Matemáticas, A.C.

1. Introducción

Considere un problema epidemiológico donde la variable de interés se mide en una escala ordinal. Donde la propagación de la enfermedad está correlacionada al estado de los *vecinos*. Formalmente, decimos que el sitio j $(\neq i)$ es un vecino del sitio i si y solo si la forma funcional de $\Pr(x_i \mid x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$ depende de la variable x_j. Si inversamente, tenemos una medida de probabilidad que defina una estructura de vecinos, obtenemos un *Campo Aleatorio Markoviano* (MRF) (Besag, 1974). Un ejemplo importante de un MRF para este trabajo son los modelos autologísticos. El modelo autologístico base fue desarrollado por Besag (1974) . Besag (1977) usa un modelo auto-logístico para describir la incidencia de la enfermedad de una planta como función de la enfermedad en plantas vecinas tanto en el tiempo anterior como actual. Combinando los conceptos de campos markovianos para la dependencia espacial y de cadenas de Markov para la dependencia temporal. Una extensión para datos ordenados categóricamente, aparece en Strauss (1992) . En el modelo de Besag (1977) las plantas tienen o no la enfermedad así que este trabajo puede considerarse como una extensión para el caso de datos ordenados categóricamente. En el caso del modelo autologístico, la verosimilitud es intratable excepto para los casos más simples, así que son necesarios métodos alternativos a máxima verosimilitud para la estimación de parámetros.

En este trabajo, proponemos un modelo para el análisis e inferencia de datos ordinales con dependencia espacial y temporal e investigamos las condiciones para le existencia y unicidad del EMV así como sus propiedades asintóticas. Analizamos un conjunto de datos de Agave.

[1]Trabajo realizado con apoyos de la CGIP-UAS y CONACYT: SEP-2004-C01-45974-F
[2]fperaza@uas.uasnet.mx
[3]farias@cimat.mx

2. Modelación

Sea $\mathcal{R}$ el conjunto de sitios de una rejilla rectangular que por lo pronto supondremos regular. Sea n el número total de sitios, τ el total de periodos observados y supongamos que en cada sitio $i \in \mathcal{R}$ y cada tiempo $t \in \{0, 1, 2, ..., \tau\}$ cada observación z_{it} toma uno de E valores en el conjunto ordenado $S = \{1, 2, 3, ..., E\}$. Y que $z_{i1}, ..., z_{i\tau}$ es no decreciente.

Sea $\mathbf{Z}_{\partial_i,t}$, el conjunto de vecinos de la planta i al tiempo t. Combinando la suposición de dependencia Markoviana en el tiempo y el espacio, supondremos que la distribución condicional de Z_{it} depende funcionalmente solo del estado de sus vecinos y del mismo sitio i al tiempo anterior, de la siguiente manera,

$$\Pr\left\{ Z_{it} \mid \{Z_{js},\ j \in \mathcal{R}; Z_{i,s}\}_{s=0}^{t-1} \right\} = \Pr\left\{ Z_{it} \mid Z_{i,t-1}, \mathbf{Z}_{\partial_i,t} \right\} = G\left(Z_{it}; Z_{i,t-1}, \mathbf{Z}_{\partial_i,t} \right) \tag{1}$$

donde G es una función de densidad -función de transición de probabilidades- discreta. Es claro que esta modelación no es un Campo Aleatorio Markoviano, dado que existe dependencia con respecto al pasado. Denotando por $\mathbf{Z}^t = (Z_{1t}, Z_{2t}, ..., Z_{nt})$ las observaciones en la rejilla al tiempo t con las suposición adicional que las observaciones son condicionalmente independientes dados sus vecinos, podemos escribir,

$$P\left(\mathbf{Z}^0 = \mathbf{z}^0, \mathbf{Z}^1 = \mathbf{z}^1, ..., \mathbf{Z}^\tau = \mathbf{z}^\tau \right) = P\left(\mathbf{Z}^0 = \mathbf{z}^0 \right) \prod_{t=1}^{\tau} \prod_{i=1}^{n} G\left(Z_{it}; \mathbf{Z}_{\partial_i,t}, Z_{i,t-1} \right). \tag{2}$$

3. Inferencia

Para parametrizar la función de transición G, usamos la función de distribución logística. La distribución logística para datos espaciales ha sido ampliamente utilizada, por ejemplo en Besag (1974, 1977) y Strauss (1992). Para $t \in \{1, 2, ...\tau\}$, $i \in \{1, 2, ...n\}$ y $s_1, s_2 \in \{1, 2, ..., E\}$, sea $u_{i,t}(e)$ el número de vecinos de la observación i, que se encuentre en el estado e, al tiempo t. Y sea $\mathbf{x}_{i,t-1} = \left(u_{i.t-1}(1), ..., u_{i.t-1}(E) \right)^T$. Supondremos que para cada t, las probabilidades de transición siguen un modelo logístico,

$$P\left(Z_{it} = e_2 \mid Z_{i,t-1} = e_1, \mathbf{Z}_{\partial_i,t-1} = \mathbf{z}_{\partial_i t-1} \right) = \begin{cases} G\left(\mathbf{x}_{i,t-1}^T \boldsymbol{\beta}_{e_1} \right), & e_2 = e_1 + 1 \\ 1 - G\left(\mathbf{x}_{i,t-1}^T \boldsymbol{\beta}_{e_1} \right), & e_2 = e_1 \\ 1, & e_1 = e_2 = E \end{cases} \tag{3}$$

Para la estimación, consideremos la distribución conjunta (2) y la parametrización (3), y para $e = 1, 2, ..., E$, sea $l_e(\boldsymbol{\beta}_e)$ la verosimilud correspondiente a los datos cuyas transiciones de probabilidad son del tipo (e, e) o $(e, e+1)$, entonces podemos escribir la verosimilitud como $l(\boldsymbol{\beta}) = \log P\left(\mathbf{Z}^0 = \mathbf{z}^0\right) + l_1(\boldsymbol{\beta}_1) + l_2(\boldsymbol{\beta}_2) + ...l_{E-1}\left(\boldsymbol{\beta}_{E-1}\right)$ donde

$$l_e(\boldsymbol{\beta}_e) = \sum_{i=1}^{n}\sum_{t=1}^{\tau} 1_e(z_{i,t-1})\left\{1_{e+1}(z_{i,t})\log G\left(\mathbf{x}_{t-1}^T\boldsymbol{\beta}_e\right) + 1_e(z_{i,t})\log\left[1 - G\left(\mathbf{x}_{t-1}^T\boldsymbol{\beta}_e\right)\right]\right\}.$$

Así, los parámetros $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_{E-1}$ son funcionalmente independientes. Las condiciones para la existencia y unicidad del EMV se muestran en Peraza (2004), y se basan en los resultados de Silvapulle (1981). Para demostrar las propiedades asintóticas del modelo, usamos las ideas de los Modelos Markovianos Parcialmente Ordenados (POMM) definidos en Huang and Cressie, (2000). Como comentamos antes nuestro modelo no es MRF, por tanto no es un POMM. Sin embargo explotamos las ideas de orden parcial en la rejilla definidos como en un POMM. En Peraza (2004) se muestran las condiciones para que el EMV sea único, consistente y asintóticamente normal.

