

Memorias del XIX Foro Nacional de Estadística



www.inegi.gob.mx

Memorias del XIX Foro Nacional de Estadística



www.inegi.gob.mx

**DR © 2005, Instituto Nacional de Estadística,
Geografía e Informática
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 XIX Foro Nacional de Estadística

**Impreso en México
ISBN 970-13-4513-4**

Presentación

El XIX Foro Nacional de Estadística se llevó a cabo del 4 al 8 de octubre de 2004 en el Instituto Tecnológico de Estudios Superiores de Monterrey (ITESM), campus Monterrey. Siendo el ITESM quien organizó el evento.

En estas memorias se presentan algunos de los resúmenes de las contribuciones presentadas en este foro. Todos los resúmenes recibidos fueron incluidos sin un proceso de arbitraje, aunque con una detallada revisión.

La Asociación Mexicana de Estadística agradece al ITESM su 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:

Karim Anaya Izquierdo

Alberto Contreras Cristán

Jesús Armando Domínguez Molina

Elida Estrada Barragán

Contenido

Presentación	I
Un análisis estadístico de solvencia	1
<i>Adaya, J.M., Gómez, J., y Mendoza, M.</i>	
Análisis de sensibilidad en regresión: funciones lineales de parámetros	9
<i>Alvarado, V.M., Díaz, J.A. y González-Farías, G.</i>	
Inferencia automática en modelos estimados mediante la norma L_1	15
<i>Aguirre, V. y Domínguez, M.</i>	
Estrategias estadísticas para generar productos confiables	21
<i>Carballo, C.A. y Domínguez, J.</i>	
Aplicación de la metodología seis sigma a una empresa de transporte	29
<i>Centeno, A., Domínguez, J. y González, A.</i>	
Inferencia Bayesiana para la media y desviación estándar de una población normal cuando sólo se observan el tamaño, la media y el rango muestrales	39
<i>de Alba, E., Fernández-Durán, J.J. y Gregorio-Domínguez, M.</i>	
Reducción de dimensión en regresión a través de gráficas y métodos de regresión inversa	47
<i>de la Vega, J.</i>	

Análisis de costos en la experimentación industrial <i>Domínguez, J.</i>	55
Efectos aleatorios en modelos de elección discreta <i>Domínguez, J.R. y Ramos-Quiroga, R.</i>	63
Desagregación y causalidad en una teoría para la predicción económica <i>Espasa, A. y Albacete, R.</i>	69
Muestreo por seguimiento de nominaciones con muestra inicial de sitios seleccionada secuencialmente <i>Félix-Medina, M. y Monjardin, P.</i>	77
Estimación en el modelo de regresión logística en presencia de datos separados y colinealidad <i>Godínez, F. y Ramírez, G.</i>	83
Densidades conjuntas condicionadas a estadísticas suficientes y aplicaciones <i>González-Barrios, J.M., O'Reilly, F. y Rueda, R.</i>	91
Valuación de Opciones Asiáticas mediante Monte Carlo con reducción de varianza <i>Ibarra, V.H. y Saavedra, P.</i>	97
Planes óptimos para pruebas de vida acelerada con esfuerzos escalonados <i>Jiménez, J.C. y Villa, E.</i>	103

Una prueba para exponencialidad basada en la razón de dos estimadores	109
<i>Kantún, M.D. y Villaseñor, J. A.</i>	
Cálculo de estimadores no lineales y de sus varianzas estimadas a partir de información de la muestra del censo nacional de población 2000	115
<i>López, E., Padilla, A., Real, R., Trejo, M. y Eslava, G.</i>	
Una investigación sobre dificultades del proceso de enseñanza aprendizaje relacionadas con distribuciones de probabilidad continuas	121
<i>López, J.A. y Albert, J.A.</i>	
Algoritmos genéticos en la discriminación	129
<i>Montano, A., Cantú, M.</i>	
Estratificación óptima para el índice de desarrollo humano	139
<i>Muller, F., Sánchez, F.J. y Padrón-Corral, E.</i>	
Una generalización de los modelos frailty	145
<i>Nieto-Barajas, L.E. y Walker, S.G.</i>	
La estadística multivariada como análisis de datos de guayule	153
<i>Padrón-Corral, E., Méndez-Ramírez, I., Sánchez, F.J. y Olivares, E.</i>	
Mejora del curado adhesivo QMI505MT en el encapsulado PDIP de los circuitos integrados	161
<i>Pérez-Abreu, R. y Maynes, O.</i>	

Clasificación usando análisis de regresión de Gini: Una alternativa a las máquinas de vector soporte	169
<i>Pérez, B.R. y de los Cobos, S.</i>	
Tablas del tamaño de muestra y la potencia de la prueba UMPI para demostrar la equivalencia de medias de dos distribuciones normales	175
<i>Ramírez, C. y Sotres, D.</i>	
Un procedimiento para selección de los modelos logit mixtos	181
<i>Ruiz, J.J. y González-Farías, G.</i>	
Estudio estadístico de algunas variables climatológicas en una ciudad del estado de Veracruz	187
<i>Galván-Martínez, R., Sánchez-Galván, I.R. y Cruz-Kuri, L.</i>	
Pruebas de bondad de ajuste para el movimiento Browniano	193
<i>Villaseñor, J. A. y González-Estrada, E.</i>	
Identifying sectors of deviations from multinormality	201
<i>von Eye, A. y Anne-Bogat, G.</i>	

Un Análisis Estadístico de Solvencia

J.M. Adaya¹

J. Gómez²

Maestría en Finanzas, Instituto Tecnológico Autónomo de México

M. Mendoza³

División Académica de Estadística, Instituto Tecnológico Autónomo de México

1. Insolvencia Económica

Uno de los principales conflictos en los que se han centrado gran parte de los estudios sobre problemas financieros ha sido la predicción de quiebras. Otro problema que afecta a las empresas es la insolvencia, que difiere del primero en que ésta no implica necesariamente la disolución de la empresa, ya que puede reestructurarse, venderse o fusionarse antes de llegar a su liquidación total. La insolvencia económica se define como la incapacidad de una empresa para hacer frente a sus obligaciones de corto o largo plazo.

La idea central de este trabajo consiste en utilizar las razones financieras de las empresas para clasificarlas en solventes o insolventes con uno o varios períodos de tiempo de anticipación. Se consideran los estados financieros de 91 empresas que cotizan en la Bolsa Mexicana de Valores y de las que se dispone de información suficiente hasta el año 2002, la información se obtuvo del sistema de información *Bloomberg*.

2. Base de Datos

Para obtener un modelo representativo de las empresas del país se seleccionaron, para el ajuste, a las empresas que satisficieran dos criterios, uno financiero y otro técnico. En el primer caso, se buscó que las empresas mostraran información homogénea o típica. En el segundo caso, debido a que con la información se construirán las variables del modelo mediante un Análisis de Componentes Principales (ACP), se buscó que las empresas fueran

¹amig29@hotmail.com

²julgomez@alumnos.itam.mx

³mendoza@itam.mx

RAZONES FINANCIERAS CONSIDERADAS				Índices	Abreviatura
Razón	Abreviatura	Definición	Unidades		
Razón de deuda a Capital	RDC	Pasivo Total/Capital Contable	Porcentaje		
Razón de Apalancamiento	RAP	Pasivo Total/Activo Total	Porcentaje		
Razón de Cobertura de Intereses	RCI	Utilidad Operativa/Gastos por Intereses	Voces		
Razón de flujo de efectivo a Deuda	RFED	Flujo de Efectivo Operativo/Pasivo Total	Voces		
Razón de Circulante	RCI	Activo Circulante/Pasivo Circulante	Voces		
Prueba Ácida	PA	(Activo Circulante-Inventario)/Pasivo Circulante	Voces		
Razón de Efectivo	RE	(Efectivo+Inversiónes Temporales)/Pasivo Circulante	Voces		
Razón de Capital de Trabajo a Efectivo	RCTA	Capital de trabajo/Activo Total	Porcentaje		
Razón de Posición de Efectivo	RPE	(Efectivo+Inversiónes Temporales)/Activo Circulante	Porcentaje		
Rotación de Cuentas por Cobrar	RCC	Ventas Netas/Promedio de Cuentas por Cobrar	Voces		
Rotación de Inventarios	RI	Costo de Ventas/Inventario Promedio	Voces		
Rotación de Activo	RAP	Ventas Netas/Activo Total	Voces		
Margen Operativo	MO	Utilidad Operativa/Ventas Netas	Porcentaje		
Rentabilidad del Activo	RDA	Utilidad Operativa/Activo Total	Porcentaje		
Rentabilidad del Capital	ROE	Utilidad Operativa/Capital Contable	Porcentaje		
Rentabilidad Acumulada	RAC	Utilidades Acumuladas/Activo Total	Porcentaje		

The diagram illustrates the classification of financial ratios into four main categories: Apalancamiento (ACP), Liquidez (ACP), Rentabilidad (ACP), and Antigüedad. Each category is associated with a specific index and its abbreviation:

- Apalancamiento ACP:** Estructura de Capital (82.4%)
- Liquidez ACP:** Ausencia de Liquidez (71.3%)
- Rentabilidad ACP:** Eficiencia de Activos (92.7%)
- Antigüedad:** Rentabilidad Acumulada (RAC)

Figura 1: Se aplicó Análisis de Componentes Principales (ACP) dentro de las clases de razones financieras de apalancamiento, liquidez y rentabilidad. En el lado derecho del cuadro se muestra, entre paréntesis, el porcentaje de variación que de las variables originales preservaron los índices seleccionados

capaces de proporcionar índices interpretables y con sentido financiero. Las empresas que no cumplieron con alguno de los dos criterios fueron 24, y se agruparon dentro de un grupo denominado de Empresas Atípicas. El análisis se centra en las 67 empresas restantes, con las cuales se formó al azar dos grupos: El Grupo de Ajuste del modelo, con 39 empresas; y el Grupo de Pronóstico, con 28. Los integrantes de cada grupo pueden consultarse en el cuadro 2.

Se utilizó un modelo de regresión logístico cuya variable respuesta, es una variable binaria que representa de forma indirecta la situación de insolvencia de una empresa. ésta variable utiliza la información contable correspondiente al año 2002 y se define por medio de la utilidad neta de cada empresa como se muestra en la ecuación 2.1

$$Y = \begin{cases} 1, & \text{Utilidad Neta} \leq 0 \quad (\text{Insolvencia}) \\ 0, & \text{Utilidad Neta} > 0 \quad (\text{Solvencia}) \end{cases} \quad (2,1)$$

Se consideraron 16 razones financieras calculadas a partir de los estados financieros que reportaron las empresas en el año 2001, éstas se definen del lado derecho de la Figura 1. Las razones financieras se construyen a partir de información contable y se agrupan en clases muy conocidas en el mundo financiero: Apalancamiento, Liquidez, Rentabilidad y Antigüedad. La Figura 1, muestra estas razones, sin embargo, se puede definir un mayor número, cada una de ellas describiendo aspectos diferentes, por lo que las correspondientes unidades de medición pueden ser diferentes. Considerando lo anterior, se realizó ACP dentro las cuatro clases mencionadas y, como lo muestra la Figura 1, de 15 razones financieras se obtuvieron 6 índices. De los índices obtenidos más la variable de la clase de antigüedad, se seleccionaron como variables explicativas del modelo aquellas que pudieran distinguir la ubicación de los

subconjuntos de datos de empresas solventes e insolventes, esto se realizó por medio de una prueba no paramétrica de medianas con 95 % de confianza. Las variables seleccionadas fueron las denominadas como EDC (X_1), DEUDA (X_2) y $EO(X_3)$.

3. Ajuste del Modelo

El modelo inicial se especificó considerando una distribución Bernoulli para la variable respuesta, ecuación 3.1, y la función de verosimilitud indicada en la ecuación 3.2

$$Y_i \sim \text{Bernoulli} (q_i) \quad (3,1)$$

$$p(Y|a) = \prod_{i=1}^n [q_i]^{Y_i} [1 - q_i]^{1-Y_i} \quad (3,2)$$

En estas expresiones, q representa la probabilidad de insolvencia y n el número de empresas dentro del Grupo de Ajuste. Se especificó una liga logística, ecuación 3.3, donde el predictor esta dado como lo indica la ecuación 3.4

$$q_i = F(x'_i \alpha) = e^{x'_i \alpha} / (1 + e^{x'_i \alpha}) \quad (3,3)$$

$$x'_i \alpha = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} , \quad i = 0, 1, \dots, 39 \quad (3,4)$$

Para el análisis del modelo anterior se consideró un enfoque Bayesiano, en el cual se asignaron distribuciones iniciales uniformes para los parámetros (coeficientes) del modelo. El ajuste del modelo, se realizó utilizando el programa WinBugs y se realizaron diversas pruebas de convergencia en la obtención de las muestras de cada parámetro. Para la selección de variables se eliminaron aquellas cuyos coeficientes, a posteriori, incluían el valor cero en la región de máxima densidad 0.95.

4. Resultados

Se obtuvo una muestra de tamaño 50,000 de la distribución final conjunta de los coeficientes del modelo final. Las funciones de densidad de probabilidad marginal final de los dos

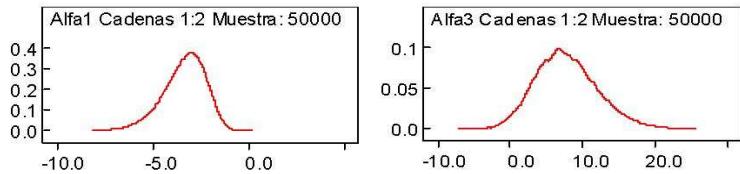


Figura 2: Funciones de densidad final de los parámetros del modelo final ajustado

PARÁMETRO	PROMEDIO	DESV. EST.	CUANTIL 2.5%	CUANTIL 5%	MEDIANA	CUANTIL 95%	CUANTIL 97.5%
α_1	-3.375	1.121	-5.925	-5.424	-3.248	-1.768	-1.537
α_3	7.818	4.307	0.213	1.273	7.502	15.430	17.200

Cuadro 1: Estadísticos descriptivos de las densidades marginales de los parámetros del modelo final ajustado

parámetros estimados, que corresponden a las variables X_1 y X_3 , se presentan en la Figura 2, y el Cuadro 1 muestra las estadísticas descriptivas correspondientes.

La probabilidad de insolvencia, para un vector de covariables \underline{x} , está dada por la ecuación 4.1

$$q = F(x'\alpha) = e^{x'\alpha} / (1 + e^{x'\alpha}) \text{ donde } x'_i\alpha = \alpha_1 X_{1i} + \alpha_3 X_{3i} \quad (4.1)$$

A partir de la distribución final de α se puede obtener, para cada \underline{x} , una distribución para q , en particular, si una de las covariables se fija en su valor medio, se puede observar el impacto de la otra en la distribución, esto se ilustra en la Figura 3.

La variable X_1 , cuyo coeficiente es α_1 , corresponde al primer componente principal del grupo de apalancamiento; se le llamó Estructura de Capital (EDC), y mientras mayor es el valor de este índice, para una empresa, menor es la importancia de la deuda como fuente de su financiamiento. Manteniendo a la otra variable constante, incrementos en esta variable se asocian con disminuciones en la probabilidad de insolvencia. Este resultado es consistente con lo que se podría esperarse desde una perspectiva intuitiva, ya que una empresa con poco endeudamiento, como proporción de su financiamiento total, muy probablemente enfrente menos problemas de insolvencia. La sigmoide del lado izquierdo de la gráfica anterior, muestran cinco diagramas de caja y brazos para diferentes valores de la variable EDC (-1,-0.5,0,0.5 y 1), con los cuales es posible apreciar la dispersión de la probabilidad de insolvencia. Es claro que las regiones con mayor dispersión se encuentran alrededor de los valores -0.5 y 0.5. Por otro lado, la dispersión es menor en las colas y es mínima alrededor del cero. En el histograma se aprecia que la mayoría de las empresas mexicanas tienen valores correspondientes a EDC dentro de la región (-1.3, 1.1] y alrededor del 25 % se encuentran en la zona de variabilidad mínima. Los resultados sugieren que, aproximadamente el 15 % de las empresas

no enfrentan, de forma clara, problemas de insolvencia, al menos respecto a esta variable. Sin embargo, las gráficas sugieren que un número considerable de empresas, aquellas en el intervalo $(-1.3, 1.1]$, podrían prevenir problemas de insolvencia mediante la mejora de las razones financieras dentro de este índice.

Por otro lado, el coeficiente α_3 , pondera el peso de la variable EO, que es el primer componente del grupo de rentabilidad y es un indicador de la eficiencia operativa de las empresas. En particular, el indicador aumenta su valor cuando empeora la eficiencia operativa de la empresa. En este caso, manteniendo constante a la otra variable, incrementos en EO se asocian con una mayor probabilidad de insolvencia. Lo anterior tiene sentido, ya que una empresa ineficiente, en términos operativos, debiera ser más susceptible de enfrentar problemas de insolvencia. Como se puede corroborar en el lado derecho de la Figura 3, el incremento en EO tiene asociado un aumento en la probabilidad de que la empresa enfrente problemas de insolvencia económica. La sigmoide se presenta con ocho diagramas de caja y brazos para diferentes valores de la variable $(-0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4 \text{ y } 0.5)$ e indican, como en el caso anterior, menor dispersión cerca de la cola derecha y alrededor del cero. El histograma muestra que la mayor parte de las empresas, aproximadamente el 80%, se encuentran en las clases $(-0.2, 0]$ y $(0, 0.2]$. Aunque la mayoría se encuentra en una zona de poca variabilidad,

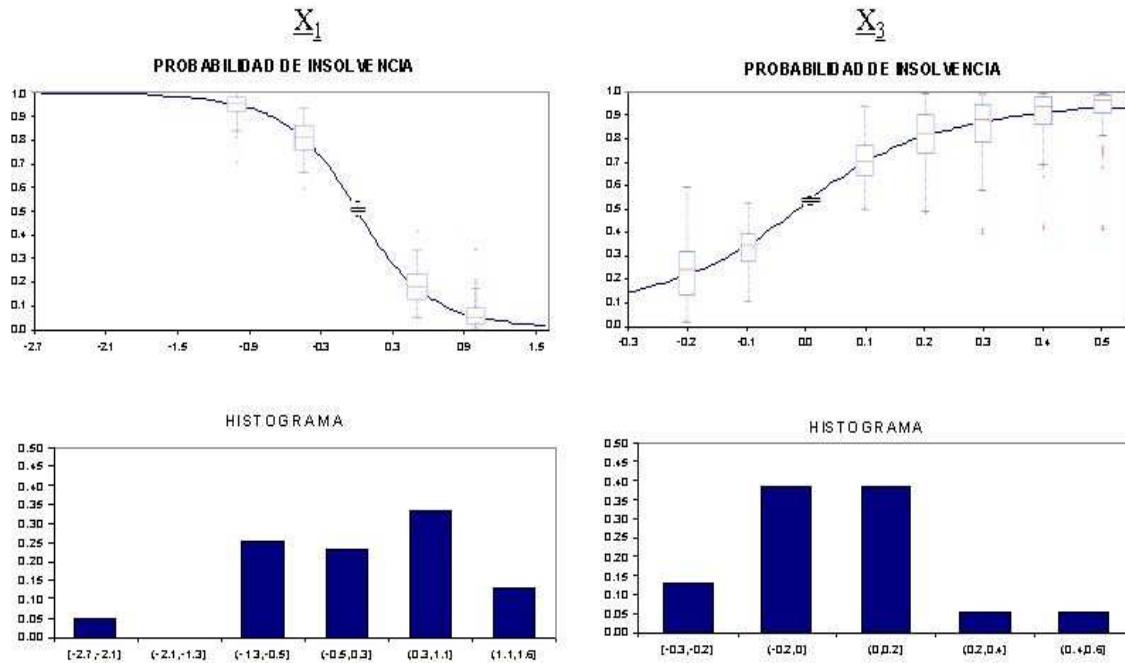


Figura 3: Efectos de la estructura de Capital X_1 y la Eficiencia Operativa X_2

GRUPO DE PRONÓSTICO					
EMPRESA	PROBAILIDAD	EMPRESA	PROBAILIDAD	EMPRESA	PROBAILIDAD
ALFAA	0.9178	GEOB	0.0854	NUTRISA	0.0175
ALSEA	0.0149	GEUPECB	0.0000	PARRAS	0.9986
BACHOCO	0.0000	GMODERN	0.0029	POSADASL	0.5009
BAFARB	0.0002	HERDEZ	0.1175	PYPB	1.0000
BENAVIDESB	0.9940	HILASALA	0.0225	REGIOEMB	0.6939
CEL V	0.9737	IASASA	0.9768	SANLUISA	0.9982
CIEB	0.1513	IMSAUBC	0.1801	SAVIAA	0.9695
COVERA	0.1706	MEXCHEMA	0.0000	TVAZTCPO	0.4316
ELEKTRA	0.3723	MINSAA	0.9967		
GCORVIU	0.9561	MOVILAB	1.0000		

GRUPO DE EMPRESAS ATÍPICAS					
EMPRESA	PROBAILIDAD	EMPRESA	PROBAILIDAD	EMPRESA	PROBAILIDAD
APASCO	0.0001	EKCO	0.0634	LIVERPOL1	0.0012
ARA	0.0003	FEMSAUBD	0.0003	MASECAB	0.0000
ARGOSB	0.0001	FRAGUAB	0.0007	NADROB	0.0000
BIMBOA	0.0003	GAMB	1.0000	QUM MAA	0.9714
CONTAL	0.0000	GMODELLOC	0.0000	SORIANA	0.0000
DIXON	0.1089	HOGARB	0.9880	SYNKROA	0.9999
ECE	1.0000	KIMBERA	0.0001	TELMEXL	0.0062
EDOARDOB	0.0000	KOFL	0.0634	WALMEXC	0.0000

Cuadro 2: Probabilidad de insolvencia para el grupo de pronóstico y el grupo de empresas atípicas

todas ellas pueden prevenir la probabilidad de enfrentar problemas de insolvencia por medio de la mejora del rubro de rentabilidad.

La probabilidad predictiva esta dada por la ecuación 4.2 y puede aproximarse con la muestra conjunta obtenida para los parámetros, por medio de la expresión 4.3, de forma que puede obtenerse un pronóstico de la probabilidad de insolvencia para cada una de las empresas de la muestra.

$$\tilde{q}(\underline{x}) = \int q(\underline{x}, \underline{\alpha} | \text{Datos}) d\underline{\alpha} = \int p(Y = 1 | \underline{x}, \underline{\alpha}) p(\underline{\alpha} | \text{Datos}) d\underline{\alpha} \quad (4.2)$$

$$\int p(Y = 1 | \underline{x}, \underline{\alpha}) p(\underline{\alpha} | \text{Datos}) d\underline{\alpha} = \frac{1}{N} \sum_{i=1}^N p(Y = 1 | \underline{x}, \underline{\alpha}_i) \quad (4.3)$$

Se realizaron pronósticos para el Grupo de Pronóstico así como para el Grupo de Empresas Atípicas, los cuales se presentan en el Cuadro 2.

Con las probabilidades predictivas, se puede definir una regla de clasificación, la cual se podría definir con criterios de optimalidad. Como ejemplo se consideró: Solvente ($\tilde{q} \leq 0,3$); Insolvente ($\tilde{q} \geq 0,7$); Indecisión en otro caso. Los resultados se presentan en el Cuadro 3

Referencias

Ambrose J. M. y Carroll A. M. (1994). Using Best's Ratings in Life Insurer Insolvency Prediction. *The Journal of Risk and Insurance*, **61**, No. 2, Tort Reform Symposium, pp. 317-327.

BarNiv R. y Hathorn J. (1999). Confidence Intervals for the Probability of Insolvency in the Insurance Industry. *The Journal of Risk and Insurance*, **66**, No. 1 pp. 125-137.

BarNiv R. y Hershberger R. A. (1990). Classifying Financial Distress in the Life Insurance Industry. *The Journal of Risk and Insurance*, **57**, No. 1, pp. 110-136.

Bernardo J.M. y Smith A.F.M. (1994). *Bayesian Theory*. John Wiley & Sons, LTD.

Congdon P. (2001). *Bayesian Statistical Modeling*. Wiley Series in Probability and Statistics.

Gamerman D. (1997). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman & Hall.

Gelman A., Carlin B. J., Stern S. H. y Rubin B. D. (1995). *Bayesian Data Analysis*. Chapman & Hall.

Spiegelhalter D., Thomas A., Best N. y Lunn D. (2003). *WinBUGS User manual version*

SOLVENTES (Total de empresas: 18)			INSOLVENTES (Total de empresas: 10)		
	Empresas	Porcentaje		Empresas	Porcentaje
Clasificación Correcta	12	66.67%	Clasificación Correcta	9	90.00%
Clasificación Incorrecta	2	11.11%	Clasificación Incorrecta	1	10.00%
Indecisión	4	22.22%	Indecisión	0	0.00%

BASE DE DATOS COMPLETA (Total de empresas: 28)		
	Empresas	Porcentaje
Clasificación Correcta	21	75.00%
Clasificación Incorrecta	3	10.71%
Indecisión	4	14.29%

Cuadro 3: Clasificación de los pronósticos para las empresas solventes e insolventes, así como para la base de datos completa

1.4, <http://www.mrc-bsu.cam.ac.uk/bugs>

Zellner A. (1997). *Bayesian Analysis in Econometrics and Statistics*. Cheltenham, UK: Edward Elgar Publishing Co.

Zellner A. y Rossi P. (1984). Bayesian Analysis of Dichotomous Quantal Response Models. *Journal of Econometrics*, **25**, pp. 365-393.

Análisis de Sensibilidad en Regresión: Funciones lineales de parámetros

Víctor M. Alvarado Castro¹

Centro de Investigación en Matemáticas, A.C. y Universidad Autónoma de Guerrero.

José A. Díaz García²

Universidad Autónoma Agraria Antonio Narro.

Graciela González Farías³

Centro de Investigación en Matemáticas, A.C.

1. Detectando Observaciones Influyentes

Considere el modelo de regresión multivariado

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

donde $\mathbf{Y} : n \times p$ es matriz de respuestas, $\mathbf{X} : n \times q$ es la matriz de regresión de rango q , $\boldsymbol{\beta} : q \times p$ es la matriz de parámetros de regresión y $\boldsymbol{\varepsilon} : n \times p$ es la matriz de errores con distribución $\mathcal{N}_{n \times p}(0, \boldsymbol{\Sigma} \otimes I_n)$, con $\boldsymbol{\Sigma}$ definida positiva. Ahora, considere el modelo lineal general multivariado modificado, el cual se obtiene de (1) eliminando el i -ésimo renglón de \mathbf{Y} , \mathbf{X} y $\boldsymbol{\varepsilon}$, esto es, eliminando la i -ésima observación. Sean $\widehat{\boldsymbol{\beta}}_{(i)}$ y $\widehat{\boldsymbol{\Sigma}}_{(i)}$ los correspondientes estimadores de verosimilitud máxima para este modelo. Así, setiene que

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i\widehat{\boldsymbol{\varepsilon}}'_i}{1 - p_{ii}}, \quad (2)$$

donde p_{ii} es el i -ésimo elemento diagonal del proyector ortogonal $\mathbf{P} = \mathbf{XX}^-$, con A^- el inverso de Moore-Penrose de A y $\widehat{\boldsymbol{\varepsilon}} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = (\mathbf{I} - \mathbf{P})\mathbf{Y}$, ver Chatterjee y Hadi (1988) y Díaz-García y González-Farías (2004).

Teorema 1.1. *Considere el modelo lineal general multivariado dado en (1). Bajo el supuesto de normalidad de la matriz de errores y considerando las matrices $\mathbf{N} : l \times q$ y $\mathbf{M} : p \times s$, de*

¹alvarado@cimat.mx

²jadiaz@uaaan.mx

³farias@cimat.mx

rangos l y s respectivamente. Entonces, la distancia de Cook modificada, \mathcal{AC}_i , para detectar una observación outlier en funciones lineales de los parámetros $\beta, \mathbf{N}\beta\mathbf{M}$, puede ser escrita como:

$$\mathcal{AC}_i = \begin{cases} \text{vec}'(\mathbf{N}(\widehat{\beta} - \widehat{\beta}_{(i)})\mathbf{M}) \left(\widehat{\text{Cov}} \left(\text{vec}(\mathbf{N}(\widehat{\beta} - \widehat{\beta}_{(i)})\mathbf{M}) \right) \right)^{-1} \text{vec}(\mathbf{N}(\widehat{\beta} - \widehat{\beta}_{(i)})\mathbf{M}), \\ (1 - p_{ii})^{-1} \text{vec}'(\mathbf{Y})(\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}\mathbf{M}' \otimes \mathbf{P}_i\mathbf{P}_i') \text{vec}(\mathbf{Y}), \\ (1 - p_{ii})^{-1} \text{tr}(\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}\mathbf{M}'\mathbf{Y}'\mathbf{P}_i\mathbf{P}_i'\mathbf{Y}), \\ (1 - p_{ii})^{-1} (\mathbf{M}'\widehat{\epsilon}_i)'(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}(\mathbf{M}'\widehat{\epsilon}_i). \end{cases} \quad (3)$$

Sea $I = \{i_1, i_2, \dots, i_k\}$ un conjunto de tamaño k de $\{1, 2, \dots, n\}$, tal que $(n - k) \geq q$. Ahora, para el modelo(1) denote $\mathbf{X}_{(I)}$, $\mathbf{Y}_{(I)}$ y $\boldsymbol{\varepsilon}_{(I)}$ las matrices de regresión, de datos y de errores, respectivamente, obtenidas después de borrarlas correspondientes observaciones de acuerdo a los subíndices I y sean $\widehat{\beta}_{(I)}$ y $\widehat{\Sigma}_{(I)}$ los correspondientes estimadores de verosimilitud máxima para este modelo.

Teorema 1.2. *Considere el modelo lineal general multivariado dado en (1). Bajo el supuesto de normalidad de la matriz de errores y considerando las matrices $\mathbf{N} \in \mathbb{R}^{l \times q}$ y $\mathbf{M} \in \mathbb{R}^{p \times s}$, de rangos l y s , respectivamente. Entonces la distancia de Cook modificada, \mathcal{AC}_I , para detectar k outliers en funciones lineales de los parámetros $\beta, \mathbf{N}\beta\mathbf{M}$, puede ser escrita como:*

$$\mathcal{AC}_I = \begin{cases} \text{vec}'(\mathbf{N}(\widehat{\beta} - \widehat{\beta}_{(I)})\mathbf{M}) \left(\widehat{\text{Cov}}(\text{vec}(\mathbf{N}(\widehat{\beta} - \widehat{\beta}_{(I)})\mathbf{M})) \right)^{-1} \text{vec}(\mathbf{N}(\widehat{\beta} - \widehat{\beta}_{(I)})\mathbf{M}) \\ \text{vec}'(\widehat{\beta} - \widehat{\beta}_{(I)}) (\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}\mathbf{M}' \otimes \mathbf{N}'((\mathbf{N}\mathbf{R}\mathbf{N}')^{-1}\mathbf{N})) \text{vec}(\widehat{\beta} - \widehat{\beta}_{(I)}) \\ \text{tr}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1} \left(\mathbf{N}(\widehat{\beta} - \widehat{\beta}_{(I)})\mathbf{M} \right)' ((\mathbf{N}\mathbf{R}\mathbf{N}')^{-1} \left(\mathbf{N}(\widehat{\beta} - \widehat{\beta}_{(I)})\mathbf{M} \right)) \\ \text{tr}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1} \mathbf{M}'\widehat{\epsilon}_I'(\mathbf{I}_k - \mathbf{P}_I)^{-1}\widehat{\epsilon}_I \mathbf{M}. \end{cases} \quad (4)$$

2. Distribuciones Asociadas con las Distancias Modificadas

Las distribuciones exactas de \mathcal{AC}_i y \mathcal{AC}_I , las cuales se obtienen utilizando estadísticos pivotales, se muestran en el siguiente resultado.

Teorema 2.1. Considera el modelo lineal general multivariado (1) y las definiciones de \mathcal{AC}_i y \mathcal{AC}_I . Suponiendo que $\boldsymbol{\varepsilon} \sim \mathcal{N}_{n \times p}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$, se tiene,

1.

$$\frac{\mathcal{AC}_i}{n - q} \sim \beta(s/2, (n - q - s)/2), \quad (5)$$

donde $\beta(s/2, (n - q - s)/2)$ denota una distribución beta centrada con parámetros $s/2$ y $(n - q - s)/2$.

2.

$$\frac{\mathcal{AC}_I}{n - q} \sim \mathcal{P}(w, m, h), \quad (6)$$

donde $\mathcal{P}(w, m, h)$ denota la distribución centrada para el estadístico Pillai con parámetros w , m , y h , ver Seber (1984) o Rencher (1995). Los parámetros son definidos como $w = \min(s, k)$, $m = (|s - k| - 1)/2$ y $h = (n - q - s - 1)/2$.

Del Teorema 2.1, dado un nivel de significancia α , podemos escribir las siguientes reglas de decisión:

1. \mathbf{Y}_i , $i = 1, 2, \dots, n$, es una observación influyente en $\mathbf{N}\boldsymbol{\beta}\mathbf{M}$ si

$$\frac{\mathcal{AC}_i}{n - q} \geq \beta_{\alpha:s/2,(n-q-s)/2}, \quad (7)$$

donde $\beta_{\alpha:s/2,(n-q-s)/2}$ es el correspondiente $\alpha - percentil$ superior de una distribución β con parámetros $s/2$ y $(n - q - s)/2$.

2. $\mathbf{Y}_{i_1}, \mathbf{Y}_{i_2}, \dots, \mathbf{Y}_{i_k}$, es un conjunto de observaciones influyentes en $\mathbf{N}\boldsymbol{\beta}\mathbf{M}$ si

$$\frac{\mathcal{AC}_I}{n - q} \geq P_{\alpha:w,m,h}, \quad (8)$$

donde $P_{\alpha:w,m,h}$ es el correspondiente $\alpha - percentil$ superior de una distribución Pillai con parámetros $w = \min(s, k)$, $m = (|s - k| - 1)/2$ y $h = (n - q - s - 1)/2$.

3. Aplicación

Ilustramos el uso de las pruebas exactas en la Sección 2. Se utilizaron los datos presentados en Potthoff y Roy (1964). Los datos consisten en cuatro mediciones de la distancia (en milímetros) del centro de la pituitaria a la fisura maxilar, realizadas en edades de 8, 10, 12 y 14 años en 11 niñas y 16 niños. El modelo a considerar es:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde $\mathbf{Y} : 27 \times 4$, es la matriz respuesta con y_{igt} = distancia de la pituitaria a la fisura maxilar del i individuo, del grupo g (niñas=1 o niños=2) al tiempo t , con tiempos $t = (t_1, t_2, t_3, t_4) = (8, 10, 12, 14)$. $\mathbf{X} : 27 \times 3$ es la matriz de covariables, su primer columna está formada por unos correspondientes al intercepto, las dos columnas restantes tienen unos y ceros los cuales indican a que grupo g pertenece el individuo i , y $\boldsymbol{\beta} : 3 \times 5$ es la matriz de parámetros. Un

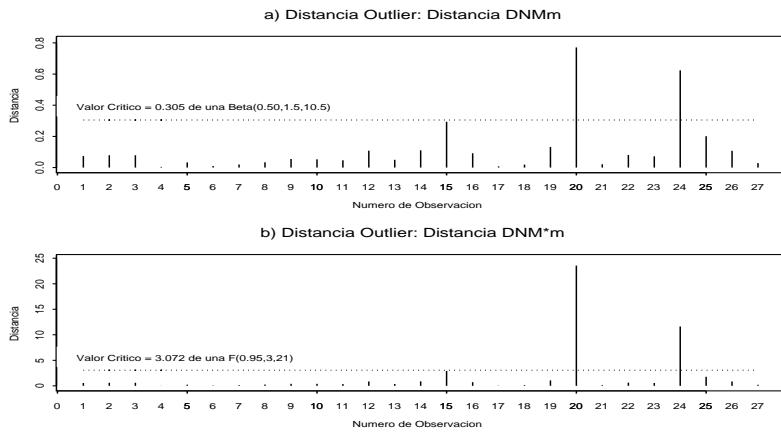


Figura 1: Identificación de observaciones influyentes basada en a) la distancia DNM_{m_i} y b) la distancia $DNM_{m_i}^*$.

problema, que surge de manera natural, es ver si existe diferencia significativa en la distancia de la pituitaria a la fisura maxilar en los diferentes tiempos; esto se puede investigar estimando las funciones paramétricas $\boldsymbol{\beta}\mathbf{M}$, donde la primera columna contrastaría las distancias para las edades 8 y 10, la segunda contrastaría las edades 10 y 12, y la última las edades 12 y 14. Se utilizaron las métricas obtenidas en el Teorema 1.2 para probar si las observaciones 15, 20 y 24 son conjuntamente influyentes en la combinación lineal $\hat{\boldsymbol{\beta}}\mathbf{M}$, ver Figura 1. Ambas pruebas

Métrica	Estadístico de prueba	Valor Crítico $1 - \alpha = 0.95$
$DNM_{m_I} = 1.677703$	10.15023	2.012705
$DNM_{m_I}^* = 5.854709$	13.44415	2.034774

Tabla 1: Dos métricas para detectar un conjunto de observaciones influyentes en la combinación lineal $\hat{\beta}\mathbf{M}$.

consideran a estas observaciones como conjuntamente influyentes. La Tabla 1 resume los resultados obtenidos.

Referencias

- Besley, D., Kuh, E., and Welsch, R. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York.
- Caroni, C. (1987). Residuals and influence in the multivariate linear model, *The Statistician*, **36**, 365-370.
- Cook, R. D., and Weisberg, S. (1982). *Residual and Influence in Regression*, Chapman and Hall, London.
- Chatterjee, S., and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*, John Wiley & Sons. New York.
- Díaz-García, J.A., and González- Farías, G. (2004). A note on the Cook's distance, *J. Statist. Plan. Inference*, **120**, 119-136.
- Díaz-García, J.A., Galea, M., and Leiva- Sánchez, V. (2001). Influence diagnostics for elliptical regression linear models, *Comm. Statist. T. M.*, **32(2)**, 625-641.
- Díaz-García, J.A., Leiva- Sánchez, V., and Galea, M. (2002). Singular elliptic distribution: density and applications, *Comm. Statist. T. M.*, **31(5)**, 661-682.
- Draper, N., and Smith, H. (1981). *Applied Regression Analysis*, (2nd ed.), John Wiley & Sons, New York.

Galea, M., Paula, G., and Bolfarine, H. (1997). Local influence in elliptical linear regression models, *The Statistician*, **46**, 71-79.

Fang, K. T., and Anderson T. W. (1990). *Statistical Inference in Elliptically Contoured and Related Distributions*, Allerton Press Inc., New York.

Fang, K. T., and Zhang, Y. T. (1990). *Generalized Multivariate Analysis*, Science Press, Beijing, Springer-Verlang.

Gupta, A. K., and Varga, T. (1993). *Elliptically Contoured Models in Statistics*, Kluwer Academic Publishers, Dordrecht.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika*, **51**, 313-326.

Rencher, A. C. (1995). *Methods of Multivariate Analysis*, John Wiley & Sons, New York.

Seber, G. A. F. (1984). *Multivariate Observations*, John Wiley & Sons, New York.

Inferencia Automática en Modelos Estimados Mediante la Norma L_1

Víctor Aguirre Torres¹

Departamento de Estadística, Instituto Tecnológico Autónomo de México

Manuel Domínguez Toribio²

Departamento de Estadística y Centro de Investigación Económica, Instituto Tecnológico Autónomo de México

1. Introducción

En el presente trabajo se muestra que es posible hacer inferencia en modelos de regresión estimados por medio de la minimización de la norma L_1 sin necesidad de recurrir a la elección de un parámetro de suavizamiento. Lo anterior se presenta como una aplicación de una propiedad distribucional novedosa del remuestreo.

Para describir la situación, por simplicidad examinaremos el caso de regresión lineal simple, aunque los resultados tienen validez para regresión lineal múltiple. Consideramos que $\mathcal{Z}_n = \{(Y_i, X_i)\}_{i=1}^n$ es un conjunto de observaciones que están relacionadas por la siguiente ecuación

$$Y_i = \beta_1 + \beta_2 X_i + U_i, \quad i = 1, 2, \dots, n.$$

Bajo condiciones de regularidad, las cuales por restricciones de espacio no incluiremos aquí, Koenker y Bassett (1978) muestran que el estimador percentil de $\beta = (\beta_1, \beta_2)^T$ tiene una distribución asintótica normal bivariada. El estimador L_1 es un caso particular cuando el percentil es del 50 % y por lo tanto

$$\sqrt{n} (\hat{\beta}_2 - \beta_2) \xrightarrow{d} N(0, \sigma^2), \quad (1)$$

con

¹aguirre@itam.mx

²madt@itam.mx

$$\sigma^2 = [f_U^2(F_U^{-1}(1/2)) s_{xx}/4]^{-1}$$

donde F_U y f_U son la función de distribución y de densidad de U respectivamente y $s_{xx} = \sum(X_i - \bar{X})^2$. De lo anterior se ve que para llevar a cabo inferencia sobre los parámetros de interés del modelo se requiere de la estimación de f_U lo que a su vez necesita de la elección de un parámetro de suavizamiento, ya que típicamente la estimación de una densidad tiene la forma

$$\hat{f}_U(z) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{\hat{U}_i - z}{h_n}\right)$$

siendo K la función medular, h_n el parámetro de suavizamiento y \hat{U}_i $i = 1, \dots, n$, los residuos del modelo ajustado. Para una discusión acerca de lo controversial que puede resultar el desempeño de este tipo de estimadores, el lector puede consultar Bowman (1985) y Cao, Cuevas y González-Manteiga (1994).

Como una alternativa al problema de suavizamiento en inferencia en regresión percentil, se puede usar el resultado de Hahn (1995), quien demostró que bajo las mismas condiciones de regularidad que dan lugar a (1) se tiene que

$$\sqrt{n} (\hat{\beta}_2^* - \hat{\beta}_2) \rightarrow_{d^*} N(0, \sigma^2), \quad \text{en probabilidad,} \quad (2)$$

donde $\hat{\beta}_2^*$ es el estimador de remuestreo que se obtiene de estimar el parámetro del modelo cuando se genera una pseudo muestra con $Y_i^* = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{U}_i^*$, $i = 1, 2, \dots, n$. Los pseudo errores \hat{U}_i^* se obtienen remuestreando con igual probabilidad y con reemplazo de los residuos \hat{U}_i $i = 1, \dots, n$. Es conveniente clarificar el sentido condicional de la convergencia en (2), lo que quiere decir es que si $T_n^* = \sqrt{n}(\hat{\beta}_2^* - \hat{\beta}_2)$, $F_{T_n^*|Z_n}(r)$ es la función de distribución acumulada de T_n^* condicional en la muestra observada y $N(r, 0, \sigma^2)$ es la distribución normal acumulada evaluada en r , entonces

$$\left| F_{T_n^* | \mathcal{Z}_n} (r) - N(r, 0, \sigma^2) \right| \rightarrow_p 0 \text{ cuando } n \rightarrow \infty$$

De aquí se ve que por ejemplo, usando el remuestreo crudo, con un número grande de replicaciones de remuestreo (B) se puede hacer inferencia sobre β_2 . Esto se obtendría calculando $\{\widehat{\beta}_{2,b}^*\}$ para $b = 1, \dots, B$ y después, para conseguir un intervalo de confianza a nivel α por ciento, se estimaría como

$$\left(\widehat{\beta}_{2,(r)}^*, \widehat{\beta}_{2,(s)}^* \right),$$

calculando los respectivos percentiles de $\alpha/2$ y $1 - \alpha/2$ porciento. Un hecho bien conocido en la literatura es que es superior usar el llamado remuestreo-t que hace uso de una cantidad pivotal en función del parámetro de interés, la cual en este caso sería

$$t_b^* = \frac{\sqrt{n}(\widehat{\beta}_{2,b}^* - \widehat{\beta}_2)}{\widehat{\sigma}_b^*}$$

para $b = 1, \dots, B$ y después estimar el intervalo de confianza para β_2 como

$$\left(\widehat{\beta}_2 - t_{(r)}^* \widehat{\sigma}, \widehat{\beta}_2 + t_{(s)}^* \widehat{\sigma} \right)$$

con $t_{(r)}^*$ y $t_{(s)}^*$ los respectivos percentiles de $\alpha/2$ y $1 - \alpha/2$ porciento. Note que este procedimiento necesariamente requiere del suavizamiento en cada replicación de remuestreo para el cálculo de $\widehat{\sigma}_b^*$, que es precisamente lo que se quiere evitar. En la siguiente sección presentaremos una aplicación novedosa del remuestreo que permite usar B arbitrariamente pequeño y que además evita la selección de un parámetro de suavizamiento.

2. Convergencia no Condicional del Remuestreo

Bajo ciertas condiciones de regularidad, específicas de cada situación, se tiene que típicamente el operador de interés T_n converge en distribución a una variable aleatoria que denotaremos por T . Ya vimos anteriormente que usualmente, bajo esas mismas condiciones de regularidad, si T_n^* denota el mismo operador pero evaluado sobre los datos generados por remuestreo, entonces $T_{n,b}^* \rightarrow_{d^*} T$, $n \rightarrow \infty$ en probabilidad para cada $b = 1, \dots, B$. Ver por ejemplo Bickel y Freedman (1981) para el caso i.i.d.

En Domínguez y Aguirre (2004) se demuestra que bajo esta situación, manteniendo B fijo

$$(T_n, T_{n,1}^*, \dots, T_{n,B}^*) \rightarrow_d \underline{T}_B, \quad n \rightarrow \infty \quad (3)$$

donde \underline{T}_B es un vector aleatorio con $1 + B$ variables aleatorias i.i.d. Varios hechos son importantes de notar respecto al resultado (3). En primer lugar, para B fijo, la convergencia es no condicional de esta manera este resultado difiere de casi todos los resultados que sobre remuestreo se encuentran en la literatura. En segundo lugar, (3) establece que T_n y $T_{n,b}^*$ son asintóticamente independientes para $b = 1, \dots, B$ por lo que este resultado brinda una aproximación para la distribución conjunta del estadístico de interés y de sus contrapartes obtenidas por remuestreo. Finalmente, cabe mencionar que este resultado es bastante general y que únicamente depende de que se tenga la convergencia débil de T_n y $T_{n,b}^*$.

3. Aplicación a Inferencia en Regresión Percentil

En este caso se aplica el resultado de la sección anterior a un operador que es asintóticamente normal. Es decir, supongamos que T tiene distribución $N(0, \sigma_T)$. Entonces por (3) es fácil ver que para B fijo

$$t_B^* = \frac{T_n}{\left(\frac{1}{B} \sum_{b=1}^B [T_{n,b}^*]^2\right)^{1/2}} \rightarrow_d t_B, \quad n \rightarrow \infty \quad (4)$$

donde t_B es la distribución t de Student con B grados de libertad. En particular si consideramos $T_n = \sqrt{n} (\hat{\beta}_2 - \beta_2)$ y $T_n^* = \sqrt{n} (\hat{\beta}_2^* - \hat{\beta}_2)$ tenemos que (4) se convierte en

$$t_B^* = \frac{(\hat{\beta}_2 - \beta_2)}{\left(\frac{1}{B} \sum_{b=1}^B [\hat{\beta}_{2,b}^* - \hat{\beta}_2]^2 \right)^{1/2}} \rightarrow_d t_B, \quad n \rightarrow \infty$$

con lo cual se puede hacer inferencia sobre β_2 sin necesidad de seleccionar un parámetro de suavizamiento para estimar σ . Si denotamos por $s^2 = \frac{1}{B} \sum_{b=1}^B [\hat{\beta}_{2,b}^* - \hat{\beta}_2]^2$ entonces el intervalo de confianza se obtendría como

$$(\hat{\beta}_2 - t_{B,\alpha/2}s, \hat{\beta}_2 + t_{B,1-\alpha/2}s)$$

De hecho en Domínguez y Aguirre (2004), se demuestra que si de desea probar la hipótesis $H_0 : \beta_2 = \beta_{2,0}$ versus $H_1 : \beta_2 \neq \beta_{2,0}$ haciendo $t_{B,0}^* = \frac{(\hat{\beta}_2 - \beta_{2,0})}{\left(\frac{1}{B} \sum_{b=1}^B [\hat{\beta}_{2,b}^* - \hat{\beta}_2]^2 \right)^{1/2}}$ entonces

$$\lim_{n \rightarrow \infty} \Pr [|t_{B,0}^*| \leq t_{B,1-\alpha/2}] = \alpha \text{ bajo } H_0,$$

$$\lim_{n \rightarrow \infty} \Pr [|t_{B,0}^*| \leq t_{B,1-\alpha/2}] = 0 \text{ bajo } H_1,$$

con lo que se tiene un procedimiento consistente de tamaño α .

4. Consideraciones

Nótese que (4) parece indicar que el denominador está estimando la desviación estándar del numerador, pero no es así. El denominador no tiene porqué converger a σ_T de manera alguna. Lo que se está aprovechando es la convergencia conjunta del numerador y del denominador a

un vector normal multivariado con entradas independientes. Típicamente se usa el remuestreo para estimar la desviación estándar del operador de interés, pero eso requiere de condiciones de regularidad adicionales.

El enfoque mostrado en (4) se puede generalizar a un concepto que se podría denominar como el de transformaciones asintóticamente pivotales, en las cuales se combina el operador de interés con sus contrapartes de remuestreo. Es decir, hay que constuir

$$\psi(T_n, T_{n,1}^*, \dots, T_{n,B}^*) \quad (5)$$

tal que (5), para cada B fijo, sea asintóticamente pivotal.

Referencias

- Bickel, P. y Freedman, D. (1981), Some Asymptotic Theory for the Bootstrap, *Annals of Statistics*, 9, 1196-1217.
- Bowman, A. (1985), A Comparative Study of Some Kernel-based Nonparametric Density Estimators, *Journal of Statistical Computation and Simulation*, 21 , 313-327.
- Domínguez, M. y Aguirre, V. (2004), Broadening the Scope of the Bootstrap in Complex Problems, Reporte Técnico DE-C04.5, Departamento de Estadística, ITAM.
- Cao, R., Cuevas, A. y González-Manteiga, W. (1994), A Comparative Study of Several Smoothing Methods in Density Estimation, *Computational Statistics and Data Analysis*, 17 , 153-176.
- Hahn, J. (1995), Bootstrapping Quantile Regression Estimators, *Econometric Theory*, 11, 105-121.
- Koenker, R. y Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46, 33-50.

Estrategias Estadísticas para Generar Productos Confiables

Carlos A. Carballo Monsivais¹

Jorge Domínguez Domínguez²

Centro de Investigación en Matemáticas, A.C.

