Matemática: Una publicación del FCNM– ESPOL
2017, Vol.15, No. 2 **Publicado: 2017-10-02**

ANÁLISIS COMPARATIVO DE REGRESIÓN ORTOGONAL COMO ALTERNATIVA A LA REGRESIÓN ORDINARIA

ORTOGONAL COMPARATIVE REGRESSION ANALYSIS AS ALTERNATIVE TO REGULAR REGRESSION

Colon Mario Celleri Mujica[1]

Resumen: *El objetivo de este trabajo es comparar los dos procedimientos de regresión, ordinaria y ortogonal o también llamada total, y mostrar, que de existir variables regresoras no fijas, el método de regresión ordinaria no es el adecuado para determinar los valores de la variable de respuesta a partir de las variables regresoras, más bien, el método de mínimos cuadrados totales o regresión ortogonal es el más indicado.*

Palabras Claves: Regresión Ortogonal, Regresión Lineal, Mínimos Cuadrados totales,

Abstract: *The objective of this work is to compare the two regression procedures, ordinary and orthogonal or also called total, and to show that if there are non-fixed regressor variables, the ordinary regression method is not adequate to determine the values of the response variable from the regressor variables, rather, the method of total least squares or orthogonal regression is the most indicated..*

Keywords: Orthogonal Regression, Linear Regression, Total Square Minima.

Recibido: Enero 2017
Aceptado: Septiembre 2017

INTRODUCCIÓN

El método de regresión lineal es un método que permite ajustar una recta, la "mejor" recta, a una colección de puntos. El término "mejor" se refiere a que se minimiza la suma total de los errores o suma de los errores individuales. El método de regresión más comúnmente utilizado es el llamado de los "mínimos cuadrados" (Gillard, 2006), el cual consiste en medir la distancia existente entre el punto que corresponde a la observación y el punto sobre la recta (o subespacio) hipotético, aquella recta (o subespacio) que produzca la mínima suma es la elegida.

Posteriormente en 1877 Adcock (Gillard, 2006), utiliza un método de ajuste que se conoce con el nombre de "regresión ortogonal", modelo con error en las variables" o también llamado con "errores en la medición", pero que en la actualidad también se conoce con el nombre de método de los "mínimos cuadrados totales".

[1] *Colon Celleri, MSc., Profesor de la Escuela Superior Politécnica del Litoral (ESPOL); (e-mail: ccelleri@espol.edu.ec).*

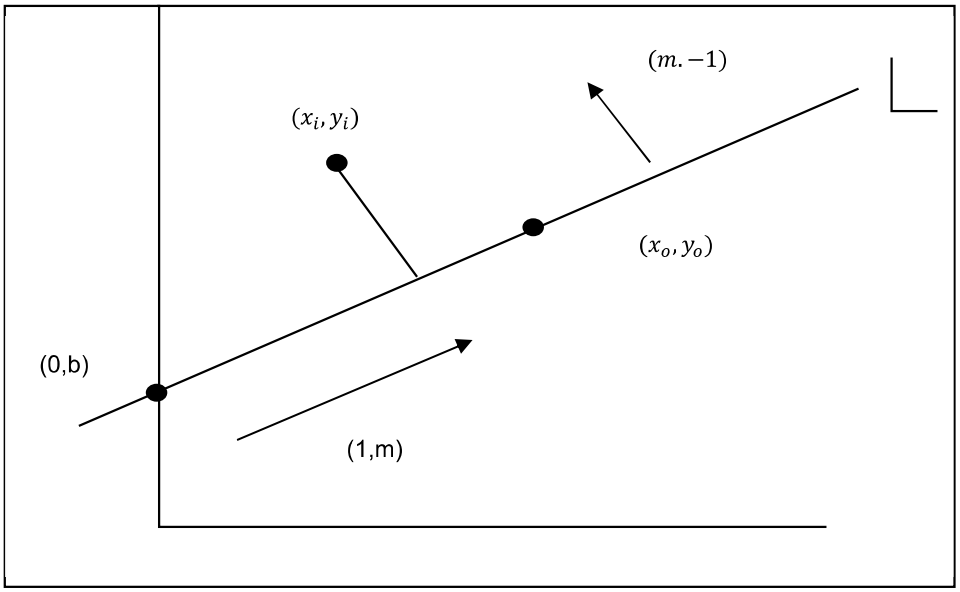
Más recientemente, el método de mínimos cuadrados totales también ha despertado el interés fuera de las estadísticas. En el campo del análisis numérico, este problema se estudió por primera vez por Golub y Van Loan (Golub & Van Loan, 1980). Su análisis, así como su algoritmo, se basa en el procedimiento de descomposición de valores singulares. Una visión geométrica de las propiedades de la descomposición de valores singulares se presentó por Staar (Golub & Van Loan, 1980) de forma independiente. El papel clave del método de los mínimos cuadrados en el análisis de regresión lineal es el mismo que el del método de los mínimos cuadrados totales en los modelos de error en variables (Golub & Van Loan, 1980).