4. Predicción

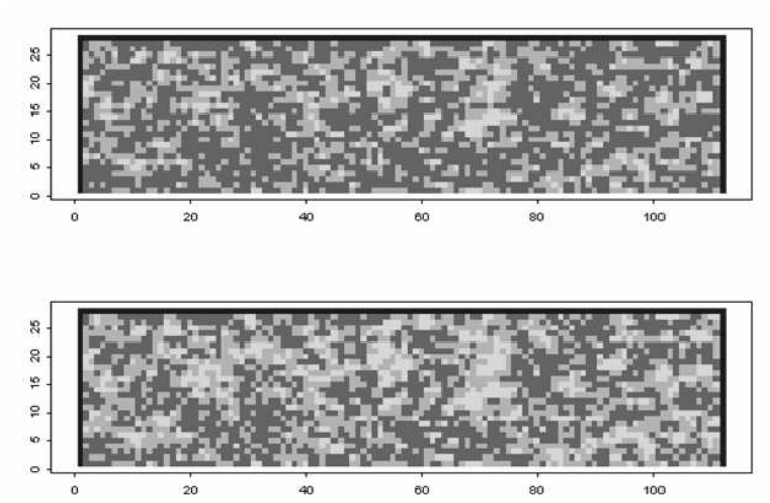
Para el caso de predecir nuevas observaciones en los mismos sitios y en los tiempos futuros, suponga datos en un rejilla $\mathcal{R}$ observados en los tiempos $t = 0, 1, ..., \tau$ y sea $\hat{\boldsymbol{\beta}}$ el EMV. Dada la suposición markoviana en el tiempo y la suposición que las observaciones son condicionalmente independientes dados sus vecinos, obtenemos el estimador MV de $\hat{Z}_{i,\tau+1}$, dado por $\hat{Z}_{i,\tau+1}^{MV} = \sum_{e=1}^{E-1} 1_e(z_{i,\tau})\left[e + G\left(\mathbf{x}_{i,\tau}^T\hat{\boldsymbol{\beta}}_e\right)\right]$, si $z_{i,\tau} \neq E$ y $\hat{Z}_{i,\tau+1}^{MV} = E$ si $z_{i,\tau} = E$.

5. Ejemplo: *Agave Tequilana Weber*

El agave es una planta que crece en ciertas regiones de México de la cual se extrae el tequila. En 1998 el 22 % de las plantaciones comenzaron a presentar síntomas de una enfermedad que produce la marchitez de la planta dejándola inutilizable para la producción del tequila. Presentamos aquí un conjunto de datos provenientes de un estudio que se realizó para una empresa dedicada a procesar la planta de agave para producir tequila. Para considerar la posibilidad de mayor frecuencia de

contagio dentro del mismo surco, modificamos el modelo asignándole una ponderación de 5/6 si la planta vecina provenía del mismo surco y de 1/6 en otro caso. Una vez que el hongo está presente no es factible eliminarlo por lo que el efecto de los agroquímicos debería ser el retardo de la presencia de daño severo en la hoja. En la práctica, se mide el daño visual registrando los valores de 1 (hoja sana) , 2 y 3 (pérdida total). Se hacen mediciones en distintos periodos de tiempo (5 en total, aproximadamente cada 2 meses) para ver el avance de la enfermedad. Para el análisis consideramos los datos de una parcela de 28 surcos con 112 plantas cada uno. Los datos han sido modificados o tomados de partes estratégicas para preservar la confidencialidad de la información. El EMV para las transiciones del estado 1 al 2 es $\hat{\boldsymbol{\beta}}_1 = (-0{,}737, -2{,}129, 0{,}969, 2{,}711)$, y el EMV para las transiciones del estado 2 al 3 es $\hat{\boldsymbol{\beta}}_2 = (-1{,}121, -1{,}572, 0{,}558, 2{,}247)$.

Finalmente, la siguiente figura muestra la parcela observada al tiempo 4, y su predicción al tiempo 5. Los valores mas claros indican plantas con mayor daño. El error de predicción estimado es de 0,135.



6. Conclusiones y trabajo futuro

Presentamos un modelo para datos ordinales que incorporan dependencias espaciales y temporales bajo un enfoque de máxima verosimilitud. Este modelo permite probar, de manera unilateral los diferentes tipos de dependencia. Para futuro trabajo es conveniente considerar otras covariables adicionales a los vecinos, por ejemplo para incluir los efectos de los tratamientos. En este trabajo

consideramos que los parámetros no se modifican con el tiempo, la cual puede ser una suposición no realista para algunos casos, actualmente exploramos esta situación utilizando MCMC. La investigación del modelo desde un punto de vista bayesiano, forma también parte de nuestra investigación futura.

7. Referencias

Besag, J.E. (1977) . Some methods of statistical analysis for spatial data. *Proceedings of the 41st Session of the International Statistical Institute* (New Delhi, 1977), **2**. With discussion. *Bull. Inst. Internat. Statist.* **47**, no. 2, 77–91, 138–147.

Besag, J.E. (1974) . Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statist. Soc. B*, **36**, 192-236.

Huang, H. and Cressie, N. (2000) . Asymptotic properties of maximum (composite) likelihood estimators for partially ordered Markov models. *Statistica Sinica*, **10**, 1325-1344.

Strauss, D. J. (1992) . Clustering on coloured lattices. *Journal of Applied Probability*, **14**, 135-143.

Peraza-Garay, F. (2004) . *A model for longitudinal and ordinal data with spatial dependency* Disertación Doctoral. CIMAT, México.

Silvapulle, M. J. (1981) . On the existence of maximum likelihood estimators for the binomial reponse models. *J.R. Statist. Soc. B*, **43**, no. 3, 310-313.

