38

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

METODOS NO PARAMETRICOS PARA ESTIMAR TRANSFORMACIONES EN REGRESION

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

OMAR GUTIERREZ ARREOLA

DIRECTOR DE TESIS:
DR. JOSE RODOLFO MENDOZA BLANCO

2001

UNIVERSIDAD NACIONAL AVTÓNOMA DE MEXICO

M. EN C. ELENA DE OTEYZA DE OTEYZA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo de Tesis:

"Métodos no paramétricos para estimar transformaciones en regresión"

realizado por Omar Gutiérrez Arreola

con número de cuenta 9036298-9 , pasante de la carrera de Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dr. José Rodolfo Mendoza Blanco	José Rodolfo Mendoza Blanco.
Propietario	M. en A.P. María del Pilar Alonso Reyes	
Propietario	M. en C. José Antonio Flores Díaz	
Suplente	M. en I. María del Carmen Hernández Ayuso	María del Carmen Hdez Ayuso
Suplente	Act. Jaime Vázquez Alamilla	

Consejo Departamental de Matemáticas

M. en C. José Antonio Flores Díaz

Agradecimientos

De manera general agradezco profundamente a todas las personas -- tanto a aquellas que me recibieron en este mundo como a las que se han incorporado en mi vida posteriormente; a las que continúan hoy a mi lado y a las que han partido, ambas habiendo formado lo que hoy constituye un preciado legado -- que de alguna manera han compartido su esfuerzo, felicidad, amor y paciencia, pues sin duda ustedes han sido los responsables de muchas de mis alegrías y satisfacciones.

Me considero además muy afortunado por los obstáculos que se han presentado en mi vida para poder llegar a este punto, y de que se me hayan otorgado la capacidad y las condiciones que permitieron sobrellevarlos, pues es de esta forma como los logros han adquirido un significado más especial.

Índice

Prólogo 2

Capítulo 1. Introducción a la Regresión
1.1 Análisis de Regresión 3
1.2 Transformaciones de Variables 9

Capítulo 2. Método ACE
(Esperanzas Condicionales Alternadas)
2.1 Introducción 19
2.2 Teorema de Proyección y Esperanza Condicional 21
2.3 Algoritmo ACE 31
2.4 Transformaciones Óptimas 34
2.5 Implantación del ACE en el caso discreto 41
2.6 Convergencia del ACE en el caso discreto 45
2.7 Ejemplos 49

Capítulo 3. Método AVAS
(Aditividad y Estabilización de Varianza)
3.1 Observaciones sobre el ACE 60
3.2 Modificaciones al Método ACE: Método AVAS 72
3.3 Normalización de Familias y Estabilización de Varianza 73
3.4 Algoritmo AVAS 80
3.5 Observaciones sobre el AVAS 82
3.6 Ejemplos 85

Conclusiones 94

Anexo 1. Observaciones sobre Matrices 100

Anexo 2. Tipos de Convergencia 101

Bibliografía 102

Prólogo

En años recientes se han desarrollado diversos trabajos referentes a la técnica de regresión, y más específicamente a la tarea de reemplazar las variables dependiente e independientes $Y, X_1, X_2, ..., X_p$ por transformaciones $\theta(Y), \phi_1(X_1), \phi_2(X_2), ..., \phi_p(X_p)$ tales que describan lo mejor posible un modelo de la forma $\theta(Y) = \phi_1(X_1) + \phi_2(X_2) + ... + \phi_p(X_p)$, contando únicamente con una serie de datos $\{(y_k, x_{k1}, x_{k2}, ..., x_{kp}), 1 \leq k \leq N\}$.

En este trabajo se describen dos métodos para la estimación no paramétrica de transformaciones en regresión conocidos como ACE (Esperanzas Condicionales Alternadas) y AVAS (Aditividad y Estabilización de Varianza), los cuales son más flexibles que procedimientos como el de Box-Cox, pues permiten casi cualquier tipo de transformación.

La distribución de la obra es la siguiente: después de mencionar el propósito, extensiones y cuidados en general de la técnica de regresión lineal en el primer capítulo, se prosigue en el segundo presentando el principio perseguido por el ACE: maximizar la correlación existente entre $\theta(Y)$ y $\sum_{i=1}^{p} \phi_i(X_i)$. Para ello se utilizan las propiedades de la esperanza condicional, las cuales son revisadas en seguida. Como paso siguiente se establece el algoritmo, se estudian las condiciones bajo las cuales existen transformaciones óptimas y cuándo el método converge a ellas. A continuación se discute su implantación en el caso discreto, mediante el uso de técnicas de suavización, así como su convergencia. Se concluye este tema mostrando una serie de ejemplos.

El capítulo final inicia ilustrando algunas críticas del ACE que motivaron el desarrollo de una variación del mismo. De esta manera se menciona la finalidad del denominado AVAS, el cual busca maximizar la correlación requiriendo estabilidad en la varianza del modelo transformado. Posteriormente se introduce la llamada transformación asintótica de varianza, la cual se incorpora en el ACE para obtener el nuevo método. Finalmente se desarrollan varios ejemplos para analizar su comportamiento y se efectúan comparaciones prácticas de ambos algoritmos.

Cabe mencionar que la parte teórica (salvo las bases del teorema de proyección y de la transformación asintótica de varianza) se basó en los trabajos originales de los algoritmos. Las demostraciones fueron simplificadas y detalladas en lo posible, incluyéndose las referencias para su consulta. Por último, para generar las gráficas se utilizó el paquete estadístico S-Plus, el cual incluye ambos algoritmos programados dentro de sus rutinas.

Capítulo 1

Introducción a la Regresión

1.1 Análisis de Regresión

El análisis de regresión es una técnica estadística cuyo propósito es identificar y/o modelar la relación existente entre las variables de un sistema. Una variable puede ser cualquier factor categórico o mensurable de éste. Esta libertad le otorga al análisis de regresión la cualidad de ser una herramienta sumamente útil en diversas ciencias, como la biología, la economía y las ciencias sociales entre otras, siendo actualmente uno de los principales métodos de análisis de datos. El objetivo de realizar una regresión es tratar de explicar el comportamiento de una variable (la variable de interés) con base en una o varias variables explicativas.

Antes de llevar a cabo un análisis es importante identificar el tipo de inferencias que desean obtenerse, pues tanto el proceso para la estimación como el modelo a utilizar dependen de éste. En primera instancia podría pensarse que aquel que mejor aproxime a los datos debería ser adoptado para cualquier propósito. Sin embargo, un modelo ofrece soluciones al aspecto particular que lo generó, mas puede no ser recomendable en otros casos.

Entre los usos más comunes del análisis de regresión se encuentran los siguientes:

- Proporciona un modelo que aproxima el comportamiento de la variable de interés de acuerdo a las condiciones de regresores, lo cual permite predecir el comportamiento de la primera conociendo el de las dependientes. Esto puede ser un resumen más útil que una tabla de datos o una gráfica, incluso. Por ejemplo, una compañía puede estimar la relación de la demanda de su producto con respecto a la presentada en meses anteriores, el sueldo mínimo vigente o cualquier otra variable, y mensualmente utilizar este modelo para prepararse para los niveles de demanda futuros basándose en las estimaciones.

- Puede ser un instrumento para comparar el grado de importancia de cada factor, y así explicar el comportamiento de la variable de interés. Supóngase que se quiere mejorar el sabor de cierta bebida comercial. El número de ingredientes de la bebida es demasiado numeroso para modificar cada ingrediente, además de que el sabor se vería alterado radicalmente. Por ello se lleva a cabo una prueba, utilizando diversas variantes de la bebida, alterando en cada una la combinación en algunos ingredientes y se estudia qué tanto mejoró el sabor. De esta forma se puede llegar a identificar los elementos que pueden contribuir a mejorarlo más que otros para pruebas posteriores.

- En ocasiones, cuando el tipo de relación que guardan las variables es conocido, el análisis representa una herramienta para estimar los parámetros desconocidos. Para ejemplificarlo tómese un caso de las finanzas. La volatilidad que una acción muestra (con respecto a un mercado en general) puede ser capturada en un coeficiente de sensibilidad. Este parámetro se obtiene al realizar una regresión de los rendimientos que ha presentado la acción con respecto a los rendimientos correspondientes del mercado. De esta forma, el valor obtenido proporciona una "representación porcentual promedio" de la sensibilidad de la acción a cambios en el entorno en general.

Ejemplo 1.1.1 *Un estudiante de medicina desea establecer una relación entre el peso y la estatura de los habitantes masculinos de una localidad. Para ello obtiene los datos de 15 personas al azar y los recopila en la siguiente tabla:*

Estatura(m)	Peso(kg)	Estatura(m)	Peso(kg)
1.77	75	1.84	88
1.68	73	1.63	65
1.75	67	1.62	78
1.68	73	1.76	79
1.62	72	1.73	81
1.87	83	1.75	78
1.65	76	1.85	74
1.67	80		

Tabla (1.1). Estaturas y pesos.

Realizando un ajuste por mínimos cuadrados se da cuenta que la relación $Peso = 17.1 + 34.2 \times estatura$ describe bien los datos tomados en principio, como lo detalla la gráfica siguiente.

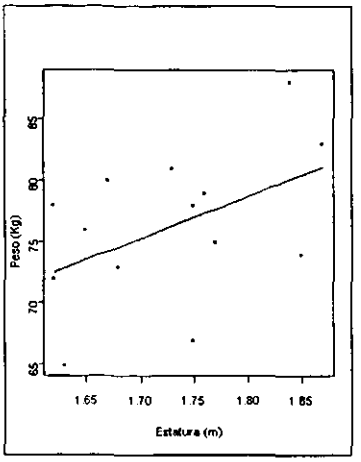

Fig. (1.1). Recta de mínimos cuadrados.

Algunos aspectos a considerar al llevar a cabo un análisis de regresión son mencionados a continuación:

- En general la ecuación que explica la relación entre las variables involucradas puede no ser válida fuera de la región que contiene los datos observados de las variables explicativas. Por ello si es necesario extrapolar valores debe verificarse si las condiciones existentes en el rango de datos son similares en la región donde se extrapolará.

- El análisis de regresión es parte de un amplio proceso para aproximar el comportamiento del sistema. Esto es, el obtener una ecuación puede no ser el objetivo principal del estudio llevado a cabo. El comprender el proceso que genera los datos puede considerarse más valioso.

- La veracidad del análisis depende directamente de la calidad de la información inicial. Si los datos en los que éste se basa son dudosos, las conclusiones lo serán aún más; más aún, deben contener los aspectos más importantes del aspecto a estudiar del sistema. Por ello, es importante el origen de la muestra, como un experimento controlado o una simulación, y estar concientes de su influencia en las conclusiones. Sin embargo, aún cuando se haya identificado la existencia de algún tipo de asociación, no siempre puede garantizarse que la muestra la haya captado de manera clara, independientemente de la forma de obtener los datos.

En el ejemplo expuesto se escogió un modelo para intentar explicar la relación existente entra las variables. Un modelo es lo que el analista percibe como el mecanismo que genera los datos, pudiendo establecer ciertas condiciones con respecto a ellos (supuestos). En este caso el modelo seleccionado tiene la forma

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \ i = 1, 2, ..., 15,$$

donde Y_i representa el peso y X_i la estatura del i-ésimo individuo. Los términos ε_i, $i = 1, 2, ..., 15$, corresponden a las variaciones de $\beta_0 + \beta_1 X_i$ con respecto a Y_i, $i = 1, 2, ..., 15$, y son tales que

$$\begin{aligned} E(\varepsilon_i) &= 0 \quad \forall i, \\ Cov(\varepsilon_i, \varepsilon_j) &= \begin{cases} 0 & \text{si } i \neq j, \\ \sigma^2 & \text{si } i = j. \end{cases} \end{aligned}$$

En este modelo la variable Y es llamada dependiente y X independiente o regresor. Puede ser conveniente pensar en ε como una variable aleatoria que representa el error estadístico, en el sentido de que contiene la falla del modelo de adecuarse a los datos de manera exacta. Finalmente β_0 y β_1 son cantidades desconocidas llamadas coeficientes de regresión. Generalmente los procedimientos llevados a cabo en un análisis buscan conclusiones con respecto a los éstos últimos, de acuerdo a criterios establecidos.

El modelo utilizado en el ejemplo es conocido como modelo de regresión lineal simple, y es el más sencillo debido a que contiene sólo un regresor. El término lineal se refiere a la linealidad en los coeficientes de cada variable, y no a la linealidad de cada variable en sí. Para formalizar lo anterior se expondrá el modelo de regresión lineal múltiple, del cual el simple es un caso particular.

1.1.1 Modelo de Regresión Lineal Múltiple

Considérese una relación de la forma:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, 2, ..., N.$$

Los datos son proporcionados por una muestra aleatoria $(Y_i, X_{i1}, X_{i2}, ..., X_{ip})$, $i = 1, 2, ...N$. Si se define

$$\epsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{1}_{N\times 1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \mathbf{X}_{(j)} = \begin{pmatrix} X_{1j} \\ \vdots \\ X_{Nj} \end{pmatrix}, \quad j = 1, 2, ...p,$$

$$\mathbf{X}_{N\times(p+1)} = (1, \mathbf{X}_{(1)}, ..., \mathbf{X}_{(p)}) = (X_{ij}),$$

entonces la forma matricial del modelo está dada por:

$$\mathbf{Y}_{N\times 1} = \mathbf{X}_{N\times(p+1)}\, \beta_{(p+1)\times 1} + \epsilon_{N\times 1}$$

bajo las siguientes hipótesis:

$$\begin{aligned} E(\varepsilon_i) &= 0 \quad \forall i, \\ Cov(\varepsilon_i, \varepsilon_j) &= \begin{cases} 0 & si \quad i \neq j, \\ \sigma^2 & si \quad i = j, \end{cases} ; \\ Rango(X) &= p. \end{aligned}$$

De esta forma se buscan los valores $\widehat{\beta}^t = (\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_p})$ que minimicen la suma de cuadrados (criterio establecido) $\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$, donde $\widehat{Y}_i = \widehat{\beta}^t \mathbf{X}_i$ y $\mathbf{X}_i^t = (1, X_{i1}, X_{i2}, ..., X_{ip})$.

Al estimador que minimiza la suma se le conoce como estimador de mínimos cuadrados, y es uno de los criterios más aceptados para considerar a los valores de los parámetros como "óptimos". Para llegar a él se utilizarán algunas definiciones y propiedades de matrices, las cuales se encuentran enunciadas en el Apéndice 1 al final de este trabajo.

El objetivo buscado es el siguiente: sea $\varepsilon_i = Y_i - \beta^t \mathbf{X}_i$. Se desea encontrar los valores de β que minimicen $\sum_{i=1}^{N} \varepsilon_i^2$.

Inicialmente se observa que

$$\begin{aligned} \sum_{i=1}^{N} \varepsilon_i^2 &= \epsilon^t \epsilon = (\mathbf{Y} - \mathbf{X}\beta)^t (\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}^t\mathbf{Y} - \beta^t\mathbf{X}^t\mathbf{Y} - \mathbf{Y}^t\mathbf{X}\beta + \beta^t\mathbf{X}^t\mathbf{X}\beta \\ &= \mathbf{Y}^t\mathbf{Y} - 2\mathbf{Y}^t\mathbf{X}\beta + \beta^t\mathbf{X}^t\mathbf{X}\beta. \end{aligned}$$

Ahora, como X^tX es simétrica, por las observaciones (A1.2) y (A1.3) se cumple

$$\frac{\partial \sum_{i=1}^{N} \varepsilon_i^2}{\partial \beta} = -2\mathbf{X}^t\mathbf{Y} + 2\mathbf{X}^t\mathbf{X}\beta.$$

Igualando a cero se obtiene

$$2\mathbf{X}^t\mathbf{X}\widehat{\beta} = 2\mathbf{X}^t\mathbf{Y}.$$

Como $rango(X^tX) = rango\,(X) = p$, se deduce que X^tX es invertible, por lo que de la última igualdad se tiene

$$\widehat{\beta} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}.$$

Para finalizar, al volver a derivar $\sum_{i=1}^{N} \varepsilon_i^2$ con respecto a $\widehat{\beta}^t$ se tiene

$$\frac{\partial^2 \sum_{i=1}^{N} \varepsilon_i^2}{\partial \widehat{\beta}^t \partial \widehat{\beta}} = 2\mathbf{X}^t\mathbf{X}$$

la cual es fácil mostrar que es definida positiva, por lo que la función alcanza un mínimo en $\widehat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$.

El proceso de estimar los parámetros del modelo de regresión no concluye la tarea del analista; representa el primer paso que puede llamarse el adecuar el modelo a los datos. La siguiente fase consiste en verificar si el grado de adecuación es aceptable, y determinar si son necesarios otros ajustes. De aquí que una regresión puede implicar un proceso iterativo, en el cual una muestra conduce a un modelo en primera instancia, y posteriormente éste se modifica hasta adecuarse a los datos satisfactoriamente.

Con el fin de poder modelar estadísticamente el comportamiento del error y efectuar pruebas de hipótesis, intervalos de confianza, etc., suelen agregarse supuestos, debiendo tenerse presente el papel trascendental que cualquiera de ellos juega en el análisis. Si éstos no se cumplen puede acarrear inestabilidad, en el sentido de que muestras distintas pueden llevar a modelos muy diferentes, con conclusiones contradictorias. De aquí la importancia del segundo paso mencionado anteriormente.

Entre los supuestos más examinados se encuentran los de simetría o normalidad de la distribución de los errores y la estabilidad de la varianza de los mismos. Dentro de las técnicas para verificar su validez, y del modelo en general, se encuentran el uso de estadísticos como el t o el F, el coeficiente de variabilidad explicada R^2, así como el análisis de residuales por medio de métodos gráficos como la "gráfica normal", entre otros. En caso que los resultados obtenidos indiquen que el modelo es inapropiado, se deberá considerar algún tipo de ajuste para mejorarlo.

1.2 Transformaciones de Variables

No obstante que la regresión lineal es una herramienta sumamente socorrida en un sinnúmero de situaciones, el alcance del mismo puede verse ampliado mediante la incorporación de procedimientos en los datos o evaluando si transformaciones de las variables pueden mejorar el poder explicativo del modelo. Aunque en diversas ocasiones éstas se escogen empíricamente, existen varias técnicas formales que pueden ayudar a especificar algún tipo en particular. A continuación se discutirán dos procedimientos analíticos para encontrar transformaciones en las variables del modelo.

1.2.1 Transformaciones en la Variable Dependiente

Considérese una relación entre las variables dependiente e independientes de la forma

$$Y^{\lambda} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon. \tag{1.1}$$

Claramente el modelo de regresión lineal múltiple es un caso particular, con $\lambda = 1$. Una simplificación del método de Box-Cox[1] sugiere encontrar el valor óptimo a través de una serie de regresiones que involucran varios valores del parámetro, comparando para tal efecto la suma de residuales $SR = \sum_{i=1}^{N} \{Y_i - \widehat{Y}_i(\lambda)\}^2$ para cada valor asignado a λ. De esta forma el valor óptimo será aquel cuya suma de residuales sea (aproximadamente) mínima. Este valor puede encontrarse analizando la gráfica de SR como función de λ. Como una segunda fase se puede refinar el área donde se percibe el mínimo de la suma de residuales para encontrar un valor más aproximado al mínimo.

El trabajo original de Box-Cox indica utilizar la función de máxima verosimilitud valuada para los distintos valores de λ e identificar gráficamente el valor máximo (como se mencionó antes), para posteriormente obtener una aproximación del intervalo de confianza.

Adicionalmente los autores presentaron resultados para transformaciones de la forma

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda \widetilde{Y}^{\lambda-1}}, & \lambda \neq 0. \\ \widetilde{Y} \ln Y, & \lambda = 0. \end{cases} \qquad \text{con } \widetilde{Y} = \exp(\frac{\sum Y_i}{N})$$

y

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(Y), & \lambda = 0 \end{cases}$$

entre otras.

[1] Box, G. E. P. y Cox, D. R. (1964). "An Analysis of Transformations". Journal of the Royal Statistical Society, 26.

Ejemplo 1.2.1 *Una compañía de electricidad desea obtener un modelo que refleje la relación entre el consumo de energía en la hora "pico" del día con el consumo total del cliente al mes. La importancia del estudio radica en que, si bien los consumidores pagan los servicios por el uso total, la planta debe ser capaz de satisfacer la demanda incluyendo la de las horas pico. Los datos se muestran en la tabla siguiente:*

X (kwh)	Y (kw)	X (kwh)	Y (kw)	X (kwh)	Y (kw)
679	.79	745	.77	770	1.74
292	.44	435	1.39	724	4.10
1012	.56	540	.56	808	3.94
493	.79	874	1.65	790	.96
582	2.70	1543	5.28	783	3.29
1156	3.64	1029	.64	406	.44
997	4.73	710	4.00	1242	3.24
2189	9.50	1434	.31	658	2.14
1097	5.34	837	4.20	1746	5.71
2078	6.85	1748	4.88	468	.64
1818	5.84	1381	3.48	1114	1.90
1700	5.21	1428	7.58	413	.51
747	3.25	1255	2.63	1787	8.33
2030	4.43	1777	4.99	3560	14.94
1643	3.16	370	.59	1495	5.11
414	.50	2316	8.19	2221	3.85
354	.17	1130	4.79	1526	3.93
1276	1.88	463	.51		

Tabla (1.2). Consumo de energía por hora y consumo total.

Al realizar una regresión lineal simple, el modelo obtenido por mínimos cuadrados está dado por

$$\widehat{Y} = -0.8283 + 0.00368X.$$

El ajuste que este modelo presenta con relación a los datos se ilustra enseguida.

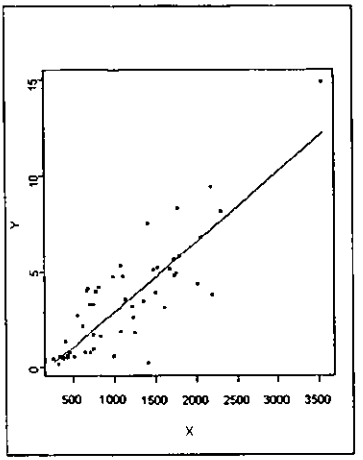

Fig. (1.2). Recta de mínimos cuadrados.

Los resultados de los estadísticos al realizar la regresión son $R^2 = 0.7046$ y $F = 160.26$ (significativo al 1%). Aunque no reflejan la necesidad de modificar el modelo, examinando la gráfica de residuales $e_i = \widehat{y_i} - y_i$ contra x_i se observa que el modelo no presenta estabilidad en la distribución de sus residuales. Por ello se decide verificar si una transformación en la variable explicativa es recomendable, mediante el proceso de Box-Cox.

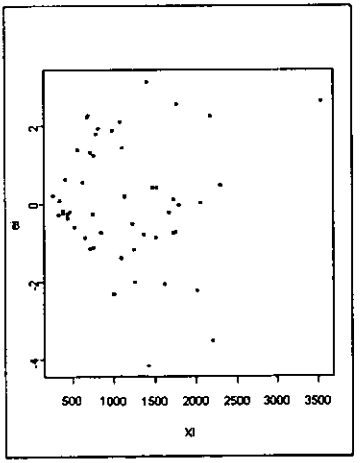

Fig. (1.3). Gráfica de residuales.

La tabla a continuación muestra la suma de residuales evaluada utilizando distintos valores de λ.

λ	SR(λ)
-2	34,101.03
-1	986.04
-0.5	291.58
0	134.09
0.5	96.94
1	126.86
2	1,275.55

Tabla (1.3).Valores de λ y SR.

Al parecer, el mínimo de la suma de residuales se encuentra entre los valores de 0 y 1 para λ. Por ello se hace un refinamiento en los incrementos, para precisar mejor un valor semióptimo. Los resultados son los siguientes:

λ	SR(λ)
0	134.09
0.125	118.19
0.25	107.20
0.375	100.2561
0.5	96.94
0.625	97.28
0.75	101.68
1	126.86

Tabla (1.4). Refinamiento de valores de λ y SR.

El mínimo de SR parece alcanzarse entre los valores de 0.375 y 0.625. Como este intervalo no contiene al valor 0, el método Box-Cox sugiere que una transformación de la variable dependiente es necesaria. De acuerdo a los valores obtenidos, la mejor estimación para el exponente es el valor de 0.5, o equivalentemente $\sqrt{Y}$. La gráfica siguiente ilustra el ajuste superior del modelo transformado, graficado utilizando rombos, al alcanzado por el modelo original, en puntos.

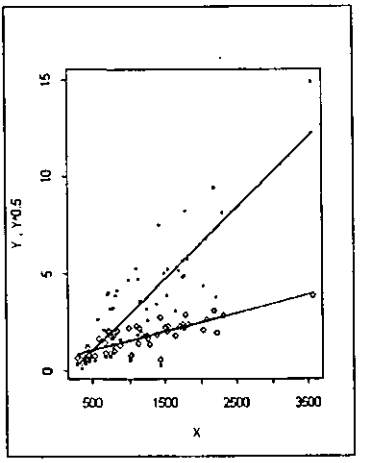

Fig. (1.4). Rectas de mínimos cuadrados para Y y $\sqrt{Y}$.

1.2.2 Transformaciones en las Variables Explicativas

Como complemento a (1.1) se puede pensar en un modelo de la forma:

$$Y = \beta_0 + \beta_1 X_1^{\alpha_1} + \beta_2 X_2^{\alpha_2} + ... + \beta_p X_p^{\alpha_p} + \varepsilon. \tag{1.2}$$

Box y Tidwell[2] proporcionaron un procedimiento para determinar el valor óptimo de los exponentes $\alpha_1, \alpha_2, ..., \alpha_p$, el cual se discutirá a continuación.

Definiendo

$$W_i = \begin{cases} X_i^{\alpha_i} & \text{si } \alpha_i \neq 0 \\ \ln(X_i) & \text{si } \alpha_i = 0 \end{cases} \qquad i = 1, 2, ..., p,$$

el modelo toma la forma:

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 ... + \beta_p W_p + \varepsilon.$$

[2]Box, G. E. P. y Tidwell, P. W. (1962). "Transformations of the Independent Variable". Technometrics, 4.

El método estima exponentes para uno o más regresores, basándose en la expansión de Taylor. El proceso requiere de un valor inicial para $\boldsymbol{\alpha} = (\alpha_{1,...,}\alpha_p)$ del cual se parte para mejorarlo gradualmente. Generalmente se selecciona $\hat{\alpha}_i = 1$, $i = 1, 2, ..., p$ para dicho valor, por lo cual se inicia con una regresión lineal múltiple. De esta forma, si se denota

$$f(\alpha_{1,...,}\alpha_p) = \beta_0 + \beta_1 W_1 + ... + \beta_p W_p,$$

la expansión de Taylor alrededor de $\boldsymbol{\alpha}_0 = (\alpha_{1,0}, ..., \alpha_{p,0})$, ignorando los términos superiores al primer orden, será:

$$E(Y) \cong [f(\alpha_{1,...,}\alpha_p)]_{\alpha=\alpha_0} + (\alpha_1 - \alpha_{1,0}) \left[\frac{\partial f}{\partial \alpha_1}\right]_{\alpha=\alpha_0} + \cdots + (\alpha_p - \alpha_{p,0}) \left[\frac{\partial f}{\partial \alpha_p}\right]_{\alpha=\alpha_0}.$$

Utilizando el valor $\boldsymbol{\alpha}_0 = 1$, la ecuación anterior adquiere la forma

$$E(Y) \cong \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + (\alpha_1 - 1)\beta_1 X_1 \ln X_1 + \cdots + (\alpha_p - 1)\beta_p X_p \ln X_p.$$

Tomando $\gamma_i = (\alpha_i - 1)\beta_i$, $Z_i = X_i \ln X_i$, $i = 1, 2, \ldots, p$, el modelo se reduce a:

$$E(Y) \cong \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \gamma_1 Z_1 + ... + \gamma_p Z_p. \tag{1.3}$$

De esta forma el proceso para estimar el valor de $\boldsymbol{\alpha}$ es el siguiente:

1. Se lleva a cabo una regresión lineal múltiple del modelo $E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$. Denótese $\widehat{\beta}_i$ al estimador de mínimos cuadrados de β_i.

2. Por separado se realiza una regresión lineal múltiple del modelo $E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \gamma_1 Z_1 + ... + \gamma_p Z_p$. Denótese $\widehat{\gamma}_i$ al estimador encontrado de γ_i.

3. Calcular los valores $\widehat{\alpha}_i = \frac{\widehat{\gamma}_i}{\widehat{\beta}_i} + 1, \quad i = 1, 2, \ldots, p.$

El papel del desarrollo de Taylor en el proceso anterior es el siguiente: (1.3) es un modelo de regresión lineal múltiple aumentado con términos en las Z. La parte extra corresponde a la desviación de la linealidad en las X del modelo original. Como resultado, valores de γ_i distintos a 0 sugieren la necesidad de un ajuste en el término X_i. Por el contrario, valores de γ_i cercanos a 0 indican que $\alpha_i - 1 \cong 0$, y por tanto no hay evidencia que sugiera ajustes. El proceso puede indicar transformaciones cuadráticas ($\alpha_i = 2$), logarítmicas ($\alpha_i = 0$), recíprocas ($\alpha_i = -1$), etc.

Lo anterior puede verse como una iteración que "mejora" el valor anterior de α, que inicialmente era 1. De esta forma, el procedimiento puede repetirse hasta que las nuevas estimaciones sean suficientemente parecidas a las anteriores. Para ello se puede utilizar el siguiente procedimiento de la segunda fase en adelante:

1. Hacer $\widehat{W}_i = x_i^{\widehat{\alpha}_i}, \quad i = 1, 2, ..., p$, y ajustar el modelo $E(Y) = \beta_0 + \beta_1 \widehat{W}_1 + ... + \beta_p \widehat{W}_p$. Denotar por $\widehat{\beta}_i$ al estimador de $\beta_i, \quad i = 1, 2, ..., p$.

2. Definir $\widehat{Z}_i = \widehat{W}_i \ln \widehat{W}_i$.

3. Ajustar el modelo $E(Y) = \beta_0 + \beta_1 \widehat{W}_1 + ... + \beta_p \widehat{W}_p + \gamma_1 \widehat{Z}_1 + ... + \gamma_p \widehat{Z}_p$. Denotar como $\widehat{\gamma}_i$ a los estimadores de γ_i.

4. Definir el nuevo valor de $\widehat{\alpha}_i$:

$$\widehat{\alpha}_i = \left(\frac{\widehat{\gamma}_i}{\widehat{\beta}_i} + 1 \right) \cdot (\textit{valor anterior de } \widehat{\alpha}_i), \quad i = 1, 2, \ldots, p. \tag{1.4}$$

Posteriormente se sigue el mismo proceso que en la primera iteración, asignando a $\hat{\alpha}_i$ el valor $\tilde{\alpha}_i \left(\frac{\widehat{\gamma}_i}{\widehat{\beta}_i} + 1 \right)$, donde $\tilde{\alpha}_i$ es la estimación del paso anterior y la expresión $\left(\frac{\widehat{\gamma}_i}{\widehat{\beta}_i} + 1 \right)$ se encuentra mediante una nueva regresión. De esta forma el coeficiente estimado de X_i después de la primera iteración será $X_i^{\tilde{\alpha}_i (\widehat{\gamma}_i / \widehat{\beta}_i + 1)}$ de acuerdo a (1.4).

