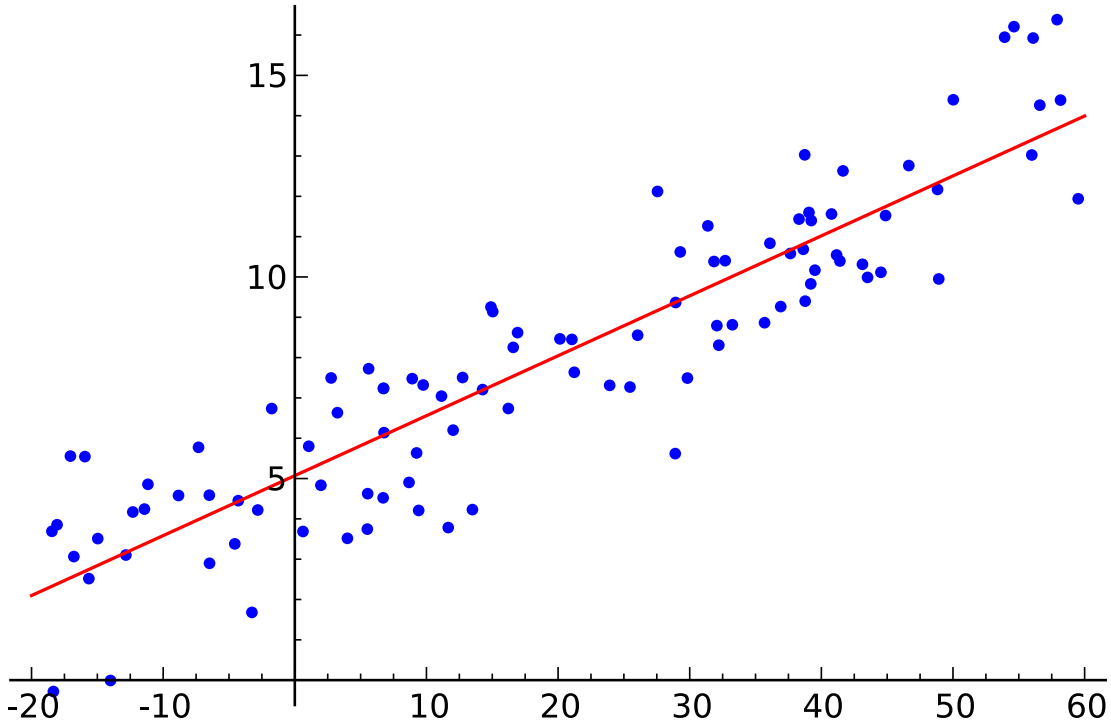
Introduccion a Modelos de Regresion

Índice general

Capítulo 1

Regresión lineal



Ejemplo de una regresión lineal con una variable dependiente y una variable independiente.

En estadística la **regresión lineal** o **ajuste lineal** es un método matemático que modela la relación entre una variable dependiente *Y*, las variables independientes *Xi* y un término aleatorio ε. Este modelo puede ser expresado como:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Y_t : variable dependiente, explicada o regresando.

$X_1, X_2, \cdots, X_p$: variables explicativas, independientes o regresores.

$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$: parámetros, miden la influencia que las variables explicativas tienen sobre el regresando.

donde β_0 es la intersección o término "constante", las β_i $(i > 0)$ son los parámetros respectivos a cada variable independiente, y p es el número de parámetros independientes a tener en cuenta en la regresión. La regresión lineal puede ser contrastada con la regresión no lineal.

1.1 Historia

La primera forma de regresión lineal documentada fue el método de los mínimos cuadrados que fue publicada por Legendre en 1805,[1] y en dónde se incluía una versión del teorema de Gauss-Márkov.

1.1.1 Etimología

El término *regresión* se utilizó por primera vez en el estudio de variables antropométricas: al comparar la estatura de padres e hijos, donde resultó que los hijos cuyos padres tenían una estatura muy superior al valor medio, tendían a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media; es decir, "regresaban" al promedio.[2] La constatación empírica de esta propiedad se vio reforzada más tarde con la justificación teórica de ese fenómeno.