Sin embargo, hay una gran cantidad de confusión en los campos del análisis numérico y estadísticas sobre el principio de mínimos cuadrados totales y su relación con los modelos con error en variables. Las ventajas computacionales de los algoritmos de los mínimos cuadrados totales aún no se conocen en la comunidad estadística, mientras que el concepto de modelado con error en variables no penetra lo suficientemente en el campo de las matemáticas computacionales e ingeniería. Una descripción general dedicada a los mínimos cuadrados totales se presentan en los textos de Van Huffel y Vanderwalle (Van Huffel, 2004).

MÉTODO DE REGRESIÓN ORTOGONAL (MÍNIMOS CUADRADOS TOTALES)

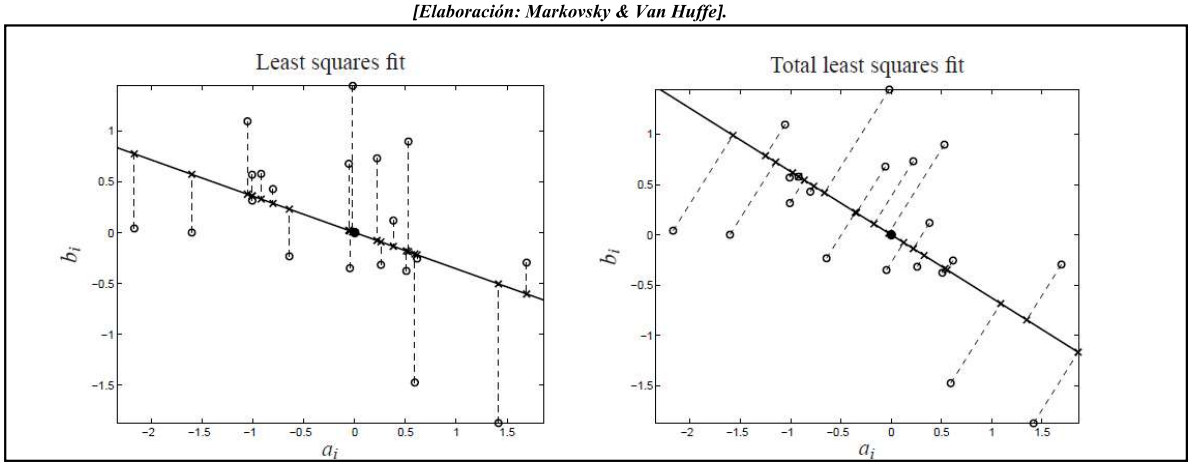
En el método de regresión ortogonal se cambia la definición del error (distancia desde el punto observado a la recta de regresión), de tal modo que la distancia se mide de manera perpendicular del punto a la recta estimada.

FIGURA 1.
Distancia en la regresión ortogonal. [Fuente: Casella & otros: Statistical Inference].[Elaboración: autor].



En el método de regresión ortogonal (véase figura 1) se cambia la definición del error (distancia desde el punto observado a la recta de regresión), de tal modo que la distancia se mide de manera perpendicular del punto a la recta estimada.

FIGURA 2
Comparación entre mínimos cuadrados y mínimos cuadrados totales. [Fuente: Overview of total least squares methods]. [Elaboración: Markovsky & Van Huffe].