Ejemplo 1.2.2 *Se lleva a cabo un estudio para estimar la relación existente entre la velocidad del viento en determinado tiempo con el voltaje que generan las aspas de un molino diseñado para tal efecto. Los datos recolectados se muestran en la tabla a continuación:*

Vel. viento (MPH)	Voltaje	Vel. viento (MPH)	Voltaje
5.00	1.582	5.80	1.737
6.00	1.822	7.40	2.088
3.40	1.057	3.60	1.137
2.70	0.500	7.85	2.179
10.00	2.236	8.80	2.112
9.70	2.386	7.00	1.800
9.55	2.294	5.45	1.501
3.05	0.558	9.10	2.303
8.15	2.166	10.20	2.310
6.20	1.866	4.10	1.194
2.90	0.653	3.95	1.144
6.35	1.930	2.45	0.123
4.60	1.562		

Tabla (1.5). Velocidad del viento y voltaje.

Al ajustarse una regresión lineal simple a los datos, se obtiene el modelo:

$$\widehat{Y} = 0.1309 + 0.2411X.$$

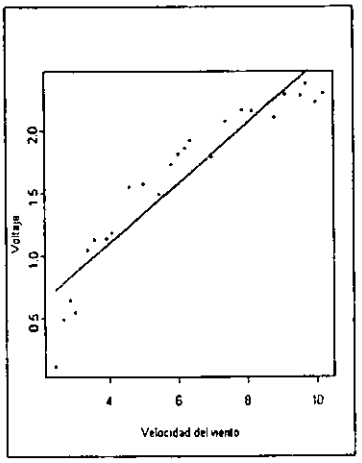

Fig. (1.5). Recta de mínimos cuadrados.

En este modelo $R^2 = 0.8745$, $SRC = .0557$ y $F_0 = 160.26$ (significativo al 1%). Revisando nuevamente la gráfica de residuales e_i contra los valores de X_i se detecta que el estado actual del modelo es inapropiado para satisfacer los supuestos de los errores, por lo que se lleva a cabo el proceso de Box-Tidwell para mejorarlo.

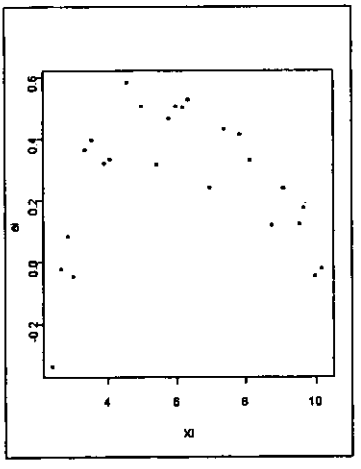

Fig. (1.6). Gráfica de residuales.

Ya que el primer paso del proceso es llevar a cabo una regresión lineal simple, se procede a mejorar el valor inicial $\hat{\alpha}_0 = 1$.

Defínase el valor $W = X \ln X$ y llevando a cabo una regresión para

$$Y = \beta_0 + \beta_1 X + \gamma W$$

el modelo estimado resulta

$$Y = -2.4168 + 1.5344X - 0.4626W.$$

De (1.4) se calcula

$$\hat{\alpha}_1 = \frac{\widehat{\gamma}}{\widehat{\beta}_1} + 1 = \frac{-0.4626}{0.2411} + 1 = -0.92$$

para el nuevo valor de $\hat{\alpha}$.

Una nueva iteración del procedimiento con la variable explicativa $\hat{W} = X^{-0.92}$ lleva a un modelo de la forma

$$Y = \beta_0 + \beta_1 \hat{W} = 3.1039 - 6.6784\hat{W}.$$

Incorporando el segundo regresor $\hat{Z} = \hat{W} \ln \hat{W}$ se obtiene

$$Y = \beta_0 + \beta_1 \hat{W} + \gamma \hat{Z} = 3.2409 - 6.445\hat{W} + 0.5994\hat{Z}.$$

La modificación de la estimación de α está dada entonces por:

$$\hat{\alpha}_2 = \left(\frac{\widehat{\gamma}}{\widehat{\beta}_1} + 1\right)\hat{\alpha}_1 = \left(\frac{0.5994}{-6.6784} + 1\right)(-0.92) = -0.8374.$$

Llevando a cabo un par de iteraciones más, los resultados obtenidos son $\hat{\alpha}_3 = -0.83342$ y $\hat{\alpha}_4 = -0.83334$. De esta forma, graficando los datos $X^{-0.83334}$ recomendado por el proceso de Box - Tidwell contra Y y la recta de mínimos cuadrados $Y = 3.26 - 6.4684X^{-0.83334}$ se puede observar una mejor adecuación del modelo.

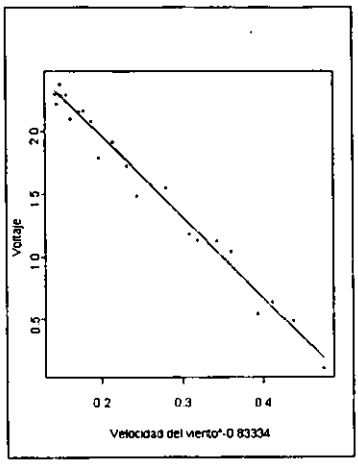

Fig. (1.7). Recta de mínimos cuadrados para $X^{-0.83334}$.

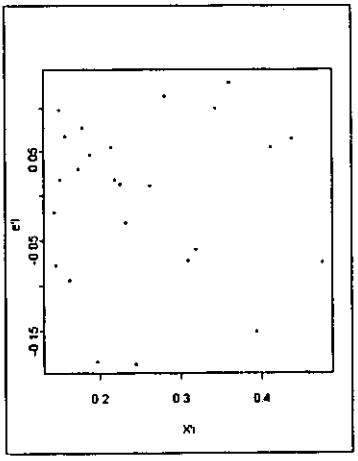

Fig. (1.8). Gráfica de residuales para $X^{-0.83334}$.

Como corolario, con los métodos anteriores puede estimarse una relación de la forma

$$Y^{\lambda} = \beta_0 + \beta_1 X_1^{\alpha_1} + \beta_2 X_{21}^{\alpha_2} + ... + \beta_p X_p^{\alpha_p} + \epsilon.$$

Una vez introducidos el concepto, utilidad, limitaciones y extensión de la regresión lineal mediante el uso de transformaciones, se presentarán dos métodos para estimar transformaciones generales, mediante los algoritmos ACE (el cual busca aquellas que maximicen la correlación del modelo) y AVAS (que provee además estabilidad en la varianza).

Capítulo 2

Método ACE

2.1 Introducción

En un análisis de regresión la variable dependiente Y y los regresores $X_1, ..., X_p$ frecuentemente pueden ser reemplazados por transformaciones $\theta(Y)$ y $\phi_1(X_1), ..., \phi_p(X_p)$, con la finalidad de encontrar el modelo aditivo que mejor aproxime a los datos. El método llamado ACE (Alternate Conditonal Expectations) es un procedimiento no paramétrico desarrollado por Leo Breiman y Jerome Friedman (1985), que tiene por finalidad estimar funciones "óptimas" (como se detalla más adelante) $\theta^*(Y)$ y $\phi_1^*(X_1), ..., \phi_p^*(X_p)$, utilizando una muestra de las variables $Y, X_1, ..., X_p$, imponiendo mínimas condiciones a la distribución de los datos, sin requerir que las transformaciones pertenezcan a alguna familia particular, ni aún que sean monótonas.

Este método puede ser aplicado en casos donde se incluyan mezclas arbitrarias de variables continuas y categóricas. $\theta(Y), \phi_1(X_1), ..., \phi_p(X_p)$ son funciones con codominio real, por lo que si la variable correspondiente es categórica se le asigna un valor real distinto a cada elemento de la imagen.

Para formalizar un poco, sean $Y, X_1, ..., X_p$ variables aleatorias con Y la variable dependiente y $X_1, ..., X_p$ las variables explicativas. Sean $\theta(Y), \phi_1(X_1), ..., \phi_p(X_p)$ funciones reales arbitrarias, medibles en sus respectivos espacios y con esperanza cero. Entonces el porcentaje de variabilidad no explicada por una regresión de $\theta(Y)$ en $\sum_{i=1}^{p} \phi_i(X_i)$ está dado por

$$e^2(\theta(Y), \phi_1(X_1), ..., \phi_p(X_p)) = \frac{E[\{\theta(Y) - \sum_{i=1}^{p} \phi_i(X_i)\}^2]}{E[\theta(Y)^2]}.$$

Se dice que las transformaciones $\theta^*(Y), \phi_1^*(X_1), ..., \phi_p^*(X_p)$ (o abreviando $\theta^*, \phi_1^*, ..., \phi_p^*$) son óptimas en regresión si

$$e^2(\theta^*, \phi_1^*, ..., \phi_p^*) = \min_{\theta,\phi_1,...,\phi_p} \{e^2(\theta, \phi_1, ..., \phi_p)\},$$

donde el mínimo se toma con respecto a todas las funciones medibles.

El método se basa en las propiedades de la esperanza condicional. Básicamente es un proceso iterativo que utiliza esperanzas condicionales bivariadas, independientemente del número de variables explicativas. Es posible demostrar que las funciones óptimas existen y que el algoritmo converge a ellas. En el caso práctico, donde las transformaciones se estiman de una serie de datos finita, el resultado es una estimación gráfica de las transformaciones.

En el caso univariado, las transformaciones óptimas $\theta^*(Y)$ y $\phi^*(X)$ deben satisfacer

$$\rho(\theta^*, \phi^*) = \max_{\theta,\phi} \{\rho(\theta, \phi)\},$$

donde ρ es el coeficiente de correlación entre las variables correspondientes. El valor de la correlación máxima es utilizado para medir el grado de dependencia entre variables. De esta forma, en el caso p=1 el ACE representa un proceso que permite conocer este valor, además de proveer una estimación de las transformaciones que alcanzan el máximo.

El método es aplicable a los siguientes casos:

1. Modelos aleatorios.

2. Serie de tiempo ergódicas estacionarias.

3. Diseños controlados.

En el primero, se asume que los datos $(y_k, \mathbf{x_k})$ $k = 1, 2, ..., n$ (con $\mathbf{x_k}$ un vector p-dimensional) provienen de muestras aleatorias independientes de las distribuciones de Y, X_1, $..., X_p$. En el segundo, los datos corresponden a una serie de tiempo ergódica estacionaria con esperanza cero. Esto es:

-La distribución de la i-ésima variable depende de las anteriores.

-Distribuciones idénticas independientemente del tiempo.

-$E(X_i) = 0, \forall i$.

En este caso las transformaciones óptimas son aquellas que minimizan

$$e^2 = \frac{E[\{\theta(X_{p+1}) - \sum_{i=1}^{p} \phi_i(X_i)\}^2]}{E[\theta(X_{p+1})^2]},$$

y la muestra consiste en n+p observaciones consecutivas $X_1, ..., X_{n+p}$. Para incluir este caso en la forma estándar del algoritmo se define a $Y_k = X_{k+p}$, $\mathbf{X}_k = (X_{k+p-1}, ..., X_k)$, $k = 1, 2, ..., n$. Finalmente, en el caso de diseños controlados, para cada vector $\mathbf{X} = (X_1, ..., X_p)$ se especifica una distribución $P(Y \mid \mathbf{X})$ para la variable dependiente en el espacio correspondiente. El "diseño" de n-ésimo orden consta de n puntos $\mathbf{x}_1, ..., \mathbf{x}_n$ junto con la variable dependiente correspondiente $Y_1, ..., Y_n$ obtenida de cada punto. Supóngase que los datos $\{Y_k\}$ son independientes, con Y_k tomado de la distribución $P(Y_k \mid \mathbf{X}_k)$. Sea $\widehat{P}_n(\mathbf{X})$ la distribución empírica que asigna el valor de $1/n$ a cada punto $\mathbf{X}_1, ..., \mathbf{X}_n$ (con $\mathbf{X}_i \neq \mathbf{X}_j$ si $i \neq j$). Adicionalmente, supóngase que $\widehat{P}_n \to P$, $n \to \infty$ donde $P(\mathbf{X})$ es la probabilidad en el espacio correspondiente. Entonces $P(Y \mid \mathbf{X})$ y $P(\mathbf{X})$ determinan la distribución conjunta de las variables aleatorias $Y, X_1, ..., X_p$ y las transformaciones óptimas cumplen la definición dada.

2.2 Teorema de Proyección y Esperanza Condicional

En esta sección se describen algunas propiedades e implicaciones de la esperanza condicional, las cuales son heredadas del llamado teorema de proyección. Para ello será necesario introducir algunas definiciones y proposiciones. Adicionalmente se citan varios tipos de convergencia, cuya definición se incluye en el Apéndice 2 al final de la obra.

Definición 2.2.1 *Sean Ω un conjunto y $\Im$ una familia de subconjuntos de Ω. Se dice que $\Im$ es una σ-álgebra de Ω si cumple:*

i) $\emptyset \in \Im$.
ii) $A \in \Im \Rightarrow A^c \in \Im$.
iii) $\{A_n\}_{n=1}^{\infty} \in \Im \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \Im$.

Entonces $(\Omega, \Im)$ forman un ***espacio medible****, y cualquier elemento de $\Im$ se denomina conjunto $\Im$-medible.*

Definición 2.2.2 *Una* ***medida*** *en una σ-álgebra $\Im$ es una función $\mu : \Im \to \overline{R} = [-\infty, +\infty]$ que cumple:*

i) $\mu(\emptyset) = 0$.
ii) $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} A_n$ *si* $A_i \bigcap A_j = \emptyset$, $\forall\, i \neq j$.

Entonces $(\Omega, \Im, \mu)$ forman un ***espacio de medida****. Si además μ cumple $\mu(A) \geq 0$ $\forall A \in \Im$, y $\mu(\Omega) = 1$, entonces μ es una* ***probabilidad*** *y $(\Omega, \Im, \mu)$ es un* ***espacio de probabilidad.***

Definición 2.2.3 *Sea Ω un espacio vectorial. Un producto interior es una función $\langle \cdot, \cdot \rangle : \Omega^2 \rightarrow \Re$ que satisface:*

i) $\langle x, y \rangle = \langle y, x \rangle \ \forall x, y \in \Omega$.
ii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle \ \forall x, y, z \in \Omega$.
iii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle \ \forall x, y \in \Omega \ y \ \alpha \in \Re$.
iv) $\langle x, x \rangle \geq 0 \ \forall x \in \Omega$.
v) $\langle x, x \rangle = 0 \Leftrightarrow x = 0$.

Entonces $(\Omega, \langle , \rangle)$ forman un ***espacio producto interior.***

Ejemplo 2.2.1 *Si $\Omega = \Re^n$, $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i$ corresponde al producto interior común en $\Re^n$.*

Ejemplo 2.2.2 *Sea $\Omega = \{f : \Re \rightarrow \Re : \int_{-\infty}^{\infty} f^2(x)dx < \infty\}$, con el producto interior definido por $\langle f, g \rangle = E[fg] = \int_{-\infty}^{\infty} f(x)g(x)dx$.*

Definición 2.2.4 *Una* ***norma*** *en un espacio vectorial Ω es una función $\| \cdot \| : \Omega \rightarrow \Re^+$ que satisface:*

i) $\|x + y\| \leq \|x\| + \|y\| \ \forall x, y \in \Omega$.
ii) $\|\alpha x\| = \mid \alpha \mid \|x\| \ \forall x \in \Omega, \alpha \in \Re$.
iii) $\|x\| \geq 0 \ \forall x \in \Omega$.
iv) $\|x\| = 0 \ \Leftrightarrow x = 0$.

Es fácil probar que cualquier producto interior $\langle \cdot, \cdot \rangle$ genera una norma si se define $\|x\| = \langle x, x \rangle^{1/2}$. En adelante se entenderá que la norma en un espacio producto interior es la inducida por el producto interior.

Proposición 2.2.1 *Una norma sobre un espacio producto interior $(\Omega, \langle , \rangle)$ cumple con las siguientes propiedades $\forall x_n, x, y \in \Omega$:*

i) $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$ *(Ley del paralelogramo).*
ii) *Si $\|x_n - x\| \rightarrow 0$ ent. $\|x_n\| \rightarrow \|x\|$* *(Continuidad).*
iii) $|\langle x, y \rangle| \leq \|x\| \ \|y\|$ *(Desigualdad de Cauchy-Schwarz),*

y la última igualdad se da si y sólo si x y y son linealmente dependientes.

Demostración:

i) Sean $x, y \in \Omega$. Entonces

$$\begin{aligned} \|x+y\|^2 + \|x-y\|^2 = \\ \langle x+y, x+y\rangle + \langle x-y, x-y\rangle = \\ \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle + \langle x,x\rangle - 2\langle x,y\rangle + \langle y,y\rangle = \\ 2\langle x,x\rangle + 2\langle y,y\rangle = 2\|x\|^2 + 2\|y\|^2. \end{aligned}$$

ii) Por la propiedad i) de la Definición (2.2.4) se tiene que $\|x_n\| \leq \|x_n - x\| + \|x\|$ y $\|x\| \leq \|x - x_n\| + \|x_n\|$. De aquí se tiene que $\|x_n - x\| \geq |\|x_n\| - \|x\||$, por lo que se sigue que $\|x_n\| \to \|x\|$.

iii) Suponiendo que x y y son linealmente independientes, entonces $\forall \lambda \in \Re$

$$0 < \langle x - \lambda y, x - \lambda y\rangle = \langle x,x\rangle - 2\lambda\langle x,y\rangle + \lambda^2\langle y,y\rangle .$$

Sea $\lambda = \langle x,y\rangle / \langle y,y\rangle$; entonces

$$0 < \langle x,x\rangle - \frac{2\langle x,y\rangle^2}{\langle y,y\rangle} + \frac{\langle x,y\rangle^2}{\langle y,y\rangle},$$

por lo que

$$|\langle x,y\rangle|^2 < \langle x,x\rangle\langle y,y\rangle ,$$

de donde se sigue la desigualdad.

Suponiendo ahora que x y y son linealmente dependientes, si $y = 0$ la igualdad es inmediata. Sea ahora $\lambda \neq 0$ tal que $x = \lambda y$.

$$|\langle x,y\rangle| = |\langle \lambda y, y\rangle| = |\lambda|\,|\langle y,y\rangle| = |\lambda|\,\|y\|\,\|y\| = \|\lambda y\|\,\|y\| = \|x\|\,\|y\| .\blacksquare$$

Definición 2.2.5 *Se dice que una sucesión $\{x_n\} \subseteq (\Omega, \langle,\rangle)$ es una sucesión de Cauchy si*

$$\|x_n - x_m\| \to 0,\ n, m \to \infty.$$

Definición 2.2.6 *Un espacio $(\Omega, \langle,\rangle)$ es un **espacio de Hilbert** si es completo, esto es, si toda sucesión de Cauchy $\{x_n\}$ converge en norma (ver Def. A2.2) a un elemento $x \in \Omega$.*

Ejemplo 2.2.3 *Retomando el Ejemplo (2.2.2), con la norma definida por $\|f\|^2 = E[f^2] = \int_{-\infty}^{\infty} f(x)^2 dx$, $(\Omega, \langle,\rangle)$ es un espacio completo y por tanto un espacio de Hilbert, el cual es conocido como espacio L^2.*

Proposición 2.2.2 *Sea* $\{x_n\} \to x$ *puntualmente (ver Def. A2.1), con* $x_n \in (\Omega, \langle,\rangle)$ *un espacio de Hilbert. Entonces* $\{x_n\}$ *es una sucesión de Cauchy.*

Demostración:

Sea $\varepsilon > 0$. Como $\{x_n\} \to x$, sea $n_{\varepsilon/2}$ tal que si $n' > n_{\varepsilon/2}$ entonces $\|x_{n'} - x\| < \varepsilon/2$. Por el inciso i) de la Definición (2.2.4) se observa que, para $n, m > n_{\varepsilon/2}$ se tiene que $\|x_n - x_m\| = \|x_n - x + x - x_m\| \leq \|x_n - x\| + \|x_m - x\| < \varepsilon$.■

Definición 2.2.7 *Sean* $x, y \in (\Omega, \langle,\rangle)$ *un espacio de Hilbert. Se dice que* x *es ortogonal a* y *(denotado* $x \perp y$*) si se cumple:*

$$\langle x, y\rangle = 0.$$

Definición 2.2.8 *Sean* $(\Omega, \langle,\rangle)$,*un espacio de Hilbert y* $\Gamma \subseteq \Omega$. *Se dice que* Γ *es cerrado si contiene a todos sus puntos límite, i.e.:*

$$x_n \in \Gamma \; y \;\; \| \, x_n - x \, \| \to 0 \;\; \Rightarrow x \;\in \Omega.$$

Definición 2.2.9 *Sea* $\Gamma \subseteq \Omega$, *con* $(\Omega, \langle,\rangle)$ *un espacio de producto interior. El complemento ortogonal de* Γ *(denotado* $\Gamma^{\perp}$*) es el conjunto de elementos de* Ω *que son ortogonales a cada elemento de* Γ, *i.e.:*

$$\Gamma^{\perp} = \{x \in \Omega : x \perp y \; \forall y \in \Gamma\} = \{x \in \Omega : \langle x, y\rangle = 0 \; \forall y \in \Gamma\}.$$

Proposición 2.2.3 *El complemento ortogonal de cualquier subespacio* $\Gamma \subseteq (\Omega, \langle,\rangle)$ *es un subespacio cerrado de* $(\Omega, \langle,\rangle)$.

Demostración:

Subespacio:

Como $\langle 0, x\rangle = 0 \; \forall x \in \Gamma$ se tiene que $0 \in \Gamma$.

Ahora, sea $x = \alpha_1 x_1 + \alpha_2 x_2$, con $\alpha_1, \alpha_2 \in \Re$ y $x_1, x_2 \in \Gamma^{\perp}$. Para toda $y \in \Gamma$ se tiene que

$$\langle x, y\rangle = \langle \alpha_1 x_1 + \alpha_2 x_2, y\rangle = \alpha_1 \langle x_1, y\rangle + \alpha_2 \langle x_2, y\rangle = 0 + 0 = 0.$$

Cerrado:

Si $x_n \in \Gamma^{\perp}$ y $\|x_n - x\| \to 0$, por el inciso iii) de la Proposición (2.2.1) se tiene que

$$\begin{aligned} 0 \;\; &\leq \;\; |\langle x_n - x, y\rangle| \leq \|x_n - x\| \, \|y\| \to 0 \\ &\Rightarrow \;\; \langle x_n - x, y\rangle = \langle x_n, y\rangle - \langle x, y\rangle = -\langle x, y\rangle \to 0 \\ &\Leftrightarrow \;\; \langle x, y\rangle = 0.\blacksquare \end{aligned}$$

Antes de establecer el teorema de proyección se ilustrará el uso del mismo en espacios de Hilbert particulares mediante los siguientes ejemplos.

Ejemplo 2.2.4 *Sean* $\boldsymbol{y}$ $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ *vectores en* $\Re^3$ *dados por:*

$$\begin{aligned} \mathbf{y} &= (\tfrac{2}{3}, 0, 1), \\ \mathbf{x_1} &= (1, 1, 0), \\ \mathbf{x_2} &= (1, -1, 0). \end{aligned}$$

El problema consiste en encontrar la combinación lineal de $\mathbf{x_1}$ *y* $\mathbf{x_2}$, $\widehat{\mathbf{y}} = \alpha_1\mathbf{x_1} + \alpha_2\mathbf{x_2}$, *más cercana a* $\mathbf{y}$; *es decir, encontrar el elemento* $\widehat{\mathbf{y}}$ *del subespacio generado por* $\mathbf{x_1}$ *y* $\mathbf{x_2}$ *que minimice la función* $S = \|\mathbf{y} - \alpha_1\mathbf{x_1} - \alpha_2\mathbf{x_2}\|^2$, *con* $\|\mathbf{z}\|^2 = \sum_{i=1}^{3} z_i^2$.

Un camino es utilizar cálculo para minimizar S. Sin embargo es posible resolverlo utilizando un enfoque geométrico: obsérvese que el vector buscado $\widehat{\mathbf{y}}$ cumple con que $(\mathbf{y} - \widehat{\mathbf{y}})$ es ortogonal al plano generado por $\mathbf{x_1}$ y $\mathbf{x_2}$, como lo muestra la figura siguiente:

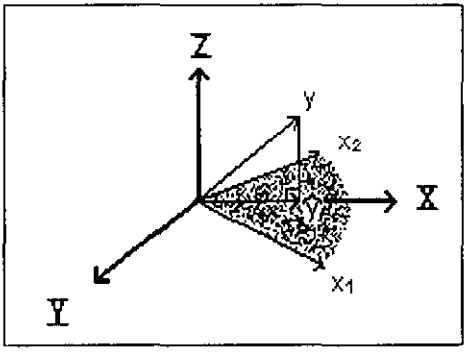

Fig. (2.1). La proyección ortogonal.

La condición de ortogonalidad puede expresarse como:

$$\langle \mathbf{y} - \alpha_1\mathbf{x_1} - \alpha_2\mathbf{x_2}, \mathbf{x_i} \rangle = 0, \; i = 1, 2,$$

o equivalentemente:

$$\begin{aligned} \alpha_1 \langle \mathbf{x_1}, \mathbf{x_1} \rangle + \alpha_2 \langle \mathbf{x_2}, \mathbf{x_1} \rangle &= \langle \mathbf{y}, \mathbf{x_1} \rangle, \\ \alpha_1 \langle \mathbf{x_1}, \mathbf{x_2} \rangle + \alpha_2 \langle \mathbf{x_2}, \mathbf{x_2} \rangle &= \langle \mathbf{y}, \mathbf{x_2} \rangle. \end{aligned}$$

Para los valores especificados de los vectores y de acuerdo a la definición de producto interior en $\Re^2$, las ecuaciones correspondientes son:

$$\begin{aligned} 2\alpha_1 &= \frac{2}{3}, \\ 2\alpha_2 &= \frac{2}{3}, \end{aligned}$$

por lo que los valores buscados son $\alpha_1 = \alpha_2 = \frac{1}{3}$, y por tanto $\widehat{\mathbf{y}} = (\frac{2}{3}, \frac{2}{3}, 0)$.

Ejemplo 2.2.5 *Sea $(\Omega, \Im, \mu)$ un espacio de probabilidad. Sean X_1, X_2 y Y variables aleatorias sobre $L^2(\Omega, \Im, \mu)$. Se desea estimar el valor de Y utilizando combinaciones lineales $\widehat{Y} = \alpha_1 X_1 + \alpha_2 X_2$ que minimicen:*

$$S = E[(Y - \alpha_1 X_1 - \alpha_2 X_2)^2] = \| Y - \alpha_1 X_1 - \alpha_2 X_2 \|^2 .$$

Nuevamente existen dos formas para resolver el problema. La primera es simplificar:

$$S = E[Y^2] + \alpha_1^2 E[X_1^2] + \alpha_2^2 E[X_2^2] - 2\alpha_1 E[Y X_1] - 2\alpha_2 E[Y X_2] + 2\alpha_1\alpha_2 E[X_1 X_2],$$

y minimizar S, igualando a cero las derivadas correspondientes. Por otra parte es posible emplear una analogía del ejemplo anterior; se desea encontrar un elemento $\widehat{Y}$ en el conjunto

$$\Gamma = \{X \in L^2(\Omega, \Im, \mu) : X = \alpha_1 X_1 + \alpha_2 X_2, \ con\ \alpha_1, \alpha_2 \in \Re\},$$

cuya distancia de Y, $\|Y - \widehat{Y}\|$, sea mínima. Al igual que en el ejemplo anterior, es claro que la variable $(Y - \widehat{Y})$ debe ser "ortogonal" a cualquier elemento de $X \in \Gamma$, utilizando el producto interior en el espacio para definir ortogonalidad. Esto es, $(Y - \widehat{Y})$ es ortogonal a X si se cumple:

$$\left\langle Y - \widehat{Y}, X \right\rangle = E[(Y - \widehat{Y})X] = 0 \ \forall X \in \Gamma.$$

Usando el hecho $X \in \Gamma$ y la linealidad del producto interior, la ecuación anterior puede expresarse como:

$$\langle Y - \alpha_1 X_1 - \alpha_2 X_2, X_i \rangle = 0, \ i = 1, 2.$$

Estas ecuaciones son las mismas que las derivadas previamente, aunque con diferente definición de producto interior. Con la definición del mismo en $L^2(\Omega, \Im, \mu)$ las ecuaciones toman la forma:

$$\begin{aligned} \alpha_1 E[X_1^2] + \alpha_2 E[X_2 X_1] &= E[Y X_1], \\ \alpha_1 E[X_1 X_2] + \alpha_2 E[X_2^2] &= E[Y X_2], \end{aligned}$$

de donde los valores α_1 y α_2 pueden encontrarse.

Teorema 2.2.1 *(Teorema de proyección). Sea Γ un subespacio cerrado de Ω, un espacio de Hilbert, donde $x \in \Omega$. Entonces:*

i) $\exists!\ \widehat{x} \in \Gamma \ \ni \| x - \widehat{x} \| = \inf_{y \in \Gamma} \| x - y \|$.

ii) Sea $\widehat{x} \in \Gamma$. *Entonces* $\| x - \widehat{x} \| = \inf_{y \in \Omega} \| x - y \| \ \Leftrightarrow \ (x - \widehat{x}) \in \Gamma^{\perp}$.

Al elemento $\widehat{x}$ se le conoce como la proyección (ortogonal) de x sobre Γ.

Demostración.

i) Sea $d = \inf_{y \in \Gamma} \| x - y \|^2$. Entonces existe una sucesión $\{y_n\}$ de elementos de $\Gamma \ \ni$ $\| y_n - x \|^2 \to d$.

Utilizando la propiedad de subespacio vectorial, se tiene que $(y_m + y_n)/2 \in \Gamma \ \forall n, m$. Ahora aplíquese la ley del paralelogramo para tener la igualdad:

$$0 \leq \| (y_m - x) + (y_n - x) \|^2 + \| (y_m - x) - (y_n - x) \|^2 = 2\{\| y_m - x \|^2 + \| y_n - x \|^2\}$$

Despejando al segundo sumando del lado izquierdo se tiene:

$$\begin{aligned} \| y_m - y_n \|^2 &= - \| (y_m - x) + (y_n - x) \|^2 + 2\{\| y_m - x \|^2 + \| y_n - x \|^2\} \\ &= - \| y_m + y_n - 2x \|^2 + 2\{\| y_m - x \|^2 + \| y_n - x \|^2\} \\ &= - \| 2\{(y_m + y_n)/2 - x\} \|^2 + 2\{\| y_m - x \|^2 + \| y_n - x \|^2\} \\ &= -4 \| \{(y_m + y_n)/2 - x\} \|^2 + 2\{\| y_m - x \|^2 + \| y_n - x \|^2\} \\ &\leq -4d + 2\{\| y_m - x \|^2 + \| y_n - x \|^2\} \\ &\to -4d + 2(d + d) = 0 \text{ cuando } n, m \to \infty. \end{aligned}$$

Por el criterio de Cauchy, $\exists\ \widehat{x} \in \Gamma$ tal que $\| y_n - \widehat{x} \| \to 0$. Como Γ es cerrado $\widehat{x} \in \Gamma$, y por la continuidad del producto interior se cumple que:

$$\| x - \widehat{x} \|^2 = \lim_{n \to \infty} \| x - y_n \|^2 = d.$$

Para probar la unicidad, sea $\widehat{y} \in \Gamma$ tal que $\| x - \widehat{y} \|^2 = \| x - \widehat{x} \|^2 = d$. Aplicando la ley del paralelogramo se obtiene:

$$\begin{aligned} 0 \ &\leq \ \| \widehat{x} - \widehat{y} \|^2 = -4 \| (\widehat{x} + \widehat{y})/2 - x \|^2 + 2\{\| \widehat{x} - x \|^2 + \| \widehat{y} - x \|^2\} \\ &\leq \ -4d + 4d = 0, \end{aligned}$$

y por las propiedades de la norma $\widehat{y} = \widehat{x}$.

ii) $\Leftarrow$) Si $\widehat{x} \in \Gamma$ y $(x - \widehat{x}) \in \Gamma^{\perp}$ entonces $\widehat{x}$ es el elemento único de Γ definido en i), ya que para cada $y \in \Gamma$ se cumple:

$$\begin{aligned} \| x - y \|^2 &= \langle x - \widehat{x} + \widehat{x} - y, x - \widehat{x} + \widehat{x} - y \rangle \\ &= \| x - \widehat{x} \|^2 + \| \widehat{x} - y \|^2 \\ &\geq \| x - \widehat{x} \|^2 . \end{aligned}$$

y la igualdad se cumple $\Leftrightarrow y = \widehat{x}$.