Comparación de concentraciones medias de contaminantes usando una prueba de razón de verosimilitud

Fidel Ulín-Montejo[1]

División Académica de Ciencias Básicas, Universidad Juárez Autónoma de Tabasco

Humberto Vaquera-Huerta [2]

Estadística Colegio de Postgraduados

1. Introducción

Cuando la concentración de un contaminante es tan pequeña que no puede cuantificarse se reporta un *nondetect*, dato-censurado-por-la-izquierda (Helsel, 2004). Se ha reportado que las concentraciones de ciertos contaminantes son lognormales (Ott, 1990), de donde la normalidad se asume logtransformando los datos y se realiza inferencia sobre las medianas (Stolline, 1992), sin embargo, El-Shaarawi y Viveros (1997); y la US EPA discuten lo anterior y la necesidad de obtener resultados sobre las concentraciones medias, ya que éstas tienen una interpretación fisicoquímica. Este trabajo propone una prueba para comparar concentraciones medias de k poblaciones lognormales con σ común en presencia de *nondetects*, basados en los métodos de Meeker y Escobar (1998). La potencia de la prueba es observada, bajo el efecto de censura, mediante simulación Monte Carlo.

2. Metodología

Sean $w_{11}, \ldots, w_{1n_1}; \cdots; w_{k_1}, \cdots; w_{k_{n_k}}$ muestras independientes lognormales de las k poblaciones, con mediana y media, $M_i = exp(\mu_i)$ y $E_i = exp(\mu_i + \sigma_i^2/2)$, $i = 1, \cdots, k$, respectivamente. Se supone que las primeras r_i realizaciones para la muestra i son *no-censuradas*; es decir $y_{ij} = log(w_{ij})$, $j = 1, \cdots, r_i$; por el contrario, las últimas $n_i - r_i$ son *nondetects* debidas a un límite de detección (LD), es decir $y_{ij} = log(LD_{ij})$, $j = r_i + 1, \cdots, n_i$. Luego, se define μ como un modelo de regresión con variables *dummy* $X = (1, x_1, \cdots, x_{k-1})$ y parámetros $\beta = (\beta_0, \beta_1, \cdots, \beta_{k-1})$:

[1] fidel@colpos.mx

[2] hvaquera@colpos.mx

$$\mu(X\beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 +, \cdots, \beta_{k-1} x_{k-1}, \quad \sigma_i = \sigma \quad \forall i. \tag{1}$$

2.1. Modelo Reducido

La función de verosimilitud para el modelo $H_1 : \mu_i = \mu, \quad \forall i$ es

$$L_1(\beta, \sigma) = \prod_{i=1}^{k} \left[\prod_{j=1}^{r_i} \frac{1}{\sigma} \phi(z_{ij}) \prod_{j=r_i+1}^{n_i} \Phi(z_{ij}) \right]. \tag{2}$$

Donde $z_{ij} = (y_{ij} - \beta_0)/\sigma$, para $i = 1, 2, \ldots, k$ y $j = 1, \ldots, n_i$; $\phi(y)$ $\Phi(y)$, fdp y fda de la $n(0, 1)$, respectivamente. Ahora, con $W(z) = \frac{\phi(z)}{\Phi(z)}$ y empleando nlminb de R, los estimadores de máxima verosimilitud (EMV) para β_0 y σ se obtienen como soluciones simultáneas de:

$$\frac{\partial log L_1(\beta_0, \sigma)}{\partial \beta_0} = \sum_{1=1}^{k} \sum_{j=1}^{r_i} z_{ij} - \sum_{i=1}^{k} \sum_{j=r_i+1}^{n_i} W(z_{ij}) = 0, \tag{3}$$

$$\frac{\partial log L_1(\beta_0, \sigma)}{\partial \sigma} = \sum_{i=1}^{k} \left[-r_i + \sum_{j=1}^{r_i} z_{ij}^2 - \sum_{r_i+1}^{n_i} z_{ij} W(z_{ij}) \right] = 0 \tag{4}$$

2.2. Modelo Completo

Sea $z_{ij} = (y_{ij} - \mu_i)/\sigma$ y $\mu_i = \beta_0 + \beta_i$, entonces la función de verosimilitud para $H_2 : \mu_i \neq \mu_j$ es

$$L_2(\beta, \sigma) = \prod_{i=1}^{k} \left[\prod_{j=1}^{r_i} \frac{1}{\sigma} \phi(z_{ij}) \prod_{j=r_i+1}^{n_i} \Phi(z_{ij}) \right]. \tag{5}$$

Análogamente al modelo reducido los EMV de $\beta_0, \cdots, \beta_{k-1}$ y σ se obtienen de:

$$\frac{\partial log L_2(\beta, \sigma)}{\partial \beta_0} = \sum_{i=1}^{r_1} z_{1i} - \sum_{j=r_i+1}^{n_1} W(z_{1i}) = 0, \quad i = 1, \cdots, k \tag{6}$$

$$\frac{\partial logL_2(\beta,\sigma)}{\partial\sigma} = -\sum_{i=1}^{k} r_i + \sum_{i=1}^{k}\left[\sum_{j=1}^{r_i} z_{ij}^2 - \sum_{j=r_i+1}^{n_i} z_{ij}W(z_{ij})\right]. \tag{7}$$

2.3. Prueba de Razón de Verosimilitud e Intervalo de Confianza

Sea $R(\theta_0) = L_2/L_1$, entonces un modelo reducido ajusta los datos tan bien como un modelo completo si $-2log[R(\theta_0)] \sim \mathcal{X}^2_{(s-r)}$. Entonces, la comparación de las concentraciones medias de las k poblaciones se realiza comparando ambos modelos:

$$\mathcal{X}^2 = -2\left[logL_1(\hat{\beta}_0,\hat{\sigma}) - logL_2(\hat{\beta}_0,\cdots,\hat{\beta}_{k-1},\hat{\sigma})\right] > \mathcal{X}^2_{\alpha,(k+1)-2}. \tag{8}$$