El término *lineal* se emplea para distinguirlo del resto de técnicas de regresión, que emplean modelos basados en cualquier clase de función matemática. Los modelos lineales son una explicación simplificada de la realidad, mucho más ágiles y con un soporte teórico mucho más extenso por parte de la matemática y la estadística.

Pero bien, como se ha dicho, podemos usar el término lineal para distinguir modelos basados en cualquier clase de aplicación.

1.2 El modelo de regresión lineal

El modelo lineal relaciona la variable dependiente *Y* con *K* variables explícitas X_k *(k = 1,...K)*, o cualquier transformación de éstas que generen un hiperplano de parámetros β_k desconocidos:

$$(2)\ Y = \sum \beta_k X_k + \varepsilon$$

donde ε es la perturbación aleatoria que recoge todos aquellos factores de la realidad no controlables u observables y que por tanto se asocian con el azar, y es la que confiere al modelo su carácter estocástico. En el caso más sencillo, con una sola variable explícita, el hiperplano es una recta:

$$(3)\ Y = \beta_1 + \beta_2 X_2 + \varepsilon$$

El problema de la regresión consiste en elegir unos valores determinados para los parámetros desconocidos β_k , de modo que la ecuación quede completamente especificada. Para ello se necesita un conjunto de observaciones. En una observación i-ésima *(i= 1,... I)* cualquiera, se registra el comportamiento simultáneo de la variable dependiente y las variables explícitas (las perturbaciones aleatorias se suponen no observables).

$$(4)\ Y_i = \sum \beta_k X_{ki} + \varepsilon_i$$

Los valores escogidos como estimadores de los parámetros $\hat{\beta}_k$, son los coeficientes de regresión sin que se pueda garantizar que coincida n con parámetros reales del proceso generador. Por tanto, en

$$(5)\ Y_i = \sum \hat{\beta}_k X_{ki} + \hat{\varepsilon_i}$$

Los valores $\hat{\varepsilon_i}$ son por su parte estimaciones o errores de la perturbación aleatoria.

1.3 Hipótesis del modelo de regresión lineal clásico

1. Esperanza matemática nula.

$E(\varepsilon_i) = 0$

Para cada valor de X la perturbación tomará distintos valores de forma aleatoria, pero no tomará sistemáticamente valores positivos o negativos, sino que se supone tomará algunos valores mayores que cero y otros menores que cero, de tal forma que su valor esperado sea cero.

2. Homocedasticidad

$Var(\varepsilon_t) = E(\varepsilon_t - E\varepsilon_t)^2 = E\varepsilon_t^2 = \sigma^2$ para todo t

Todos los términos de la perturbación tienen la misma varianza que es desconocida. La dispersión de cada ε_t en torno a su valor esperado es siempre la misma.

3. Incorrelación.

$Cov(\varepsilon_t, \varepsilon_s) = (\varepsilon_t - E\varepsilon_t)(\varepsilon_s - E\varepsilon_s) = E\varepsilon_t\varepsilon_s = 0$ para todo t,s con t distinto de s

Las covarianzas entre las distintas pertubaciones son nulas, lo que quiere decir que no están correlacionadas. Esto implica que el valor de la perturbación para cualquier observación muestral no viene influenciado por los valores de las perturbaciones correspondientes a otras observaciones muestrales.

4. Regresores no estocásticos.

5. No existen relaciones lineales exactas entre los regresores.

6. $T > k + 1$ Suponemos que no existen errores de especificación en el modelo, ni errores de medida en las variables explicativas

7. Normalidad de las perturbaciones $\varepsilon \sim N(0, \sigma^2)$

1.4 Supuestos del modelo de regresión lineal

Para poder crear un modelo de regresión lineal es necesario que se cumpla con los siguientes supuestos:[3]

1. Que la relación entre las variables sea lineal.
2. Que los errores en la medición de las variables explicativas sean independientes entre sí.
3. Que los errores tengan varianza constante. (Homocedasticidad)
4. Que los errores tengan una esperanza matemática igual a cero (los errores de una misma magnitud y distinto signo son equiprobables).
5. Que el error total sea la suma de todos los errores.