$\Rightarrow$) Sea $\widehat{x} \in \Gamma$ con $(x - \widehat{x}) \notin \Gamma^{\perp}$. Entonces $\widehat{x}$ no es el elemento de Γ más cercano a x, ya que $\widetilde{x} = \widehat{x} + ay/\|y\|^2$ (con y y a tales que $\langle x - \widehat{x}, y \rangle \neq 0, a = \langle x - \widehat{x}, y \rangle$) cumple:

$$\begin{aligned} \|x - \widetilde{x}\|^2 &= \langle x - \widehat{x} + \widehat{x} - \widetilde{x}, x - \widehat{x} + \widehat{x} - \widetilde{x} \rangle \\ &= \|x - \widehat{x}\|^2 + a^2/\|y\|^2 + 2\langle x - \widehat{x}, \widehat{x} - \widetilde{x} \rangle \\ &= \|x - \widehat{x}\|^2 - a^2/\|y\|^2 \\ &< \|x - \widehat{x}\|^2 . \blacksquare \end{aligned}$$

Corolario 2.2.1 *(La proyección de Ω en Γ). Si Γ es un subespacio cerrado de un espacio de Hilbert Ω e I es la función identidad en Ω, entonces existe una única función $P_{\Gamma} : \Omega \to \Gamma$ tal que $I - P_{\Gamma}$ mapea a Ω en $\Gamma^{\perp}$.*

Demostración.

Por el teorema de proyección se sabe que $\forall x \in \Omega \; \exists! \widehat{x} \in \Gamma \ni x - \widehat{x} \in \Gamma^{\perp}$.

La función P_{Γ} es la definida por:

$$P_{\Gamma}(x) = \widehat{x}. \blacksquare$$

Proposición 2.2.4 *Sean Ω un espacio de Hilbert y P_{Γ} la proyección de Ω en Γ, un subespacio cerrado de Ω. Sean $x, x_n, y \in \Omega$ y $\alpha, \beta \in \Re$. Entonces se cumple:*

i) $P_{\Gamma}(\alpha x + \beta y) = \alpha P_{\Gamma}(x) + \beta P_{\Gamma}(y)$.
ii) $\|x\|^2 = \|P_{\Gamma}(x)\|^2 + \|(I - P_{\Gamma})(x)\|^2$.
iii) $\forall x \in \Omega$ *existe una única representación como suma de un elemento en Γ y otro en $\Gamma^{\perp}$, dada por:*
$x = P_{\Gamma}(x) + (I - P_{\Gamma})(x)$.
iv) $\|x_n - x\| \to 0 \Rightarrow P_{\Gamma}(x_n) \to P_{\Gamma}(x)$.
v) $x \in \Gamma \Leftrightarrow P_{\Gamma}(x) = x$.
vi) $x \in \Gamma^{\perp} \Leftrightarrow P_{\Gamma}(x) = 0$.
vii) $\Gamma_1 \subseteq \Gamma_2 \Leftrightarrow P_{\Gamma_1} P_{\Gamma_2}(x) = P_{\Gamma_1}(x) \quad \forall x \in \Omega$.

Demostración.

i) Como Γ es un subespacio, $\alpha P_\Gamma(x) + \beta P_\Gamma(y) \in \Omega$. Además, como $\Gamma^\perp$ es un subespacio cerrado, se cumple:

$$\alpha x + \beta y - (\alpha P_\Gamma(x) + \beta P_\Gamma(y)) = \alpha(x - P_\Gamma(x)) + \beta(y - P_\Gamma(y)) \in \Gamma^\perp.$$

Por el Corolario (2.2.1), estas dos propiedades identifican a $\alpha P_\Gamma(x) + \beta P_\Gamma(y)$ como la proyección de $\alpha x + \beta y$ sobre Γ.

ii) Es consecuencia directa de la ortogonalidad de $P_\Gamma(x)$ y $(I - P_\Gamma)(x)$.

iii) Es claro que $P_\Gamma(x) + (I - P_\Gamma)(x) = P_\Gamma(x) + x - P_\Gamma(x) = x$.

Sean $y \in \Gamma$, $z \in \Gamma^\perp$ tales que $x = y + z$. Esto implica que:

$$\begin{aligned}
&\Rightarrow \quad P_\Gamma(x) + (I - P_\Gamma)(x) = y + z \\
&\Rightarrow \quad P_\Gamma(x) - y + (I - P_\Gamma)(x) - z = 0 \\
&\Rightarrow \quad \langle P_\Gamma(x) - y + (I - P_\Gamma)(x) - z, P_\Gamma(x) - y \rangle = 0 \\
&\Rightarrow \quad \langle P_\Gamma(x) - y, P_\Gamma(x) - y \rangle + \langle (I - P_\Gamma)(x) - z, P_\Gamma(x) - y \rangle = 0. \\
&\qquad \text{Como } (I - P_\Gamma)(x) - z \in \Gamma^\perp : \\
&\qquad \|P_\Gamma(x) - y\|^2 = 0 \\
&\Rightarrow \quad P_\Gamma(x) = y,\ (I - P_\Gamma)(x) = z.
\end{aligned}$$

iv) Se sigue de lo establecido en ii), pues $\|P_\Gamma(x_n - x)\|^2 \leq \|x_n - x\|^2 \to 0$ si $\|x_n - x\|^2 \to 0$.

v) y vi) son triviales.

vii) Se ha probado que:

$$\begin{aligned}
&x = P_{\Gamma_2}(x) + (I - P_{\Gamma_2})(x). \\
\Rightarrow \quad &P_{\Gamma_1}(x) = P_{\Gamma_1} \circ P_{\Gamma_2}(x) + P_{\Gamma_1} \circ (I - P_{\Gamma_2})(x). \\
&\text{Así:} \\
&P_{\Gamma_1}(x) = P_{\Gamma_1} \circ P_{\Gamma_2}(x) \\
\Leftrightarrow \quad &P_{\Gamma_1}(y) = 0, \forall y \in \Gamma_2^\perp \\
\Leftrightarrow \quad &\Gamma_1^\perp \subseteq \Gamma_2^\perp \\
\Leftrightarrow \quad &\Gamma_2 \subseteq \Gamma_1. \blacksquare
\end{aligned}$$

Más aún, el teorema de proyección establece que el elemento de Γ más cercano a x es el elemento único $\widehat{x}$ que cumple:

$$\langle x - \widehat{x}, y \rangle = 0 \ \forall y \in \Gamma. \tag{2.1}$$

Una vez que se cuenta con la teoría de espacios de Hilbert, es posible definir la esperanza condicional como un caso particular. En el resto de la sección Ω será el espacio real de Hilbert $L^2(\Omega, \Im, P)$ con el producto interior $\langle X, Y \rangle = E[XY]$.

Definición 2.2.10 *Sean $X_n, X \in L^2$. Entonces X_n converge a X en media cuadrática si:*

$$\lim_{n\to\infty} \|X_n - X\|^2 = \lim_{n\to\infty} E[(X_n - X)^2] = 0. \tag{2.2}$$

Reescribiendo las propiedades ya establecidas para la convergencia en norma al caso particular de convergencia en media cuadrática se tiene la siguiente

Proposición 2.2.5 *Sean $\{X_n\}, \{Y_n\}, X, Y \in L^2$. Entonces se cumple:*

i) X_n converge en media cuadrática $\Leftrightarrow E[(X_n - X_m)^2] \to 0$ conforme $n, m \to \infty$.

ii) Si $X_n \to X$ y $Y_n \to Y$, entonces conforme $n \to \infty$,

(a) $E[X_n] \to E[X]$,
(b) $E[X_n^2] \to E[X^2]$,
(c) $E[X_n Y_n] \to E[XY]$.

Definición 2.2.11 *Si Γ es un subespacio de L^2 y $Y \in L^2$ entonces la mejor predicción (con respecto a la media cuadrática) es el elemento $\widehat{Y} \in \Gamma$ que cumple:*

$$\left\|Y - \widehat{Y}\right\| = \inf_{Z\in\Gamma} \|Y - Z\| = \inf_{z\in\Gamma} E[(Y - Z)^2].$$

El teorema de proyección identifica a la mejor predicción de Y como $P_\Gamma(Y)$. Al incorporar más estructura al subespacio Γ, de (2.2) se obtiene el concepto de esperanza condicional.

Definición 2.2.12 *Sea Γ un subespacio cerrado de L^2 que contiene las funciones constantes. Para $X \in L^2$ se define la* ***esperanza condicional*** *de X dada Γ como la proyección*

$$E[X \mid \Gamma] = P_\Gamma(X).$$

Utilizando la definición de producto interior en L^2 y (2.1) es posible definir la esperanza condicional $E[X \mid \Gamma]$ como el único elemento de Γ que cumple:

$$E(Y \cdot E[X \mid \Gamma]) = E(YX) \ \forall Y \in \Gamma.$$

Proposición 2.2.6 *La esperanza condicional cumple con las propiedades:*

i) $E[\alpha X + \beta Y \mid \Gamma] = \alpha E[X \mid \Gamma] + \beta E[Y \mid \Gamma]$, $\alpha, \beta \in \Re$.

ii) $X_n \to X \Rightarrow E[X_n \mid \Gamma] \to E[X \mid \Gamma]$ *en media cuadrática.*

iii) $\Gamma_1 \subseteq \Gamma_2 \Rightarrow E(E[X \mid \Gamma_2] \mid \Gamma_1) = E[X \mid \Gamma_2]$.

iv) $E[1 \mid \Gamma] = 1$.

v) Si Γ_0 *es el subespacio cerrado compuesto por todas las funciones constantes entonces*

$$E[X \mid \Gamma_0] = E[X].$$

Demostración:

Se sigue de las propiedades de la Proposición (2.2.4).■

Definición 2.2.13 *Sea* Z *una variable aleatoria en un espacio de probabilidad* $(\Omega, \Im, P)$ *y* $X \in L^2(\Omega, \Im, P)$. *Entonces la esperanza condicional de* X *dado* Z *está definida por:*

$$E[X \mid Z] = E_{\Gamma(Z)}(X),$$

donde $\Gamma(Z)$ *es el subespacio cerrado de* L^2 *formado por todas las variables aleatorias en* L^2 *de la forma* $\phi(Z)$, *con* $\phi : \Re \to \Re$ *una función boreliana.*

No es difícil mostrar que $E[X \mid Z]$ cumple con las propiedades de $E[X \mid \Gamma]$.

El concepto anterior puede extenderse de manera natural al caso de varias variables: sean $\{Z_\lambda, \lambda \in \Lambda\}$ variables aleatorias en $(\Omega, \Im, P)$ y $X \in L^2$. Entonces se define

$$E[X \mid Z_\lambda, \lambda \in \Lambda] = E_{\Gamma(Z_\lambda, \lambda \in \Lambda)}(X),$$

donde $\Gamma(Z_\lambda, \lambda \in \Lambda)$ es el subespacio cerrado de L^2 formado por todas las variables de la forma $\phi(Z_1, Z_2, ..., Z_n)$, con $\phi : \Re^n \to \Re$ una función boreliana y $\Gamma(Z_\lambda, \lambda \in \Lambda)$ cumple con las propiedades anteriores.

Por el teorema de proyección, la esperanza condicional $E[X \mid Z_\lambda, \lambda \in \Lambda]$ es la función de $\{Z_\lambda, \lambda \in \Lambda\}$ que mejor aproxima a X en términos de media cuadrática.

2.3 Algoritmo ACE

El procedimiento para estimar las transformaciones $\theta^*(Y)$ y $\phi_i^*(X_i), i = 1, 2, ..., p$, se basa en un algoritmo iterativo, el cual se lleva a cabo hasta converger, considerándose entonces las distribuciones finales como óptimas. Como ya se ha mencionado, la función objetivo es minimizar el porcentaje de variabilidad no explicada entre una función cualquiera θ de la variable dependiente Y y una combinación lineal de transformaciones cualesquiera ϕ_i de las variables explicativas X_i, $i = 1, 2, ..., p$. Sin pérdida de generalidad se puede suponer que $E(\theta^2) = 1$ y que las transformaciones tienen media cero, para asegurar la unicidad de las funciones óptimas.

Supóngase inicialmente que las distribuciones de las variables Y, X_i son conocidas. Se introducirá primero el algoritmo en el caso p=1 y posteriormente el caso general.

En el caso univariado se tiene que:

$$e^2(\theta, \phi) = E[(\theta - \phi)^2]. \tag{2.3}$$

Para ϕ fija, por las propiedades de la esperanza condicional se sabe que el mínimo de (2.3) con respecto a θ se alcanza en

$$\theta_0 = E[\phi \mid Y].$$

Imponiendo la condición $E(\theta^2) = 1$, el mínimo se alcanza en

$$\theta_0 = \frac{E[\phi \mid Y]}{\|E[\phi \mid Y]\|}, \tag{2.4}$$

donde la definición de norma está dada por:

$$\begin{aligned} E(\cdot) &= \int_{-\infty}^{\infty} f(\cdot) d(\cdot), \\ \|(\cdot)\| &= \{E[\cdot^2]\}^{1/2}. \end{aligned}$$

Ahora considérese el problema de minimizar (2.3) con respecto a ϕ para θ fija. Análogamente la solución es

$$\phi_0 = E[\theta \mid X]. \tag{2.5}$$

(2.4) y (2.5) forman la base del proceso iterativo de optimización del método llamado Esperanzas Condicionales Alternadas (ACE). El algoritmo en el caso univariado es el siguiente:

Dadas dos variables aleatorias X y Y con distribuciones conocidas:

-Asignar $\theta(Y) = Y / \|Y\|$;

-Mientras $e^2(\theta, \phi)$ disminuya, defínase:
$\phi_0(X) = E[\theta(Y) \mid X]$;
Asignar $\phi(X) = \phi_0(X)$;
$\theta_0(Y) = E[\phi(X) \mid Y] / \|E[\phi(X) \mid Y]\|$;
Asignar $\theta(Y) = \theta_0(Y)$;

-θ y ϕ son las soluciones θ^*, ϕ^*;
-Fin del algoritmo.

En cada iteración del algoritmo, el valor de (2.3) disminuye, inicialmente manteniendo fija a θ y sustituyendo ϕ, y posteriormente fijando la nueva ϕ y sustituyendo θ. El ACE finaliza al no disminuir (2.3) después de una iteración. Para fines prácticos, tal condición puede modificarse por decrecer menos de cierta cantidad (por ejemplo 10^{-3}) en un número de iteraciones predefinido. Para asegurar que no se caerá en un bucle infinito, en secciones siguientes se estudiarán la existencia del mínimo y la convergencia del algoritmo.

Considerando ahora el caso multivariado, el objetivo es minimizar la función

$$e^2(\theta, \phi_1, ..., \phi_p) = E\left[\theta - \sum_{i=1}^{p} \phi_i\right] \tag{2.6}$$

con $E[\theta^2] = 1$ y $E[\theta] = E[\phi_i] = 0$. Utilizando el principio anterior, para un conjunto de ϕ_i, $i = 1, ..., p$ fijas, la solución a minimizar (2.6) con respecto a θ está dada por:

$$\theta_0(Y) = \frac{E\left[\sum_{i=1}^{p} \phi_i(X_i) \mid Y\right]}{\left\| E\left[\sum_{i=1}^{p} \phi_i(X_i) \mid Y\right] \right\|}. \tag{2.7}$$

El siguiente paso es minimizar (2.6) con respecto a $\phi_1, ..., \phi_p$. Para ello considérese fijas tanto a θ como a ϕ_j ($j \neq i$ para i fija). Entonces la solución de (2.6) con respecto a ϕ_i es:

$$\phi_{i,0}(X_i) = E\left[\theta(Y) - \sum\nolimits_{j \neq i} \phi_j(X) \mid X_i\right]. \tag{2.8}$$

De esta forma, el algoritmo general está dado por:

Dadas las distribuciones de $Y, X_1, X_2, ..., X_p$:

-Asignar $\theta(Y) = Y / \|Y\|$;

-Mientras $e^2(\theta, \phi_1, ..., \phi_p)$ disminuya:

-Para $i = 1, 2, ..., p$ obtener la transformación de X_i:
$\phi_{i,0}(X_i) = E[\theta(Y) - \sum_{j \neq i} \phi_j(X_j) \mid X_i)]$;
Asignar $\phi_i(X_i) = \phi_{i,0}(X_i)$;
-Obtener la transformación de Y:
$$\theta_0(Y) = E\left[\sum_{j=1}^{p} \phi_j \mid Y/\right] / \left\| E\left[\sum_{j=1}^{p} \phi_j \mid Y\right] \right\|;$$
Asignar $\theta(Y) = \theta_0(Y)$;

-$\theta, \phi_1, ..., \phi_p$ son las soluciones $\theta^*, \phi_1^*, ..., \phi_p^*$;
-Fin del algoritmo.

En el algoritmo anterior, cada iteración del bucle interior minimiza (2.8) con respecto a la función ϕ_i, $i = 1, 2, ..., p$, dejando fijas las demás con su valor de la iteración anterior (o inicial). La función del bucle exterior es minimizar (2.8) con respecto a θ, estableciendo como condición para detener el algoritmo que el valor de e^2 no decrezca después de una iteración completa, o que la variación tras varias iteraciones sea mínima.

2.4 Transformaciones Óptimas

El propósito de esta sección es identificar condiciones bajo las cuales existen soluciones óptimas al problema de maximizar la correlación en un modelo así como estudiar la unicidad, para en una sección posterior analizar la convergencia del algoritmo.

Definición 2.4.1 *Sean Y, $X_1, ..., X_p$ variables aleatorias con dominio real o un conjunto numerable. Una transformación de un modelo es un conjunto de funciones medibles $(\theta, \phi) = \{\theta(Y), \phi_1(X_1), ..., \phi_p(X_p)\}$ definidas en el rango de su variable correspondiente que cumplen:*

$$\begin{aligned} E[\theta(Y)] &= 0, \quad E[\phi_i(X_i)] = 0, \quad i = 1, 2, ..., p. \\ E[\theta^2(Y)] &= 1, \quad E[\phi_i^2(X_i)] < \infty, \quad i = 1, 2, ... p. \end{aligned} \tag{2.9}$$

En adelante se denotará:

$$\overline{\phi} = \overline{\phi}(\mathbf{X}) = \sum_{i=1}^{p} \phi_i(X_i).$$

Por último se entenderá por Θ al conjunto de todas las transformaciones.

Definición 2.4.2 *Una transformación (θ^*, ϕ^*) es óptima en regresión si $E[(\theta^*)^2] = 1$ y*

$$\begin{aligned} (e^*)^2 &= E[(\theta^* - \overline{\phi^*})^2] \\ &= \inf_{\Theta} \{E[(\theta - \overline{\phi})^2]; E[\theta^2] = 1\}. \end{aligned}$$

Definición 2.4.3 *Una transformación (θ^*, ϕ^*) es óptima en correlación si $E[(\theta^*)^2] = 1$, $E[(\overline{\phi^*})^2] = 1$ y*

$$\begin{aligned} \rho^* &= E[\theta^* \overline{\phi^*}] \\ &= \sup_{\Theta} \{E[\theta \overline{\phi}]; E[\theta^2] = 1, E[\overline{\phi}^2] = 1\}. \end{aligned}$$

Teorema 2.4.1 *(θ, ϕ) es óptima en regresión si y sólo si $(\theta, \frac{\phi}{\rho})$ es óptima en correlación. Adicionalmente $(e^*)^2 = 1 - (\rho^*)^2$.*

Demostración:

$$\begin{aligned} E[(\theta - \overline{\phi})^2] &= E[\theta^2] - 2E[\theta\overline{\phi}] + E[\overline{\phi}^2] \\ &= 1 - 2E[\theta\overline{\phi}] + E[\overline{\phi}^2]. \end{aligned}$$

Sea $\widehat{\phi} = \overline{\phi}/\sqrt{E[\overline{\phi}^2]}$, por lo que

$$\begin{aligned} E[(\theta - \overline{\phi})^2] &= 1 - 2E\left[\theta\widehat{\phi}\sqrt{E[\overline{\phi}^2]}\right] + E[\overline{\phi}^2] \\ &= 1 - 2\sqrt{E[\overline{\phi}^2]}E[\theta\widehat{\phi}] + E[\overline{\phi}^2] \\ &\geq 1 - 2\sqrt{E[\overline{\phi}^2]}\rho^* + E[\overline{\phi}^2], \end{aligned}$$

y se cumple la igualdad si y sólo si $E[\theta\widehat{\phi}] = \rho^*$. Además, el mínimo del lado derecho de la desigualdad con respecto a $E[\overline{\phi}^2]$ se alcanza en $E[\overline{\phi}^2] = (\rho^*)^2$, donde se cumple $(e^*)^2 = E[(\theta - \overline{\phi})^2] = 1 - (\rho^*)^2$. Por lo tanto $E[(\theta - \overline{\phi})^2]$ alcanza el ínfimo sobre Θ si y sólo si $2E[\theta\overline{\phi}]$ alcanza el supremo sobre Θ.■

El paso siguiente será mostrar la existencia de transformaciones óptimas. Para ello se necesitarán suponer las condiciones descritas a continuación.

Supuesto 2.4.1 *Sean $\theta(Y), \phi_i(X_i)$, $i = 1, 2, ..., p$, que cumplan (2.9). Entonces $\theta(Y) + \sum_{i=1}^{p} \phi_i(X_i) = 0$ c.s. $\Leftrightarrow \theta(Y) = 0$ c.s., $\phi_i(X_i) = 0$ c.s., $i = 1, 2, ..., p$, donde "c.s." denota convergencia casi segura (ver Definición A2.3).*

Definición 2.4.4 *Denótese al espacio de Hilbert $H_2(Y)$, como el conjunto formado por las funciones medibles $\theta(Y)$ que satisfacen (2.9) y con el producto interior usual. Análogamente defínase los espacios $H_2(X_1), ..., H_2(X_p)$.*

Supuesto 2.4.2 *Los operadores de esperanza condicional*

$$\begin{aligned} E[\theta(Y) \mid X_i] &: H_2(Y) \to H_2(X_i), \\ E[\phi_j(X_j) \mid X_i] &: H_2(X_j) \to H_2(X_i),\ i \neq j, \\ E[\phi_i(X_i) \mid Y] &: H_2(X_i) \to H_2(Y) \end{aligned}$$

son compactos.

Los autores aseguran que el supuesto anterior se satisface en la mayoría de los casos de interés. Asimismo, presentan una condición suficiente para ello[1]:

Sean X, Y variables aleatorias con densidad conjunta f_{XY} y densidades marginales f_X, f_Y. Entonces una condición suficiente para que el operador de esperanza condicional $H_2(Y) \to H_2(X)$ sea compacto es que:

$$\int\int \left[\frac{f_{XY}^2}{f_X f_Y}\right] dxdy < \infty.$$

Para poder probar el teorema de existencia de las transformaciones óptimas se procederá de la siguiente forma:

1. Se verificará que el conjunto de transformaciones de la forma $\{\theta(Y), \bar{\phi}(\mathbf{X})\}$ con el producto interior $\langle g, f\rangle = E[fg]$ es un espacio de Hilbert, y los conjuntos de transformaciones $\theta(Y), \bar{\phi}(\mathbf{X})$ y $\phi_i(X_i)$ son subespacios cerrados.
2. Se mostrará que una sucesión de transformaciones de $\{Y_n, \bar{\phi}_n(\mathbf{X})\}_{n\in\mathbb{N}}$ converge si y sólo si las correspondientes sucesiones $\{Y_n\}_{n\in\mathbb{N}}$ y $\{\bar{\phi}_n(\mathbf{X})\}_{n\in\mathbb{N}}$ convergen.
3. Con lo anterior se probará que la esperanza condicional de Y dadas $X_1, ..., X_p$ y la esperanza de $X_1, ..., X_p$ dada Y son compactas.

Proposición 2.4.1 *El conjunto de transformaciones de la forma*

$$f(Y, \mathbf{X}) = \theta(Y) + \sum_{i=1}^{p} \phi_i(X_i), \quad \theta \in H_2(Y), \quad \phi_i \in H_2(X_i), \quad i = 1, 2, ..., p,$$

con el producto interior y norma:

$$\begin{aligned} \langle g, f\rangle &= E[gf], \\ \|f\|^2 &= E[f^2], \end{aligned}$$

es un espacio de Hilbert, denotado por H_2. El conjunto de transforaciones $\bar{\phi}$ de la forma

$$\bar{\phi}(\mathbf{X}) = \sum_{i-1}^{p} \phi_i(X_i), \quad \phi_i \in H_2(X_i), \tag{2.10}$$

es un subespacio cerrado de $H_2(\mathbf{X})$. De igual forma los subespacios $H_2(Y)$, $H_2(X_1), ...,$ $H_2(X_p)$.

Demostración:

[1]Breiman, L. y Freidman, J. H. (1985). "Estimating Optimal Transformations for Multiple Regression and Correlation". Journal of the American Statistical Association, 80, pág. 590.

Esta proposición se sigue de la siguiente

Proposición 2.4.2 *Bajo los Supuestos (2.4.1) y (2.4.2), existen constantes* $0 < c_1 \leq c_2 < \infty$ *que cumplen*

$$c_1 \left(\|\theta(Y)\|^2 + \sum_{i=1}^{p} \|\phi_i(X_i)\|^2 \right) \leq \left\| \theta(Y) + \sum_{i=1}^{p} \phi_i(X_i) \right\|^2 \leq c_2 \left(\|\theta(Y)\|^2 + \sum_{i=1}^{p} \|\phi_i(X_i)\|^2 \right).$$

Demostración[2]*:*

La desigualdad de la derecha es inmediata.

Suponiendo que la desigualdad de la izquierda no se cumple, entonces existe una sucesión $f_n = \theta_n + \sum_{i=1}^{p} \phi_{ni}$ tal que $\|\theta_n\|^2 + \sum_{i=1}^{p} \|\phi_{ni}\|^2 = 1$ pero $\|f_n\|^2 \to 0$. Además existe una subsucesión $f_{n'}$ tal que $\theta_{n'} \to \theta$ débilmente en $H_2(Y)$ y $\phi_{n'i} \to \phi_i$ débilmente en $H_2(X_i)$, $i = 1, 2, ..., p$.

Haciendo

$$E\left[\phi_{n'i}(X_i)\phi_{n'j}(X_j)\right] = E\left[\phi_{n'i}(X_i)E\left[\phi_{n'j}(X_j) \mid X_i\right]\right], \quad i \neq j,$$

se observa que el Supuesto (2.4.2) implica que $E\left[\phi_{n'i}(X_i)\phi_{n'j}(X_j)\right] \to E\left[\phi_i(X_i)\phi_j(X_j)\right]$, y análogamete para $E\left[\theta_{n'}\phi_{n'j}\right]$.

Adicionalmente $\|\theta\| \leq \liminf \|\theta_{n'}\|^2$ y $\|\phi_i\| \leq \liminf \|\phi_{n'i}\|^2$. Definiendo $f = \theta + \sum_{i=1}^{p} \phi_i$ se tiene que

$$\|f\|^2 = \left\| \theta + \sum_{i=1}^{p} \phi_i \right\|^2 \leq \liminf \|f_{n'}\|^2 = 0,$$

por lo que el Supuesto (2.4.1) implica que $\theta = \phi_1 = \phi_2 = ... = \phi_p = 0$. Por otra parte,

$$\|f_{n'}\|^2 = \|\theta_{n'}\|^2 + \sum_{i=1}^{p} \|\phi_{n'i}\|^2 + 2\sum_{i=1}^{p} \langle \theta_{n'}, \phi_{n'i} \rangle + 2\sum_{i \neq j} \langle \phi_{n'i}, \phi_{n'j} \rangle.$$

Esto último implica que si $f = 0$ entonces $\liminf \|f_{n'}\| \geq 1$!■

Corolario 2.4.1 $f_n \to f$ *débilmente en* H_2 *si y sólo si* $\theta_n \to \theta$ *débilmente en* $H_2(Y)$ *y* $\phi_{n,i} \to \phi_i$ *débilmente en* $H_2(X_i)$ $i = 1, 2, ..., p$.

Demostración[3]*:*

[2] *Ibidim*, pág. 590.
[3] *Ibidim*, pág. 590.

Suponiendo que $f_n = \theta_n + \sum_{i=1}^{p} \phi_{ni} \rightarrow f = \theta + \sum_{i=1}^{p} \phi_i$, entonces por la Proposición (2.4.2) se tiene que $\limsup \|\theta_n\| < \infty$ y $\limsup \|\phi_{ni}\| < \infty$. Es posible entonces tomar una subsucesión $f_{n'}$ tal que $\theta_{n'} \rightarrow \theta'$ para alguna θ' y $\phi_{n'i} \rightarrow \phi_i'$ para algunas ϕ_i', $i = 1, 2, ..., p$. Haciendo $f' = \theta' + \sum_{i=1}^{p} \phi_i'$ se tiene que para toda $g \in H_2$ se cumple $\langle g, f_n' \rangle \rightarrow \langle g, f' \rangle$ débilmente, por lo que $\langle g, f' \rangle = \langle g, f \rangle$ para toda g.

La implicación contraria es obvia.■

Definición 2.4.5 *Denótese a los operadores de proyección:*

$$P_Y : H_2 \rightarrow H_2(Y) = E[\theta(Y) + \sum_{i=1}^{p} \phi_i(X_i) \mid Y],$$

$$P_{X_i} : H_2 \rightarrow H_2(X_i) = E[\theta(Y) + \sum_{i=1}^{p} \phi_i(X_i) \mid X_i],$$

$$P_{\mathbf{X}} : H_2 \rightarrow H_2(\mathbf{X}) = E[\theta(Y) + \sum_{i=1}^{p} \phi_i(X_i) \mid X_1, ..., X_p].$$

En $H_2(X_j)$, P_{X_i} $(i \neq j)$ es el operador de esperanza condicional. Análogamente para P_Y.