Así la suma de los errores se expresa como:

$$S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(\frac{(mx_i - y_i + b)^2}{m^2 + 1} \right)$$

$$= \frac{1}{m^2 + 1} \sum_{i=1}^{n} (mx_i - y_1 + b)^2$$

Como se observa una expresión que representa una ecuación cuadrática en términos de la variable *m* y cuya solución es:

$$m = \frac{(S_{yy} - S_{xx}) \pm \sqrt{\left[-(S_{yy} - S_{xx})\right]^2 - 4S_{xy}(-S_{xy})}}{2S_{xy}}$$

$$m = \frac{S_{yy} - S_{xx} \pm \sqrt{\left(S_{yy} - S_{xx}\right)^2 + 4\, S_{xy}^2}}{2S_{xy}}$$

ANÁLISIS DEL MÉTODO DE MÍNIMOS CUADRADOS TOTALES (TLS) MULTIVARIABLE

Aunque el nombre de mínimos cuadrados totales fue dado por Golub, (1980), el método originalmente fue interpretado como un procedimiento de resolución numérica, pero en el campo de la estadística es conocido como regresión ortogonal, método para error en variables e incluso método para error en la medida.

El método de mínimos cuadrados totales es una de las varias técnicas de estimación mediante el uso de parámetros lineales en el que se pretenden compensar los efectos de errores en las variables.

El algoritmo de resolución que se presenta en este trabajo está basado en la descomposición de valores singulares (Singular Value Decomposition) (SVD) (Van Huffel, 2004) en sus siglas inglesas, el cual es el más aceptado por su eficiencia, versatilidad y robustez. Desde un punto de vista ingenieril el método de mínimos cuadrados totales se considera una aplicación orientada o dirigida a casi la mayoría de aplicaciones ingenieriles donde los datos se ven contaminados por ruido.
De aquí es más sencillo observar la idea detrás del método de mínimos cuadrados totales, ahora hay que considerar las perturbaciones tanto en el vector de respuesta b, como en la matriz de datos A, esto es:

$$\text{Minimizar } \left\|\left[\tilde{E}|\underline{\tilde{r}}\right]\right\|_F$$
$$\text{E, } \underline{r}$$

$$\text{Sujeto } a\text{: } \quad \underline{\boldsymbol{b}} + \underline{\boldsymbol{r}} \in \ Imagen\ (A + E)$$

Una vez que $\left[\tilde{E}|\underline{\tilde{r}}\right]$ ha sido encontrado y es el valor mínimo, entonces cualquier valor x que satisfaga:

$$(A + \tilde{E})\underline{\boldsymbol{x}} = \underline{\boldsymbol{b}} + \underline{\tilde{\boldsymbol{r}}}$$

Se dice que es una solución del problema de mínimos cuadrados totales. Con $\left[\hat{A}|\underline{\hat{\boldsymbol{b}}}\right] = U\hat{\Sigma}V^T$ siendo $\hat{\Sigma} = diag(\sigma_1, \dots \sigma_n, 0)$ y la solución del problema de mínimos cuadrados totales, existe, es única y viene dada por:

$$\underline{\hat{x}} = -\frac{1}{v_{n+1,n+1}}\left[v_{1,n+1,} \ \dots v_{p,p+1}\right]^T$$

A continuación se muestra el algoritmo para la resolución básica del problema de mínimos cuadrados totales:

Algoritmo:

$$\text{Dado } A \in \mathbb{R}^{n \times p}; \underline{\boldsymbol{b}} \in \mathbb{R}^n$$

Paso 1: Calcule la descomposición de valores singulares de:

$$\left[A|\underline{\boldsymbol{b}}\right] = U\Sigma V^T$$

Paso 2: Si $v_{p+1,p+1} \neq 0$

Entonces:

$$\underline{\hat{\boldsymbol{x}}} = -\frac{1}{v_{p+1,p+1}}\left[v_{1,p+1,\dots}, v_{p,p+1}\right]^T$$

Se puede fácilmente observar que para el caso univariado $(p = 1)$, se obtiene:

$$\widehat{x_1} = -\frac{v_{12}}{v_{22}}$$

Y luego se obtiene el valor de $\widehat{x_o}$. (una vez obtenido $\widehat{x_1}$).