1.5 Tipos de modelos de regresión lineal

Existen diferentes tipos de regresión lineal que se clasifican de acuerdo a sus parámetros:

1.5.1 Regresión lineal simple

Sólo se maneja una variable independiente, por lo que sólo cuenta con dos parámetros. Son de la forma:[4]

$$(6)\ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

donde ε_i es el error asociado a la medición del valor X_i y siguen los supuestos de modo que $\varepsilon_i \sim N(0, \sigma^2)$ (media cero, varianza constante e igual a un σ y $\varepsilon_i \perp \varepsilon_j$ con $i \neq j$).

Análisis

Dado el modelo de regresión simple, si se calcula la esperanza (valor esperado) del valor *Y*, se obtiene:[5]

$$(7)\ E(y_i) = \hat{y}_i = E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i)$$

Derivando respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$ e igualando a cero, se obtiene:[5]

$$(9)\ \frac{\partial \sum (y_i - \hat{y}_i)^2}{\partial \hat{\beta}_0} = 0$$

$$(10)\ \frac{\partial \sum (y_i - \hat{y}_i)^2}{\partial \hat{\beta}_1} = 0$$

Obteniendo dos ecuaciones denominadas ecuaciones normales que generan la siguiente solución para ambos parámetros:[4]

$$(11)\ \hat{\beta}_1 = \frac{\sum x \sum y - n \sum xy}{(\sum x)^2 - n \sum x^2} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$(12)\ \hat{\beta}_0 = \frac{\sum y - \hat{\beta}_1 \sum x}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

La interpretación del parámetro medio β_1 es que un incremento en Xi de una unidad, Yi incrementará en β_1

1.5.2 Regresión lineal múltiple

La regresión lineal permite trabajar con una variable a nivel de intervalo o razón. De la misma manera, es posible analizar la relación entre dos o más variables a través de ecuaciones, lo que se denomina **regresión múltiple** o **regresión lineal múltiple**.

Constantemente en la práctica de la investigación estadística, se encuentran variables que de alguna manera están relacionadas entre sí, por lo que es posible que una de las variables puedan relacionarse matemáticamente en función de otra u otras variables.

Maneja varias variables independientes. Cuenta con varios parámetros. Se expresan de la forma:[6]

$$(13)\ Y_i = \beta_0 + \sum \beta_i X_{ip} + \varepsilon_i$$

donde ε_i es el error asociado a la medición i del valor X_{ip} y siguen los supuestos de modo que $\varepsilon_i \sim N(0, \sigma^2)$ (media cero, varianza constante e igual a un σ y $\varepsilon_i \perp \varepsilon_j$ con $i \neq j$).

1.6 Rectas de regresión

Las rectas de regresión son las rectas que mejor se ajustan a la nube de puntos (o también llamado diagrama de dispersión) generada por una distribución binomial. Matemáticamente, son posibles dos rectas de máximo ajuste:[7]

- La recta de regresión de *Y* sobre *X*:

$$(14)\ y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

- La recta de regresión de *X* sobre *Y*:

$$(15)\ x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

La correlación (“*r*”) de las rectas determinará la calidad del ajuste. Si *r* es cercano o igual a 1, el ajuste será bueno y las predicciones realizadas a partir del modelo obtenido serán muy fiables (el modelo obtenido resulta verdaderamente representativo); si *r* es cercano o igual a 0, se tratará de un ajuste malo en el que las predicciones que se realicen a partir del modelo obtenido no serán fiables (el modelo obtenido no resulta representativo de la realidad). Ambas rectas de regresión se intersecan en un punto llamado centro de gravedad de la distribución.

1.7 Aplicaciones de la regresión lineal

1.7.1 Líneas de tendencia

Una *línea de tendencia* representa una tendencia en una serie de datos obtenidos a través de un largo período. Este tipo de líneas puede decirnos si un conjunto de datos en particular (como por ejemplo, el PBI, el precio del petróleo o el valor de las acciones) han aumentado o decrementado en un determinado período.[8] Se puede dibujar una línea de tendencia a simple vista fácilmente a partir de un grupo de puntos, pero su posición y pendiente se calcula de manera más precisa utilizando técnicas estadísticas como las regresiones lineales. Las líneas de tendencia son generalmente líneas rectas, aunque algunas variaciones utilizan polinomios de mayor grado dependiendo de la curvatura deseada en la línea.