1. Introducción

Una actividad importante hoy en día es diseñar productos o componentes que no fallen, esta práctica cae en el contexto del análisis de confiabilidad. Existen varias características que dan lugar a que un producto sea confiable, por ejemplo la vida de anaquel, la degradación, el tiempo de vida, periodos de falla, entre otras. El diseño de experimentos juega un papel relevante para crear estrategias que permitan obtener productos que sean confiables. En estudios de confiabilidad existen datos censurados por ello se requieren procedimientos eficientes para estimar parámetros de los modelos. El objetivo del presente trabajo es describir los procedimientos de Hamada-Wu y Hahn-Morgan-Schmee para la estimación de los parámetros en datos censurados y luego mediante un ejemplo comparar sus resultados. El primero se basa en el principio de Máxima Verosimilitud, el segundo en mínimos cuadrados iterados. La ventaja de este último estrive en la familiaridad que tiene la aplicación de los mínimos cuadrados en la industria. La aplicación de los resultados de estos métodos tienen como finalidad: 1. Establecer características de los equipos o componentes para determinar periodos de garantía. 2. Predecir la confiabilidad de un producto. 3. Precedir los costos de garantía. 4. Asegurar la calidad, mantenimiento y seguridad de los productos demandados por los clientes.

¹abraham@cimat.mx

²jorge@cimat.mx

2. Descripción y planteamiento del Análisis de Confabilidad

La confiabilidad es la mejor medida cuantitativa de la integridad de una parte diseñada, componente, producto, o sistema. Así la confiabilidad se define como la probabilidad de que un sistema opere (aptitud para realizar una función determinada) por un periodo de tiempo. El periodo de tiempo es un intervalo de longitud t , a saber $[0, t)$. En este caso la confiabilidad: $R(t) = P[\text{sistema opere durante } [0, t))$. Una característica adicional en el análisis de confiabilidad surge porque algunas unidades experimentales pueden no fallar durante un periodo establecido. Esto da lugar a datos incompletos para el análisis estadístico a esta situación se le denomina censura. Existen varios tipos de censura: censura por la derecha, censura por la izquierda, y censura por intervalo.

El análisis estadístico que se trabajará en este resumen se referirá a la censura por la derecha. Esta se define como sigue: considere un periodo de observación C , entonces una unidad observa un tiempo de falla T_i si $T_i \leq C$ en tal caso la unidad no se censura, en caso contrario la unidad se censura. El tiempo de observación de la unidad es $T_o = \min(T_i, C)$. Para analizar la confiabilidad de los datos experimentales se recurre al modelo de regresión:

$$Y = \log(T) = x^t \beta + \varepsilon, \quad (1)$$

donde $x^t = (x_1, \dots, x_k)$ son los factores que afectan la confiabilidad del producto, β es un vector de parámetros, finalmente ε es una distribución de probabilidad, en este estudio se propone la distribución Gumbel. Cabe observar la relación entre las variables T y Y en el modelo (1), mientras que T tiene una distribución Weibull, $\log(T)$ sigue una distribución Gumbel.

El principio de verosimilitud permite estimar los parámetros del modelo (1) y el planteamiento de este se presenta en el contexto del tipo de censura es decir:

$$L(\theta) = -\sum_i^n \delta_i \log(\sigma) + \sum_{i=1}^n \delta_i \log f(z_i) - \sum_{i=1}^n (1 - \delta_i) \log R(z_i), \quad (2)$$

donde $f(z_i)$ y $R(z_i)$ son respectivamente las funciones de densidad y de confiabilidad de la

Gumbel (2), se expresan como:

$$f(z) = \frac{1}{\sigma} \exp \{z - \exp z\}, \quad R(z) = \exp [-\exp z],$$

en términos del modelo (1) $z = \frac{Y - x^t \beta}{\sigma}$ y δ_i es la indicadora, es decir $\delta_i = 1$, si $t_i \leq C$, no hay censura y $\delta_i = 0$, si $t_i > C$, el dato se censura.

3. Métodos

En este trabajo se presentarán dos métodos para ajustar un modelo de regresión. En la siguiente subsección se mostrará el algoritmo del método HW propuesto por Hamada y Wu (1991). A continuación se describe el método HMS propuesto por Hanh, Morgan y Schmee (1981). Mediante un ejemplo discutido por Davis (1995), se presentan los resultados que generan ambos métodos y finalmente se incluyen conclusiones con comentarios generales de los resultados.

3.1. Método Hamada y Wu (HW)

Hamada y Wu (1991) proponen un método iterativo como una alternativa muy simple y flexible considerando muchos modelos simultáneos. El procedimiento del método HW consiste en lo siguiente:

1. Fase A: Existe una selección del modelo, un modelo inicial de especificación, lo llamamos el modelo 0, que incluye los efectos, interacciones que pueden ser potencialmente importantes. $\mu = X_0 \beta_0$.
2. Despues se ajusta el modelo (lo llamados modelo i) usando el criterio de Máxima Verosimilitud (2), el trato de una observacion censurada y un tiempo de falla es distinto en esta fase. $\mu = X_i \beta_i$.
3. En la fase de imputación, se imputan los datos censurados por su esperanza condicional. Puesto que es de interés identificar la localización de los efectos, usamos la esperanza condicional como un valor típico. $E[(h)|y \in (a, b)] = x_i \beta_i + \sigma f(z)/F(z)$.

4. Se selecciona el modelo. Informalmente se aplica una técnica estándar para seleccionar el modelo, se detiene la iteración cuando el modelo i actual es el mismo al anterior, puesto que $X_i\beta_i = X_{i-1}\beta_{i-1}$. Se repiten los pasos [2], [3] y [4] hasta la selección del modelo adecuado.
5. Fase B: Se determina un modelo. El modelo final en la Fase A es utilizado para un análisis formal. Este es juzgado por su simplicidad, estructura adecuada y significancia científica. Hay que analizar los residuales para determinar supuestos distribucionales.
6. Fase C: Recomendación del nivel Factor. Hay que seleccionar niveles donde se mejoran considerablemente los tiempos de falla que se usan para predecir respuestas con el modelo final seleccionado. Una vez utilizado las predicciones de todas las combinaciones, hay que determinar la mejor combinación y realizar corridas confirmatorias.

3.2. Método Hahn-Morgan-Schmee (HMS)

Schmee y Hahn(1979) sugieren un simple método de suma de cuadrados iterativos para analizar datos censurados. Posteriormente Hahn-Morgan-Schmee (1981) lo aplican para analizar experimentos fraccionados con datos censurados. Los pasos para la aplicación de un algoritmo de solución son los siguientes.

1. Se realiza una estimación con mínimos cuadrados de la manera común tratando los datos censurados como tiempo de falla completos.

2. Se calcula el tiempo de falla esperado, con la siguiente ecuación:

$$\mu_x^* = \mu_x + \sigma f(z)/R(z) \text{ donde } z = (t - \mu_x)/\sigma \text{ y } R(z) = 1 - F(z). \quad (3)$$

3. Se calculan los valores iniciales en la iteración 0 utilizando (3), y posteriormente en la iteración 1 se sustituyen en lugar de la censura en el modelo para lograr un mejor ajuste en las iteraciones sucesivas.
4. Por último se siguen estos pasos hasta alcanzar la convergencia.

4. Ejemplo

Se presenta un caso de un estudio real de la Compañía Ford Motor (Davis, 1995), el enfoque es utilizar un modelo de regresión Weibull y comparar los métodos HW y HMS. El experimento trata sobre los ejes de transmisión automática de un cierto vehículo, se provoca fatiga bajo esfuerzo prolongado hasta que una falla es observable por un agrietamiento del eje. El eje es un componente clave de la transferencia de energía por rotación para la impulsión del motor. Se determinaron los siguientes factores (Figura 1):

Factor	Descripción	Nivel		
		1	2	3
A	Perfil del Extremo	esferico	acanalado	---
B	Tiempo de exp.	30 min.	1 hr.	4 hr.
C	Diametro del eje	16.1 mm	17.7 mm	18.8mm
D	Intensidad del tiro	3A	6A	9A
E	Covertura	200%	400%	600%
F	Temperatura de Op.	140	160	180
G	Agostación de tiro	sin	con	

Figura 1: Factores y sus niveles

El propósito del experimento es encontrar la mejor combinación de niveles de los factores que se podrían utilizar para aumentar la vida útil del eje. El experimento consiste en utilizar un arsenal ortogonal comúnmente conocido como L_{18} , que posee 18 corridas experimentales cada combinación se réplica dos veces, dando un total de 36 ejes de transmisión en el experimento. Se presentan los resultados del experimento (Figura 2), la censura por la derecha se denota con el signo + en la derecha

Se analizan estos resultados por ambos métodos HW y HMS, además se incluye el análisis de Máxima Verosimilitud, se comparan los resultados para discutir la eficiencia de los métodos.

Configuración	Factores							Tiempos de Fallas		
	A	B	C	D	E	F	G	e	Replica 1	Replica 2
1	1	1	2	2	2	2	1	2	322	2000
2	1	2	2	1	1	1	2	1	95	95.4
3	1	3	2	3	3	3	3(1)	3	2000+	125
4	1	1	1	2	1	3	3(1)	1	747	414
5	1	2	1	1	3	2	1	3	821	192
6	1	3	1	3	2	1	2	2	2000+	2000+
7	1	1	3	1	2	1	3(1)	2	972	2000+
8	1	2	3	3	1	3	1	1	2000+	1920
9	1	3	3	2	3	2	2	3	2000+	2000+
10	2	1	2	3	3	1	1	3	739	285
11	2	2	2	2	2	3	2	2	1080	634
12	2	3	2	1	1	2	3(1)	1	2000+	1940
13	2	1	1	1	3	3	2	3	2000+	1790
14	2	2	1	3	2	2	3(1)	2	2000+	617
15	2	3	1	2	1	1	1	1	2000+	2000+
16	2	1	3	3	1	2	2	1	1380	1110
17	2	2	3	2	3	1	3(1)	3	2000+	2000+
18	2	3	3	1	2	3	1	2	2000+	2000+

Figura 2: Resultados al realizar el experimento

5. Conclusiones

Se presenta la tabla con los resultados finales (Figura 3) aplicando los métodos HW, HWS.

Coeficiente	Valores del Contraste basados en		
	HMS	HW	m.l.e.
β_0	5.449	6.497	6.477
β_1	0.388	0.506	0.502
β_2	0.702	0.99	0.952
β_{22}	0.84	1.043	1.02
β_3	0.211	0.574	0.564
β_{33}	1.424	1.782	1.736
β_4	0.081	-0.316	-0.136
β_{44}	-0.407	-0.683	-0.645
β_5	0.061	0.151	0.142
β_{55}	-0.514	-0.695	-0.689
β_6	-0.066	0.028	0.046
β_{66}	0.163	0.21	0.199
β_7	0.062	0.15	0.151
σ	0.882	0.924	0.861

Figura 3: Estimación de los parámetros del modelo, por los métodos de HMS, HW y MV.

Se comenta que la aplicación del método HMS en el análisis de los datos experimentales proporciona a los ingenieros una mayor flexibilidad y habilidad en la estimación de parámetros

para representar estos procedimientos estadísticos, debido a su facilidad de razonar su complejidad. Por otro lado, el método HW da resultados no parecidos si se aplica directamente Máxima Verosimilitud, este hecho da lugar a explorar las características de estas diferencias por técnicas de simulación.

En estudios futuros es necesario realizar una comparación formal de estos métodos en otros escenarios, por ejemplo cambiar el tipo de censura, la distribución de los datos. Es claro que el diseño de experimentos es una herramienta de gran utilidad para obtener productos con mayor confiabilidad, además que utilizar este tipo de herramientas tiene un alto impacto económico en el negocio sobre la generación de productos confiables.

Referencias

- Davis, T.P., (1995). Analysis of an Experiment Aimed At Improving the Reliability of Transmission Centre Shafts, *Life Time Data Analysis*, I, 275-306.
- Hahn, G. J., Morgan, C.B., y Schmee J. (1981). The Analysis of a Fractional Experiment With Censored Data Using Iterative Least Squares. *Technometrics*, 23, 33-36.
- Hamada, M. y Wu C.F. (1991). Analysis of Censored Data From Highly Fractionated Experiments, *Technometrics*, 33, 25-38.
- Hamada, M. (1995). Using Statistically Designed Experiments to Improve Reliability and to Achieve Robust Reliabilitly, *IEEE Transactions on Reliability*, 44, 206-215.
- Schemee J. y Hahn G. (1979). A simple Method for Regression Analysis With Censored Data. *Technometrics*, 21, 417-432.

Aplicación de la Metodología Seis Sigma a una Empresa de Transporte

Adriana Centeno Gil¹

Universidad de las Américas, Puebla

Jorge Domínguez Domínguez²

Centro de Investigación en Matemáticas, A.C.

Antonio González Fragoso³

Universidad de las Américas, Puebla

1. Introducción

A través del tiempo se han propuesto metodologías para evaluar la calidad de los bienes y servicios. En lo que se refiere a servicios se ha presentado frecuentemente el problema de que las evaluaciones de los procesos no son objetivas, no logrando una confianza total cuando se aplican algunas de éstas técnicas.

La mayoría de los esfuerzos para el mejoramiento de la calidad son dirigidos a manufactura debido a que en los procesos de servicios, en muchas ocasiones, existe cierta dificultad para realizar la medición de sus variables por la naturaleza de las mismas.

El objetivo principal de este trabajo es analizar la calidad en una empresa que ofrece servicios de transporte urbano, en uno de sus procesos, bajo la metodología de Seis Sigma. Siendo el Despliegue de la Función de Calidad (Quality Function Deployment-QFD) la metodología fundamental para el desarrollo de este trabajo. Esta última, es de vital utilidad para realizar el análisis de las necesidades del cliente.

¹adri_mexico@yahoo.com.m

²jorge@cimat.mx

³antonio.gonzalez@udlap.mx

2. Metodología Seis Sigma

Seis Sigma es una metodología que evalúa a las empresas de tal modo que se identifiquen las fallas en éstas y de esa forma se busca una mejora continua. Se dirige a tres áreas principales: mejorar la satisfacción del cliente, reducir el tiempo de ciclo y reducir los defectos.

Las aplicaciones de la metodología Seis Sigma se encuentran dentro de un marco simple de funcionamiento, mejor conocido como ciclo DMAMC: definir, medir, analizar, mejorar y controlar. Siempre sustentándose cada una de estas cinco etapas, de diversas herramientas estadísticas.

Seis Sigma técnicamente significa no tener más de 3.4 defectos por cada millón de oportunidades, en todo proceso, entendiéndose por defecto un cliente insatisfecho. Pero más importante que la definición técnica anterior, es que Seis Sigma es una metodología con el propósito fundamental de lograr una mejora continua en los procesos.

3. El Despliegue de la Función de Calidad

El despliegue de la función de calidad (Quality Function Deployment, QFD) es una metodología de administración de la calidad dirigido por el cliente cuyo objetivo es crear una mejor satisfacción del producto o servicio que el cliente adquiere.

QFD está compuesto por cuatro fases que despliegan las necesidades del cliente a través de procesos de planeación, véase Figura 1.

La primera fase es conocida como la “Casa de calidad” (House of quality, HOQ), en esta fase se traducen las necesidades del cliente (QUE's) a medidas técnicas (COMO's).

La siguiente fase es el despliegue de partes, con lo que se traduce las medidas técnicas clave (nuevos QUE's) determinadas en la fase anterior, en partes características (COMO's).

La tercera fase es el planeamiento del proceso, es decir, traduce las partes características clave (nuevos QUE's) obtenidos en la etapa anterior, en procesos de operación (COMO's).

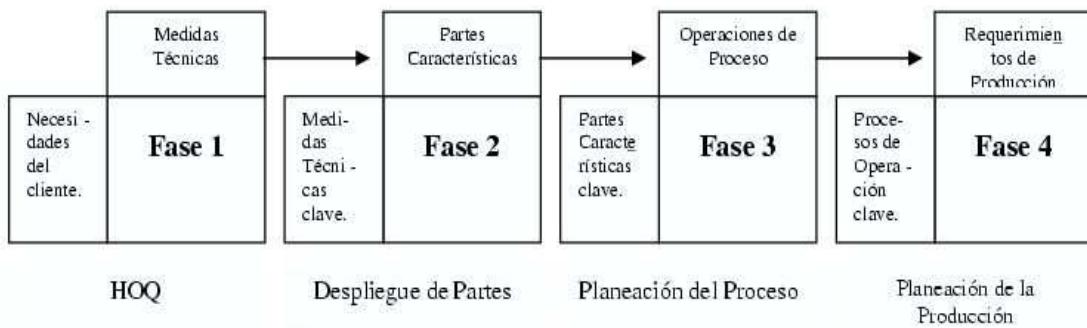


Figura 1: La cuatro fases del Despliegue de la Función de Calidad.

Finalmente la última fase es el planeamiento de la producción, que traduce los procesos de operación clave (nuevos QUE's) en requerimientos de producción día a día (COMO's).

4. El Caso de Estudio: Aplicación de Seis Sigma a una Empresa de Transporte

La empresa a la cual irá enfocada la aplicación de la metodología Seis Sigma, es un grupo que cuenta con servicios de transporte foráneo, transporte especializado y servicios de envío. En trabajo solo se enfoca al servicio de transporte foráneo.

El requerimiento de la empresa fue el de conocer sus áreas de oportunidad en uno de sus procesos, el proceso de traslado de sus clientes de la Ciudad de Puebla, terminal CAPU a la Ciudad de México terminal TAPO.

En la siguiente Figura se resume la aplicación de la metodología Seis Sigma, bajo el esquema DMAMC, al proceso de servicio considerado en este análisis. Es importante mencionar que el alcance de este trabajo fue la aplicación las tres primeras etapas: definir, medir y analizar. Para las otras dos fases solo se dieron recomendaciones.

En la etapa de definición, la empresa contaba ya bases de datos construidas por encuestas propias anteriores, es por eso que a través del análisis de los resultados y de la definición de los requerimientos de la empresa (por medio de pláticas con el personal de la empresa

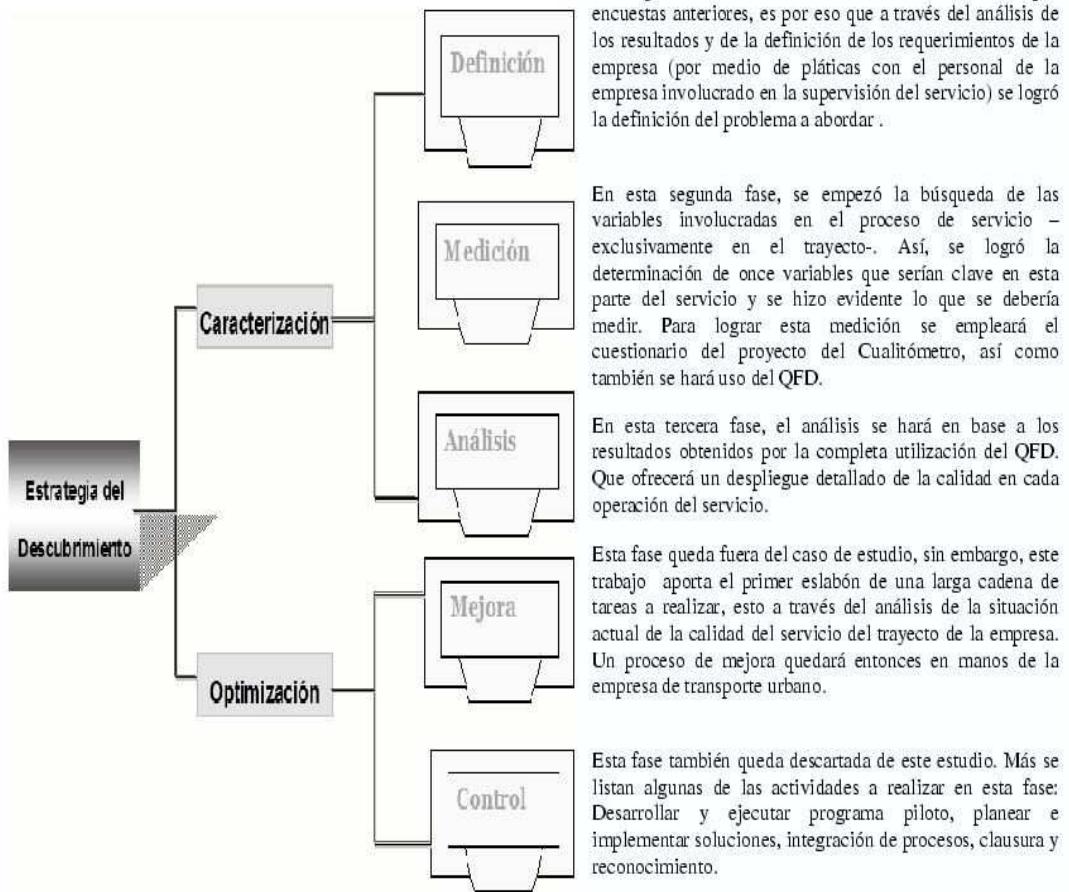


Figura 2: Ciclo DMAMC para la empresa de transporte.

1er. Nivel	2o. Nivel
Tangibles	Condiciones del autobús
Confort	Asientos cómodos, espaciosos
Limpieza	En el autobús, interiores y exteriores
Organización	Todo listo para partir
Puntualidad	Salidas en el tiempo indicado
Cortesía	Respeto, amabilidad
Comunicación	Información entendible y habilidad de escuchar
Conocimiento del cliente	Entendimiento de las necesidades del cliente
Higiene personal	Pulcritud en la apariencia del conductor
Orden	El autobús se nota ordenado
Seguridad	Conducción responsable, portación de gafetes

Figura 3: Carta de calidad demandada.

involucrado en la supervisión del servicio) se logró la definición del problema a abordar.

Una vez que se decidieron once variables como las más importantes a medir, se procedió a elaborar un cuestionario, en base al del proyecto del Cualitómetro, ver Fiorenzo - Rosetto (1998). En este se mide la calidad esperada, la calidad percibida y la importancia de cada una de las once variables de importancia. Se aplicó a los clientes de forma auto-administrada y se cuidaron detalles de aleatoriedad tanto en los horarios de los traslados, como en la selección de los clientes dentro del autobús.

Una vez comprendida las necesidades de los clientes, se procedió a la construcción de la carta de calidad (Figura 3), la cual consiste en listar las variables que constituyen las necesidades de los clientes y a su vez, se definieron los conceptos que pueden ser medibles dentro de las variables de dicha carta (Figura 4). A continuación se presentan ambas cartas.

Con la combinación de las dos cartas anteriores, se obtuvo la carta de calidad para la empresa de transporte, sin embargo, para proseguir en la aplicación de QFD fue necesario determinar cuáles variables eran las preferentes. Esto llevó a un problema de toma de decisiones con múltiples atributos. La tarea de la toma de decisiones con múltiples atributos (MADM, siglas en inglés) se refiere a las puntuaciones y a la selección de la o las alternativas por medio de la construcción de un orden de preferencias en los atributos considerados.

Para este estudio se eligió la técnica: para el orden de preferencias por similaridad a la solución ideal (TOPSIS, siglas en inglés) ver, Lai - Wu. (1998) TOPSIS aplica un principio intuitivo: la alternativa seleccionada deberá tener la distancia más corta a la mejor solución,

1er. Nivel	2o. Nivel	3er. Nivel
Grado del trayecto	Grado de las condiciones tangibles del autobús	Grado de limpieza de los pisos Grado de limpieza de las ventanillas Grado de limpieza de los asientos Grado de limpieza de las cortinas y protectores de sol Grado de limpieza de los baños Grado de limpieza de la cajuela Grado del buen estado de los asientos Grado del buen estado de los cinturones de seguridad Grado del buen estado de los pisos Grado del buen estado de los portabultos de mano Grado del buen estado de las luces internas Grado del buen estado del audio Grado del buen estado del video Grado del buen estado del baño Grado del buen estado de los vidrios de las ventanillas

Figura 4: Fragmento de la carta de elementos de calidad.

y la distancia más grande a la peor. Para aplicar esta técnica de optimización, fue necesario definir para cada QUE el nivel de relación existente con cada COMO.

El siguiente paso fue la elaboración de la carta del diseño de calidad, la cual implicó la definición, por parte de la empresa de transporte foráneo, de los lineamientos de calidad para las doce variables elegidas después de la aplicación de la técnica de optimización antes descrita. Un fragmento de la carta se presenta a continuación.

El siguiente paso fue la elaboración de la carta de despliegue de la operación del servicio, dicha carta se puede entender como una mezcla de un diagrama de flujo de la operación del servicio con los lineamientos establecidos previamente en la carta del diseño de calidad. Un fragmento de ésta se presenta a continuación.

La carta final se denomina carta de control de calidad para los procesos de servicio, ésta es de suma importancia para la empresa de transporte foráneo ya que viene a ser un mapa de verificación para el conductor del autobús y la gerencia. Ésta establece los objetivos que se procuraran lograr en la operación del servicio para ofrecer un servicio con calidad.

QUE's	COMO's
Limpieza de los pisos	<ul style="list-style-type: none"> • No debe haber basura. • No se deben observar manchas.
Limpieza de las ventanillas	<ul style="list-style-type: none"> • Los cristales deberán verse sin manchas. • La vista a través de ellos deberá ser clara.
Limpieza de los asientos	<ul style="list-style-type: none"> • Las vestiduras deberán estar limpias. • No se verá basura (o migajas) en las vestiduras.
Limpieza de los baños	<ul style="list-style-type: none"> • El sanitario deberá verse limpio.
Limpieza de la carrocería exterior	<ul style="list-style-type: none"> • La pintura deberá mantenerse intacta, no se observarán raspones. • Deberá observarse limpio, sin manchas.
Buen estado del baño	<ul style="list-style-type: none"> • El sanitario deberá estar bien cuidado. • No se permitirá que se encuentre roto. • Deberá funcionar bien.
Buena señalización interior, funcional y suficiente	<ul style="list-style-type: none"> • Las señalizaciones de las salidas de emergencia deberán ser notorias. • Deberán ser visibles a todos los pasajeros. • Deberán mantenerse en buen estado (sin rayones).

Figura 5: Fragmento de la carta del diseño de calidad.

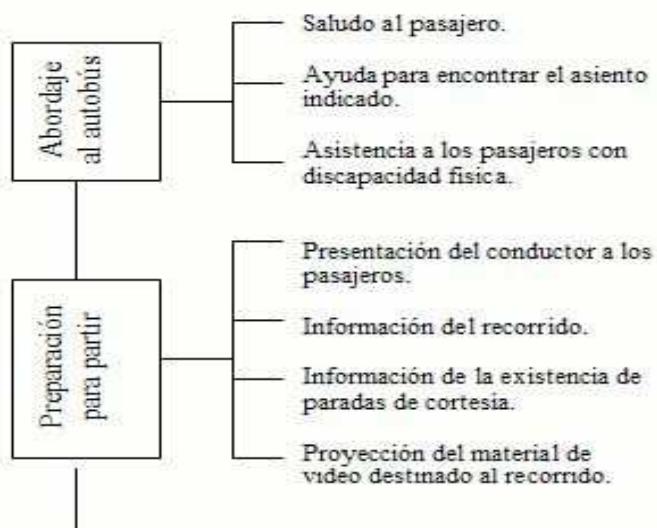


Figura 6: Fragmento de la carta del despliegue de la operación del servicio.

Calidad	Índices de Calidad	Clasificación	Objetivo	Límite de control	Definición
Del autobús.	Tasa de quejas por una deficiente limpieza.	Pisos	0		Número de quejas por falta de limpieza.
		Ventanillas	0		Número de quejas por falta de limpieza.
		Asientos	0		Número de quejas por falta de limpieza.
		Baño	0		Número de quejas por falta de limpieza.
		Carrocería exterior	0		Número de quejas por falta de limpieza.
	Tasa de quejas por el estado de las instalaciones.	Baño	0		Número de quejas.

Figura 7: Fragmento de la carta de control de calidad para los procesos de servicio.

5. Conclusiones

La metodología Seis Sigma ha permitido aplicar dos técnicas estadísticas para evaluar el servicio que presta una empresa de transporte. Se pueden destacar dos aspectos relevantes:

1. Se pudo detectar que los clientes no perciben la calidad que esperan. La otra es que la empresa tiene una metodología para planear y evaluar su desempeño.
2. La tarea de evaluar la calidad de cualquier proceso de servicio no es nada fácil.
3. Conduce a buenos resultados siempre que se cuente con buenas y eficientes técnicas de evaluación (por buenas y eficientes se entienden a todas las metodologías que ayuden a eliminar la subjetividad en el proceso de evaluación).
4. Siempre que se elabore en conjunto con personal de la empresa de servicios dará a la evaluación creatividad e iniciativa, que en adición a lo anterior, ayudará a obtener mejores resultados.

Referencias

Fiorenzo, F. and Rossetto, S (1998). On-line Service Quality Control: the Qualitometro Method. *Quality Engineering*, **10**(4), pp 633 – 643.

Lai Ch.,K and Wu, M., L.(1998). Prioritizing the technical measures in Quality Function Deployment. *Quality Engineering*, **10**(3), pp 467 - 479.

Inferencia Bayesiana para la media y desviación estándar de una población normal cuando sólo se observan el tamaño, la media y el rango muestrales

Enrique de Alba¹

División Académica de Actuaría, Instituto Tecnológico Autónomo de México

Juan José Fernández-Durán²

División Académica de Estadística, Instituto Tecnológico Autónomo de México

Ma. Mercedes Gregorio-Domínguez³

División Académica de Matemáticas, Instituto Tecnológico Autónomo de México

1. Introducción

El valor de mercado del jugo de naranja o concentrado depende de su contenido de azúcar, el cual se expresa en grados BRIX: °BRIX. La Codex Alimentarius Commission de la FAO (ONU) es el organismo internacional responsable de determinar el valor mínimo estándar de °BRIX para las transacciones comerciales. El Grupo de Trabajo Intergubernamental para Frutas y Hortalizas, Ad-Hoc para el Codex Alimentarius, acordó un procedimiento para el análisis del BRIX. Se estableció que el estándar se debe expresar en términos de la media de °BRIX correspondiente a jugo de expresión directa. Además, al determinar este valor se debe tomar en cuenta el volumen de producción correspondiente a la información de ° BRIX presentada.

En el grupo de trabajo, algunas delegaciones manifestaron que el rango tan amplio de valores °BRIX, especialmente en el caso de la naranja, hacía que fuera difícil establecer una metodología para evaluar esta información y establecer un estándar que fuera equitativo para todos los países.

¹dealba@itam.mx

²jfdez@itam.mx

³mercedes@itam.mx

El comité de México considera que para determinar el °BRIX estándar es muy importante considerar la variabilidad de los valores que presentan los países; se deben analizar los datos incorporando a la elaboración del estándar los siguientes criterios estadísticos:

1. Desviación estándar de las medias.
2. Límites de confianza de la media.
3. Cálculo de la incertidumbre específica.

Sin embargo, la única información disponible actualmente es limitada. La información de °BRIX para jugo de naranja fresco en muchos países consiste en: tamaño de muestra, media y rango. En muchas situaciones se reporta el rango como medida de variabilidad. Aunque la metodología es aplicable a cualquier fruta en esta nota nos concentraremos en la información para naranjas. En este artículo proponemos un modelo que permite aprovechar la información disponible para establecer un estándar de ° BRIX, que tome en cuenta los criterios mencionados. La metodología propuesta se aplica a datos reales, que corresponden al contenido °BRIX en el jugo de naranjas producidos por diversos países. Los resultados son útiles para fijar límites al contenido de azúcar en las naranjas de exportación y así establecer la cantidad que puede exportar cada uno de los países productores.

2. El modelo

En el caso de la información sobre grados °BRIX es razonable suponer normalidad. Considerese una muestra de una distribución normal cuya media y desviación estándar se desconocen. Sólo se han registrado la media, el rango y el tamaño de la muestra. Se requiere hacer inferencia acerca de la media y la desviación estándar de la población, así como obtener la distribución predictiva. En esta nota se presenta la estimación de la media y la varianza, desde el punto de vista Bayesiano mediante un algoritmo basado en MCMC para generar observaciones de la distribución posterior conjunta de la media y la desviación estándar. No existe una expresión analítica para dicha distribución.

Sea X_1, \dots, X_n m.a. de $N(\mu, \sigma^2)$ donde $X_{(i)}, i = 1, \dots, n$, son las estadísticas de orden. A partir de este supuesto se tienen los siguientes resultados:

1. La media muestral (\bar{X}) y el rango $R = X_{(n)} - X_{(1)}$, son independientes.

Sigue de que R se puede escribir como

$$R = \max\{X_1 - \bar{X}, \dots, X_n - \bar{X}\} - \min\{X_1 - \bar{X}, \dots, X_n - \bar{X}\}. \quad (1)$$

2. Si sólo observamos \bar{X} y R la distribución conjunta de \bar{X} y R dados μ y σ^2 , cumple

$$f_{\bar{X}, R}(\bar{x}, r | \mu, \sigma^2) = f_{\bar{X}}(\bar{x} | \mu, \sigma^2) f_R(r | \mu, \sigma^2) \quad (2)$$

donde $\bar{X} | \mu, \sigma^2 \sim N(\mu, \frac{\sigma^2}{n})$.

3. La distribución marginal de R dados μ, σ se puede derivar de la distribución conjunta del rango y semi-rango: $R = X_{(n)} - X_{(1)}$ y $T = \frac{X_{(n)} + X_{(1)}}{2}$.

En general para una m.a. X_1, \dots, X_n , de una población con distribución, F_X , y densidad, f_X , con parámetro, θ :

$$f_{R,T}(r, t | \theta) = n(n-1) \left[F_X \left(t + \frac{r}{2} | \theta \right) - F_X \left(t - \frac{r}{2} | \theta \right) \right]^{n-2} f_X \left(t - \frac{r}{2} | \theta \right) f_X \left(t + \frac{r}{2} | \theta \right)$$

para $r > 0$ y $-\infty < t < \infty$.

En el caso específico de que $X_i \sim N(\mu, \sigma^2)$

$$\begin{aligned} f_{R,T}(r, t | \theta) &= f_{R,T}(r, t | \mu, \sigma^2) = \\ n(n-1) &\left[\int_{t-\frac{r}{2}}^{t+\frac{r}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \right]^{n-2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(t-\frac{r}{2}-\mu)^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(t+\frac{r}{2}-\mu)^2} \\ & \text{y} \\ f_R(r | \mu, \sigma^2) &= \\ \int_{-\infty}^{\infty} n(n-1) &\left[\int_{t-\frac{r}{2}}^{t+\frac{r}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \right]^{n-2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(t-\frac{r}{2}-\mu)^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(t+\frac{r}{2}-\mu)^2} dt \end{aligned} \quad (3)$$

que es equivalente a

$$\begin{aligned}
f_R(r \mid \mu, \sigma^2) &= \\
n(n-1) \frac{1}{\sqrt{2\pi}\sqrt{2}\sigma} e^{-\frac{1}{4\sigma^2}r^2} \int_{-\infty}^{\infty} &\left[\Phi\left(\frac{1}{\sqrt{2}}z + \frac{r}{2\sigma}\right) - \Phi\left(\frac{1}{\sqrt{2}}z - \frac{r}{2\sigma}\right) \right]^{n-2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \\
n(n-1) N_R(0, 2\sigma^2) E_Z &\left[\left[\Phi\left(\frac{1}{\sqrt{2}}Z + \frac{r}{2\sigma}\right) - \Phi\left(\frac{1}{\sqrt{2}}Z - \frac{r}{2\sigma}\right) \right]^{n-2} \right]
\end{aligned} \tag{4}$$

donde $Z \sim N(0, 1)$ y $\Phi(\cdot)$ es la cdf de una Normal estándar. Se cumple

$$f_R(r \mid \mu, \sigma^2) = f_R(r \mid \sigma^2). \tag{5}$$

La verosimilitud es

$$p(\bar{x}, r \mid \mu, \sigma^2) = N_{\bar{X}}(\mu, \frac{\sigma^2}{n}) n(n-1) N_R(0, 2\sigma^2) E_Z \left[\left[\Phi\left(\frac{1}{\sqrt{2}}Z + \frac{r}{2\sigma}\right) - \Phi\left(\frac{1}{\sqrt{2}}Z - \frac{r}{2\sigma}\right) \right]^{n-2} \right], \tag{6}$$

o usando la precisión $\tau = \frac{1}{\sigma^2}$

$$p(\bar{x}, r \mid \mu, \tau) = N_{\bar{X}}(\mu, \frac{1}{n\tau}) n(n-1) N_R(0, \frac{2}{\tau}) E_Z \left[\left[\Phi\left(\frac{1}{\sqrt{2}}Z + \frac{r\sqrt{\tau}}{2}\right) - \Phi\left(\frac{1}{\sqrt{2}}Z - \frac{r\sqrt{\tau}}{2}\right) \right]^{n-2} \right]. \tag{7}$$

Suponemos independencia a-priori de los parámetros

$$p(\mu, \tau) = p(\mu)p(\tau)$$

con

$$p(\mu) \propto 1 \quad p(\tau) \propto \frac{1}{\tau}.$$

La distribución posterior conjunta es:

$$p(\mu, \tau \mid \bar{x}, r) \propto e^{-\frac{\tau}{2} \left(n(\mu - \bar{x})^2 + \frac{r^2}{2} \right)} E_Z \left[\left[\Phi\left(\frac{1}{\sqrt{2}}Z + \frac{r\sqrt{\tau}}{2}\right) - \Phi\left(\frac{1}{\sqrt{2}}Z - \frac{r\sqrt{\tau}}{2}\right) \right]^{n-2} \right]. \tag{8}$$

Nótese que es complicado obtener los valores del siguiente término “intratable”:

$$E_Z \left[\left[\Phi\left(\frac{1}{\sqrt{2}}Z + \frac{r\sqrt{\tau}}{2}\right) - \Phi\left(\frac{1}{\sqrt{2}}Z - \frac{r\sqrt{\tau}}{2}\right) \right]^{n-2} \right]. \tag{9}$$

En la siguiente sección se presenta un algoritmo de simulación por Monte Carlo con cadenas de Markov (MCMC).

3. Algoritmo MCMC

Las condicionales posteriores completas son

$$p(\mu \mid \bar{x}, r, \tau) \propto e^{-\frac{n\tau}{2}(\mu - \bar{x})^2} \propto N(\bar{x}, \frac{1}{n\tau}) \quad (10)$$

y

$$p(\tau \mid \bar{x}, r, \mu) \propto e^{-\frac{\tau}{2}(n(\mu - \bar{x})^2 + \frac{r^2}{2})} E_Z \left[\left[\Phi \left(\frac{1}{\sqrt{2}}Z + \frac{r\sqrt{\tau}}{2} \right) - \Phi \left(\frac{1}{\sqrt{2}}Z - \frac{r\sqrt{\tau}}{2} \right) \right]^{n-2} \right]. \quad (11)$$

La condicional posterior de τ involucra a los términos intratables ya mencionados y la constante de normalización se puede resolver incluyendo un “Metropolis-Hastings dentro del Gibbs”. El valor esperado se obtiene mediante la siguiente aproximación:

$$\hat{p}(\tau \mid \bar{x}, r, \mu) \propto e^{-\frac{\tau}{2}(n(\mu - \bar{x})^2 + \frac{r^2}{2})} \left[\frac{1}{M} \sum_{k=1}^M \left[\Phi \left(\frac{1}{\sqrt{2}}Z_k + \frac{r\sqrt{\tau}}{2} \right) - \Phi \left(\frac{1}{\sqrt{2}}Z_k - \frac{r\sqrt{\tau}}{2} \right) \right]^{n-2} \right] \quad (12)$$

donde Z_1, \dots, Z_M son i.i.d. de $N(0, 1)$ y el término intratable se obtiene mediante integración Monte Carlo directa. Se tiene entonces el siguiente algoritmo MCMC:

1. Especificar el número total de iteraciones del muestreador de Gibbs (TGS), el número total de iteraciones de Metropolis-Hastings (TMH) y el número total de normales estándar a usar en la aproximación del valor esperado que aparece en la condicional posterior de τ (TEV). También valores iniciales para μ y τ : μ_0 y τ_0 .
2. El siguiente valor de μ se obtiene muestreando de $N(\bar{x}, \frac{1}{n\tau})$ en donde τ se reemplaza por el valor anterior de τ en la cadena.
3. Se instrumenta la subcadena Metropolis-Hastings para muestrear el siguiente valor de τ . Esta tiene las siguientes características:
 - a) Distribución de propuesta: $q(\tau_{prop} \mid \tau_{pre}) = GamaInversa(\alpha, \beta)$, i.e., una propuesta Gama Inversa independiente, donde τ_{prop} representa el siguiente valor de τ propuesto, que se muestrea de esta distribución y τ_{pre} es el valor actual de τ en la subcadena.
 - b) La probabilidad de aceptación, $a(\tau_{prop}, \tau_{pres})$, está dada por

$$a(\tau_{prop}, \tau_{pres}) = \left(\frac{\hat{p}(\tau_{prop} | \bar{x}, r, \mu)}{\hat{p}(\tau_{pre} | \bar{x}, r, \mu)} \right) \left(\frac{q(\tau_{pre} | \tau_{prop})}{q(\tau_{prop} | \tau_{pre})} \right). \quad (13)$$

Se usan los mismos valores de Z_1, \dots, Z_M en el numerador y el denominador de la fracción que define la probabilidad de aceptación.

- 1) Si $a(\tau_{prop}, \tau_{pres}) > 1$, se acepta τ_{prop} .
- 2) Si no, se genera una v.a. uniforme U .
- 3) Si $U < a(\tau_{prop}, \tau_{pres})$ se acepta τ_{prop} .
- 4) Si no, se rechaza y se queda τ_{pres} .

El nuevo valor de τ en el muestreador de Gibbs corresponde al último valor en la subcadena de Metropolis-Hastings.

4. Aplicación

El objetivo principal de este trabajo es obtener estimaciones puntuales e intervalos de credibilidad Bayesianos para la media y la desviación estándar de una población normal a partir del tamaño de muestra, la media y el rango observados, así como predecir la siguiente observación de la distribución original. Los intervalos Bayesianos de credibilidad, del 90 % y 95 %, se obtienen a partir de los cuantiles 0.025, 0.05, 0.95 y 0.975. Utilizamos el paquete de cómputo *gibbsit* desarrollado por Raftery, A. E. y Lewis, S.M. (1992, 1992b and 1995). En particular se usó la versión de *S* que está disponible de forma gratuita en la librería *Statlib*. Entre los resultados que produce el paquete (output) se encuentra le número de iteraciones requeridas para alcanzar la precisión que se haya especificado, la cual en este caso se estableció en $r=0.0075$ con probabilidad $s=0.95$. Adicionalmente se obtiene la razón N_{min} (*I_RL*) que permite detectar problemas de convergencia; por experiencia Raftery y Lewis (1992, 1992b y 1995) han encontrado que valores de *I_RL* mayores a 5 son una clara indicación de la existencia de problemas de convergencia. El ejercicio de MCMC se corrió con $TGS=5000$, $TMH=1000$ y $TEV=500$. La muestra piloto en MCMC para el *gibbsit* fue 1665. En la Tabla 1 se presentan los datos °BRIX disponibles para varios países, con las estimaciones para la media y la desviación estándar. La parte superior de la tabla corresponde a las estimaciones de la media, mientras que la inferior corresponde a la desviación estándar, así como los cuantiles 0.025, 0.05, 0.95 y 0.975. Los valores máximos de *I_RL* y *Nprec* para dichos cuantiles se presentan en la sexta y séptima columnas, respectivamente. Los valores de *I_RL* no indican que pueda haber

Tabla 1

Country	Sample			MCMC Estimates	I_{RL}	$Nprec$	Quantiles			
	Size n	Mean \bar{X}	Range r				0.025	0.05	0.95	0.975
Brazil	1257	10.60	7.30	10.5978	1.08	3509	10.5352	10.5444	10.6492	10.6595
Cuba	24	12.00	4.00	12.0042	1.14	3689	11.5951	11.6705	12.3507	12.4183
Italy	72	11.89	3.10	11.8921	1.06	3422	11.7440	11.7686	12.0162	12.0430
Spain	670	11.60	5.00	11.6005	1.06	3254	11.5387	11.5487	11.6523	11.6625
Argentina	900	10.28	4.30	10.2802	0.96	3065	10.2373	10.2437	10.3165	10.3240
USA (Florida)	66513	11.07	6.10	11.0683	1.20	3472	11.0628	11.0637	11.0729	11.0738
South Africa	264	11.00	4.40	11.0003	1.09	3472	10.9054	10.9221	11.0799	11.0946
Turkey	932	12.06	4.30	12.0643	1.01	3258	12.0215	12.0281	12.1000	12.1077
Costa Rica	17040	11.04	6.54	11.0423	0.98	3174	11.0299	11.0319	11.0527	11.0547
Brazil	1257	10.60	7.30	1.0981	1.17	3509	0.9320	0.9602	1.2254	1.2500
Cuba	24	12.00	4.00	0.9907	1.07	3472	0.7004	0.7378	1.3043	1.4002
Italy	72	11.89	3.10	0.6411	1.04	3386	0.4922	0.5115	0.7840	0.8181
Spain	670	11.60	5.00	0.7950	1.04	3386	0.6663	0.6867	0.9008	0.9201
Argentina	900	10.28	4.30	0.6654	1.06	3422	0.5628	0.5773	0.7497	0.7654
USA (Florida)	66513	11.07	6.10	0.7073	1.01	3174	0.6427	0.6525	0.7575	0.7662
South Africa	264	11.00	4.40	0.7688	1.06	3422	0.6224	0.6453	0.8911	0.9131
Turkey	932	12.06	4.30	0.6632	1.10	3560	0.5618	0.5781	0.7460	0.7613
Costa Rica	17040	11.04	6.54	0.8197	1.13	3662	0.7310	0.7474	0.8862	0.8977

problemas de convergencia del algoritmo de MCMC propuesto y los valores de $Nprec$ indican que se requieren menos de 5000 iteraciones efectivas del algoritmo MCMC. Como un ejemplo de la aplicación de las distribuciones predictivas de este modelo se presenta la probabilidad de que la producción de cada país se pueda comercializar, suponiendo que se establezca como estándar °BRIX que sólo es aceptable la producción con valores °BRIX entre 11 y 12. En la Tabla 2 se presenta la probabilidad de que esto se cumpla para cada país. Claramente Italia, España y USA (Florida) resultan beneficiados, con relación a los demás.

Tabla 2	
País	Probabilidad predictiva
Brazil	0.2510
Cuba	0.3382
Italia	0.4868
España	0.4558
Argentina	0.1354
E.U.A. (Florida)	0.4424
Sudáfrica	0.4052
Turquía	0.4006
Costa Rica	0.4016

5. Agradecimientos

Los autores agradecen el apoyo de la Asociación Mexicana de Cultura A.C. para la realización de este proyecto.

Referencias

Raftery, A.E. and Lewis, S.M. (1992) How many iterations in the Gibbs sampler ? In *Bayesian Statistics, Vol. 4* (Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M., eds.) Oxford University Press, pp. 763-773.

Raftery, A.E. and Lewis, S.M. (1992b) One long run diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science*, Vol. 7, pp. 493-497.

Raftery, A.E. and Lewis, S.M. (1995) The number of iterations, convergence diagnostics and generic Metropolis algorithms. In *Practical Markov Chain Monte Carlo* (Gilks, W.R., Spiegelhalter, D.J. and Richardson, S. eds.) Chapman and Hall.

Reducción de dimensión en regresión a través de gráficas y métodos de regresión inversa.

Jorge de la Vega Góngora¹

Banco de México

1. Introducción

En la práctica muchos problemas de regresión involucran un número grande de predictores, así que uno de los primeros problemas a los que se enfrenta el analista es el de cómo reducir la dimensión de su modelo. Lo ideal sería reducir el problema de tal forma que sea posible capturar toda la información disponible en uno o dos predictores para poder visualizar la relación entre las variables en una gráfica bi o tridimensional. Cook y Weisberg proveen un marco teórico en donde las gráficas juegan un papel muy importante en el proceso de reducir la dimensión de un modelo.

Adicionalmente, se cuenta con varias técnicas numéricas que permiten hacer inferencia sobre la dimensión de un problema. Entre los métodos usados , se cuenta con SIR (sliced inverse regression), SAVE (slice average variance estimation) y pHd (principal Hessian directions). En esta platica se introducirán algunas ideas relevantes del marco teórico y se discutirán los métodos numéricos en el contexto de Reducción de Dimensión, además de algunos problemas que requieren mayor estudio. También se mostrará cómo implementar las ideas sugeridas en un sistema para análisis de regresión, basado en Lisp-Stat, llamado Arc.

2. Regresión

Regresión es el estudio de la distribución condicional de una respuesta y dados los valores de los predictores x_1, \dots, x_p :

$$F(y|x_1, \dots, x_p) = F(y|x)$$

La variable de respuesta puede ser discreta o continua, al igual que los predictores. Típicamente

¹jvega@banxico.org.mx

interesan los momentos de F , particularmente la *función media* $E(y|x)$ o la *función varianza* $\text{VAR}(y|x)$. Otras posibilidades son la función cuantíl $t_\alpha(x) = P(Y > \alpha|x)$ o la mediana. En términos generales, no se asume ningún modelo paramétrico en particular para estas funciones.

La reducción de dimensión es estudiada bajo el siguiente esquema: Se quiere encontrar una matriz de rango completo B de dimensiones $p \times q$ con $q << p$ tal que

$$F(y|x) = F(y|B'x) \quad \forall x \in \mathcal{X}$$

Esto es equivalente a pedir que $y \perp x|B'x$. La dimensión del problema de regresión es q . Las columnas de $B'x$ son los *predictores suficientes*. Una gráfica de y versus las columnas de $B'x$ es una *gráfica sumaria suficiente*. Idealmente un problema será D0 (y y x son independientes), D1 o D2.

Varios modelos entran dentro de estas categorías. Por ejemplo los modelos de regresión lineal simple $E(Y|x) = \beta'x$ con varianza constante $Var(y|x) = \sigma^2$ o con varianza no constante y proporcional a $\beta'x$:

$$E(Y|x) = \beta'x, \quad Var(y|x) = \sigma^2(\beta'x)$$

También los modelos de regresión no lineal $E(Y|x) = g(\beta'x, \theta)$ con g conocida, o bien los modelos aditivos $E(Y|x) = \alpha_0 + \sum_{i=1}^p g_i(\beta'x)$ con g_i desconocidas. Otro caso es el denominado *Projection pursuit*, o bien los *Single index models*: $E(Y|x) = g(\beta'x)$ con g desconocida.

Una idea similar se puede encontrar en el método de Componentes Principales (CP), pero hay diferencias importantes:

En CP el objeto del problema es encontrar una matriz Q tal que $Q'x|y$ capture la máxima variabilidad en los predictores originales. Sin embargo,

- No hay garantía de que $F(y|x) = F(y|Q'x)$
- Requiere que $x|y$ sea normal.
- No toma en cuenta la información proporcionada por y .
- Utiliza un criterio para reducir dimensión basado en la variabilidad de x .

Lo que se propone es una metodología más general para reducir dimensión en problemas condicionales.

3. Subespacios de reducción de dimensión

El subespacio de reducción de dimensión (SRD) es el espacio vectorial generado por las columnas de B ; se denota por $\mathcal{S}(B)$.