ANÁLISIS DE LA SIMULACIÓN Y SUS RESULTADOS

Para el análisis se supondrá el modelo en que se tiene una variable de respuesta *y* con dos variables regresoras z_1 y z_2, pero estas variables regresoras no son las realmente medidas sino más bien unas variables observadas x_1 y x_2, las cuales son variables aleatorias que representan a las variables originales una vez que se ha introducido el error, el cual se supone normal con media cero y varianza $\sigma_{z_1}^2$ y $\sigma_{z_2}^2$. Esto es:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \varepsilon_y$$

Dónde:

$$x_1 = z_1 + \varepsilon_1 \quad \text{y} \quad x_2 = z_2 + \varepsilon_2$$

Y además se supondrá que:

$$\begin{aligned} z_1 &\sim N\ \left(\mu_1, \sigma_{z_1}^2\right) \\ z_2 &\sim N\ \left(\mu_1, \sigma_{z_2}^2\right) \\ \varepsilon_1 &\sim N\ \left(0, \sigma_{\varepsilon_1}^2\right) \\ \varepsilon_2 &\sim N\ \left(0, \sigma_{\varepsilon_2}^2\right) \\ \varepsilon_y &\sim N\ \left(0, \sigma_y^2\right) \end{aligned}$$

Es decir, en forma matricial:

$$\begin{bmatrix} z_1 \\ z_2 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{z_1}^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{z_2}^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\varepsilon_1}^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\varepsilon_2}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_y^2 \end{bmatrix}\right)$$

Un resultado interesante que permite observar muy rápidamente el comportamiento de los estimadores de regresión lineal ordinarios y ortogonales es si se supone los siguientes parámetros: $\mu_1 = \mu_2 = 0\,; \beta_0 = 0\,, \beta_1 = \beta_2 = 1$, lo que resulta reemplazando en la expresión del valor esperado como:

$$E(y/(x_1, x_2)) = \frac{\sigma_{z_1}^2}{\sigma_{z_1}^2 + \sigma_{\varepsilon_1}^2}\ x_1 \ + \ \frac{\sigma_{z_2}^2}{\sigma_{z_2}^2 + \sigma_{\varepsilon_2}^2}\ x_2$$

CASO 1:

$$\boldsymbol{\mu_1 = 0, \mu_2 = 0, \sigma_{z_1}^2 = 1, \sigma_{z_2}^2 = 1, \sigma_{\varepsilon_1}^2 = 1, \sigma_{\varepsilon_2}^2 = 1}$$

$$\boldsymbol{\beta_0 = 0\,, \beta_1 = 1\,, \beta_2 = 1}$$

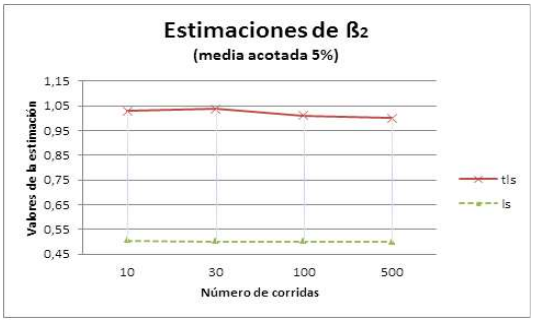
FIGURA 3
Caso 1
[Elaboración: Autor].

CAS0 1:
Nsimul = 10000

	MEDIA ARITMETICA β_2		% MEDIA	ERROR CUADRATICO MEDIO	
	TLS	**LS**	**ACOTADA**	**TLS**	**LS**
n = 10	1,03420	0,50016	3	3,89596	0,36754
	1,02877	0,50564	5	2,28191	0,35314
	0,99314	0,49335	8	1,40568	0,33689
n = 30	1,04854	0,49842	3	0,30301	0,27901
	1,03673	0,50229	5	0,23341	0,27082
	1,02658	0,49956	8	0,19602	0,26906
n = 100	1,00845	0,50021	3	0,04865	0,25742
	1,01051	0,50183	5	0,04386	0,25467
	1,01031	0,50091	8	0,04122	0,25423
n = 500	1,00111	0,50001	3	0,00891	0,25144
	0,99945	0,49941	5	0,00808	0,25179
	1,00091	0,49962	8	0,00754	0,25141

En este primer caso se puede apreciar que el estimador de regresión ortogonal presenta mejor característica en la estimación de los coeficientes de las variables regresoras, además se aprecia el sesgo que involucra a los estimadores de los coeficientes de regresión ordinaria, también se puede indicar que con pocos datos ya se puede hacer una estimación muy buena en regresión ortogonal.

FIGURA 4
Estimaciones de β_2.[Elaboración: Autor]



CONCLUSIONES

Uno de los principales resultados para un problema de dos variables es que en la regresión ortogonal no interesa cual variable sea considerada la regresora ya que se obtiene la misma recta de regresión, lo cual se mostró no ocurre en la regresión ordinaria.