Proposición 2.4.3 *El operador P_Y es compacto en $H_2(\mathbf{X}) \rightarrow H_2(Y)$ y $P_{\mathbf{X}}$ es compacto en $H_2(Y)$ y $H_2(\mathbf{X})$.*

Demostración[4]*:*

Sea $\overline{\phi_n} \in H_2(\mathbf{X})$, $\overline{\phi_n} \rightarrow \overline{\phi}$ débilmente. Por el Corolario (2.4.1) se sabe que $\phi_{n,i} \rightarrow \phi_i$ débilmente. Por el Supuesto (2.4.2), $P_Y \phi_{n,i} \rightarrow P_Y \phi_i$, por lo que $P_Y \overline{\phi_n} \rightarrow P_Y \overline{\phi}$.

Ahora sean $\theta \in H_2(Y)$, $\overline{\phi} \in H_2(\mathbf{X})$; entonces $\langle \theta, P_Y \bar{\phi} \rangle = \langle \theta, \bar{\phi} \rangle = \langle P_{\mathbf{X}} \theta, \bar{\phi} \rangle$. Por consiguiente $P_{\mathbf{X}} : H_2(Y) \rightarrow H_2(\mathbf{X})$ es el operador adjunto de P_Y y por lo tanto compacto.■

Se está ahora en posibilidades de presentar el teorema de existencia.

Teorema 2.4.2 *Bajo los Supuestos (2.4.1) y (2.4.2), existen transformaciones óptimas $(\theta^*, \overline{\phi}^*)$.*

Demostración:

[4] *Ibidim*, pág. 591.

Por el teorema de proyección se sabe que para toda transformación $(\theta, \overline{\phi})$ con $\|\theta\|^2 = 1$ se satisface:

$$\left\|\theta - \overline{\phi}\right\|^2 \geq \|\theta - P_{\mathbf{X}}(\theta)\|^2 .$$

Si $\|\theta^* - P_{\mathbf{X}}(\theta^*)\|^2 = \min\{\|\theta - P_{\mathbf{X}}(\theta)\|^2 : \|\theta\|^2 = 1\}$ entonces $(\theta^*, P^*_{\mathbf{X}}(\theta))$ es una transformación óptima en regresión. Además, por el hecho de que $\|\theta\|^2 = 1$, en el Teorema (2.4.1) se mostró que una transformación óptima satisface:

$$\|\theta - P_{\mathbf{X}}(\theta)\|^2 = 1 - \|P_{\mathbf{X}}(\theta)\|^2 .$$

Ahora denótese por $s = \sup\{\|P_{\mathbf{X}}(\theta)\| : \|\theta\|^2 = 1\}$. Se mostrará que existe una transformación que alcanza el supremo de $\|P_{\mathbf{X}}(\theta)\|^2$ y por tanto el ínfimo de $\|\theta - P_{\mathbf{X}}(\theta)\|^2$.

Si $s = 0$ entonces $\|\theta - P_{\mathbf{X}}(\theta)\|^2 = 1$ para toda θ con $\|\theta\|^2 = 1$ y todas las transformaciones $(\theta, \mathbf{0})$ son óptimas.

Supóngase que $s > 0$. Sea θ_n con $\|\theta_n\|^2 = 1$, tal que $\theta_n \to \theta$ y $\|P_{\mathbf{X}}(\theta_n)\| \to s$. Como $P_{\mathbf{X}}$ es compacto entonces $\|P_{\mathbf{X}}(\theta_n)\| \to \|P_{\mathbf{X}}(\theta)\|^2 = s$. Además $\|\theta\| \leq 1$. Para mostrar que se cumple la igualdad, si $\|\theta\| < 1$, entonces al hacer $\theta' = \theta / \|\theta\|$ se tiene que $\left\|P_{\mathbf{X}}(\theta')\right\|^2 = \|P_{\mathbf{X}}(\theta)\|^2 / \|\theta\|^2 > \|P_{\mathbf{X}}(\theta)\|^2 = s!$ ■

Una vez probada la existencia de transformaciones óptimas, a continuación se verificará cuándo es única. Para ello se denotarán a los operadores $U : H_2(Y) \to H_2(Y)$ y $V : H_2(\mathbf{X}) \to H_2(\mathbf{X})$:

$$\begin{aligned} U(\theta) &= P_Y \circ P_{\mathbf{X}}(\theta) \\ V(\overline{\phi}) &= P_{\mathbf{X}} \circ P_Y(\overline{\phi}) \end{aligned}$$

y se sigue la siguiente

Proposición 2.4.4 *U y V tienen los mismos valores propios y existe una relación uno a uno entre los espacios propios dado un valor propio positivo, especificado por:*

$$\overline{\phi} = P_{\mathbf{X}}(\theta) / \|P_{\mathbf{X}}(\theta)\| , \quad \theta = P_Y(\overline{\phi}) / \left\|P_Y(\overline{\phi})\right\| .$$

Demostración:

Ver Breiman, L. y Freidman, J. H. (1985): "Estimating Optimal Transformations for Multiple Regression and Correlation". Journal of the American Statistical Association, 80.

Sea $\overline{\lambda}$ el mayor valor propio de los operadores $(\overline{\lambda} = \|U\| = \|V\|)$. Si se cumple que existe θ_0 tal que $\|P_{\mathbf{X}}(\theta_0)\| > 0$ entonces $\overline{\lambda} = 0$ y se da el siguiente

Teorema 2.4.3 *Si* $(\theta^*, \overline{\phi}^*)$ *es una transformación óptima en regresión entonces*

$$\begin{aligned} \overline{\lambda}\theta^* &= U(\theta^*), \\ \overline{\lambda}\,\overline{\phi}^* &= V(\overline{\phi}^*), \end{aligned}$$

y la implicación contraria también se da: si θ *es tal que* $\|\theta\| = 1$ *y* $\overline{\lambda}\theta = U(\theta)$, *entonces* $(\theta, P_X\theta)$ *es óptima en regresión.*

Demostración[5]*:*

Sea $(\theta^*, \overline{\phi}^*)$ una transformación óptima; por lo tanto $\overline{\phi}^* = P_{\mathbf{X}}(\theta^*)$ y $\theta^* = P_Y(\overline{\phi}^*)$. Descomponiendo

$$\left\|\theta^* - \overline{\phi}^*\right\|^2 = 1 - 2\left\langle \theta^*, \overline{\phi}^* \right\rangle + \left\|\overline{\phi}^*\right\|^2,$$

y $\left\langle \theta^*, \overline{\phi}^* \right\rangle = \left\langle \theta^*, P_Y(\overline{\phi}^*) \right\rangle \leq \left\|P_Y(\overline{\phi}^*)\right\|$, cumpliéndose la igualdad si y sólo si $\theta^* = P_Y(\overline{\phi}^*)/\left\|P_Y(\overline{\phi}^*)\right\|$. Esto implica que

$$\begin{aligned} \left\|P_Y(\overline{\phi}^*)\right\|\theta^* &= P_Y(\overline{\phi}^*) \\ &= P_Y(P_{\mathbf{X}}(\overline{\phi}^*)) \\ &= U(\theta^*) \end{aligned}$$

y

$$\left\|P_Y\overline{\phi}^*\right\|\overline{\phi}^* = V\overline{\phi}^*,$$

de manera que $\left\|P_Y(\overline{\phi}^*)\right\|$ es un valor propio λ^* de U, V. Sustituyendo, $\left\|\theta^* - \overline{\phi}^*\right\|^2 = 1 - \lambda^*$. Tomando ahora cualquier función propia θ de U correspondiente al valor propio $\overline{\lambda}$ (con $\|\theta\| = 1$) sea $\overline{\phi} = P_{\mathbf{X}}(\theta)$; entonces $\left\|\theta - \overline{\phi}\right\|^2 = 1 - \overline{\lambda}$. Lo anterior muestra que $(\theta^*, \overline{\phi}^*)$ no son óptimas a menos que $\lambda^* = \overline{\lambda}$, de donde se sigue el resto de la demostración.■

De esta forma, inmediatamente se concluye el siguiente

Corolario 2.4.2 *Si* $\overline{\lambda}$ *tiene multiplicidad uno, entonces la transformación óptima es única, salvo por un cambio de signo.*

Demostración:

Ver Breiman, L. y Freidman, J. H. (1985): "Estimating Optimal Transformations for Multiple Regression and Correlation". Journal of the American Statistical Association, 80.

[5] *Ibidim*, pág. 591.

2.5 Implantación del ACE en el caso discreto

En el caso práctico normalmente se presenta que la distribución de los datos es desconocida. Como herramienta de trabajo se cuenta con una muestra aleatoria $(y_k, x_{k1}, ...x_{kp}), k = 1, 2, ..., N$, de las variables $Y, X_1, ..., X_p$. Se tiene como objetivo estimar las transformaciones óptimas $\theta(Y), \phi_1(X_1), ..., \phi_p(X_p)$ a partir de estos datos.

En el caso de variables categóricas, la esperanza condicional se puede estimar de la siguiente forma. Sea (x_k, y_k), $k = 1, 2, ..., N$, una muestra aleatoria con Y categórica. Se define entonces

$$\widehat{E}[X \mid Y = y] = \frac{\sum\limits_{(x_k, y_k): y_k = y} x_k}{N},$$

donde X toma valores reales, y la suma del numerador se toma sobre los datos cuyo valor de Y sea precisamente y.

Para variables que no sean categóricas, la estimación de la esperanza condicional se basa en técnicas llamadas **suavizadores**, que se explicarán a continuación.

Retomando primero el caso univariado, el modelo es:

$$E[Y \mid X] = \phi(X).$$

Para estimar $\phi(X)$ de los datos de la muestra se puede usar cualquier estimación razonable de $E[Y \mid X = x]$. Se analizará primero el que puede considerarse como el suavizador más básico: el promedio móvil.

Sea (x_i, y_i), $i = 1, 2, ..., N$, una muestra aleatoria. Se define el suavizador de promedios móviles como

$$s(x_i) = Prom_{i \in N_i}(y_i),$$

es decir, el promedio de los valores de la variable explicativa contenidos en una cierta vecindad N_i de x_i. Estas vecindades por lo general son consideradas simétricas. Asociado a cada suavizador de este tipo se encuentra la amplitud del intervalo w, la cual representa la proporción del numero total de puntos contenidos en cada vecindad. Específicamente, sea $[\cdot]$ la función parte entera y supóngase que $[wN]$ es impar. Entonces una vecindad simétrica de amplitud w contendrá $[wN]$ puntos: $\frac{[wN]-1}{2}$ puntos de cada lado. Suponiendo que la muestra aleatoria ha sido ordenada, la definición formal de una vecindad simétrica alrededor de x_i está dada por el intervalo abierto:

$$V_i = \left[\max\left(i - \frac{[wN]-1}{2}, 1\right), ..., i-1, i, i+1, ..., \min\left(i + \frac{[wN]-1}{2}, N\right)\right].$$

Nótese que la vecindad de un dato suficientemente cerca de los extremos estará truncada en la misma dirección. El rol del valor de w es esencial en el suavizador, pues entre mayor sea w mayor sera el efecto de suavizar los datos. Más adelante se mencionará una forma de determinar un valor de w dependiendo de los datos que compongan a la muestra.

Cabe mencionar que se observa cierta ineficiencia en este suavizador, ya que los puntos cerca de los extremos están sumamente sesgados, y por ello no reproduce satisfactoriamente rectas (i.e., si los datos corresponden exactamente a una recta, el estimador no respeta la forma). Un ajuste en el suavizador soluciona este problema.

Defínase el suavizador de rectas móviles como

$$\widehat{s}(x_i) = \widehat{\beta}_{0i} + \widehat{\beta}_{1i} x_i,$$

donde $\widehat{\beta}_{0i}$ y $\widehat{\beta}_{1i}$ son los estimadores de mínimos cuadrados para los puntos de la vecindad N_i :

$$\begin{aligned} \widehat{\beta}_{1i} &= \frac{\sum_{j \in N_i} (x_j - \overline{x}_i) y_j}{\sum_{j \in N_i} (x_j - \overline{x}_i)^2}, \\ \widehat{\beta}_{0i} &= \overline{y}_i - \widehat{\beta}_{1i} \overline{x}_i, \end{aligned}$$

y $\overline{x}_i = (1/N) \sum_{j \in N_i} x_j$, $\overline{y}_i = (1/N) \sum_{j \in N_i} y_j$. Adicionalmente, una estimación de $s(z)$ con z distinto a las x_i se puede obtener mediante interpolación.

El suavizador de rectas móviles es la generalización natural de la recta de mínimos cuadrados. Si $w = 2$ (es decir, cada vecindad contiene a todos los datos), el estimador corresponde a la recta misma (nótese que si $w = 1$, las vecindades correspondientes a los datos cercanos a los extremos contendrán sólo aproximadamente la mitad de los datos). En este caso el estimador presenta una varianza reducida y gran efecto de suavización. En el otro caso extremo, si $w = 1/N$, $\widehat{s}(x_i)$ será sencillamente y_i, y la varianza será enorme, además que no existirá efecto de suavización. De esta forma debe seleccionarse un valor entre $1/N$ y 2 para la amplitud de la vecindad a utilizar. Unicamente se dirá que suele escogerse el valor de w que minimice:

$$\sum_{i=1}^{N} \frac{(y_i - \widehat{s}_w^{-i}(x_i))^2}{N},$$

donde $\widehat{s}_w^{-i}(x_i)$ corresponde al suavizador de rectas móviles de amplitud w, sustrayendo el valor de x_i al estimar los parámetros $\widehat{\beta}_{0i}$ y $\widehat{\beta}_{1i}$ para efectos de independencia entre y_i y $\widehat{s}_w^{-i}(x_i)$. Este criterio para seleccionar w pondera el efecto "sesgo - varianza" basándose en los datos de la muestra.

Aunque la naturaleza de este suavizador es sumamente simple, los resultados que produce son bastante aceptables y cuenta con la ventaja de que el estimador en una vecindad puede calcularse utilizando el correspondiente a la vecindad anterior. Como consecuencia, el suavizador puede implantarse en un algoritmo cuyo numero de pasos es proporcional al número de datos (característica conocida como $O(N)$).

Una vez que se ha introducido una noción básica de suavizadores univariados, se pasará a definir formalmente un suavizador para concluir con ejemplos en el caso multivariado.

Denótese por D a un conjunto de N datos p-dimensionales $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, esto es: $\mathbf{x}_k = (x_{k1}, x_{k2}, ..., x_{kp})$. Sea $\wp_N$ el conjunto que tiene como elementos a todos los conjuntos "D". Para un elemento de D fijo sea $F(\mathbf{x})$ el espacio de todas las funciones de dominio real ϕ que están definidas para cada $x_i \in \mathbf{x}$; esto es, $\phi \in F(\mathbf{x})$ está definida por los N números reales $\{\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_N)\}$. Por último sea $F(x_j)$, $j = 1, 2, ..., p$, el conjunto de todas las funciones reales definidas en el conjunto de componentes $\{x_{1j}, x_{2j}, ..., x_{Nj}\}$.

Definición 2.5.1 *Un suavizador S de $\mathbf{x}$ en x_j es una función $S : F(\mathbf{x}) \rightarrow F(x_j)$ definida para cada $D \in \wp_N$. Sea $S(\phi \mid x_j)$ el elemento correspondiente en $F(x_j)$.*

A continuación se mencionan algunos ejemplos de suavizadores multivariados.

1. Histograma. Sea $\{I_l\}$ un conjunto de intervalos tal que es una partición de $\Re$ (o del rango de la muestra). Para $x_k \in I_l$ sea:

$$S(\phi \mid x_j) = \frac{1}{N_l} \sum_{\mathbf{x}_m, x_j \in I_l} \phi(\mathbf{x}_m),$$

   con N_l el número de observaciones $\mathbf{x}_i$ con $x_{ij} \in I_l$.

2. Vecindad más cercana. Sea $M < N/2$. Sean $\{\mathbf{x}_j\}$ tales que $x_1 < x_2 < ..., < x_N$, $1 \leq j \leq p$ (suponiendo que no existen empates). Se define el suavizador de vecindad más cercana como:

$$S(\phi \mid x_j) = \frac{1}{2M} \sum_{\substack{i=-M, \\ i \neq 0}}^{i=M} \phi(\mathbf{x}_{j+i}).$$

3. Kernel. Sea $K(x)$ definida en los reales con máximo en $x = 0$. Se define el kernel como el suavizador:

$$S(\phi \mid x_j) = \frac{\sum_i \phi(\mathbf{x}_i) K(x_i - x_j)}{\sum_i K(x_i - x_j)}.$$

De esta manera, en el algoritmo se reemplazan los operadores $E(\cdot \mid X_j), E(\cdot \mid Y)$ con suavizadores de los datos correspondientes. Adicionalmente, vale la pena resaltar que el rol que el suavizador juega en el algoritmo no es trascendental, salvo por la rapidez de convergencia del último.

En el caso del paquete estadístico utilizado para desarrollar los ejemplos, se utiliza un suavizador que emplea rectas móviles, variando la amplitud de w, conocido como el supersuavizador (*supersmoother* de Friedman y Stueltze).

Para finalizar con el problema de la implantación del ACE a datos, otros elementos necesarios se reemplazan por sus versiones muestrales:

$$\begin{aligned} E(X) &\leftarrow \overline{x} = \sum_{i=1}^{N} \frac{x_i}{N}, \\ Var(X) &= \sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{N(N-1)}, \\ e^2(\theta, \phi_1, ..., \phi_p) &= \frac{1}{N} \sum_{k=1}^{N} \left[\theta(y_k) - \sum_{j=1}^{p} \phi_j(x_{kj}) \right]^2. \end{aligned}$$

Sea $f(y, x_1, ..., x_p)$ una función real definida para todos los valores de la muestra. Entonces

$$\|f\|_N^2 = \sum_{k=1}^{N} \frac{f^2(y_k, x_{k1}, ..., x_{kp})}{N^2}.$$

Ahora bien, el papel del algoritmo es encontrar evaluaciones óptimas para $\theta^*, \phi_1^*, ..., \phi_p^*$ en cada punto $(y_k, x_{k1}, ..., x_{kp})$, $k = 1, 2, ..., N$ (es decir, encontrar el valor óptimo de θ^* para cada valor $\{y_k\}$, $k = 1, 2, ..., N$, al igual que para cada ϕ_j^* en cada punto $\{x_{kj}\}$, $k = 1, 2, ..., N$, $j = 1, 2..., p$). La apariencia de cada transformación se puede apreciar graficando a $\theta^*(y_k)$ contra y_k, y a $\phi_j^*(x_{kj})$ contra x_{kj}.

2.6 Convergencia del ACE en el caso discreto

En la presente sección se denotará a una muestra como $(y_k, \mathbf{x}_k) = (y_k, x_{k1}, x_{k2}, ...x_{kp})$, $k = 1, 2, ..., N$. Supóngase que $\overline{y} = \overline{x}_1 = \overline{x}_2 = \cdots = \overline{x}_p = 0$. Defínase los suavizadores $S_Y, S_1, S_2, ..., S_p$, donde $S_Y : F(y, \mathbf{x}) \rightarrow F(y)$ y $S_i : F(y, \mathbf{x}) \rightarrow F(x_i)$. Nuevamente sea $H_2(y, \mathbf{x})$ el conjunto de funciones en $F(y, \mathbf{x})$ con media cero, y $H_2(y), H_2(x_i)$, $i = 1, 2, ..., p$, los correspondientes subespacios.

Recordando las restricciones impuestas a las transformaciones para ser óptimas, es necesario modificar los suavizadores de manera tal que las funciones estimadas tengan también media cero. Esto se obtiene simplemente restando la media al suavizador:

$$\widetilde{S}_i(\cdot) = S_i(\cdot) - Prom\{S_j(\cdot)\}.$$

De esta forma, en adelante se supondrá que los suavizadores utilizados S cumplen con:

- $Prom(S)$=0.
- Preservan constantes; esto es, que para toda $\psi \in F(\mathbf{x})$ tal que $\phi \equiv c$ constante, se tiene que $S(\psi) = c$ (aunado al primer punto implica que los suavizadores mapean a constantes en cero).
- Son lineales; esto es, que para toda $\psi_1, \psi_2 \in F(\mathbf{x})$ y para toda constante α se cumple $S(\alpha\psi_1 + \psi_2) = \alpha S(\psi_1) + S(\psi_2)$.

En el caso discreto el algoritmo ACE está dado por lo siguiente:

1. $\theta^{(0)}(y_k) = y_k, \quad \phi_i^{(0)}(x_{ki}) \equiv 0.$

   Bucle interior:

2. En el n-ésimo paso del bucle externo empezar con $\theta^{(n)}, \phi_i^{(n)}$. Para $m \geq 1$ e $i = 1, 2, ..., p$, hacer

$$\phi_i^{(m+1)} = S_i\left(\theta^{(n)} - \sum_{j<i} \phi_j^{(m+1)} - \sum_{j>i} \phi_j^{(m)}\right)$$

   e incrementar m hasta converger.

   Bucle exterior:

3. Hacer $\theta^{(n+1)} = S_Y(\sum_{i=1}^p \phi_i) / \left\|S_Y(\sum_{i=1}^p \phi_i)\right\|_N$. Regresar al bucle interior con $\phi_i^{(0)} = \phi_i$ y continuar hasta converger.

Para formalizar el algoritmo se introducirá el espacio $H_2(\theta, \phi)$ con elementos $(\theta, \phi_1, \phi_2, ..., \phi_p)$, $\theta \in H_2(y)$, $\phi_i \in H_2(x_i)$, $i = 1, 2, ..., p$, así como los subespacios $H_2(\dot{\theta})$ y $H_2(\dot{\phi})$ con elementos $\dot{\theta}$ =$(\theta, 0, 0, ..., 0)$ y $\dot{\phi}$ =$(0, \phi_1, \phi_2, ..., \phi_p)$ respectivamente.

Para $\mathbf{f} = (f_0, f_1, f_2, ..., f_p) \in H_2(\theta, \phi)$ sea $S_i : H_2(\theta, \phi) \to H_2(\theta, \phi)$ definido por

$$\{S_i(\mathbf{f})\}_j = \begin{cases} 0, & i \neq j, \\ f_i + S_i\left(\sum_{j \neq i} f_j\right), & i = j. \end{cases}$$

Iniciando con $\dot{\theta} = (\theta, 0, 0, ..., 0)$, $\dot{\phi}^{(m)} = (0, \phi_1^{(m)}, \phi_2^{(m)}, ..., \phi_p^{(m)})$, un ciclo completo del bucle interior es:

$$\dot{\theta} - \dot{\phi}^{(m+1)} = (I - S_p)(I - S_{p-1}) \cdots (I - S_1)(\dot{\theta} - \dot{\phi}^{(0)}). \tag{2.11}$$

Denótese por $\widehat{T} : H_2(\theta, \phi) \rightarrow H_2(\theta, \phi)$ al operador producto introducido en (2.11). Entonces la ecuación mencionada toma la forma:

$$\dot{\phi}^{(m)} = \dot{\theta} - \widehat{T}^m(\dot{\theta} - \dot{\phi}^{(0)}). \tag{2.12}$$

Para una $\dot{\theta}$ fija el bucle interior converge. Entonces el límite satisface

$$S_i(\dot{\theta} - \dot{\phi}) = 0, \quad i = 1, 2, ..., p. \tag{2.13}$$

Esto es, el suavizador de los residuales de cada variable explicativa es igual a cero. Al agregar el hecho de que

$$\dot{\theta} = S_Y(\dot{\phi}) / \left\| S_Y(\dot{\phi}) \right\|_N \tag{2.14}$$

a (2.13) se obtiene un conjunto de ecuaciones que deben ser cumplidas por las estimaciones de las transformaciones óptimas.

Por el supuesto de linealidad de los suavizadores, (2.13) toma la forma

$$S_i(\dot{\theta}) = S_i(\dot{\phi}), \quad i = 1, 2, ..., p. \tag{2.15}$$

Defínase ahora como $sp(S_i)$ a la descomposición espectral de la matriz S_i y además asúmase que $1 \notin sp(S_i)$ (el número 1 es la descomposición espectral de los suavizadores que preservan constantes, pero no de los suavizadores modificados). Se definen las matrices $A_i = S_i(I - S_i)^{-1}$, $i = 1, 2, ..., p$, y la matriz $A = \sum_{i=1}^{p} A_i$. Por último supóngase que $-1 \notin sp(A)$. Entonces (2.15) tiene una única solución dada por

$$\phi_i = A_i(I + A)^{-1}\theta, \; i = 1, 2, ..., p. \tag{2.16}$$

Al elemento formado por las ecuaciones de la forma (2.16) dado por $\dot{\phi} = (0, \phi_1, \phi_2, ..., \phi_p)$ se denotará como $\widehat{P}(\dot{\theta})$. Utilizando el hecho que $(1 - \widehat{T})(\dot{\theta} - \widehat{P}(\dot{\theta})) = 0$, (2.12) se puede escribir como

$$\dot{\phi}^{(m)} = \widehat{P}(\dot{\theta}) - \widehat{T}^{(m)}(\widehat{P}(\dot{\theta}) - \dot{\phi}^{(0)}).$$

De esta forma, si se puede mostrar que $\widehat{T}^{(m)}(f) \rightarrow 0$ para toda $f \in H_2(\dot{\phi})$, entonces el bucle interior converge. El siguiente teorema muestra una condición equivalente a lo anterior.

Teorema 2.6.1 *Si* $\det(I + A) \neq 0$ *y si el radio espectral de* S_i, $i = 1, 2, ..., p$ *son todos menores a uno, entonces* $\widehat{T}^{(m)}(f) \to 0 \quad \forall f \in H_2(\phi) \Longleftrightarrow$

$$\det\left[\lambda I - \prod_{i=1}^{p}\left(I - \frac{S_i}{\lambda}\right)^{-1}(I - S_i)\right] \tag{2.17}$$

no tiene ceros en $|\lambda| \geq 1$ *excepto* $\lambda = 1$.

Demostración:

Para que $\widehat{T}^{(m)}(f) \to 0$, con $f \in H_2(\dot{\phi})$, es necesario y suficiente mostrar que el radio espectral de $\widehat{T}$ es menor que uno. La ecuación $\widehat{T}(f) = \lambda f$ se puede escribir como

$$\lambda f_i = -S_i\left(\lambda_i \sum_{j<i} f_j + \sum_{j>i} f_j\right), \quad i = 1, 2, ..., p. \tag{2.18}$$

Sea $s = \sum_{j=1}^{p} f_j$. Así, (2.18) toma la forma

$$(\lambda I - S_i) f_i = S_i\left((1 - \lambda)\sum_{j<i} f_j - s\right). \tag{2.19}$$

Si $\lambda = 1$, (2.19) se simplifica a $(I - S_i)f_i = -S_i s$ y más aún $s = -As$. Pero por el supuesto sobre A esto implica que $s = 0$ y por tanto $f_i = 0\ \forall i$. Esto descarta la posibilidad de que $\lambda = 1$ sea un valor propio de $\widehat{T}$.

Suponiendo que $\lambda \neq 1$, pero mayor al máximo del radio espectral de S_i, $i = 1, 2, ..., p$, defínase $g_i = (1 - \lambda)\sum_{j<i} f_i - s$. Entonces $f_i = (g_{i+1} - g_i)/(1 - \lambda)$ y

$$(\lambda I - S_i)(g_{i+1} - g_i) = (1 - \lambda)S_i g_i,$$

o

$$g_{i+1} = \left(I - \frac{S_i}{\lambda}\right)^{-1}(1 - S_i)g_i. \tag{2.20}$$

Como $g_{p+1} = -\lambda s$, $g_1 = -s$; entonces (2.20) lleva a

$$\lambda s = \left(I - \frac{S_p}{\lambda}\right)^{-1}(I - S_p)\cdots\left(I - \frac{S_1}{\lambda}\right)^{-1}(I - S_1)\, s. \tag{2.21}$$

Si (2.21) no tiene soluciones no triviales, entonces $s = 0$, $g_i = 0$ para $i = 1, 2, ..., p$, lo que implica que $f_i = 0\ \forall i$. Recíprocamente, si (2.21) tiene una solución $s \neq 0$ entonces esto conduce a una solución de (2.18).

Por desgracia la condición (2.17) es difícil de verificar en general en suavizadores lineales. Si S_i son autoadjuntas, definidas no negativas y tales que los elementos de la matriz suavizadora no modificada son no negativos, entonces todos los radios espectrales de S_i son menores que 1 y puede verificarse que (2.17) se cumple verificando que

$$|\lambda| \leq \prod_{i=1}^{p} \left\| \left(I - \frac{S_i}{\lambda} \right)^{-1} (I - S_i) \right\|$$

no tenga soluciones λ con $|\lambda| > 1$ y posteriormente descartando soluciones con $|\lambda| = 1$.

Suponiendo que el bucle interior converge a $\widehat{P}\dot{\theta}$, el paso del bucle exterior está dado por

$$\dot{\theta}^{(n+1)} = S_Y \widehat{P} \dot{\theta}^{(n)} / \left\| S_Y \widehat{P} \dot{\theta}^{(n)} \right\|_N .$$

Sustituyendo la matriz $S_Y \widehat{P} = \widehat{U}$ se tiene

$$\dot{\theta}^{(n+1)} = \widehat{U} \dot{\theta}^{(n)} / \left\| \widehat{U} \dot{\theta}^{(n)} \right\|_N .$$

Si el valor propio $\widehat{\lambda}$ de $\widehat{U}$ que tiene el mayor valor absoluto es real y positivo entonces $\dot{\theta}^{(n+1)}$ converge a la proyección de $\dot{\theta}^{(0)}$ en el espacio propio de $\widehat{\lambda}$. Para el límite $\dot{\theta}$, $\widehat{P}\dot{\theta}$ es una solución de (2.13) y (2.14). Si $\widehat{\lambda}$ no es real y positiva entonces $\dot{\theta}^{(n)}$ oscila y no converge. Si los suavizadores son autoadjuntos y definidos no negativos entonces $S_Y \widehat{P}$ es el producto de dos matrices autoadjuntas y definidas no negativas; por lo tanto tienen únicamente valores propios reales no negativos. Sin embargo, hasta la fecha de la publicación del algoritmo no se habían podido hallar condiciones que garantizaran tales hechos para suavizadores en general.

No obstante los autores aseguran que no es difícil mostrar que con algunas modificaciones en los puntos cerca de los límites de las muestras, el suavizador de vecindad más cercana satisface las condiciones anteriores. Su investigación indicaba además que otros suavizadores comunes pueden ser modificados para ser autoadjuntos y definidos no negativos, con elementos matriciales no negativos, por lo que lo establecido anteriormente también sería válido para ellos.

Finalmente, los autores declaran que el ACE ha convergido utilizando un gran número de suavizadores que no son autoadjuntos, con excepción de un suavizador de tipo kernel. Para finalizar conjeturan que para la mayoría de conjuntos de datos, un suavizador "razonable" es lo bastante "cercano" a ser autoadjunto para que su mayor valor propio sea real, positivo y menor a uno.

2.7 Ejemplos

En esta sección se presentará una serie de aplicaciones del algoritmo a datos concretos, con el objeto de ilustrar varios aspectos de su comportamiento. Para obtener las estimaciones y generar las gráficas se utilizó el paquete de estadístico S-Plus, el cual incluye una programación del ACE.

Ejemplo 2.7.1 *Ejemplo introductorio.*

Se iniciará con un ejemplo de un modelo univariado. Como primer paso se genera una muestra de tamaño 200 de una distribución $U(0, 2\pi)$ para utiizarla como valores de la variable independiente. Adicionalmente se obtiene una muestra aleatoria independiente del mismo tamaño de una distribución $N(0, 1)$ para que juegue el papel de error. Por último se utiliza el modelo

$$Y = \exp(sen(X) + Z)$$

para obtener los valores de la variable dependiente a partir de los datos anteriores.