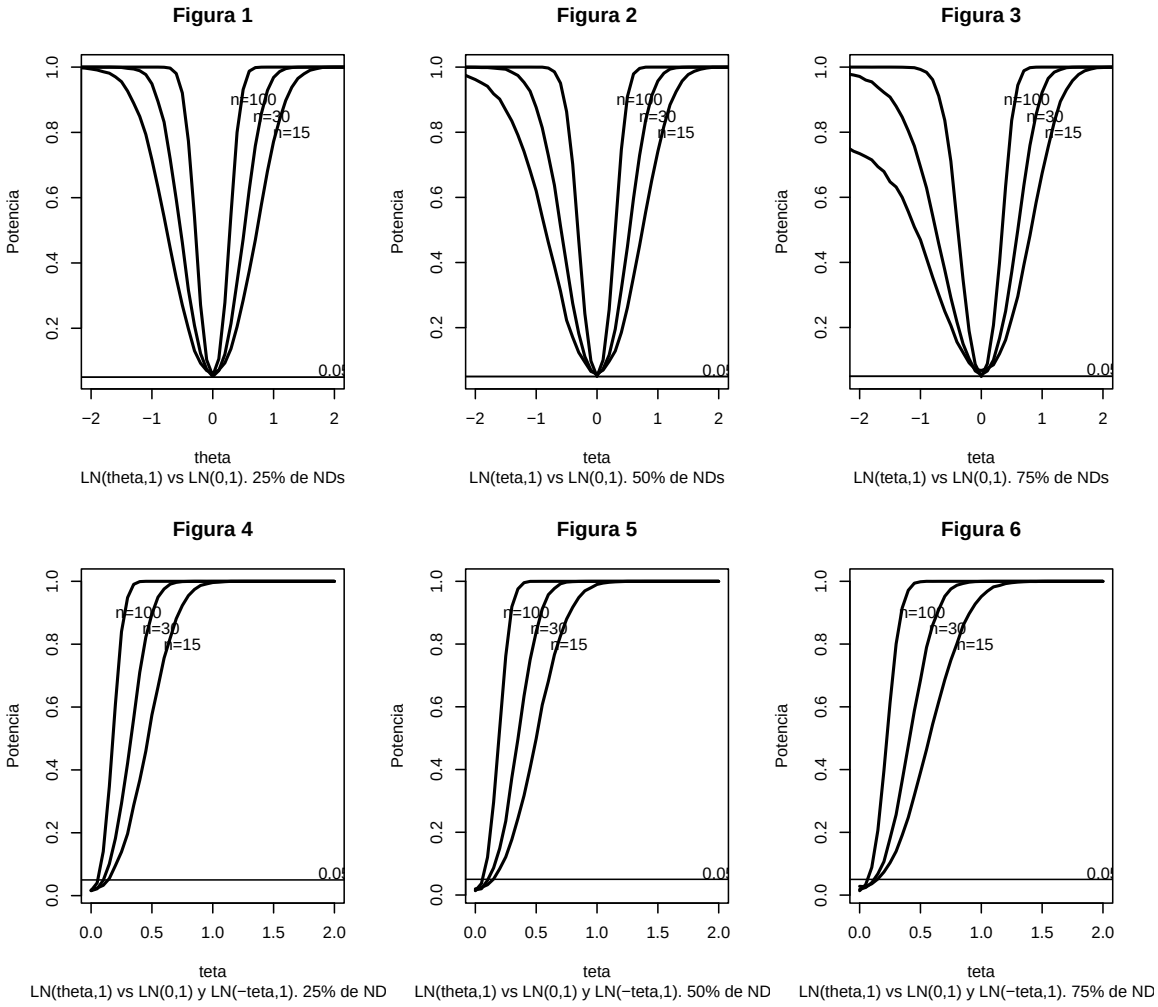
Ahora, sea g una función de β, entonces con la aproximación normal $R(\theta_0)$ y la matriz observada de Fisher (Díaz-Francés y Sprott, 2000), un IC aprox. del $100(1-\alpha)\,\%$ para g se construye así:

$$[g_L(\beta), g_U(\beta)] = \hat{g}(\beta) \pm z_{(1-\alpha/2)}\hat{se}_{\hat{g}(\beta)} \tag{9}$$

3. Potencia de la prueba

La función de potencia se obtiene de la $P_{\underline{\beta}}$ (muestra aleatoria $\in$ región de rechazo). Luego, para obtener una prueba de nivel α, se elige una constante c, tal que $sup_{\beta\in B_0}P_{\underline{\beta}}[R(\underline{\beta_0}) \leq c] \leq \alpha$.

Las Figuras 1, 2 y 3, muestra las potencias estimadas al comparar muestras aleatorias de dos poblaciones lognormales, $LN(\theta,1)$ y $LN(0,1)$, con $n_1 = n_2 = 15, 30, 100$ y $\alpha = 0{,}05$. Se consideran porcentajes totales de *nondetects*(NDs) del $25, 50, 75\,\%$ con 10,000 simulaciones en R. Las Figuras 4, 5 y 6 muestran las potencias para *tres poblaciones*, $LN(\theta,1), LN(0,1)$ y $LN(-\theta,1)$.



4. Aplicaciones

Ejemplo 4.1. De Millard y Deverel (1988), en la Tabla 1 se reportan niveles de cobre en mantos freáticos de dos áreas en California, USA. Alrededor del 20 % son *nondetects*, denotados por " < ".

Tabla 1. Datos de concentraciones de cobre

Alluvial Fan Zone	<1, <1, <1, <1, <5, <5, <5, <5, <5, <5, <5, <20, <20, 1, 1, <5, <10, <10, <10, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4 4, 4, 5, 5, 5, 7, 7, 7, 8, 9, 10, 11, 12, 16, 20
Basin-Trough Zone	<1, <1, <2, <2, <5, <5, <5, <5, <5, <10, <10,<10, <10, <15, 1, 1, 1, 1, 1, 1,1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 6, 6, 8, 9, 9,12, 14, 15, 17, 23

Con los datos $logL_2 = -224{,}3$, $logL_1 = -224{,}9$ y $p - valor = 0{,}273$, por lo que no difieren significativamente; así, no se rechaza el modelo reducido. Nótese también, en los IC aprox. de la Tabla 2, que $\mu_0 = \mu_1 = \beta_0$ y, del supuesto, $\sigma_i = \sigma$. Así, a la luz de los datos contenidos en las muestras, las poblaciones tienen la misma concentración media de cobre.

Tabla 2. EMV e IC Aprox. para los parámetros en los datos de Cobre

Modelo	Parámetro	EMV	Error Estándar	IC aprox. 95 % Conf.
Completo,H_2.	β_0	1.045	0.138	0.775, 1.315
	β_1	-0.040	0.181	- 0.395, 0.314
	σ	0.882	0.068	0.759, 1.026
Reducido,H_1.	β_0	1.021	0.090	0.844, 1.199
	σ	0.883	0.068	0.759, 1.026

Ejemplo 4.2. En EPA (1992) se reportan concentraciones de zinc en cinco pozos, los datos comprenden alrededor del 50 % de *nondetects* y se muestran en la Tabla 3.

Tabla 3. Datos de concentraciones de zinc

Well 1	<7, <7, <7, <7,<7, 10.00, 11.41, 15.00
Well 2	<7, <7, <7, <7, 10.50, 11.56, 12.59, 13.70
Well 3	<7, <7, <7, <7, 9.36, 12.00, 12.85, 14.20
Well 4	<7, <7, <7, 10.90, 11.05, 11.69, 12.22, 13.24
Well 5	<7, <7, <7, <7, 8.74, 11.15, 12.35,13.31

De la Tabla 3, $logL_2 = -78{,}5$, $logL_1 = -78{,}9$ y $p - valor = 0{,}938$, por lo que no existen diferencias significativas. Nótese también en la Tabla 4, que los IC para los β_i contienen el cero. De modo que $\mu_i = \beta_0, i = 1, \cdots, 5$. Entonces, las poblaciones tienen la misma concentración media de zinc.