La existencia se garantiza trivialmente, uno es simplemente $\mathcal{S}(I_p)$. Claro, no muy útil.

los SRD no son únicos en todos los casos; cuando lo es para un problema específico y se satisfacen ciertas condiciones, al SRD se le llama el subespacio central y se le denota como $\mathcal{S}_{y|x}$.

Cuando sólo se tiene interés en la función media, se considera sólo es subespacio medio central $S_{E(y|x)}$.

Algebraicamente, el problema de estimación del subespacio central se reduce a encontrar una base para el subespacio y así poder obtener una gráfica sumaria suficiente.

4. Métodos para estimar el espacio central

Hay cuatro métodos primarios para estimar una base para $\mathcal{S}_{y|x}$:

- Mínimos cuadrados y otros métodos basados en funciones objetivo convexas.
- Sliced inverse regression (SIR, Li 1991 y Cook 1998),
- Sliced average variance estimation (SAVE, Cook y Weisberg, 1991) y
- Principal Hessian directions (pHd) (pHd, Li, 1992).

Además de la existencia de $\mathcal{S}_{y|x}$ se requieren las siguientes condiciones:

Condición de linealidad Para todos los predictores x_j ,

$$E(x_j|B'x) = a_j + b'_j(B'x)$$

Siendo B desconocido, en la práctica se puede asegurar la condición para todas las B posibles. Esta es la condición de *predictores linealmente relacionados* (PLR). Esta condición es requerida por todos los métodos.

Además, SIR y SAVE requieren de la siguiente condición para hacer inferencia:

Condición de covarianza constante La condición de varianza constante es requerida por algunos métodos. Para todos los predictores x_i, x_j

$$\text{Cov}(x_j, x_k|B'x) = \sigma_{j,k}$$

Ambas condiciones son satisfechas cuando los predictores tienen una distribución normal, pero también pueden satisfacerse bajo otros tipos de distribuciones, como distribuciones elípticas. Las condiciones se pueden aproximar buscando transformaciones que normalicen los predictores.

¿Cuál es el papel de los métodos de estimación tradicionales en este enfoque? Si el modelo 1D

$$E(y|x) = M(\beta'x) \text{ y } \text{Var}(y|x) = V(\beta'x)$$

se cumple, *aún sin conocer* M y V , y los predictores satisfacen PLR, entonces Resultado de Li-Duan (1989):

$$\beta_{OLS} \in \mathcal{S}_{y|x}$$

A partir de la gráfica de y versus \hat{y} se puede decir algo sobre M y V .

Los métodos numéricos restantes para estimar el SRD se basan en el siguiente procedimiento: encontrar un estimador consistente \hat{M} de una matriz poblacional M que dependa de algún procedimiento particular, con la propiedad de que

$$\mathcal{S}(M) \subseteq \mathcal{S}_{y|x}.$$

Las inferencias de al menos una parte de $\mathcal{S}_{y|x}$ se pueden basar en \hat{M} , siendo sus columnas elementos del subespacio central y la inferencia se basa casi siempre en la suma de funciones de los valores propios de \hat{M} .

¿Cómo es que la regresión inversa $x|y$ puede ser informativa para $y|x$? Si PRL se satisface, entonces $\mathcal{S}_{E(x|y)} \subseteq \mathcal{S}_{y|x}$. Se puede probar que $\mathcal{S}(\text{VAR}(E(x|y))) = \mathcal{S}_{E(x|y)}$, por lo que $M = E(\text{VAR}(x|y))$. M se puede estimar consistentemente usando suavizamiento:

1. Divide el rango de y en h subintervalos y reemplaza y por una versión discretizada \tilde{y}
2. En cada subintervalo s , calcula la media muestral de x , \bar{x}_s
3. Define $\hat{M} = \frac{1}{n} \sum_{s=1}^h n_s \bar{x}_s \bar{x}_s'$, donde n_s es el número de observaciones en el subintervalo s .

5. Pruebas de hipótesis

Para probar hipótesis del tipo

$$H_0 : d = m \text{ vs. } H_1 : d \geq m,$$

se utiliza la estadística

$$\hat{\Lambda}_m = n \sum_{j=m+1}^p \hat{\lambda}_j$$

donde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p$ son los eigenvalores de \hat{M} . La distribución asintótica de $\hat{\Lambda}_m$ es una combinación lineal de distribuciones $\chi^2_{(1)}$

Para recobrar direcciones en el subespacio central, podemos fijarnos en momentos superiores. Cook y Weisberg (1991) se basan en la siguiente relación:

$$I - \text{VAR}(x|y) = P_B(I - \text{VAR}(x|y))P_B$$

Entonces $\mathcal{S}(I - \text{VAR}(z|y)) \subseteq \mathcal{S}_{y|z}$ y $M = (I - \text{VAR}(z|y))^2$. Un estimado de M se obtiene de la siguiente fórmula

$$\hat{M} = \sum_{s=1}^h \frac{n_s}{n} (\widehat{I - \text{VAR}(z|y) \in I_s})^2.$$

Se utiliza la misma estadística de prueba $\hat{\Lambda}_m$ que en el caso de SIR. Sin embargo, no se conoce la distribución asintótica de $\hat{\Lambda}_m$, sólo se conoce en el caso de que la respuesta es binaria y en ese

caso se distribuye como una combinación lineal de χ^2 . La distribución de $\hat{\Lambda}_m$ se puede estimar indirectamente usando pruebas de permutación.

Se pueden aplicar pruebas de permutación no paramétricas para d . Para probar la hipótesis

$$H_0 : d = 0 \text{ vs. } H_1 : d \geq 1,$$

- i. los n valores de la respuesta y se permutan al azar C veces.
- ii. Para cada permutación, se calcula el valor de $\hat{\Lambda}_0^{(k)}$ de la estadística de prueba del método en cuestión con $m = 0$, para $k = 1, \dots, C$.
- iii. Se calcula el p -value como la fracción de $\hat{\Lambda}_0^{(k)}$'s que exceden $\hat{\Lambda}_0$, el valor de la estadística con los datos originales.

El procedimiento para probar $d = k$ vs. $d > k$ es el mismo excepto que ahora los índices i de los vectores $(y_i, \hat{\beta}_1' x_i, \dots, \hat{\beta}_k' x_i)$ se permutan C veces.

Cuando $x \sim \mathcal{N}(\mu, \Sigma)$, es posible asociar la matriz Hessiana $H(x) = \frac{\partial E(y|x)}{\partial x \partial x'}$ al subespacio central. Como $E(y|x) = E(y|B'x)$ para $B \in \mathcal{S}_{y|x}$, se sigue que

$$H(x) = B[H(B'x)]B'$$

y los eigenvectores de $H(x)$ viven en $\mathcal{S}_{y|x}$. Como H típicamente varía con x a menos que la superficie sea de dimensión menor que 2, se sustituye H por $E(H)$, y bajo el supuesto de normalidad, $E(H) = M$, donde $M = E((y - E(y)) \times xx')$. Si \mathcal{S}_{yxx} denota $\mathcal{S}(M)$, entonces $\mathcal{S}_{yxx} \subset \mathcal{S}_{y|x}$. Un estimador consistente de M se obtiene de

$$\hat{M} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}) \hat{x}_i \hat{x}_i'.$$

Los métodos de estimación SIR,SAVE y pHd están implementados en Arc (Weisberg y Cook, 1999-2004) que es un programa que se basa en Xlisp-Stat (Tierney, 1989) para hacer análisis de regresión. El programa se puede obtener gratuitamente de www.stat.umn.edu/arc. Los métodos también se encuentran disponibles en el paquete **dr** de R y S-plus.

Referencias

- Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. New York: Wiley.
- Cook, R. D. (1998a). *Regression Graphics: Ideas for studying regressions thru graphics*. New York: Wiley.
- Cook, R. D. (1998b). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, 93, 84-100.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86, 316-342.
- Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87, 1025-1039.
- Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *Annals of Statistics*, 17, 1009-1052.

Análisis de Costos en la Experimentación Industrial

Jorge Domínguez Domínguez¹

Centro de Investigación en Matemáticas, A.C.

1. Introducción

Las técnicas estadísticas del diseño de experimentos (dde) son herramientas eficientes para adquirir de manera rápida un conocimiento adicional de productos y procesos, principalmente en las variables de calidad. Sin embargo, los costos deben ser considerados tanto en el planteamiento experimental como en el análisis. Los costos asociados en la estrategia experimental son inevitables en la práctica y estos tienen impacto en la economía y finanzas de las empresas.

Varios métodos se han propuesto para optimizar la media y la varianza, no obstante, los costos y la razón de no conformidades asociados con los niveles óptimos de los factores no han sido suficientemente explorados desde un punto de vista riguroso ni explícito. Nuestro interés en este trabajo es plantear una estrategia de optimización, tal que, permita explorar soluciones alternativas para determinar los niveles de los factores que minimicen costos de producción sin afectar las características de calidad. Dentro de este concepto cabe lo que se conoce como diseño de tolerancias.

Dado que una variedad importante de problemas reales tienen más de una respuesta, el análisis de costos se puede extender a el caso de multirrespuesta. Para redondear este resumen al final se indicarán algunas líneas de investigación relacionadas con este trabajo.

2. Descripción de los problemas a estudiar

Considere que la variable Y representa una característica de calidad de un producto y que esta se ve afectada por una serie de factores $x^t = (x_1, \dots, x_k)$ que intervienen en el proceso. Por lo general se desea que un producto tenga un valor objetivo de calidad, el que se denota por M , este valor de M puede ser un mínimo, máximo o un valor puntual, al procedimiento estadístico que permite determinar este valor se le conoce como diseño de parámetro. El primer planteamiento de la relación

¹jorge@cimat.mx

de costo en una estrategia experimental se da mediante el uso de la conocida función de pérdida, propuesta inicialmente en el esquema de diseño robusto y se expresa por:

$$P(Y(x)) = k(Y(x) - M)^2. \quad (1)$$

Donde k es el costo de calidad asociado a una unidad producida Li y Wu (1999). $Y(x)$ es el modelo de regresión en función de los factores x es decir: $Y(x) = \beta_0 + x^t\beta + x^tBx + \varepsilon$, con β_0 constante, $\beta = (\beta_1, \dots, \beta_k)$ un vector de parámetros $B = (\beta_{11}, \dots, \beta_{1k}, \beta_{k1}, \dots, \beta_{kk})$ matriz de parámetros de segundo orden, y $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. La esperanza de la expresión (1) permite que el promedio de la pérdida de calidad se descomponga en dos términos, es decir:

$$E(P(Y(x))) = k(\sigma^2(x) + (\mu(x) - M)^2). \quad (2)$$

Se observa de la expresión anterior que se puede realizar un trabajo experimental tal que $\mu(x)$ coincida con el valor objetivo y además minimice la varianza, entonces el promedio de la pérdida disminuye. En ese sentido los datos que se obtienen del experimento permitirán optimizar el modelo de regresión entorno al valor objetivo M . Así la pérdida esperada es:

$$\hat{P}(\hat{Y}(x)) = k(\hat{\sigma}^2(x) + (\hat{Y}(x) - M)^2). \quad (3)$$

Varios autores han trabajado sobre este esquema por ejemplo Kim y Cho (2000). La descripción gráfica de este planteamiento se muestra en la Figura 1.

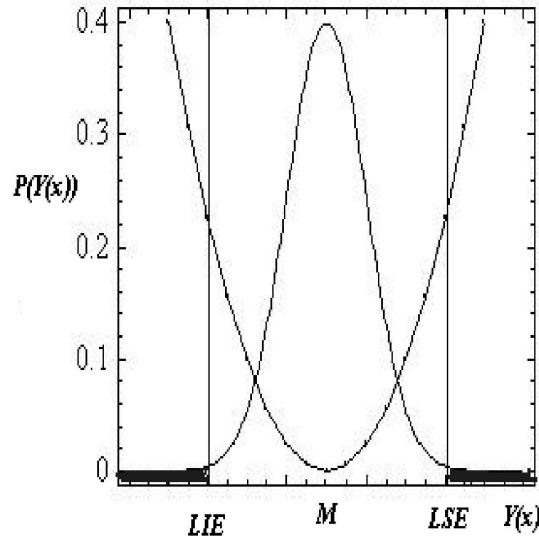


Figura 1: M es el valor objetivo, la región sombreada denota los no conformes.

El procedimiento para reducir la varianza desempeña un papel importante en la práctica, sin embargo lograr esta meta usualmente causa un incremento en el costo de manufactura porque requiere: procedimientos operacionalmente más precisos, mejores medios de operación y técnicos mejor entrenados. En ese sentido disminuir la varianza de un proceso implica tolerancias más estrechas.

Una segunda exposición en relación al costo esta en función de la tolerancia. La tolerancia se define mediante los límites de especificación, esto es $Tolerancia = LSE - LIE$, Figura 1. De tal manera los productos que caigan fuera de estos límites dan lugar a **costos** por unidades no conformes. La razón de no conformes estimada para la respuesta Y es:

$$\widehat{R} = P\left(Z < \frac{LIE - \widehat{Y}(x)}{\widehat{\sigma}(x)}\right) + P\left(Z > \frac{LSE - \widehat{Y}(x)}{\widehat{\sigma}(x)}\right), \quad (4)$$

número de no conformes por millón. La función de costo asociado a lo factores se puede plantear como:

$$FC = 0,5 \left\{ \sum_i^m C_i^+ + \sum_i^m C_i^- + \sum_i^m (C_i^+ - C_i^-) x_i \right\}, \quad (5)$$

donde C_i^+ y C_i^- son los costos asignados a los niveles 1 y 2 del diseño factorial 2^k y $m = 2^{k-1}$ Kraus et al. (2000).

Se pueden modelar las tolerancias para reducir costos, considerando este planteamiento se da una solución conjunta al diseño de parámetro y diseño de tolerancias. En esa dirección es suficiente con extender la expresión (3) tal que tome en cuenta el marco de la tolerancia.

$$\widehat{P}(\widehat{Y}(x)) = k(\widehat{\sigma}^2(x) + (\widehat{Y}(x) - M)^2) + \sum_i^n f(u_i), \quad (6)$$

donde $u_i = 3\sigma$, y $f(u)$ conocida como función de transferencia denominadas modelos de costos. Algunas formas matemáticas de estos modelos son: $f(u) = c_0 + c_1 u^r$, $r = 1, 2$. o $f(u) = c_0 + c_1 e^{c_2 u}$. En ese sentido estos modelos sugieren que los costos de manufactura están en función de la varianza del proceso. Es frecuente que en los procesos de producción exista más de una variable de respuesta por lo tanto la expresión anterior se extiende de manera natural al caso multivariado, Ribeiro et al. (2001) y Romano et al. (2004).

3. Situaciones para la optimización de costos

Los planteamientos sobre los casos de costos realizados en el apartado anterior se formalizarán en esta sección dentro del contexto de la la metodología de superficie de respuesta, Myers (1999). En resumen se puede decir que existen cinco etapas en los estudios de optimización experimental de costos: 1. identificar el problema, 2. planear y efectuar el experimento, 3. modelar la media, la varianza y/o alguna función relacionada con costos, 4. seleccionar la función de utilidad y el criterio de optimización, y 5. aplicar el procedimiento de optimización.

3.1. Diseño de parámetro minimizando la función de pérdida

Se puede emplear la expresión (3) para estudiar los costos de no calidad en cualquier proceso industrial y optimizar la función de pérdida. El procedimiento en forma de algoritmo es como sigue:

1. Determinar el valor de k , si la característica de calidad Y difiere del valor nominal M . Para ello se establece la pérdida en pesos, se despeja k de la expresión (1), denomine ese valor k_o .
2. Realizar el experimento siguiendo un diseño apropiado, por ejemplo un diseño factorial fraccionado 2^{k-p} , un diseño central compuesto o un diseño factorial incompleto o completo con factores en tres niveles entre otros.
3. Modelar los resultados experimentales. En ese sentido obtener los modelos de regresión para la media $\hat{Y}_1(x)$, y para la desviación estándar $\hat{Y}_2(x)$ evaluar cada modelo estadísticamente, es decir analizar la significancia, falta de ajuste y estimar el coeficiente de determinación.
4. Optimizar la función de pérdida para minimizar costos, $k_o(\hat{Y}_2(x) + (\hat{Y}_1(x) - M_c)^2)$. El planteamiento de optimización es:

$$\begin{aligned} & \text{Minimizar } \hat{Y}_2(x) \\ & \text{Sujeto a } \hat{Y}_1(x) = M_c \\ & x \in R(x), \end{aligned}$$

donde M_c es mínimo, máximo o valor objetivo, Domínguez (2004).

Observación. Para modelar $\hat{Y}_2(x)$ se presentan varios escenarios, el caso directo es cuando en los experimentos hay réplicas. Si en el trabajo experimental no existen réplicas una alternativa para

modelar $\widehat{Y}_2(x)$ es mediante la siguiente expresión: $\log(\text{abs}(Y_1(x) - \widehat{Y}_1(x)))$. Otra situación que se presenta en los estudios experimentales es la existencia de factores de ruido $z^t = (z_1, \dots, z_q)$ en ese situación el modelo de $\widehat{Y}_2(x)$ se obtiene mediante: $\widehat{\text{Var}}(\widehat{Y}_1(x|z)) = (\widehat{\gamma} + \widehat{D}x)(\widehat{\gamma} + \widehat{D}x)^t + \widehat{\sigma}_e^2$, donde $\widehat{\gamma}$ es el estimador de los parámetros del factor de ruido z , bajo el supuesto de que $z \sim N(0, 1)$, \widehat{D} es la matriz cuyos elementos son los estimadores de los efectos de interacción entre los factores x y z , Romano et al. (2004).

3.2. Optimización de costos en el caso de no conformes

Al planear y efectuar el experimento para el análisis de costos en cada tratamiento también se contemplan las unidades experimentales que son no conformes. En ese sentido ahora se tendrán cuatro respuestas a saber: 1. la media $Y_1(x)$, 2. la desviación estándar $Y_2(x)$, 3. el costo $Y_3(x)$ y no conformes $Y_4(x)$. El modelo de optimización para estas condiciones se puede dar en varios escenarios, se plantearán 2 posibles.

1. Optimizar el costo:

$$\begin{aligned} &\text{Minimizar } \widehat{Y}_3(x) \\ &\text{Sujeto a } \widehat{Y}_1(x) = M_c \\ &\quad \widehat{Y}_2(x) = T \\ &\quad \widehat{Y}_4(x) < 0,001 \\ &\quad x \in R(x). \end{aligned}$$

2. Optimizar no conformes

$$\begin{aligned} &\text{Minimizar } \widehat{Y}_4(x) \\ &\text{Sujeto a } \widehat{Y}_1(x) = M_c \\ &\quad \widehat{Y}_2(x) = T \\ &\quad \widehat{Y}_3(x) = S \\ &\quad LIE = t_o \text{ y } LSE = t_1, \quad x \in R(x). \end{aligned}$$

donde M_c , T , S , t_o y t_1 son valores de referencia.

3.3. Optimización en la tolerancia relacionada a costos

Nuevamente se retoma la estrategia de los cinco pasos citados, aunque en este procedimiento es más general porque integra el diseño de parámetro y el de tolerancias. Además incorpora los costos, la estructura de optimización queda de la siguiente manera:

$$\begin{aligned} &\text{Minimizar } \widehat{P} = k(\widehat{Y}_2(x, u) + (\widehat{Y}_1(x, u) - M)^2) + \sum_i^n f_i(u_i) \\ &\text{Sujeto a } \widehat{Y}(x, u) \leq T_{\text{máx}} \\ &\quad u_i^I \leq u_i \leq u_i^S \\ &\quad x \in R(x). \end{aligned}$$

donde $\mathbf{u} = 3\sigma$, u_i^I y u_i^S son los límites inferior y superior para u_i y T_{max} representa el máximo permitible para la tolerancia. Una particularidad importante aquí es la generación de los modelos: $\widehat{Y}_2(x, u)$ y $(\widehat{Y}_1(x, u)$, estos se obtienen mediante una aplicación del método de expansión de Taylor. Esto es: $\widehat{Y}_1(x, u) = \widehat{Y}(x) + \frac{1}{2} \sum_i^n \frac{1}{9} \widehat{\beta}_{ii} u_i^2$, y $\widehat{Y}_2(x, u) = \widehat{Y}(x, u) + \sigma_\varepsilon^2$, donde $\widehat{Y}(x, u) = \sum_i^n \frac{1}{9} (\widehat{\beta}_i + 2\widehat{\beta}_{ii} x_i + \sum_{j=i+1}^n \widehat{\beta}_{ij} x_j + \sum_{j=1}^{i-1} \widehat{\beta}_{ji} x_j)^2 u_i^2$.

4. Discusión

Sin duda estos procedimientos de optimización resultan de mucha utilidad en la práctica porque permite sustanciales ahorros a los empresarios, además sin sacrificar las otras variables de calidad del proceso. Falta ilustrar la aplicación de estos planteamientos de optimización a casos reales, por cuestiones de espacio en este resumen no es posible. Sin embargo a lector interesado puede solicitar la referencias a casos prácticos al correo electrónico mencionado. Una vez realizada esta presentación se indican posibles extenciones y líneas de estudio al esquema de optimización que aquí se han trabajado. Un extensión es considerar restricciones de aleatorización en el planteamiento experimental. Una línea de investigación es tomar en cuenta el caso multivariado desde un enfoque de la metodología de superficie de respuesta, Allen y Yu (2002).

Agradecimientos: Al Consejo de Ciencia y Tecnología de Guanajuato por su apoyo al proyecto por su apoyo al proyecto GTO-2002-C01-6137.

5. Referencias

Allen, T. and Yu, L. (2002). “Low-Cost Response Surface Method from Simulation Optimization”. *Qual. Reliab. Engng. Int.*, 18, pp.5-17.

Domínguez, D.J. (2004). “Estrategia Experimental una Opción para la Mejora Continua de Procesos Industriales”. *Gaceta del Concyteg*.

Ribeiro, J. L., Fogliatto, F. S. and Caten, C. S. (2001). “Minimizing Manufacturing and Quality Costs in Multiresponse Optimization”. *Quality Engineering*, 13(4), pp. 559-569.

Kim, Y.J. and Cho, B.R. (2000). "Economic Considerations on Parameter Design". *Qual. Reliab. Engng. Int*, 16, pp.501-514.

Kraus, P.D., Benneyan, J.C. and Mackertich, N. A. (2000). "Use of Mathematical Programming in the Analysis of Constrained and Unconstrained Industrial Experimental". *Quality Engineering*, 12(3), pp 395-406.

Li, W. and Wu C. F. J. (1999). "An Integrated Method of Parameter Design and Tolerance Design". *Quality Engineering*, 11(3), pp 417-425.

Myers, R. H. (1999). "Response Surface Methodology- Current Status and Future Directions". *Journal of Quality Technology* 31, pp. 30-44.

Romano, D., Varetto, M. and Vicario G. (2004). "Multiresponse Robust Design: A General Framework Based on Combined Array". *Journal of Quality Technology* 36(1), pp. 27-37.

Efectos Aleatorios en Modelos de Elección Discreta

José Ramón Domínguez Molina¹

Rogelio Ramos Quiroga²

Centro de Investigación en Matemáticas, A.C.

1. Introducción

En este trabajo se describen los modelos de elección discreta con efectos aleatorios que más se utilizan actualmente en la modelación de datos de panel y variación en los gustos; discutiremos además algunos aspectos de estimación de parámetros mediante el método de máxima verosimilitud. Los modelos descritos se basan en el principio de Utilidad Aleatoria que consiste en que un individuo selecciona un producto entre K alternativas de forma tal que maximiza su utilidad, esto significa que si U_{nj} es la utilidad que el individuo n percibe en la alternativa j , entonces n elige i si

$$U_{ni} > U_{nj}, \quad \forall j \neq i.$$

Sin embargo, el investigador no observa directamente las utilidades del individuo n , sólo la decisión tomada por el individuo y una *proxy* de su utilidad, V_{ni} , definida por los atributos de los productos en competencia. Un modelo simple pero suficientemente general, es el dado por proxys lineales,

$$U_{ni} = V_{ni} + \varepsilon_{ni} = x_{ni}\beta + \varepsilon_{ni} \tag{1}$$

donde los supuestos distribucionales de ε y de β originan todos los modelos de este trabajo.

2. Modelo Logit Multinomial

McFadden (1974) demostró que si la parte fija de la función de utilidad es lineal como en la ecuación (1) y los errores son independientes e idénticamente distribuidos Valor Extremo tipo I, entonces la probabilidad de que la persona n elija el producto i es,

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \tag{2}$$

¹jrguez@cimat.mx

²rramosq@cimat.mx

Este modelo tiene el supuesto implícito de que no hay variaciones en los gustos del individuo n , eso es, dado un vector de atributos x_{ni} , siempre tendrá una probabilidad fija P_{ni} de elegir la alternativa i ; dada la estructura tan restrictiva que se asume en los errores no es posible captar variación en los gustos. Una propuesta para modelar datos de panel, descrita en Train (2002), consiste en modelar la utilidad del individuo n ante la alternativa i en su t -ésima situación de elección para $t = 1, \dots, T$, como

$$U_{nit} = V_{nit} + \varepsilon_{nit} \quad \forall i, t. \quad (3)$$

Si ε_{nit} se distribuye de tipo valor extremo independientemente de n, i y t , entonces las probabilidades de elección están dadas por

$$P_{nit} = \frac{e^{V_{nit}}}{\sum_j e^{V_{njt}}}$$

Para capturar el efecto de observaciones repetidas sobre un mismo individuo se propone que la parte fija de la utilidad sea

$$V_{nit} = \alpha y_{ni(t-1)} + x_{nit}\beta$$

donde $y_{ni(t-1)}$ es igual a 1 si la persona eligió el producto i en la elección anterior. Por lo tanto, la utilidad del producto elegido en la ocasión $t - 1$ se incrementará en α en la elección al tiempo t , esto para $\alpha > 0$; cuando $\alpha < 0$ tendremos individuos en búsqueda de *variedad*, esto es, su utilidad es mayor cuando eligen alternativas diferentes a la escogida en la elección anterior.

2.1. Estimación

Bajo el modelo logit multinomial (2), la contribución de la persona n a la verosimilitud puede ser expresada como,

$$\prod_{i=1}^K P_{ni}^{y_{ni}}$$

donde $y_{ni} = 1$ si la persona n elige el producto i y cero en otro caso. Ahora, asumiendo que cada persona elige de manera independiente la verosimilitud queda dada por,

$$L(\beta) = \prod_{n=1}^N \prod_{i=1}^K P_{ni}^{y_{ni}}, \quad (4)$$

donde β es el vector de parámetros contenidos en el modelo dado por (1) y P_{ni} está dado por (2). La función de logverosimilitud es entonces,

$$l(\beta) = \sum_{n=1}^N \sum_{i=1}^k y_{ni} \ln P_{ni}$$

McFadden(1974) demostró que si la función de utilidad es lineal como en (1) entonces la verosimilitud es cóncava, por lo que su máximo es único, además como el cálculo del gradiente para esta función es directo, los métodos de maximización son fácilmente implementables y con convergencia rápida.

El proceso de estimación para el modelo (3) es similarmente implementable, la inclusión de variables con retraso no acarrea estimadores inconsistentes siempre y cuando asumamos independencia con los errores; por otro lado, el supuesto de independencia *entre* los errores es muy fuerte considerando que el modelo (3) implica observaciones repetidas en un mismo individuo; en este caso se recomienda el modelo Probit que a continuación se describe.

3. Modelo Probit

El modelo Probit asume que el vector aleatorio ε_n que contiene los errores de cada alternativa, se distribuye normal multivariado, con una matriz de covarianza general, esto es,

$$U_{ni} = V_{ni} + \varepsilon_{ni} \quad (5)$$

donde $\varepsilon_n = (\varepsilon_{n1}, \dots, \varepsilon_{ni}, \dots, \varepsilon_{nk})' \sim N(\mathbf{0}, \Sigma)$. El hecho de permitir una estructura general en la matriz de varianzas y covarianzas hace que este modelo sea flexible para poder modelar variación en los gustos y datos de panel, la manera de trabajar en este caso es muy semejante a los modelos usados en el área de datos longitudinales.

Modelando variación en los gustos. La manera de modelar variación en los gustos es suponer que los parámetros asociados a los pesos de los atributos tienen una distribución normal con media \mathbf{b} y matriz de varianzas y covarianzas \mathbf{W} .

$$\begin{aligned} U_{ni} &= x_{ni}\beta + \varepsilon_{ni}, \quad \beta_n \sim N(b, W), \quad \varepsilon_n \sim N(0, \Delta) \\ U_{ni} &= x_{ni}b + x_{ni}\tilde{\beta}_n + \varepsilon_{ni}, \quad \tilde{\beta} = b - \beta_n \\ U_{ni} &= x_{ni}b + \eta_{ni}, \quad Var(\eta_n) = x_{ni}'Wx_{ni} + \Delta = \Sigma \end{aligned}$$

Este caso se puede simplificar para tener un modelo con parámetros fijos aunque con una estructura de error un poco más compleja. La probabilidad de elección está dada por

$$\begin{aligned} P_{ni} &= \Pr(U_{ni} > U_{jn} \quad \forall j \neq i) = \Pr(V_{ni} + \varepsilon_{ni} > V_{jn} + \varepsilon_{jn} \quad \forall j \neq i) \\ P_{ni} &= \int I(V_{ni} + \varepsilon_{ni} > V_{jn} + \varepsilon_{jn} \quad \forall j \neq i) \phi(\varepsilon_n) d\varepsilon_n \end{aligned} \quad (6)$$

Modelando datos de panel. En este caso se supone que la utilidad observada por el individuo n sobre el producto i en la situación de elección t , esta dada por,

$$U_{nit} = V_{nit} + \varepsilon_{nit} \quad (7)$$

donde los errores son correlacionados en el tiempo, por lo tanto la matriz de covarianza de $(\varepsilon_{n11}, \varepsilon_{n21}, \dots, \varepsilon_{nK1}, \varepsilon_{n12}, \varepsilon_{n22}, \dots, \varepsilon_{nK2}, \dots, \varepsilon_{n1T}, \varepsilon_{n2T}, \dots, \varepsilon_{nKT})$ es

$$\text{Cov}(\varepsilon_{it}, \varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2 + \sigma_t^2 & \text{para } i = j \text{ y } t = s \\ \sigma_\varepsilon^2 & \text{para } i \neq j \text{ y } t = s \\ \sigma_t^2 & \text{para } i = j \text{ y } t \neq s \\ 0 & \text{para } i \neq j \text{ y } t \neq s \end{cases}$$

las probabilidades de elección correspondientes tienen la misma estructura que en (6).

3.1 Estimación

En este caso la logverosimilitud es la misma que en el caso dado por el modelo logit multinomial (4), pero la probabilidad de elección está dada por (6), esta expresión no es cerrada y es difícil de evaluar analíticamente, típicamente en la literatura se utilizan métodos de integración numérica tipo Montecarlo.

4. Modelo Logit Mixto

El modelo Logit Mixto es altamente flexible, puede aproximarse a cualquier modelo de utilidad aleatoria (McFadden y Train, 2000). En este modelo es obvio que se pueden modelar variaciones de los gustos y datos de panel, además no está restringido a que los parámetros de la utilidad sean normales como en el modelo probit. La utilidad esta dada por

$$U_{ni} = V_{ni} + \varepsilon_{ni} \quad (8)$$

donde $V_{ni} = x_i \beta$ y además $\beta \sim \phi(b, W)$.

Modelando variación en los gustos. En este caso es muy natural la modelación de variación en los gustos ya que desde la definición de la utilidad se da por hecho que los parámetros pueden ser aleatorios, en este caso,

$$P_{ni|\beta} = \frac{e^{x_{ni}\beta}}{\sum_j e^{x_{nj}\beta}} \quad \text{y} \quad P_{ni} = \int \left(\frac{e^{x_{ni}\beta}}{\sum_j e^{x_{nj}\beta}} \right) \phi(\beta|b, W) d\beta \quad (9)$$

donde esta integral típicamente es evaluada vía simulación.

Modelando datos de panel. Una especificación muy simple del modelo logit mixto que sirve para modelar datos de panel consiste en asumir que los coeficientes de la utilidad varían entre las personas pero permanecen constantes en cada elección que hacen las personas en el tiempo. La utilidad que una persona n observa sobre un producto i en su situación de elección t es,

$$U_{njt} = x_{njt}\beta_n + \varepsilon_{njt}$$

donde ε_{njt} se asume *iid* sobre las personas, las alternativas y el tiempo. Supongamos ahora que una persona hace una elección en cada uno de T escenarios, entonces la probabilidad de que la persona seleccione el producto i , dado los parámetros esta dada por,

$$P_{ni|\beta} = \prod_{t=1}^T \left[\frac{e^{x_{nit}\beta}}{\sum_j e^{x_{njt}\beta}} \right]$$

la probabilidad sin condicionar sobre los parámetros está dada por la siguiente integral,

$$P_{ni} = \int \prod_{t=1}^T \left[\frac{e^{x_{nit}\beta}}{\sum_j e^{x_{njt}\beta}} \right] \phi(\beta|b, W) d\beta \quad (10)$$

4.1 Estimación

La logverosimilitud es la misma que en el caso dado por el modelo logit multinomial (4), pero la probabilidad de elección está dada por (9) o por (10) y, como en el caso del modelo Probit, no se tienen expresiones cerradas y por lo tanto, se recurre a métodos de integración numérica.

5. Conclusiones

El modelo Logit Mixto es construido a partir de asumir términos de error adicionales que implican una matriz de covarianza del error heterocedastica y correlacionada. El modelo Probit solamente asume distribuciones de tipo normal para los parámetros, mientras que el Logit Mixto puede asumir cualquier distribución. El modelo Probit no tiene una expresión cerrada para las probabilidades de elección, por lo que necesita ser aproximada mediante simulación. El modelo Logit Mixto tiene una expresión cerrada para la probabilidad de elección dados los parámetros, lo cual la hace fácilmente interpretable. La dimensión del número de integrales numéricas en el caso Probit es $J - 1$, mientras que en el caso Logit Mixto es K , por lo que si $K < J - 1$, entonces existiría una ventaja de estimación para el modelo Logit Mixto en comparación del Probit.

Para modelar variación en los gustos es posible utilizar un modelo Probit o Logit Mixto. Si se desea modelar heterocedasticidad, entonces se debe recurrir a modelos cuya matriz de covarianza acepte elementos distintos en la diagonal, como lo hacen los modelos Logit Mixto y Probit. Mediante simulación es posible probar que si no se modelan adecuadamente los modelos en presencia de correlación y/o heterocedasticidad las estimaciones de los parámetros resultan con una gran cantidad de sesgo.

Referencias

Jordan J. Louviere, David A. Hensher y Joffre D Swait (2000). *Stated Choice Methods, Analysis and Application*. Cambridge University Press.

Kenneth E. Train (2002), *Discrete Choice Methods with Simulation*. Cambridge University Press.
<http://elsa.berkeley.edu/~train>

McFadden, D. (1974) Conditional logit analysis of qualitative choice behaviour, *Frontiers in Econometrics*, Academic Press, New York, pp. 105-142.

Ricardo Álvarez Daziano y Marcela A. Munizaga, *Modelación flexible de elecciones discretas: Una revisión crítica*. <http://tamarugo/cec.uchile.cl/~dicidet/>

Desagregación y Causalidad en una Metodología para la Predicción Económica

Antoni Espasa¹

Rebeca Albacete²

Universidad Carlos III de Madrid

1. Introducción

En el Instituto Flores de Lemus de la Universidad Carlos III de Madrid se viene realizando una tarea de predicción macroeconómica desde hace once años y actualmente abarca a la Unión Europea, Estados Unidos, España y ciertas regiones de la economía española. Esta tarea consiste en el desarrollo e implementación de técnicas de predicción que proporcionen resultados precisos a la vez que estimen los factores determinantes de las variables de interés. Este trabajo teórico ha ido en paralelo con un trabajo aplicado que desde el inicio se ha plasmado en la publicación mensual, Boletín de Inflación y Análisis Macroeconómico, que actualmente se edita en castellano e inglés. Todo el esfuerzo investigador sobre el tema está orientado por el principio de que la predicción económica tiene sentido si sirve para la toma de decisiones de instituciones y agentes. Esto requiere que, concibiendo las variables económicas como variables estocásticas, la predicción se formule en un marco estadístico apropiado y se realice basada en modelos econométricos. De este modo es posible contrastar los resultados de un ejercicio mensual sistemático de predicción como el que se realiza desde el Instituto Flores de Lemus. Con estos contrastes se desarrolla una tarea de evaluación de las predicciones, que genera un proceso continuo de perfeccionamiento y ampliación de los conjuntos informativos utilizados y de los correspondientes modelos econométricos, que produce una mejora sistemática de los resultados de la predicción. La metodología desarrollada se basa en los siguientes puntos. (a) Estrecha conexión entre los conjuntos informativos relevantes para el problema de predicción contemplado y la construcción de modelos econométricos apropiados a los datos y adecuados para los fines pretendidos.

¹espasa@est-econ.uc3m.es

²albacete@est-econ.uc3m.es

(b) La desagregación de variables macroeconómicas como forma de aumentar el contenido informativo sobre las mismas. (c) La construcción de modelos econométricos teniendo en cuenta la relación de causalidad entre las variables de acuerdo con la teoría económica. (d) Todo lo anterior implica que la orientación econométrica utilizada sea la de una modelización amplia que acabe combinando los resultados de diversos modelos construidos con conjuntos informativos diferentes, distinta frecuencia temporal y diversas formulaciones dinámicas y funcionales. A continuación se comentan estos aspectos fundamentales de la metodología de predicción mencionada. La construcción de modelos econométricos está vinculada al conjunto informativo utilizado en cada caso.

2. Ampliación del conjunto de información y desagregación de datos

El conjunto informativo mínimo que se puede considerar es el que incluye exclusivamente el presente y pasado de la variable de interés, al que denominaremos conjunto univariante básico. A partir de él es posible construir modelos univariantes ARIMA, propuestos por Box & Jenkins (1970), en los que el valor presente de la variable de interés viene explicado por su relación con sus valores pasados. El conjunto informativo univariante básico se puede ampliar (véase Albacete 2004) en diferentes direcciones que no son excluyentes entre sí y se pueden clasificar de la siguiente forma: (a) ampliación frecuencial, incorporando datos más frecuentes en el tiempo; (b) ampliación por desagregación funcional y geográfica de una variable agregada; (c) ampliación con otras series con las que se detecta una relación empírica de dependencia; y (d) ampliación con otras series temporales con las que se postula una relación teórica. En la predicción a corto y medio plazo es, en general, preferible la mayor desagregación frecuencial disponible, tal y como demuestran Albacete y Espasa (2005) para el caso de la inflación en la zona del euro. El análisis de una variable agregada, como la inflación, plantea la cuestión de si se obtienen mejores predicciones modelizando directamente el agregado o, por el contrario, desagregando y obteniendo las predicciones para el agregado a través de los componentes. En la literatura se han efectuado análisis desagregados sobre variables macroeconómicas y empresariales basados en criterios alternativos, sectoriales o geográficos, con resultados, en general, favorables a la desagregación.

Espasa et al. (2002) proponen una desagregación del índice de Precios al Consumo Armonizado, IPCA, total de la UME en cinco sectores, alimentos elaborados, bienes industriales no energéticos, servicios, alimentos no elaborados y energía. Por otro lado, Espasa y Albacete (2004a) desagregan el IPCA total de la UME en las siguientes cinco áreas geográficas, Alemania, Francia, Italia, España

y una quinta zona que agrega a todos los demás países que componen la zona euro. Estos trabajos muestran que la inflación en la UME no está plenamente cointegrada ni por sectores ni por países, existiendo tanto relaciones de cointegración entre los componentes así como pluralidad de factores tendenciales comunes. Por tanto, la desagregación es un modo de incrementar la información sobre los diferentes tendencias que afectan a los precios, que se puede explotar econométricamente siempre y cuando se disponga de datos adecuados a nivel desagregado - como es el caso de los precios al consumo europeos - y sea posible obtener modelos razonablemente aceptables para los componentes. En tales circunstancias, un modelo vectorial desagregado es el marco conveniente para considerar las restricciones de largo plazo entre los diferentes subíndices de precios, en línea con los trabajos de Clements & Hendry (1999).

Otro aspecto importante, tal y como muestran Espasa y Albacete (2004a), es que los IPCA's por sectores y países son variables agregadas que sufren frecuentes efectos especiales puntuales en algunos de sus componentes, que hacen referencia a: cambios metodológicos, como la introducción de los precios rebajados en el cálculo del IPCA o la entrada del euro; variaciones de tipo impositivo en el caso de los precios administrados, como los precios del tabaco o del gas y la electricidad; crisis del petróleo en el caso de los precios de combustibles y carburantes; o adversas condiciones climatológicas y epidemias que afectan a los precios de los alimentos no elaborados. Ante esta situación, la orientación propuesta en Albacete (2004) consiste en estimar los efectos agregados a partir de los efectos identificados y estimados en los subíndices afectados en cada caso. El análisis desagregado propuesto en este trabajo se basa en la consideración de la modelización simultánea de los componentes, como forma de captar tanto interrelaciones a largo plazo como la dependencia temporal entre sus variaciones estacionarias. Estos modelos simultáneos incorporan además un tratamiento adecuado de los efectos especiales que han influido en las observaciones de los índices de precios y permiten la incorporación de indicadores individuales en las ecuaciones de componentes o indicadores generales que puedan tener efectos diferenciados por componentes. Documentos de trabajo muy recientes de otras instituciones estudian también si prediciendo el IPCA a través de sus componentes se obtienen predicciones más ajustadas que a partir de un modelo agregado. Los resultados que obtienen, descritos en la tabla 1 son que, en general, tal posible ventaja en el enfoque desagregado no existe o es pequeña y limitada a horizontes de predicción cortos. Todos esos trabajos ni tienen en cuenta la modelización simultánea de los componentes ni incorporan un tratamiento adecuado de los efectos especiales. Ambas características propuestas en Espasa y Albacete (2004), Albacete (2004) y Albacete y Espasa (2005) se muestran como aportaciones de gran interés si se quieren obtener predicciones más precisas con la desagregación.

Tabla 1: Diferentes estudios sobre la predicción de la tasa de inflación anual total de la UME					
	Tesis (2004) 2000(1)-2003 (7)(a)	Hubrich (2003) 1998(2)-2001 (12)	den Reijer & Vlaar (2003) 1998(1)-2002 (12)	Benalal et al (2004) 1998(1)-2002 (6)	Hendry & Hubrich (2004) 1998(2)-2001(12) (12)
Reducción ó ampliación del RECM a 12 períodos con la predicción desagregada	Reducción(b)	Ampliación	Ampliación	Ampliación	Ampliación
(a) Período de evaluación de las predicciones.					
(b) La diferencia en el RECM es significativamente distinta de cero según el estadístico de Diebold & Mariano.					

Comparando el desempeño predictivo de diferentes modelos mensuales, Espasa y Albacete (2004a) llegan a la conclusión de que la mejor estrategia a nivel mensual para predecir la inflación en la zona euro consiste en desagregar el IPCA total en diez componentes, dos sectores - subyacente y residual - en las cinco áreas geográficas mencionadas anteriormente, construir un modelo VEqCM incorporando la restricción de diagonalidad por bloques entre las ecuaciones correspondientes a los índices de precios subyacentes y residuales e incorporar los precios del crudo tipo brent en euros como indicador adelantado, pero únicamente para los horizontes 1 y 2.

3. Modelos congruentes para predicción

Las predicciones mensuales proporcionan un buen ajuste, al incorporar la información más reciente sobre precios y una desagregación funcional y geográfica importante, pero no ofrecen una explicación de los factores determinantes de la predicción. En este sentido, resulta importante avanzar en el conjunto informativo utilizado y considerar variables explicativas que muestran una relación de causalidad fundamentada en la teoría económica a través de la elaboración de modelos econométricos congruentes, es decir, derivados a partir de la teoría económica y acordes con los datos. Estos modelos se formulan a nivel trimestral dado que factores determinantes de la inflación como los costes laborales unitarios sólo se observan con una periodicidad trimestral.

Estos modelos congruentes, siguiendo a Hendry (2001), incorporan como factores determinantes de la inflación en la zona euro los desequilibrios en diferentes mercados, bienes y servicios, laboral, monetario e internacional, incluyendo de este modo las teorías más relevantes en el análisis de la inflación. De estos modelos se deriva un análisis de la inflación en función de sus factores determinantes, entre los que se distinguen cuatro clases: (1) los factores que componen la dinámica

transitoria, entre los que se encuentran la inflación retardada, las variaciones de los costes laborales unitarios, de la masa monetaria, del PIB y del exceso de demanda y los cambios en los precios de importación y en los precios del crudo; (2) los desequilibrios a largo plazo, constituidos por relaciones de cointegración entre los precios agregados y las otras variables económicas sugeridas por la teoría económica; (3) factores que incorporan el efecto de variables artificiales que capturan la estacionalidad determinista y la incorporación de los precios rebajados en la construcción del IPCA a partir del año 2000; y (4) finalmente un factor residual. A partir de esta clasificación de factores se pueden calcular los efectos de éstos en la inflación en cada momento, muestral o futuro, y se puede interpretar la política monetaria seguida o extraer pautas sobre su implementación futura.

El resultado final que se alcanza en Albacete y Espasa (2005) demuestra que si se combinan las predicciones derivadas del modelo mensual desagregado vectorial diagonal por bloques, debidamente trimestralizadas, con las predicciones procedentes del modelo agregado trimestral congruente vectorial que incluye restricciones de largo plazo, se obtiene el mejor ajuste para la predicción de la inflación, tal y como muestra la tabla 2.

Tabla 2: Raíz del error de predicción cuadrático medio para la tasa de inflación anual en la UME. Período de predicción 2000(I)-2003(II).						
Horizonte	Modelo vectorial agregado trimestral	Modelos vectoriales desagregados diagonales por bloques mensuales			Combinación de Predicciones	
		Tres meses desconocidos	Primer mes conocido	Dos primeros meses conocidos	Tres meses desconocidos	Dos primeros meses conocidos
1p	0.12	0.12	0.06	0.02	0.09	0.06
2p	0.18	0.18	0.17	0.13	0.14	0.11
3p	0.21	0.21	0.20	0.19	0.17	0.15
4p	0.28	0.24	0.21	0.18	0.22	0.20

En negrita aparece el menor RECM en cada caso.

Estas predicciones combinadas son las que se proponen para analizar la inflación en el área del euro. A partir de ellas habrá que ajustar las contribuciones de las variables económicas a la inflación para dar una explicación precisa de los factores que determinan las predicciones finales. Así mismo, habrá que ajustar las predicciones de los diferentes subíndices de precios por países y sectores para ofrecer un mapa desglosado sectorial y geográficamente de los valores futuros estimados para la inflación en la UME. Esta conclusión final recoge las aportaciones metodológicas y teóricas relevantes aparecidas en la literatura para estudiar la inflación y se desarrolla en una orientación cuantitativa moderna como lo es la modelización gruesa propuesta por Granger y Jeon (2004). De todos los resultados anteriores se deriva una propuesta metodológica, descrita en Espasa y Albacete (2004b), para la predicción de indicadores o variables macroeconómicas. Los puntos básicos de

dicha metodología se centran en trabajar con el máximo nivel frecuencial con el que se observan la variable de interés y aquellas otras que por estar teóricamente relacionadas con ella se incluyen en el análisis; utilizar una versión desagregada que realmente aumente el conocimiento de los factores tendenciales de la variable de interés; construir modelos simultáneos como forma de recoger la dependencia dinámica a corto y largo plazo entre los componentes de la variable de interés y entre ésta y las demás variables económicas que intervienen en la formulación del modelo; incluir en la medida de lo posible indicadores realmente adelantados que incorporen al modelo lo más novedoso que se ha observado en el sistema. Con frecuencia, no es posible obtener un único modelo con todas estas características, por lo que el analista debe construir varios pocos modelos, de modo que cada uno capta facetas relevantes sobre la variable en cuestión: máxima frecuencia temporal en los datos, aspectos tendenciales en sus componentes, relaciones con otras variables económicas, los efectos de la información más novedosa presente en indicadores adelantados, etc. En tales casos habrá que predecir con todos ellos y producir como predicción final una combinación de todas ellas.

4. Conclusiones

Se puede concluir considerando que la metodología descrita se basa en el principio de aumento progresivo del conjunto informativo relevante con un tratamiento estadístico adecuado en cada caso, que viene determinado por la mejora en la predicción que dicha orientación progresiva conlleva.

Referencias

- Albacete, R. (2004). Modelización de la inflación a nivel europeo con fines de predicción y diagnóstico a corto plazo. *Tesis doctoral*. Departamento de Estadística. Universidad Carlos III de Madrid.
- Albacete, R. y Espasa, A. (2005). Forecasting inflation in the euro area using monthly time series models and quarterly econometric models. Documento de trabajo. Universidad Carlos III de Madrid.
- Clements, M.P. y Hendry, D.F. (1999). *Forecasting non-stationary economic time series*. Cambridge, Mass: MIT Press.
- Espasa, A., Senra, E. y Albacete, R. (2002). Forecasting Inflation in the European Monetary Union:

a disaggregated approach by countries and by sectors. *The European Journal of Finance*, **8**, 402-421.

Espasa, A. y Albacete, R. (2004a). Econometric Modelling for Short-Term Inflation Forecasting in the EMU. Working Paper 03-43. *Statistics and Econometric Series 09*. July 2004. Dpto. Estadística. Universidad Carlos III de Madrid.

Espasa, A. y Albacete, R. (2004b). Consideraciones sobre la predicción económica: metodología desarrollada en el Boletín de Inflación y Análisis Macroeconómico. Publicado en el libro *Estudios en Homenaje a Luis ángel Rojo*, Vol. I, Políticas, Mercados e Instituciones Económicas, editado por José Pérez, Carlos Sebastián y Pedro Tedde. Editorial Complutense, S.A. Diciembre 2004.

Granger, C.W.J y Jeon, Y. (2004). Thick modeling. *Economic Modelling*, **21**, 323-343.

Hendry, D. F. (2001). Modelling UK Inflation, 1875-1991. *Journal of Applied Econometrics*, **16**, 255-275.

Muestreo por Seguimiento de Nominaciones con Muestra Inicial de Sitios Seleccionada Secuencialmente¹

Martín H. Félix Medina²

Pedro E. Monjardin³

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 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. Félix Medina y Thompson (2003) desarrollaron una variante del Muestreo por Seguimiento de Nominaciones (MSN) en la cual el supuesto de una muestra inicial Bernoulli se substituye por una muestra aleatoria simple de sitios, la cual se selecciona de un marco muestral que cubre sólo una parte de la población de interés. Sin embargo, esta variante no permite al investigador controlar el número de personas nominadas ni controlar el tamaño de la muestra final. En este trabajo modificamos la variante propuesta por Félix-Medina y Thompson (2004) mediante el uso de una muestra inicial de sitios seleccionados secuencialmente en lugar de una muestra inicial seleccionada mediante muestreo aleatorio simple (MAS). Nuestra variante permite al investigador tener cierto control del tamaño de la muestra final, del número de personas nominadas, o de la precisión de los estimadores.