Una gráfica de los valores de X contra Y se muestra a continuación:

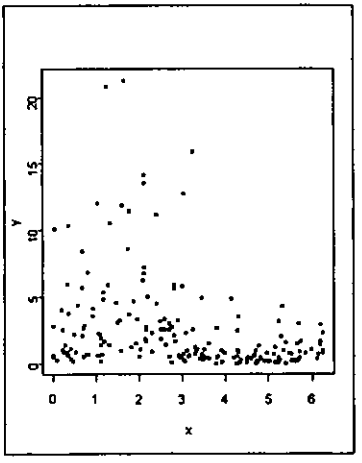

Fig. (2.2). $Y = \exp[sen(X) + Z]$.

Como se ilustra en este sencillo caso el estimar transformaciones óptimas sin la ayuda de una herramienta adecuada puede ser una tarea complicada.

Por otra parte, al conocer la relación que guardan estas variables es posible comparar las transformaciones $\phi^*(X)$ y $\theta^*(Y)$ sugeridas por el ACE contra las originales $sen(X)$ y $\log(Y)$, resaltando que las últimas pueden no generar la máxima correlación posible, debido al factor del error. Sin embargo se puede decir que son transformaciones "satisfactorias".

Una vez aplicado el algoritmo, las gráficas de los datos transformados óptimos según el ACE son las siguientes:

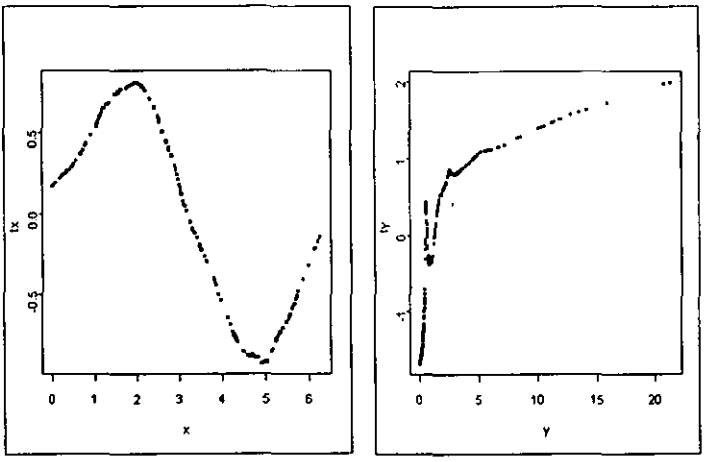

Fig. (2.3). a) $\phi^*(X)$; b) $\theta^*(Y)$.

La primera gráfica corresponde a la transformación sugerida para la variable explicativa mientras que la última corresponde a la dependiente.

En efecto, las transformaciones sugeridas por el ACE corresponden a la forma de las funciones que se utilizaron en el modelo para generar los datos. Este hecho se puede confirmar en las siguientes gráficas, donde se vuelven a presentar las transformaciones $\theta^*(X)$ y $\phi^*(Y)$ además de las funciones originales, graficadas en línea sólida.

El desplazamiento de las soluciones del algoritmo con respecto a las funciones originales se explica debido a que las primeras son ajustadas para que su esperanza en el rango de la muestra sea igual a cero.

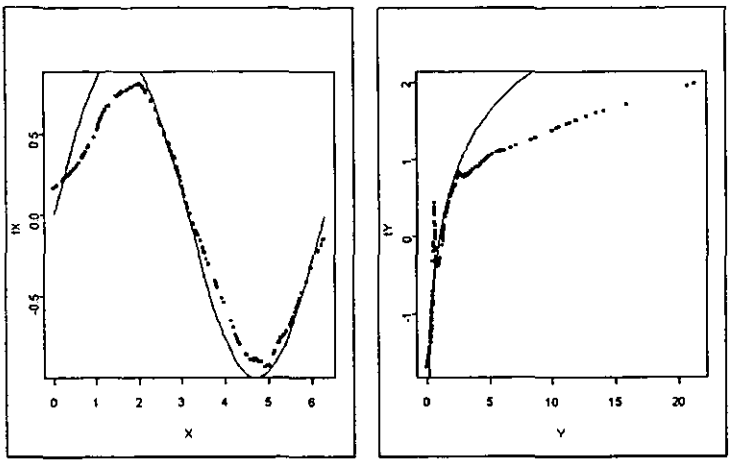

Fig.(2.4). a) $\phi^*(X)$ y $sen(X)$; b) $\theta^*(Y)$ y $\log(Y)$.

Por último, es posible verificar la linealidad que alcanza el modelo transformado, mediante una gráfica de los valores de $\theta^*(X)$ contra los valores de $\phi^*(Y)$.

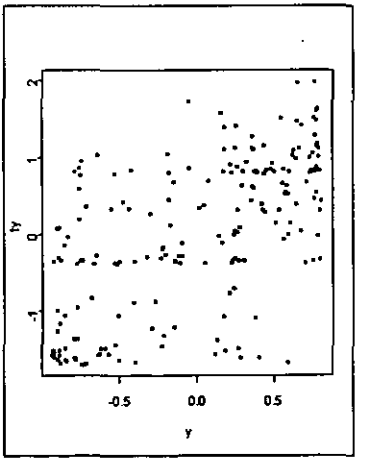

Fig. (2.5). $\phi^*(X)$ vs $\theta^*(Y)$.

Como se puede apreciar, la linealidad alcanzada parece aceptable. No obstante es importante observar la variabilidad que los valores de $\phi^*(Y)$ alcanzan en cada pequeño intervalo de $\theta^*(X)$, lo cual será retomada y discutido con mayor detalle más adelante.

Ejemplo 2.7.2 *Caso multivariado.*

Para este ejemplo las variables independientes (X_1 y X_2) son generadas por una muestra de tamaño 200 de una distribución $U(-2\pi, 2\pi)$ y $U(-1, 1)$ respectivamente. Nuevamente se usa la distribución normal estándar para generar el vector de errores.

El modelo para obtener el valor de la variable dependiente es

$$y = \log(4 + \cos(X_1) + (X_2)^3 + Z),$$

por lo que se puede asumir que una transformación acertada para la variable dependiente sería similar a una exponencial.

A continuación se muestran las gráficas de las transformaciones de X_1, X_2 y Y una vez que aplicado el método a los datos.

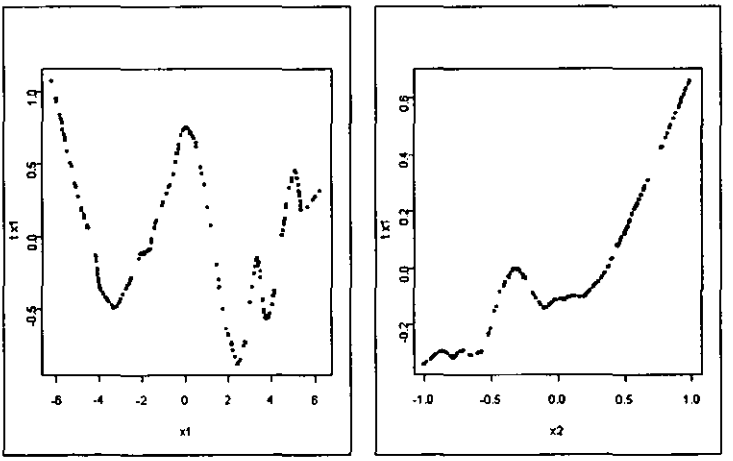

Fig. (2.6.1). a) $\phi_1^*(X_1)$; b) $\phi_2^*(X_2)$.

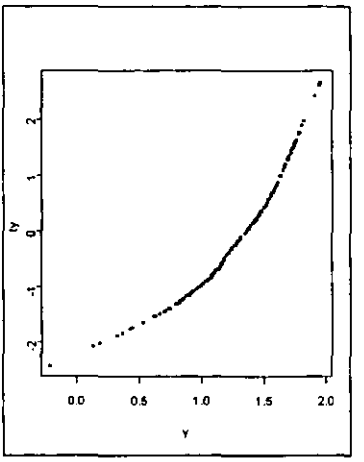

Fig. (2.6.2). $\theta^*(Y)$.

Al observar las estimaciones por primera vez, parece que las transformaciones sugeridas no son tan claras. La transformación $\phi_1^*(X_1)$ tiene una forma que podría ser algún tipo de polinomio no muy obvio. Por su parte, la transformación $\phi_2^*(X_2)$ presenta cierta anomalía en la parte central del rango de la muestra de X_2 que puede hacer pensar en una función algo compleja. Por último, la gráfica de $\theta^*(Y)$ parece no presentar problemas.

Recordando que al aplicar el algoritmo al modelo presentado, el resultado es estimar no dos sino tres transformaciones, se puede esperar que las estimaciones presenten algo más de "ruido" provocado por el error introducido en el modelo.

A la luz del comentario anterior, analícense un poco más a fondo las gráficas, fijando la atención en la forma general de las estimaciones.

En el caso de $\phi_1^*(X_1)$ se puede notar una especie de periodo entre el cambio de tendencia general, al moverse la variable X_1 aproximadamente dos o tres unidades. Además es notorio que la amplitud del rango entre el máximo y mínimo de una "racha grande" es aproximadamente constante. Estas dos características sugieren que la transformación sugerida está relacionada de alguna manera con alguna función sinusoidal. Observando el valor que ϕ_1^* toma en $X_1 = 0$ y el cambio de tendencia aproximadamente en $X_1 = -3.5, 0$ y 2.5 puede inclinar la balanza hacia $\phi_1^*(X_1) = \cos(X_1)$. Observando que los valores en el extremo derecho de la gráfica no definen una tendencia clara, es factible que en efecto esta función es buen candidato, aunque pueden existir algunas dudas que podrían analizarse aún más.

En el caso de la transformación para X_2, a pesar de que la función X^3 parece probable, por la forma que presenta la gráfica en la parte central podría pensarse también en una función definida por segmentos. Este ejemplo ilustra una decisión que comúnmente debe realizarse: el valorar si se deben ignorar algunos detalles y utilizar una función "sencilla" (X^3), o no ignorarlos, lo cual llevaría a una función que puede complicar el modelo.

Por último, la transformación para Y presenta otra disyuntiva. La función exponencial parece un buen candidato. Sin embargo, si existen otras alternativas que aproximen a la función óptima dentro del rango de los datos respectivos (como aquí podría ser el caso de una cuadrática), puede suceder que al estimar la transformación el algoritmo incline (por el factor error) a seleccionar la función equivocada. Ahora bien, si el modelo estimado es utilizado para predicción y se usan datos dentro del rango con el que se desarrolló el modelo, al aproximar la transformación escogida a la original no se tendrán grandes problemas. Sin embargo, si se desea predecir un valor para la variable fuera del rango de la muestra, la cercanía de las funciones puede no existir más, y una diferencia enorme puede existir. Esto es sólo recordatorio de una observación hecha anteriormente, para cualquier extrapolación en general.

A continuación se presentan las gráficas de las estimaciones anteriores junto con las funciones óptimas para mostrar qué tan bien se aproximan las sugerencias obtenidas al aplicar el algoritmo.

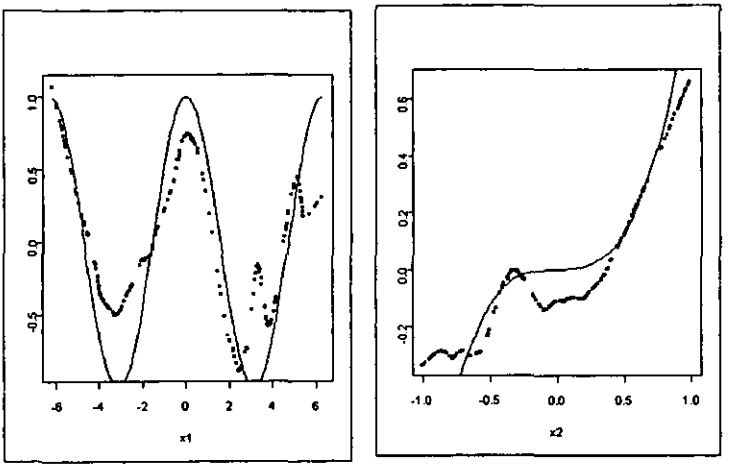

Fig. (2.7.1). a) $\phi_1^*(X_1)$ y $\cos(X_1)$; b) $\phi_2^*(X_2)$ y (X_2).

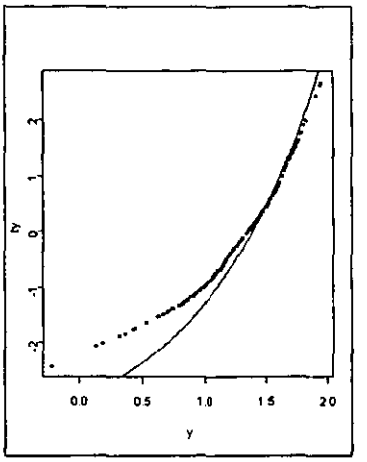

Fig. (2.7.2). $\theta^*(Y)$ y $\exp(Y)$.

Como última observación, nótese que en los últimos dos ejemplos se cuenta con el conocimiento de las variables que generan el modelo y sus valores, sin errores de medición. Estas ventajas se presentan rara vez en la práctica. Sin embargo, aún con el sencillo modelo utilizado, algunas dudas no son fáciles de despejar para tomar una decisión. Lo que es más, aún cuando se conoce que el algoritmo está presentando una buena estimación, al final todo queda en las manos y experiencia del usuario para tomar la decisión que puede significar enormes diferencias. De aquí que el algoritmo es una herramienta que requiere un buen juicio humano.

Ejemplo 2.7.3 *Variables categóricas.*

El modelo en este ejemplo es un modelo univariado, donde la variable dependiente únicamente toma los valores 1 o -1. Se utilizó un vector de tamaño 200, donde los primeros 100 componentes se fijaron en 1 y los demás en -1. El error es una muestra de tamaño 200 $N(0,1)$ y el modelo para obtener la variable dependiente fue

$$Y = \exp(X + Z).$$

La gráfica que resulta al aplicar el algoritmo se muestra a continuación, con la transformación obtenida con el método graficada con puntos

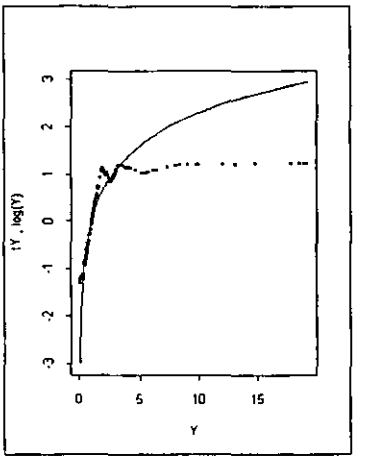

Fig. (2.8). $\theta^*(Y)$ y $log(Y)$.

la cual parece ser una buena aproximación a la función logarítmica, presentada en línea sólida en la misma gráfica. De esta manera, se puede observar que el algoritmo cumple una buena tarea en este caso de variables categóricas.

Como se podrá haber notado, hasta ahora se han utilizado transformaciones similares para la variable dependiente. La constante ha sido transformaciones monótonas para Y, lo cual asegura la invertibilidad de la función para aplicarla a ambos lados del modelo y poder despejar la variable independiente. En el siguiente ejemplo se utiliza un modelo que no presenta esta característica.

Ejemplo 2.7.4 *Transformaciones no monótonas.*

Se utilizarán dos muestras independientes idénticamente distribuidas $U(-1,1)$ de tamaño 200 cada una para generar los datos de las variables explicativas X_1 y X_2. Posteriormente se generan los valores de la variable dependiente de acuerdo al modelo

$$Y = X_1 X_2.$$

Conociendo el modelo que genera los valores, se puede intentar obtener las transformaciones $\theta(Y), \phi_1(X_1)$ y $\phi_2(X_2)$ que permitan obtener un modelo de la forma $\theta(Y) = \phi_1(X_1) + \phi_2(X_2)$.

Inicialmente pueden considerarse transformaciones idénticas de la forma

$$\theta(\cdot) = \phi_1(\cdot) = \phi_2(\cdot) = \log(\cdot).$$

Sin embargo, considerando que las tres variables llegan a tomar valores negativos tal proposición necesita ajustarse. Sustituyendo las transformaciones por

$$\theta(\cdot) = \phi_1(\cdot) = \phi_2(\cdot) = \log|\cdot|$$

se soluciona el problema. En este caso, la transformación $\theta(Y) = \log|Y|$ es una función no monótona, por lo que el modelo servirá para el propósito.

Las transformaciones sugeridas por el algoritmo para X_1, X_2 y Y se presentan a continuación, junto con las funciones $\log|X_1|, \log|X_2|$ y $\log|Y|$ en línea sólida.

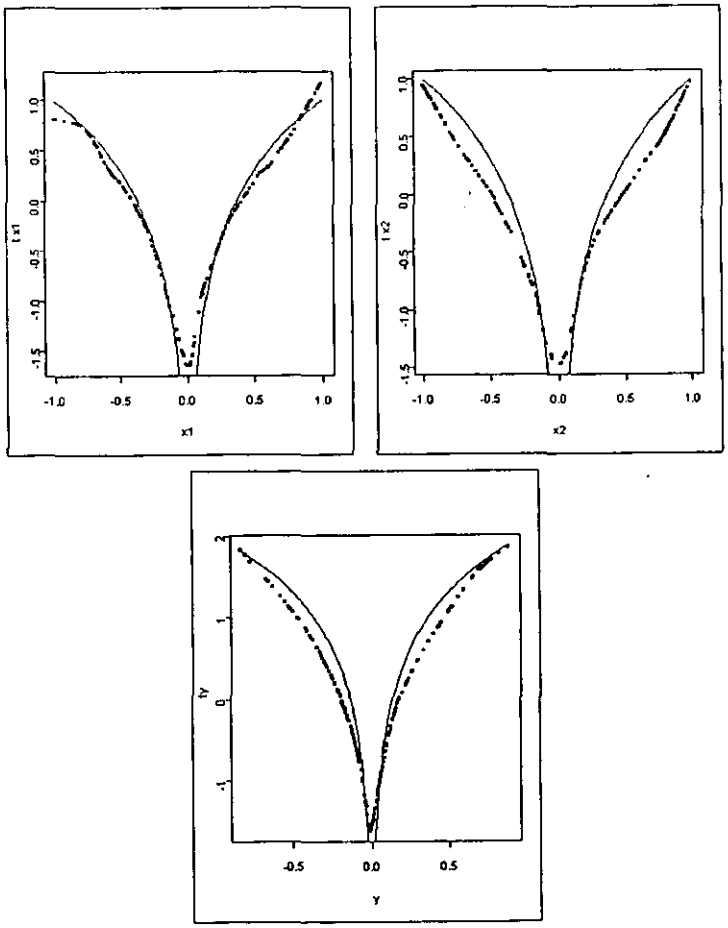

Fig. (2.9.1). a) $\phi_1^*(X_1)$ y $\log|X_1|$; **b)** $\phi_2^*(X_2)$ y $\log|X_2|$; c) $\theta^*(Y)$ y $\log|Y|$.

Como se puede apreciar, el algoritmo refleja bastante bien la forma de las transformaciones originales, a excepción tal vez de los datos cerca del origen, debido a que el ACE se aplica a un número finito de datos, y el suavizador no puede reproducir la discontinuidad de la función más allá de cierto punto. Sin embargo hay que destacar que el presente modelo no incluyó factor alguno de error. Aún así es posible concluir que, al menos en este ejemplo en particular, el algoritmo fue capaz de manejar una transformación no monótona en la variable dependiente.

Ejemplo 2.7.5 *Comparación de la correlación.*

En el articulo donde se presenta el algoritmo, Breiman y Friedman estudian la posibilidad de "forzar" una alta correlación entre la transformación de la variable dependiente y la combinación lineal de las transformaciones de los regresores, debido a la constante aplicación de los suavizadores.

Utilizando una serie de 100 simulaciones independientes, los autores obtienen la correlación que resulta al aplicar el algoritmo en cada experimento. Al mismo tiempo registran la correlación muestral entre las transformaciones reales que se utilizaron para generar los datos. El objetivo es analizar las diferencias que el algoritmo genera sobre las medidas ρ y R^2.

En el presente ejemplo se retoma la idea anterior aplicada al modelo

$$Y = \exp(sen(X) + Z)$$

efectuando el "experimento" en 300 muestras para X y Z de tamaño 200 cada una, tomadas independientemente, con $X \sim U(-\pi, \pi)$ y $Y \sim N(0, 1)$.

Al aplicar el algoritmo a cada una de las muestras, se registra el porcentaje de variación no explicada por las transformaciones del ACE (R^{2*}), a partir del cual se obtiene el coeficiente de correlación correspondiente al ACE (ρ^*) utilizando la igualdad $R^{2*} = 1 - \rho^{*2}$.

Al mismo tiempo se registran los mismos estadísticos al aplicar las transformaciones originales que generan los datos, definidos simplemente por

$$\begin{aligned} \rho &= \frac{1}{N}\sum_{k=1}^{N} \log(y_k) sen(x_k), \\ R^2 &= 1 - \rho^2, \end{aligned}$$

y se compararán los resultados obtenidos.

La siguiente tabla muestra las medias y sus desviaciones estándar registradas en el experimento.

	Media	Desviación estándar
ρ (modelo)	.8089	.0433
ρ^* (ACE)	.8845	.0488
	Media	**Desviación estándar**
R^2 (modelo)	.6563	.0688
R^{*2} (ACE)	.7847	.0839
	Media	**Desviación estándar**
$\rho^* - \rho$	.0755	.0471
$R^{*2} - R^2$	.1284	.0787

Tabla (2.1). R^2 y ρ del modelo original y del modelo sugerido por el ACE.

Se puede percibir que el algoritmo forza la correlación de los datos de una manera hasta cierto punto considerable. Sin embargo, al tratar de aproximar la estimación sugerida con una función conocida, se compensará en parte este problema.

Por último es recomendable señalar que los resultados presentados por los autores con un número menor de simulaciones presentan diferencias menos notorias entre las transformaciones sugeridas por el algoritmo y las originales.

Se concluye de esta manera la exposición del primer método, la teoría detrás de él así como una serie de ejemplos donde se muestra principalmente un comportamiento aceptable del mismo. Sin embargo, es claro que toda técnica tiene defectos o debilidades; algunos de los detectados en el ACE se discutirán al inicio del siguiente capítulo, incluyendo uno en particular que servirá de punto de enlace entre este algoritmo y su modificación, estudiada posteriormente.

Capítulo 3

Método AVAS

3.1 Observaciones sobre el ACE

En el capítulo anterior se discutieron las características del ACE, así como algunos ejemplos que muestran su comportamiento. No obstante se debe recordar que esta metodología es sólo una herramienta, por lo que no ofrece una solución por sí misma. El resultado del algoritmo, primeramente, se basa en ciertos criterios, y aún más, está sujeto a interpretaciones; es en este momento cuando el juicio del usuario puede desembocar en una aproximación aceptable o tal vez errónea. Por ello es necesario considerar algunos aspectos tanto empíricos como teóricos atribuibles al ACE, mismos que se irán discutiendo a través de diversos ejemplos.

3.1.1 La correlación como objetivo

Como se ha descrito, la construcción del ACE se basa en la idea de maximizar la correlación que puede existir entre una transformación de la variable dependiente Y y la combinación lineal de transformaciones de las variables independientes X_i. De esta forma, el primer punto a considerar es la consecuencia de utilizar esta función objetivo.

Para comenzar, considérese el caso univariado. Sean X y Y variables aleatorias y ϕ y θ funciones medibles en el respectivo espacio de cada variable. Sea $\rho(\phi,\theta) = E[\phi(X)\theta(Y)]$ el coeficiente de correlación muestral alcanzado por el modelo $\theta(Y) = \phi(X)$. Denótese por $\rho^* = \max_{\phi,\theta}\{\rho(\phi,\theta)\}$. Las propiedades que ρ^* debe cumplir están dadas por

i) $0 \leq \rho^* \leq 1$.
ii) $\rho^* = 0 \Leftrightarrow X$ y Y son independientes.
iii) $\rho^* = 1 \Leftrightarrow \exists\, \theta, \phi$ (con $var(\theta) > 0 < var(\phi)$) $\ni \theta(Y) = \phi(X)$.

Por las anteriores propiedades, suele relacionarse a valores de ρ^* cercanos a uno con un estado de "alta dependencia" entre las variables Y y X. Sin embargo, hay casos en donde tal intuición es injustificable, como se muestra a continuación. Sean X y Y variables aleatorias cuya función de distribución cumpla el siguiente modelo:

$$(X,Y) = \begin{cases} (X, f(X)) \text{ si } X \in S, \\ (X, Y_1) \text{ si } X \notin S, \end{cases}$$

donde f es una función estrictamente monótona definida para todo valor de X, S es un subconjunto medible del rango de X y Y_1 es una variable independiente de X. La intuición sugiere que si la medida del subconjunto S tendiese a cero, el valor de ρ^* debería converger a cero de igual forma. Sin embargo éste no es el caso. Para cualquier subconjunto S con medida positiva, la transformación $\phi(X) = f(X)I_{[S]}$ es una función no trivial. Por lo tanto, las transformaciones $\phi(X) = f(X)I_{[S]}$, $\theta(Y) = Y I_{[f(S)]}$, con I la función característica usual, obligan a que ρ^* sea igual a uno, sin importar la medida del conjunto S.

La misma observación aplicaría si la variable dependiente y alguna de las explicativas (X_j para alguna j), cambian de signo al mismo tiempo. Se podría asegurar una máxima correlación de uno prediciendo el signo de X_j y haciendo caso omiso a las demás variables.

De esta forma se concluye que la correlación máxima no es una medida eficaz de la independencia entre variables, pues puede no responder a cambios generales conforme las distribuciones de éstas pasan de un estado de alta dependencia a una independencia casi total. Esto implica que el tratar de maximizar la correlación puede llevar a transformaciones que "desperdician" información, debido a que al considerarla la correlación disminuiría. Un ejemplo sencillo se ilustra enseguida:

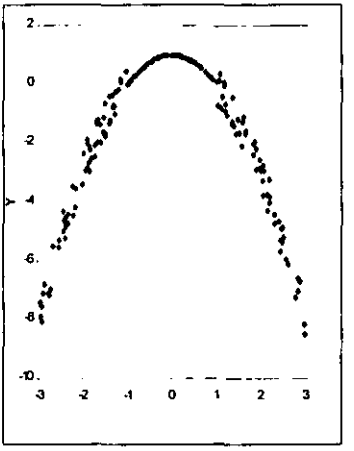

Fig. (3.1). $Y \approx -X^2 + 1$.

El modelo exacto está dado por:

$$Y = -X^2 + ZI_{[-2,-1]\cup[1,2]} + 1.$$

De esta forma, para el objetivo de maximizar la correlación, la transformación $\phi^*(X) = (-X^2 + 1)\, I_{[-1,1]}$ genera una correlación perfecta $\rho^* = 1$, siendo que la transformación que intuitivamente es más conveniente es simplemente $-X^2 + 1$.

Este "síntoma" de despreciar zonas informativas y seleccionar transformaciones poco interesantes puede representar un punto crítico del algoritmo, pues no hay que olvidar que un punto esencial de una regresión puede ser el explicar la variabilidad de Y y no el de cualquier transformación arbitraria $\phi(Y)$.

Aunque bien es cierto que el considerar a la correlación como un punto de partida teórico deficiente para la construcción del algoritmo está sujeto a discusión. En casos como el descrito anteriormente donde existen "regímenes" separados, la correlación se enfoca en la existencia de ésta división en los datos como la característica más sobresaliente, e ignora otro tipo de estructura más fina. Sin embargo no puede argumentarse definitivamente que tal característica refleja una insensibilidad del procedimiento. Como los mismos creadores han señalado, el principal problema es que "aún no se ha llegado a entender lo suficiente al algoritmo en el caso discreto".

Se pasará ahora al segundo punto a discusión, el cual está relacionado con el primero. Es bien sabido que la máxima correlación es una de las principales medidas (aunque no única) para "calificar" qué tan bien se ajusta un modelo a los datos. Sin embargo se observa que no se incluyen más factores en el desarrollo del ACE para obtener transformaciones óptimas. Esto implica, obviamente, que aquellos aspectos de un modelo que no alcancen a ser considerados por la correlación serán insensibles al algoritmo. Para ilustrar un poco lo anterior considérese el siguiente ejemplo.

Ejemplo 3.1.1 *Sea un modelo univariado donde las variables guardan la relación*

$$Y = X + (.1X)Z,$$

y los datos para la variable explicativa provienen de una muestra de tamaño 200 de una distribución $U(0,1)$, *mientras que* Z *se distribuye* $N(0,1)$ *representando el error.*

En este ejemplo se trata de ilustrar la inconveniencia de considerar únicamente la correlación, e ignorando otros factores no menos importantes, como es la estabilidad en la varianza alcanzada por el modelo final. Una figura de los 200 datos aleatorios se muestra en la siguiente figura.

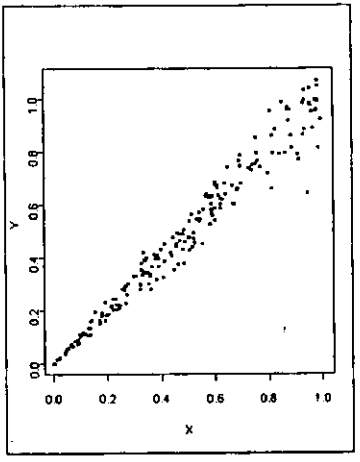

Fig. (3.2). $Y = X + (.1X)Z$.

Al aplicar el ACE a los datos, la transformación para la variable independiente que resulta es casi una línea recta, como resultaría al realizar una regresión lineal común. Sin embargo es necesario recordar que el algoritmo está diseñado para considerar todo tipo de transformaciones, por lo que existen más posibilidades para poder aproximar este modelo, y de hecho para poder estabilizar la varianza que artificialmente fue alterada.

Obviamente la estabilidad alcanzada por la transformación casi lineal es sumamente criticable, como se muestra a continuación

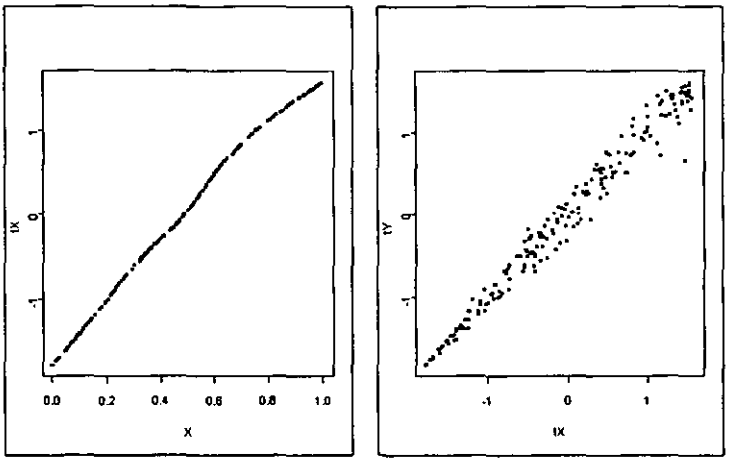

Fig. (3.3). a) $\phi^*(X)$; b) $\phi^*(X)$ vs. $\theta^*(Y)$.