Tabla 4. EMV e IC Aprox. para los parámetros en los datos de Zinc

Modelo	Parámetro	EMV	Error Estándar	IC aprox. 95 % Conf.
Completo,H_2.	β_0	1.994	0.197	1.609, 2.380
	β_1	- 0.108	0.278	-0.654, 0.438
	β_2	0.040	0.271	- 0.491, 0.571
	β_3	0.037	0.271	-0.494, 0.568
	β_4	0.154	0.267	- 0.369, 0.677
	σ	0.494	0.090	0.346, 0.707
Reducido,H_1.	β_0	2.019	0.097	1.829, 2.210
	σ	0.501	0.091	0.351, 0.717

5. Conclusiones

Los métodos basados en máxima verosimilitud son muy útiles en problemas ambientales con información censurada. La prueba que se propone provee una alternativa a los procedimientos no

paramétricos, donde se ignora la información censurada o se asignan cantidades arbitrarias; y también para los procedimientos basados en la normalización, lo cual es discutible considerando las características fisicoquímicas de los contaminantes. Por otro lado, la prueba muestra una potencia aceptable bajo tamaños de muestra relativamente pequeños y altos porcentajes totales de censura; además, con los intervalos de confianza aproximados, pueden establecerse criterios de comparación y agrupación.

6. Referencias

Díaz-Francés, E. y Sprott, D. A. (2000). The Use of The Likelihood Function in The Analysis of Environmental Data. *Environmetrics* **11**, 75-79.

El-Shaarawi, A. H. y Viveros, R. (1997). Inferences About The Mean In Log-Regression With Environmental Applications. *Environmetrics* **8**, 569-582.

EPA (1992). Statistical Training Course for Ground-Water Monitoring Data Analysis. *U.S. Environmental Protection Agency Office of Solid Waste,* EPA530-R-93-003.

Helsel, D. R. (2004). *Nondetects And Data Analysis.* New York: Wiley.

Meeker, W. Q. y Escobar, L. A. (1998). *Failure-Time Regression Analysis. In: Statistical Methods for Reliability Data.* New York: Wiley.

Millard, S. P. y S. J. Deverel (1988). Nonparametric statistical methods for comparing two sites. *Water Resources Research* **24**, 2087-2098.

Ott, W. R. (1990). A physical explanation of the lognormality of pollutant concentrations. *Journal of the Air and Waste Management Association* **40**, 1378-83.

Stoline, M. R. (1993). Comparison of Two Medians Using A Two-Sample Lognormal Model in Environmental Contexts. *Environmetrics* **4**, 323-339.

Estimación del área bajo la curva ROC

Carlos Cuevas Covarrubias[1]
Universidad de Anáhuac

1. Introducción

La curva ROC es una herramienta de análisis gráfico diseñada para evaluar la calidad de cualquier índice de riesgo o función discriminante. Sus primeras aplicaciones prácticas se dieron durante la década de 1940 conjuntamente con el desarrollo de los sistemas de radar. El objetivo inicial era evaluar la capacidad del operador para distinguir entre ruido y señal; probablemente, el término ROC proviene de este contexto[2]. Poco más tarde, aparecieron nuevas aplicaciones en la psicología experimental (Bamber 1975). Actualmente, las curvas ROC constituyen una herramienta estándar en la evalución del diagnóstico médico (Zweig y Cambell 1993, Campbell y Machin 1990). En este trabajo presentamos una nueva metodología que permite estimar curvas ROC con kérneles de suavizamiento. Algunos de los resultados que aquí presentamos se explican con mayor detalle en Cuevas-Covarrubias y Copas (2001).

2. Índices de riesgo y curvas ROC

Con frecuencia, el diagnóstico médico puede apoyarse en diversas técnicas de clasificación estadística. Dado un índice de riesgo S y un umbral numérico t, se aplica una regla de decisión como la siguiente:

$$\text{clasificar en } \begin{cases} \Omega_0 & \text{si } S < t \\ \Omega_1 & \text{si } S \geq t, \end{cases}$$

siendo Ω_0 y Ω_1 los pacientes sanos y enfermos respectivamente [3]. La curva ROC de un índice de riesgo S se define como

$$\text{ROC} = \{(x, y) | x = 1 - F_0(t), y = 1 - F_1(t), t \in \Re\},$$

en donde $F_0 = Pr[S \leq t|\Omega_0]$ y $F_1 = Pr[S \leq t|\Omega_1]$. Es decir, la curva ROC es el conjunto de todas las combinaciones de sensibilidad y tasa de falsos positivos que podemos obtener al tomar a S como

[1] ccuevas@anahuac.mx
[2] Receiver Operating Characteristic
[3] Valores grandes de S sugieren que el paciente padece la enfermedad objeto del diagnóstico

índice de riesgo; es por lo tanto, un resumen gráfico de su desempeño. El área bajo la curva ROC nos brinda un resumen objetivo sobre la calidad de S. Si definimos a X y Y como valores de S en elementos de Ω_0 y Ω_1 respectivamente, entonces se cumple que

$$A = \int_{-\infty}^{\infty} F_0(t) dF_1(t) = P(X \leq Y).$$

En otras palabras, el área bajo la curva ROC es la probabilidad de que al seleccionar aleatoriamente un individuo de Ω_0 y otro de Ω_1, el índice S los ordene correctamente conforme a la regla de diagnóstico.

3. La curva ROC empírica

Dadas dos muestras aleatorias de $X = S : \Omega_0 \to \Re$ y $Y = S : \Omega_1 \to \Re$, la curva ROC empírica se define como

$$\widehat{ROC} = \{(x, y) | x = 1 - \widehat{F_X}(t), y = 1 - \widehat{F_Y}(t), t \in \Re\}$$

en dónde $\widehat{F_X}(t)$ y $\widehat{F_Y}(t)$ son funciones empíricas de distribución;

$$\widehat{F_X}(t) = \sum_{i=1}^{n_x} \mathrm{I}_{(-\infty,t]}(x_i) \qquad \widehat{F_Y}(t) = \sum_{i=1}^{n_y} \mathrm{I}_{(-\infty,t]}(y_i).$$

Es importante notar que la curva ROC empírica es una sucesión finita de puntos; es decir, es un estimador discreto para una curva continua. Generalmente, este problema se resuelve interpolando linealmente los puntos de la curva. Esta técnica no sólo es conveniente desde el simple punto de vista gráfico. El área bajo una curva $\widehat{ROC}$ interpolada es un estimador insesgado de $A = Pr[X \leq Y]$ que además coincide con la estadística de *Mann-Whitney*.