Tal como se ha verificado en las tablas de resultados, los estimadores de regresión ortogonal resultaron ser más adecuados por su exactitud respecto a los valores que se obtuvieron por el método de regresión ordinaria.

A pesar que se utilizó pocos datos (n = 10), ya se tiene una buena aproximación a los resultados con los cuales se generaron los datos para la regresión. Aunque también en esta situación aparecen valores del error cuadrático medio mayores que cuando se utilizó los otros valores para n.

Si bien es cierto, al incrementar el valor de la varianza de la variable que representaría el ruido blanco (proceso estocástico de media nula) en las variables originales del modelo, ambos métodos se vieron afectados en su exactitud, el resultado de la regresión ortogonal sigue siendo de mejor rendimiento, aunque cuando se asignaron valores diferentes a la varianza de los elementos de error ambas regresiones resultaron inexactas en su estimación.

Se pudo también verificar que los sesgos que ocurren en la regresión ordinaria y que son el principal motivo para no usar dicho método se cumplieron tal y cual se obtuvo en la derivación teórica.

Así también los intervalos de confianza que se calcularon para ambos procedimientos, muestran el mejor comportamiento del método de regresión ortogonal a los que se han obtenido por el método de regresión ordinaria.

Por otro lado, hay algunos aspectos teóricos que aun necesitan revisarse y se augura hacerlo en otros trabajos, con lo cual se sugiere nuevos análisis especialmente en los casos en que no hubo mayor exactitud en los resultados.

Otros aspectos también a tomar en consideración, son las aplicaciones prácticas que se derivan de este trabajo. Es muy común la aplicación de la regresión ordinaria en diversos aspectos y campos de la ciencias sociales y aun en las ciencias duras, sin tomar en consideración los supuestos que deben cumplirse y respetarse, especialmente en las ciencias económicas donde se aplica indiscriminadamente la regresión ordinaria, la cual como se ha demostrado no obtiene los mejores resultados cuando el supuesto de la no aleatoriedad de las variables regresoras no se cumple.

Finalmente, en muchas aplicaciones de regresión, los datos no son obtenidos a partir de experimentos diseñados, sino que son observacionales, y no se ha fijado ninguna variable regresora. Si el interés es predecir una variable en términos de otras, la regresión ordinaria es aceptable. Pero si lo que interesa es la relación en sí, y ninguna de las variables involucradas es fijada de antemano, es conveniente utilizar regresión ortogonal.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. **Al-sharadqah, A., Chernov, N., & Huang, Q. (2011)**. *Errors-In-Variables regression and the problem of moments*. Recuperado el 2012 de Julio, de Brazilian Journal of Probability and Statistics: http://www.math.uab.edu/ chernov/cl

[2]. **Azarang, M. R., & García, E.** (1996). *Simulación y análisis de Modelos Estocásticos*. México: McGraw-Hill.

[3]. **Baker, K. (2005).** *Singular Value Decomposition Tutorial*. Recuperado el Agosto de 2012, de Ohio State University: http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf

[4]. **Blyth, T. S., & Robertson, E. F**. (2002). *Further Linear Algebra*. Londres: Springer.

[5]. **Boggs, P. T., & Rogers, J. E.** (1990). *Orthogonal Distance Regression*. Recuperado el Julio de 2012, de Center for Computing and Applied Mathematics: http://docs.scipy.org/doc/external/odr_ams.pdf

[6]. **Casella, G., & Berger, R. L**. (2002). *Statistical Inference.* Thomson Learning.

[7]. **Davidov, O. (2004).** *Estimating the slope in measurement error models—a different perspective.* Recuperado el 2012 de Julio, de Statistics & Probability Letters: www.elsevier.com/locate/stapro

[8]. **De Groen, P. (1996)**. *An Introduction to Total Least Squares.* Recuperado el Julio de 2012, de Vrije Universiteit Brussel, Department of Mathematics.: arXiv:math/9805076v1

[9]. **Di Ciccio, T., & Tibshirani, R.** (1991). *Department of Statistics, University of Toronto*. Recuperado el 10 de enero de 2014, de http://www.utstat.toronto.edu/wordpress/WSFiles/technicalreports/9107.pdf

[10]. **Dobson, A. J.** (2002). An *Introduction to Generalized Linear Models*. New York: Chapman & Hall/CRC.