De esta forma, parece recomendable que el algoritmo considere otros aspectos para encontrar una transformación óptima, y de hecho tal observación fue la base para el desarrollo de una "segunda generación" del algoritmo.

3.1.2 Soluciones Subóptimas

El siguiente aspecto a recalcar del ACE es que, si bien las transformaciones que se obtuvieron en los ejemplos iniciales fueron satisfactorias, existen situaciones en las que el algoritmo no es tan preciso. La serie de ejemplos presentada en el capítulo anterior es una clásica serie de ejemplos teóricos los cuales no suelen presentarse en la vida real. Más aún, factores aparentemente irrelevantes pueden hacer que la claridad en los resultados se vea perturbada notablemente, aún en el caso teórico.

Ejemplo 3.1.2 *Retomándo el modelo del Ejemplo (2.7.4) dado por* $Y = X_1X_2$ *con* $X_1, X_2 \sim U(-1,1)$, *modifíquese la distribución de* X_2 *a una* $U(0,1)$.

En este caso las transformaciones buscadas no cambian, y como X_2 no toma valores negativos se puede decir que las transformaciones siguientes son óptimas:

$$\begin{aligned} \theta(Y) &= \log|Y|, \\ \phi_1(X_1) &= \log|X_1|, \\ \phi_2(X_2) &= \log(X_2). \end{aligned}$$

Sin embargo, de igual forma las transformaciones $\theta(Y) = I_{[0,\infty)}$, $\phi_1(X_1) = I_{[0,\infty)}$ y $\phi_2(X_2) \equiv 0$ son igualmente óptimas en correlación, aunque puede pensarse que carecen de interés en comparación a las transformaciones inicialmente mencionadas, aún cuando explican el modelo con sólo un regresor.

Las transformaciones sugeridas por el ACE a una serie de datos aleatoria de tamaño 200 se muestran a continuación:

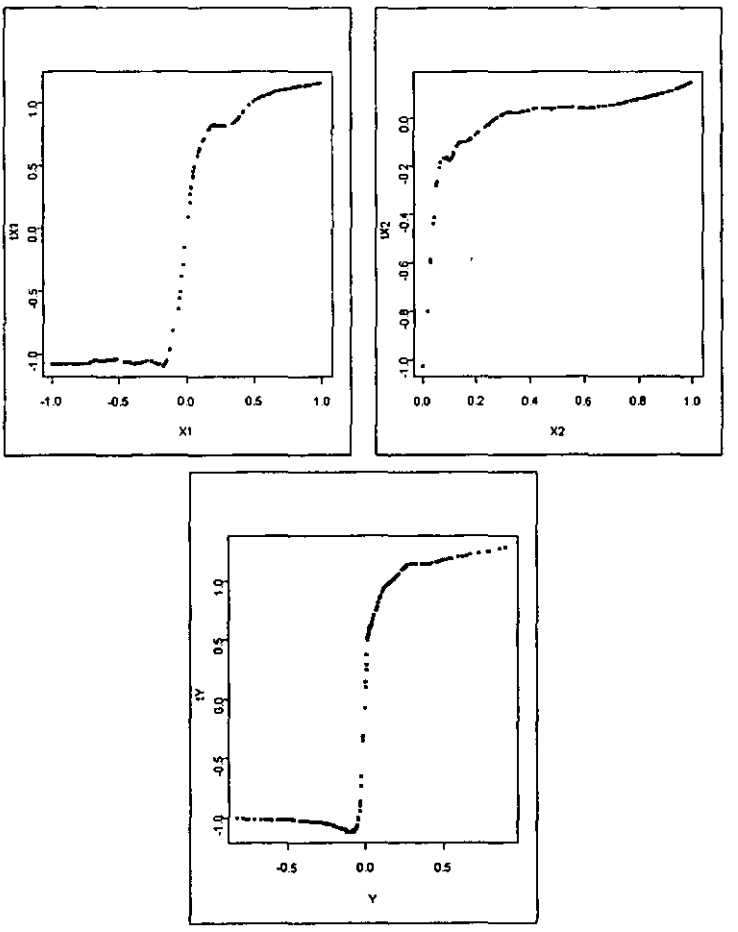

Fig. (3.4). a) $\phi_1^*(X_1)$; b) $\phi_2^*(X_2)$; c) $\theta(Y)$.

Como se observa, las transformaciones no reflejan tan claramente el resultado buscado.

En estos casos se desearía entender cómo el algoritmo selecciona la transformación óptima de entre el conjunto de transformaciones óptimas posibles. Al respecto, los autores mencionan[1] que sólo se requiere una modificación al algoritmo para que sea capaz de presentar una serie de transformaciones subóptimas: la estandarización de la transformación $\theta(Y)$ es reemplazada por una ortogonalización más general en la cual sus proyecciones en las transformaciones "óptimas" anteriores de Y son sustraidas.

Implementando la anterior modificación al ACE, Breiman y Friedman presentan las gráficas de varias transformaciones subóptimas sugeridas con el ACE de un ejemplo de características similares, lo cual desafortunadamente no es posible reproducir en este trabajo por no disponer del algoritmo modificado.

3.1.3 Comportamiento en presencia de bajas asociaciones.

Ejemplo 3.1.3 *Considérese el modelo*

$$Y = X_1 + X_2 + ... + X_{10} + Z,$$

con Z *y* X_i *v.a.i.i.d.* $N(0,1)$.

En un inicio, el modelo parece no presentar complicación alguna a pesar del alto número de variables explicativas, debido a que las transformaciones que deberían resultar óptimas son la identidad misma. Sin embargo se observa que los resultados no son así. En la siguiente figura se muestran las transformaciones obtenidas para cuatro variables usando una muestra de tamano 300. Como se puede apreciar, la linealidad queda sugerida en ellas.

[1] *Ibidim*, pag. 616.

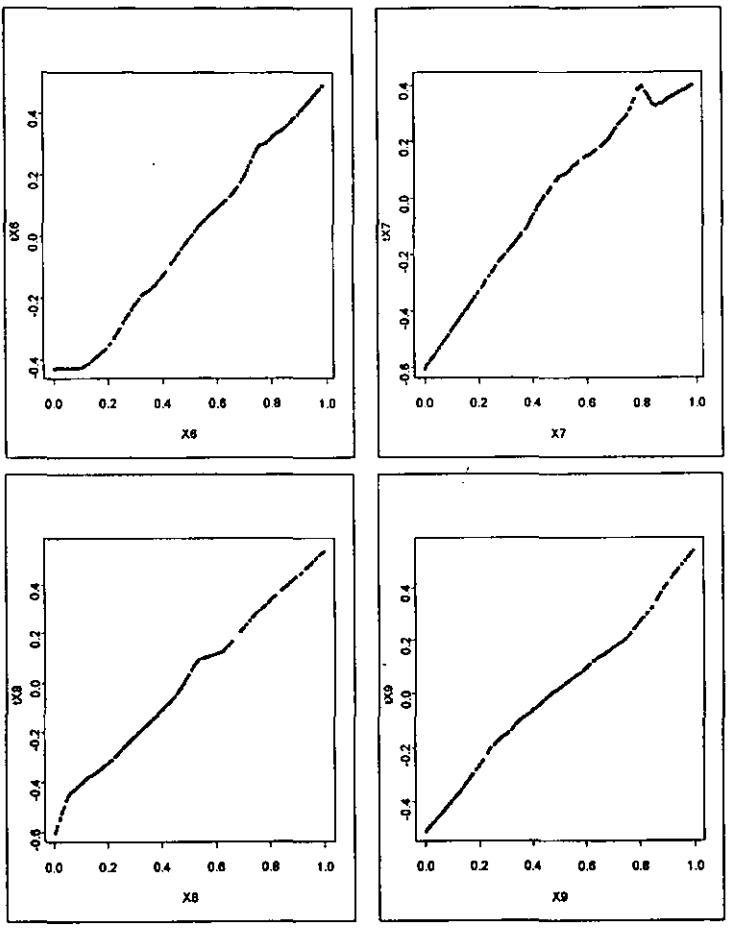

Fig. (3.5). a) $\phi_6^*(X_6)$; b) $\phi_7^*(X_7)$; c) $\phi_8^*(X_8)$; d) $\phi_9^*(X_9)$ con N=300.

Sin embargo, considérese el mismo ejercicio al aplicarlo a una muestra de tamaño 100. A continuación se muestran las mismas variables para el caso en turno. Como se puede apreciar, la linealidad sugerida por el algoritmo ha disminuido considerablemente.

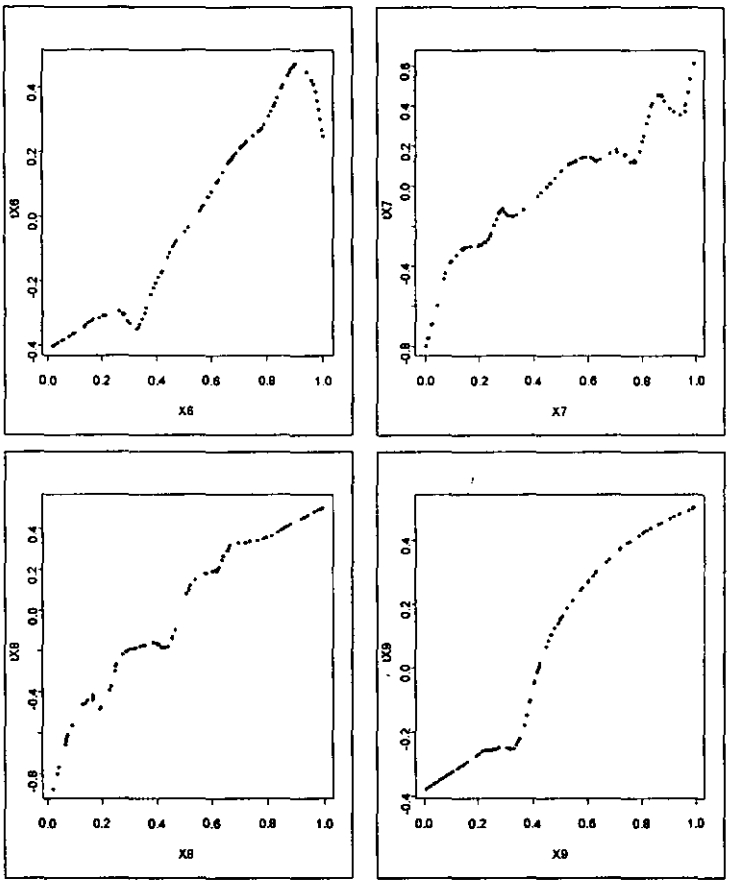

Fig. (3.6). a) $\phi_6^*(X_6)$; b) $\phi_7^*(X_7)$; c) $\phi_8^*(X_8)$; d) $\phi_9^*(X_9)$ con N=100.

Los autores atribuyen este comportamiento a lo que se conoce como una escasa asociación entre variables. En este caso, de una manera algo atrevida se podría considerar que la contribución que cualquier variable explicativa aporta al valor de la variable dependiente es en promedio menor al 10 por ciento. De esta forma, como sucede al utilizar cualquier modelo, es recomendable no basar conclusiones fuertes en los resultados obtenidos por el ACE cuando se cuenta con un escaso poder explicativo. Al respecto, sólo se mencionará que es posible mejorar considerablemente el desempeño del algoritmo en estos casos al incrementar la amplitud de la ventana del suavizador para las variables cuya asociación con $\theta(Y)$ es escasa en las primeras iteraciones del procedimiento, e incrementarla de igual forma para estimar $\theta(Y)$ si se encuentra que el valor de e^{*2} es grande, como los autores señalan.

3.1.4 Sensibilidad a Datos Aberrantes.

Ejemplo 3.1.4 *La siguiente tabla presenta los datos de 62 especies de mamíferos correspondientes a su peso corporal en kilogramos* (X) *y peso cerebral en gramos* (Y).

Especie	X	Y	Especie	X	Y
Zorro ártico	3.39	44.50	Hombre	62.00	1320.00
Mono	.48	15.50	Elefante africano	6654.00	5712.00
Castor de montaña	1.35	8.10	Erizo de roca	3.50	3.90
Vaca	465.00	423.00	Mono	6.80	179.00
Lobo gris	36.33	119.50	Canguro	35.00	56.00
Cabra	27.66	115.0	Marmota	4.05	17.00
Venado	14.83	98.20	Hamster	0.12	1.00
Cerdo de Guinea	1.04	5.50	Ratón	0.023	0.40
Verveta	4.19	58.00	Murciélago café	0.10	0.25
Chinchilla	0.43	6.40	Lemur	1.40	12.50
Ardilla	0.10	4.00	Okapi	250.00	490.00
Ardilla ártica	0.92	5.70	Conejo	2.50	12.10
Rata africana	1.00	6.60	Oveja	55.50	175.00
Musaraña cola corta	.005	0.14	Jaguar	100.00	157.00
Topo	.06	1.00	Chimpancé	52.16	440.00
Armadillo	3.50	10.80	Mandril	10.55	179.50
Hyrax de árbol	2.00	12.30	Erizo de desierto	0.55	2.40
Zarigüeya americana	1.70	6.30	Armadillo gigante	60.00	81.00
Elefante de asia	2547.00	4603.00	Hyrax de roca	3.60	21.00
Murciélago	0.023	0.30	Mapache	4.29	39.20
Burro	187.10	419.00	Rata	0.28	1.90
Caballo	521.00	655.00	Topo americano	0.075	1.20
Erizo europeo	0.785	3.50	Topo rata	0.122	3.00
Mono	10.00	115.00	Muzaraña de almizcle	0.048	0.33
Gato	3.30	25.60	Cerdo	192.00	180.00
Jineta	0.20	5.00	Equidna	3.00	25.00
Gálago	1.41	17.50	Tapir	160.00	169.00
Jirafa	529.00	680.00	Tenrec de Madagascar	0.90	2.60
Gorila	207.00	406.00	Marsupial	1.62	11.40
Foca gris	85.00	325.00	Musaraña de árbol	0.104	2.50
Hyrax	0.75	12.30	Zorro rojo	4.235	50.40

Tabla (3.1). Datos de pesos corporal y cerebral.

De esta forma, se aplica el ACE para establecer una relación entre estas dos variables.

La gráfica de los datos originales se presenta a continuación.

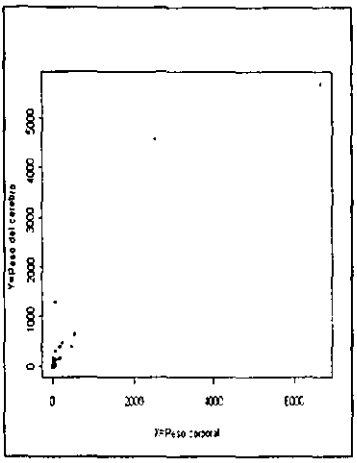

Fig. (3.7). Datos observados.

La gráfica a primera vista sugiere inmediatamente tomar logarítmos, como regla empírica. Sin embargo las transformaciones obtenidas con el ACE son las siguientes:

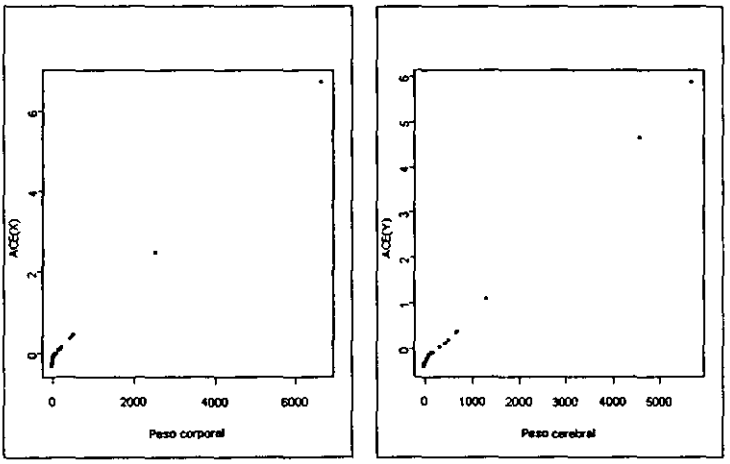

Fig. (3.8). a) $\phi^*(X)$; b) $\theta^*(Y)$.

Graficando los datos transformados se observa que la estabilidad del modelo dista mucho de la alcanzada por las transformaciones logarítmicas, como se ilustra en seguida.

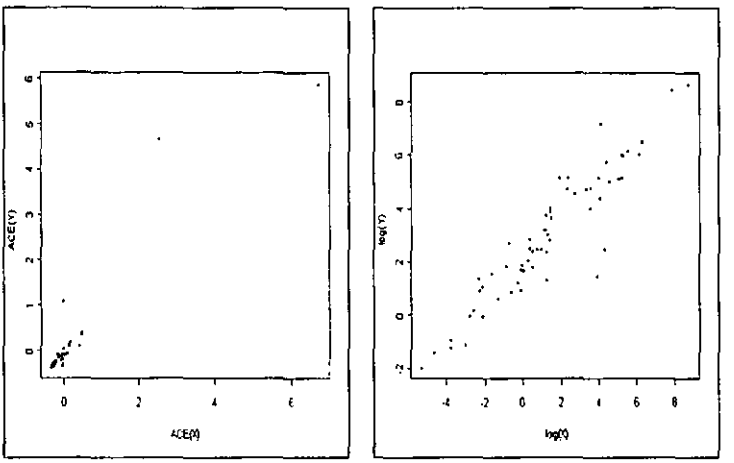

Fig. (3.9). a) $\phi^*(X)$ vs $\theta^*(Y)$; b) $log(X)$ vs $log(Y)$.

Este comportamiento es provocado por la existencia de dos datos aberrantes, los cuales alejan las sugerencias del ACE de las transformaciones logarítmicas. Si se eliminan los datos aberrantes los resultados alcanzados por el ACE son los siguientes:

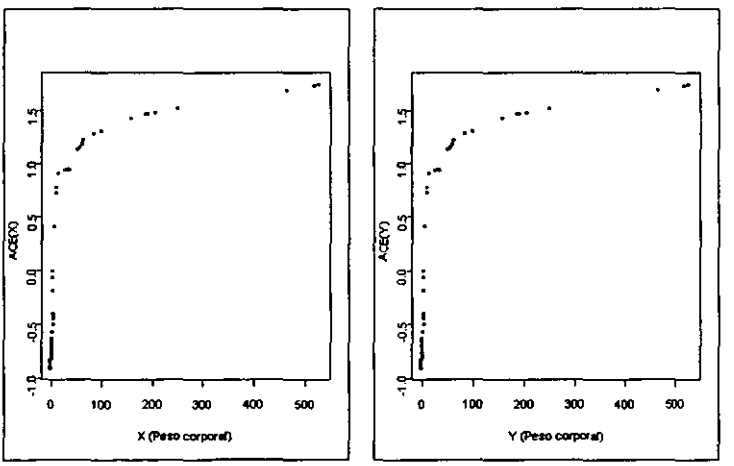

Fig. (3.10.1). a) $\phi^*(X')$; b) $\theta^*(Y')$.

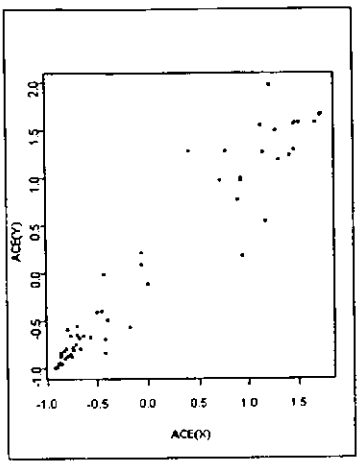

Fig. (3.10.2). $\phi^*(X')$ vs $\theta^*(Y')$.

que presentan una notable mejoría en la gráfica de variables transformadas. De esta forma se debe tener un especial cuidado al identificar datos extremos en una muestra a estudiar, de forma similar que con otros métodos tradicionales de regresión, debido a la extrema sensibilidad que el algoritmo presenta hacia ellos.

3.2 Modificaciones al Método ACE: Método AVAS

Como se ha mencionado anteriormente, el método ACE representa un avance significativo como herramienta para estimar transformaciones generales en regresión; si bien la idea del proceso se comprende mejor en el caso de variables aleatorias con una distribución conocida, su implantación a una serie de datos finita puede ser útil en un gran número de ocasiones. Sin embargo, por varias razones este método parece ser más recomendable para el análisis de correlación entre datos que para una regresión.

En principio, considérese un modelo fijo de la forma

$$\theta_0(Y) = \phi_0(X) + \varepsilon, \tag{3.1}$$

donde θ_0 es estrictamente monótona y ε es independiente de X, además de tener esperanza cero. Si se asume a θ_0 invertible, las distribuciones específicas de X y ε determinan una distribución conjunta de X y Y, digamos $f(X,Y)$. Una vez dada la distribución conjunta, las transformaciones $\phi^*(X)$ y $\theta^*(Y)$ que maximizan $corr[\phi(X), \theta(Y)]$ generalmente no son $\phi_0(Y)$ y $\theta_0(X)$. De esta manera el algoritmo no reproduce las transformaciones del modelo, debido a que existen otras transformaciones que mejoran el criterio de la correlación.

Adicionalmente, varios estadistas han señalado defectos del algoritmo como los presentados anteriormente, además de otros más complejos como la simetría del algoritmo en X y Y, la sensibilidad a la forma de la distribución marginal de X y el colapsamiento de conglomerados separados en la distribución conjunta de X y Y, los cuales no se detallarán.

Es importante conocer las limitaciones del algoritmo para poderlo aplicar correctamente a casos reales, aunque indudablemente el tratar de corregir los defectos es lo ideal.

La idea del ACE fue modificada por Robert Tibshirani, quien tres años después de que el ACE fuera dado a conocer, presentó su trabajo consistente en un algoritmo similar llamado AVAS (Additivity and Variance Stabilization), el cual no tiene como objetivo único el maximizar la correlación final, sino que requiere hacerlo entre las transformaciones que presenten una varianza estable. Este método presenta las virtudes alcanzadas por el ACE y al mismo tiempo está diseñado específicamente para regresión.

Recordando el proceso del ACE, la estimación de la transformación para X estaba dada inicialmente por $E[\theta(Y) \mid X]$ mientras que para la transformación de Y se utilizaba $E[\phi(X) \mid Y] \;/\; \|E[\phi(X) \mid Y]\|$ iterativamente. El algoritmo AVAS difiere escencialmente del ACE en que para estimar la transformación de Y utiliza la llamada "transformación asintótica de estabilización de varianza".

Esta modificación parece solucionar los problemas, en cuanto a regresión se refiere, atribuidos al ACE. En particular, para un modelo de la forma (3.1) las transformaciones $\theta_0(Y)$ y $\phi_0(X)$ son puntos fijos del AVAS. Adicionalmente, este algoritmo se comporta aceptablemente en otros problemas presentados por el ACE, tanto en datos generados como en reales.

3.3 Normalización de Familias y Estabilización de Varianza[2]

En esta sección se comentará el trabajo de Bradley Efron (1982), donde se presenta un resultado que se incorporará en la estructura del algoritmo ACE con el propósito de corregir el problema de inestabilidad de varianza en las transformaciones finales sugeridas.

En su artículo el autor plantea el siguiente problema: dada una familia de variables aleatorias, ¿es posible encontrar una transformación monótona tal que ser aplicada a cada elemento de la familia, la distribución que resulta sea normal? Aunque debe señalarse que el objetivo que interesa para este trabajo es obtener la "función de estabilización de varianza" más que los resultados de normalización.

[2] A lo largo de esta parte, por cuestión de nomenclatura estándar, se denotará por $\phi(\cdot)$ a la función de densidad de la distribución $N(0,1)$; no confundir con la anterior denotación de $\phi_i(\cdot)$ una transformación de la i-ésima variable explicativa de un modelo, con $\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_p)$ o con $\overline{\phi}(\cdot) = \sum_{i=1}^{p} \phi_i(\cdot)$.

Para iniciar, sea $\Im$ una familia de distribuciones indizadas por un parámetro λ, y sea Λ el espacio de parametrización de $\Im$, el cual puede ser un intervalo de la recta real. El objetivo es encontrar una transformación monótona $g(\cdot)$ tal que para cada $X_\lambda \in \Im$

$$Y_\lambda = g(X_\lambda) \sim N(v_\lambda, 1) \tag{3.2}$$

A lo largo de la sección se utilizará que para una variable aleatoria $Z \sim N(0,1)$ se cumple que

$$\begin{aligned} \phi(z) &= (2\pi)^{-1/2} \exp(\frac{-z^2}{2}), \\ \Phi(z) &= \int_{-\infty}^{z} \phi(u)du, \end{aligned}$$

donde ϕ es su función de densidad y Φ es su función de distribución acumulativa.

Con el propósito de poder apreciar el grado con el que una familia de distribuciones cumple con la forma de transformación normal dada por (3.2) y cuándo es posible encontrar $g(\cdot)$, respectivamente, se definirán varias familias ilustrando algunos casos particulares, así como una función diagnóstico para poder estudiar la existencia.

3.3.1 Tipos de Familias Transformables

Definición 3.3.1 *Se dice que $\Im$ es una Familia Normal Transformable (NTF) si existe una función $g(\cdot)$ estrictamente monótona tal que*

$$g(X_\lambda) \sim v_\lambda + Z,$$

con $Z \sim N(0,1)$ y v_λ (la media de la distribución normal de $g(X_\lambda)$) una función diferenciable con respecto a λ.

Definición 3.3.2 *Se dice que $\Im$ es una Familia General Transformable (GTF) si adicionalmente existe una función $q(\cdot)$ estrictamente creciente y diferenciable que cumpla con*

$$\begin{aligned} q(0) &= 0, \\ \dot{q}(0) &= 1, \end{aligned} \tag{3.3}$$

tal que

$$g(X_\lambda) \sim v_\lambda + q(Z).$$

Definición 3.3.3 *Se dice que* $\Im$ *es una Familia Normal Escalada Transformable (NSTF) si existe además una función diferenciable (con respecto a* λ*)* $\sigma_\lambda > 0$ *tal que*

$$g(X) \sim v_\lambda + \sigma_\lambda Z,$$

y en adelante se asumirá que $\dot{v}_\lambda = \frac{\partial v}{\partial \lambda} \neq 0$*, salvo en un número finito de valores de* λ*.*

Definición 3.3.4 *Se dice que* $\Im$ *es una Familia General Escalada Transformable (GSTF) si*

$$g(X) \sim v_\lambda + \sigma_\lambda q(Z). \tag{3.4}$$

3.3.2 Función Diagnóstico

Dada $\Im$, para una variable aleatoria $X_\lambda \in \Im$ sea

$$F_\lambda(x) = Prob_\lambda\{X_\lambda \leq x\}$$

la función de distribución acumulativa de X_λ.

En adelante se supondrá que existen las parciales

$$\begin{aligned} \dot{F}_\lambda(x) &= \frac{\partial F_\lambda(x)}{\partial \lambda}, \\ f_\lambda(x) &= \frac{\partial F_\lambda(x)}{\partial x}. \end{aligned}$$

Para poder definir la función diagnóstico, para $0 < \alpha < 1$ sea

$$x_{\alpha,\lambda} : F_\lambda(x_{\alpha,\lambda}) = \alpha,$$

por lo que $x_{\alpha,\lambda}$ es el $(100 \times \alpha)\%$ percentil de X bajo F_λ. En particular $x_{.5,\lambda}$ es la media de X. Es posible introducir ahora la **función diagnóstico**

$$D(z,\lambda) = \frac{\dot{F}_\lambda(x_{\Phi(z),\lambda})}{\dot{F}_\lambda(x_{.5,\lambda})} \frac{\phi(0)}{\phi(z)}. \tag{3.5}$$

Como se puede apreciar, $D(z,\lambda)$ está definida únicamente en términos de $F_\lambda(x)$, por lo que es posible evaluarla sin tener que conocer $g(x)$, o inclusive sin suponer que ésta existe.

La definición de $D(z,\lambda)$ puede expresarse en función de la llamada **transformación local de normalidad,** dada por

$$t_\lambda(x) \equiv \Phi^{-1}\{F_\lambda(x)\}. \tag{3.6}$$

Bajo el valor λ del parámetro, t_λ se distribuye $N(0,1)$. De esta forma (3.5) toma la forma

$$D(z,\lambda) = \frac{\dot{t}_\lambda(x_{\Phi(z),\lambda})}{\dot{t}_\lambda(x_{.5,\lambda})}, \tag{3.7}$$

donde $\dot{t}_\lambda(x) \equiv \frac{\partial t_\lambda(x)}{\partial \lambda}$. Explicando brevemente lo anterior, $D(z,\lambda)$ mide qué tan rápido $D(z,\lambda)$ cambia a medida que λ varía.

Ahora es posible pasar al siguiente

Teorema 3.3.1 *Sea $\Im$ una GSTF. Entonces la función diagnóstico cumple*

$$D(z,\lambda) = \frac{1 + q(z)\epsilon_\lambda}{q'(z)},$$

donde

$$\epsilon_\lambda = \frac{\dot{\sigma}_\lambda}{\dot{v}_\lambda} = \frac{\partial \sigma_\lambda/\partial\lambda}{\partial v_\lambda/\partial\lambda}.$$

Demostración:

Se sabe que la función de distribución acumulativa de $\tilde{Z} = q(Z)$ es

$$\tilde{\Phi}(\tilde{z}) = \Phi(q^{-1}(\tilde{z})). \tag{3.8}$$

Si z_α es el $(100 \times \alpha)\%$ percentil normal, entonces

$$\Phi(z_\alpha) = \alpha,$$

y $\tilde{z}_\alpha = q(z_\alpha)$ es el correspondiente percentil de $(100 \times \alpha)\%$ de $\tilde{Z}$. En particular $\tilde{Z}$ tiene media $q(z_{.5}) = q(0) = 0$, por (3.3). La función de densidad de $\tilde{Z}$, $\tilde{\phi}(\tilde{z}) = \tilde{\Phi}'(\tilde{z})$ satisface

$$\tilde{\phi}(\tilde{z}) = \frac{\phi(z)}{q'(z)}, \tag{3.9}$$

y por (3.3)

$$\tilde{\phi}(0) = \phi(0) = \frac{1}{\sqrt{2\pi}}. \tag{3.10}$$

Más aún, por (3.4) $F_\lambda(x) = \tilde{\Phi}(\frac{g(x)-v_\lambda}{\sigma_\lambda})$, lo cual implica que

$$\dot{F}_\lambda(x) = -\tilde{\phi}\left(\frac{g(x)-v_\lambda}{\sigma_\lambda}\right)\left(\frac{\dot{v}_\lambda}{\sigma_\lambda} + \frac{g(x)-v_\lambda}{\sigma_\lambda}\frac{\dot{\sigma}_\lambda}{\sigma_\lambda}\right).$$

Usando nuevamente (3.4) y el hecho de que los percentiles son mapeados de la manera obvia bajo transformaciones monótonas, es decir

$$z_\alpha = \frac{g(x_{\alpha,\lambda}) - v_\lambda}{\sigma_\lambda},$$

se cumple

$$\dot{F}_\lambda(x_{\alpha,\lambda}) = -\tilde{\phi}(\tilde{z}_\alpha)\left(\frac{\dot{v}_\lambda}{\sigma_\lambda} + \tilde{z}_\alpha\frac{\dot{\sigma}_\lambda}{\sigma_\lambda}\right) = -\frac{\phi(z_\alpha)}{q(z_\alpha)}\left(\frac{\dot{v}_\lambda}{\sigma_\lambda} + q(z_\alpha)\frac{\dot{\sigma}_\lambda}{\sigma_\lambda}\right), \tag{3.11}$$

por (3.9).