$$\hat{A} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \psi(y_j - x_i).$$

En donde

$$\psi(u) = \begin{cases} 0 & \text{if } u < 0 \\ \frac{1}{2} & \text{if } u = 0 \\ 1 & \text{if } u > 0. \end{cases}$$

La varianza de $\widehat{A}$ está dada por

$$\frac{A(1-A) + (n_Y - 1)Var(F_Y(X)) + (n_X - 1)Var(F_X(Y))}{n_x n_y}$$

4. Curvas ROC suavizadas

Es posible definir un estimador suave de la curva ROC a partir de los estimadores suavizados de F_X y F_X; i.e.

$$\widetilde{ROC} = \{(x,y)|x = 1 - \widetilde{F_X}(t), y = 1 - \widetilde{F_Y}(t), t \in \Re\};$$

siendo

$$\widetilde{F_X}(t) = \int_{-\infty}^{x} \widetilde{f_X}(t)dt = \frac{1}{n_X}\sum_{i=1}^{n_X} \Phi\Big(\frac{x - x_i}{h_X}\Big).$$

y $\widetilde{F_Y}(t)$ definida de forma equivalente. Dado que Φ es un kérnel Gaussiano, entonces la primer convolución de la distribución normal nos permite expresar el área bajo la curva $\widetilde{ROC}$ como

$$\tilde{A} = \frac{1}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} \Phi\Big(\frac{y_j - x_i}{h}\Big)$$

en donde

$$h = \sqrt{h_X^2 + h_Y^2}.$$

Nuestra propuesta es obtener el estimador $\tilde{A}$ con mínimo error cuadrático medio, y luego construir un estimador suave $\widetilde{ROC}$ condicionado al estimador óptimo del área.

5. Calibración del modelo

El error cuadrático medio de $\tilde{A}$ es una función continua del parámetro de suavizamiento. La siguiente expresión es un polinomio de Taylor que permite aproximarlo para valores pequeños de h:

$$M(h) \approx Var(\hat{A}) - \frac{\kappa f_D(0)}{n_X n_Y}h + \frac{V^*}{n_X n_Y}h^2 + \frac{\{f_D'(0)\}^2}{4}h^4. \quad (1)$$

Siendo

$$V^* = \{(n_Y - 1)cov(f_Y'(X), F_Y(X)) + (n_X - 1)cov(f_X'(Y), F_X(Y))\}$$

y f_D la densidad de $D = Y - X$ ($\kappa \approx 0{,}5642$). El parámetro óptimo de suavizamiento h^*, puede aproximarse resolviendo $M'(h) = 0$; i.e.,

$$h^3 + ph + q = 0. \quad (2)$$

Los coeficientes p y q de la ec. (2), se obtienen directamente de la ec. (1). Es posible demostrar que esta ecuación tiene una y sólo una solución en los reales positivos, y que el orden de magnitud de h^* depende del signo de p (ver Cuevas-Covarrubias y Copas 2001).

p	h^*
< 0	$O(n^{-\frac{1}{2}})$
$= 0$	$O(n^{-\frac{2}{3}})$
> 0	$O(n^{-1})$

Algunos autores recomiendan ocupar los niveles de suavizamiento óptimos para estimar funciones de densidad o funciones de distribución; i.e., $h^* = O(n^{-\frac{1}{5}})$ o $h^* = O(n^{-\frac{1}{3}})$. Sin embargo, nuestro trabajo demuestra que para estimar $\widetilde{A}$ necesitamos un menor nivel de suavizamiento (Cuevas-Covarrubias y Copas 2001).

6. Estimación de la curva ROC

Dado un t fijo, la distancia esperada entre sus puntos correspondientes en ROC y $\widetilde{ROC}$ es

$$\nabla^2 = E(F_X(t) - \tilde{F}_X(t))^2 + E(F_Y(t) - \tilde{F}_Y(t))^2$$

Podemos entonces minimizar ∇^2 sujeto a $h^* = (h_X^2 + h_Y^2)^{1/2}$. Utilizando un polinomio de Taylor para aproximar el error cuadrático medio de F_X, h_X se obtiene resolviendo

$$a_X h_X^2 + \frac{b_X}{4h_X} = a_Y(h^{*2} - h_X^2) + \frac{b_Y}{4\sqrt{h^{*2} - h_X^2}}.$$

Dados h y h_X, el parámetro h_Y se calcula directamente.

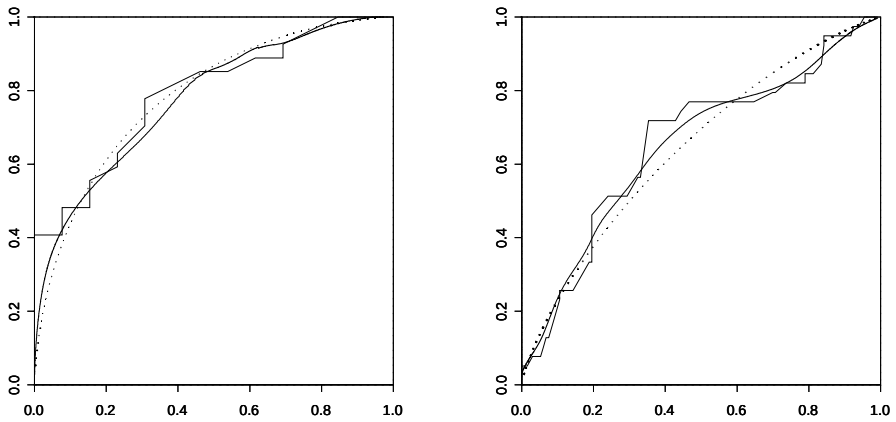
7. Ejemplos

Finalmente, presentamos dos ejemplos prácticos que ilustran la aplicación y funcionamiento de la técnica propuesta. El primero se refiere a la evaluación del volumen de espiración forzada (VEF) como base en el diagnóstico de la neumoconiosis; para esto, se toma una base datos con el VEF para

30 mineros de carbón (27 de ellos padecen la enfermedad). El segundo ejemplo evalúa la capacidad del índice de razonamiento verbal de la *Prueba de Aptitutd Académica* para detectar alumnos con alto rendimiento académico durante su primer año de estudios universitarios. Los resultados se muestran en la tabla siguiente:

PARÁMETRO	VEF	IRV
$\hat{A}$	0.792	0.655
$\tilde{A}$	0.765	0.649
$\overline{A}$	0.786	0.643

Se muestran tres estimadores de A: $\hat{A}$ es el valor de la estadística de Mann-Whintney; $\tilde{A}$ corresponde al estimador suavizado con mínimo error cuadrático medio; y finalmente, $\overline{A}$ es el valor del estimador máximo verosimil para A bajo el modelo binormal. Con base en los resultados, podemos decir que el VEF es más capaz al diagnosticar la neumoconiosis que el IRV detectando estudiantes de alto rendimiento. Las gráficas siguientes muestran los resultados obtenidos al estimar las curvas correspondientes por tres métodos distintos. La gráfica de la izquierda corresponde al Volumen de Espiración Forzada; la de la derecha al Índice de Razonamiwento Verbal. En ambos cuadros, la línea punteada indica el modelo binormal estimado por máxima verosimilitud; las líneas continuas muestran las curvas $\widehat{ROC}$ y $\widetilde{ROC}$. Al comparar los resultados es posible concluir que, en el caso del VEF el modelo binormal muestra un buen ajuste comparable al ofrecido por $\widetilde{ROC}$. Sin embargo, para el caso del Índice de Razonamiento Verbal el modelo binormal muestra un ajuste pobre claramente superado por el método propuesto.