¹Trabajo realizado con apoyos parciales de los proyectos UASIN-EXB-01-01 y PIFI-2003-25-28 de la SEP y del proyecto PAFI-UAS-2002-I-MHFM-0 dela UAS

²mhfelix@uas.uasnet.mx

³pemo@uas.uasnet.mx

2. Diseño muestral

Al igual que en Félix-Medina y Thompson (2004) supondremos que una población U de un número desconocido τ de personas de difícil detección está dividida en dos partes U_1 y U_2 de tamaños desconocidos τ_1 y $\tau_2 = \tau - \tau_1$. Asimismo, supondremos que U_1 está cubierta por un marco muestral de N sitios A_1, \dots, A_N , tales como parques, hospitales o cruceros de calles. De este marco muestral se selecciona un sitio mediante MAS sin reemplazo. Se identifican los miembros de la población que pertenecen al sitio seleccionado y 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 lo nomina. Enseguida, se verifica si se satisface o no una regla de terminación del muestreo previamente definida. En el caso afirmativo el procedimiento muestral termina, en caso contrario el procedimiento anterior se repite hasta que se satisfaga la regla de terminación del muestreo. Nótese que al final del proceso muestral, tendremos una muestra inicial ordenada $S_0 = (A_1, \dots, A_n)$ de los n sitios que fueron seleccionados secuencialmente. Denominaremos al diseño que da lugar a S_0 muestreo aleatorio simple secuencial (MASS).

3. Estimadores e intervalos de confianza

En el caso de la variante del MSN con muestra inicial seleccionada por MAS, Félix-Medina y Thompson (2004) proponen estimadores máximo verosímiles (EMV) de los tamaños poblacionales. Posteriormente, Félix-Medina y Monjardin (2004) proponen estimadores que derivan bajo el enfoque Bayesiano. Los supuestos bajo los cuales obtienen los EMV son los siguientes. Se supone que el número m_i de miembros de la población que pertenecen al sitio A_i es la realización de una variable aleatoria Poisson con media λ_1 . Las m_i 's, $i = 1, \dots, N$, se suponen independientes. Asimismo, se supone que la variable $X_{ij}^{(k)}$, que vale 1 si la persona $u_j \in U_k - A_i$, $k = 1, 2$, es nominada por el sitio $A_i \in S_0$, y 0 en otro caso, es una variable Bernoulli con media $p_i^{(k)}$, $i = 1, \dots, n$; $k = 1, 2$. Para la obtención de los estimadores derivados bajo el enfoque Bayesiano, se supone además que τ_1 y τ_2 tienen distribución Poisson con medias $N\lambda_1$ y λ_2 , y que λ_1 y λ_2 son variables con distribuciones Gamma con parámetros conocidos. Finalmente, se supone que el logit $\alpha_i^{(k)} = \ln[p_i^{(k)} / (1 - p_i^{(k)})]$ de $p_i^{(k)}$ tiene distribución normal con media θ_k y varianza conocida σ_k^2 , y que θ_k es una variable normal con media y varianzas conocidas μ_k y γ_k^2 . Como en los casos anteriores, se supone que los $\alpha_i^{(k)}$'s son independientes. Puesto que nuestra variante de MSN difiere de la

propuesta por Félix-Medina y Thompson (2004) únicamente en el hecho de que la muestra inicial se selecciona secuencialmente en lugar de mediante MAS, por el principio de la regla de terminación del muestreo, tanto los EMV como los derivados bajo el enfoque Bayesiano se pueden usar con nuestro diseño. Por restricciones de espacio no presentaremos las expresiones anaíticas de los estimadores de τ_1 , τ_2 and τ . En este trabajo obtenemos intervalos de confianza para los tamaños poblacionales mediante el procedimiento Bootstrap usado por Félix-Medina y Monjardin (2004). Sin embargo, en lugar de usar el método percentil, usamos el método básico (Ver Davison y Hinkley, 1977, p. 194). Cabe aclarar que aún en el caso de los estimadores derivados bajo el enfoque Bayesiano, al igual que en Félix-Medina y Monjardin (2004), las inferencias las hacemos bajo el enfoque frecuentista.

4. Estudio Monte Carlo

Se generaron dos poblaciones de $N = 250$ valores de m_i 's a partir de la distribuciones Poisson con media 7.2 y binomial negativa con media 7.2 y varianza 2.4. En la primera población $\tau_1 = \sum_1^N m_i = 1828$, mientras que en la segunda $\tau_1 = 1861$. En ambos casos τ_2 se fijó en 700. Las probabilidades de nominación se generaron mediante el modelo $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, donde los valores de β_k fueron tales que se tuvieron dos casos. Caso 1: $(\mathbf{E}(p_i^{(1)}), \mathbf{E}(p_i^{(2)})) \approx (0.05, 0.03)$ y Caso 2: $(\mathbf{E}(p_i^{(1)}), \mathbf{E}(p_i^{(2)})) \approx (0.01, 0.006)$. Los valores de los parámetros de las distribuciones iniciales se fijaron como en Félix-Medina y Monjardin (2004). Se consideraron las variantes del MSN con muestra inicial secuencial y con muestra inicial aleatoria simple. En la primer variante la selección de sitios terminó cuando el número R_2 de nominados en U_2 alcanzó 250 o por primera vez excedió ese número. En el caso de la otra variante, el tamaño n de la muestra inicial se fijó igual al valor promedio, basado en 2000 replicaciones, de la muestra inicial secuencial.

Tabla 1. Sesgos relativos y raíces cuadradas de errores cuadráticos medios relativos de estimadores de los tamaños poblacionales de la población con m_i 's con distribución Poisson. Resultados basados en 2000 iteraciones.

Muestra inicial secuencial		$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$
$E(p_i^{(1)}) \approx 0.05$	$E(n) = 14.56$.001	.051	.015	.001	.044	.013
$E(p_i^{(2)}) \approx 0.03$	$E(R_2) = 263.3$.035	.129	.044	.034	.114	.040
$E(p_i^{(1)}) \approx 0.01$	$E(n) = 68.96$	-.000	.088	.024	-.001	.074	.020
$E(p_i^{(2)}) \approx 0.006$	$E(R_2) = 257.2$.022	.153	.045	.022	.131	.039
Muestra inicial aleatoria simple							
$E(p_i^{(1)}) \approx 0.05$	$E(n) = 15$	-.001	.010	.002	-.001	.007	.001
$E(p_i^{(2)}) \approx 0.03$	$E(R_2) = 259.7$.033	.115	.040	.033	.103	.038
$E(p_i^{(1)}) \approx 0.01$	$E(n) = 69$	-.000	.016	.004	-.001	.009	.002
$E(p_i^{(2)}) \approx 0.006$	$E(R_2) = 241.2$.023	.131	.039	.023	.114	.035

El primer renglón de cada celda contiene los sesgos relativos de los estimadores; el segundo renglón contiene las raíces cuadradas de los errores cuadráticos medios relativos de los estimadores. $\tilde{\tau}_k$, estimador máximo verosímil; $\hat{\tau}_k$, estimador obtenido bajo el enfoque Bayesiano. Valores esperados de n y R_2 obtenidos mediante simulación.

Por restricciones de espacio en las Tablas 1 y 2 sólo se presentan los resultados para la primera población, pero comentaremos los resultados para ambas poblaciones. Con respecto al desempeño de los estimadores de los tamaños poblacionales, se observó que con nuestra variante del MSN, los errores cuadráticos medios relativos de los estimadores fueron relativamente mayores que los obtenidos con la variante del MSN con muestra inicial seleccionada mediante MAS. Es decir, nuestra variante del MSN produjo una ligera pérdida en la eficiencia de los estimadores con respecto a la segunda variante. Se observó también que los estimadores derivados bajo el enfoque Bayesiano fueron ligeramente más eficientes que los EMV. Finalmente, cada uno de los estimadores mostró ser robusto a la desviación de la distribución Poisson de los m_i 's. En cuanto al desempeño de los intervalos de confianza, los intervalos que se obtuvieron con nuestra variante del MSN fueron más cortos que los obtenidos con la variante con muestra inicial seleccionada mediante MAS. Además en nuestra variante, las probabilidades de cobertura fueron cercanas al valor nominal, 0.95, mientras que en la otra variante, los intervalos que se obtuvieron a partir del estimador $\hat{\tau}_1$, obtenido bajo el enfoque Bayesiano, no fueron tan cercanas a 0.95. Intervalos basados en estimadores obtenidos bajo el enfoque Bayesiano tuvieron mejores desempeños que los basados en EMV. La desviación del supuesto de la distribución Poisson de los m_i 's incrementó las longitudes de los intervalos para τ_1 and τ , y alejó ligeramente sus probabilidades de cobertura del valor nominal. Las longitudes y probabilidades de cobertura de los intervalos para τ_2 no se afectaron por la desviación del supuesto de la distribución Poisson. Los resultados obtenidos son alentadores e indican que nuestra variante es una alternativa a tomarse en cuenta para el muestreo de poblaciones de difícil detección.

Tabla 2. Longitudes promedio y probabilidades de cobertura de intervalos de confianza bootstrap del 95 % para los tamaños poblacionales de la población con m_i 's con distribución Poisson . Resultados basados en 2000 iteraciones.

Muestra inicial secuencial		$\bar{\tau}_1$	$\bar{\tau}_2$	$\bar{\tau}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$
$E(p_i^{(1)}) \approx 0.05$	$E(n) = 14.56$	254.9	360.8	446.1	254.9	304.8	398.1
$E(p_i^{(2)}) \approx 0.03$	$E(R_2) = 263.3$.947	.946	.961	.945	.932	.946
$E(p_i^{(1)}) \approx .01$	$E(n) = 68.96$	178.4	402.2	442.5	177.2	322.1	368.3
$E(p_i^{(2)}) \approx 0.006$	$E(R_2) = 257.2$.949	.963	.975	.948	.948	.965
Muestra inicial aleatoria simple							
$E(p_i^{(1)}) \approx 0.05$	$E(n) = 15$	218.2	336.7	414.5	218.2	290.7	364.9
$E(p_i^{(2)}) \approx 0.03$	$E(R_2) = 259.7$.942	.930	.959	.896	.913	.924
$E(p_i^{(1)}) \approx 0.01$	$E(n) = 69$	162.4	359.5	396.7	135.6	306.0	335.6
$E(p_i^{(2)}) \approx 0.006$	$E(R_2) = 241.2$.925	.929	.949	.831	.905	.914

El primer renglón de cada celda contiene las longitudes promedios; el segundo renglón contiene las probabilidades de cobertura. Muestras bootstrap de tamaños igual a 1000.

Referencias

Booth, J. G., Butler, R. W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, **89**, 1282-1289.

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. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population: a Bayesian assisted approach. En revisión en *Survey Methodology*.

Estimación en el Modelo de Regresión Logística en Presencia de Datos Separados y Colinealidad

Flaviano Godínez Jaimes¹

Unidad Académica de Matemáticas de la Universidad Autónoma de Guerrero

Gustavo Ramírez Valverde²

Especialidad de Estadística. ISEI. Colegio de Postgraduados

1. Introducción

Consideremos datos en regresión binaria $\{(Y_i, x_i^T) : i = 1, \dots, n\}$ donde las Y_i son variables aleatorias independientes con distribución Bernoulli con probabilidad de éxito desconocida $\pi_i = P(Y_i = 1)$ y $x_i^T = (x_{i1}, \dots, x_{ip})$ son vectores no estocásticos de dimensión p ($p < n$). Sea \mathbf{X} la matriz diseño de $n \times (p + 1)$ cuyos renglones son $(1 \ x_i^T)$ y $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. Los datos siguen el modelo de regresión logística si $\pi_i = P(Y_i = 1) = e^{x_i^T \beta} / (1 + e^{x_i^T \beta})$ donde $\beta = (\beta_0 \ \beta_1 \ \dots \ \beta_p)^T$ es el vector de parámetros desconocidos. La existencia del estimador de máxima verosimilitud (EMV), $\hat{\beta}$, del modelo de regresión logística depende de la configuración de los datos. Hay separación en los datos si existe un $\theta \in \mathbb{R}^{p+1}$ tal que $x_i^T \theta > 0$ cuando $Y_i=1$ y $x_i^T \theta < 0$ cuando $Y_i=0$, para $i = 1, \dots, n$. Hay quasi separación en los datos si existe un $\theta \in \mathbb{R}^{p+1} \setminus \{0\}$ tal que $x_i^T \theta \geq 0$ cuando $Y_i=1$ y $x_i^T \theta \leq 0$ cuando $Y_i=0$, para todo i , y además existe $j \in \{1, \dots, n\}$ tal que $x_j^T \theta = 0$ y hay traslape si no existe separación ni quasi separación en los datos. Recientemente Christmann y Rousseeuw (2001) propusieron el procedimiento NCOMPLETE que determina de manera exacta el número mínimo de observaciones que hay que eliminar para que en el resto haya separación en los datos. El EMV del modelo de regresión logística no existe cuando hay separación o quasi separación en los datos y éste existe y es único cuando hay traslape en los datos (Albert y Anderson, 1984).

Lesaffre y Marx (1993) señalaron que la matriz de información, $X^T \hat{V} X$, usada para estimar el EMV del modelo de regresión logística puede ser singular si: a) X es de rango incompleto, b) el vector de parámetros estimado se acerca a la frontera del espacio de parámetros, y c) ambas condiciones se

¹fgodinezj@colpos.mx

²gramirez@colpos.colpos.mx

presentan. Aun cuando \mathbf{X} sea de rango completo pueden existir dependencias lineales cercanas, esto es, $c_1X_1 + \dots + c_pX_p \approx 0$. Entre más cerca se este a una dependencia lineal exacta, \mathbf{X} se acerca mas a la singularidad. Este fenómeno es conocido como colinealidad entre las variables explicatorias o x-colinealidad. Cuando hay separación o quasi separación en los datos entonces al menos una componente del vector de parámetros tiende a $\pm\infty$ que es la frontera del espacio de parámetros. La matriz de informacion se acerca a la singularidad por el efecto combinado de x-colinealidad y la cercanía a la separación en los datos, este problema es conocido como mv-colinealidad. Tanto la x-colinealidad como la mv-colinealidad son problemas de grado pues siempre están presentes en menor o mayor grado.

En presencia de separación en los datos dos estimadores han sido propuestos: el estimador de Firth y el estimador de Rousseeuw y Christmann, mientras que los estimadores ridge han sido usados cuando hay x-colinealidad en los datos.

2. Estimadores Estudiados

Firth (1993) propuso un estimador con el fin de reducir el sesgo cuando se usan muestras pequeñas en el modelo lineal generalizado. Heinze y Schemper (2002) encontraron que el estimador de Firth también existe cuando hay separación en los datos. El estimador de Firth puede considerarse como un estimador penalizado donde la función penalty es la *apriori Invariante de Jeffreys* para este problema. La función de log-verosimilitud modificada de este estimador es $\log L(\beta)^* = \log L(\beta) + 1/2 \log |I(\beta)|$ donde $L(\beta)$ es la verosimilitud del modelo de regresión logística. Las ecuaciones de score son:

$$U(\beta_r)^* = \sum_{i=1}^n \{y_i - \pi_i + h_i(1/2 - \pi_i)\}x_{ir} = 0 \quad (r = 0, \dots, p)$$

donde las h_i 's son el *i-ésimo* elemento

en la diagonal de la matriz 'hat', $H = V^{1/2}X(X'VX)^{-1}X'V^{1/2}$, con $V = \text{diag}\{\pi_i(1 - \pi_i)\}$. El estimador de Firth, $\hat{\beta}_F$, que denotaremos por F, puede ser obtenido iterativamente usando el método de Newton-Raphson $\beta_F^{(s+1)} = \beta_F^{(s)} + I^{-1}(\beta_F^{(s)})U(\beta_F^{(s)})^*$. Este estimador no existe en el caso extremo de separación en los datos en que todas las respuestas son del mismo tipo.

Rousseeuw y Christmann (2003) propusieron el modelo de regresión logística escondido. En este modelo se supone que el verdadero status T , con valores éxito (s) y falla (f), no se puede observar debido a un mecanismo estocástico adicional. Sin embargo, hay una variable binaria observada Y fuertemente relacionada con T en la siguiente forma: si el verdadero status es $T = s$ observamos $Y = 1$ con $P(Y = 1|T = s) = \delta_1$ y por tanto hay una clasificación incorrecta con prob-

abilidad $P(Y = 0 | T = s) = 1 - \delta_1$. Análogamente, si el verdadero status es $T = f$ observamos $Y = 0$ con probabilidad $P(Y = 0 | T = f) = 1 - \delta_0$ y la clasificación incorrecta con probabilidad $P(Y = 1 | T = f) = \delta_0$. Suponiendo que la probabilidad de observar el verdadero status es mayor al 50 %, entonces, $0 < \delta_0 < 0.5 < \delta_1 < 1$. El estimador $\hat{\beta}_{RC}$, denotado por RC, de el modelo de regresión logística escondido se obtiene ajustando por máxima verosimilitud el modelo de regresión logística a las pseudo-observaciones, $\tilde{y}_i = (1 - y_i) \delta_0 + y_i \delta_1$. Suponiendo que $\delta_0 \neq 1 - \delta_1$, $\hat{\beta}_{RC}$ se estima con el siguiente algoritmo

1. Calcular $\hat{\pi} = \max(\delta, \min(1 - \delta, \bar{\pi}))$, donde $\bar{\pi} = \frac{1}{n} \sum_{i=1}^n y_i$ y $\delta = 0.01$,
2. Calcular $\delta_0 = \frac{\delta \hat{\pi}}{1 - \delta}$ y $\delta_1 = \frac{1 + \delta \hat{\pi}}{1 - \delta}$,
3. Calcular las pseudo-observaciones $\tilde{y}_i = (1 - y_i) \delta_0 + y_i \delta_1$,
4. Ajustar el modelo de regresión logística a las pseudo-observaciones.

Los estimadores ridge en regresión logística fueron propuestos para reducir el tamaño de $\hat{\beta}$ ocasionado por la existencia de x-colinealidad. Schaefer et al (1984) mostraron que siempre es posible encontrar un valor para el parámetro ridge, k , de manera que el error cuadrático medio del estimador ridge logístico, $\hat{\beta}_R$, sea menor que el de $\hat{\beta}$.

El estimador ridge logístico (ERL) esta dado por $\hat{\beta}_R(k) = [X' \hat{V} X + kI]^{-1} X' \hat{V} X \hat{\beta}$ donde $\hat{\beta}$ es el EMV del modelo de regresión logística. Varias formas de seleccionar k han sido propuestas ($1/\beta' \beta$, $(p+1)/\beta' \beta$, traza $(X^T V X)/\beta' X^T V X \beta$, $1 / \max_j (v'_j \beta)^2$ donde v'_j es un vector propio de $X^T \hat{V} X$) pero todas dependen de β . El ERL hereda los problemas del EMV, esto es, no existe cuando hay separación en los datos y tiene pobre desempeño cuando hay problemas de x-colinealidad.

Estimadores Alternativos

El estimador RC se calcula usando en todos los casos la constante $\delta = 0.01$. Denotaremos por RCA al estimador $\hat{\beta}_{RCA}$ que se obtiene cuando δ se selecciona como el valor que minimiza la media del cuadrado del error, $\frac{1}{n} \sum_i (Y_i - \hat{\pi}_i)^2$. Denotaremos por RCS, al estimador $\hat{\beta}_{RCS}$, que se obtiene al igualar las probabilidades de mala clasificación en el RCA, esto es, $\delta_0 = 1 - \delta_1 = \gamma$, donde γ

se selecciona como el valor que minimiza la media del cuadrado del error. En la simulación δ y γ toman los valores 0.0001, 0.0006, ..., 0.4996.

Liu (2003) para el modelo de regresión lineal propuso calcular k con $k_L = (\lambda_1 - 100\lambda_p)/99$ donde λ_1 y λ_p son el mayor y el menor de los valores propios de $X^T X$. Cuando hay separación en los datos $\hat{\beta}_F$ y $\hat{\beta}_{RC}$ pueden usarse para obtener estimadores Ridge logísticos. Los estimadores resultantes se denotan por $\hat{\beta}_{RLF}$, $\hat{\beta}_{RLRCA}$ y $\hat{\beta}_{RLRCS}$, se denominan por RLF, RLRCA, y RLRCS y se obtienen con $\hat{\beta}_{RLF}(k_L) = [X^* \hat{V}_F X^* + k_L I]^{-1} X^* \hat{V}_F X^* \hat{\beta}_F$

$$\hat{\beta}_{RLRCA}(k_L) = [X^* \hat{V}_{RCA} X^* + k_L I]^{-1} X^* \hat{V}_{RCA} X^* \hat{\beta}_{RCA}$$

$$\hat{\beta}_{RLRCS}(k_L) = [X^* \hat{V}_{RCS} X^* + k_L I]^{-1} X^* \hat{V}_{RCS} X^* \hat{\beta}_{RCS}$$

donde $X^* = [\sqrt{n} \mathbf{1} (X_{i1} - \bar{X}_1) / \sum (X_{i1} - \bar{X}_1)^2 \cdots (X_{ip} - \bar{X}_p) / \sum (X_{ip} - \bar{X}_p)^2]$, y

$$\hat{V}_F = \text{diag} \left\{ p_i(\hat{\beta}_F) (1 - p_i(\hat{\beta}_F)) \right\}, \quad \hat{V}_{RCA} = \text{diag} \left\{ p_i(\hat{\beta}_{RCA}) (1 - p_i(\hat{\beta}_{RCA})) \right\} \text{ y}$$

$$\hat{V}_{RCS} = \text{diag} \left\{ p_i(\hat{\beta}_{RCS}) (1 - p_i(\hat{\beta}_{RCS})) \right\}.$$

3. Simulación

Se hizo una simulación del proceso con dos variables explicatorias y los siguientes factores:

1. Se consideraron dos grados de correlación muestral (CM): alta ($r=0.95$) y severa($r=0.99$).
2. Se consideraron muestras (TM) de 20 y 40 observaciones.
3. Se consideraron dos orientaciones para β , las dadas por los vectores propios (VP) asociados al valor propio mayor y menor (VP1, VP3) de $X^T X$.
4. Se usaron tamaños del vector propio (TVP) de 1 y 9 unidades (TVP1 y TVP9), los cuales se obtienen multiplicando el vector propio correspondiente por 1 y 3.
5. El porcentaje de traslape (PT) se midió para los datos generados y se clasificó en cuatro categorías: PT0, PTI, PTII, PTIII y PTIV en las cuales los porcentajes de traslape están en los intervalos 0, (0, 10], (10, 20], (20, 30], (30, 40].

Las matrices diseño, $\mathbf{X} = [\mathbf{1} \ X_1 \ X_2]$, se generaron de orden 3x20 y 3x40. Con $X_1 \sim U[0, 1]$ y X_2 se obtuvo con $X_2 = X_1 + cu$, donde $u \sim U[0, 1]$, y c toma valores que permitieron obtener CM aproximadas de 0.95 y 0.99. Y fue calculado usando $Y_i = 1$ si $\pi_i > w$ y $Y_i = 0$ en otro caso; con $w \sim U[0, 1]$.

Los estimadores fueron comparados considerando su error cuadrático medio:

$ECM = \frac{1}{R} \sum_{r=1}^R (\tilde{\beta}_r - \beta)^T (\tilde{\beta}_r - \beta)$, donde $\tilde{\beta}_r$ es uno de los estimadores de β en la r -ésima repetición.

Se investigó cada una de las combinaciones de CM x TM x VP x TVP. En cada escenario se hicieron 5000 repeticiones, mismas que se separaron de acuerdo al PT obtenido. Este fue calculado usando una versión en SAS del programa NCOMPLETE. El EMV del modelo de regresión logística será denotado por L en el resto del trabajo.

Resultados

Los estimadores L, F, RCA y RCS generalmente tienen menor ECM con el VP1, lo contrario ocurre solo en 6 de 68 casos, mientras que los estimadores RLF, RLRCA y RLRCS siempre tienen menor ECM con el VP1. El cociente de el ECM obtenido con VP3 con el de VP1 varía entre 0.81 y 4.57 para los estimadores L, F, RCA y RCS y entre 1.07 y 6.67 para los estimadores RLF, RLRCA y RLRCS.

Todos los estimadores tienen el mayor ECM cuando hay separación en los datos donde L no existe así como tampoco algunos casos de F. Mientras que el ECM de L, F, RCA Y RCS disminuye monótonamente al incrementar el PT, el ECM de RLF, RLRCA y RLRCS disminuye muy poco.

Hay un fuerte efecto de la CM en el ECM de L, F, RCA y RCS pues el ECM en C99 es entre 3.83 y 13.56 veces mayor que en C95. Para los estimadores RLF, RLRCA y RLRCS, el cociente del ECM obtenido con CM 99 con el de CM 95 varia entre 0.42 y 1.25, y en 67 de 130 casos el ECM con CM 99 es menor. Este tipo de estimadores Ridge son mejores cuando hay mayor CM (mayor x-colinealidad).

Con frecuencia, el ECM de todos los estimadores es menor con TVP9 y esto es más fuerte con TM 20 y/o VP3. El ECM de los estimadores L, F, RCA y RCS con el TVP9 representa entre 0.15 y 8.62 veces que el ECM con TVP1. De 82 casos solo en 28 es mayor el ECM con TVP9. Para los

estimadores RLF, RLRCA y RLRCS el ECM con el TVP9 es entre 0.12 y 10.35 que el ECM con TVP1. De 105 casos solo en 23 es mayor el ECM con TVP9.

Acorde con resultados asintóticos, el ECM de los estimadores es mejor con TM 40 que con TM 20. El ECM de L, F, RCA y RCS con TM 20 representa entre 0.43 y 12.10 veces que con TM 40 y solo en 3 de 90 veces es mayor el ECM con TM 40. El ECM de RLF, RLRCA y RLRCS con TM 20 es entre 0.86 y 3 veces que el obtenido con TM 40. En estos estimadores solo en 16 de 115 veces el ECM con TM 40 es mayor, pero la diferencia no es muy grande. Este tipo de estimadores Ridge son buenos en muestras pequeñas y no mejoran sustancialmente cuando se incrementa el TM.

Si el ECM se usa como criterio para comparar los estimadores, F tiene menor ECM que L, RCA y RCS, pero menor ECM se obtiene con cualquiera de los estimadores RLF, RLRCA y RLRCS (ver Figura 1).

4. Conclusiones

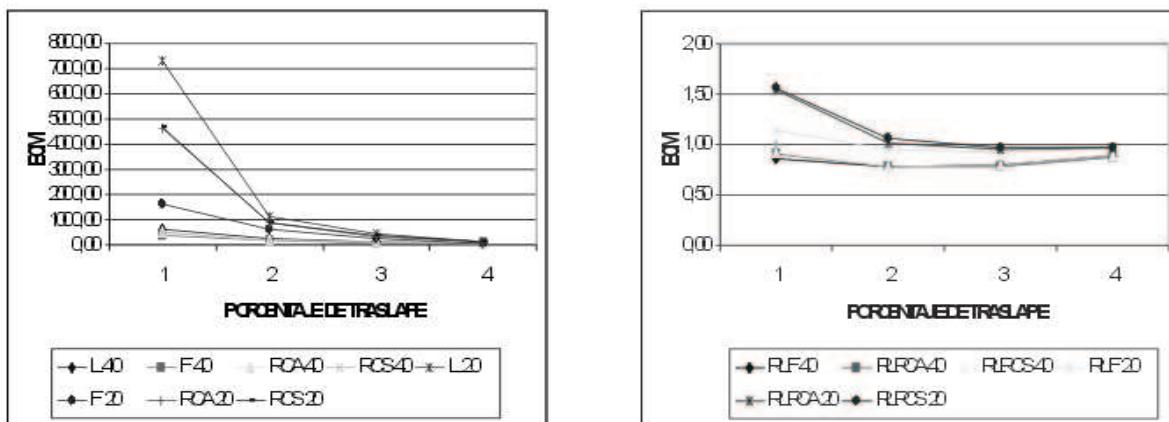


Figura 1: Efecto del Tamaño muestral en el error cuadrático medio de los estimadores a) Ridge-Liu y b) L, F, RCA y RCS, con CM 99, VP1 y TVP1.

El estimador de Firth y el de Rousseeuw y Christmann tienen la ventaja de que existen en casos en que hay separación en los datos pero son afectados negativamente cuando hay x-colinealidad. Los estimadores RLF, RLRCA y RLRCS propuestos en este trabajo son mejores, en términos de error cuadrático medio, en condiciones no ventajosas para el modelo.

Referencias

- Albert, A. and J. A. Anderson. 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 71:1-10.
- Christmann, A., and P. J. Rousseeuw. 2001. Measuring overlap in logistic regression. *Computational Statistics and Data Analysis*. 37:65-75.
- Firth, D., 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80, 27-38.
- Heinze, G., and M. Schemper. 2002. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21: 2409-2419.
- Lesaffre, E. and B. D. Marx. 1993. Collinearity in generalized linear regression. *Communications in Statistics-Theory and Methods*. 22(7):1933-1952.
- Liu, K., 2003. Using Liu-type estimator to combat collinearity. *Communications in Statistics-Theory and Methods*, Vol **32**, No 5, pp 1009-1020.
- Rousseeuw, P. J., and A. Christmann. 2003. Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis* 43, 315-332.
- Schaefer, R. L., L. D. Roi, and R. A. Wolfe, 1984. A ridge logistic estimator. *Communications in Statistics-Theory and Methods*. 13(1): 99-113.

Densidades conjuntas condicionadas a estadísticas suficientes y aplicaciones

José M. González-Barrios¹

Federico O'Reilly²

Raúl Rueda³

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,

Universidad Nacional Autónoma de México

1. Introducción

Sea X_1, X_2, \dots, X_n una muestra aleatoria cuya distribución pertenece a una familia paramétrica $\{F(x; \theta) \mid \theta \in \Theta\}$ donde $\Theta \subset \mathbb{R}^k$. Sea $T_n : \mathbb{R}^n \rightarrow \mathbb{R}^k$ una estadística suficiente para θ , entonces toda la información acerca de θ está incluida en T_n .

Si X_1, \dots, X_n son variables aleatorias discretas con valores en un conjunto numerable $M \subset \mathbb{R}$, cuya densidad pertenece a la familia $\{f(x; \theta) \mid \theta \in \Theta\}$ donde $\Theta \subset \mathbb{R}$, y T_n es una estadística suficiente para θ , definimos la **Densidad Condicional discreta**, mediante:

$$\hat{f}_n(x_1, x_2, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n \mid T_n = t),$$

donde $(x_1, \dots, x_n) \in M^n$. De acuerdo con la teoría de Suficiencia los valores de \hat{f}_n no dependen de θ , y en muchos ejemplos, ver Teorema 2.1, podemos encontrar expresiones exactas para estas densidades.

Propondremos como aplicaciones de la densidad condicional, pruebas de bondad de ajuste, pruebas de hipótesis y selección de modelos.

¹gonzaba@sigma.iimas.unam.mx

²federico@sigma.iimas.unam.mx

³pinky@sigma.iimas.unam.mx

2. Resultado principal

En esta sección usaremos las variables aleatorias discretas más comunes, recordemos que X tiene densidad geométrica con parámetro $0 < p < 1$, denotado por $X \sim G(p)$ si y sólo si

$$P(X = x) = (1 - p)^x p \quad \text{si } x = 0, 1, \dots$$

Decimos que X tiene densidad binomial con parámetros $k \geq 1$ y $0 < p < 1$, denotado por $X \sim \text{Bin}(k, p)$ si y sólo si

$$P(X = x) = \binom{k}{x} p^x (1 - p)^{k-x} \quad \text{si } x = 0, 1, \dots, k.$$

Decimos que X tiene densidad Poisson con parámetro $\lambda > 0$, denotado por $X \sim P(\lambda)$ si y sólo si

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{si } x = 0, 1, \dots$$

Decimos que X tiene densidad binomial negativa con parámetros $r \geq 1$ y $0 < p < 1$, denotado por $X \sim \text{NB}(r, p)$ (si $r = 1$, $X \sim G(p)$) si y sólo si

$$P(X = x) = \binom{r + x - 1}{x} p^r (1 - p)^x \quad \text{si } x = 0, 1, \dots$$

Observemos que las distribuciones de arriba están incluidas en una clase muy amplia de distribuciones llamada las distribuciones de series de potencias, ver Johnson *et al.* (1992), donde la función de densidad puede escribirse como

$$P(X = x) = \frac{a_x \theta^x}{\eta(\theta)} \quad \text{para } x = 0, 1, 2, \dots,$$

donde $\theta > 0$, $a_x \geq 0$ y $\eta(\theta) = \sum_{x=0}^{\infty} a_x \theta^x$. La función $\eta(\cdot)$ es llamada la *función serie*. Es bien conocido, ver Johnson *et al.* (1972), que la suma de n variables aleatorias mutuamente independientes teniendo cada una la misma distribución de series de potencias, tiene una distribución de la misma clase con función serie $[\eta(\theta)]^n$.

Ahora presentamos el resultado principal de este trabajo, el cual da una caracterización de las densidades condicionales de la familia de las distribuciones de series de potencias.

Teorema 2.1 Sea X_1, X_2, \dots, X_n una muestra aleatoria de la distribución de series de potencias

$$P(X = x) = \frac{a_x \theta^x}{\eta(\theta)} \quad \text{para } x = 0, 1, 2, \dots, \quad (1)$$

donde $\theta > 0$, $a_x \geq 0$ y $\eta(\theta) = \sum_{x=0}^{\infty} a_x \theta^x$. Sea $T_n = \sum_{i=1}^n X_i$. Entonces

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | T_n = t) = \frac{a_{x_1} a_{x_2} \cdots a_{x_n}}{\hat{a}_t},$$

donde \hat{a}_t corresponde a la densidad de T_n , y para cada $0 \leq x_i, i = 1, 2, \dots, n$ tal que $\sum_{i=1}^n x_i = t$. Recíprocamente, si X_1, \dots, X_n tiene la distribución condicional de arriba dado $T_n = \sum_{i=1}^n X_i = t$ para cualesquiera n y t enteros no negativos, entonces X_i tiene la distribución dada en la ecuación (1).

Observación 2.2 Primero notemos que las observaciones son condicionalmente intercambiables, es decir, dada cualquier permutación σ del conjunto $\{1, 2, \dots, n\}$

$$P(X_1 = x_1, \dots, X_n = x_n | T_n = t) = P(X_1 = x_{\sigma(1)}, \dots, X_n = x_{\sigma(n)} | T_n = t),$$

para cada $t \geq 0$ y cualquier vector (x_1, \dots, x_n) tal que $\sum_{i=1}^n x_i = t$.

Observación 2.3 Ya que T_n es una estadística suficiente para θ en las distribuciones de series de potencias, entonces las densidades conjuntas de (X_1, X_2, \dots, X_n) dado $T_n = t$ no dependen del parámetro para el cual T_n es suficiente.

Observación 2.4 En el caso de variables aleatorias independientes geométricas con parámetro p , se tiene que $a_{x_i} = 1$, para $i = 1, 2, \dots, n$, y \hat{a}_t es $\binom{t+n-1}{n-1}$. Entonces la densidad condicional del vector (X_1, \dots, X_n) dado $T_n = t$ es

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | T_n = t) = \frac{1}{\binom{t+n-1}{n-1}},$$

es decir, una distribución uniforme sobre todas las posibles formas de sumar n enteros no negativos y obtener el valor de t . Por cierto, se puede obtener el número de formas de hacer lo anterior, y el número es

$$\binom{t+n-1}{n-1}. \quad (2)$$

Para ver una prueba combinatoria de este hecho, ver por ejemplo Theorem 3.4, Bose and Manvel (1984). Llamaremos un *arreglo de* (t, n) a un vector $(x_n, x_{n-1}, \dots, x_2, x_1)$ si $x_n \geq x_{n-1} \geq \dots \geq x_2 \geq x_1$ con $\sum_{i=1}^n x_i = t$. En teoría combinatoria un arreglo es a veces llamado una partición entera. En vista de la observación 2.2, notemos que el número de arreglos de (t, n) , es considerablemente más pequeño que el número de formas en las que al sumar n enteros no negativos obtenemos t , (en el Cuadro 1 comparamos el número de arreglos de $(t, 10)$ y el número de formas en las que al sumar 10 enteros obtenemos el valor de t para $t = 0, 1, 2, \dots, 15$). La idea de usar arreglos facilita encontrar los valores de la densidad de (X_1, X_2, \dots, X_n) dado cualquier valor fijo t de T_n .

Cuadro 1: Número de formas en las cuales 10 enteros suman t y el número de arreglos

t	$\binom{t+10-1}{10-1}$	arrangements
0	1	1
1	10	1
2	55	2
3	220	3
4	715	5
5	2002	7
6	5005	11
7	11440	15
8	24310	22
9	48620	30
10	92378	42
11	167960	55
12	293930	75
13	497420	97
14	817190	128
15	1307504	164

Hasta donde sabemos, no existe una fórmula cerrada para encontrar el número de arreglos de (t, n) . Sin embargo si $t = n$, tenemos la fórmula recursiva de Euler, que dice que si p_n es el número de arreglos de (n, n) , entonces

$$p_n = \sum_{m=1}^{\infty} (-1)^{m+1} (p_{n-m(3m-1)/2} + p_{n-m(3m+1)/2}),$$

donde $p_0 = 1$ y $p_k = 0$ si $k < 0$. Esta fórmula recursiva puede usarse para evaluar p_n ya que solamente $O(\sqrt{n})$ términos son distintos de cero, ver Skiena (1990).

Observación 2.5 En el caso de variables aleatorias independientes binomiales con parámetros $k \geq 1$, que se supondrá conocida, y $0 < p < 1$, tenemos que $a_{x_i} = \binom{k}{x_i}$, para $i = 1, 2, \dots, n$, y $\hat{a}_t = \binom{nk}{t}$. Entonces la densidad de (X_1, X_2, \dots, X_n) dado que $T_n = t$ es

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | T_n = t) = \frac{\binom{k}{x_n} \binom{k}{x_{n-1}} \cdots \binom{k}{x_1}}{\binom{nk}{t}},$$

que corresponde a una versión “multivariada” de la distribución hipergeométrica, consistente en n clases cada una de tamaño k .

Observación 2.6 En el caso de la distribución Poisson con parámetro λ , $a_{x_i} = 1/x_i!$, para cada $i = 1, 2, \dots, n$, y $\hat{a}_t = 1/t!$. Entonces la densidad de (X_1, X_2, \dots, X_n) dado $T_n = t$ es

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | T_n = t) = \binom{t}{x_n} \binom{t - x_n}{x_{n-1}} \binom{t - x_n - x_{n-1}}{x_{n-2}} \cdots \binom{t - x_n - \cdots - x_2}{x_1} \left(\frac{1}{n}\right)^t = \frac{t!}{x_1! \cdots x_n!} \left(\frac{1}{n}\right)^t,$$

que corresponde a una distribución multinomial donde $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$, ver por ejemplo Johnson *et al.* (1997).

Observación 2.7 En el caso de binomiales negativas independientes con parámetros $r \geq 1$ que suponemos conocido, y $0 < p < 1$, $a_{x_i} = \binom{r + x_i - 1}{r - 1}$, para $i = 1, 2, \dots, n$, y $\hat{a}_t = \binom{nr + t - 1}{nr - 1}$. Entonces la densidad de (X_1, X_2, \dots, X_n) dado $T_n = t$ es

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | T_n = t) = \frac{\binom{r + x_n - 1}{r - 1} \binom{r + x_{n-1} - 1}{r - 1} \cdots \binom{r + x_1 - 1}{r - 1}}{\binom{nr + t - 1}{nr - 1}},$$

que es similar a una densidad hipergeométrica “multivariada” con n clases, pero en este caso el tamaño de cada clase varía de acuerdo a los valores de las x'_i s, y el número de observaciones para cada clase, $r - 1$, está fijo.

3. Aplicaciones

Utilizando la densidad condicional y el teorema de caracterización de las distribuciones, así como el concepto de arreglos, es posible dar una nueva prueba de bondad de ajuste para $H_0 : X$ tiene una distribución de series de potencias con función serie $\eta(\theta)$ vs $H_1 : X$ no tiene una distribución de series de potencias con función serie $\eta(\theta)$, basada en una región de máxima densidad condicional. Así mismo, es posible hacer una prueba de hipótesis del tipo de Neyman-Pearson para $H_0 : X$ tiene una distribución de series de potencias con función serie $\eta(\theta)$ vs $H_1 : X$ tiene una distribución de series de potencias con función serie $\zeta(\theta)$, donde $\zeta(\theta)$ es otra función serie fija. El estudio de estas pruebas se puede encontrar en González-Barrios *et al.* (2005).

Tambien se puede utilizar la densidad condicional para realizar selección de modelos que pertenezcan a la familia de distribuciones de series de potencias, ver Contreras y González-Barrios (2005).

Referencias

- Bose, R.C. and Manvel, B. (1984), *Introduction to combinatorial theory*. Ed. John Wiley & Sons, New York.
- Contreras, A. and González-Barrios, J.M. (2005), Model selection using the conditional density. (preprint).
- González-Barrios, J.M., O'Reilly, F. and Rueda R. (2005), Goodnees of fit for discrete random variables using the conditional density. (preprint).
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997), *Discrete multivariate distributions*. Ed. Wiley, New York.
- Johnson, N.L., Kotz, S. and Kemp, A.W. (1992), *Univariate discrete distributions*. Ed. Wiley, New York.
- Skiena, S.S. (1990), *Implementing Discret Mathematics, Combinatorics and Graph Theory with Mathematica*, Ed. Addison-Wesley Publ. Co., Redwood City, Cal.

Valuación De Opciones Asiáticas Mediante Monte Carlo Con Reducción De Varianza

Víctor Hugo Ibarra Mercado¹

Universidad Anáhuac y ESFM-I.P.N.

Patricia Saavedra Barrera²

Universidad Autónoma Metropolitana

1. Introducción

El mercado de energéticos tiene varias características que lo distinguen del mercado de dinero. Entre ellas está el que el subyacente es un bien cuya posesión misma da al consumidor una ventaja, como el gas, el petróleo y la electricidad. Otra es que el mercado es mucho más volátil y menos líquido que el mercado de dinero. Y por último el mercado es estacional. Estas características han obligado a que los modelos y el tipo de opciones que se utilizan, para cubrirse del riesgo de las fluctuaciones de precio del subyacente, sean distintas a las que se usan en el mercado de dinero, véase Pilopovic (1997). En particular, se prefiere utilizar opciones asiáticas, cuya función de pago depende del promedio aritmético de los valores que toma el subyacente a lo largo de la vigencia de la opción.

Como se sabe, Pemex importa de Estados Unidos parte del gas que se consume en México. La mayor parte de éste lo adquiere al precio spot, precio diario, en el mercado de Houston y ya en México lo vende a los distribuidores al precio de referencia. Éste es fijado mensualmente por la CRE, Comisión Reguladora de Energía, tomando en cuenta el precio futuro mensual del gas del NYMEX. A Pemex le interesa cubrirse de la diferencia entre el precio futuro mensual y el precio spot. Para ello, se usan las opciones tipo swing que dependen, tanto del valor promedio del valor spot y del precio futuro del NYMEX como del volumen.

En este trabajo se presenta una aplicación del método de Monte Carlo, con la técnica de reducción de varianza, para estimar el valor de una opción asiática, que depende del precio spot, que se denotará como $\{S_t, t \geq 0\}$ y del precio futuro mensual del gas, $\{R_t, t \geq 0\}$. Esta es una opción cuyo

¹vibarra@anahuac.mx

²psb@xanum.uam.mx

precio de ejercicio es igual a R_T , con una madurez de un mes, o sea $T = 1/12$, y función de pago $\max\{\frac{1}{T} \int_0^T S_t dt - R_T, 0\}$.

Debido a que en el mercado de energéticos, el subyacente presenta reversión a la media, es decir, el valor del subyacente tiende a un valor de equilibrio, modelarlo por medio de un browniano geométrico no es lo más conveniente, pero los resultados obtenidos en trabajos anteriores, véase Saavedra e Ibarra (2005), muestran que para opciones con tiempo de maduración de un mes y menores, el browniano geométrico es una buena aproximación. Por la misma razón, no se tomará en cuenta la estacionalidad del precio. En este trabajo haremos las siguientes suposiciones:

- i. Se tiene un espacio de probabilidad (Ω, \mathcal{F}, P) dotado con una filtración $\{F_t, t \geq 0\}$.
- ii. El mercado es viable y completo.
- iii. El precio spot del gas satisface

$$S(0) = S_0, \quad dS_t = \mu S_t dt + \sigma_1 S_t dW_t^1,$$

con W_t^1 un browniano estándar, y el precio futuro mensual del Nymex lo denotaremos por R_t que satisface

$$R(0) = R_0, \quad dR_t = \alpha R_t dt + \sigma_2 R_t dW_t^2,$$

con W_t^2 un browniano correlacionado con W_t^1 .

2. Valuación mediante Monte Carlo

Como el mercado es viable y completo existe una única probabilidad equivalente P^* , la probabilidad de riesgo neutro, bajo la cual los precios descontados $\tilde{S}_t = e^{-rt} S_t$ y $\tilde{R}_t = e^{-rt} R_t$ son martingalas, véase Lamberton y Lapeyre (1996), con r la tasa libre de riesgo.

Bajo la probabilidad de riesgo neutro los procesos que siguen el precio spot y el precio Nymex mensual, respectivamente, son los siguientes: dado $S(0) = S_0$ y $R(0) = R_0$

$$dS_t = r S_t dt + \sigma_1 S_t dB_t^1, \tag{1}$$

$$dR_t = r R_t dt + \sigma_2 R_t dB_t^2, \quad (2)$$

B_t^1, B_t^2 brownianos correlacionados.

El valor de la opción se obtiene al calcular la siguiente esperanza condicionada a la información al tiempo t , $V_t = E^* \left(e^{-r(T-t)} \left(\frac{1}{T} \int_0^T S_t dt - R_T \right)_+ | F_t \right).$

No es posible determinar una solución analítica, debido a que la integral de un browniano geométrico, no es un browniano geométrico, por lo que la única alternativa es aproximar la solución numéricamente por medio de árboles binarios, por Monte Carlo o approximando la solución de una ecuación en derivadas parciales. Los árboles binarios no son muy recomendables para opciones asiáticas, pues se deben generar todas las trayectorias, lo que rápidamente se convierte en un problema computacional. Si se opta por transformar el problema del cálculo de una esperanza condicionada en un problema en derivadas parciales, se obtendrá una ecuación en tres dimensiones, dado que el valor de la opción depende de S_t , R_t y de $I_t = \int_0^t S_u du$. Así que, desde el punto de vista numérico, el método más eficiente y sencillo de implementar es el método Monte Carlo. Para ello es necesario aproximar cada trayectoria que pueden seguir S_t y R_t , lo que haremos usando la solución exacta de las ecuaciones (1) y (2). Dado $N \in \mathcal{N}$ sea $h = 1/N$ definamos como $t_0 = 0$, $t_i = ih$ y $t_N = T$.

Dadas S_0 y R_0 , denotemos como $S_i = S(t_i)$ entonces $S_i = e^{(r-\frac{\sigma_1^2}{2})h+\sigma_1\sqrt{h}\varepsilon_i^1}$, con $\varepsilon_i^1 \sim N(0, 1)$ y $R_i = e^{(r-\frac{\sigma_2^2}{2})h+\sigma_2\sqrt{h}z_i}$, con $z_i = \sqrt{\rho}\varepsilon_i^1 + \sqrt{1-\rho}\varepsilon_i^2$, y $\varepsilon_i^2 \sim N(0, 1)$. Para aproximar el valor de la integral, usaremos un trapecio compuesto, es decir $I_T = \int_0^T S_u du \approx I_N = \sum_{i=0}^N S_i [1 + \frac{rh}{2} + \frac{\sigma[W_i - W_{i-1}]}{2}]$. Por último, al aplicar Monte Carlo se obtiene que

$$V_0 \approx V_0^{M,N} = \frac{e^{-rT}}{T M} \sum_{j=1}^M (I_N^j - R_N)_+,$$

con M el número total de trayectorias.

El teorema de límite central nos permite obtener un intervalo de confianza de nivel α para el valor de V_0 , a saber $V_0 \in \left(V_0^{M,N} - \frac{\sigma z_\alpha}{\sqrt{M}}, V_0^{M,N} + \frac{\sigma z_\alpha}{\sqrt{M}} \right)$.

Observemos que el estimador $V_0^{M,N}$ no es un estimador insesgado, o sea no satisface que $E(V_0^{M,N}) = V_0$. Esto se debe a que estimamos la integral. Sin embargo, el error que introducimos es pequeño $o(1/N^{3/2})$, en comparación del error de Monte Carlo: $o(\frac{1}{\sqrt{M}})$, véase Lapeyre y Temam (2001).

Uno de los inconvenientes al utilizar el método de Monte-Carlo es la lentitud de su convergencia. Con el fin de acelerarla se utilizan varias técnicas para la reducción de varianza, véase Glasserman (2004). En este trabajo se usará la técnica de reducción de varianza por medio de una variable de control.

3. Técnica de variable de control para reducción de varianza.

Este método tiene como base la idea siguiente: si queremos estimar $E(X)$ y al hacerlo deseamos reducir lo más posible la varianza, definimos una nueva variable $Z = X + \beta[Y - E(Y)]$, con Y una variable aleatoria, denominada variable de control, cuya esperanza es conocida y que esté correlacionada con X .

Observemos que $E(Z) = E(X)$ pero

$$Var(Z) = Var(X) + \beta^2 Var(Y) + 2\beta Cov(X, Y). \quad (3)$$

Para reducir la varianza de Z basta obtener el valor de β que minimiza a $Var(Z)$. Es decir, $\beta^* = \frac{-Cov(X, Y)}{Var(Y)}$.

Por lo que al substituir β^* en (3), se obtiene: $Var(Z) = Var(X) - \frac{Cov^2(X, Y)}{Var(Y)}$. De esta forma hemos obtenido un procedimiento que permite estimar la $E(X)$ reduciendo la varianza. En nuestro caso utilizamos como variable de control la función de pago evaluada en la media geométrica del precio spot a lo largo del mes. Es decir, $Y = (e^{\frac{1}{T} \int_0^T \ln S_t dt} - R_T)_+$.

El valor esperado de Y se puede determinar en forma exacta porque $\exp\left(\frac{1}{T} \int_0^T \ln S_t dt\right)$ es un browniano geométrico, y se puede obtener una fórmula del estilo de Black-Scholes para calcularlo, véase Ibarra (2004). Así que el problema se reduce a estimar por Monte Carlo la esperanza de Z . Denotemos con $RV_0^{M,N}$ la estimación de $E(Z)$.

4. Resultados numéricos

Dada una opción con $R_0 = 6.515$, $S_0 = 6.515$, $r = 0.05$, $\sigma_1 = .5518$, $\sigma_2 = .4402$, $\rho = .8596$ y $T = 1/12$, es decir, un mes. A continuación se presentan los resultados obtenidos con y sin reducción de varianza.