Utilizando (3.11) y (3.3) en el caso de $x_{.5,\lambda}$ se obtiene:

$$\dot{F}_\lambda(x_{.5,\lambda}) = -\frac{\phi(z_{.5})}{q(z_{.5})}\left(\frac{\dot{v}_\lambda}{\sigma_\lambda} + q(z_{.5})\frac{\dot{\sigma}_\lambda}{\sigma_\lambda}\right) = -\phi(0)\left(\frac{\dot{v}_\lambda}{\sigma_\lambda}\right). \tag{3.12}$$

Sustituyendo (3.11) y (3.12) en (3.7) se llega a lo siguiente:

$$\begin{aligned} D(z,\lambda) &= \frac{\dot{t}_\lambda(x_{\Phi(z),\lambda})}{\dot{t}_\lambda(x_{.5,\lambda})} = \frac{\dot{F}_\lambda(x_{\Phi(z),\lambda})}{\dot{F}_\lambda(x_{.5,\lambda})}\frac{\phi(0)}{\phi(z)} = \frac{-\frac{\phi(z)}{q(z)}\left(\frac{\dot{v}_\lambda}{\sigma_\lambda} + q(z)\frac{\dot{\sigma}_\lambda}{\sigma_\lambda}\right)}{-\phi(0)\frac{\dot{v}_\lambda}{\sigma_\lambda}}\frac{\phi(0)}{\phi(z)} = \\ &= \frac{\frac{1}{q(z)}\left(\frac{\dot{v}_\lambda}{\sigma_\lambda} + q(z)\frac{\dot{\sigma}_\lambda}{\sigma_\lambda}\right)}{\frac{\dot{v}_\lambda}{\sigma_\lambda}} = \frac{\left(\frac{\dot{v}_\lambda}{\sigma_\lambda} + q(z)\frac{\dot{\sigma}_\lambda}{\sigma_\lambda}\right)/\frac{\dot{v}_\lambda}{\sigma_\lambda}}{q(z)} = \frac{1 + q(z)\frac{\dot{\sigma}_\lambda}{\dot{v}_\lambda}}{q(z)}, \end{aligned}$$

que es el resultado buscado.■

3.3.3 Función de Estabilización de Varianza

Como objetivo principal de esta sección, se desea encontrar $g(X)$ en una GTF que cumpla con

$$g(X) = v_\lambda + \sigma_\lambda q(Z).$$

Primeramente, sean x_1, x_2 dos valores arbitrarios de X tales que $x_1 < x_2$. Sea $\lambda_{1,2}$ el valor del parámetro λ tal que $F_{\lambda_{1,2}}(x_1) = 1 - F_{\lambda_{1,2}}(x_2)$, digamos

$$\alpha = F_{\lambda_{1,2}}(x_1) = 1 - F_{\lambda_{1,2}}(x_2). \tag{3.13}$$

En seguida defínase $\lambda_x = \mu^{-1}(x)$ como el valor de λ tal que x es la media de X :

$$F_{\lambda_x}(x) = .5,$$

y sea

$$f_\lambda(x) = F'_\lambda(x)$$

la función de densidad de X.

Teorema 3.3.2 *En una GTF se cumple*

$$g'(x) = \frac{f_{\lambda_x}(x)}{\phi(0)}. \tag{3.14}$$

Demostración:

Recordando que $F_\lambda(x) = \tilde{\Phi}\left(\frac{g(x)-v_\lambda}{\sigma_\lambda}\right)$, donde $\tilde{\Phi}(\tilde{z}) = \Phi(q^{-1}(\tilde{z}))$ como se menciona en (3.8), y derivando con respecto a x se obtiene que

$$f_\lambda(x) = \tilde{\phi}\left(\frac{g(x)-v_\lambda}{\sigma_\lambda}\right)\left(\frac{g'(x)}{\sigma_\lambda}\right).$$

Evaluando $f_\lambda(x)$ en $x_{\alpha,\lambda}$ se tiene

$$f_\lambda(x_{\alpha,\lambda}) = \tilde{\phi}(\tilde{z}_\alpha)\left(\frac{g'(x_{\alpha,\lambda})}{\sigma_\lambda}\right) = \frac{\phi(z_\alpha)g'(x_{\alpha,\lambda})}{q'(z_\alpha)\sigma_\lambda}. \tag{3.15}$$

Tomando $\theta = \mu^{-1}(x) = \lambda_x$ y $\alpha = .5$ se tiene que $x = x_{\alpha,\lambda}$ y $z_\alpha = 0$. Como $\sigma_\lambda = 1$ en una GTF y $q'(0) = 1$ por (3.3), (3.15) implica que $g'(x) = f_{\lambda_x}(x)/\phi(0)$ como se requería. ■

Una interpretación intuitiva de la fórmula (3.14) en términos de la transformación local de normalidad $t_\lambda(x) = \Phi^{-1}\{F_\lambda(x)\}$ es la siguiente: en una NTF, $X = g^{-1}(v_\lambda + Z)$, se tiene $t_{\lambda_0}(x) = \Phi^{-1}\{\Phi(g(x) - v_{\lambda_0})\} = g(x) - v_{\lambda_0}$, por lo que

$$t'_{\lambda_0}(x) = g'(x). \tag{3.16}$$

Lo anterior significa que no importa el valor λ_0 que se escoja, en una NTF $t'_{\lambda_0}(x)$ siempre coincide con el valor de $g'(x)$. En otras palabras, cualquier transformación local de normalidad normaliza globalmente una NTF.

No obstante la necesidad del supuesto de tratarse de una NTF para que (3.16) se cumpla, es posible tratar de escoger entre las diferentes transformaciones $t'_{\lambda_x}(x)$ seleccionando el valor de λ más apropiado para cada x. Una elección obvia es $\lambda = \lambda_x = \mu^{-1}(x)$, con $t_\lambda(x)$ cuya derivada $t'(x)$ cumpla

$$t'_{\lambda_x}(x) = f_{\lambda_x}(x)/\phi(0),$$

que coincide con (3.14). Es decir, $\bar{g}(x)$ es la transformación que tiene la misma derivada (con respecto a X) que $t_\lambda(x)$, evaluada en el valor λ para el cual x es la media de X.

La razón por la cual $\bar{g}(x)$ es conocida como transformación estabilizadora de varianza es la siguiente: en una GTF, donde es posible una perfecta estabilización, $\bar{g}(x)$ la alcanza. $Y = \bar{g}(X) = v_\lambda + Z$ es una familia de traslación con varianza constante.

El siguiente corolario muestra que $\bar{g}(x)$ trata de estabilizar la varianza en un contexto más general que una GTF.

Corolario 3.3.1 *Si $\Im$ es una GSTF, $X = g^{-1}(v_\lambda + \sigma_\lambda \tilde{Z})$ entonces*

$$\bar{g}'(x) = g'(x)/\sigma_\lambda. \tag{3.17}$$

Demostración:

Tomando $\alpha = .5$ en (3.15) se tiene

$$f_{\lambda_x}(x) = \frac{g'(x)}{\sigma_{\lambda_x}}\phi(0),$$

y el corolario se sigue directamente de (3.14).■

La interpretación de (3.17) es la siguiente. Inicialmente hay que obtener la transformación $Y = g(X)$, la cual produce que $Y = v_\lambda + \sigma_\lambda \tilde{Z}$. Posteriormente se aplica (3.14) a esta familia. Como Y tiene densidad $\frac{1}{\sigma_\lambda}\phi\{\frac{y-v_\lambda}{\sigma_\lambda}\}$, (3.14) obtiene la transformación de Y, $h(y)$, con derivada en $y = g(x)$ igual a

$$h'(y) = 1/\sigma_\lambda.$$

Aquí se ha usado (3.10) y el hecho de que $Prob_\lambda\{X < x\} = .5$ (i.e. si $\lambda = \mu^{-1}(x)$); entonces $Prob_\lambda\{Y < y\} = .5$ (i.e. $\lambda = v^{-1}(y)$). De acuerdo a (3.17), la transformación $\bar{g}(x)$ es la composición $h \circ g(x)$.

En el caso de una NSTF, $X = g^{-1}(v_\lambda + \sigma_\lambda Z)$, $Y = g(X)$ es perfectamente normal, $Y \sim N(\lambda, \sigma_\lambda^2)$,pero con varianza no constante. Entonces (3.14) realiza una transformación adicional $W = h(Y)$ donde $h'(y) = 1/\sigma_\lambda$, que descompone la normalidad pero tiende a producir una varianza más constante.

La anterior propiedad es la que motivó a incorporar $h(\cdot)$ en el algoritmo AVAS. Sin tener más sentido para efectos de este trabajo comentar el resto de la discusión realizada por Efron, se procederá a enunciar el algoritmo en sí.

3.4 Algoritmo AVAS

Para ser consistentes, inicialmente se discutirá el algoritmo en el caso de distribuciones conocidas y se proseguirá con la implementación del método a datos.

Tomando primero el caso univariado, dadas dos variables aleatorias X y Y el problema ahora consiste en encontrar dos transformaciones reales medibles $\phi(X)$ y $\theta(Y)$ que satisfagan las condiciones

$$E[\theta(Y) \mid X] = \phi(X) \tag{3.18}$$

$$Var[\theta(Y) \mid \phi(X)] = k \tag{3.19}$$

con k constante y $\theta(Y)$ estrictamente monótona. Sin pérdida de generalidad se puede asumir que $\theta(Y)$ es estrictamente creciente.

Al final del proceso se requiere llegar a un modelo de la forma $\theta(Y) = \phi(X) + \varepsilon$. Por el teorema de proyección, la condición (3.18) implica que para cualquier otra transformación $\theta_0(X)$, $corr[\theta_0(Y), \phi(X)] \leq corr[\theta(Y), \phi(X)]$. Por su parte la condición (3.19) asegura la homogeneidad en la varianza del modelo.

La idea básica detrás del algoritmo es la siguiente: si $\theta(Y)$ es conocida, una estimación de $\phi(X)$ para cumplir la primera condición es $\phi(X) = E[\theta(Y) \mid X]$. Para inicializar el algoritmo, simplemente se utiliza Y para encontrar la primera estimación de $\phi(X)$.

Para cumplir la segunda condición, se utilizan los resultados de la sección anterior: sea $\Im$ una familia de distribuciones con $v(\cdot)$ la esperanza de la familia y $\sigma^2(v)$ su varianza. Entonces la transformación estabilizadora de varianza para una variable aleatoria X_λ, $\lambda \in \Lambda$ el conjunto de parámetros de la familia, está dada por:

$$h(t) = \int_0^t 1/\sqrt{\sigma^2(v)}dv. \tag{3.20}$$

Como observación se tiene que si $\theta(Y)$ es solución al problema, entonces $Var[\theta(Y) \mid \phi(X)]$ es constante, y $h(\cdot)$ es igual a la identidad, por lo que $\theta(Y)$ permanece sin cambio. De otra forma la estimación que resulta, que puede denotarse como $h(\theta(Y))$, debería tener una varianza más estable como función de v que $\theta(Y)$.

Al iterar los dos pasos anteriores se obtiene el algoritmo AVAS. Sin embargo se debe tener en cuenta que si $\theta(Y)$ y $\phi(X)$ son soluciones, entonces $a + b\theta(Y)$ y $a + b\phi(X)$ también lo son, por lo que además se imponen las condiciones

$$\begin{aligned} E[\phi(X)] &= 0, \\ E[\theta(Y)] &= 0, \\ Var[\theta(Y)] &= 1, \end{aligned}$$

para asegurar la unicidad.

Una vez que se ha dado la idea general, se procede a enunciar el algoritmo.

Dadas dos variables aleatorias X y Y con distribuciones conocidas:

-Asignar $\theta(Y) = (Y - E[Y])/\|Y\|$;

-Mientras $e^2(\theta, \phi)$ disminuya:
-Asignar $\phi(X) = E[\theta(Y) \mid X]$;
-Obtener la transformación de Y :
Calcular $\sigma^2(v) = var[\theta(Y) \mid \phi(X) = v]$;
Calcuar la transformación estabilizadora de varianza $h(t) = \int_0^t [\sigma^2(v)]^{-1/2} dv$;
Asignar $\theta(t) = h(\theta(t))$;
Estandarizar $\theta(t) = (\theta(t) - E[\theta(Y)]) / \|\theta(Y)\|$;

-θ y ϕ son las soluciones θ^* y ϕ^*;
-Fin del algoritmo.

De esta manera, en cada paso del algoritmo inicialmente se aumenta la correlación entre las transformaciones al estimar la transformación de la variable dependiente y posteriormente, con la estimación para la transformación de la variable independiente, el modelo alcanza una variabilidad aproximadamente homogenea en las transformaciones.

Para el caso multivariado se utiliza el mismo proceso del ACE, al incluir un bucle interno dentro del bucle principal del algoritmo, utilizando en éste el paso de estabilización de varianza en la estimación de cada variable independiente. Asimismo, el método AVAS también puede incorporar variables independientes continuas o categóricas, o una mezcla de las dos. Suponiendo que las variables independientes son $X_1, X_2, ..., X_p$ asúmase de nueva cuenta un modelo aditivo de estas variables para la esperanza de $\theta(Y)$. De esta forma modelo en el caso multivariado toma la forma

$$\theta(Y) = \sum_{i=1}^{p} \phi_i(X_i) + \varepsilon,$$

con $\theta(Y)$ estrictamente creciente y ε con media cero e independiente de X_i, con las restricciones

$$\begin{aligned} E[\phi_i(X_i)] &= 0,\ i = 1, 2, ..., p, \\ E[\theta(Y)] &= 0, \\ Var[\theta(Y)] &= 1. \end{aligned}$$

Asimismo, las soluciones deben cumplir con

$$Var[\theta(Y) \mid \sum_{i=1}^{p} \phi_i(X_i)] = k.$$

Para incluir la posibilidad de múltiples variables explicativas sólo es necesario modificar el paso donde se estima la transformación de X, añadiendo el ciclo para estimar la transformación de cada X_i de la misma forma que en el algoritmo ACE. De esta forma, en cada paso del "bucle interior" se fija una variable, digamos X_j, y se obtiene

$$\phi_j(X_j) = E[\theta(Y) - \sum_{i \neq j} \phi_i(X_i) \mid X_j = x_j], \quad j = 1, 2, ..., p.$$

Para simplicidad de notación, denótese nuevamente a $\overline{\phi}(\mathbf{X}) = \sum_{i=1}^{p} \phi_i(X_i)$. Así, el algoritmo para múltiples regresores está dado por:

Dadas las distribuciones de $Y, X_1, X_2, ..., X_p$:

-Asignar $\theta(Y) = (Y - E[Y])/\|Y\|$;

-Mientras $e^2(\theta, \phi_1, ..., \phi_p)$ disminuya:

-Para $i = 1, 2, ..., p$ obtener la transformación de X_i:
Definir $\phi_{i,0}(X_i) = E[\theta(Y) - \sum_{j \neq i} \phi_j(X_j) \mid X_i)]$;
Asignar $\phi_i(X_i) = \phi_{i,0}(X_i)$;

-Obtener la transformación de Y :
Calcular $\sigma^2(v) = var[\theta(Y) \mid \overline{\phi}(\mathbf{X}) = v]$;
Calcuar la transformación estabilizadora de varianza $h(t) = \int_0^t [\sigma^2(v)]^{-1/2} dv$;
Asignar $\theta(t) = h(\theta(t))$;
Estandarizar $\theta(t) = (\theta(t) - E[\theta(Y)]) \,/\, \|\theta(Y)\|$;

-θ y ϕ_i son las soluciones θ^* y ϕ_i^*, $i = 1, 2, ..., p$;
-Fin del algoritmo.

3.5 Observaciones sobre el AVAS

Debido a que la teoría anterior no es del todo formal, a continuación se mencionan los supuestos asumidos y algunos aspectos a tomar en cuenta al implantar el método.

1. Al introducir el algoritmo, se asume que la varianza $\sigma^2(v)$ existe, y además cumple $0 < \sigma^2(v) < \infty$. Adicionalmente se asume que los momentos $E[\theta(Y) \mid X_i]$, $E[\theta(Y)]$ y $var[\theta(Y)]$ existen y son finitos. Desafortunadamente, como lo menciona el autor, parece complicado encontrar un conjunto de condiciones que puedan asegurar que tales supuestos se cumplen en alguno de los pasos del AVAS.

2. Hay que hacer notar que el algoritmo ACE no impone que la función $\theta(Y)$ sea monótona. Esto hace sentido debido a que la transformación óptima para Y dada la distribución conjunta de las X_i y Y puede ser no monótona. En el presente algoritmo, sin embargo, se asume que el modelo

$$\theta(Y) = \sum_{i=1}^{p} \phi_i(X_i) + \varepsilon$$

se cumple, con $\theta(Y)$ estrictamente creciente. Acertadamente la estimación obtenida por el AVAS cumple esta condición, debido a que es la integral de una función positiva.

3. Una ventaja del AVAS radica en que produce soluciones independientemente de la distribución marginal de X. En particular, si a X_i se le aplica una transformación monótona $f(\cdot)$, entonces, ignorando escalas, la transformación sugerida por el AVAS en cada iteración $\phi_i(X_i)$ es mapeada de la manera esperada, $\phi_i(f(X_i))$, lo cual no sucede en el ACE. Nótese que en el caso discreto esta equivarianza sólo se cumple aproximadamente, debido al uso de suavizadores..

4. Puede presentarse que existan valores de v para los cuales $\sigma(v)$ no esté definida. En el caso de que $\overline{\phi}(\mathbf{X})$ esté acotada y Y tenga varianza finita, entonces la fórmula para la transformación asintótica $h(t)$ involucra valores de v fuera del rango de definición de $\sigma^2(v)$. En este caso se define $\sigma^2(v)$ por interpolación o extrapolación lineal del integrando $[\sigma^2(v)]^{1/2}$. Esto cobra extrema importancia cuando X toma sólo valores discretos, pues la fórmula para $h'(t) = [\sigma^2(t)]^{1/2}$ sólo se cumple en valores t para los cuales $\sigma^2(t)$ está definida. Suponiendo que la variable X es binaria, por ejemplo, la derivada de h sólo está definida en dos puntos. Es necesario entonces asumir que la derivada $h'(t)$ es lineal entre los puntos donde está definida.

5. El AVAS busca transformaciones tales que $var[\theta^*(Y) \mid \sum_{i=1}^{p} \phi_i^*(X_i)]$ sea constante. Si $\theta(Y) = \sum_{i=1}^{p} \phi_i(X_i)$ tiene una distribución que no involucra a las X_i entonces $var[\theta(Y) \mid X_1, X_2, ...X_p]$ y $var[\theta(Y) \mid \sum_{i=1}^{p} \phi_i(X_i)]$ son constantes. Como la implicación contraria no es obligada, es conveniente graficar $\theta^*(Y)$ contra una o más $\phi_i(X_i)$ y verificar si $var[\theta^*(Y) \mid X_1, X_2, ...X_p]$ es constante. De ser así, habrá una fuerte sugerencia de que el modelo es adecuado.

6. Para implantar el AVAS al caso discreto se procede análogamente al ACE. Para reemplazar las esperanzas condicionales se utiliza el supersuavizador mencionado en el capítulo anterior, lo cual permite comparaciones con el ACE.

   Las esperanzas y varianzas necesarias son sustituidas por sus versiones muestrales.

Para calcular la función $\sigma^2(v)$, se requiere de un proceso más complicado[3], el cual sólo se mencionará. Inicialmente, se suaviza la versión muestral de los residuales $r^2 = E[\theta(Y) \mid \sum_{i=1}^{p} \phi_i(X_i)]^2$ contra $v = \sum_{i=1}^{p} \phi_i(X_i)$, y posteriormente se les aplica la función exponencial, utilizando un suavizador de rectas móviles, el cual es menos suceptible a la correlación en los residuales que el supersuavizador, según comenta el autor. Se escoge la escala logarítmica pues a su juicio es más natural para la varianza que la escala original, además de que asegura que la función $\sigma^2(v)$ será no negativa.

Finalmente, la integral es reemplazada por la ley del trapezoide, la cual presenta buenos resultados, aunque pueden utilizarse métodos de integración más sofisticados.

Por último se presentan algunas observaciones[4] referentes a la relación que guardan el AVAS y la función estabilizadora de varianza.

1. El trabajo de Efron se relaciona con el AVAS al considerar a $\Im$ como la familia de transformaciones indizadas por su media $E[Y|X = x]$, por lo que la definición de $\lambda_y = \mu(y)$ se tiene que $\lambda_y = y$. De esta forma el algoritmo puede corresponder al método de estabilización de varianza de esta familia particular.

2. Dada una distribución conjunta $\mathcal{L}(X_i, Y)$, transformaciones que satisfagan

$$\theta(Y) = \sum_{i=1}^{p} \phi_i(X_i) + \varepsilon,$$

donde $E[\varepsilon] = 0$ y ε es independiente de las X_i, no necesariamente existen, pero de existir son únicas. Lo anterior es consecuencia directa del trabajo de Efron donde se observa que no todas las GSTF son GTF, y que en una GTF la transformación $g(\cdot)$ es única. Ahora considérese la pregunta ¿dada una distribución conjunta $\mathcal{L}(X_i, Y)$ existen transformaciones que satisfagan

$$\begin{aligned} E\left[\theta(Y) - \sum_{i \neq j} \phi_i(X_i) | X_j = x_j\right] &= \phi_j(x_j), \\ Var\left[\theta(Y) | \sum_{i=1}^{p} \phi_i(X)\right] &= k, \end{aligned} \tag{3.21}$$

con $E[\theta(Y)] = 0$ y $Var[\theta(Y)] = 1$? En dado caso, ¿son únicas? Para contestar lo anterior considérese un ejemplo donde X toma los valores 0 o 1 y la distribución conjunta $\mathcal{L}(X, Y)$ permite que Y tome los valores de 0 o 1 con igual probabilidad para $X = 0$, y que tome los valores de 0, 1 ó 2 con igual probabilidad para $X = 1$. Entonces para toda transformación $\theta(\cdot)$ estrictamente creciente, $E[\theta(Y)|X = 1] > E[\theta(Y)|X = 0]$. Por tanto, si $\phi(\cdot)$ es solución de (3.21), $\phi(1) > \phi(0)$.

[3]Tibshirani, R. (1988). "Estimating Transformations for Regression Via Additivity and Variance Stabilization". Journal of the American Statistical Association, 83, pag. 397.

[4]Ibidim, pag. 402.

Pero es claro que $var[\theta(Y)|\phi(X) = \phi(1)] = var[\theta(Y)|X = 1] > var[\theta(Y)|\phi(X) = \phi(0)] = var[\theta(Y)|X = 0]$, y en este caso no existe solución. Sin embargo a continuación se planteará una conjetura alterna a la existencia de soluciones.

3. Intuitivamente el método AVAS parece minimizar de manera general

$$E\left[\theta(Y) - \sum_{i=1}^{p} \phi_i(X)\right]^2 \tag{3.22}$$

sujeto a las condiciones $Var\left[\theta(Y)|\sum_{i=1}^{p} \phi_i(X)\right] = k$, $E[\theta(Y)] = 0$, $Var[\theta(Y)] = 1$.

La estimación de las $\phi_i(X_i)$ minimiza (3.22) mientras que la estimación de $\theta(Y)$ procura las condiciones anteriores. En cualquier ciclo, $Var\left[\theta(Y)|\sum_{i=1}^{p} \phi_i(X)\right]$ no es necesariamente constante, pero si el método converge, $Var\left[\theta(Y)|\sum_{i=1}^{p} \phi_i(X) = u\right]$ debe ser constante.

El autor menciona en el trabajo original que ha identificado una serie de condiciones para $\mathcal{L}(X_i, Y)$ bajo las cuales existe una solución única a (3.22), sin haber podido demostrar tal conjetura a la fecha de publicación.

3.6 Ejemplos

Ejemplo 3.6.1 *Ejemplo introductorio.*

Para este ejemplo se tomará un sencillo modelo para mostrar el comportamiento del AVAS, considerando una relación de la forma

$$\log(Y) = X^2 + Z,$$

con X una muestra tamaño 200 generada de una distribución U(-2,2) y Z una muestra tamaño 200 de la distribución N(0,1).

Nótese que los valores que pueden tomar X^2 y Z puede provocar que el valor de $\log(Y)$ quede indefinido, por lo que el modelo es ajustado a

$$\log(Y) = X^2 + Z + 2.$$

La figura siguiente muestra las transformaciones sugeridas por el AVAS, donde se aprecia la forma de las transformaciones originales $\phi(X) = X^2$ y $\theta(Y) = \log(Y)$.

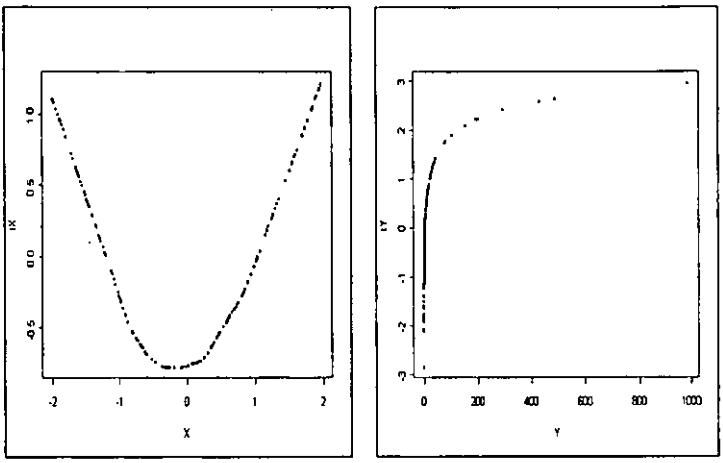

Fig. (3.11). a) $\phi^*(X)$; b) $\theta^*(Y)$.

A continuación se muestran las gráficas de los datos transformados que sugieren tanto el algoritmo AVAS como el ACE, respectivamente. Como se puede observar, en la primera gráfica los datos transformados presentan una mayor homogeneidad en la varianza, como resultado de la modificación del algoritmo para alcanzar este objetivo.

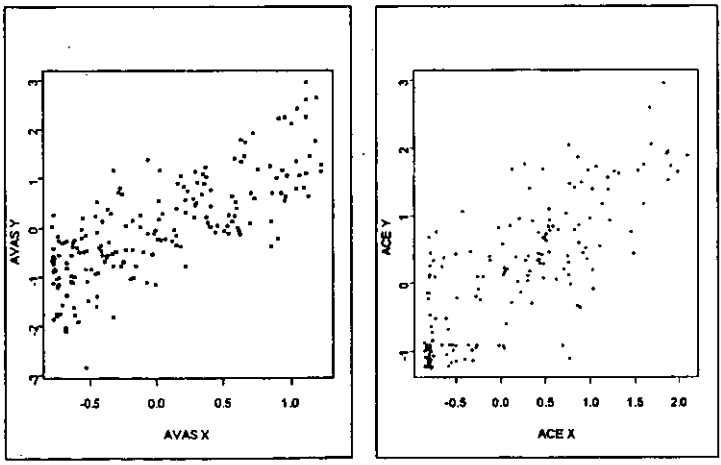

Fig. (3.12). $\phi^*(X)$ vs. $\theta^*(Y)$. a) AVAS; b) ACE.

Ejemplo 3.6.2 *Varianza inestable.*

Retomando el Ejemplo (3.1.1), un modelo de la forma

$$Y = X + (.1X)Z,$$

con X una muestra aleatoria de tamaño 200 de la distribución $U(0,1)$ y Z una muestra del mismo tamaño de la distribución normal estándar. En el caso del ACE, las transformaciones sugeridas eran similares a un linea recta, lo cual daba como resultado en que la varianza en el modelo transformado no era homogenea. Al aplicar el AVAS a la misma serie de datos las transformaciones sugeridas no son rectas, como se muestra a continuación.

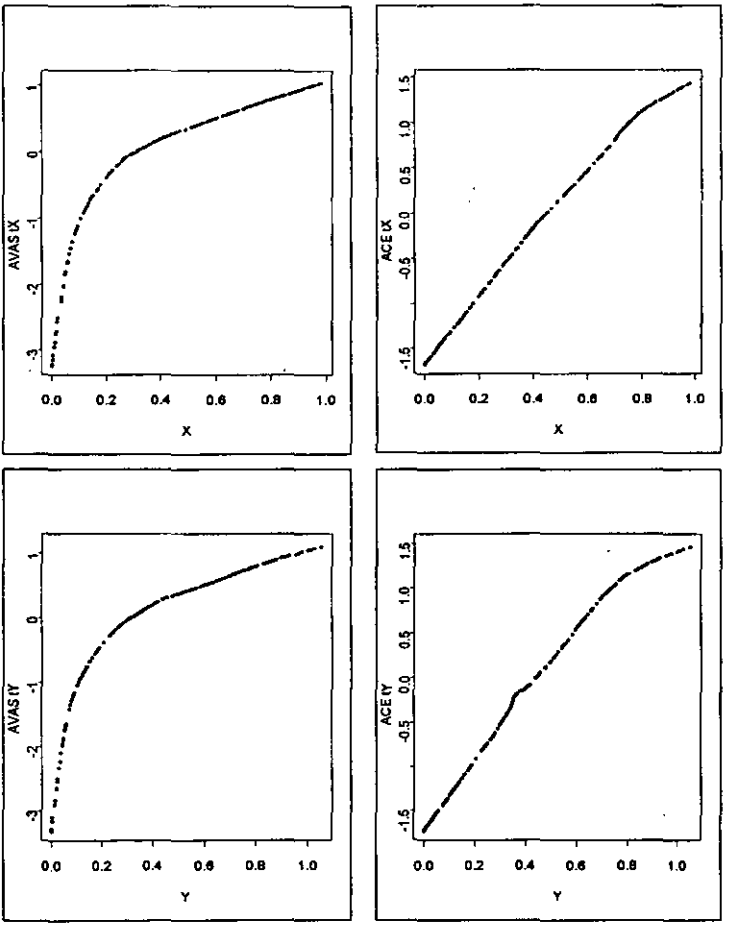

Fig. (3.14). $\phi^*(X)$ y $\theta^*(Y)$. **Izquierda:** AVAS. **Derecha:** ACE.

La estabilidad en la varianza alcanzada con estas transformaciones puede observarse en la gráfica de los datos ajustados. La siguiente figura presenta las gráficas de θ^* contra ϕ^* sugeridas por el AVAS y el ACE.

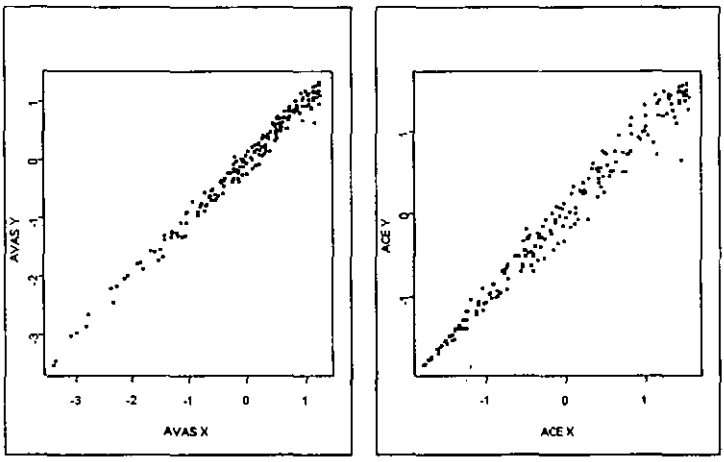

Fig. (3.15). $\phi^*(X)$ vs. $\theta^*(Y)$. a) AVAS; b) ACE.