8. Referencias

Bamber, D. (1975). The area above the ordinal dominance graph and the area bellow the ROC curve. *Journal of Mathematical Psychology*, **26**:1-12.

Campbell. M. J., Machin D. (1990). Medical Statistics, a Common Sense Approach; *John Wiley & Sons*: Chichester; pp 36-37.

Cuevas-Covarrubias C. y Copas, J. (2001). Using Gaussian Kernels to estimate the under the ROC; *Research Report* 386*, Department of Statistics, University of Warwick*

Zweig, M. H. y Cambell, G. (1993). Receiver Operating Characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**:561-577.

Estudio del índice extremo en procesos de varianza estocástica[1]

Inder Tecuapetla Gómez[2]
CIMAT A.C.
Graciela González Farías[3]
CIMAT A.C.

1. Introducción

La literatura econométrica llama *hechos estilizados* a un conjunto de características compartidas por muchas series de tiempo financieras. Una de estas serie es la de log-returns (*log-rendimientos*), dada su importancia en los mercados financieros. Desde inicios de los sesentas, se ha incrementado el conocimiento de las causas y consecuencias de los hechos estilizados. El vertiginoso crecimiento en la capacidad de almacenamiento y manejo de la información hace posible que tengamos trabajos como los de Cont (2001), donde hallamos un compendio enciclopédico de los hechos estilizados más comunes. Quizás el hecho estilizado que ha despertado mayor interés en diversos grupos de investigación, es el correspondiente a los cúmulos de variabilidad, *volatility clustering*. Este fenómeno es asociado a la volatilidad de un activo financiero, es decir, la desviación estándar del cambio en el valor de un instrumento financiero con un horizonte temporal específico, junto con los cúmulos que se originan después de un fuerte e inesperado cambio en el nivel de dicho activo. Comentarios como los de Shepard (1996), destacan la importancia de describir el cambio en el tiempo del valor de un activo financiero. El manejo de riesgo, por ejemplo, requiere un estimado de la volatilidad del precio de un activo para poder fijar su valor futuro. A pesar de las críticas recibidas a partir de su presentación por Bollerslev (1986), los modelos GARCH representan la más útil y popular herramienta para describir la evolución en el tiempo del precio de un activo en los mercados financieros. Mikosch y Starica (2000) ponen particular atención en la incapacidad de estos modelos, específicamente del GARCH(1,1), para capturar apropiadamente la dependencia en las colas de las series de rendimientos. Una primera lectura a estas líneas no parecen conectar los cúmulos de variabilidad con la dependencia en las colas de una serie temporal. Sin embargo, la teoría de valores extremos aclara esta duda. Muchos son los textos que estudian la teoría clásica de valores extremos, es decir, la distribución asintótica del máximo de una sucesión de variables aleatorias bajo el supuesto de independencia. En Leadbetter

[1]Trabajo parcialmente apoyado por CONACYT-México y Ministerio de Ciencia y Tecnología-España, a través de los proyectos 45974-F y SEJ2004-04101-ECON, respectivamente
[2]edz_yave@cimat.mx
[3]farias@cimat.mx

et al. (1983) encontramos un estudio fundamental en la extensión de esta teoría al caso de procesos con dependencia débil. El parámetro clave en la extensión de esta teoría es el denominado *índice extremo*, inicialmente presentado por Loynes (1965). La caracterización dada por Leadbetter (1983) de este número, nos permite ligar dependencia en las colas de procesos estacionarios con cúmulos alrededor de un punto de la serie. En este trabajo estudiamos, vía simulación, el comportamiento del índice extremo en el más popular de los procesos de varianza estocástica, los modelos GARCH(1,1). Utilizamos para este propósito el estimador por *niveles moderadamente altos* dado por Olmo (2006) y una de las series temporales más utilizadas en esta clase de estudios en la pasada década, los log-returns de la tasa de cambio *diaria* entre el yen japonés y el dolar americano, JPYUSD.

2. Índice Extremo

El Teorema de Familias de Distribuciones de Valor Extremo puede ser extendido casi naturalmente al caso de procesos con dependencia débil. En Leadbetter (1983) se hace uso de una condición de mezcla apropiada denotada por $D(u_n)$ que permite particionar el proceso entero en bloques asintóticamente independientes.

Con la condición $D(u_n)$ podemos establecer que si $\{X_i\}_{i=1}^n$ es un proceso estrictamente estacionario con función de distribución marginal F, θ un número no negativo, además supongamos que para cada $\tau > 0$ existe una sucesión $\{u_n(\tau)\}$ tal que

$$\lim_{n\to\infty} n(1 - F(u_n(\tau))) = \tau \tag{1}$$

$$\lim_{n\to\infty} \Pr(M_n \leq u_n(\tau)) = e^{-\theta\,\tau}, \tag{2}$$

entonces θ es llamado el índice extremo del proceso $\{X_n\}$. Aquí $M_n = \max\{X_1, \ldots, X_n\}$.

Consideremos el número de excedencias de $u_n(\tau)$ dentro de un bloque de tamaño $r_n = [n/k_n]$. Este evento define la variable aleatoria

$$B_{r_n}^{u_n(\tau)} = \sum_{i=1}^{r_n} \mathbf{1}_{\{X_i > u_n(\tau)\}},$$

de donde obtenemos

$$E[B_{r_n}^{u_n(\tau)} \mid B_{r_n}^{u_n(\tau)} \geq 1] \overset{n\to\infty}{\to} \theta^{-1}. \tag{3}$$

Bajo la condición $D(u_n(\tau))$, y teniendo en mente la estacionariedad del proceso $\{X_i\}_{i=1}^n$, argumentamos que el inverso del índice extremo mide el tiempo promedio de permanencia dentro de la serie de una perturbación de la misma. Esto bien puede interpretarse como el grado de dependencia observada en la serie a partir del evento que provocó la perturbación. Por tanto un estimador del índice extremo proporciona una idea del grado de dependecia en la serie, bien de corto plazo (θ muy cercano a uno) o de largo plazo (θ cercano a cero).