Tabla 1: Con N=20

M	Monte Carlo Crudo			Con reducción de varianza		
	Intervalo al 95 %	$V_0^{M,N}$	Desv	Intervalo al 95 %	$RV_0^{M,N}$	Desv
1000	[0.19402, 0.23151]	0.21277	0.00956	[0.20781, 0.23324]	0.22053	0.00649
10000	[0.20543, 0.21727]	0.21135	0.00302	[0.20710, 0.21526]	0.21118	0.00208
100000	[0.21211, 0.21588]	0.21400	0.00096	[0.21244, 0.21502]	0.21373	0.00066
1000000	[0.21293, 0.21413]	0.21353	0.00030	[0.21308, 0.21390]	0.21349	0.00021
10000000	[0.21320, 0.21357]	0.21339	0.00010	[0.21328, 0.21354]	0.21341	0.00007

Tabla 2: Con N=40

M	Monte Carlo Crudo			Con reducción de varianza		
	Intervalo al 95 %	$V_0^{M,N}$	Desv	Intervalo al 95 %	$RV_0^{M,N}$	Desv
1000	[0.19956, 0.23738]	0.21847	0.00965	[0.21077, 0.23726]	0.22402	0.00676
10000	[0.21097, 0.22301]	0.21699	0.00307	[0.21054, 0.21880]	0.21467	0.00211
100000	[0.21013, 0.21392]	0.21203	0.00096	[0.21133, 0.21391]	0.21262	0.00066
1000000	[0.21287, 0.21407]	0.21347	0.00030	[0.21303, 0.21385]	0.21344	0.00021
10000000	[0.21327, 0.21365]	0.21346	0.00010	[0.21327, 0.21353]	0.21340	0.00007

Los resultados numéricos nos muestran que con reducción de varianza se mejora la eficiencia de los métodos Monte Carlo. Este método requiere de una “buena” variable de control, a fin de reducir de manera significativa la varianza. Se recomienda, para disminuir el tiempo de máquina que el cálculo de la varianza y covarianza se haga de manera recursiva, véase Ross (2003). Para más sobre esta técnica consultar Ibarra (2004) y Lapeyre y Temam (2001). En el mercado el valor de esta opción es de 0.15. Es difícil reproducir los valores del mercado dado que en nuestro caso, las varianzas y covarianzas son las históricas. Sin embargo varias referencias, véase Zhang (1998), nos indican que el valor del mercado se estima aproximando la integral por el promedio geométrico; en este trabajo este valor se usa como variable de control.

5. Agradecimientos

1. La media geométrica subestima el valor de la opción.

2. Los estimadores $V_0^{M,N}$ y $RV_0^{M,N}$ de la opción no son insesgados, debido a que la integral se aproxima numéricamente. La varianza de $RV_0^{M,N}$ es menor a $V_0^{M,N}$ por lo que es un mejor estimador.

3. En todos los casos, con reducción de varianza, la disminución del intervalo de confianza fue de más del 50%.

Referencias

Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer.

Ibarra Mercado V.H. (2004). *Reducción de varianza en la valuación de opciones euroasiáticas por medio del método Monte Carlo*. Tesis de Maestría en Métodos Matemáticos en Finanzas. Universidad Anáhuac.

Lapeyre B. y Temam E. (2001). Competitive Monte Carlo methods for the Pricing of Asian Options. *Journal of Computational Finance*, 5-1.

Lamberton, D. y Lapeyre B. (1996). *Introduction to Stochastic Calculus Applied to Finance*. Chapman and Hall.

Pilopovic, Dragana (1997) *Energy Risk*. Mc Graw Hill.

Ross, S. (2003). *Simulation*. Tercera edición. Prentice Hall.

Saavedra P. e Ibarra V.H. (2005). *Valuación de opciones asiáticas en el mercado de energéticos*. Trabajo presentado en el VII Congreso Franco-Latinoamericano en Santiago de Chile.

Zhang P. (1998). *Exotic Options: A guide to second generation options*. World Scientific, Segunda edición.

Planes Optimos para Pruebas de Vida Acelerada con Esfuerzos Escalonados

José del Carmen Jiménez Hernández¹

Universidad Tecnológica de la Mixteca

Enrique R. Villa Diharce²

Centro de Investigación en Matemáticas, A.C.

1. Introducción

Las pruebas de vida acelerada de un producto o material son usadas para obtener información rápida de su distribución de vida. Las unidades en prueba se someten a niveles de esfuerzo alto y fallan más tempranamente que en condiciones de diseño. La información que se obtiene en condiciones aceleradas se analiza en términos de un modelo y después se extrapola en condiciones de diseño para estimar la distribución de vida. Tales pruebas reducen tiempo y costo. Algunos niveles de esfuerzo alto involucran, temperatura, voltaje, presión, o algunas combinaciones de estos.

Los esfuerzos se pueden aplicar de diferentes maneras, los métodos más comunes son esfuerzo constante y esfuerzo escalonado. En una prueba de vida acelerada (PVA) bajo esfuerzo constante, cada unidad se somete a esfuerzo constante durante el estudio, y la vida a nivel de diseño se estima por métodos de regresión. En una prueba con esfuerzo escalonado, una unidad se somete sucesivamente a esfuerzos de nivel creciente. Primero una unidad se somete a un nivel de esfuerzo constante por una duración de tiempo especificada, si no falla, se somete a otro nivel de esfuerzo más alto por otra duración de tiempo especificada y así sucesivamente. De esta forma, el esfuerzo sobre una unidad incrementa paso a paso hasta que falla. Usualmente todas las unidades siguen el mismo patrón de niveles de esfuerzo y tiempos de prueba especificados.

El objetivo del presente trabajo, principalmente es obtener planes óptimos para PVA con esfuerzos escalonados suponiendo que los tiempos de falla de unidades expuestas a esfuerzos constantes siguen una distribución Weibull.

¹jcjim@mixteco.utm.mx

²villadi@cimat.mx

Notación

n	— Tamaño total de la muestra
x_0, x_1, x_2	— Esfuerzos transformados, de diseño, bajo y alto, respectivamente
β_0, β_1	— Parámetros de la función loglineal entre el esfuerzo y la vida característica η
γ_0, γ_1	— Parámetros del modelo reparametrizado
ξ	— Factor de extrapolación
τ	— Longitud de tiempo al esfuerzo bajo
τ^*	— Tiempo de prueba óptimo al nivel x_1
T	— Tiempo de censura
p_1^*, p_2^*, p_c^*	— Proporción de fallas esperada al nivel x_1, x_2 y después de T

2. El modelo

Para analizar datos de una prueba bajo esfuerzo escalonado, uno necesita un modelo que relacione la distribución de vida bajo esfuerzo escalonado y la distribución de vida bajo esfuerzo constante, el modelo usado aquí es el expuesto por Nelson (1990), en el cual supone que la vida restante de las unidades solo depende de la fracción de fallas acumuladas y del esfuerzo actual y que las unidades que no han fallado en este nivel de esfuerzo, lo harán de acuerdo a la función de distribución acumulada de este esfuerzo pero empezando previamente en la fracción de fallas que se ha acumulado, por tal razón se le llama, modelo de daño acumulado, el cual se usa para la estimación de los parámetros.

3. Planes óptimos modelo log-lineal Weibull

Suposiciones básicas.

1. Sólo se usan dos niveles de esfuerzo x_1 y x_2 ($x_1 < x_2$), los cuales están dados.
2. Para cualquier nivel de esfuerzo, la distribución del tiempo a la falla es Weibull con parámetro de escala $\eta(x) = \exp(\beta_0 + \beta_1 x)$ y parámetro de forma β constante y es independiente del esfuerzo, es decir, los log-tiempos a la falla siguen una distribución de valores extremos (VE) para mínimos

con parámetro de localización $\mu(x) = \ln \eta(x) = \beta_0 + \beta_1 x$ y parámetro de escala $\sigma = 1/\beta$.

3. El modelo de daño acumulado es válido.

Criterio de optimización.

El criterio de optimización usado aquí es minimizar la varianza asintótica del estimador de máxima verosimilitud (EVM) del cuantil P de la distribución a un nivel de diseño especificado.

Planteamiento de la prueba.

Supóngase que se tienen n unidades al inicio de la prueba y trabajan hasta un tiempo τ . Si la unidad no ha fallado, se cambia al esfuerzo x_2 y la prueba continua hasta que todas las unidades fallan ó hasta un tiempo de censura especificado.

Modelo reparametrizado.

Es conveniente reparametrizar el modelo, definamos el factor de extrapolación como, $\xi_j = \frac{(x_j - x_2)}{(x_0 - x_2)}$, $j = 0, 1, 2$, donde x_0 y x_2 son los niveles de esfuerzo de diseño y alto, respectivamente. Nótese que para el nivel de esfuerzo alto $x = x_2$, $\xi_2 = 0$; y para el nivel de diseño $x = x_0$, $\xi_0 = 1$. Entonces, el parámetro de localización $\mu(x)$ de la distribución de VE puede escribirse en términos de ξ_j como

$$\mu(\xi_j) = \gamma_0 + \gamma_1 \xi_j, \quad j = 0, 1, 2. \quad (1)$$

donde los nuevos parámetros γ_0 y γ_1 se relacionan con β_0 y β_1 por $\gamma_0 = \beta_0 + \beta_1 x_2$ y $\gamma_1 = \beta_1(x_0 - x_2) = \mu(x_0) - \mu(x_2)$. El EMV del cuantil P de la distribución es $\hat{y}_P = \hat{\gamma}_0 + \hat{\gamma}_1 \xi + u_P \hat{\sigma}$, por lo que a nivel de diseño x_0 , es decir, cuando $\xi_0 = 1$, es $\hat{y}_P = \hat{\gamma}_0 + \hat{\gamma}_1 + u_P \hat{\sigma}$, donde $\hat{\gamma}_0$, $\hat{\gamma}_1$ y $\hat{\sigma}$ son los EMV y $u_P = \ln[-\ln(1 - P)]$ es el cuantil P de la distribución de VE estándar. La varianza asintótica de éste estimador es el valor de la forma cuadrática

$$\text{Var}(\hat{y}_P) = [1, 1, u_P] \Sigma [1, 1, u_P]', \quad (2)$$

donde la ' denota el vector transpuesto y Σ es la matriz de varianzas y covarianzas de los parámetros $\hat{\gamma}_0$, $\hat{\gamma}_1$ y $\hat{\sigma}$, la cual se obtiene de la manera usual, tomando la inversa de la matriz de información de Fisher. Los elementos de esta matriz se obtienen tomando las esperanzas negativas a las segundas derivadas parciales de la función de log-verosimilitud del modelo reparametrizado. La varianza

asintótica del EMV del cuantil P de la distribución a nivel de diseño dada en (2) toma la forma,

$$\begin{aligned}\text{Var}(\hat{y}_P) &= \text{Var}(\hat{\gamma}_0) + \text{Var}(\hat{\gamma}_1) + u_P^2 \text{Var}(\hat{\sigma}) + 2\text{Cov}(\hat{\gamma}_0, \hat{\gamma}_1) + \\ &+ 2u_P \text{Cov}(\hat{\gamma}_0, \hat{\sigma}) + 2u_P \text{Cov}(\hat{\gamma}_1, \hat{\sigma}),\end{aligned}\quad (3)$$

la cual es función de τ , T , ξ_1 , ξ_2 , γ_0 , γ_1 y σ .

Para el plan óptimo considerado aquí, se quiere el τ óptimo que minimiza (3) del cuantil \hat{y}_P estimado en condiciones de diseño, dados los valores de x_0 , x_1 , x_2 , γ_0 , γ_1 , σ , T y P . La probabilidad P corresponde al cuantil y_P de interés, como sabemos, los valores de los parámetros del modelo de vida acelerada γ_0 , γ_1 y σ se suponen conocidos y los niveles de esfuerzo de diseño x_0 y los de prueba x_1 y x_2 se fijan previamente, al igual que el tiempo de duración de la prueba. Una vez que encontramos el valor de τ que minimiza (3), hemos determinado el plan óptimo para el modelo log-lineal Weibull.

4. Ejemplo

Con el interés de obtener la información necesaria para determinar un plan óptimo para una PVA con esfuerzos escalonados, para un tipo de interruptores, se llevó a cabo una prueba piloto acelerando la falla con temperatura, iniciando con $120^\circ C$ y terminando con $200^\circ C$. Lo anterior ya que el ingeniero de confiabilidad que realizó la prueba, comentó que con esfuerzo mayor se podrían generar otros modos de falla diferentes al de diseño. El esfuerzo de prueba se elevó en forma escalonada cada 10 Kciclos y posteriormente cada 5 Kciclos, la prueba terminó a los 125 Kciclos. El modelo de vida acelerada ajustado fue el Arrhenius-Weibull, los parámetros estimados fueron, $\hat{\beta}_0 = 0.99$, $\hat{\beta}_1 = 0.13$ y $\hat{\sigma} = 0.22$. De acuerdo a la experiencia del estudio piloto, elegimos los siguientes valores de las temperaturas para la determinación del plan óptimo, $T_0 = 80^\circ C$, $T_1 = 145^\circ C$ y $T_2 = 200^\circ C$, T_j , $j = 0, 1, 2$ son las temperaturas de diseño, primer y segundo nivel de esfuerzo, con una duración de la prueba de 200 Kciclos, es decir, el tiempo de censura es $T = 200$. En el modelo de vida acelerada log-lineal requerimos el esfuerzo transformado dado por, $x_j = (11605)/(276.15 + T_j^\circ C)$, $j = 0, 1, 2$. Así, se tiene $x_0 = 32.86$, $x_1 = 27.75$ y $x_2 = 24.53$. De x_0 , x_1 y x_2 se obtienen los factores de extrapolación y los nuevos parámetros, dados por, $\xi_1 = 0.39$ y $\xi_2 = 0$, y $\hat{\gamma}_0 = 4.23$ y $\hat{\gamma}_1 = 1.12$.

Esta información se introduce en un código programado en Splus que nos dá la varianza del cuantil \hat{y}_P como función de τ y se procede a minimizar $\text{Var}(\hat{y}_P)$ para diferentes valores de P . En la Tabla

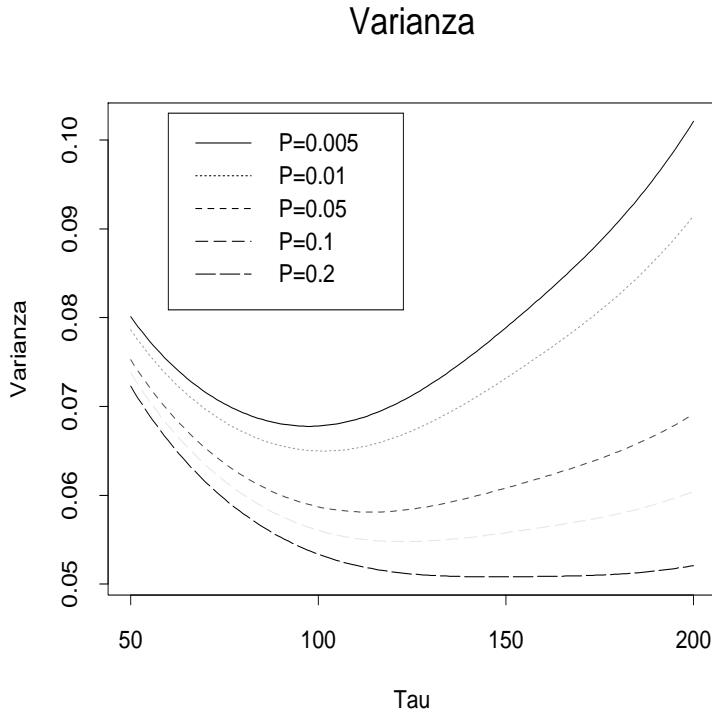


Figura 1: Efecto de la varianza para el diferentes valores de P

1 se exhibe el valor de τ^* de τ que minimiza la varianza y también exhibe $\text{Var}(\hat{y}_P)$, para diferentes valores de P , así como las proporciones de fallas esperadas en los niveles de esfuerzo bajo, alto y después de T .

P	τ^*	$\text{Var}(\hat{y}_P)$	p_1^*	p_2^*	p_c^*
0.005	92.00	0.071	0.31	0.68	0.001
0.010	95.79	0.068	0.37	0.56	0.70
0.050	106.70	0.061	0.54	0.24	0.22
0.100	113.65	0.058	0.64	0.11	0.25
0.200	125.70	0.054	0.80	0.016	0.18

Tabla 1: Tiempos de Permanencia Optimos

Los resultados anteriores nos indican que por ejemplo, si se está interesado en el cuantil 0.1, dada la información previa, las unidades en prueba estarían en el primer nivel de esfuerzo hasta aproximadamente 114 Kciclos para después cambiar al segundo nivel de esfuerzo y continuar la prueba hasta el tiempo de censura, es decir, hasta $T = 200$. En la Tabla 1 observamos que al aumentar el valor de P , el tiempo óptimo de permanencia de las unidades en prueba en el nivel de esfuerzo bajo

aumenta, mientras que con la varianza del cuantil estimado ocurre lo contrario, va disminuyendo. En la Figura 1, se muestra la relación entre el tiempo de permanencia de las unidades en el nivel de esfuerzo bajo y la varianza del cuantil estimado. En todos los casos tenemos una función convexa donde el mínimo de va desplazando a la izquierda a menudo que crece la probabilidad P del cuantil \hat{y}_P .

5. Conclusiones

Se desarrolla un plan óptimo para una prueba acelerada con esfuerzos escalonados, para el modelo log-lineal Weibull. En este caso se obtiene el tiempo óptimo de aplicación del esfuerzo bajo de prueba, para el caso en que la prueba tiene una duración T , predeterminada.

Referencias

- Abdulla A. Alhadeed and Shei-Shein Yang. (2002), Optimum Simple Step-Stress Plan for Khamis-Higgins Model, *IEEE Transactions on Reliability*, **R-51**, 212-215.
- Bai D. S., Cha M. S., and Chung S. W. (1992), Optimum Simple Ramp-Test for Weibull Distribution and Type-I Censoring, *IEEE Transactions on Reliability*, **R-41**, 407-412.
- Meeker, W. Q., and Escobar L. A. (1998), *Statistical Methods for Reliability Data*, John Wiley & Sons, Inc.
- Miller R., and Nelson, W. B. (1983), Optimum Simple Step-Stress Plans for Accelerated Life Testing, *IEEE Transactions on Reliability*, **R-29**, 103-108.
- Nelson, W. B. and Kielpinski, T. (1980), Accelerated Life Testing - Step Stress Model and Data Analyses, *IEEE Transactions on Reliability*, **R-29**, 103-108.
- Nelson, W. B. (1990), *Accelerated Testing: Statistical Models Test Plans, and Data Analyses*, New York: John Wiley & Sons.

Una Prueba para Exponencialidad Basada en la Razón de dos Estimadores

María Diódora Kantún Chim¹

Universidad Autónoma de Yucatán

José Aurelio Villaseñor Alva²

Colegio de Postgraduados

1. Introducción

A partir de 1960, véase por ejemplo Spurrier (1984), se han propuesto pruebas de bondad de ajuste para la distribución exponencial, ya que esta distribución es de mucha importancia en el análisis del tiempo de vida. Se han desarrollado pruebas para exponencialidad para casos de datos completos, datos censurados o para ambos casos. Algunas de las pruebas propuestas se han construido considerando distribuciones establecidas en la hipótesis alternativa. El objetivo del presente estudio es proponer una prueba basada en la razón de un estimador para la media y otro para la desviación estándar, ambos son función del máximo y mínimo de una muestra aleatoria y compararla contra otras consideradas entre la literatura que en general proporcionan mayor potencia.

2. Distribución R

Una variable aleatoria X tiene una distribución exponencial si tiene una función de densidad de la forma:

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} I_{(x,\infty)} \quad \beta > 0$$

donde β es el parámetro de escala. Sea X_1, \dots, X_n una muestra aleatoria de tamaño n de la distribución exponencial, obsérvese que:

$$\mu = E(X_i) = \beta$$

$$\sigma = \sqrt{Var(X_i)} = \beta$$

¹kchim@tunku.ady.mx

²jvillasr@colpos.mx

Sean:

$$Y_1 = \min(X_1, \dots, X_n) \quad Y_n = \max(X_1, \dots, X_n)$$

y considérese como estimadores de μ y σ las estadísticas:

$$\hat{\mu} = \frac{Y_1 + Y_n}{2} \quad \hat{\sigma} = Y_n - Y_1$$

semirango y rango respectivamente. Ahora definase la estadística R como la razón del rango entre dos veces el semirango:

$$R = \frac{\hat{\sigma}}{2\hat{\mu}} = \frac{Y_n - Y_1}{Y_n + Y_1}$$

Para probar las hipótesis:

$H_0 : X_1, \dots, X_n$ proviene de una distribución exponencial

vs

$H_A : X_1, \dots, X_n$ no proviene de una distribución exponencial.

Se propone la prueba R para rechazar H_0 si $R < r_1$ o $R > r_2$, donde para una prueba de tamaño α^* los valores r_1 y r_2 son tales que:

$$P(r_1 < R < r_2) = 1 - \alpha^*$$

Una manera de seleccionar a r_1 y r_2 es :

$$P(R < r_1 | H_0) = P(R > r_2 | H_0) = \frac{\alpha^*}{2}$$

donde la función de densidad de R resulta ser:

$$f_R(r) = 2n(n-1) \sum_{k=0}^{n-2} \binom{n-2}{k} \frac{(-1)^{n-2-k}}{(r(n-2-2k)+n)^2} \quad 0 < r < 1$$

con parámetro n que es precisamente el tamaño de la muestra. La prueba es restringida para tamaños de muestra $n \leq 45$ por limitaciones en el cálculo numérico. En el Cuadro 1 se presenta los cuantiles para los diferentes tamaños de muestra para la prueba R , donde la aproximación numérica es con un error de $\delta=0.0001$

CUADRO 1. Cuantiles estimados de R con error $\delta=0.0001$ para diferentes α^*								
n/α^*	0.005	0.0125	0.025	0.05	0.95	0.975	0.9875	0.995
2	0.005004883	0.01245117	0.02490234	0.05004883	0.9499512	0.9750977	0.9875488	0.9949951
4	0.187500000	0.25390625	0.31933594	0.39990234	0.9860840	0.9931030	0.9965820	0.9986572
6	0.390625000	0.46679688	0.53222656	0.60546875	0.9925232	0.9963074	0.9981689	0.9992676
8	0.535156250	0.60253906	0.65820312	0.71704102	0.9950562	0.9975586	0.9987793	0.9995117
10	0.630859375	0.68945312	0.73583984	0.78393555	0.9963684	0.9982147	0.9991150	0.9996490
12	0.699218750	0.74902344	0.78759766	0.82739258	0.9971619	0.9985962	0.9993057	0.9997253
14	0.748046875	0.79101562	0.82421875	0.85766602	0.9976959	0.9988632	0.9994354	0.9997711
16	0.784179688	0.82226562	0.85083008	0.87963867	0.9980621	0.9990463	0.9995270	0.9998093
18	0.812500000	0.84570312	0.87109375	0.89624023	0.9983368	0.9991837	0.9995956	0.9998398
20	0.834960938	0.86474609	0.88696289	0.90905762	0.9985504	0.9992867	0.9996452	0.9998589
22	0.853515625	0.87939453	0.89965820	0.91937256	0.9987183	0.9993668	0.9996872	0.9998741
24	0.868164062	0.89208984	0.90991211	0.92773438	0.9988556	0.9994354	0.9997177	0.9998894
26	0.879882812	0.90209961	0.91845703	0.93469238	0.9989624	0.9994888	0.9997482	0.9999008
28	0.890625000	0.91064453	0.92578125	0.94042969	0.9990578	0.9995346	0.9997673	0.9999084
30	0.899414062	0.91796875	0.93188477	0.94537354	0.9991341	0.9995728	0.9997883	0.9999161
32	0.907226562	0.92431641	0.93713379	0.94964600	0.9992027	0.9996071	0.9998055	0.9999237
34	0.914062500	0.92993164	0.94165039	0.95330811	0.9992599	0.9996357	0.9998188	0.9999275
36	0.919921875	0.93457031	0.94567871	0.95654297	0.9993114	0.9996605	0.9998322	0.9999332
38	0.924804688	0.93896484	0.94921875	0.95935059	0.9993553	0.9996834	0.9998436	0.9999371
39	0.927246094	0.94091797	0.95080566	0.96063232	0.9993763	0.9996929	0.9998474	0.9999390
40	0.929687500	0.94262695	0.95239258	0.96185303	0.9993973	0.9997025	0.9998512	0.9999409
42	0.934082031	0.94604492	0.95520020	0.96411133	0.9994316	0.9997196	0.9998608	0.9999447
44	0.937500000	0.94894409	0.95777893	0.96612549	0.9994621	0.9997349	0.9998684	0.9999475
45	0.939453125	0.94995117	0.95877075	0.96710205	0.9994812	0.9997412	0.9998723	0.9999456

3. Comparación de pruebas

Con base en los estudios realizados por Spurrier (1984) y Ascher (1990), se concluye que las pruebas de Cox y Oakes (1984) y la HP de Hollander y Proschan (1972) resultan ser de las más potentes. Una prueba relacionada también con estadísticas de orden es propuesta por Wong y Wong (1979). Estas pruebas, incluyendo la prueba R, son consideradas en un estudio de comparación mediante simulación Monte Carlo. Para este estudio se utilizan algunas familias de distribuciones con la característica de tener a la distribución exponencial como un caso particular, estas son: Gamma, Weibull y Pareto Generalizada. Es de interés comentar que las distribuciones Gamma y Weibull tienen función de riesgo monótona, característica que Hollander y Proschan (1972) al igual que Cox y Oakes (1984) toman en consideración en la hipótesis alternativa para el desarrollo de sus pruebas.

Cuadro 2A. Potencias de las pruebas R, Q, Cox y HP cuando la H.A. es Gamma, Weibull o Pareto generalizada (1000 repeticiones)										
Potencias estimadas de las pruebas R, Q, Cox, y HP con H.A. Gamma($\alpha, \beta = 1$), $n = 20$ y $\alpha^*=0.05$.										
PRUEBAS/ α 0.2 0.5 0.7 0.8 1.0 1.5 2.0 4.0 8.0										
R	0.995	0.447	0.162	0.104	0.054	0.185	0.440	0.949	1.000	
Q	0.997	0.592	0.255	0.148	0.056	0.001	0.000	0.000	0.000	
COX	1.00	0.716	0.261	0.135	0.053	0.235	0.583	0.995	1.000	
HP	0.997	0.569	0.210	0.127	0.051	0.181	0.466	0.981	1.000	
Potencias estimadas de las pruebas R, Q, Cox, y HP con H.A. Weibull($\alpha, \beta = 1$), $n = 20$ y $\alpha^*=0.05$.										
PRUEBAS/ α 0.2 0.5 0.7 0.8 1.0 1.2 1.5 2.0 4.0 8.0										
R	1.000	0.700	0.225	0.116	0.051	0.107	0.393	0.788	1.000	
Q	1.000	0.818	0.358	0.186	0.063	0.014	0.002	0.000	0.000	
COX	1.000	0.965	0.553	0.239	0.043	0.156	0.586	0.970	1.000	
HP	1.000	0.896	0.459	0.197	0.051	0.125	0.484	0.935	1.000	
Potencias estimadas de las pruebas R, Q, Cox, y HP con H.A. Pareto Generalizada($\alpha, \beta = 1$), $n = 20$ y $\alpha^*=0.05$.										
PRUEBAS/ α -5.0 -3.0 -1.0 -0.3 -0.1 -0.01 0.01 0.3 1.0 3.0 5.0										
R	0.995	0.964	0.278	0.069	0.047	0.046	0.049	0.103	0.323	0.684
Q	0.998	0.988	0.428	0.115	0.059	0.050	0.051	0.035	0.018	0.001
COX	1.000	1.000	0.798	0.190	0.062	0.052	0.053	0.122	0.598	0.963
HP	0.999	0.979	0.532	0.106	0.067	0.056	0.056	0.100	0.685	0.997

Cuadro 3B. Potencias estimadas de las pruebas R, Q, Cox y HP con diferentes distribuciones en la H.A con $n=30$ y $\alpha^* = 0,05$				
DISTRIBUCIONES	R	Q	Cox	HP
Beta (0.2,0.2)	0.963	0.977	0.929	0.224
Beta (0.3,0.3)	0.733	0.806	0.678	0.214
Beta (0.4,0.4)	0.462	0.554	0.358	0.249
Lognormal(0,0.8)	0.624	0.000	0.484	0.461
Lognormal(0,0.9)	0.339	0.000	0.224	0.224
Lognormal(5,0.8)	0.598	0.000	0.473	0.497
Lognormal(5,0.9)	0.302	0.000	0.218	0.230

En el Cuadro 2A se observa que cuando la H.A. es Gamma (α, β) con $\alpha \leq 0.2$ ó $4 \leq \alpha$ las pruebas R, Cox y HP presentan potencias altas. Si la H.A. es una distribución Weibull (α, β) con $\alpha \leq 0.5$ ó $2 \leq \alpha$ las pruebas R, Cox y HP presentan potencias altas. Ahora, si la H.A. es la distribución Pareto Generalizada (α, β), las pruebas de Cox y HP tienen potencias similares, y más altas, sin embargo, cuando $\alpha \leq -3$ ó $3 \leq \alpha$ las potencias de la prueba R son tan altas como las de HP y Cox.

En el Cuadro 2B se presentan distribuciones en las hipótesis alternativas en las cuales la prueba R resulta ser más potente que las pruebas de Cox y HP.

4. Conclusiones

La prueba R para exponencialidad no está enfocada hacia alguna distribución alternativa en particular. Los resultados comparativos obtenidos dan evidencias de la robustez de R al compararla con las pruebas de Hollander y Proschan y la de Cox y Oakes, las cuales han sido desarrolladas precisamente para distribuciones alternativas que tienen función de riesgo monótona. La prueba Q no muestra evidencia en general de tener potencia considerable para la hipótesis planteada. La prueba R presenta evidencias de mayor potencia que la de Cox y HP en algunas distribuciones Beta y lognormal.

Referencias

- Ascher, S. (1990). A Survey of Test for Exponentiality *Commun. Statist. Theory Methods*, 19(5): pp. 1811-1825.
- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Champan and Hall. First edition. pp: 13-47.
- D'Agostino, R. B. and Stephens, M. A. (1984). *Goodness-Of-Fit Techniques*. Marcel Dekker. pp: 421-453.
- Galambos, J. (1987). *The asymptotic Theory of Extreme Order Statistics*. Krieger Drive. Second Edition. pp: 1-86.
- Hollander, M. and Proschan, F. (1972). Testing Whether New is Better than Used. *The Annals of Mathematical Statistics* Vol. 43 No. 4: pp. 1136-1146.
- Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, Wiley-Interscience. Second Edition. pp: 495-517.
- Spurrier, J. D. (1984). An Overview of test for exponentiality. *Commun. Statistic. Theor. Meth.*,

13(13): pp 1635-1654.

Wong, P. G. and Wong, S. P. (1979). An Extremal Quotient Test for Exponential Distribution. *Metrika* Volume 26: pp.1-4.

Cálculo de Estimadores no Lineales y de sus Varianzas Estimadas a partir de Información de la Muestra del Censo Nacional de Población 2000.

Emilio López Escobar¹

Alberto Padilla Terán¹

Rigoberto Real Miranda¹

Martha Trejo González¹

Guillermrina Eslava Gómez²

Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México

1. Introducción

Con el fin de proporcionar datos actualizados de las principales características demográficas, sociales y económicas del país, durante el levantamiento del XII Censo General de Población y Vivienda se extrajo una muestra probabilística con las características generales siguientes.

- a) Se seleccionó una muestra de 2.3 millones de viviendas, que representa aproximadamente un 10 % del total de viviendas del país y permite generar información a nivel municipal en la mayor parte de los indicadores del cuestionario ampliado.
- b) Se aplicó un cuestionario ampliado a los integrantes de las viviendas en la muestra.
- c) El diseño fue estratificado (43,326 estratos) y en una sola etapa, con conglomerados o unidades primarias de muestreo de diferente tamaño (92,877).
- d) La unidad primaria de muestreo varió según el tipo de localidad, rural o no rural.
- e) Se incluyó a todos los municipios del país y las 16 delegaciones del D.F, cada uno fue considerado como un dominio de estudio.

¹Maestría en Ciencias Matemáticas del postgrado de la UNAM

²eslava@lya.fciencias.unam.mx

El uso de un método de aproximación de la varianza es apropiado cuando se tiene la expresión del estimador de la varianza, pero no se cuenta con la suficiente información para calcularla. Esto sucede con frecuencia ya que generalmente sólo se reporta el factor de expansión o peso muestral de cada unidad muestral, es decir, el inverso de la probabilidad de selección de cada unidad.

Los métodos de aproximación de varianza también se emplean cuando no existe una expresión específica para el estimador de la varianza. Como ejemplo de lo anterior, se tiene para el caso de un estimador puntual que sea una función no lineal de otros estimadores, como el estimador de razón el cual es un cociente de variables aleatorias. Para estimar su varianza, el estimador puntual se aproxima por Taylor con un estimador lineal, posteriormente se determina la expresión del estimador de la varianza, o bien se emplean técnicas de remuestreo. Se presentan las estimaciones puntuales del ingreso medio, mediana del ingreso y la proporción de viviendas con teléfono. Por otra parte, las estimaciones de la varianza se efectuaron por los tres métodos siguientes.

- i) Aproximación de Taylor, Woodruff (1971).
- ii) Jackknife (método de remuestreo).
- iii) Balanced repeated replications (BRR), también llamado Balanced Half-Samples (método de remuestreo).

Para el índice de concentración de Gini se emplearon los estimadores puntual (estimador de razón) y de varianza propuestos por Sandström, Wretman y Walden (1988). La construcción es para diseños arbitrarios y los estimadores puntual y de varianza son fórmulas explícitas que sólo dependen de los factores de expansión de las unidades.

Se compararon estimaciones puntuales y varianzas estimadas del ingreso medio mensual, la mediana del ingreso mensual y el índice de concentración de Gini (Stuart y Ord, 1994) para el ingreso medio mensual por hogar, así como de la proporción de viviendas con teléfono, empleando los diferentes tipos de estimadores y métodos de estimación de varianzas mencionados. Asimismo, se cuantificó el error que se comete al ignorar el diseño complejo y se hicieron observaciones generales sobre el uso adecuado de información proveniente de encuestas complejas en el proceso de estimación.

Cabe mencionar que para cada una de las variables se efectuaron estimaciones a nivel nacional, por entidad federativa, por área rural (localidades con un máximo de 2500 habitantes) y no rural y, a excepción del índice de concentración de Gini, se obtuvieron estimaciones considerando los diseños siguientes.

1. Diseño complejo; considera la estratificación y la conglomeración.
2. Ignorando diseño complejo; como si la selección de las unidades hubiera sido extraída por muestreo aleatorio simple (MAS).
3. Postestratificación en dos grupos, rural y no rural, considerando el diseño complejo.
4. Dominios planeados, urbano y rural, con diseño complejo.
5. Dominios no planeados, a nivel rural y no rural y considerando diseño complejo.

2. Resultados

Existe una gran cantidad de resultados acerca de los métodos de remuestreo; sin embargo, aquí sólo se incluyeron las propiedades deseables de un estimador de varianza: a) Insesgado o aproximadamente insesgado. b) Estable, es decir, se requiere que la varianza del estimador de varianza sea pequeña. c) Que produzca intervalos de confianza que cubran al parámetro de interés con una probabilidad aproximadamente igual a la establecida en el nivel de confianza.

En las tres cuadros que aparecen al final del texto se encuentran resultados numéricos de valores estimados. En el Cuadro 1 se muestran estimaciones calculadas a nivel nacional y usando la aproximación de Taylor, considerando cinco posibilidades: Bajo diseño, ignorando diseño (MAS), usando una postestratificación (rural y no rural), sólo para el dominio rural, y finalmente sólo para el dominio no rural. En el Cuadro 2 se presentan estimaciones puntuales y por intervalo; las estimaciones de las varianzas se hicieron considerando el diseño pero con tres métodos distintos: aproximación de Taylor, Jackknife, Balanced Half samples. En cuanto al índice de concentración de Gini, éste sólo se calculó para 5 estados de la república mexicana y se empleó el método de Sandström, Wretman y Walden (1988) para aproximar la varianza, los valores estimados se presentan en el Cuadro 3.

3. Conclusiones

Conclusiones en cuanto al tipo de estimador.

- a) Los resultados de las estimaciones nacionales mostraron los errores que se cometían al estimar

Cuadro 1: Estimaciones a nivel nacional usando la aproximación de Taylor para el cálculo de varianzas estimadas

	Diseño complejo	Ignorando diseño	Post-estratificación	Dominio rural	Dominios no rural
Ingreso medio	6005 ± 137	5258 ± 116	5944 ± 137	3106 ± 376	6807 ± 139
Mediana del Ing.	3085 ± 18	$2571 \pm .04$	3000 ± 9.3	1436 ± 33	3856 ± 1.40
Teléfono	$0.37 \pm .004$	$0.28 \pm .001$	$0.37 \pm .002$	$0.06 \pm .004$	$.46 \pm .004$

la varianza ignorando el diseño complejo, en general para diseños que involucran conglomeramiento las varianzas se subestiman.

- b) El uso de la postestratificación incrementó la precisión o fue igual que el estimador que considera solamente el diseño complejo.
- c) La reducción en la estimación de varianza usando postestratificación sólo se aprecia a nivel estatal o municipal y no a nivel nacional.
- d) Al postestratificar es necesario usar información auxiliar externa actualizada, con el fin de no introducir un sesgo en el estimador.
- e) Prácticamente no se encontraron diferencias entre el estimador con dominios planeados y no planeados. Esto se debió a que el tamaño de muestra en los dominios rural y no rural era grande a nivel estatal y nacional.
- f) Lo anterior generalmente no sucede en la práctica, ya que en el caso de dominios no planeados el tamaño de muestra es una variable aleatoria y los tamaños de muestra totales son pequeños.

Finalmente en cuanto al método de estimación de varianza podemos decir lo siguiente. Primero, la aproximación por Taylor produjo varianzas estimadas menores que las correspondientes a jackknife o a BRR. Segundo, una particularidad del método BRR comparado con el jackknife es que se aplica a un diseño específico: estratificado con dos unidades por estrato. En caso de que no se tengan dos unidades por estrato, como es el caso en este estudio, es necesario adecuar la base de datos colapsando o particionando estratos o unidades de primera etapa.

Cuadro 2: Estimaciones a nivel nacional usando la aproximación de Taylor para el cálculo de estimadores puntuales y tres métodos distintos para las varianzas estimadas

	Taylor	Jackknife	BRR
Ingreso medio	6005 ± 137	6005 ± 153	6005 ± 183
Mediana del Ingreso	3085 ± 18	–	3085 ± 98
Teléfono	$0.37 \pm .004$	$0.37 \pm .01$	$0.37 \pm .008$

Cuadro 3: Valores estimados del índice de Gini para 5 estados

Baja California	Hidalgo	Morelos	Querétaro	Yucatán
$.59 \pm .017$	$.57 \pm .016$	$.54 \pm .023$	$.69 \pm .021$	$.56 \pm .014$

Referencias

Sandström, A., Wretman, J. H. and Walden, B. (1988). Variance Estimation of the Gini Coefficient-Probability Sampling. *Journal of Business & Economic Statistics*. **6**, 113-119.

Stuart, A. and Ord, K. (1994). *Kendall's Advanced Theory of Statistics, Sixth Edition, Volume I, Distribution Theory*. Arnold Publishers.

Una Investigación sobre Dificultades del Proceso de Enseñanza Aprendizaje Relacionadas con Distribuciones de Probabilidad Continua

Juan Antonio López Esquivel¹

José Armando Albert Huerta²

Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Monterrey

1. Introducción

Las evidencias en los índices de reprobación de los estudiantes en los cursos de Estadística, las fallas interpretativas de usuarios a diversos niveles, las dificultades que muchos de nosotros experimentamos al estudiar un concepto, la falta de fluidez para lograr las transferencias de las ideas, la confusión sobre las diferentes interpretaciones que puede tener un concepto, entre otros, nos hacen ver la necesidad de abordar esta problemática de una forma más científica. Este trabajo forma parte de un estudio más completo en el que mostramos, desde el punto de vista de la Teoría de Situaciones Didácticas (TSD) de Brousseau (1986), la forma en que se puede abordar la problemática en torno a las distribuciones de probabilidad continua. En particular, mostraremos algunos resultados relevantes de los análisis preliminares epistemológico y didáctico.

2. Metodología

La metodología que se sigue en este trabajo es la Ingeniería Didáctica y es cualitativa principalmente. Consiste de buscar hacer diseños justificados con el propósito de hacer seguimientos de cómo es que los estudiantes aprenden, en interacción con sus iguales y con el profesor en situación escolar. La justificación de los diseños se compone de un análisis preliminar y formulación de hipótesis y sugerencias para el diseño. El análisis preliminar se realiza desde tres componentes: epistemológico, didáctico y cognitivo. Resulta de gran importancia el Análisis preliminar porque a través de

¹juanantonio.lopez@itesm.mx

²albert@itesm.mx

él pueden identificarse dificultades y obstáculos de diversa naturaleza que permitan observar con mayor detalle las trayectorias cognitivas que los estudiantes siguen, con el propósito de que el profesor cuente con conocimientos sobre qué es lo que sus alumnos hacen, qué tipo de ayuda debe darles y diseñar las mejores actividades para sus alumnos.

3. Importancia de la función de distribución acumulada

La Función de Distribución Acumulada es de amplio uso en estadística ya sea explícita o implícitamente: en principio, el cálculo de la probabilidad de que una variable aleatoria tome un valor entre un número y otro es mediante el uso de la FDA. La evaluación directa en la FDA nos suministra las probabilidades. Las tablas de percentiles de distribuciones de probabilidad que se usa en las teorías de prueba de hipótesis e intervalos de confianza son tablas de acumuladas. Aún cuando en ocasiones el investigador está preocupado inicialmente por asociar una FDP a determinada situación, hace uso de la FDA para calcular probabilidades. Incluso en niveles avanzados la FDA aparece en técnicas específicas, como su uso para hallar la distribución de probabilidad en la transformación de una variable aleatoria o en la generación de muestras aleatorias de una distribución particular (teorema de la transformación integral de probabilidad).

En los procesos de tiempo de falla una variable aleatoria puede ser caracterizada mediante cuatro formas alternativas: a) la FDA, denotada por $F(x)$, b) la FDP, denotada por $f(x)$, c) la función de sobrevivencia (confiabilidad), denotada por $S(t)$, y d) la función de riesgo, denotada por $h(x)$, como establecen Meeker y Escobar (1998). Burr (1942), anima al uso directo de la FDA aprovechando sus ventajas teóricas. En su artículo él señala que el proceso de encontrar la FDP a partir de la FDA es, al menos teóricamente, mucho más simple que a la inversa.

Actualmente la FDA juega un rol menor en el discurso de la enseñanza de la Probabilidad y Estadística. Esta investigación busca explorar la potencialidad teórica y didáctica de la FDA.

4. Análisis epistemológico histórico

Según nuestra investigación, históricamente primero surgen y se desarrollan las FDP a mediados del siglo XVIII y comienzos del siglo XIX. Sin embargo, hay indicios de que a finales del siglo XIX

Francis Galton (1822-1911) fue el primero que utilizó la idea de ordenar datos en orden ascendente graficándolos contra rangos y en usar esta técnica con el fin de clasificarlos; esta forma de proceder fue aconsejada por Galton por 1875, Stigler (1986). La Figura 2 muestra esbozos de gráficas que aparecen en el trabajo de Galton:

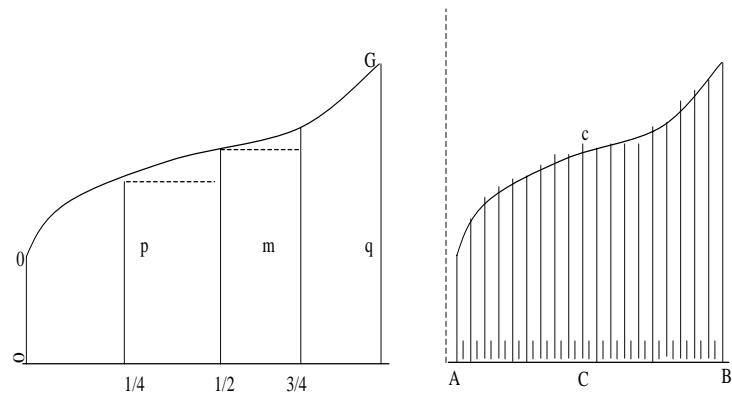


Figura 1: Gráficas de Galton. Ambas en Sitgler (1986)

Respecto a la idea de Galton, habría una curva ideal a la cual se ajustarían los datos si fueran homogéneos (esta curva corresponde a la FDA normal inversa). Él propuso el nombre de ojiva al gráfico resultante. Es de hacerse notar que hasta ese tiempo, Adolphe Quetelet (1796-1874) usaba la distribución normal para demostrar homogeneidad (al que se le criticaba por exagerar su uso), de hecho se piensa que a Quetelet se le ocurrió que la “constante” aparición de la distribución normal podría justificar el utilizarla para resolver el problema, planteado por el Barón de Keverberg, de decidir cuando un grupo de observaciones pueden ser tratadas como homogéneas. Galton ideó una nueva forma de verificar tal homogeneidad, aunque fue más allá, utilizando el concepto de acumu-

lada, que no llamaba así, sino ojiva, concepto tomado por él de la arquitectura, para un análisis que llamó “estadística por intercomparación”, Stigler (1986). En Galton (1880) se puede observar como el autor muestra la forma de obtener la ojiva y sugiere usarla para comparar poblaciones a detalle. Igualmente en Galton (1875), nos da una descripción detallada de como encontrar los percentiles y cuál es su utilidad. Y también en Galton (1886), concluye que la ojiva debe de poder tener representación matemática y que en particular corresponde, bajo ciertas condiciones, con la curva que se obtendría integrando la FDP normal. En otras palabras, Galton construye la FDA y se conecta, posteriormente, con la FDP. Es digno de recalcar que Galton llama ojiva, no al polígono de frecuencias como usualmente se hace en la literatura actual sino a la curva de la FDA, y además en la figura que menciona traza los percentiles en el eje vertical. Al parecer el artículo de 1875, es el primer artículo en el que Galton introdujo la ojiva, será comentado en un trabajo posterior. Según se ve, puede ser más natural graficar en unos ejes cartesianos la acumulada contra los valores de la variable aleatoria, que la razón de cambio de la acumulada contra esos valores, así, si el eje vertical proporciona la probabilidad acumulada para un valor de la variable aleatoria, y se tiene la acumulada, sólo se necesita evaluar la FDA. A su vez, si interesa obtener la razón de cambio de la FDA, sólo se tiene que derivar o usar los métodos numéricos necesarios para hallarla. Aunado a ello, si establecemos la interrogante: ¿qué hacer si no se dispone de la acumulada, pero sí de su razón de cambio (la FDP)? es decir: ¿cómo encontrar la FDA conociendo la FDP? pues sería entonces cuando se vería la necesidad de realizar un proceso de integración para obtener la cantidad de interés que finalmente es la FDA. Algo similar se puede hacer en el caso de estadística descriptiva, si tengo las frecuencias absolutas (o relativas) acumuladas, ¿cómo se obtienen las frecuencias absolutas (relativas)? O los histogramas correspondientes. Y esto podría servir para hacer la transición del análisis exploratorio al teórico. Hay, por tanto, indicios de que la FDP surgió primero en la historia pero ello no implica que así debería de ser en el aula, de hecho según las observaciones que más adelante se muestran, la acumulada, irónicamente, es un obstáculo en el estudio de la estadística.

5. Análisis didáctico

El análisis didáctico busca aquella evidencia que hay sobre formas institucionalizadas de que los estudiantes aprendan el tema: libros de texto, programas de estudio, entrevistas con profesores, actividades escolares con esa intención. En este apartado mostraremos el análisis de un libro de texto universitario, que en cierto sentido, consideramos representativo: Devore, J. L. (2004) .

Es notorio como en el texto primero se trabaja con la FDP, mediante una visualización con his-

togramas pero estableciendo como dogma que un área bajo la FDP representa una probabilidad y que esta se calcula integrando la curva suave obtenida, es decir, a la FDA. Hasta aquí todo desligado de la FDA o de un histograma de frecuencias relativas acumuladas. El texto trae consigo unos “applets” que ayudan a visualizar diversos conceptos, en particular, el “applet” que apoya a los conceptos de la FDP y la FDA tiene la propiedad de sombrear las áreas en la FDP y mover el punto sobre la curva de la FDA al mismo tiempo, el cual se encuentra recalculado. Las curvas parecen pertenecer a la normal estándar pero no lo indica. El texto trae otro applet para visualizar sombreados y valores para cualquier normal (estándar o no) pero ya no incluye el correspondiente para la FDA. La definición de los percentiles que da no parece ayudar a diferenciar qué es el percentil y cómo se conecta con su probabilidad asociada, siendo que en una ojiva esto es más o menos claro. Si al estudiante se le dice que sólo tiene que igualar la FDA con la probabilidad correspondiente lo hace, pero la definición por sí misma es difícil de manejar: cuando en clase se le pide al estudiante calcular la mediana a partir de una FDP o FDA, no entiende como hacerlo. Primero, es un concepto que manejó en estadística descriptiva, segundo no conecta que es el 50avo percentil, y tercero no sabe como utilizar ambas funciones. En el caso de la normal estándar es notorio también que ya no reproduce la gráfica de la FDA de la normal estándar y que las tablas sólo representan la tabulación de la integral de la FDP, nos parece que es demasiado breve la conexión de estas tablas con la FDA a la que por cierto asigna otra simbología (que nos parece apropiado). Resulta extraño como al parecer se pretende recalcar la FDA de la normal estándar asignándole su propio símbolo, pero sin graficar ni abundar más que decir que lo que está en tablas es la FDA. Sobre la notación z_α , nos parece un desaprovechamiento del concepto de la FDA, el estudiante tiene que manejar esta notación casi como un concepto nuevo ya que a pesar de decirse que es un percentil, es una idea que no termina de quedarse en él. Incluso al tener el extracto de los valores principales de z_α : pues para qué pensar en recordar o asociar con la FDA, si ya tiene los valores que se van a estar utilizando. Aunque usa la FDA para tratar de conectar los percentiles de normales estándar y no estándar, creemos que debería abundar en ello ya que la visualización de este procedimiento puede quedar oculta para el estudiante. Es interesante como el texto siempre que hace ejemplos con normales maneja en el cálculo de probabilidades la notación $\Phi(z)$, lo cual nos parece una práctica sana que puede mover al estudiante a analizar el significado de tal notación.

6. Análisis cognitivo

Respecto a las dificultades que el estudiante tiene en la apropiación del concepto de la FDA, al parecer, no hay mucha literatura relacionada. Sin embargo, Batanero (2001) al ocuparse de

los estadísticos de orden, entre los cuales sobresalen los cuartiles, y señalando su importancia en estadística, menciona que hay dos niveles de dificultades: procedural y conceptual. Aunque la parte procedural comenta que tiene que ver con el hecho de manejar datos agrupados y no agrupados. En particular, distingue el caso de la mediana (50avo percentil), en la que señala dificultades surgidas de su propia definición y de la forma de calcularla mediante las gráficas de frecuencias acumuladas (ojivas). Indica la dificultad de los alumnos para obtener la mediana a partir de la gráfica de frecuencias acumuladas a no ser que se haya enfatizado explícitamente la forma de hacerlo. Para mayor referencia, en el documento aparecen las gráficas de dos acumuladas pero recorridas una con respecto a la otra y señala las dificultades de cálculo de la mediana. Generamos unas gráficas similares a las discutidas ahí en la figura 2.