Se verifica que la estabilidad de la varianza alcanzada por las transformaciones no lineales del AVAS es notablemente superior a la alcanzada por las rectas del ACE.

Ejemplo 3.6.3 *Escasa relación de las variables dependientes.*

Retomando otro ejemplo presentado con anterioridad, se genera una muestra aleatoria de las variables independientes X_i, $i = 1, 2, ..., 10$, y de Z de tamaño 100, donde cada variable se distribuye $N(0, 1)$. Como se vió anteriormente, al aplicar el ACE al modelo dado por

$$Y = X_1 + X_2 + ... + X_{10} + Z$$

podía generar transformaciones poco lineales, sobre todo cuando el tamaño de la muestra se reducía.

Los resultados presentados a continuación comparan las funciones dadas por el AVAS y por el ACE de cuatro variables independientes.

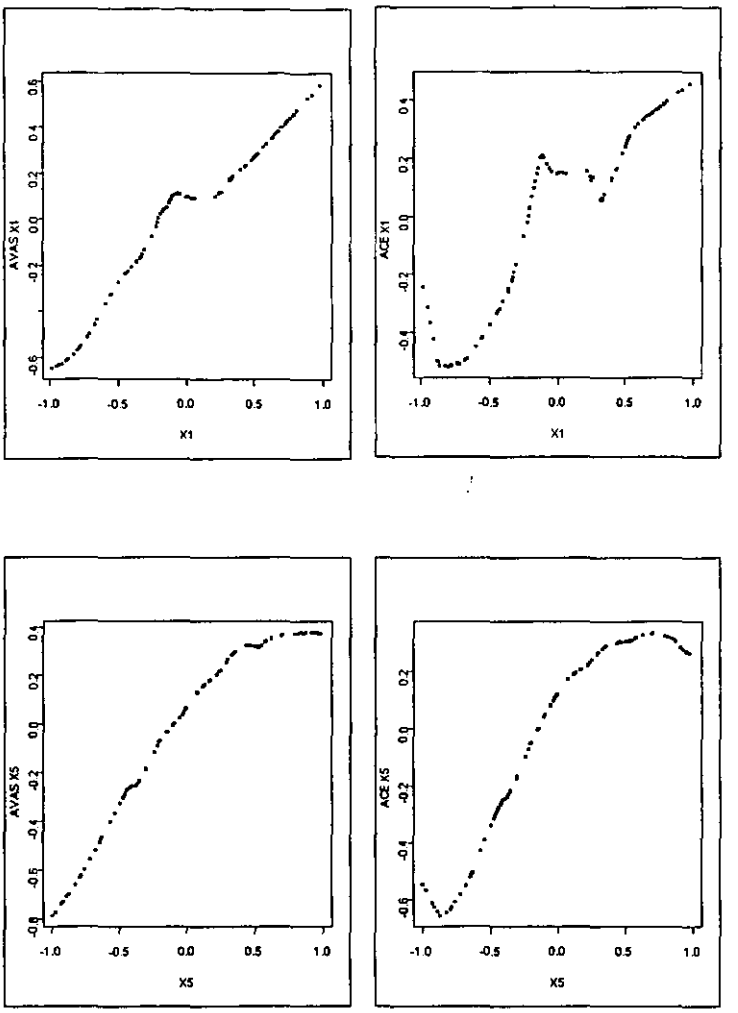

Fig. (3.16.1). $\phi_i^*(X_i)$, $i = 1, 5$. **Izquierda:** AVAS. **Derecha:** ACE.

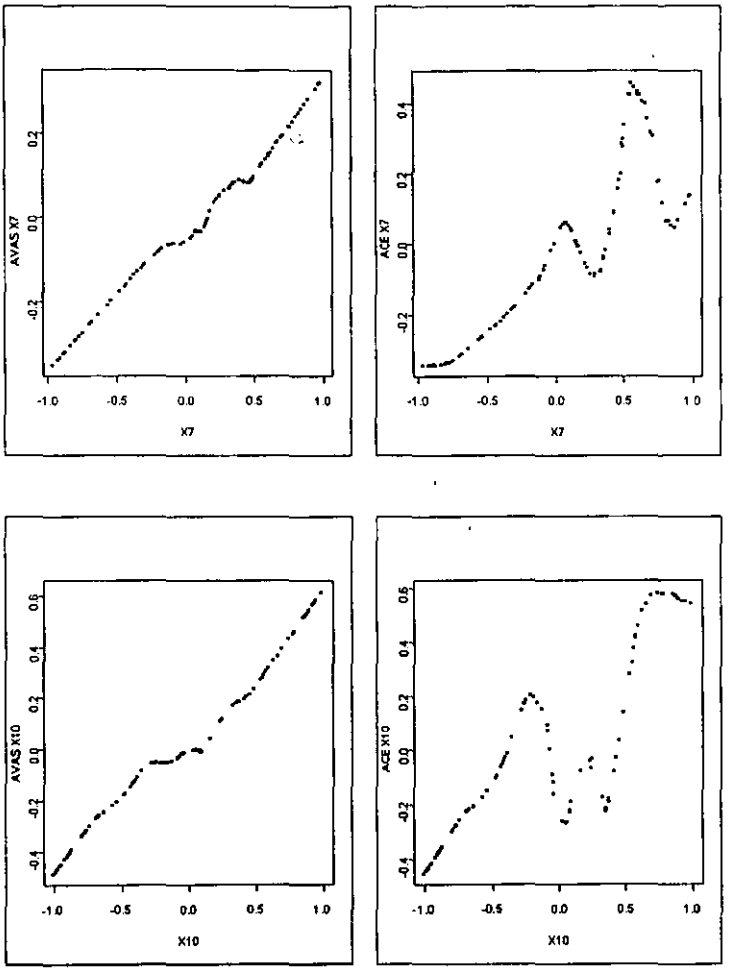

Fig. (3.16.2). $\phi_i^*(X_i)$, $i = 7, 10$. **Izquierda:** AVAS. **Derecha:** ACE.

Se percibe que al buscar que la varianza sea más estable, el AVAS presenta transformaciones más lineales que aquellas del ACE, por lo que el primero se comporta mejor al existir una baja relación entre las variables independientes en este caso.

Ejemplo 3.6.4 *Sensibilidad a datos aberrantes.*

En el Ejemplo (3.1.4) se aplicó el ACE a una serie de datos correspondientes al peso corporal y del cerebro de 62 mamíferos para establecer una relación entre ellos. Al analizar los resulados se notó que el modelo final sugerido por el algoritmo presentaba una exagerada variabilidad en los residuales, mientras que empíricamente una transformación de la forma log – log parecía muy adecuada. Según los autores del ACE, este comportamiento obedece a la presencia de dos observaciones aberrantes; al remover estos datos la mejoría era notable.

Al aplicar el AVAS a los datos completos se obtuvieron las gráficas incluidas en la figura a continuación:

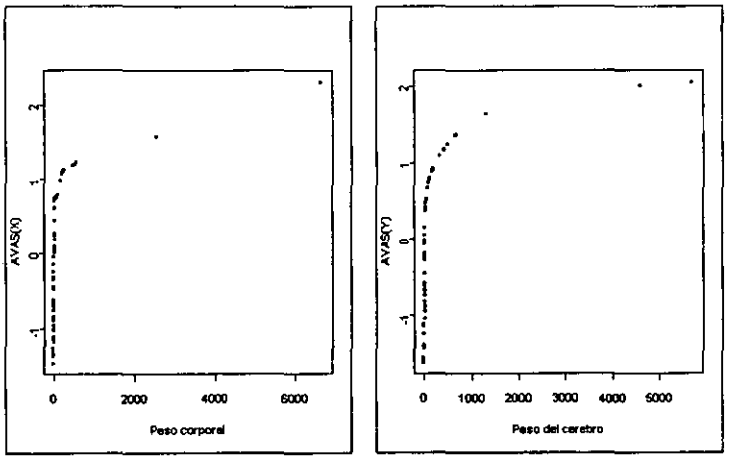

Fig. (3.17). a) $\phi^*(X)$; b) $\theta^*(Y)$.

La forma de las transformaciones en este ejemplo es bastante similar a la logarítmica, presentando por consiguiente una buena estabilidad en la varianza, ilustrada en la siguiente figura:

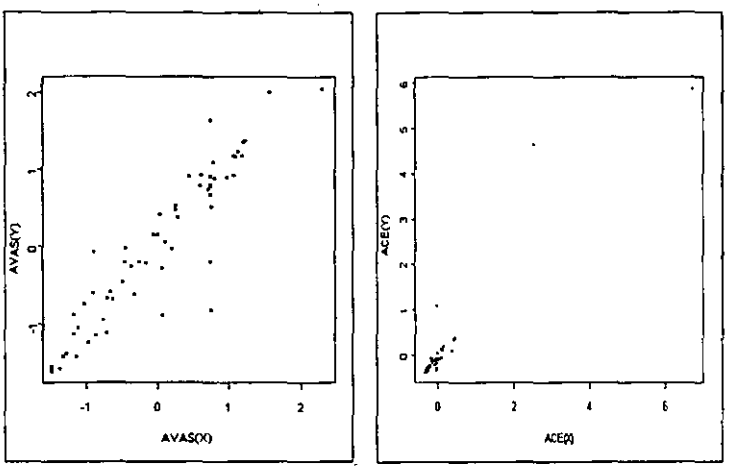

Fig. (3.18). $\phi^*(X)$ vs. $\theta^*(Y)$. a) AVAS; b) ACE.

Se aprecia entonces que este método presenta una menor sensibilidad a este tipo de datos que su predecesor.

Ejemplo 3.6.5 *Transformaciones no monótonas de las variables independientes.*

En los supuestos del AVAS se establece que la transformación de la variable independiente debe ser estrictamente monótona. Sin embargo, en este ejemplo se analizarán los resultados del AVAS para funciones no monótonas de las variables dependientes. Los datos siguen la relación:

$$Y = sen(X_1) + (X_2)^3 + Z,$$

donde $X_1 \sim U(0, 2\pi)$, $X_2 \sim U(-\pi, \pi)$, $Z \sim N(0, 1)$ y se utilizaron muestras indepen-dientes de tamaño 200. Los resultados se muestran a continuación.

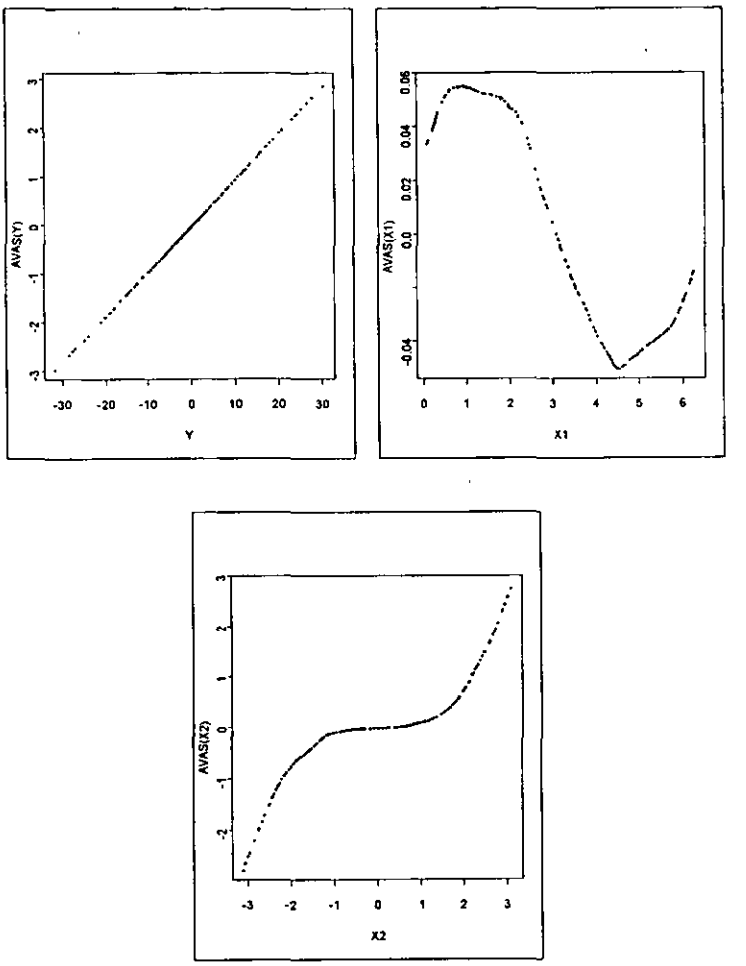

Fig. (3.19). a) $\theta^*(Y)$; b) $\phi_1^*(X_1)$; c) $\phi_2^*(X_2)$.

Se considera que las soluciones son por demás aceptables. Esto sugiere que el algoritmo no presenta problemas para reproducir modelos con transformaciones de variables independientes no monótonas, al igual que el ACE.

Se concluye así la descripción del algoritmo AVAS para pasar a mencionar algunas conclusiones, a la luz de lo expuesto hasta aquí.

Conclusiones

Esta sección final tiene como propósito resumir lo expuesto a lo largo del trabajo, así como discutir algunos aspectos generales de regresión y la situación que los algoritmos guardan con respecto a ellos.

Principales características de los algoritmos.

De manera general, es posible sintetizar las características del ACE y del AVAS citando los siguientes puntos:

- Ambos algoritmos tratan el problema de encontrar de manera no paramétrica, el conjunto de transformaciones óptimas, cada una según un criterio distinto, para explicar un modelo de la forma $\theta(Y) = \sum_{i=1}^{p} \phi_i(X_i)$, $i = 1, 2, ..., p$.
- El ACE considera que un conjunto de transformaciones (θ^*, ϕ^*) es óptimo si éste cumple con que $e^2(\theta^*, \phi^*) = \min\{e^2(\theta, \phi)\}$, donde $e^2(\theta, \phi_1, ..., \phi_p) = E[\theta - \sum_{i=1}^{p} \phi_i]$, y el mínimo es con respecto al conjunto de todas las transformaciones.
- La modificación incorporada por el AVAS considera el mínimo con respecto al conjunto de transformaciones que cumplan con que el modelo resultante $\theta(Y) = \sum_{i=1}^{p} \phi_i(X_i)$ tenga varianza constante.
- La base de la construcción del ACE son las propiedades de la esperanza condicional.
- El AVAS incorpora además la transformación asintótica de varianza.
- En el algoritmo ACE se cuenta con teoría matemática que demuestra que, en el caso continuo, bajo algunos supuestos existen transformaciones ACE-óptimas y que el algoritmo converge a ellas.
- El AVAS utiliza resultados que muestran que la transformación asintótica en general mejora la estabilidad de la varianza.

- La aplicación de los algoritmos se da en el caso discreto, donde se cuenta con una muestra de las variables $(X, \mathbf{Y})$ para determinar las transformaciones, y las sugerencias de los algoritmos se obtienen de manera gráfica, mediante comparaciones de los valores de las variables originales contra los valores transformados.

- En el caso discreto del ACE, bajo algunos supuestos, se demuestra la convergencia del algoritmo y unicidad del resultado.

Para finalizar se presentan los comentarios y observaciones, con el objeto de ir más allá de un simple resumen.

Propósitos de una regresión.

En general, se identifica que los modelos paramétricos de regresión tienen dos propósitos: proporcionar información acerca del comportamiento de un sistema y especificar supuestos bajo los cuales los parámetros del modelo pueden ser simple y efectivamente estimados.

Por lo regular, de presentarse un conflicto entre los puntos anteriores, puede decirse que es mejor prestar mayor atención al segundo punto, ya que una inferencia aproximada del comportamiento del sistema es preferible a una inferencia "exacta" cuya definición sea un tanto artificial o compleja. De aquí se deduce la ventaja que representa el que los supuestos utilizados en estos algoritmos son mínimos.

Supuestos de modelos.

Las técnicas usuales para el análisis de modelos lineales, como lo son el análisis de varianza y la regresión misma, generalmente son justificados asumiendo[5]:

1. Simplicidad en la estructura de Y.
2. Estabilidad en la varianza de los errores.
3. Normalidad de las distribuciones.
4. Independencia de las observaciones.

En aplicaciones de estas técnicas, como el ACE y el AVAS, un ejemplo muy importante del supuesto 1 es el supuesto de aditividad, i.e. ausencia de interacción de las variables explicativas, siendo las más elemenales las combinaciones lineales de las variables explicativas. Por su parte, si los supuestos 2 y 3 no se cumplen en términos de las observaciones originales, una transformación no lineal puede proporcionar una solución.

[5]Box, G. E. P. y Cox, D. R. (1964). "An Analysis of Transformations". Journal of the Royal Statistical Society, 26.

Cada una de las consideraciones 1 a 3 pueden ser y han sido utilizadas para seleccionar candidatos de familias de transformaciones paramétricas en distintos trabajos. Al respecto los algoritmos presentados pretenden, de manera no paramétrica, proporcionar un alto poder explicativo de forma tal que los supuestos 1 y 2 se satisfagan, mientras que el punto 4 debe ser asumido en la labor del muestreo previo al análisis, como en cualquier otro análisis estadístico de este tipo.

Limitaciones de Modelos Paramétricos y ventajas del ACE y AVAS.

A través de los años se han propuesto una serie de modelos que permitan una mayor libertad de uso o generalización a la técnica de regresión, a través del empleo de transformaciones que no impongan a los modelos estudiados una estructura particular. Sin embargo, una gran parte de estos procedimientos son paramétricos, lo cual aún limita los sistemas a ciertas familias de transformaciones, o requieren identificar gráficamente el tipo de transformaciones, contando con nada más que los datos originales, mediante análisis de residuales u otras técnicas. Lo anterior por supuesto no es una tarea sencilla, además de ser altamente susceptible al criterio del usuario.

De ello caben destacar dos aspectos de los algoritmos:

- Están diseñados para permitir relaciones no lineales tanto en las variables dependiente e independientes, e incluso estas transformaciones no tienen que corresponder a cierta familia parametrizable. Adicionalmente, las transformaciones de la variable dependiente pueden ser no monótonas, permitiendo incorporar al análisis variables continuas reales ordenadas, reales periódicas, discretas ordenadas o categóricas.

- En cuanto a la subjetividad, si bien es necesario cierto grado de ella (o tal vez más correctamente debería llamársele participación del analista)[6], sí existe una diferencia en la forma en que el usuario debe tomar decisiones. Mientras que el uso de algunos métodos paramétricos exigen al analista desde seleccionar el tipo de familia a utilizar para proceder a la estimación de parámetros y especificar así las transformaciones, al utilizar los algoritmos ACE o AVAS la participación del usuario para poder generar las sugerencias es mínima; sólo se requiere para interpretarlas. Esta diferencia entre selección de familias - interpretación de resultados representa una ventaja importante, debido a que puede minimizar las fallas por error humano o el tiempo de trabajo.

Análisis multivariado.

Como complemento a la obtención automática de sugerencias, otro punto destacable de los algoritmos es que proporcionan transformaciones para modelos multivariados, los

[6]El sentido común puede indicar que una técnica que pretenda cubrir una amplia gama de opciones debe hasta cierto punto utilizar el criterio, capacidad o conocimiento de un analista. De aquí el porqué es necesaria la participación de éste en los algoritmos que utilizan un mínimo grado de restricciones.

cuales suelen presentar complicaciones adicionales, debido a que entre más variables interactúen mayor dificultad existirá para poder dividir, mediante análisis tradicionales, el comportamiento general del sistema en componentes correspondientes a cada una de las variables.

Cuándo tiene sentido utilizar transformaciones de variables.

Como se ha mencionado anteriormente, mientras que en algunos problemas de regresión múltiple puede ser deseable el producir el modelo de regresión más simple posible, debe tenerse en cuenta que el conjunto de transformaciones deberían estar en una métrica en cuyos términos los resultados obtenidos para el problema en cuestión puedan ser útiles.

Con relación a lo anterior, es posible clasificar a las variables de un sistema en dos categorías: variables extensivas y no extensivas[7]. Las primeras se caracterizan por contar con una propiedad de aditividad física, mientras que las últimas carecen de ésta. De esta forma, la cantidad de productos fabricados en un periodo de tiempo determinado es extensivo. El tiempo de falla de un componente electrónico puede ser considerado extensivo si la característica principal a considerar es el número de componentes utilizados en un periodo. Propiedades como temperatura, viscosidad o la calidad del producto no son extensivas. Remitiéndonos inicialmente al caso de variables dependientes, suponiendo que el objeto del análisis sea la capacidad de predicción o explicativa[8], la opinión de analistas y estadistas coincide en que no existe razón para preferir la forma original de una variable no extensiva a una transformación, aunque no monótona, de ella. Tratándose de variables extensivas, sin embargo, es posible que la media de la población de Y sea un parámetro determinante en el comportamiento a largo plazo del sistema (como los modelos de reversión a la media). Así, en los ejemplos de variables extensivas anteriores, tanto el número de artículos producidos como la cantidad total de los componentes electrónicos utilizados en un gran periodo de tiempo están determinados por el promedio de artículos producidos y por el tiempo de falla promedio respectivamente, independientemente de la distribución.

De esta forma, aunque en un sentido práctico (que puede o puede no ser algo limitado) el interés radica en la media de la población Y, no de una transformación de Y. Por ello se debe

[7] Box, G. E. P. y Cox, D. R. (1964). "An Analysis of Transformations". Journal of the Royal Statistical Society, 26.

[8] Recordando, es primordial distinguir entre los análisis de datos cuyo interés principal es:

- la transformación particular, escencialmente de la variable dependiente, como en el caso de un análisis de datos para verificar si una muestra de una población capturó esta propiedad de la población en general, y
- el estudio de la estructura del modelo en general.

Por supuesto, el segundo caso es el más común. Sin embargo pueden haber más casos reales que los que a primera vista pueden imaginarse, como por ejemplo el estudio de un sistema con dos variables explicativas cuya interacción no está bien identificada; posiblemente el objeto del análisis es encontrar una transformación particular , si existe, para la cual no exista interacción entre las variables independientes.

identificar si el sistema en cuestión requiere un análisis lineal de datos de Y sin transformar o, de habérsele aplicado una transformación para obtener un análisis más eficiente y válido, transformar las conclusiones de vuelta a su escala original.

Para efecto de los algoritmos, el contar con los datos en su métrica original es equivalente a obtener los valores $\theta^{*-1}\left(\sum_{i=1}^{p}\phi_i^*(x_i)\right)$. Para ello, en el caso del ACE, es posible proceder de la siguiente forma: una vez que se han obtenido las transformaciones $\phi_i^*(\cdot)$, $i=1,2,...,p$, se puede definir a $Z=\sum_{i=1}^{p}\phi_i^*(x_i)$, y como se sabe que la mejor predicción de Y en términos de transformaciones $\chi(Z)$ está dada por $E(Y \mid Z)$, todo lo que se requiere es suavizar los valores de y en los valores de $\sum_{i=1}^{p}\phi_i^*(x_i)$. En el caso del AVAS no es posible utilizar el mismo procedimiento debido la estabilidad de varianza que se aplica en cada ciclo del algoritmo.

Utilidad de los Algoritmos para revisión de Modelos.

Un uso adicional de estos algoritmos es que representan una alternativa para calificar si un modelo previamente ajustado de transformaciones es óptimo (o no hay grandes indicaciones por parte de los algoritmos que es posible mejorar el modelo sensiblemente), mediante el uso de los valores de las variables pretransformadas y aplicando el algoritmo correspondiente, según el objetivo buscado[9], para calificar al sistema en cuestión. De esta forma, el obtener un bajo grado de linealidad en las transformaciones al aplicar el algoritmo sugerirfa que es posible modificar las transformaciones para alcanzar un mayor poder de ajuste.

Experiencia vs. Técnica: ¿competencia o complemento?

En estudios de regresión donde exista una fuerte evidencia empírica o un razonamiento previo (como podrían ser el caso de sistemas relacionados con fuerzas físicas o reacciones químicas) que sugiera cierto procedimiento o familia explícita, una manera recomendable de proceder es el llevar a cabo el procedimiento o transformación en cuestión y posteriormente considerar si transformaciones adicionales son necesarias.

Por otra parte, en cuanto al uso de criterios subjetivos o heurísticos como el mencionado en el Ejemplo (3.1.4), se debe tomar en cuenta que la hábilidad de un analista es adicionalmente el resultado de un razonamiento heurístico, y no de un entrenamiento puramente teórico. Es cierto que en ocasiones este primer tipo de reglas no pueden ser justificadas de manera formal, pero sí con la experiencia.

Como medida puede utilizarse conjuntamente el uso del razonamiento heurístico junto con herramientas soportadas con bases científicas, las cuales más que poder reemplazar una a la otra son complementarias.

[9]Para correlación únicamente el ACE puede ser de utilidad, mientras que si se desea considerar la estabilidad de varianza también, se puede emplear el AVAS.

Requerimiento de tiempo.

Finalmente, el tiempo requerido para la obtención de los resultados, una vez que se han identificado las tentativas variables explicativas y muestreado los datos, requiere de unos cuantos segundos gracias a los avances tecnológicos con que se cuentan actualmente. Posiblemente esta última característica hoy en día es considerada como un requisito más que como una ventaja, en cuyo caso el propósito de mencionarla es indicar que los algoritmos cumplen con él.

Anexo 1. Observaciones sobre Matrices

En esta sección se introducirán algunas definiciones y se harán algunas observaciones relacionadas con matrices, las cuales son empleadas al presentar el modelo de regresión lineal múltiple en el primer capítulo de este trabajo.

Sea H la representación matricial de una función lineal $h : \Re^N \rightarrow \Re^N$ y $\mathbf{z}$ un vector n-dimensional.

Definición A1.1 *La parcial de H con respecto a $\mathbf{z}$ está dada por:*

$$\frac{\partial H(\mathbf{z})}{\partial \mathbf{z}} = \begin{pmatrix} \frac{\partial H(\mathbf{z})}{\partial z_1} \\ \vdots \\ \frac{\partial H(\mathbf{z})}{\partial z_N} \end{pmatrix}.$$

Observación A1.1 *La segunda parcial de H con respecto a $\mathbf{z}$ está dada por:*

$$\frac{\partial^2 H(\mathbf{z})}{\partial \mathbf{z}^t \partial \mathbf{z}} = \frac{\partial}{\partial \mathbf{z}^t}\left\{\frac{\partial H(\mathbf{z})}{\partial \mathbf{z}}\right\} = \left(\frac{\partial^2 H(\mathbf{z})}{\partial z_i \partial z_j}\right).$$

Observación A1.2 *Si $\mathbf{a}$ es un vector N-dimensional, entonces:*

$$\frac{\partial \mathbf{a}^t \mathbf{z}}{\partial \mathbf{z}} = \left(\frac{\partial \sum_{j=1}^N a_j z_j}{\partial z_i}\right) = (a_i) = \mathbf{a}.$$

Observación A1.3 *Si A es una matriz simétrica de dimensión $N \times N$, entonces:*

$$\begin{aligned} \frac{\partial \mathbf{z}^t A \mathbf{z}}{\partial \mathbf{z}} &= \left(\frac{\partial \sum_{i=1}^N \sum_{j=1}^N a_{ij} z_i z_j}{\partial z_k}\right) = \left(2a_{kk}z_k + \sum_{j \neq k} a_{kj} z_j + \sum_{i \neq k} a_{ik} z_i\right) \\ &= 2\left(\sum_{j=1}^N a_{kj} z_j\right) = 2A\mathbf{z}. \end{aligned}$$

Observación A1.4 *Si A es una matriz simétrica de dimensión $N \times N$, entonces:*

$$\frac{\partial \mathbf{z}^t A \mathbf{z}}{\partial \mathbf{z}^t \partial \mathbf{z}} = \frac{\partial}{\partial \mathbf{z}^t}\left\{\frac{\partial \mathbf{z}^t A \mathbf{z}}{\partial \mathbf{z}}\right\} = \frac{\partial}{\partial \mathbf{z}^t}\{2A\mathbf{z}\} = 2A.$$

Anexo 2. Tipos de Covergencia

A continuación se describen los tipos de convergencia utilizados a lo largo de la obra, escencialmente al presentar la esperanza condicional. Únicamente se omite la definición en media cuadrática la cual se presentó en esa sección, debido a su importancia en la misma.

Sea $(\Omega, \langle , \rangle , \Im, \mu)$ un espacio producto interior y además un espacio de medida, con $\|, \|$ la norma inducida por el producto interior, y $x \in \Omega$. Finalmente sean $f_n, f : \Omega \to \Re$, $n \in \aleph$.

Definición A2.1 *Se dice que f_n converge puntualmente (o simplemente converge) a f si $\forall\ \varepsilon > 0$ y $\forall\ x \in \Omega \quad \exists\ n_{x,\varepsilon} > 0$ tal que*

$$N > n_{x,\varepsilon} \Rightarrow |f_N(x) - f(x)| < \varepsilon.$$

Definición A2.2 *Se dice que f_n converge en norma a f si $\forall\ \varepsilon > 0\ \exists\ n_\varepsilon > 0$ tal que*

$$N > n_\varepsilon \Rightarrow \|f_N(\cdot) - f(\cdot)\| < \varepsilon.$$

Definición A2.3 *Se dice que f_n converge casi seguramente a f si $\mu(A^c) = 0$, donde*

$$A = \{x \in \Omega : f_n(x) \text{ converge a } f(x)\}.$$

Bibliografía

- Bartle, Robert G. (1966). The Elements of Integration. John Wiley & Sons.
- Bibby, J. M., Kent, J. T. y Mardia, K. V. (1982). Multivariate Analysis. Academic Press, Inc.
- Box, G. E. P. y Cox, D. R. (1964). "An Analysis of Transformations". Journal of the Royal Statistical Society, 26, 211-252.
- Box, G. E. P. y Hunter, W. G. (1962). "A Useful Method for Model Building". Technometrics, 4, 301-318.
- Box, G. E. P. y Tidwell, P. W. (1962)."Transformations of the Independent Variable". Technometrics, 4, 531-550.
- Breiman, L. y Freidman, J. H. (1985). "Estimating Optimal Transformations for Multiple Regression and Correlation". Journal of the American Statistical Association, 80, 580-597.
- Brockwell, P. J. y Davis, R. A. (1990). Time Series: Theory and Models. Springer-Verlag.
- Draper, N. R. y Smith, H. (1998). Applied Regression Analysis. John Wiley & Sons.
- Efron, B. (1982). "Transformation Theory: How Normal is a Family of Distributions?". The Annals of Statistics, 10, 323-339.
- Hinkley, D. y Runger, G. (1984). "The Analysis of Transformed Data". Journal of the American Statistical Association, 79, 309-328.
- Pregibon, D. y Vardi, Y. (1985). Comentario de "Estimating Optimal Transformations for Multiple Regression and Correlation". Journal of the American Statistical Association, 80, 598-601.
- Tibshirani, R. (1988). "Estimating Transformations for Regression Via Additivity and Variance Stabilization". Journal of the American Statistical Association, 83, 394-405.

- Wonnacott, R. J. y Wonnacott, T. H. (1981). Regression: A Second Course in Statistics. R. E. Krieger.