2.1. Estimador por niveles moderadamente altos

Consideremos otra sucesión de umbrales $v_n(\tau) \geq u_n(\tau)$, tal que

$$E[B_{r_n}^{v_n(\tau)} \mid B_{r_n}^{u_n(\tau)} \geq 1] \overset{n\to\infty}{\to} 1. \tag{4}$$

Sean k_n y r_n como arriba, definamos

$$M_j^{k_n} = \text{máx}\,\{X_k : (j-1)\,r_n + 1 \leq k \leq j\,r_n\} \quad j = 1, \ldots, k_n,$$

los máximos 'locales' del proceso $\{X_i\}_{i=1}^n$ a través de la partición de k_n bloques de tamaño r_n. Ahora consideremos los máximos locales (*block maxima*) que excedan el nivel u_n. Así, definimos la variable aleatoria

$$Z_{r_n}^{u_n(\tau)} = \sum_{i=1}^{k_n} \mathbf{1}_{\{M_j^{k_n} > u_n(\tau)\}}.$$

Con este procedimiento se definen sendos *procesos puntuales de posición conglomerada* $N_t^{u_n(\tau)}$ sobre el intervalo $(0, 1]$ consistente de los elementos de $Z_{r_n}^{u_n(\tau)}$ indexado por $t = j/k_n, j = 1, \ldots, k_n$. Cuando n crece, este proceso puntual converge a un proceso de Poisson N con media $\theta\,\tau$. Análogamente, definimos el proceso $N_t^{v_n(\tau)}$, considerando la variable aleatoria $Z_{r_n}^{v_n(\tau)}$, este proceso convergerá a otro proceso de Poisson N' de media $\theta^2\,\tau$. Por tanto otro estimador del índice extremo es

$$\lim_{n\to\infty} \frac{E[N_t^{v_n(\tau)}]}{E[N_t^{u_n(\tau)}]} = \theta. \tag{5}$$

Utilizando los procesos definidos por u_n y v_n, Olmo (2006) define el siguiente estimador,

$$\hat{\theta} = \frac{Z_{r_n}^{v_n}}{Z_{r_n}^{u_n}}.$$

Puede probarse que este estimador es asintóticamente insesgado y consistente.

3. Dependencia en las colas de GARCH(1,1)

JPYUSD: contiene observaciones diarias del Japan Yen/United States Dollar Foreign Exchange Rate. Contamos con 7344 observaciones de esta serie. Ajustamos un GARCH(1,1) a este conjunto de datos, con los parámetros obtenidos simulamos 1000 conjuntos de datos. Estimamos el índice extremo de cada una de estas series por el método de Olmo. De acuerdo a la Figura 1, en el límite, el índice extremo calculado sobre el modelo GARCH(1,1) es mayor que los homólogos en la serie real. Esto implicaría que utilizando el GARCH(1,1), subestimamos el tamaño medio de los conglomerados en series de este tipo. Esta serie fue estudiada por Mikosch y Starica (2000), en aquel estudio se reporta una sobreestimación por parte de los GARCH(1,1). Sin embargo, tal conducta puede obedecer a la escala de medición tomada, los autores mencionados consideraron una escala de tiempo más pequeña a mediciones diarias. En un contexto en donde nos interese estudiar la serie diaria, nuestros resultados parecen sensatos.

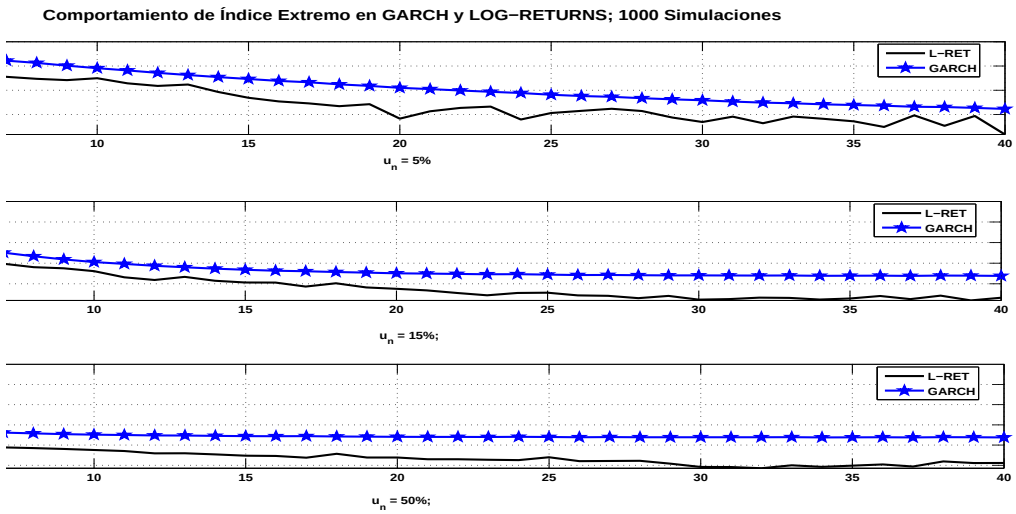


Figura 1: De arriba a abajo: umbrales de 5 %, 15 % y 50 %. Es claro que los estimados del índice extremo calculados sobre los correspondientes GARCH(1,1) ajustados a la serie JPYUSD son mayores a los calculados sobre la serie original. Estudios de simulación previamente hechos, nos sugieren utilizar un número de bloques en el conjunto $\{25, \ldots, 50\}$ para observar convergencia adecuada. Aquí usamos 40 bloques para calcular los estimados.

4. Referencias

Leadbetter, M.R., Lindgren, G. y Rootzen, H. (1983). *Extremes and related properties of random sequences and processes*. Springer-Verlag. New York.

Shepard, N. (1996). *Capítulo 1 en Time Series Models In econometrics, finance and other fields*. Editado por D.R. Cox, D.V. Hinkley y O.E. Barndorff-Nielsen. Chapman & Hall.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*. **31**. 307–327.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*. **1**. 223-236.

Leadbetter, M.R.(1983). Extremes and local dependence in stationary sequences. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*. **65**. 291–306.

Olmo, J. (2006). A new family of estimators for the extremal index. *Working Paper. Department of Economics. City University, London.*

Esta publicación consta de 410 ejemplares y se terminó de imprimir en el mes de octubre de 2007 en los talleres gráficos del
Instituto Nacional de Estadística, Geografía e Informática
Av. Héroe de Nacozari Sur Núm. 2301, Puerta 11, Nivel Acceso
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México