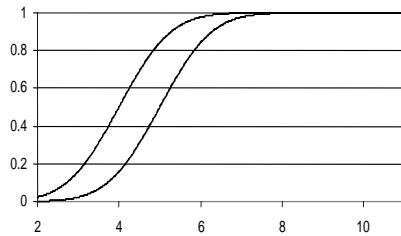


Figura 2: Dificultad para hallar la mediana a partir de las gráficas de frecuencias acumuladas

A este respecto, creemos que debería ser natural la obtención, mediante la gráfica, de la mediana y que esto también es indicio de que esta carencia posteriormente se vuelve un obstáculo de compren-

sión hacia la FDA. Es posible que si desde el análisis exploratorio de datos pudieran los alumnos construir tales gráficas y ubicar los percentiles ayudaría a vencer el obstáculo.

Batanero también cita a Estepa (1990), trabajo en el que se reporta las dificultades de los alumnos al interpretar las gráficas de frecuencias acumuladas para variables discretas. También cita a Barr, G. V., (1980), el cual menciona la falta de comprensión de los estudiantes sobre la mediana para estudiantes entre 17 y 21 años.

7. Conclusiones

El conocimiento de nuestras variables de estudio: epistemológico, didáctico y cognitivo, nos da pautas a posibles propuestas de enseñanza. Galton, aunque fue un hombre no muy versátil en matemáticas, desarrolló una gran capacidad de resolver problemas a través de esta idea de la FDA. A este respecto, se sugiere diseñar para los estudiantes actividades que consideren un análisis exploratorio, a partir del histograma de frecuencias relativas acumuladas o de la ojiva, en lugar del histograma de frecuencias relativas, con miras a una comprensión más fluida de la FDA y sus consecuentes ventajas de uso; junto con la inclusión temprana de los percentiles parece ser más natural y simple. El trabajo de un alumno debería ser similar a la actividad científica en cierto modo; buscar la solución de un problema, divagar, proponer, probar, establecer.

Referencias

Batanero, C., (2001). *Didáctica de la Estadística* Departamento de didáctica de la matemática Universidad de Granada. Pág. 91-93

Brousseau, G., (1986). Fundamentos y Métodos de la Didáctica de las Matemáticas, Universidad de Burdeos I. Traducción de Fondements et Méthodes de la didactique des mathématiques, *Recherches en Didactique des Mathématiques*, Vol. 7, n.2, pp.33- 115.

Barr, G.V., (1980). Some students ideas on the median and mode. *Teaching Statistics*, 2, 38-41.

Burr, I.W., (1942). Cumulative Frequency Functions, *The Annals of Mathematical Statistics*, Vol. 13, No. 2. pp. 215-232.

Devore, J.L, (2004). *Probability and Statistics for Engineering and the Sciences*, 6th Edition, Thomson-Brooks/Cole.

Estepa, A., (1990). Enseñanza de la Estadística basada en el uso de ordenadores. Un Estudio exploratorio. *Memorias de Tercer Ciclo. Universidad de Granada*: Departamento de didáctica de la matemática.

Galton, F., (1875). Statistics by Intercomparison, *Phil. Mag., Jan.*, 4th series, 49:33-46

Galton, F., (1880). Statististics of Mental Imagery, *Mind*, 5, 301-318.

Galton, F. (1886). Family Likeness in Stature, *Proceedings of the Royal society of London*, Vol 40. pp. 42-73

Hald, A., (1998). *A History of Mathematical Statistics. From 1750 to 1930*, Wiley Series in Probability and Statistics.

Meeker, W.Q., Escobar, L.A., (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons Inc.

Stigler, S.M., (1986). *The History of Statistics. The Measurement of Uncertainty before 1900*, The Belknap Press of Harvard University Press.

Algoritmos Genéticos en la Discriminación

Aurora Montano Rivas¹

Facultad de Estadística e Informática, Universidad Veracruzana

Mario Cantú Sifuentes

Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro

1. Introducción

La clasificación puede cubrir cualquier contexto en el que se toma una decisión o se realiza una predicción considerando la información disponible en ese momento. Cuando se sabe que existen clases y se desea conocer una regla para poder clasificar nuevas observaciones en las clases ya existentes se usa alguno de los tres métodos siguientes: el análisis discriminante, la regresión logística o los algoritmos genéticos.

El procedimiento del análisis discriminante permite obtener una regla de decisión tal que minimice la probabilidad de clasificación errónea. El análisis de regresión logística es una técnica utilizada para predecir un resultado binario de un juego de predictores que no son todos continuos. Por otra parte, los algoritmos genéticos son métodos adaptativos que se pueden usar para resolver problemas de búsqueda, se basan en el problema genético de los organismos vivos, tomando como entrada los ejemplares y retornan como salida aquellos que deben generar descendencia para la nueva generación.

Independientemente de los supuestos de estos métodos, se tiene un problema de optimización, donde la función es estimada a partir de una muestra de entrenamiento y es no diferenciable. Debido a que existen pocos estudios donde los algoritmos genéticos son una alternativa para encontrar mejores funciones discriminantes, se consideró conveniente presentar una nueva técnica de clasificación, la que se espera reporte una confiabilidad mayor que la de los métodos tradicionales. Para ello se propone el uso de remuestreo, la aplicación del análisis discriminante o la regresión logística y posteriormente los algoritmos genéticos.

¹ julmontano@uv.mx

2. Metodología

Análisis Discriminante.

La discriminación para el caso de dos grupos se basa en una muestra de entrenamiento, la cual está conformada de una variable dependiente dicotómica y varias variables independientes continuas.

El objetivo del análisis discriminante es encontrar una combinación de las variables que logren maximizar la separación de las poblaciones y minimizar la proporción de malas clasificaciones.

Si la población Π_j se distribuye $N_p(\mu_j, \Sigma)$, para $j = 1, 2$ y $\Sigma_1 = \Sigma_2 > 0$ la función discriminante para dos poblaciones es lineal.

La regla de decisión lineal para el caso de dos grupos se puede definir como:

$$\begin{cases} x \in \pi_1 & \text{si } a'x > 0 \\ x \in \pi_2 & \text{si } a'x \leq 0, \end{cases}$$

donde el vector a se encuentra al minimizar

$$f(a) = 1/2P(a'x > 0|x \in \pi_2) + 1/2P(a'x \leq 0|x \in \pi_1).$$

Por lo que se considera haber encontrado la función discriminante óptima para propósitos de clasificación.

Análisis de Regresión Logística.

Al igual que el análisis discriminante, la regresión logística es una técnica utilizada para predecir un resultado binario a través de un conjunto de variables independientes que no son todas continuas. Debido a esto, el error sigue una distribución binomial en vez de la normal, ocasionando la violación del supuesto de normalidad.

Uno de los objetivos de la regresión logística es predecir la probabilidad de que se produzca un suceso o acontecimiento definido como $Y = 1$ en función de los valores de las variables independientes.

Una alternativa para la discriminación y la clasificación es una especificación del parámetro de las probabilidades a posteriori $P(g_1|X_u)$ y $P(g_2|X_u)$, el cual puede ser estimado por medio de una muestra de entrenamiento.

La regla es: Se asigna la unidad u a la población g_1 si $P(g_1|X_u) > P(g_2|X_u)$. Por defecto asigna a los sujetos al grupo para el que su probabilidad de pertenencia es mayor que 0.50.

Algoritmos Genéticos.

Los Algoritmos Genéticos son algoritmos de optimización con búsqueda aleatoria sobre el espacio de posibles soluciones inspirados en el mecanismo de genética y selección natural.

Las soluciones posibles, también conocidas como cromosomas son un conjunto de modelos generados por la aplicación de alguna técnica (regresión lineal múltiple, análisis cluster, análisis discriminante, regresión logística, etc.) a la muestra de entrenamiento. Dicho conjunto es conocido como población inicial o población cero, siendo esta generalmente de tamaño de 20 a 100 cromosomas. Como el modelo es un cromosoma, entonces los parámetros son los genes.

Ya generada la población inicial se procede a la aplicación de un método de selección, el cual elegirá los progenitores de acuerdo a una probabilidad basada en la función objetivo, es decir, los individuos que tienen un valor de ajuste mejor tienen mayor oportunidad de ser seleccionados.

Método de Selección.

Existen muchos métodos para la selección de individuos, en este caso se usó el método de la ruleta, el cual considera que el individuo i es seleccionado y copiado a la nueva población si cumple con la siguiente condición: $C_{i-1} < U(0, 1) < C_i$; donde $C_i = \sum_{j=1}^i P_j$ y $U(0,1)$ es un número aleatorio.

Operadores Genéticos.

Los operadores genéticos proporcionan los mecanismos de búsqueda de los algoritmos genéticos, son usados para crear nuevas soluciones, basados en el conjunto de los mejores individuos elegidos previamente por un método de selección.

Básicamente son tres operadores que se usan en los algoritmos genéticos, pero en este caso solo se usaron dos, cruce aritmético y mutación; debido a que el cromosoma es de tipo continuo.

Cruce aritmético: Es definido como una combinación lineal de dos vectores. Si se tienen x_1 y x_2 vectores que serán cruzados, los resultados de las descendencias se generan como: $x'_1 = ax_1 + (1-a)x_2$ y $x'_2 = ax_2 + (1 - a)x_1$; donde a es un valor aleatorio entre 0 y 1.

Mutación: Se realiza después de un cruce, con la finalidad de prevenir posibles fallas de las soluciones en la población. En este operador un gene es seleccionado aleatoriamente de una cadena, el cual sufre una alteración y se produce una nueva solución única.

3. Técnica propuesta

Al aplicar el análisis discriminante y el análisis de regresión logística se comete el mínimo error de clasificación, por lo que surge la pregunta: Existirá una técnica que permita minimizar más ese error de clasificación?

La propuesta es generar por medio de remuestreo una población inicial de entre 20 y 100 combinaciones lineales discriminantes o logísticas con coeficientes de tipo continuo; a las que se les aplicará la técnica de Algoritmos Genéticos y posteriormente determinar cuál o cuáles de ellas son las que presentan el menor error de clasificación con respecto al reportado por el uso de técnicas tradicionales. Cuando esto ocurre decimos que encontramos la función discriminante óptima.

Para explicar el procedimiento de la técnica se usó como muestra de entrenamiento dos especies (las que presentan interacción) de la base de datos de Flores de Iris (Fisher, 1936); la cual está constituida por 100 individuos, 50 de cada especie y 4 variables independientes: $Y = \text{Especies}$ (Versicolor y Virginica), $X_1 = \text{Longitud del sépalo}$, $X_2 = \text{Ancho del sépalo}$, $X_3 = \text{Longitud del pétalo}$ y $X_4 = \text{Ancho del pétalo}$.

Como primer paso, se aplica el análisis discriminante de Fisher, a la base completa para obtener el modelo original y observar el error promedio, el cual se toma como valor de tolerancia; ya que al aplicar la técnica propuesta se espera obtener una o varias funciones con error promedio menor que el del método tradicional.

Con la idea de validar la técnica, de la base de datos completa se generan dos muestras de tamaño 50, de manera que cada una de éstas contenga 25 individuos de cada especie; una de estas muestras se usa para generar los modelos y la otra para evaluar aquellos que cumplieron con la tolerancia.

Estructura del programa.

Para llevar a cabo el proceso de la Técnica Propuesta, se realizó un programa en el paquete Estadístico S-PLus (Ver.6.1); en el cual la muestra de entrenamiento es captura y la variable dependiente

debe ser declarada de tipo factor.

El programa realiza como primer paso el análisis discriminante de Fisher o la regresión logística, según la opción elegida; ya que cualquiera de las dos técnicas proporciona la función discriminante. Posteriormente se realiza el remuestreo, ejecutando en cada muestra el análisis elegido para formar la población inicial de funciones discriminantes, después los vectores de X's se sustituyen en cada modelo para evaluar cada uno de éstos contabilizando el número de individuos mal clasificados en ambos grupos, y así determinar la proporción de error de clasificación.

Ya evaluados los modelos, a cada uno se le asigna un valor de fitness (probabilidad de mejor ajuste), el cual se calcula cómo sigue:

1.- Obtener para cada función la proporción de individuos mal clasificados en cada grupo, se suman y se dividen entre dos, obteniéndose el error promedio de clasificación.

2.- Ahora se realiza la suma generándose el error promedio total del problema.

3.- El fitness es igual al error promedio de clasificación entre el error promedio total del problema.

Obtenido el fitness de cada modelo se procede a calcular el fitness acumulado. Para la selección del mejor modelo, se usa el método de la ruleta, aplicando después el operador de cruce aritmético, el cual usa dos modelos para generar dos nuevos.

Posteriormente, al conjunto de modelos reportados por el cruce se les aplica mutación normal y después mutación uniforme, con la finalidad de observar cual de estos dos aporta mejores soluciones. La primera consiste en la selección de una variable aleatoria, de la cual se obtiene la media y la varianza para generar un valor aleatorio de una distribución normal. En la segunda se selecciona la variable de forma aleatoria y se localizan los valores mínimo y máximo y en ese rango se genera un valor aleatorio con distribución uniforme.

El valor aleatorio generado por cualquiera de estas dos distribuciones se sustituye en la columna seleccionada.

Los modelos del cruce como los obtenidos por las mutaciones son evaluados con los datos originales que generaron la población inicial. Y es aquí donde obtenemos el error de discriminación y se procede con la validación.

4. Resultados

Como primera etapa se aplicó el análisis discriminante de Fisher para la base completa, obteniéndose el modelo original (Cuadro 1) con error promedio de 0.03.

Cuadro 1. Función discriminante obtenida por el método de Fisher.

Long	Ancho	Long	Ancho	Térmo	Error	Error	Error
Sépal	Sépal	Pétalo	Pétalo	Const.	Grup1	Grup2	Promedio
3.556303	5.578621	-6.970128	-12.38604	16.66309	0.04	0.02	0.03

Como se espera encontrar una o varias funciones con error promedio menor que 0.03, éste es tomado como la tolerancia en el programa.

Se generaron 20 muestras de las cuales se obtuvieron las funciones discriminantes que formaron la población inicial, se evalúan y se aplica el método de la ruleta para la selección de las mejores, se emplea el cruce aritmético y las mutaciones normal y uniforme, obteniéndose lo siguiente (ver Cuadro 2 y 3):

Cuadro 2. Funciones generadas por Mutación (Distribución Normal).

Long	Ancho	Long	Ancho	Térn	Error
Sépal	Sépal	Pétalo	Pétalo	Const.	Promedio
4.389463	4.311353	-7.96471	-12.08190	18.80822	0.02
3.937811	7.306391	-7.96471	-13.45358	16.23628	0.02

Cuadro 3. Aplicación de Mutación (Distribución Uniforme).

Long	Ancho	Long	Ancho	Térn	Error
Sépal	Sépal	Pétalo	Pétalo	Const.	Promedio
4.367464	6.900467	-8.984654	-14.31159	21.73694	0.02
4.367464	5.653616	-8.683885	-11.99341	18.91315	0.02

En este caso se tienen dos modelos por cada mutación, los cuales discriminan mejor que la función proporcionada por el método tradicional.

Ahora, al aplicar la regresión logística a los datos originales se encontró que el error promedio es muy alto, por lo que difícilmente un modelo de regresión logística logrará cumplir con la tolerancia

de 0.03 y mucho menos reporte una función con error igual a 0.02 (ver Cuadro 4); de manera que se decidió probar con una tolerancia de $e=0.70$ generando $n=20$ remuestreos.

Cuadro 4. Función de regresión logística de la base de datos original .

Long	Ancho	Long	Ancho	Térn	Error	Error	Error
Sépalo	Sépalo	Pétalo	Pétalo	Const.	Grup1	Grup2	Promedio
-2.46521	-6.68065	9.42901	18.2855	-42.63567	0.98	0.98	0.98

Aplicando la técnica propuesta se tiene que por mutación normal 19 funciones reportan un error menor que la tolerancia, en 15 de ellas disminuyó hasta 0.50. Para el caso de mutación uniforme todos los modelos reportan un error de 0.50, lo que implica que sólo un 50 % de los individuos logran ser clasificados adecuadamente.

Validación usando el método de datos propuestos para análisis discriminante.

Como se mencionó anteriormente, la base completa es dividida en dos muestras del mismo tamaño, la muestra 1 se usa para generar los modelos y la muestra 2 para evaluar aquellos que cumplieron con la tolerancia.

La función discriminante de los datos de la muestra 1 es:

Cuadro 5. Función discriminante de Fisher de la muestra 1.

Long	Ancho	Long	Ancho	Térn	Error	Error	Error
Sépalo	Sépalo	Pétalo	Pétalo	Const.	Grup1	Grup2	Promedio
3.91284	5.190329	-7.632298	-12.72481	19.52356	0.04	0	0.02

En este caso lo ideal es encontrar una función con error promedio de discriminación menor o igual que 0.02. Aplicando el remuestreo se encontró un modelo con error promedio de cero. (ver Cuadro 6)

Cuadro 6. Solución inicial con error cero.

Long	Ancho	Long	Ancho	Térn.
Sépalo	Sépalo	Pétalo	Pétalo	Const.
2.556413	3.513466	-7.55458	-11.84034	31.50987

Por lo que se encontró la función óptima, lo que nos llevaría a detener el proceso, pero como la idea es validar se continua la ejecución del programa.

Al llegar a la etapa de mutar las variables se obtuvieron dos modelos para cada caso con errores promedio de discriminación de 0.00 y 0.02. (ver Cuadro 7 y 8)

Cuadro 7. Soluciones por Mutación por distribución normal.

N de Función	Long Sépalo	Ancho Sépalo	Long Pétalo	Ancho Pétalo	Térn Const.	Error Promedio
1	2.7516019	4.954462	-7.793511	-14.16819	31.99538	0.00
2	4.8587625	4.954462	-9.235785	-13.31772	22.73018	0.02

Cuadro 8. Soluciones por Mutación por distribución uniforme.

N de Función	Long Sépalo	Ancho Sépalo	Long Pétalo	Ancho Pétalo	Térn Const.	Error Promedio
1	5.324897	5.386912	-9.576163	-13.61683	20.95264	0.02
2	4.858762	5.386912	-9.235785	-13.31772	22.73018	0.00

Los modelos obtenidos por mutación normal se evalúan en la segunda muestra y reportan errores promedio de clasificación de 0.08 y 0.04 respectivamente; mientras que los obtenidos por mutación uniforme presentan un error de clasificación similar (0.0016).

Por otra parte, la validación de modelos generados por regresión logística no funcionó, ya que los errores de clasificación fueron muy grandes.

5. Conclusiones

Los métodos de clasificación comúnmente utilizados son el discriminante y el de regresión logística, y se recomienda emplearlos cuando cumplen o no respectivamente el supuesto de normalidad. Recientemente se aplican los métodos de algoritmos genéticos y al combinarlos con el análisis discriminante o con regresión logística se obtienen modelos de clasificación más eficientes. Para lo anterior se realizó el programa Dislog en el paquete estadístico Splus, en el cual se ejecutó una base de datos de Flores de Iris con 100 observaciones, cuatro variables independientes y una dependiente o de agrupación. Se obtuvieron los modelos de clasificación y sus respectivos errores de tolerancia y en relación a ella se concluye que la técnica de algoritmos genéticos aplicada a las funciones proporcionadas por el análisis discriminante, genera los mejores modelos discriminantes con el mínimo error de clasificación; es decir, los modelos obtenidos son más eficientes que los reportados por el método de Fisher.

Referencias

- Alfaro, Esteban, Gámez, Matias, y García, Noelia (2002). Una revisión de los métodos de clasificación aplicables a la economía. *Artículo en internet*. Universidad de Castilla-La Mancha, España.
- Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of finance*, Vol. 23 (september) pp. 589-609.
- Back, B., Laitinen, T., Sere K. y Wezel, M. (1996). Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms. *Articulo en internet*
- Banzhaf, W. and Reeves, C. (1999). *Fundations of genetic algorithms*, Morgan Kaufmann Publishers, Inc., San Francisco, California.
- Beaver, W. (1966). Financial ratios as predictors of failures. In empirical research in accounting, selected studies, 1966 in supplement to the *Journal of accounting research*, Vol 5. pp. 71- 111.
- Dallas, E. J. (1998). *Métodos multivariados aplicados a los análisis de datos*; International Thomson Editores. Madrid, España.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7, 179-188.
- Hand, D. J. (1997). *Construction and assessment of classification rules*. John Wiley & Sons. Chichester.
- Holland, J. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, Michigan.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied logistic regression*; John Wiley & Sons. Inc. New York.
- Houck, C. R. and Joines, J. A. (1998). A genetic algorithm for function optimization: A Matlab Implementation. *Articulo en internet*
- Huberty, C. J. (1994). *Applied discriminant analysis*; John Wiley and Sons. Inc. New York.

Estratificación Optima para el Indice de Desarrollo Humano

Federico R. Muller Rodríguez

Facultad de Economía de la Universidad Autónoma de Coahuila

Félix de J. Sánchez Pérez¹

Centro de Investigación en Matemáticas Aplicadas de la Universidad Autónoma de Coahuila

Emilio Padrón Corral²

Centro de Investigación en Matemáticas Aplicadas de la Universidad Autónoma de Coahuila

1. Introducción

Durante décadas el objetivo principal de las economías del mundo fue el de crecer, así la idea principal que predominó fue la que anunciaba que un crecimiento, basado en la dotación cada vez más de bienes y servicios, traería consigo un aumento en la calidad de vida de la población en general; esta idea ha sido refutada en los últimos años, debido al auge que a venido cobrando el concepto de desarrollo humano, el cual, no sólo implica un crecimiento sino un desarrollo que permita una mejor calidad de vida para los habitantes de cualquier país.

Desde la perspectiva de su medición, el desarrollo económico es imprescindible para la elaboración de políticas públicas y la evolución de las mismas; por ello, los organismos internacionales consideraron de suma importancia construir índices de desarrollo con el fin de medir los crecientes márgenes de desigualdad económica que se presentan entre las naciones.

Uno de los indicadores más conocidos y popular entre los gobiernos es el Desarrollo Humano (IDH), que es una medición que incorpora, además del ingreso per cápita, la esperanza de vida de la población y los aspectos educativos de la misma; esto complementa a criterio economista, pues considera el ingreso sólo como un medio para alcanzar el desarrollo y no un fin en si mismo, Lohr (2000).

¹fel1925@yahoo.com

²epadron@cima.uadec.mx

2. Objetivo

La investigación aborda la temática económica desde la perspectiva de los indicadores de desarrollo humano y sus fronteras. Propone una metodología alterna para la definición de los límites de los estratos que clasifican el nivel de desarrollo de los países de acuerdo con el Programa de Naciones Unidas para el Desarrollo (PNUD).

3. Indices de desarrollo humano

Recientemente, los organismos internacionales (Banco Mundial, Fondo Monetario Internacional, Naciones Unidas, entre otros), incluyeron en sus acervos estadísticos, las mediciones del desarrollo, constreñidos por los crecientes márgenes de desigualdad económica que se presentaban entre las naciones del planeta.

En el caso de México el IDH ha variado ligeramente, en 1995 ocupaba el quadragésimo noveno lugar en el *ranking mundial*, mientras que en el año 2000 descendió al quincuagésimo. Y lo más grave, al interior del país, la desigualdad en los niveles de bienestar, entre las Entidades Federativas se agudiza, como lo muestra el cuadro 1.

Ante las fluctuaciones que pueden ubicar al país en diferentes estratos de desarrollo, es recomendable definir instrumentos estadísticos-matemáticos apropiados que delimiten los estratos: alto desarrollo humano; desarrollo humano mediano; y desarrollo humano bajo.

Cuadro 1.- Índice de Desarrollo Humano y sus componentes por Estados en el país

Posición Según IDH	Entidad	IDH	Índice de Esperanza de vida	Índice Escolar	Índice PIB Per cápita
1	D. F.	0.8913	0.8700	0.8975	0.9063
2	NUEVO LEÓN	0.8534	0.8633	0.8515	0.8454
3	B. C. N.	0.8401	0.8550	0.8604	0.8050
4	CHIHUAHUA	0.8355	0.8467	0.8478	0.8120
5	COAHUILA	0.8329	0.8533	0.8568	0.7885
6	B. C. S.	0.8323	0.8550	0.8567	0.7851
28	MICHOACAN	0.7516	0.8300	0.7772	0.6477
29	VERACRUZ	0.7479	0.8167	0.7760	0.6509
30	GUERRERO	0.7312	0.8050	0.7427	0.6459
31	OAXACA	0.7135	0.7917	0.7456	0.6032
32	CHIAPAS	0.7032	0.7900	0.7240	0.5957
	NACIONAL	0.8014	0.8383	0.8181	0.7479
Fuente: Informe sobre el desarrollo humano 2000. PNUD					

4. Metodología de Dalenius Hodges

En la consideración de la formación de k estratos puede utilizarse cualquier conocimiento previo para producir semejanza dentro de un mismo estrato. La situación ideal es aquella en la que tenemos disponible la distribución de X ; y se plantea la metodología de cómo se deben determinar los límites entre estratos de tamaño y variabilidad diferente.

Así, si se considera que $N \rightarrow \infty$ es suficiente minimizar $\frac{1}{n} \left(\sum_{i=1}^k P_i S_i \right)^2$ donde $P_i = \frac{N_i}{N}$.

Si los estratos son números estrechos, $f(x)$ deberá ser aproximadamente constante (rectangular) dentro de un estrato dado. Por lo tanto

$$P_i = \int_{X_{i-1}}^{X_i} f(t) dt = f_{X_i}(X_i - X_{i-1})$$

Por consiguiente, $\sum_{i=1}^k P_i S_i = \sum_{i=1}^k f_{X_i}(X_i - X_{i-1}) \frac{(X_i - X_{i-1})^2}{\sqrt{12}} = \sum_{i=1}^k \frac{[\sqrt{f_{X_i}}(X_i - X_{i-1})]^2}{\sqrt{12}}$

Sea $A_i = \int_{X_{i-1}}^{X_i} \sqrt{f(x)} dx$ entonces $A_i \cong \sqrt{f(x_i)} (X_i - X_{i-1})$

Claramente, $V(\hat{\bar{X}_e})_{Ney} = \frac{1}{n} \left(\sum_{i=1}^k P_i S_i \right)^2 = \frac{1}{12n} \left(\sum_{i=1}^k A_i^2 \right)^2$

Minimizar $V(\hat{\bar{X}_e})_{Ney}$ es equivalente hacerlo con $\sum_{i=1}^k A_i^2$. Es mínimo cuando las A_i tienen valores dados por:

$$A_i = \frac{L}{k}, \quad i = 1, 2, \dots, k$$

donde $\sum_{i=1}^k A_i = L$

Entonces los límites óptimos bajo la asignación Neyman son obtenidos tomando intervalos iguales de la acumulada de $\sqrt{f(x)}$.

5. Resultados

Para la estratificación óptima de las naciones, basada en el IDH, se buscó primero qué distribución estadística se ajustaba a los datos (índice de desarrollo relativo al género, que incorpora las mismas variables que el IDH pero ajusta a los resultados para captar las desigualdades de género), siendo una beta, como constata en la estimación de los parámetros de la función bajo el método de Máxima Verosimilitud. Así,

$$f(x) = \frac{\Gamma(10,69)}{\Gamma(8,32)\Gamma(2,37)} x^{7,32} (1-x)^{1,37} \quad 0 \leq x \leq 1$$

A continuación se aplicó la técnica de Dalenius y Hodges. Así, se procedió a determinar la integral

$$\int_0^1 \sqrt{f(x)} dx = 9,01753$$

En seguida se definieron los límites de los estratos bajo la solución del siguiente planteamiento:

$$\int_0^1 \sqrt{f(x)} dx = 3,0058$$

Encontrándose los valores de los límites de los estratos como

Estratos	Límites
Desarrollo Bajo	0-0.60
Desarrollo Medio	0.61-0.66
Desarrollo Alto	0.67-1.00

6. Conclusiones

Los límites calculados de los estratos o corresponden a los que plantea el Programa Naciones Unidas sobre el IDH, cuya estratificación es la siguiente:

Estrato	Límites
Índice de desarrollo humano bajo	0-0.499
Índice de desarrollo humano mediano	0.500-0.799
Índice de desarrollo humano alto	0.800-1.00

Las diferencias encontradas en los límites de los estratos que miden el grado de desarrollo de un país, implican no solamente distinciones en la elección del mejor método estadístico, sino que trasciende hasta el bienestar de la sociedad.

La asesoría, la asistencia técnica y financiera de los organismos multilaterales es susceptible de realizarse ó modificarse según el grado de desarrollo que presente el país. De ahí la importancia de jerarquizar con precisión los niveles reales de desarrollo humano. Aunque el concepto de desarrollo es muy rico y complejo por contemplar aspectos culturales e ideológicos que resulta difícil captarlos y confinarlos en una serie de indicadores estadísticos, los organismos internacionales requieren de este tipo de información (IDH) que puede medirse y evaluarse.

Las políticas de asistencia que destinan recursos de las agencias internacionales necesitan de diagnósticos claros que describan las necesidades sentidas de la población y una forma de lograrlo es mediante la construcción de identificadores fieles a la teoría estadística.

Referencias

Sharon L. Lohr (2000). *Muestreo. Diseño y Análisis*. México: Thomson Editores, S. A.

Una generalización de los modelos frailty

Luis E. Nieto-Barajas¹

División Académica de Estadística, ITAM

Stephen G. Walker²

University of Kent, UK

1. Introducción

En este trabajo introducimos un modelo semiparamétrico Bayesiano para datos de supervivencia bivariados y multivariados. Consideremos el caso bivariado (T_1, T_2) , donde T_i , $i = 1, 2$ son tiempos univariados de supervivencia. Para alcanzar nuestro objetivo, necesitamos construir una función de densidad bivariada aleatoria $f(t_1, t_2)$, lo cuál se hará a través de la construcción de un proceso estocástico conveniente. La ley del proceso será nuestra distribución inicial sobre f . Como consecuencia de ésto, tendremos un par de funciones $f_1(t_1)$ y $f_2(t_2)$ aleatorias, las cuales serán modelos Bayesianos no paramétricos conocidos. Aquí, por ejemplo, $f_1(t_1) = \int f(t_1, t_2) dt_2$. La parte difícil, como lo es en el caso de modelos de supervivencia paramétricos, es el crear una distribución conjunta con propiedades de dependencia satisfactorias.

Los modelos de supervivencia paramétricos recientes, incluyendo el conocido modelo frailty, están basados en representaciones de mezcla. Crowder (1989) propuso un modelo multivariado de supervivencia basado en mezclas de familias Weibull con distribuciones de mezcla gamma y estable. Por su parte Hougaard (1986), también usó como distribución de mezcla una distribución estable. Walker y Stephens (1999) también consideraron modelos de supervivencia multivariados paramétricos basados en mezclas de Weibulls con distribución de mezcla lognormal.

La estructura del artículo es la siguiente: En la Sección 2 describimos la forma general de los modelos de mezcla y ponemos nuestro modelo como un miembro de esta clase. La Sección 3 trata de las distribuciones iniciales y finales, y en la Sección 4 ilustramos el uso de nuestro modelo con un ejemplo.

¹lnieto@itam.mx

²S.G.Walker@kent.ac.uk

2. Modelos de mezcla

Consideremos primero el caso univariado. Sea T una v.a. definida en $[0, \infty)$ y suponga que su correspondiente función de densidad $f(t)$ tiene una representación de mezcla de la forma

$$f(t) = \int_{\Omega} f(t|\omega)m(\omega)d\omega, \quad (1)$$

donde Ω es el soporte de $m(\omega)$, y para toda $\omega \in \Omega$, $f(t|\omega)$ es una función de densidad en $[0, \infty)$. No es difícil probar que, por ejemplo, las densidades gamma, Weibull y Gompertz se pueden representar como en (1).

En general, sean $h(t)$ y $H(t)$ las funciones de riesgo y riesgo acumulado, respectivamente, de una v.a. T . La función de densidad $f(t)$ se puede escribir como una mezcla (1) usando h y H : Si $f(t|\omega) = h(t)/\omega I\{\omega > H(t)\}$ y $m(\omega) = \text{Ga}(\omega|2, 1)$, entonces claramente $f(t) = h(t) \int_{\omega>H(t)} \omega^{-1} \omega e^{-\omega} d\omega = h(t) \exp\{H(t)\}$.

Consideremos ahora el caso bivariado. Sea $T = (T_1, T_2)$ un vector aleatorio definido en $[0, \infty)^2$, con función de densidad conjunta $f(t_1, t_2)$. Generalizando la idea (1) al caso bivariado tenemos,

$$f(t_1, t_2) = \int_{\Omega} \int_{\Omega} f(t_1, t_2|\omega_1, \omega_2)m(\omega_1, \omega_2)d\omega_1 d\omega_2. \quad (2)$$

Walker y Stephens (1999) propusieron una nueva familia paramétrica multivariada usando (2) basándose en distribuciones de mezcla normales. En un esquema general, nuestro objetivo es asegurar que las funciones de densidad marginales pertenecen a determinadas familias, y para ello usaremos $m(\omega_1, \omega_2)$ para crear la dependencia.

Para lograr nuestro objetivo, suponga que T_j es una variable aleatoria definida en $[0, \infty)$ con función de densidad $f_j(t)$, función de riesgo $h_j(t)$ y de riesgo acumulado $H_j(t)$. Del modelo univariado, sabemos que cada una de las densidades marginales $f_j(t)$ se pueden expresar como (1), i.e., $f_j(t|\omega_j) = h_j(t)/\omega_j I\{\omega_j > H_j(t)\}$ y $\omega_j \sim \text{Ga}(2, 1)$. Entonces, para la construcción bivariada (2) tomamos $f(t_1, t_2|\omega_1, \omega_2) = f_1(t_1|\omega_1)f_2(t_2|\omega_2)$ y (ω_1, ω_2) con distribución gamma bivariada con marginales $\text{Ga}(2, 1)$. Esta construcción garantiza que T_1 y T_2 tengan distribuciones marginales con funciones de riesgo $h_1(t)$ y $h_2(t)$, respectivamente. Más aún, la estructura de dependencia entre T_1 y T_2 se logra a través de una dependencia en $\omega = (\omega_1, \omega_2)$.

El modelo frailty usual de Clayton (1978) tiene una estructura parecida a la nuestra en el sentido de que admite una representación de la forma (2) con $f(t_1, t_2 | \omega_1, \omega_2) = f_1(t_1 | \omega_1) f_2(t_2 | \omega_2)$ pero con $\omega_1 = \omega_2 = \omega$ y $f_j(t | \omega) = \omega h_j(t) \exp\{-\omega H_j(t)\}$. En este caso, las distribuciones marginales tienen, en general, funciones de riesgo distintas a $h_j(t)$.

Una vez definido el modelo, es necesario que especifiquemos la distribución gamma bivariada. De acuerdo con Johnson y Kotz (1972), existen muchas propuestas. Una de las más sencillas tiene la siguiente construcción. Tomemos $y_k \sim \text{Ga}(\gamma_k, 1)$, $k = 0, 1, 2$ independientes, y definamos $\omega_j = y_0 + y_j$, para $j = 1, 2$. Entonces, (ω_1, ω_2) tiene una distribución gamma bivariada con parámetros $(\gamma_0, \gamma_1, \gamma_2)$, denotada por $(\omega_1, \omega_2) \sim \text{BGa}(\gamma_0, \gamma_1, \gamma_2)$. Marginalmente, $\omega_j \sim \text{Ga}(\gamma_0 + \gamma_j, 1)$ y $\text{Corr}(\omega_1, \omega_2) = \gamma_0 \{(\gamma_0 + \gamma_1)(\gamma_0 + \gamma_2)\}^{-1/2} \geq 0$. En la práctica una correlación positiva es suficiente, sin embargo, si se requiere una correlación negativa, se podría utilizar otra distribución gamma bivariada (ver, por ejemplo, Johnson and Kotz, 1972).

Para lograr que $\omega_j \sim \text{Ga}(2, 1)$, $j = 1, 2$, tomamos $\omega = (\omega_1, \omega_2) \sim \text{BGa}(\gamma, 2 - \gamma, 2 - \gamma)$, donde $\gamma \in (0, 2)$ es el parámetro que controla la correlación, de hecho $\rho = \text{Corr}(\omega_1, \omega_2) = \gamma/2$. En particular, si $\gamma \rightarrow 2$ obtenemos el modelo de Clayton (1978) con una distribución $\text{Ga}(2, 1)$ para el frailty común.

3. Distribuciones iniciales y finales

En la literatura existen varias distribuciones iniciales no paramétricas para funciones de riesgo (ver por ejemplo, Nieto-Barajas y Walker 2004, y las referencias que ahí se mencionan). Para propósitos de este artículo usaremos el proceso gamma correlacionado de Nieto-Barajas y Walker (2002). Este proceso consiste en lo siguiente:

Consideremos una partición de la escala del tiempo, digamos, $0 = \tau_0 < \tau_1 < \dots$. Sea λ_k una función de riesgo constante en el intervalo $(\tau_{k-1}, \tau_k]$. Las $\{\lambda_k\}$ siguen un proceso de Markov definido a través de un proceso latente $\{u_k\}$. El proceso inicia con $\lambda_1 \sim \text{Ga}(\alpha_1, \beta_1)$, luego, para $k = 1, 2, \dots$ tomamos $u_k | \lambda_k \sim \text{Po}(c_k \lambda_k)$ y $\lambda_{k+1} | u_k \sim \text{Ga}(\alpha_{k+1} + u_k, \beta_{k+1} + c_k)$. Si α_k y β_k son constantes para todo k , entonces el proceso $\{\lambda_k\}$ es estrictamente estacionario con $\lambda_k \sim \text{Ga}(\alpha_1, \beta_1)$ y estructura de correlación $\text{Corr}(\lambda_k, \lambda_{k+1}) = c_k / (\beta_1 + c_k)$. Finalmente, la inicial para $h_j(t)$ está dada por, $h_j(t) = \sum_k \lambda_{jk} I\{\tau_{k-1} < t \leq \tau_k\}$.

Sea $T = \{(T_{11}, T_{12}), (T_{21}, T_{22}), \dots, (T_{n1}, T_{n2})\}$ una muestra de observaciones bivariadas de $f(t_1, t_2)$. Para cada par (T_{i1}, T_{i2}) necesitaremos variables $(\omega_{i1}, \omega_{i2}, y_i)$, $i = 1, \dots, n$, donde y_i es una variable latente que nos ayudará a simplificar el cálculo de la distribución posterior.

La distribución posterior se obtendrá con la ayuda de un muestreador de Gibbs, por lo que será suficiente con conocer las distribuciones finales condicionales completas. Para $j = 1, 2$, $i = 1, \dots, n$ y $k = 1, 2, \dots$, tenemos

- $m(\omega_{ij}|t, y, h) \propto \frac{(\omega_{ij}-y_i)^{1-\gamma}}{\omega_{ij}} e^{-\omega_{ij}} I[\omega_{ij} > \max\{y_i, H_j(t_{ij})\}]$
- $g(y_i|t, \omega, h) \propto \left\{ \frac{y_i}{(\omega_{i1}-y_i)(\omega_{i2}-y_i)} \right\}^{\gamma-1} e^{y_i} I\{0 < y_i < \min(\omega_{i1}, \omega_{i2})\}$
- $\pi(\lambda_{jk}|t, \omega, y, u, \lambda_{-jk}) \propto \lambda_{jk}^{\alpha_{jk}+u_{jk-1}+u_{jk}+n_{jk}-1} e^{-(\beta_{jk}+c_{jk-1}+c_{jk})\lambda_{jk}} \prod_{i=1}^n I\{\omega_{ij} > H_j(t_{ij})\}$
donde $n_{jk} = \sum_{i=1}^n I(\tau_{k-1} < t_{ij} \leq \tau_k)$ y
 $H_j(t) = \sum_{k=1}^{k^*-1} \lambda_{jk}(\tau_k - \tau_{k-1}) + \lambda_{jk^*}(t - \tau_{k^*-1})$, cuando $t \in (\tau_{k^*-1}, \tau_{k^*}]$
- $\pi(u_{jk}|t, w, y, \lambda) \propto \frac{\{c_{jk}(c_{jk}+\beta_{jk+1})\lambda_{jk}\lambda_{j,k+1}\}^{u_{jk}}}{\Gamma(u_{jk}+1)\Gamma(\alpha_{jk+1}+u_{jk})}$, $u_{jk} = 0, 1, \dots$

El parámetro γ juega un papel muy importante en la determinación de la correlación inicial entre (T_1, T_2) . Es posible asignar una distribución inicial $\pi(\gamma)$, obteniéndose

- $\pi(\gamma|t, \omega, y, h) \propto \frac{1}{\Gamma^n(\gamma)\Gamma^{2n}(2-\gamma)} \prod_{i=1}^n \left\{ \frac{y_i}{(\omega_{i1}-y_i)(\omega_{i2}-y_i)} \right\}^\gamma \pi(\gamma)$

Estas distribuciones finales condicionales se basan únicamente en una muestra de observaciones exactas. En la presencia de observaciones censuradas, éstas pueden ser incorporadas al análisis considerándolas como observaciones faltantes e incorporando la distribución predictiva en el muestreador de Gibbs. Es posible también incluir covariables al análisis, usando el modelo de riesgos proporcionales de Cox(1972).

4. Ejemplo Numérico

La base de datos se obtuvo de McGilchrist and Aisbett (1991) y consiste de observaciones bivariadas $\{(T_{i1}, T_{i2})\}_{i=1}^n$ de tiempos de infección desde el momento de la inserción de un catéter. Una vez

que el catéter se infectaba, era removido y reemplazado por otro. Estos datos se recolectaron para $n = 38$ pacientes de diálisis.

Para definir la inicial tomamos la partición $\tau_0 = 0$ y $\tau_k = \tau_{k-1} + 10$, $k = 1, \dots, 57$, para cubrir todo el rango de observaciones. Tomamos $\alpha_{jk} = \beta_{jk} = 0.0001$ y $c_{jk} = 1000$, $j = 1, 2$ y $\forall k$. Para α_{jk} y β_{jk} se tomaron valores pequeños constantes para tener relativamente poca información en las condiciones marginales iniciales de cada λ_{jk} . El valor c_{jk} se eligió grande para introducir una correlación alta entre λ 's adyacentes permitiendo que el proceso tenga trayectorias suaves. Adicionalmente, se consideró una distribución uniforme $\gamma \sim \text{Un}(0, 2)$. El muestreador de Gibbs se corrió por 50,000 iteraciones con un período de calentamiento de 5,000.

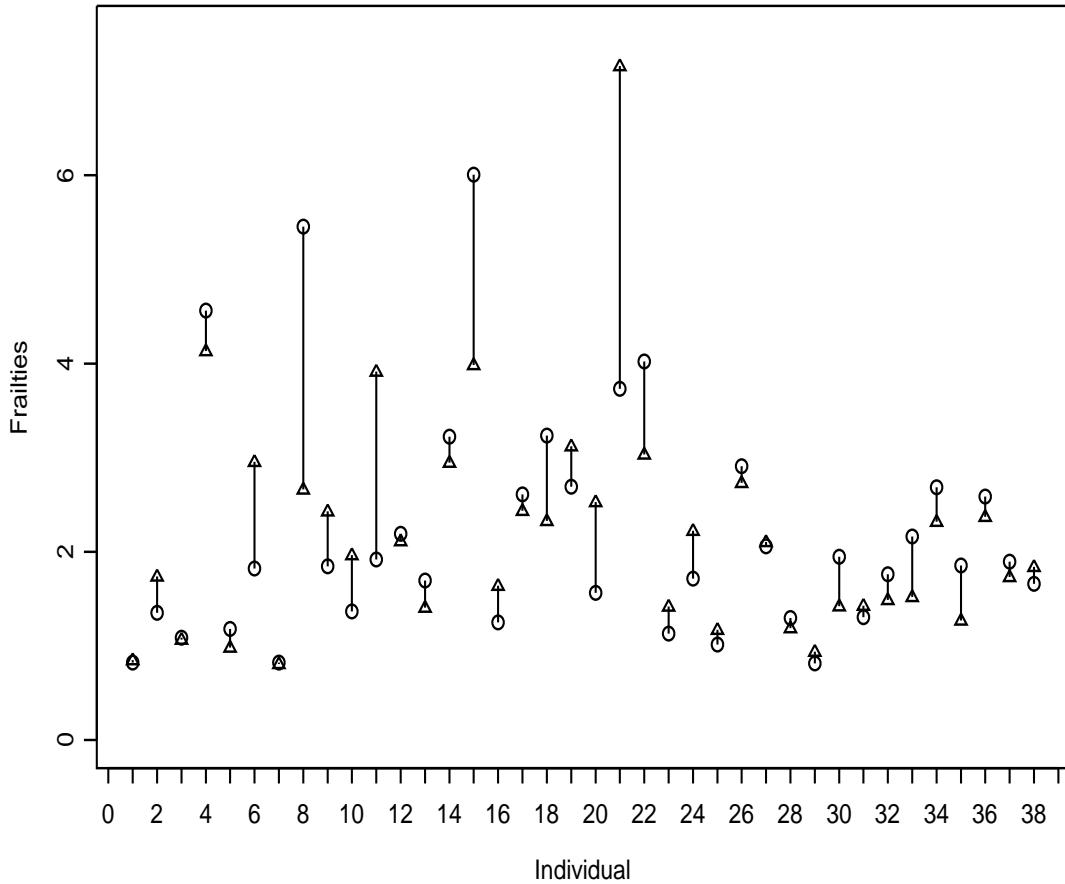


Figura 1: Estimadores posteriores de los frailties. (○) valor estimado de ω_{i1} y (△) valor estimado de ω_{i2} .

Las medias estimadas de los frailty ω_1 y ω_2 , para todos los individuos, son básicamente las mismas, i.e., 2.19 y 2.20 respectivamente. Esto podría sugerir que no hay necesidad de que los frailties sean diferentes, sin embargo, viendo los frailties individuales para cada paciente (ver Figura 1), realmente sí se observan diferencias. Los círculos y los triángulos corresponden a la media posterior de ω_{i1} y ω_{i2} , respectivamente, para $i = 1, \dots, 38$. La longitud de la línea continua que une a las medias de los frailties de cada individuo representa la diferencia entre ambos frailties. Para los individuos (1, 3, 7, 27) los estimadores de los frailties son casi iguales, pero para los individuos (8, 11, 15, 21) existe una diferencia grande entre los frailties. Más aún, para el individuo 21, $P(\omega_{21,1} < \omega_{21,2} | \text{datos}) = 0.93$ lo cual apoya fuertemente el hecho de que los frailties deben de ser diferentes.

Referencias

- Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.
- Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-202.
- Crowder, M.J. (1989). A multivariate distribution with Weibull connections. *Journal of the Royal Statistical Society, Series B* **51**, 93–107.
- Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika* **73**, 671–678.
- Johnson, N.L. and Kotz, S. (1972). *Distributions in statistics: Continuous multivariate distributions*. Wiley, New York.
- McGilchrist, C.A. and Aisbett, C.W. (1991). Regression with frailty in survival analysis. *Biometrics* **47**, 461–466.
- Nieto-Barajas, L.E. and Walker, S.G. (2002). Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics* **29**, 413–424.
- Nieto-Barajas, L.E. and Walker, S.G. (2004). Bayesian nonparametric survival analysis via Lvy driven Markov processes. *Statistica Sinica* **14**, 1127–1146.

Walker, S.G. and Stephens, D.A. (1999). A multivariate family of distributions on $(0, \infty)^p$. *Biometrika* **86**, 703–709.

La Estadística Multivariada como Análisis de Datos de Guayule

Emilio Padrón Corral¹

Universidad Autónoma de Coahuila

Ignacio Méndez Ramírez²

Universidad Nacional Autónoma de México

Félix de Jesús Sánchez Pérez³

Universidad Autónoma de Coahuila

Emilio Olivares Sáenz

Universidad Autónoma de Nuevo León

1. Introducción

Los modelos de ecuaciones estructurales son técnicas principalmente confirmatorias dentro del análisis multivariado con el fin de determinar cuando un cierto modelo es apropiado, esta clase de trabajos son frecuentemente representados gráficamente y dicha técnica implica el uso de las matrices de correlación y de covarianza, y para cada modelo se evalúa el grado de ajuste a los datos. Por lo tanto el objetivo del presente trabajo es obtener un análisis de ecuaciones estructurales que muestre la contribución de las diferentes partes de la planta sobre el rendimiento de resina y hule, así como detectar su relación con el análisis de componentes principales.

2. Metodología

En el presente análisis se muestraron 35 plantas de guayule de aproximadamente dos años de edad, provenientes de una población silvestre del Ejido Gómez Farias ubicado a 56 Km. de Saltillo, Coahuila, México. Arroyo (1999). Estas plantas fueron seccionadas en raíz, corona, ramas primarias y ramas secundarias; posteriormente se secaron en una estufa para obtener el peso seco, luego cada parte de la planta fue molida. Una muestra de 5 gramos de cada parte de tejido de las plantas fue utilizada para determinar el contenido de resina y hule, utilizando para ello tolueno y acetona como solventes. De las plantas seccionadas se estimaron las variables Pesos secos de hule y resina,

¹epadron@cima.uadec.mx

²nacho@sigma.iimas.unam.mx

³fel1925@yahoo.com

Contenidos de hule y resina, Altura, Diámetro, además de Pesos de hule y resina. Se utilizó el análisis de ecuaciones estructurales, con el fin de estimar los componentes de rendimiento de resina y hule, dentro del análisis de ecuaciones estructurales se obtuvieron los coeficientes de sendero los cuales se estimaron de acuerdo a Cox y Wermuth (1996). Los modelos saturados se ajustaron con el software EQS; Bentler (1995).

3. Resultados y Discusión

Al aplicar el análisis de ecuaciones estructurales para analizar los datos sólo se presentan las ecuaciones que contienen las componentes de rendimiento de resina y hule, la ecuación para peso de resina por planta (PR/PL) dada por el paquete estadístico es:

$$\begin{aligned} PR/PL = & \quad 0,227(PRRS) + 0,470(PPRP) \\ & + 0,247(PPCOR + 0,212(PPRAI) + 0,002(E_{15})) \end{aligned} \quad (1)$$

Dicha ecuación (1) tuvo un coeficiente de determinación de $R^2 = 0.999996$ el cual contribuye en un 99.9996 porciento de la variación explicada por dicho modelo. En el Cuadro 1 se observa que el efecto directo de PRRP(0.470), contribuye en una mayor cantidad que los efectos directos de PRRS(0.227), PPCOR(0.247) y PRRAI(0.212), con respecto a los efectos indirectos se observa que PPCOR y PRRAI, presentan altos valores a través de PRRP.