[11]. **Eaton, M. L**. (2007). *Multivariate Statistics: A vector space approach.* Ohio: Institute of Mathematical Statistics.

[12]. **Freund, J. E., Miller, I., & Miller, M.** (2000). *Estadísticas matemáticas con aplicaciones*. México: Pearson Educación.

[13]. **Fuller, W. A.** (1987). *Measurement error models*. New York: JOHN WILEY & SONS.

[14]. **Gillard, J. W.** (2006). *A Historical Overview of Linear Regression with Errors with variables*. Recuperado el Julio de 2012, de School of Mathematics, Cardiff University: http://cardiff.ac.uk/maths/resources/Gillard_Tech_Report.pdf

[15]. **Gleser, L. J.** (1981). *Estimation in a multivariate "errors in variables" regression models: large sample results.* Recuperado el 2012 de Mayo, de Annals of Statistics: www.jstor.org

[16]. **Golub, G. H., & Van Loan, C. F.** (1980). *An analysis of the total least squares problem.* Recuperado el Mayo de 2012, de Society for Industrial and Applied Mathematics: http://www.cs.cornell.edu/cv/ResearchPDF/Analysis.total.least.squares.prob.pdf

[17]. **Harville, D. A.** (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

[18]. **Maddala, G. S.** (1996). *Introducción a la Econometría.* México: Prentice-Hall Hispanoamericana, S.A.

[19]. **Markovsky, I., & Van Huffel, S.** (2007). *Overview of total least squares methods.* Recuperado el Julio de 2012, de School of Electronics and Computer Science, University of Southampton.: http://eprints.soton.ac.uk/263855/1/tls_overview.pdf

[20]. **Martin, S. B.** (1998). *An alternative method of least squares linear regression.* Recuperado el Junio de 2012, de Mc Murry University: www.mcm.edu/mathdept/sarah.pdf

[21]. **Montoya, J. A.** (2008). *La verosimilitud perfil en la Inferencia Estadistica.* Guanajuato: CIMAT.

[22]. **Nievergelt, Y.** (1994). *Total least squares: State-of-the-art regression in numerical analysis*. Recuperado el Mayo de 2012, de Society for industrila and applied mathematics: http://people.duke.edu/~hpgavin/ce200/nievergelt94.pdf

[23]. **Noble, B., & Daniel, J. W**. (1989). *Algebra Lineal Aplicada.* México: Prentice-Hall Hispanoamericana, S.A.

[24]. **Otamendi, J.** (2006). *Las etapas en la gestación del método de Montecarlo*. En A.H.E.P.E, Historia de la Probabilidad y la Estadística (III) (págs. 117 - 123). Madrid: Publicaciones Delta.

[25]. **Peña, D.** (2002). *Análisis de datos multivariantes.* Madrid: McGraw Hill.

[26]. **Petras, I., & Podlubny, I.** (2010). *Least Squares or Least Circles? A comparison of classical regression and orthogonal regression*. Recuperado el Junio de 2012, de American Statistical Association: www.amstat.org/membership/index.cfm

[27]. **Petras, I., Bednárová, D., & Podlubny, I.** (2008). *Description of behavior of national economies in state space*. Recuperado el 2012 de Juliio, de Acta Montanistica Slovaca: http://actamont.tuke.sk/pdf/2008/n1/27petras.pdf

[28]. **Schaffrin, B., & Wieser, A.** (2010). *Total least-squares adjustment of condition equations.* Recuperado el Mayo de 2012, de Springer Link: http://link.springer.com/article/10.1007/s11200-011-0032-3#page-1

[29]. **Van Huffel, S.** (2004). *Total Least Squares and Errors-in-Variables Modeling:Bridging the gap between Statistics, Computational Mathematics and Engineering.* Recuperado el Mayo de 2012, de K.U.Leuven, Dept. of Electrical Engineering: www.esat.kuleuven.ac.be/sista.

[30]. **Van Huffel, S., & Zha, H.** (1993). *The total least squares problem.* Recuperado el Junio de 2012, de Handbook of Statistics: citeseer.uark.edu:8080/citeseerx/showciting

[31]. **Venables, W. N., & Ripley, B. D.** (2002). *Modern Applied Statistics with S.* Londres: Oxford.