Cuadro1.- Efectos directos e indirectos entre PRRS, PRRP, PPCOR y PRRAI y sus correlaciones con PR/PL.

	Efectos directos e indirectos				Correlaciones de PRRS(11) PPRP(12),PPCOR(13) y PRRAI(14) con PR/PL(15)
	PRRS(11)	PPRP(12)	PPCOR(13)	PRRAI(14)	
PRRS(11)	0.227	0.270	0.103	0.103	$r_{11,15} = 0.6955^{**}$
PPRP(12)	0.130	0.470	0.209	0.164	$r_{12,15} = 0.963^{**}$
PPCOR(13)	0.094	0.397	0.247	0.156	$r_{13,15} = 0.885^{**}$
PRRAI(14)	0.110	0.364	0.181	0.212	$r_{14,15} = 0.858^{**}$

Residual=0.002 **Significativo al 1 %

En Figura 1 se puede observar que el efecto indirecto de PPCOR a través de PRRP se obtiene del producto de la correlación entre las variables predictoras PPCOR con PRRP($r=0.846$), y el efecto

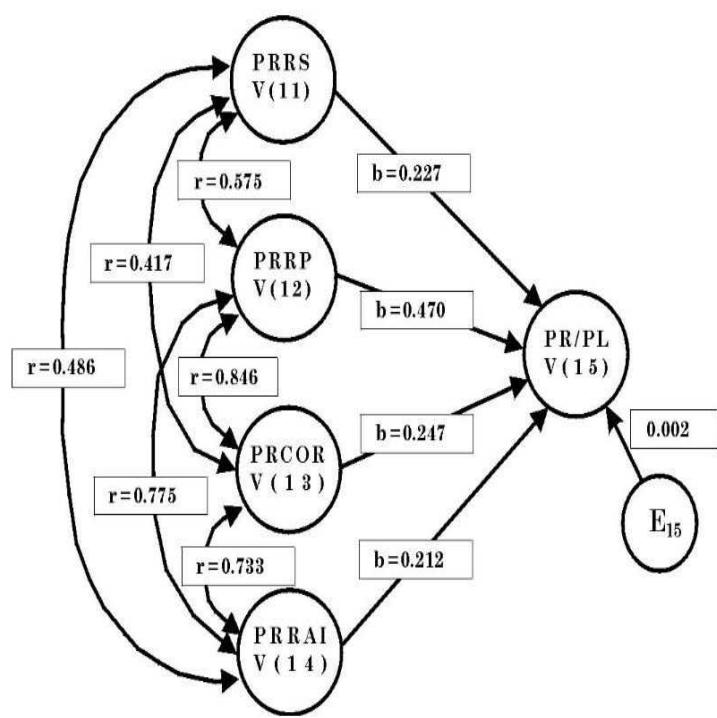


Figura 1.- Efectos directos (b)(\rightarrow) y correlación (r)(\leftrightarrow)

directo de la variable PRRP($b = 0.470$) sobre PR/PL es decir $(0.846)(0.470) = 0.397$, para mas de dos variables independientes los efectos indirectos es mejor obtenerlos vía la r menos el efecto directo. La ecuación para peso de hule por planta (PH/PL) dada por el paquete estadístico es.

$$\begin{aligned} PH/PL &= 0.187(PHRS) + 0.366(PHRP) \\ &\quad + 0.355(PHCOR) + 0.287(PHRAI) + 0.003(E_{15}) \end{aligned} \quad (2)$$

Dicha ecuación (2) tuvo un coeficiente de determinación de $R^2 = 0.999991$ el cual contribuye en un 99.9991 porciento de la variación explicada por dicho modelo. En el Cuadro 2 se observa que los efectos directos de PHRP(0.366), PHCOR(0.355) y PHRAI(0.287) contribuyen en una mayor cantidad que el efecto directo de PHRS(0.187), con respecto a los efectos indirectos se observa que PHRP, PHCOR y PHRAI presentan altos valores a través de PHCOR y PHRAI.

Cuadro2.- Efectos directos e indirectos entre PHRS, PHRP, PHCOR y PHRAI y sus correlaciones con PH/PL.

	Efectos directos e indirectos				Correlaciones de PHRS(16) PHRP(17),PHCOR(18) y PHRAI(19) con PH/PL(20)
	PHRS(16)	PHRP(17)	PHCOR(18)	PHRAI(19)	
PHRS(16)	0.187	0.194	0.075	0.086	$r_{16,20} = 0.536^{**}$
PHRP(17)	0.099	0.366	0.277	0.218	$r_{17,20} = 0.943^{**}$
PHCOR(18)	0.039	0.286	0.355	0.239	$r_{18,20} = 0.897^{**}$
PHRAI(19)	0.056	0.279	0.296	0.287	$r_{19,20} = 0.898^{**}$

Residual=0.003

Por lo tanto en Figura 2 se puede observar que el efecto indirecto de PHRAI a través de PHCOR se obtiene del producto de la correlación entre las variables PHRAI y PHCOR($r = 0.834$) y el efecto directo de la variable PHCOR($b = 0.355$) sobre PH/PL es decir $(0.783)(0.355) = 0.296$. En lo que respecta a la importancia de los componentes principales, se puede observar en el Cuadro 3 que con los cuatro primeros componentes se obtiene un 78.18 porciento de la variación explicada por los datos, de ellos, el componente principal uno (CP1) acumuló la mas alta proporción de variación explicada (49.06) porciento.

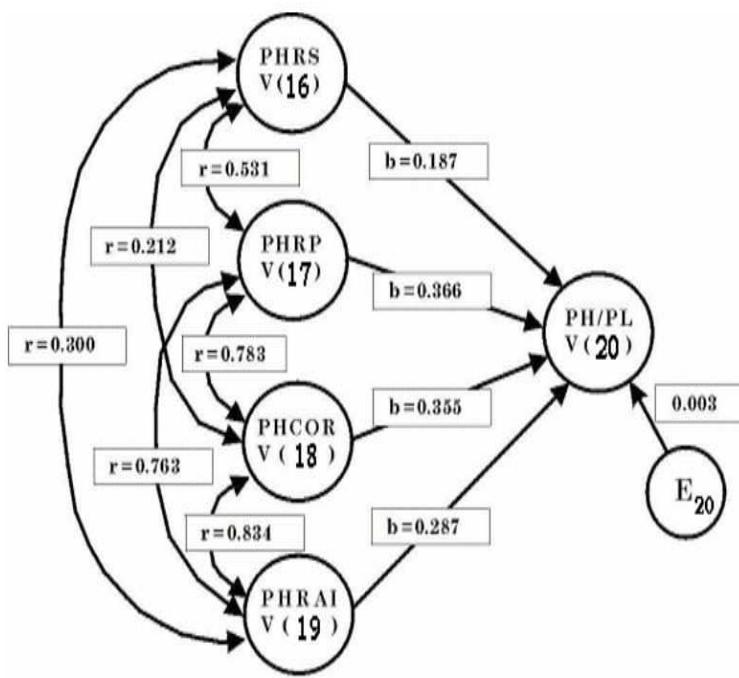


Figura 2.- Efectos directos (b)(\rightarrow) y correlación (r)(\leftrightarrow)

Cuadro 3.- Importancia de los componentes

	Comp.1	Comp.2	Comp.3	Comp.4
Desviación estándar	3.1416	1.6445	1.5640	1.3557
Proporción de varianza explicada	0.4906	0.1126	0.1019	0.0765
Proporción de varianza acumulada	0.4906	0.6033	0.7052	0.7818

4. Conclusiones

En lo que respecta a la relación entre el análisis de ecuaciones estructurales y el de componentes principales se observa que las variables peso de hule en ramas primarias, corona y raíz, los pesos de resina en ramas primarias, corona y raíz, y los pesos secos en ramas primarias, corona y raíz, en el análisis de sendero, presentaron los mas altos efectos directos sobre el rendimiento de hule y resina por planta y fueron los de más alta ponderación en la primera componente del análisis de componentes principales. De los resultados obtenidos en este trabajo, se establece que la producción de hule y resina en: ramas primarias, corona y raíz, fueron las principales estructuras de las plantas de guayule de dos años de edad que influyeron sobre el rendimiento total, ya que en estas partes de la planta se acumuló la mayor cantidad de peso. El conocimiento de la interrelación entre las diferentes partes de la planta y el uso del análisis de ecuaciones estructurales y componentes principales, fueron importantes para determinar la naturaleza entre estos componentes del rendimiento, ya que por lo general los mas altos efectos directos en el análisis de sendero presentaron las más altas ponderaciones en la componente uno del análisis de componentes principales.

Referencias

Arroyo, G.V. (1999). *Evaluación de Arbustos de Guayule (Phartenium argentatum Gray) en una Población Silvestre Regenerada Naturalmente*. Tesis Licenciatura en Fitotecnia, Universidad Autónoma Agraria Antonio Narro, Saltillo, Coahuila, México.

Bentler, M.P. (1995). *EQS Structural Equations Program Manual Multivariate Software Inc.* 49244 Balboa Blvd.# 368. Encino, California USA 91316.

Cox, D. and Wermuth, N. (1996). *Multivariate Dependencies: Model Analysis and Interpretation*.

Chapman and Hall.

Mejora del curado adhesivo QMI505MT en el encapsulado PDIP de los circuitos integrados

Rafael Pérez Abreu Carrión¹

Centro de Investigación en Matemáticas, A.C.

Omar Maynes Díaz

Texas Instruments de México, Aguascalientes

1. Resumen

El proceso de fabricación de circuitos integrado(C.I.) requiere de diferentes procesos para llegar a la producción final del CI. De manera general y en secuencia estos procesos son: Selección de la oblea, el aserrado, el corte con una sierra de alta precisión para individualizar a los dispositivos. Montaje del circuito en un armazón. La fijación de la barra con un adhesivo de alta tecnología, el curado del adhesivo y finalmente el encapsulado. Después del proceso de producción se pasa a la operación de prueba eléctrica, donde se prueba al circuito para verificar que cumpla con las funciones y parámetros eléctricos de calidad para el cual fue diseñado.

La industria constantemente busca métodos para reducir sus costos y elevar utilidades. En particular en el año 2003 Texas Instruments (TI) hizo un cambio de adhesivo en algunos procesos de CI, lo que vino a reducir los costos por el uso de adhesivos, pero a cambio resultaron problemas de operación de diferente índole, tales como CI defectuosos y menor vida útil de herramientas.

El objetivo de este estudio se centró en el comportamiento del adhesivo previo a la operación de enlace de estos circuitos integrados. A través de diseños de experimentos se logró establecer las condiciones optimas de trabajo, a fin de reducir el número de unidades defectuosas e incrementar la vida útil de algunas de las herramientas del proceso de enlace. Logrando con esto ahorros sustanciales a la empresa del orden de ciento veinte mil dólares al año.

¹rabreu@cimat.mx

2. Definición del problema

En Texas Instruments de México semiconductores de la ciudad de Aguascalientes, se implementó en 1998 un nuevo adhesivo en la operación de montaje. Dicho compuesto trajo beneficios a la planta como la mejora significativa en su costo y reducción de tiempo de curado, ya que este material se puede curar en poco tiempo, lo que permite disminuir el tiempo de ciclo en la fabricación de circuitos integrados. Sin embargo, presentó desventajas como la contaminación de las herramientas de enlazado (capilares) en el proceso de enlace. Cuando se está enlazando un circuito, el adhesivo desprende componentes volátiles (*outgassing*), este vapor se deposita sobre el capilar causando así su contaminación, ver Figura 1

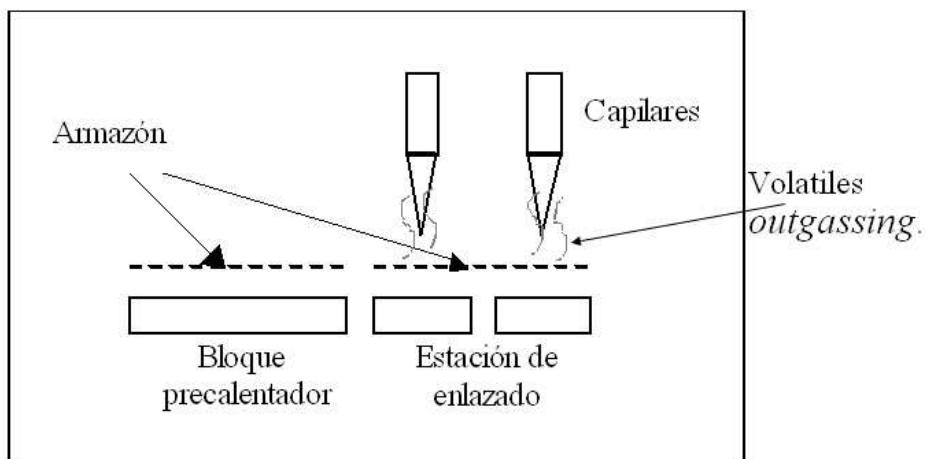


Fig. 1 Esquema de una enlazadora

Debido a que el adhesivo debe sobresalir en el perímetro del circuito con una altura aproximadamente de 30 % de la altura del circuito, esta es la fuente de donde proviene el *outgassing*.

En la operación de enlace se tienen dos armazones en proceso, uno se está curando sobre el “bloque precalentador” mientras que el otro armazón se está enlazando en la “estación de enlazado”.

3. Definición del problema

El capilar es una herramienta para colocar las conexiones entre el circuito y los pines por medio de un alambre de oro de un diámetro muy pequeño de aproximadamente 1 milésima de pulgada.

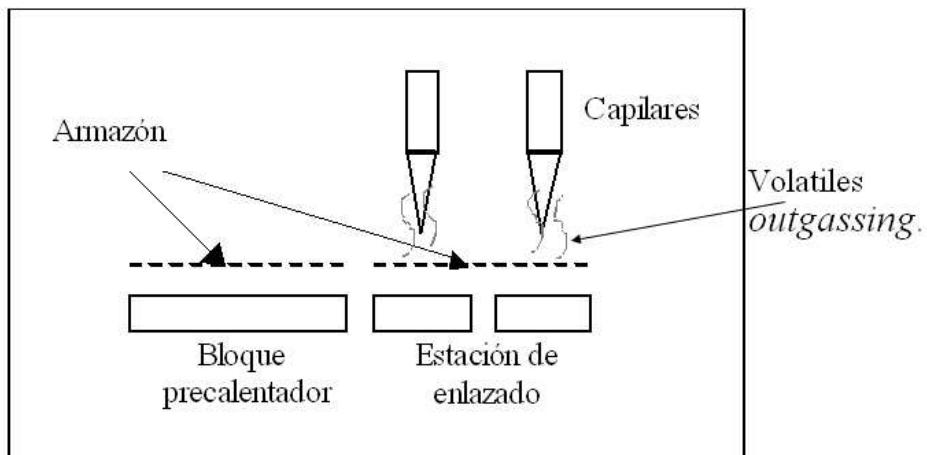


Fig. 1 Esquema de una enlazadora

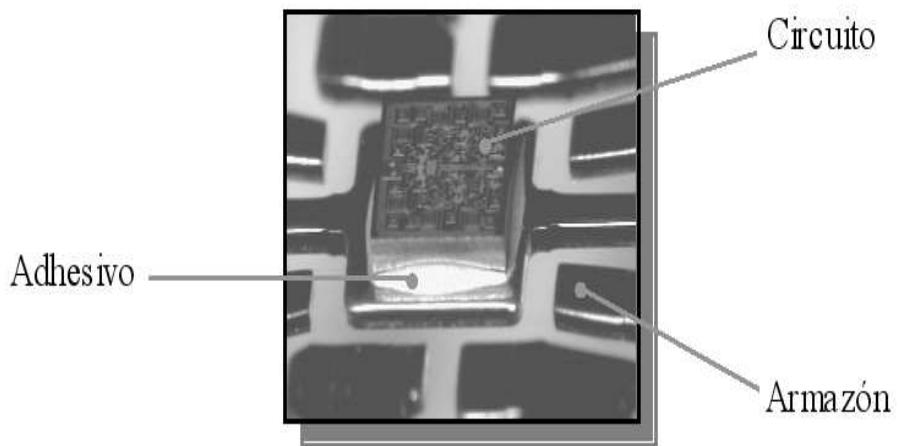
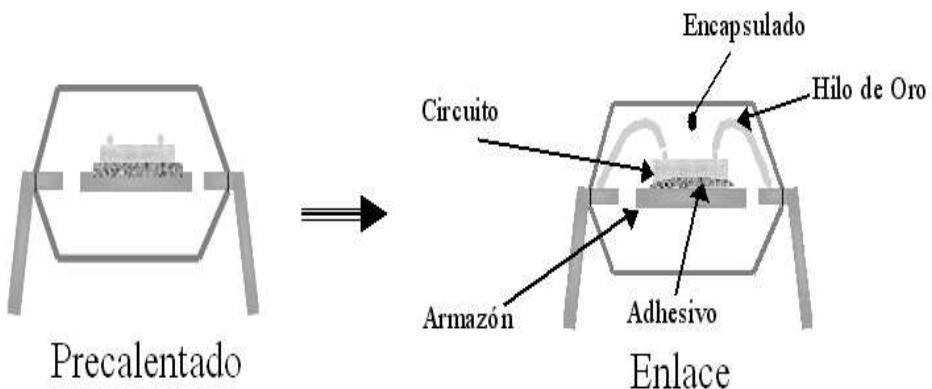


Imagen de un circuito montado (20x)

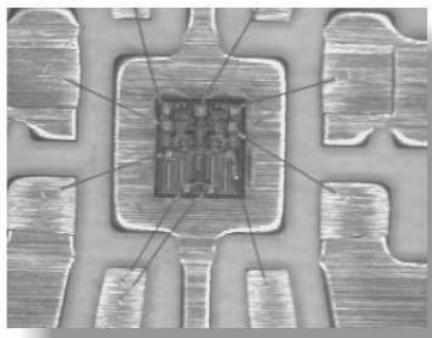
Para hacer la conexión del alambre de oro y la superficie del circuito se usa una combinación de ultrasonido (vibración del capilar en alta frecuencia), temperatura y presión del capilar.

La barra debe estar a cierta temperatura, el capilar debe ejercer presión y junto con el movimiento ultrasónico se hace una conexión mecánica y eléctricamente entre el aluminio del circuito y el alambre de oro.

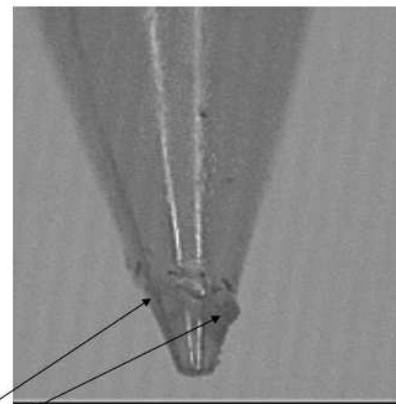


Vista transversal de circuitos integrados PDIP.

- *Vista de un circuito (15x)*



Vista lateral (35X)



Capilar contaminado.

4. Objetivos

4.1. Objetivo General

Mejorar el curado del adhesivo para minimizar la contaminación del capilar.

4.2. Objetivos particulares.

Determinar el mejor parámetro de temperatura por medio de la adherencia de la barra para evitar la contaminación del capilar (Herramienta de Enlace).

Determinar el efecto del parámetro tiempo para saber que tan conveniente es ajustar este factor.

5. Variables del Proceso

Variables de respuesta (Efecto):

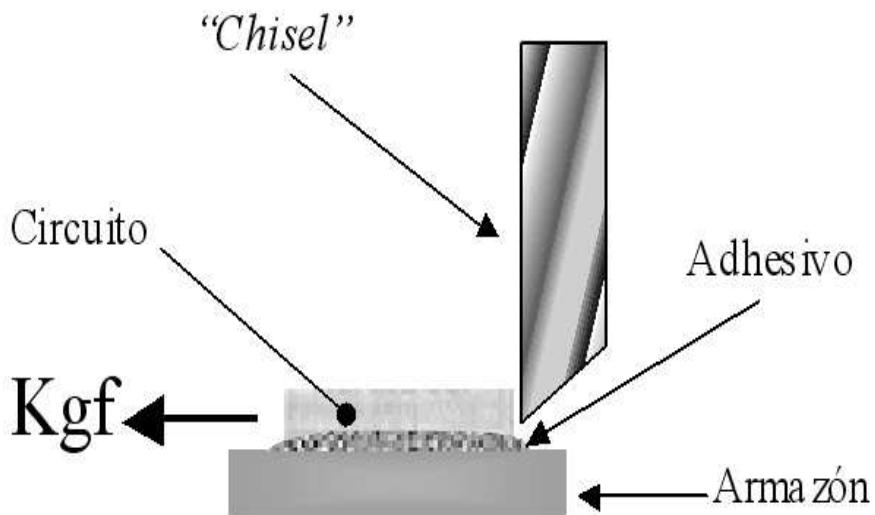
- a) **Contaminación del capilar** (Escala ordinal): Nada, Poca, Regular, Mucha o demasiada.
- b) **Porcentaje de unidades no enlazadas**, por ciclo de operación.
- c) **Adherencia de barra**. (fuerza necesaria para retirar un circuito adherido al armazón y se mide en KgF.)

Variables independientes (Causa):

Debido a que el adhesivo debe sobresalir en el perímetro del circuito con una altura aproximadamente de 30 % de la altura del circuito, esta es la fuente de donde proviene el outgassing.

- a) **Temperatura de curado**. (Temperatura a la que se fija el armazón, la cual se mide en la superficie y está dada en Grados Centígrados).
- b) **Tiempo de curado**. (Es el tiempo total en el que el armazón permanece sobre el bloque e incluye la rampa de calentamiento, el tiempo está dado en segundos).

Adherencia de barra. Fuerza necesaria para retirar un circuito adherido al armazón y se mide en KgF. (Prueba destructiva).



6. Diseños de Experimentos

Medir. Se llevó a cabo una serie de experimentos para medir y evaluar las variables del proceso. En esencia los diseños de experimentos estuvieron enfocados a encontrar la temperatura que maximizara la adherencia del circuito al armazón, y al mismo tiempo minimizara los compuestos volátiles que contaminan los capilares. (Herramientas de enlace).

$$\text{Adherencia} = f(\text{Temperatura}, \text{Tiempo}).$$

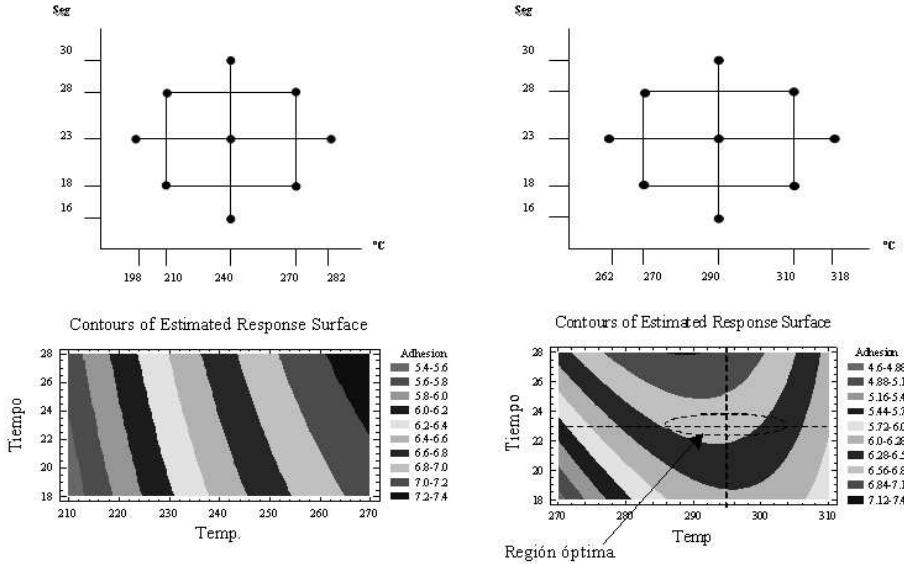
Se llevaron a cabo diseños de experimentos de superficie de respuesta.

7. Resultados y Conclusiones

Analizar. En el primer experimento se detectó una región donde el comportamiento del adhesivo aparenta ser lineal. Es decir no existe curvatura, cuando se aumenta la temperatura aumenta la adhesión. El rango de temperatura donde se determinó dicho comportamiento es de 210°C a 270°C.

El factor que tiene un efecto principal en la región del primer experimento es la temperatura.

En el segundo experimento se logró calcular cual fue la temperatura necesaria para lograr la máxima



adherencia de la barra. La temperatura fue de 295°C con un tiempo de 23 segundos, en la región analizada en un rango de 270°C a 310°C .

Una vez detectada cual es la temperatura necesaria para lograr la máxima adherencia, se logró comprobar que si el compuesto se expone a una temperatura mas alta que la óptima, la adherencia disminuye conforme se exceda de esta temperatura.

Los factores que presentan mayores efectos son el término cuadrático de la temperatura (AA) y el tiempo (B), para la región del segundo experimento.

Mejorar. Se logró determinar el efecto del parámetro tiempo para la región del segundo experimento. Por lo que no resulta conveniente aumentarlo ya que se puede aumentar la temperatura a un valor óptimo sin causar daños en el armazón.

Controlar. El nivel de temperatura donde se comienza a oxidar el armazón en un tiempo de 23 segundos es de 318°C , donde el 2.4 % de las unidades no se pudieron enlazar por causa aparentemente de oxidación en el armazón.

La contaminación observada en el capilar después de haber modificado los parámetros de operación a 295°C fue: Nada. La poca contaminación que se generó pudo haberse debido a las variaciones de temperatura en la estación de curado.



Capilares después de 800,000 enlaces. Antes del proyecto.



Capilar después de 800,000 enlaces. Después del proyecto. Ajustando la temperatura

8. Resultados

Uno de los resultados más sobresalientes fue el duplicar la vida útil de las herramientas de enlace. (capilares). De aproximadamente 800,000 a 1,700,000 enlaces.

9. Consideraciones futuras

El valor óptimo de temperatura solamente se probó en una sola maquina enlazadora, para extender el uso de las condiciones de curado, es necesario tomar una muestra representativa de las 55 máquinas enlazadoras con que cuenta actualmente la planta, y comprobar los resultados que se obtuvieron en este equipo piloto.

Clasificación Usando Análisis de Regresión de Gini: Una Alternativa a las máquinas de vector soporte

Blanca Rosa Pérez Salvador¹

Departamento de Matemáticas Universidad Autónoma Metropolitana, Iztapalapa

Sergio de los Cobos Silva ²

Departamento de Ingeniería Eléctrica

Universidad Autónoma Metropolitana, Iztapalapa

1. Introducción

Considere un conjunto de vectores $X_1, X_2, \dots, X_n \in \mathbb{R}^k$ asociados a las mediciones de algunas características de un conjunto de unidades muestrales que pertenecen a dos grupos bien definidos G_1 y G_2 . Se dice que estos vectores son linealmente separables si existe un vector constante $W \in \mathbb{R}^k$ y un escalar $a \in \mathbb{R}$ tales que se satisfacen las desigualdades

$$\begin{cases} W^T X < a & \text{si } X \in G_1 \\ y \\ W^T X > a & \text{si } X \in G_2 \end{cases} \quad (1)$$

La idea es encontrar al vector W_0 que satisfaga las condiciones dadas en (1) usando algún criterio de optimalidad. El problema ha sido tratado ampliamente usando la técnica de las máquinas de vector soporte en el reconocimiento de patrones, ver Vapnik, V. (1995) y Michie, D., et al (1994). con esta técnica se busca el hiperplano que se encuentra a la máxima distancia de los vectores más cercanos a él en cada grupo y esto se hace resolviendo el problema

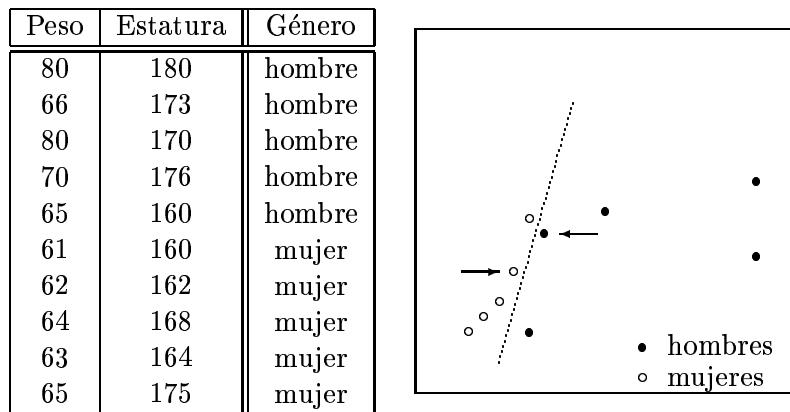
$$\min_{X \in G_1 \cup G_2} |W^T X - a| > 0,$$

la solución obtenida es un hiperplano definido por al menos un vector de cada grupo. Los vectores que definen este hiperplano se les conoce como los vectores soporte y se encuentran a la misma distancia del hiperplano solución.

¹psbr@xanum.uam.mx

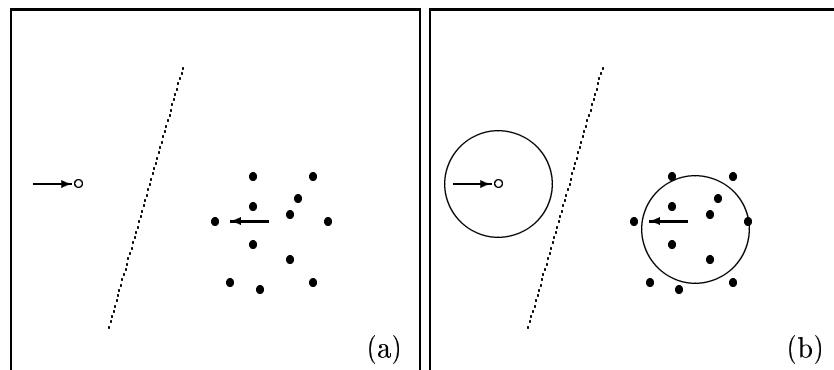
²cobos_58@terra.com.mx

Por ejemplo, observe los datos de peso y estatura de 5 hombres y 5 mujeres,



Como se puede ver, estos datos son linealmente separables. En la gráfica los vectores soporte se señalaron con una flecha.

Cuando el número de datos en ambos grupos es similar y éstos son linealmente separables, el método de la máquina de vector soporte proporciona una estimación adecuada de la ecuación de separación, pero si el número de datos en cada grupo difiere, la ecuación de separación presenta un sesgo que la aleja del grupo con más datos. Para explicar este comentario presentamos un caso extremo, en el cual se tienen dos grupos, uno de ellos tiene un sólo elemento en la muestra, mientras que el otro tiene doce elementos, como se ve en la gráfica siguiente:



En la figura (a) se dibujó el diagrama de dispersión de los datos, se trazó la recta de separación obtenida con el método de la máquina de vector soporte, y se señalaron los vectores soporte con unas flechitas.

En la figura (b) se dibujó además dos círculos, de igual tamaño, centrados en la media muestral de cada grupo de datos, esta última gráfica nos permite ver que la ecuación de separación de los datos da mayor margen de error al grupo con más datos muestrales.

Esto significa que hay más probabilidad de equivocarse al clasificar un elemento nuevo en el segundo grupo, que al clasificarlo en el primer grupo.

Esto sugiere que debería considerarse un método de estimación del hiperplano de separación que utilice toda la información de las dos submuestras. Una forma alternativa de obtener la ecuación de separación que considere tanto el tamaño de las dos submuestras, como la posición de los datos muestrales, es utilizando un modelo de regresión lineal como se describe en la siguiente sección.

2. Modelo de regresión lineal para clasificar datos

Considérese un conjunto de vectores $X_i = (X_{i1}, X_{i2}, \dots, X_{im})$, $i = 1, 2, \dots, n$, tal que

$$(X_{11}, X_{12}, \dots, X_{1m}), (X_{21}, X_{22}, \dots, X_{2m}), \dots, (X_{n1}, X_{n2}, \dots, X_{nm}) \in G_1 \cup G_2,$$

y sea Y la variable de clasificación definida por

$$Y_i = \begin{cases} 1 & \text{si } X_i \in G_1 \\ -1 & \text{si } X_i \in G_2 \end{cases},$$

bajo el supuesto que las variables X_1, X_2, \dots, X_m son linealmente separables, se tiene que utilizando el modelo de regresión lineal

$$Y_i = Y(X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_m X_{im} + \varepsilon_i,$$

es posible estimar una ecuación de clasificación, minimizando la magnitud de los errores de observación ε_i , dada por:

$$Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_m X_{im}).$$

Para los estimadores resultantes $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m$ se puede probar que

$$\hat{Y}(X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \dots + \hat{\beta}_m X_{im}.$$

satisface que $\hat{Y}(X_i) > 0$ cuando $X_i \in G_1$ y $\hat{Y}(X_i) < 0$ cuando $X_i \in G_2$.

Es muy probable que los errores de observación, ε_i en este caso, no sean variables aleatorias independientes e identicamente distribuidas de acuerdo a una ley normal con media cero y varianza σ^2 ($\varepsilon \sim N(0, \sigma^2 I)$), entonces no se puede utilizar con mucha confianza el método de los mínimos cuadrados para estimar los parámetros del modelo lineal, por lo que se debe recurrir a un método de estimación robusto.

3. Aplicando la Regresión lineal de Gini

Una alternativa para estimar los parámetros del modelo lineal propuesto es utilizar el método de regresión de Gini, ver Olkin and Yitzaki (1992), ya que los estimadores obtenidos con este método son poco sensibles a los valores extremos, además de que son insesgados y consistentes cuando $E(\varepsilon) = 0$.

La regresión de Gini se plantea de la siguiente manera:

Obtener la ecuación que minimiza la suma

$$\sum_{i=1}^n |\varepsilon_i| = \sum_{i=1}^n |Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \dots - \beta_m X_{im}|$$

La solución a este problema se encuentra con base al siguiente teorema que se presenta sin demostración.

Teorema 3.1 Sea el conjunto de vectores $\{X_1, X_2, \dots, X_n\}$ pertenecientes a dos grupos G_1 y G_2 y sea Y la variable de clasificación

$$Y_i = \begin{cases} 1 & \text{si } X_i \in G_1 \\ -1 & \text{si } X_i \in G_2 \end{cases},$$

entonces el conjunto de estimadores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_m$, que hace mínimo el valor de la función

$$\sum_{i=1}^n |\varepsilon_i| = \sum_{i=1}^n |Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \dots - \beta_m X_{im}|$$

es solución de un sistema de $m + 1$ ecuaciones de la forma

$$\begin{aligned}
\beta_0 + \beta_1 X_{i_01} + \beta_2 X_{i_02} + \beta_3 X_{i_03} + \dots + \beta_m X_{i_0m} &= Y_{i_0} \\
\beta_0 + \beta_1 X_{i_11} + \beta_2 X_{i_12} + \beta_3 X_{i_13} + \dots + \beta_m X_{i_1m} &= Y_{i_1} \\
\beta_0 + \beta_1 X_{i_21} + \beta_2 X_{i_22} + \beta_3 X_{i_23} + \dots + \beta_m X_{i_2m} &= Y_{i_2} \\
&\vdots &&\vdots &&\vdots \\
\beta_0 + \beta_1 X_{i_m1} + \beta_2 X_{i_m2} + \beta_3 X_{i_m3} + \dots + \beta_m X_{i_mm} &= Y_{i_m}
\end{aligned}$$

donde $\{X_{i_0}, X_{i_1}, \dots, X_{i_m}\} \subset \{X_1, X_2, \dots, X_n\}$.

El siguiente teorema indica que la ecuación de regresión estimada es una ecuación de clasificación de los datos muestrales.

Teorema 3.2 Si el conjunto de vectores $\{X_1, X_2, \dots, X_n\} \subset G_1 \cup G_2$ son linealmente separables, y $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_m x_m$ es la ecuación de regresión de Gini, entonces se satisfacen las siguientes condiciones:

$$\begin{aligned}
\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_m x_{im} &> 0 & \text{cuando } X_i \in G_1 \\
\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_m x_{im} &< 0 & \text{cuando } X_i \in G_2
\end{aligned}
,$$

Esta ecuación de separación es poco sensible a valores extremos, y siempre existe aunque los datos no sean linealmente separables. En este caso no se satisfacería el teorema 3.2, pero se puede establecer un escalar $a > 0$, para clasificar a un elemento, usando el correspondiente vector de observaciones X como:

$$\begin{aligned}
&\text{considerar que } X_i \in G_1 \quad \text{si } \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_m x_{im} > a \\
&\text{considerar que } X_i \in G_2 \quad \text{si } \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_m x_{im} < -a \\
&\text{quedá indefinido} \quad \text{si } -a < \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_m x_{im} < a
\end{aligned}
,$$

4. Conclusiones

La ecuación de clasificación de Gini, es una estimación robusta del modelo lineal y separa bien al conjunto de datos cuando estos son linealmente separables.

Referencias

Michie, D., D. J. Spiegelhalter, and C. C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. Englewood Cliffs, N.J.: Prentice Hall. Data available at <http://www.ncc.up.pt/liacc/ML/statlog/datasets.html>.

Olkin, I. And Yitzaki, S. (1992). Gini Regression Analysis, *International Statistical Review*, 60, 2, pp 185-192.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer-Verlag. 12

Tablas del Tamaño de Muestra y la Potencia de la Prueba UMPI Para Demostrar la Equivalencia de Medias de Dos Distribuciones Normales

Cecilia Ramírez Figueroa¹

*Instituto de Socioeconomía, Estadística e Informática del Colegio de Postgraduados. Montecillo,
Texcoco estado de México*

David Sotres Ramos²

*Instituto de Socioeconomía, Estadística e Informática del Colegio de Postgraduados. Montecillo,
Texcoco estado de México.*

1. Introducción

Las pruebas estadísticas de equivalencia son de gran utilidad en el diseño y análisis de ensayos clínicos enfocados a demostrar la equivalencia terapéutica de dos tratamientos. Por ejemplo una terapia experimental puede no ser superior al tratamiento estándar, pero en cambio puede tener un perfil de seguridad mejorado, mayor facilidad de administración, o ser mucho más barato. En estas situaciones, es conveniente realizar un experimento clínico para probar que ambos tratamientos son equivalentes en eficacia terapéutica. El objetivo de este trabajo es elaborar tablas para el tamaño de muestra, la región crítica y la potencia para realizar estas pruebas. También se muestran aplicaciones estadísticas utilizando las tablas desarrolladas.

2. Modelo estadístico de equivalencia

El modelo paramétrico estándar para demostrar la “equivalencia” de dos tratamientos con base en una variable continua supone que las observaciones correspondientes a los tratamientos representan dos muestras aleatorias independientes de dos distribuciones normales: $N(\mu_1, \sigma^2)$ y $N(\mu_2, \sigma^2)$, respectivamente; y dónde lo que se desea demostrar es la veracidad de la hipótesis alternativa en el siguiente juego de hipótesis:

¹ceciliarf@colpos.mx

²david.sotres@eca.com.mx

$$H_o : \{(\mu_1 - \mu_2) \leq -\Delta\} \quad \text{o} \quad \{(\mu_1 - \mu_2) \geq \Delta\} \quad \text{contra; } H_a : \{-\Delta < (\mu_1 - \mu_2) < \Delta\} \quad (1)$$

donde Δ es una constante fija y conocida.

Recientemente Wellek(2003) desarrolló la prueba Uniformemente Más Potente e Invariante (UMPI) para demostrar la equivalencia de medias de dos distribuciones normales, es decir para contrastar las hipótesis en (1).

3. Uso de las tablas

En la etapa de diseño de los estudios de equivalencia, se deben especificar el nivel de confiabilidad para la prueba (α) así como la diferencia máxima permitida ($\varepsilon > 0$) entre los tratamientos y también se debe especificar para qué nivel de potencia se requiere la prueba. Con todos estos elementos se determina el tamaño de muestra necesario para los estudios de equivalencia.

Ejemplo 3.1. Obtención del tamaño de muestra

Un investigador que desea probar si un medicamento en particular reduce la presión sanguínea propone hacer mediciones de la presión a un grupo de pacientes, administrar el medicamento y medir de nuevo una hora después. El investigador compara el cambio en la presión sanguínea con un medicamento estándar conveniente para el medicamento de patente. Si el investigador está buscando una diferencia entre los grupos de 1 Mm. Hg, entonces con una diferencia de medias estandarizada se considera apropiado especificar $\varepsilon = (\mu_1 - \mu_2) / \sigma = 0.50$, una confiabilidad de $\alpha=0.05$ y una potencia de 0.9, ¿ cuántos pacientes debería reclutar?

La tabla 4.6 proporciona que el investigador debería reclutar 80 pacientes para un grupo y 100 pacientes para el otro grupo a fin de obtener una potencia de la prueba de 0.90; o bien reclutar 90 pacientes para cada uno de los dos grupos con lo cual la potencia de la prueba resulta igual a 0.91

Ejemplo 3.2. Obtención del tamaño de muestra

Suponga que existe un tratamiento para infantes prematuros con deficiencia pulmonar que reduce la tasa de mortalidad a los 28 días de un 65 % a un 20 % (Lin 1995). Un producto alternativo nuevo esta siendo desarrollado y su eficacia es comparada con el tratamiento existente. Si se cree

que el nuevo producto es más efectivo que el existente y se diseña un ensayo clínico para mostrar una diferencia de 0.45 en la tasa de mortalidad, se considera apropiado especificar una diferencia de medias estandarizada $(\mu_1 - \mu_2) / \sigma = 0.40$, $\alpha = 0.05$, y una potencia de 0.76, ¿cuántos pacientes debería reclutar?

La tabla 4.7 proporciona que el investigador debería reclutar 100 pacientes para cada grupo a fin de obtener una potencia de la prueba de 0.76.

Ejemplo 3.3. Obtención de la potencia de la prueba

En un experimento con tamaño de muestra $m = n = 30$, $\alpha = 0.05$, $(\mu_1 - \mu_2) = 1$, y una diferencia de medias estandarizada $(\mu_1 - \mu_2) / \sigma = 1.00$ ¿Cuál es la potencia de este experimento?

La tabla 4.1 proporciona que la potencia de este experimento es de 0.9685.

4. Tablas del tamaño de muestra y de la potencia

Las tablas 4.1 a 4.4 proporcionan la potencia de la prueba UMPI para probar la equivalencia de dos medias al nivel $\alpha = 0.05$ para $\varepsilon = 1.0, 0.9, 0.8, 0.7, 0.6, 0.5$ y 0.4 y con tamaño de muestra $n = 10(10)100$.

n/m	10	20	30	40	50	60	70	80	90	100
10	0.4529									
20	0.6423	0.8589								
30	0.7169	0.9220	0.9685							
40	0.7555	0.9484	0.9839	0.9935						
50	0.7788	0.9620	0.9904	0.9968	0.9987					
60	0.7944	0.9699	0.9937	0.9983	0.9994	0.9998				
70	0.8055	0.9751	0.9955	0.9989	0.9997	0.9999	1.0000			
80	0.8138	0.9786	0.9966	0.9993	0.9998	1.0000	1.0000	1.0000		
90	0.8202	0.9812	0.9974	0.9995	0.9999	1.0000	1.0000	1.0000	1.0000	
100	0.8254	0.9831	0.9979	0.9997	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla 4.1 Potencia de la prueba UMPI de dos muestras con $\alpha = 0.05$ y $\varepsilon = 1.00$

n/m	10	20	30	40	50	60	70	80	90	100
10	0.3329									
20	0.5049	0.7601								
30	0.5845	0.8498	0.9269							
40	0.6281	0.8914	0.9571	0.9792						
50	0.6554	0.9145	0.9716	0.9883	0.9943					
60	0.6740	0.9288	0.9796	0.9927	0.9969	0.9985				
70	0.6874	0.9384	0.9844	0.9951	0.9982	0.9992	0.9996			
80	0.6976	0.9453	0.9875	0.9965	0.9988	0.9996	0.9998	0.9999		
90	0.7056	0.9504	0.9897	0.9974	0.9992	0.9997	0.9999	1.0000	1.0000	
100	0.7120	0.9543	0.9912	0.9994	0.9998	0.9999	1.0000	1.0000	1.0000	1.0000

Tabla 4.2 Potencia de la prueba UMPI de dos muestras con $\alpha=0.05$ y $\varepsilon=0.90$

n/m	10	20	30	40	50	60	70	80	90	100
10	0.2354									
20	0.3612	0.6187								
30	0.4307	0.7326	0.8462							
40	0.4723	0.7910	0.8979	0.9414						
50	0.4996	0.8256	0.9255	0.9625	0.9786					
60	0.5188	0.8482	0.9421	0.9739	0.9866	0.9924				
70	0.5329	0.8639	0.9528	0.9807	0.9910	0.9954	0.9974			
80	0.5437	0.8754	0.9601	0.9851	0.9936	0.9970	0.9984	0.9991		
90	0.5523	0.8842	0.9655	0.9880	0.9953	0.9979	0.9990	0.9995	0.9997	
100	0.5593	0.8911	0.9695	0.9901	0.9964	0.9985	0.9993	0.9997	0.9998	0.9999

Tabla 4.3. Potencia de la prueba UMPI de dos muestras con $\alpha=0.05$ y $\varepsilon=0.80$

n/m	10	20	30	40	50	60	70	80	90	100
10	0.1669									
20	0.2427	0.4436								
30	0.2887	0.5640	0.7081							
40	0.3184	0.6344	0.7840	0.8568						
50	0.3390	0.6791	0.8287	0.8969	0.9320					
60	0.3540	0.7096	0.8575	0.9211	0.9520	0.9686				
70	0.3653	0.7317	0.8773	0.9369	0.9643	0.9781	0.9858			
80	0.3743	0.7483	0.8916	0.9477	0.9723	0.9841	0.9902	0.9937		
90	0.3814	0.7612	0.9023	0.9555	0.9778	0.9879	0.9930	0.9957	0.9972	
100	0.3873	0.7715	0.9106	0.9613	0.9817	0.9906	0.9948	0.9970	0.9981	0.9988

Tabla 4.4. Potencia de la prueba UMPI de dos muestras con $\alpha=0.05$ y $\varepsilon=0.70$

n/m	10	20	30	40	50	60	70	80	90	100
10	0.1220									
20	0.1631	0.2799								
30	0.1879	0.3690	0.5078							
40	0.2043	0.4317	0.5976	0.6971						
50	0.2158	0.4761	0.6567	0.7591	0.8202					
60	0.2243	0.5086	0.6976	0.8003	0.8592	0.8954				
70	0.2309	0.5332	0.7273	0.8291	0.8855	0.9190	0.9402			
80	0.2361	0.5523	0.7497	0.8501	0.9041	0.9352	0.9541	0.9662		
90	0.2403	0.5677	0.7671	0.8660	0.9178	0.9467	0.9637	0.9743	0.9812	
100	0.2438	0.5802	0.7810	0.8785	0.9282	0.9551	0.9706	0.9799	0.9858	0.9896

Tabla 4.5. Potencia de la prueba UMPI de dos muestras con $\alpha=0.05$ y $\varepsilon=0.60$

n/m	10	20	30	40	50	60	70	80	90	100
10	0.0931									
20	0.1144	0.1711								
30	0.1266	0.2157	0.2979							
40	0.1346	0.2502	0.3650	0.4580						
50	0.1401	0.2769	0.4174	0.5278	0.6071					
60	0.1442	0.2980	0.4582	0.5798	0.6642	0.7232				
70	0.1474	0.3150	0.4902	0.6193	0.7063	0.7658	0.8078			
80	0.1499	0.3288	0.5158	0.6500	0.7383	0.7975	0.8385	0.8680		
90	0.1519	0.3404	0.5367	0.6743	0.7632	0.8217	0.8616	0.8897	0.9101	
100	0.1536	0.3501	0.5539	0.6941	0.7830	0.8407	0.8794	0.9061	0.9252	0.9392

Tabla 4.6 Potencia de la prueba UMPI de dos muestras con $\alpha=0.05$ y $\varepsilon=0.50$

n/m	10	20	30	40	50	60	70	80	90	100
10	0.0745									
20	0.0851	0.1107								
30	0.0909	0.1295	0.1632							
40	0.0945	0.1436	0.1918	0.2361						
50	0.0971	0.1545	0.2158	0.2752	0.3288					
60	0.0989	0.1632	0.2361	0.3090	0.3754	0.4326				
70	0.1003	0.1702	0.2532	0.3381	0.4151	0.4803	0.5339			
80	0.1014	0.1761	0.2678	0.3629	0.4487	0.5200	0.5776	0.6238		
90	0.1023	0.1810	0.2804	0.3843	0.4771	0.5530	0.6134	0.6612	0.6995	
100	0.1031	0.1851	0.2913	0.4027	0.5012	0.5807	0.6430	0.6918	0.7305	0.7616

Tabla 4.7. Potencia de la prueba UMPI de dos muestras con $\alpha=0.05$ y $\varepsilon=0.40$

Referencias

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

Wellek S. (2003). *Testing Statistical Hypotheses of Equivalence* Chapman and Hall, Florida.

Un Procedimiento para Selección de los Modelos Logit Mixtos

M.C. José de Jesús Ruiz Gallegos¹

Universidad Autónoma de Aguascalientes

Dra. Graciela González Fariás²

Centro de Investigación en Matemáticas, A.C.

1. El Modelo Logit

El modelo logit, es por mucho el modelo de elecciones discretas más simple y por ende, ampliamente utilizado. Su popularidad, es debido al hecho de que la fórmula para las probabilidades de elecciones tiene una forma cerrada y es fácilmente interpretable. Además este modelo es la base fundamental de todos los modelos de elecciones discretas.

Para deducir el modelo logit, le imponemos una distribución específica a la utilidad no observada. Un agente, etiquetado por n , enfrenta J alternativas. La utilidad que el agente n obtiene de la alternativa j se descompone en: Una parte denotada por V_{nj} , que es conocida a través de algunos parámetros, y una parte desconocida, ε_{nj} , que es tratada como aleatoria: $U_{nj} = V_{nj} + \varepsilon_{nj}, \forall j$.

El modelo logit se obtiene asumiendo que cada ε_{nj} es independiente, e idénticamente distribuido de valor extremo.

Usando lo anterior, se puede demostrar, McFadden (1974), que la probabilidad de que el agente n seleccione la alternativa i es

$$\begin{aligned} P_{ni} &= Pr(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \forall j \neq i) = Pr(\varepsilon_{nj} < \varepsilon_{ni} + V_{ni} - V_{nj} \forall j \neq i) \\ P_{ni} &= \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}, \end{aligned} \tag{1}$$

la cual es la probabilidad de elección logit.

Cuando la utilidad representativa es lineal en los parámetros: $V_{nj} = \beta' x_{nj}$, las probabilidades logit son $P_{ni} = \frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}$

¹adri_mexico@yahoo.com.m

²farias@cimat.mx

Una muestra de N agentes, se obtiene para el propósito de estimación. Como las probabilidades logit toman una forma cerrada, el procedimiento de máxima verosimilitud resulta fácilmente implementable. La probabilidad de la alternativa seleccionada por la persona n está dada por, $\prod_i (P_{ni})^{y_{ni}}$ donde $y_{ni} = 1$ si la persona n selecciona i y cero de otra manera.

Suponiendo que cada agente es independiente, la verosimilitud resulta ser

$$L(\beta) = \prod_{n=1}^N \prod_i (P_{ni})^{y_{ni}},$$

donde β es un vector que contiene los parámetros del modelo.

Por lo tanto, la función log-verosimilitud es

$$LL(\beta) = \sum_{n=1}^N \sum_i y_{ni} \log P_{ni}, \quad (2)$$

y el estimador, es el valor de β que maximiza esta función. McFadden (1974) mostró que $LL(\beta)$ es globalmente cóncava para una utilidad lineal en los parámetros.

2. El Modelo Logit Mixto

El modelo logit mixto, es un modelo altamente flexible que puede aproximar cualquier modelo de utilidad aleatoria (MacFadden y Train, 2000). Este modelo permite manejar variación aleatoria del gusto, patrones de sustitución sin restricción, y correlación en factores no observados sobre el tiempo. Su derivación es directa y la simulación de las probabilidades de las elecciones son computacionalmente simples. Sin embargo, se requiere asumir una distribución de probabilidad sobre los coeficientes de la utilidad observada. Esto conlleva a tomar la decisión sobre cuál resulta ser la más apropiada en cada caso y bajo qué criterio tomar esta decisión. Una solución a este problema es el tema fundamental de trabajo.

Un modelo logit mixto, es cualquier modelo cuyas probabilidades de elecciones puedan ser expresadas en la forma $P_{ni} = \int L_{ni}(\beta) f(\beta) d\beta$, donde $L_{ni}(\beta)$ es la probabilidad logit evaluada en los parámetros β : $L_{ni}(\beta) = \frac{e^{V_{ni}(\beta)}}{\sum_{j=1}^J e^{V_{nj}(\beta)}}$ y $f(\beta)$ es una función de densidad. $V_{ni}(\beta)$ es la porción observada de la utilidad, la cual depende de los parámetros β .

Si la utilidad es lineal en β , entonces $V_{ni} = \beta' x_{ni}$, luego

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_{j=1}^J e^{\beta' x_{nj}}} \right) f(\beta) d\beta, \quad (3)$$

la cual es la probabilidad logit mixta.

Una vez que especificamos la forma funcional de $f(\beta|\theta)$, debemos estimar los parámetros θ . Las probabilidades de las elecciones son aproximadas mediante simulación para cualquier valor dado de θ :

1. Seleccione un valor de β de $f(\beta|\theta)$ y etiquételo como $\beta^{(r)}$, con el superíndice $r = 1$ para indicar la primer selección.
2. Calcule $L_{ni}(\beta^{(r)})$.
3. Repita los pasos 1 y 2 muchas veces, y promedie los resultados.

El promedio obtenido con el procedimiento anterior es la probabilidad simulada:

$$\check{P}_{ni} = \frac{1}{R} \sum_{r=1}^R L_{ni}(\beta^{(r)}), \text{ donde } R \text{ es el número de réplicas.}$$

\check{P}_{ni} es estrictamente positiva, así que $\log \check{P}_{ni}$ está definido, lo cual es útil para aproximar la función log-verosimilitud.

\check{P}_{ni} es suave (dos veces diferenciable) en los parámetros θ y las variables x , lo cual facilita la búsqueda numérica para la maximización de la función verosimilitud y el cálculo de las elasticidades.

Las probabilidades simuladas son insertadas en la función log-verosimilitud para dar una log-verosimilitud simulada: $SLL = \sum_{n=1}^N \sum_{j=1}^J y_{nj} \log \check{P}_{nj}$, donde $y_{nj} = 1$, si el agente n seleccionó j y cero en otro caso.

El estimador de máxima verosimilitud simulada (MSLE) es el valor de θ que maximiza SLL .

3. Selección de Modelos

En esta sección se expone y se extiende el procedimiento propuesto por Pesaran y Pesaran (1993) para aplicarlo a modelos logit mixtos con distinta distribución en los coeficientes. La solución a este problema es mediante el método de simulación estocástica. Los métodos de simulación nos permiten calcular el estadístico de Cox para pruebas no anidadas, sin tener que realizar integraciones numéricas difíciles.

Considere los siguientes modelos de elecciones discretas (hipótesis no anidadas):

$$H_f : \quad P_{fni} = \int L_{ni}(\beta) f(\beta|\theta) d\beta \text{ y } H_g : \quad P_{gni} = \int L_{ni}(\beta) g(\beta|\gamma) d\beta, \text{ donde}$$

$$L_{ni}(\beta) = \frac{e^{\beta' x_{ni}}}{\sum_{j=1}^J e^{\beta' x_{nj}}} \text{ y } f \text{ y } g \text{ son dos funciones de densidad distintas, para los parámetros } \beta.$$

Por ejemplo, todos los parámetros β siguen una distribución normal, contra que algunos de los parámetros siguen una distribución normal y el resto una log-normal.

Las expresiones relevantes para la función de log-verosimilitud promedio, simuladas, bajo H_f y H_g están dadas por

$$H_f : \quad L_f(\mathbf{y}, \beta(\theta)|x) = \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^J y_{nj} \log \check{P}_{f_{nj}}, \quad (4)$$

$$H_g : \quad L_g(\mathbf{y}, \beta(\gamma)|x) = \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^J y_{nj} \log \check{P}_{g_{nj}}, \quad (5)$$

donde \mathbf{y} es el vector $N \times 1$, de observaciones de \mathbf{Y} y $\check{P}_{f_{nj}}$ y $\check{P}_{g_{nj}}$ representan las probabilidades simuladas, las cuales se obtienen usando las funciones de densidad $f(\beta|\theta)$ y $g(\beta|\gamma)$, respectivamente. Los cálculos se hacen usando las ideas de la sección anterior.

La varianza del estadístico de Cox, para la prueba de H_f contra H_g , puede ser calculada usando

$$\hat{v}_f^2 = \frac{1}{N} \mathbf{d}' \{ \mathbf{I}_N - \mathbf{R}(\hat{\theta}) [\mathbf{R}'(\hat{\theta}) \mathbf{R}(\hat{\theta})]^{-1} \mathbf{R}'(\hat{\theta}) \} \mathbf{d}, \quad (6)$$

donde $\mathbf{d}' = (d_1, d_2, \dots, d_N)$ y $\mathbf{R}(\theta)$ es la matriz de $N \times (p + 1)$:

$$\mathbf{R}(\theta) = \begin{bmatrix} 1 & \partial \log f(y_1, \theta) / \partial \theta_1 & \dots & \partial \log f(y_1, \theta) / \partial \theta_p \\ 1 & \partial \log f(y_2, \theta) / \partial \theta_1 & \dots & \partial \log f(y_2, \theta) / \partial \theta_p \\ \vdots & \vdots & & \vdots \\ 1 & \partial \log f(y_N, \theta) / \partial \theta_1 & \dots & \partial \log f(y_N, \theta) / \partial \theta_p \end{bmatrix}; \quad (7)$$

y notando que

$$d_n = \sum_{j=1}^J y_{nj} \log \left(\frac{\widehat{P}_{f_{nj}}}{\widehat{P}_{g_{nj}}} \right), \quad n = 1, 2, \dots, N. \quad (8)$$

$\widehat{P}_{f_{nj}}$ y $\widehat{P}_{g_{nj}}$ representan las probabilidades simuladas usando $f(\beta|\hat{\theta})$ y $g(\beta|\hat{\gamma})$, respectivamente. El cálculo de (6) se hace usando la función score.

El numerador del estadístico de Cox puede calcularse como

$$T_f(R) = L_f(\mathbf{y}, \beta(\hat{\theta})) - L_g(\mathbf{y}, \beta(\hat{\gamma})) - C_R(\beta(\hat{\theta}), \beta(\hat{\gamma}_*(R))), \quad (9)$$

donde,

$$L_f(\mathbf{y}, \beta(\hat{\theta})) - L_g(\mathbf{y}, \beta(\hat{\gamma})) = \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^J y_{nj} \log \left(\frac{\widehat{P}_{f_{nj}}}{\widehat{P}_{g_{nj}}} \right), \quad (10)$$

y la medida de “cercanía”, $C(\beta(\hat{\theta}), \beta(\hat{\gamma}_*(R)))$ puede calcularse usando la expresión

$$C_R(\beta(\hat{\theta}), \beta(\hat{\gamma}_*(R))) = \frac{1}{NR} \sum_{r=1}^R \sum_{n=1}^N \sum_{j=1}^J y_{njr} \log \left(\frac{\widehat{P}_{f_{nj}}}{\check{P}_{g_{nj}}^*(R)} \right), \quad (11)$$

donde, $\widehat{P}_{f_{nj}}$ y $\widehat{P}_{g_{nj}}$ se calculan usando las estimaciones máximo verosímiles de las respuestas discretas observadas. $\check{P}_{g_{nj}}^*(R)$ se calcula usando las estimaciones máximo verosímiles de las respuestas simuladas bajo H_f y y_{njr} , $n = 1, 2, \dots, N$, $j = 1, 2, \dots, J$ y $r = 1, 2, \dots, R$ son respuestas discretas simuladas independientemente bajo H_f ; con estas respuestas simuladas y usando la expresión $\hat{\gamma}_*(R, \hat{\theta}) = \frac{1}{R} \sum_{r=1}^R \hat{\gamma}_r(\hat{\theta})$, se obtienen las estimaciones de los parámetros que serán usadas en el cálculo de $\check{P}_{g_{nj}}^*(R)$. $\hat{\gamma}_r(\hat{\theta})$ es la estimación máximo verosímil de γ bajo H_g , usando observaciones de \mathbf{Y} que son simuladas suponiendo que $f(y, \hat{\theta})$ es el proceso de generación de datos.

El estadístico estandarizado de Cox simulado, digamos $S_f(R) = \sqrt{N}T_f(R)/\hat{v}_f$, puede calcularse usando el error estándar de la regresión de d_n y $\mathbf{R}(\beta(\hat{\theta}))$ como un estimador de \hat{v}_f

Referencias

- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, in P. Zarembka, ed., *Frontiers in Econometrics*, Academic Press, New York, pp. 105–142.
- McFadden, D. and Train, K. (2000). Mixed MNL models of discrete response, *Journal of Applied Econometrics* 15, 447–470.
- Pesaran, M.H. and Pesaran, B. (1993). A simulation approach to the problem of computing Cox's statistic for testing nonnested models, *Journal of Econometrics* 57, 377–392.
- Train, K. (2003). *Discrete Choice Methods with Simulation*, Cambridge University Press.

Estudio Estadístico de Algunas Variables Climatológicas en una Ciudad del Estado de Veracruz

R. Galván-Martínez¹

Instituto de Ciencias Básicas, Universidad Veracruzana

I.R. Sánchez-Galván

Facultad de Biología, Universidad Veracruzana

L. Cruz-Kuri²

Instituto de Ciencias Básicas, Universidad Veracruzana

1. Introducción

La ciudad de Xalapa ($19^{\circ} 32' N$, $96^{\circ} 55' W$), con un área de 146.00 Km^2 , limita al Norte con los municipios de Naolinco, Jilotepec y Banderilla, al Sur con Coatepec y Emiliano Zapata, al Este con Actopan y al Oeste con Rafael Lucio y Tlalnehuayocan. La vegetación predominante es de Bosque Mesófilo de Montaña, cuya comunidad arbórea es densa, que se desarrolla en sitios húmedos, con neblinas frecuentes, entre 800 - 2400 MSN. Incluye, tanto árboles perennifolios como caducifolios. Hay abundancia de líquenes, musgos, pteridofitas, fanerógamas, lianas, epifitas y helechos arborescentes. Por influencia volcánica, el tipo de suelo es Andasol, cuyo color oscuro está formado por cenizas volcánicas recientes, factible para la explotación forestal. El clima de Xalapa, según el sistema de clasificación de Köepen, modificado por Enriqueta García es: Templado con verano fresco y largo. Con temperatura anual mayor a 18°C , temperatura del mes más frío esta entre -3°C y 18°C , temperatura del mes más caliente sobre 22°C . (A)Cb Con poca oscilación Térmica.(i'). La marcha anual de temperatura tipo Ganges, es decir, la temperatura del mes más caliente es antes de junio y, se ubica en el Hemisferio Norte. (g). Precipitación del mes más seco mayor de 40 mm. C(fm). Régimen de lluvias de verano. Se presenta una pequeña temporada de sequía llamada canícula. (w"). La fórmula que representa el clima de Xalapa es (A)Cb(fm)(i')gw".

¹rgalvan@uv.mx

²kruz1111@yahoo.com.mx

2. Materiales y Métodos

Se cuenta con una base de datos digitalizada, entre otros, de datos de temperaturas máximas y mínimas ($^{\circ}\text{C}$) y totales de lluvias diarias (precipitación en $\frac{\text{mm}}{\text{cm}^2}$), desde enero de 1982 hasta diciembre del 2003, proporcionados por el Observatorio Meteorológico de Xalapa, Ver. Los procesamientos que se hacen son de tipo multivariado para las series de tiempo originales y estos también quedan apoyados en métodos gráficos; algunos de éstos son elementales, tales como los de construcción de histogramas para las temperaturas y precipitaciones, etc.. Para la ejecución de los cálculos numéricos y la elaboración de las gráficas se utilizó el programa de cómputo estadístico SPSS versión 12.0.

Las temperaturas mínimas presentan menos fluctuación alrededor de $14.07\ ^{\circ}\text{C}$. Las temperaturas máximas presentan mayor fluctuación alrededor de $23.93\ ^{\circ}\text{C}$. Siendo la variable Precipitación la que presenta mayor sesgo.

3. Análisis Estadísticos

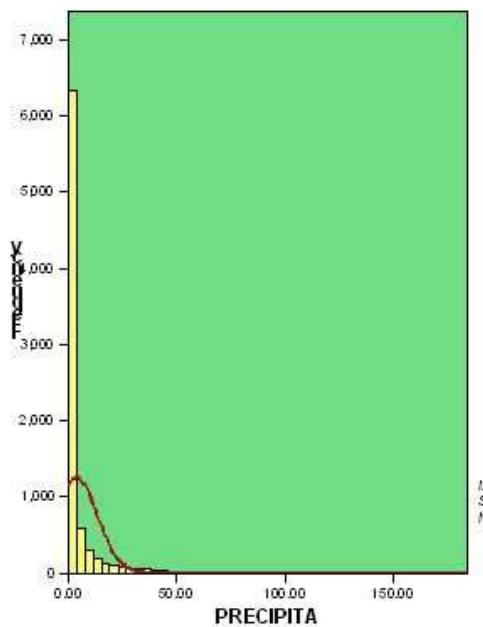
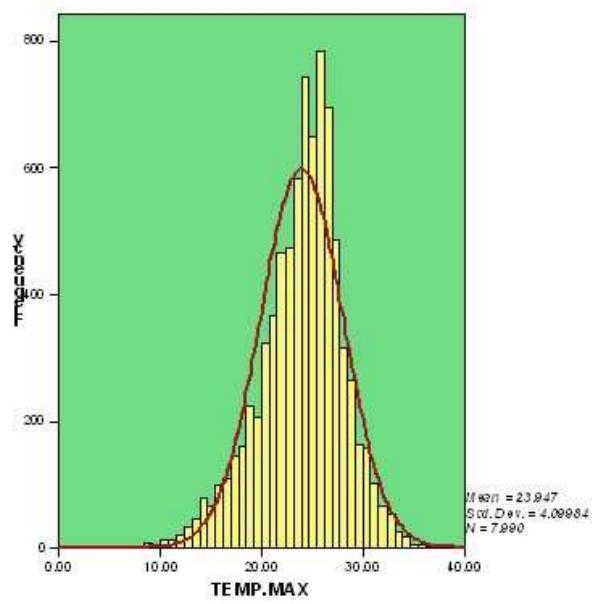
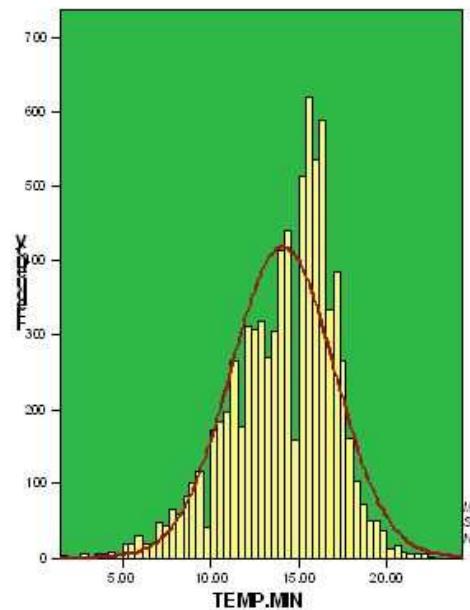
Para las series de tiempo correspondientes, se presentan gráficas de las variables suavizadas a 7 y 30 días. Medias Móviles para Temperaturas mínima, máxima y precipitación, en la parte superior están a 7 días (a la izquierda se gráfica el año 1982 y a la derecha el 2000). y en la inferior a 30 días, los años 1982 - 2003.

NOTA: La gráfica izquierda sugiere un comportamiento periódico, con un período mayor a 20 años.

4. Discusión

Se encuentran las correlaciones cruzadas ± 365 días para las tres variables. También, se calcula la serie de promedios para cada mes de las tres variables.

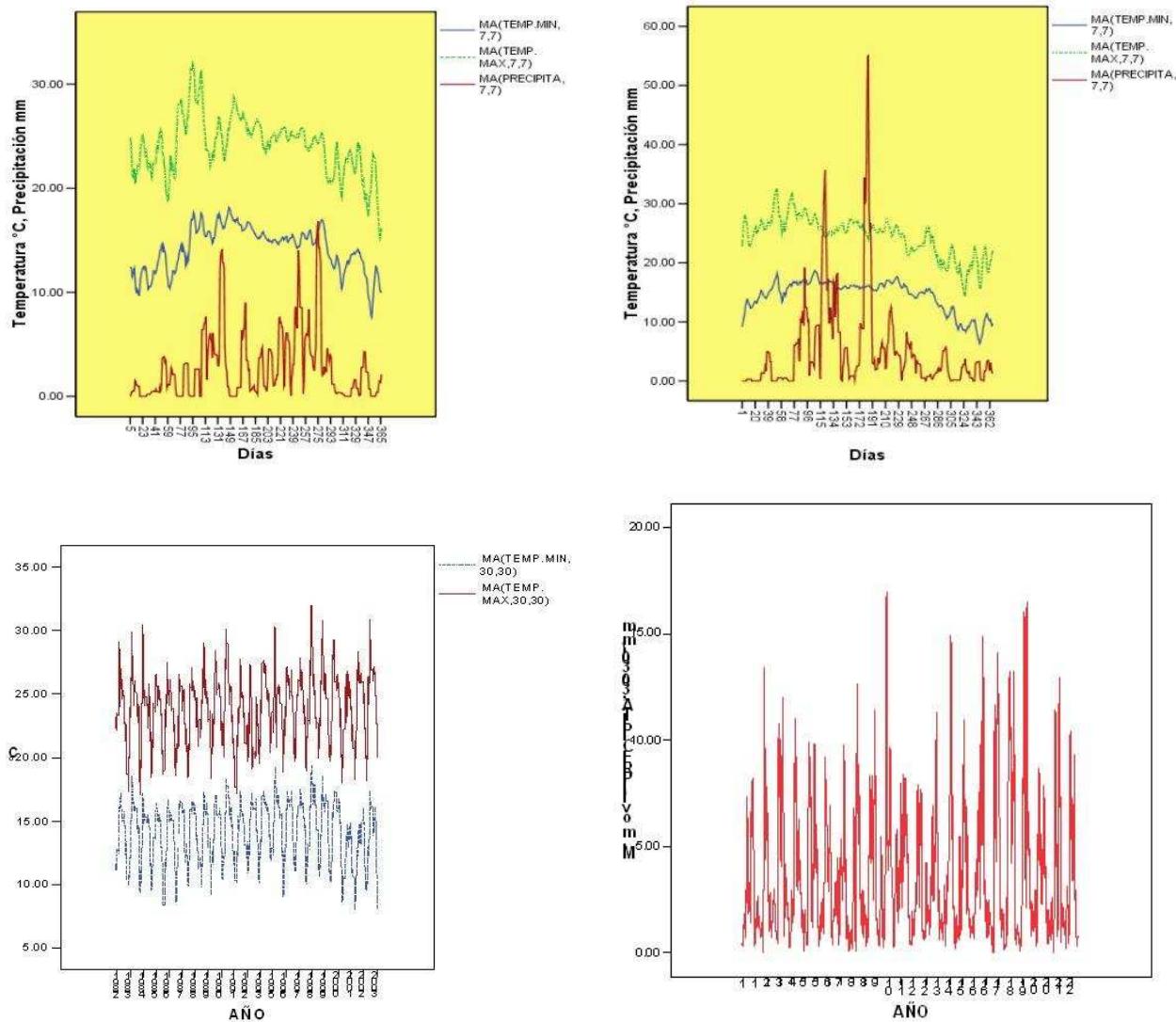
Las Figuras: 3.1, 3.2 y 3.3 presentan las correlaciones cruzadas para temperaturas máxima, mínima y precipitación para ± 365 días adelante y hacia atrás, ver las periodicidades y notar que aproximadamente la correlación es negativa la mitad del año y la otra mitad del año es positiva. Datos enero-1982 a diciembre-2003



Análisis Descriptivo. Figura 1.1 (arriba a la izquierda): Temperaturas Mínimas, Figura 1.2 (arriba a la derecha): Temperaturas Máximas, Figura 1.3 (abajo a la izquierda): Precipitación.

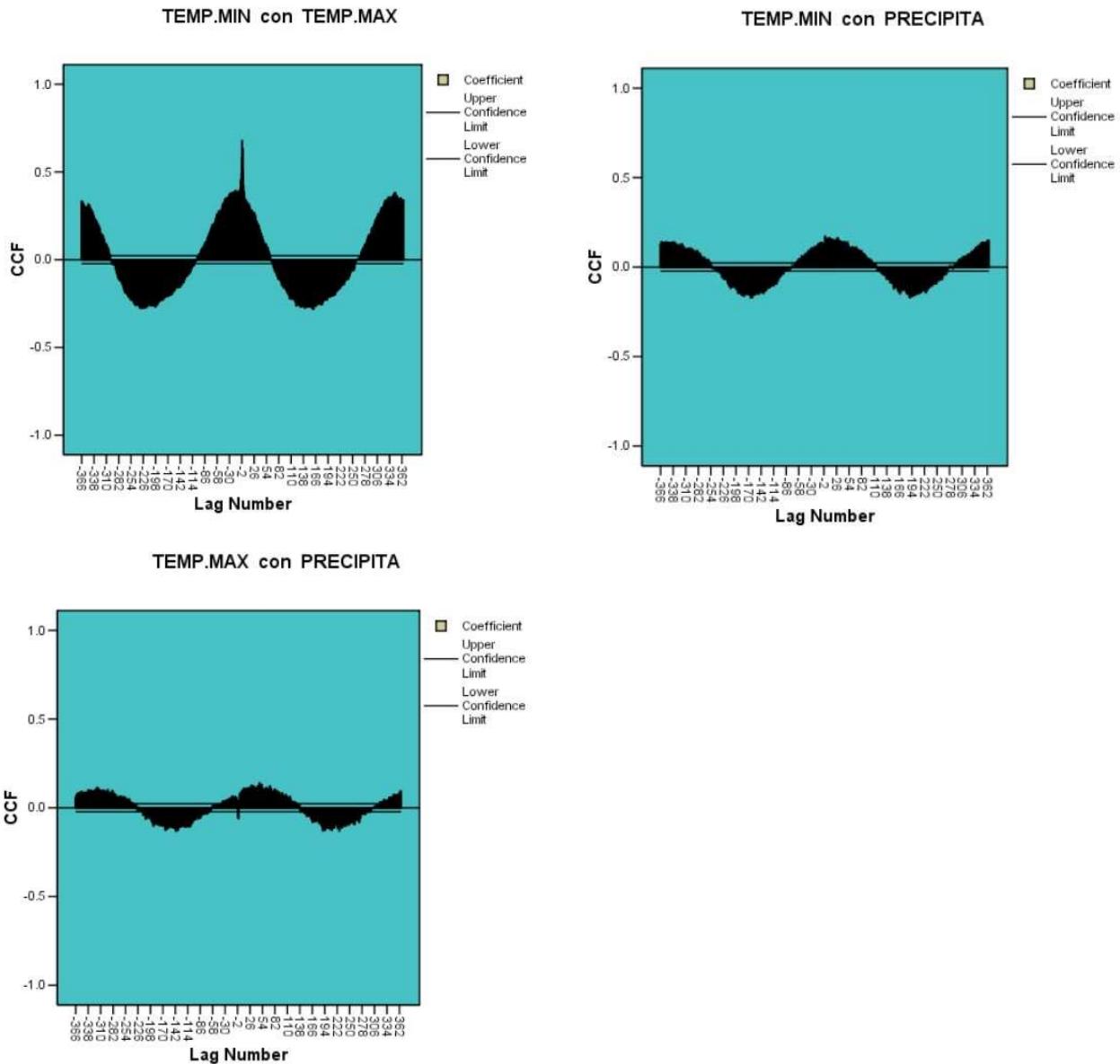
5. Conclusiones

Se desprende, tanto de los análisis a los datos originales, como de los procesamientos efectuados, lo que en particular era de esperarse, a saber, que se tienen periodicidades anuales para las vari-



Análisis Estadísticos. Figura 2.1 (arriba a la izquierda): Medias móviles temperaturas mínima, máxima y precipitación (1982), Figura 2.2 (arriba a la derecha): Medias móviles temperaturas mínima, máxima y precipitación (2000), Figura 2.3 (abajo a la izquierda): Medias móviles para temperaturas mínima, máxima (años 1982-2003), Figura 2.4 (abajo a la derecha): Medias móviles para precipitación (años 1982-2003).

ables registradas; por otra parte, al realizar análisis gráficos adicionales a las series de tiempo correspondientes, son sugeridos otras periodicidades.



Correlaciones cruzadas. Figura 3.1 (arriba a la izquierda): correlaciones temperatura mínima con temperatura máxima, Figura 3.2 (arriba a la derecha): correlaciones temperatura mínima con precipitación, Figura 3.3 (abajo a la izquierda) correlaciones temperatura máxima con precipitación

6. Agradecimientos

Los autores agradecen el valioso apoyo brindado por personal del Observatorio Meteorológico de la ciudad de Xalapa, Ver., al proporcionar la base de datos climáticos, objeto de este estudio así como

a S. Sánchez-Jácome por procesar la información.

Referencias

Diebold, F. X. (2001). *Elementos de Pronósticos*, México, D. F., Thomson Learning.

Chatfield, C. (1996). *The Analysis of Time Series: An Introduction*, 5a Edición. London: Chapman & Hall.

Pruebas de bondad de ajuste para el movimiento Browniano

José A. Villaseñor-Alva¹

Elizabeth González-Estrada

Colegio de Posgraduados.

1. Introducción

El movimiento Browniano o proceso de Wiener es uno de los modelos más conocidos de la teoría de procesos estocásticos debido a sus aplicaciones. Por definición, un proceso estocástico $\{X_t, t \geq 0\}$ es un movimiento Browniano si para $0 < s < t$ la variable aleatoria (v.a.) $X_t - X_s$ tiene distribución $N(0, \sigma^2(t - s))$, y es independiente de X_u , $0 < u < s$.

En aplicaciones en el área de finanzas, las opciones son instrumentos financieros a futuro que permiten administrar el riesgo de una acción dada en una cartera de inversión. Por esta característica favorable las opciones se ponen a la venta con un precio justo. Para el caso de las opciones de compra, Black y Scholes (1972) mostraron que, bajo el supuesto de que el logaritmo de los precios de la acción siguen el modelo del movimiento Browniano, existe un solo precio que es justo para una opción; dicho precio puede ser calculado mediante la fórmula de Black y Scholes.

Por lo tanto, es de interés contar con una metodología estadística que permita probar la hipótesis de que un proceso dado es un movimiento Browniano bajo condiciones lo más generales posibles sobre el proceso de interés.

En este trabajo se proponen tres pruebas estadísticas de bondad de ajuste para probar la hipótesis de que una realización finita de un proceso estocástico $\{X_t, t \geq 0\}$ sigue un movimiento Browniano. Estas pruebas son comparadas en términos de sus funciones de potencia por medio de simulación de Monte Carlo bajo algunas hipótesis alternativas de interés y diferentes tamaños de muestra. Dichas pruebas son aplicadas a dos conjuntos de datos reales.

¹jvillasr49@yahoo.com

2. Propiedades del Movimiento Browniano

Supóngase que la sucesión de v.a's X_1, X_2, \dots sigue un movimiento Browniano. Entonces, de acuerdo con su definición, las v.a's $X_t - X_{t-1}$ y $X_s - X_{s-1}$ para toda $s \leq t - 1$, son independientes e idénticamente distribuidas (iid) con distribución $N(0, \sigma^2)$. Es decir,

$$X_t = X_{t-1} + e_t, \quad e_t \text{ iid } N(0, \sigma^2). \quad (1)$$

El modelo en (1) también es conocido como caminata aleatoria con incrementos independientes e idénticamente distribuidos.

Para un conjunto $\{t_0, t_1, \dots, t_n\}$ de valores en el tiempo, la colección de v.a's $X_{t_0}, X_{t_1}, \dots, X_{t_n}$ es llamada una realización de tamaño $n+1$ del proceso estocástico $\{X_t, t \in T\}$.

A continuación se presentan las definiciones de algunos procesos estocásticos que están estrechamente relacionados con el movimiento Browniano.

Se dice que el proceso estocástico $\{X_t, t \in T\}$ tiene incrementos independientes si para cualesquiera tiempos $t_0 < t_1 < \dots < t_m$ las m v.a's $X_{t_1} - X_{t_0}, \dots, X_{t_m} - X_{t_{m-1}}$ son independientes y se dice que el proceso tiene incrementos estacionarios si las v.a's $X_{t+h} - X_{s+h}$ y $X_t - X_s$ tienen la misma distribución, para toda $h > 0$ y $0 < s < t$.

Se dice que una sucesión de variables aleatorias $\{X_t\}_{t=1}^n$, $n \geq 1$, es una martingala si y sólo si $E\{|X_t|\} < \infty$ y $E\{X_{t+1}|X_t, \dots, X_1\} = X_t$.

Se dice que un proceso estocástico $\{X_t, t \in T\}$ es Gaussiano si para cualquier n y cualquier subconjunto $\{t_1, t_2, \dots, t_n\}$ de valores positivos, las v.a's $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ tienen distribución conjunta normal.

Se puede demostrar que la siguiente definición es equivalente a la anteriormente expuesta.

Un proceso estocástico $\{X_t, t \geq 0\}$ es un movimiento Browniano si es tal que $X_0 = 0$,

1. tiene incrementos independientes y estacionarios, y

2. la v.a. $X_t - X_s$ tiene distribución $N(0, \sigma^2(t-s))$, $0 < s < t$.

El movimiento Browniano es un proceso estocástico con parámetro de tiempo continuo, $T = \{t : 0 \leq t < \infty\}$, y espacio de estados continuo, $S = \{s : -\infty < s < \infty\}$. Es posible verificar que un movimiento Browniano es un proceso Gaussiano y una martingala. Ross (1996) hace una revisión detallada de las propiedades del proceso Browniano.

La siguiente caracterización del movimiento Browniano es la base de la metodología que se propone en este trabajo.

Teorema. *Sean X_1, X_2, \dots, X_n observaciones del proceso estocástico $\{X_t, t \geq 0\}$ registradas en tiempos equidistantes. Las v.a's X_1, X_2, \dots, X_n siguen un movimiento Browniano si y sólo si las v.a's $e_1 = X_1 - X_0, \dots, e_n = X_n - X_{n-1}$ son iid $N(0, \sigma^2)$, donde $X_0 \equiv 0$.*

Note que las v.a's e_1, e_2, \dots, e_n son una trayectoria muestral del proceso de primeras diferencias de $\{X_t, t \geq 0\}$, denotado como $\{e_t, t \geq 0\}$, donde $e_t = X_t - X_{t-1}$, $t = 1, \dots, n$.

3. Pruebas propuestas

Para probar la hipótesis de movimiento Browniano se proponen pruebas del tipo unión-intersección las cuales consisten en probar por separado las tres hipótesis siguientes: H_{01} : las v.a's e_1, e_2, \dots, e_n son iid, H_{02} : las v.a's e_1, e_2, \dots, e_n son normales y H_{03} : las v.a's e_1, e_2, \dots, e_n tienen media cero. Se rechaza la hipótesis de movimiento Browniano si se rechaza alguna de las H_{0j} , $j = 1, 2, 3$. Para esto se consideran los tres procedimientos de prueba siguientes:

P1: compuesto por las pruebas de McLeod-Li (ML) y t .

P2: compuesto por las pruebas de rachas (R), Royston (W') y t .

P3: compuesto por las pruebas de rachas (R), Kolmogorov-Smirnov (D) y t .

donde R es una prueba para aleatoriedad, ML es una prueba para aleatoriedad y normalidad, W' (extensión de la prueba de Shapiro-Wilk) y D son pruebas para normalidad y la prueba de t es para

probar media cero. Para una discusión detallada de estas pruebas véanse por ejemplo, Campbell et al. (1997), Brockwell y Davis (1996), Royston (1992), Mood et al. (1974) y Casella y Berger (1990).

Por la desigualdad de Bonferroni (Casella y Berger, 1990), cada procedimiento de prueba es de tamaño menor o igual que α cuando la suma de los tamaños de las pruebas que integran al procedimiento sea menor o igual que α .

El procedimiento propuesto Pi para probar H_0 : las v.a's X_1, X_2, \dots, X_n siguen un movimiento Browniano es el siguiente:

1. Calcular e_1, e_2, \dots, e_n .
2. Con base en e_1, e_2, \dots, e_n , calcular las estadísticas de prueba que componen a Π .
3. Para un tamaño α dado, rechazar la hipótesis nula si al menos una de las estadísticas de prueba que componen a Π rechaza la hipótesis correspondiente.

Es conveniente hacer notar que las pruebas que componen a Π requieren la siguiente secuenciación: aleatoriedad, normalidad y media cero.

4. Comparación de las pruebas por simulación

En esta sección se discute un estudio de simulación de Monte Carlo para la comparación de las potencias de los procedimientos de prueba propuestos contra algunos procesos no brownianos y tamaños de muestra $n = 20, 30, 50, 100, 200$ y 300 .

Los procesos alternativos considerados son los siguientes:

1. El proceso autorregresivo de orden 1, AR(1): $X_t = \phi X_{t-1} + Z_t$, Z_t iid $N(0, 4)$, con $\phi = 0.1, 0.25, 0.5, 0.75, 0.9$. En este caso se pretende analizar las potencias cuando el parámetro ϕ se aleja de 1.
2. Para investigar la sensibilidad de los procedimientos de prueba cuando los incrementos del proceso no son idénticamente distribuidos se consideró el proceso: $X_t = X_{t-1} + Z_t$, donde las Z_t son v.a's independientes con distribución: (a) $N(0, (0.1t)^2)$, (b) $N(0, t/2)$ y (c) $N(0, \sigma^2)$, con σ una v.a. con distribución $U(0, 2)$.

Tablas de la Sección 4.

Resultados

Potencia estimada de P1, P2 y P3 contra la alternativa $X_t = \phi X_{t-1} + Z_t$, Z_t iid $N(0,4)$

$\alpha=0.1$.

n	P1					P2					P3				
	$\phi=0.1$	$\phi=0.25$	$\phi=0.5$	$\phi=0.75$	$\phi=0.9$	$\phi=0.1$	$\phi=0.25$	$\phi=0.5$	$\phi=0.75$	$\phi=0.9$	$\phi=0.1$	$\phi=0.25$	$\phi=0.5$	$\phi=0.75$	$\phi=0.9$
20	0.053	0.037	0.0345	0.0435	0.035	0.18	0.1345	0.0785	0.0635	0.062	0.1835	0.1355	0.0845	0.062	0.0595
30	0.065	0.0445	0.0400	0.0415	0.0445	0.31	0.191	0.1180	0.0655	0.056	0.31	0.205	0.1190	0.063	0.0575
50	0.0925	0.0585	0.0470	0.0415	0.0425	0.5555	0.4015	0.2115	0.097	0.0735	0.5555	0.408	0.2195	0.0965	0.0655
100	0.1575	0.104	0.0560	0.044	0.0545	0.8705	0.7155	0.3345	0.1275	0.073	0.869	0.7145	0.3400	0.126	0.0845
200	0.3005	0.1645	0.0745	0.0655	0.052	0.9955	0.9585	0.6280	0.221	0.112	0.9955	0.958	0.6310	0.218	0.1085
300	0.4555	0.233	0.0875	0.052	0.0535	1	0.9945	0.7965	0.2685	0.1105	1	0.995	0.7970	0.269	0.1085

Potencia estimada de P1, P2 y P3 contra la alternativa $X_t = X_{t-1} + Z_t$, $\alpha=0.1$

a) Z_t iN($0, 0.1t$)

b) Z_t iN($0, t/2$)

c) Z_t iN($0, \sigma^2$) σ U($0, 2$)

n	P1	P2	P3
20	0.0645	0.2155	0.1930
30	0.1140	0.3105	0.2860
50	0.2455	0.5030	0.4785
100	0.8780	0.8260	0.8115
200	0.9975	0.9970	0.9930
300	1	1	1

n	P1	P2	P3
20	0.0585	0.2830	0.1945
30	0.1095	0.3535	0.2275
50	0.2665	0.5640	0.3675
100	0.7205	0.8420	0.6630
200	0.9885	0.9945	0.9490
300	0.9990	1	0.9930

n	P1	P2	P3
20	0.0550	0.2500	0.2735
30	0.0590	0.3735	0.3745
50	0.080	0.5420	0.5230
100	0.0820	0.8480	0.8350
200	0.1035	0.9920	0.9920
300	0.1060	1	0.9995

Con la primer alternativa, el procedimiento P2 es más potente que P1 y P3. La potencia de los procedimientos P2 y P3 se incrementa a medida que se incrementa el tamaño de muestra y a medida que el coeficiente ϕ se aleja de 1. El procedimiento P1 no tiene potencia en ningún caso.

Con los procesos alternativos 2.(a) y 2.(b), los 3 procedimientos tienen buena potencia cuando el tamaño de muestra es mayor o igual a 100. P2 y P3 presentan buena potencia contra el proceso alternativo 2.(c) para tamaños de muestra mayores o iguales a 50.

5. Aplicaciones

La metodología propuesta se aplicó a los siguientes conjuntos de datos para probar la hipótesis de movimiento Browniano con base en una realización de tamaño n :

1. Precios de petróleo crudo durante el periodo 16/7/1996 a 21/4/1997 en los Estados Unidos ($n = 275$), Ross (1999).
2. Precios de cierre diarios de las acciones de IBM durante el periodo 17/05/1961 a 2/12/1962 ($n = 369$), Box *et al* (1998).

Considerando un nivel de significancia $\alpha = 0.1$, ninguno de los tres procedimientos de prueba rechaza la hipótesis nula de que los precios de petróleo crudo siguen un movimiento Browniano. Para los precios de cierre de las acciones de IBM, los tres procedimientos rechazan la hipótesis nula.

6. Conclusiones

El procedimiento P1 presenta muy baja potencia en general. De donde se concluye que la prueba de McLeod-Li no es eficiente para probar que un conjunto de datos forma una muestra aleatoria de la distribución normal.

El procedimiento P2, en general, presenta mayor potencia ante las alternativas consideradas, tanto para tamaños de trayectorias muestrales pequeños como grandes. Además, se observó que la potencia de P2 aumentó conforme se aumentó el tamaño de muestra en los casos estudiados, lo que indica que P2 podría ser una prueba consistente.

Por lo anterior, se recomienda usar el procedimiento P2, el cual está basado en las pruebas de rachas, Shapiro-Wilk y t.

Referencias

Black, F. and Scholes, M. (1972). The valuation of option contracts and a test of market efficiency. *Journal of Finance*, 27, 399-252.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1998). *Time series analysis: forecasting and control.* 2nd Ed. USA: Prentice Hall, Inc.

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to time series analysis and forecasting.* New York: Springer-Verlag.

Campbell, J. Y., Lo, A. W. and Mackinlay, A. C. (1997). *The econometrics of financial markets.* USA: Princeton University Press.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* California: Brooks/Cole.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the theory of statistics.* Tokyo: McGraw-Hill, Inc.

Ross, S. M. (1996). *Stochastic processes.* 2nd Ed. New York: John Wiley & Sons.

Ross, S. M. (1999). *An introduction to mathematical finance: options and other topics.* New York: Cambridge University Press.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, 2, 117-119.

Identifying Sectors of Deviations from Multinormality

Alexander von Eye¹

G. Anne Bogat²

Michigan State University, 107 D Psychology Building, East Lansing, MI 48864, USA

1. Introducción

Many parametric multivariate statistical procedures operate under the assumption that the data are a random sample from a multivariate normal population. Examples of such procedures include MANOVA and Structural Equations Modeling (SEM). For example, Bentler's SEM program EQS (Bentler, 1992) "... automatically generates estimates and test statistics ... based on multinormal theory ..." (p. 216). If the assumption of multinormality is correct, parameter estimation is simplified and the parameters have many desirable characteristic. If, however, this assumption is violated, severe bias can result, and parameter interpretation can be problematic (Hu & Bentler, 1995). This problem becomes even worse, when samples are small (Ito, 1980).

In this article, we review the two most popular methods to test whether data stem from multinormal populations, and discuss two new tests. The popular methods are Mardia's (1970, 1980) tests of multivariate skewness and kurtosis. The new methods (von Eye & Bogat, 2004; von Eye & Gardiner, 2004) (a) focus on sectors of the multivariate space, and (b) include an omnibus test. In the following sections, we first introduce a notation for the multinormal distribution. We then describe Mardia's tests, the two new tests, and give an empirical data example.

2. The multinormal distribution

The random vector $x(d \times 1)$ follows a multinormal distribution if its probability density function can be described by

$$f(x) = (2\pi)^{-d/2} |\Sigma|^{-0.5} e^{-0.5(x-\mu)' \Sigma^{-1} (x-\mu)},$$

¹voneye@msu.edu

²bogat@msu.edu

with positive definite covariance matrix Σ of rank d .

3. Mardia's tests of multivariate skewness and kurtosis

Let \bar{x} be the mean of X . then, the Mahalanobis distance of score x_i from the mean can be described by

$$r_i^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}),$$

where S is the sample covariance matrix.

The Mahalanobis angle between the vectors $x_i - \bar{x}$ and $x_j - \bar{x}$ is

$$r_{ij} = (x_i - \bar{x})' S^{-1} (x_j - \bar{x}).$$

Using these two r -measures, Mardia (1970) defines *multivariate skewness* as

$$b_{1d} = \frac{1}{N^2} \sum_i \sum_j r_{ij}^3.$$

The measure $b_{1d}/6$ is distributed as χ^2 with $d(d+1)(d+2)/6$ degrees of freedom. *Multivariate kurtosis* is defined as

$$b_{2d} = \frac{1}{N} \sum_i r_i^4.$$

This measure is asymptotically normally distributed with mean $d(d+2)$ and variance $8d(d+2)/N$. Variables are assumed to be *iid*.

The concepts of multivariate skewness and kurtosis are important because they allow one to test hypotheses that are compatible with multinormality. The tests of multivariate skewness and kurtosis therefore do not address the distributional assumption directly. The two new tests described in the following sections focus on the distribution itself. It is also important to note that, if X follows a multivariate normal distribution, then each subset of X , in particular each univariate subset, is also normally distributed. The inverse does not hold true. Variables that follow a univariate normal distribution do not necessarily follow a joint multivariate distribution, unless they are independent.

4. Two new tests of multinormality

The first of the new tests to be described in the following paragraphs (von Eye & Bogat, 2004; von Eye & Gardiner, 2005) can be seen as a multivariate extension of the well known univariate χ^2 test of normality that is described in many textbooks (e.g., Glass & Hopkins, 1984). An algorithmic description of this test can include the following 5 steps:

1. *Split each of the d variables into two or more segments.* Thus, variable j will have c_j segments, with $j = 1, \dots, d$;
2. *Cross the segmented variables* to obtain a cross-classification with $\prod_j c_j$ sectors;
3. *Calculate the probability of each sector.* Each of the $\prod_j c_j$ sectors has the boundaries x_i^1 and x_{i+1}^1 on the first variable, x_j^2 and x_{j+1}^2 on the second variable, ..., and x_l^d and x_{l+1}^d on the d th variable, where the subscripts indicate the segments and the superscripts indicate the variables. The probability of being located in the sector with these boundaries is

$$p(z_i^1 - z_{i+1}^1, z_j^2 - z_{j+1}^2, \dots, z_l^d - z_{l+1}^d) = \int_{z_i^1}^{z_{i+1}^1} \int_{z_j^2}^{z_{j+1}^2} \dots \int_{z_l^d}^{z_{l+1}^d} \Psi(z^1, z^2, \dots, z^d) dz^1 dz^2 \dots dz^d$$

A numerical solution for this expression was proposed by Genz (1992).

1. *Estimate expected sector frequencies* as $e_{i,j,\dots,l} = N p_{i,j,\dots,l}$, where $p_{i,j,\dots,l}$ is a shorthand for the expression given under (3).
2. *Perform sector-specific tests.* The null hypothesis for each of the tests is $E[o_{i,j,\dots,l}] = e_{i,j,\dots,l}$, where $o_{i,j,\dots,l}$ is the observed frequency for the sector with subscripts i,j,\dots,l . The tests can be performed using most of the residual tests from log-linear modeling or Configural Frequency Analysis (von Eye, 2002). These include the Pearson X^2 -component test, $X_{i,j,\dots,l} = (o_{i,j,\dots,l} - e_{i,j,\dots,l})^2 / e_{i,j,\dots,l}$. Because of the possibly existing dependency among tests, α -protection is advisable.

Based on the sector-specific test, an omnibus test can be devised by summing up the X^2 -components, as

$$X^2 = \sum_{i,j,\dots,l} \frac{(o_{i,j,\dots,l} - e_{i,j,\dots,l})^2}{e_{i,j,\dots,l}}.$$

This statistic can be used to test whether, overall, the cross-classification of segments follows a multinormal distribution. The test is approximately distributed as χ^2 with $(\prod_{j=1}^d c_j) - 2d - d_{cov} - 1$ degrees of freedom, where c_j is the number of segments of the j th variable. The term d_{cov} indicates the number of covariances taken into account, typically $d_{cov} = \binom{d}{2}$. That is, typically all covariances are taken into account.

5. Data example

For the following data example, we use data from the first wave of a longitudinal study examining domestic violence (defined as male violence toward their female partners) on women and their children. When the women were pregnant, they were given a structured interview (the Working Model of the Child Interview; Zeanah, Benoit, Hirshberg, Barton, & Regan, 1994) to assess their perceptions and subjective experiences of their unborn children. The interviews were audio taped, transcribed, and coded on various 5-point Likert scales. The two variables analyzed here - fear and joy - represent affective features of maternal representations (for more detail see Huth-Bocks, Levendosky, Theran, & Bogat, 2004).

One aspect of the study included the fitting of structural models with these features of the mothers' perceptions as indicators of latent variables. We now ask whether these two variables are jointly normally distributed, a precondition for the modeling approach. At the univariate level, neither of the two variables shows undue skewness or kurtosis. Table 1 displays the results.

We now ask whether the joint bivariate distribution of fear and joy is normally distributed. To answer this question, we employ the two Mardia tests, and the sector-wise and the χ^2 omnibus tests. The Mardia suggest that neither skewness nor kurtosis are problematic (skewness = 0.21, $p(\text{skewness}) = 0.14$; kurtosis = 8.40, $p(\text{kurtosis}) = 0.24$). We thus conclude that two null hypotheses that are compatible with multinormality can be retained. We now ask whether the sector and the omnibus tests suggest different conclusions. Table 2 displays the results of the sector tests.

	FEAR	JOY
N of cases	201	201
Minimum	1.000	1.000
Maximum	5.000	5.000
Mean	2.741	2.796
Standard Dev	0.673	1.016
Skewness(G1)	0.062	0.275
SE Skewness	0.172	0.172
Kurtosis(G2)	0.238	-0.585
SE Kurtosis	0.341	0.341

Table 1: Univariate analysis of the mother variables Fear and Joy

Sector	o	e	X^2	p	
11	26	16.97	4.811	0.028	
12	22	22.18	0.001	0.970	
13	21	23.01	0.175	0.676	
21	50	19.91	45.475	0.000	*
22	33	23.36	3.978	0.046	
23	29	21.5	2.62	0.106	
31	12	21.45	4.165	0.041	
32	6	22.47	12.073	0.001	*
33	2	18.37	14.59	0.000	*

Table 2: Sector tests for the Fear and Joy data

The sector tests suggest that, with α Bonferroni-protected, in Sector 2 1, there are more cases, and in Sectors 3 2 and 3 3, there are fewer cases than expected based on the assumption of binormality (the correlation of 0.10 between the two variables was taken into account). Researchers therefore have to ask why this is the case. If justifiable, resampling may be an option. In accordance with the results of the sector tests, the omnibus Pearson $X^2 = 87.89$ suggests significant deviations from binormality ($df = 7; p < 0.05$).

6. Conclusion

This research suggests new tests of multinormality. These tests are sensitive to characteristics of multinormality that Mardia's tests of skewness and kurtosis are not. Specifically, the sector tests indicate exactly, where in the distribution there are more or fewer cases than predicted based on multinormality.

References

- Bentler, P.M. (1992). *EQS. Structural Equations Program Manual*. Los Angeles, CA: BMDP Statistical Software.
- Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1, 141 - 149.
- Glass, G.V., & Hopkins, K.D. (1984). *Statistical methods in education and psychology*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall.
- Hu, L., & Bentler, P.M. (1995). Evaluating model fit. In R.H. Hoyle (ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 76 - 99). Thousand Oaks: Sage.
- Huth-Bocks, A. C., Levendosky, A. A., Theran, S. A., & Bogat, G. A. (2004). The impact of domestic violence on mothers' prenatal representations of their infants. *Infant Mental Health Journal*, 25, 79–98.
- Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519 - 530.

Mardia, K.V. (1980). Tests of univariate and multivariate normality. In P.R. Krishnaiah (ed.), *Handbook of statistics* (vol. 1; pp 279 - 320). Amsterdam: North Holland.

von Eye, A. (2002). *Configural Frequency Analysis: methods, models, and applications*. Mahwah, NJ: Lawrence Erlbaum.

Zeanah, C.H., Benoit, D., Hirshberg, L., Barton, M.L., Regan, C. (1994). Mothers' representations of their infants are concordant with infant attachment classifications. *Developmental Issues in Psychiatry and Psychology*, 1, 9–18.

von Eye, A., & Bogat, G.A. (2004). Testing the assumption of multivariate normality. *Psychology Science*, 46, 243 - 258.

von Eye, A., & Gardiner, J.C. (2004). Locating deviations from multivariate normality. *Understanding Statistics*. (in press)

Esta publicación consta de 877 ejemplares y se terminó de imprimir en el mes de agosto de 2005 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