1.7.2 Medicina

En medicina, las primeras evidencias relacionando la mortalidad con el fumar tabaco[9] vinieron de estudios que utilizaban la regresión lineal. Los investigadores incluyen una gran cantidad de variables en su análisis de regresión en un esfuerzo por eliminar factores que pudieran producir correlaciones espurias. En el caso del tabaquismo, los investigadores incluyeron el estado socio-económico para asegurarse que los efectos de mortalidad por tabaquismo no sean un efecto de su educación o posición económica. No obstante, es imposible incluir todas las variables posibles en un estudio de regresión.[10][11] En el ejemplo del tabaquismo, un hipotético gen podría aumentar la mortalidad y aumentar la propensión a adquirir enfermedades relacionadas con el consumo de tabaco. Por esta razón, en la actualidad las pruebas controladas aleatorias son consideradas mucho más confiables que los análisis de regresión.

1.7.3 Informática

Ejemplo de una rutina que utiliza una recta de regresión lineal para proyectar un valor futuro: Código escrito en PHP

<?php //Licencia: GNU/GPL $xarray=array(1, 2, 3, 4, 5); //Dias $yarray=array(5, 5, 5, 6.8, 9); //Porcentaje de ejecucion $pm=100; //Valor futuro $x2=0; $y=0; $x=0; $xy=0; $cantidad=count($xarray); for($i=0;$i<$cantidad;$i++){ //Tabla de datos print ($xarray[$i].” ---- ".$yarray[$i]."
"); //Calculo de terminos $x2 += $xarray[$i]*$xarray[$i]; $y += $yarray[$i]; $x += $xarray[$i]; $xy += $xarray[$i]*$yarray[$i]; } //Coeficiente parcial de regresion $b=($cantidad*$xy-$x*$y)/($cantidad*$x2-$x*$x); //Calculo del intercepto $a=($y-$b*$x)/$cantidad; //Recta tendencial //y=a+bx //Proyeccion en dias para un 100% de la ejecucion: if ($b!=0) $dias_proyectados=($pm-$a)/$b; else $dias_proyectados=999999; //Infinitos $dp=round($dias_proyectados,0); if($dp<=$pm) print $dp."---> Culmina antes de los $pm dias
"; if($dp >$pm) print $dp ."---> ALARMA: No culmina antes de los $pm dias
"; ?>

1.8 Véase también

- Homoscedasticidad
- Regresión logística
- Modelos de regresión múltiple postulados y no postulados
- Regresión segmentada
- Econometría
- Mínimos cuadrados
- Regularización de Tikhonov
- Cuarteto de Anscombe
- Capital Asset Pricing Model

1.9 Referencias

[1] C.F. Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae*. (1821/1823)

[2] Introduction to linear regression Curvefit.com (en inglés)

[3] "Análisis de regresión lineal", Universidad Complutense de Madrid

[4] *"Fórmulas", Probabilidad y Estadística.* Cs. Básicas. U.D.B. Matemática. Universidad Tecnológica Nacional, Facultad Regional Buenos Aires. Editorial CEIT-FRBA. (Código BM2BT2)

[5] Modelo de regresión lineal simple. EinsteinNet.

[6] Técnicas de regresión: Regresión Lineal Múltiple. *Pértega Díaz, S., Pita Fernández, S.* Unidad de Epidemiología Clínica y Bioestadística. Complejo Hospitalario de La Coruña (España)

[7] Apunte sobre Rectas de regresión. Ministerio de Educación y Ciencia. Gobierno de España.

[8] Utilización de las líneas de tendencia, Paritech (en inglés)

[9] Doll R, Peto r, Wheatley K, Gray R et al. *Mortality in relation to smoking: 40 years' observations on male British doctors* .BMJ 1994;309:901-911 (8 de octubre]

[10] "Environmental Tobacco Smoke and Adult Asthma" Division of Pulmonary and Critical Care Medicine, Division of Occupational and Environmental Medicine; Department of Medicine, Institute for Health Policy Studies; and Department of Epidemiology and Biostatistics, Universidad de California, San Francisco, California. (en inglés)

[11] Efecto del tabaquismo, los síntomas respiratorios y el asma sobre la espirometría de adultos de la Ciudad de México, Justino Regalado-Pineda; Alejandro Gómez-Gómez; Javier Ramírez-Acosta; Juan Carlos Vázquez-García

1.10 Bibliografía

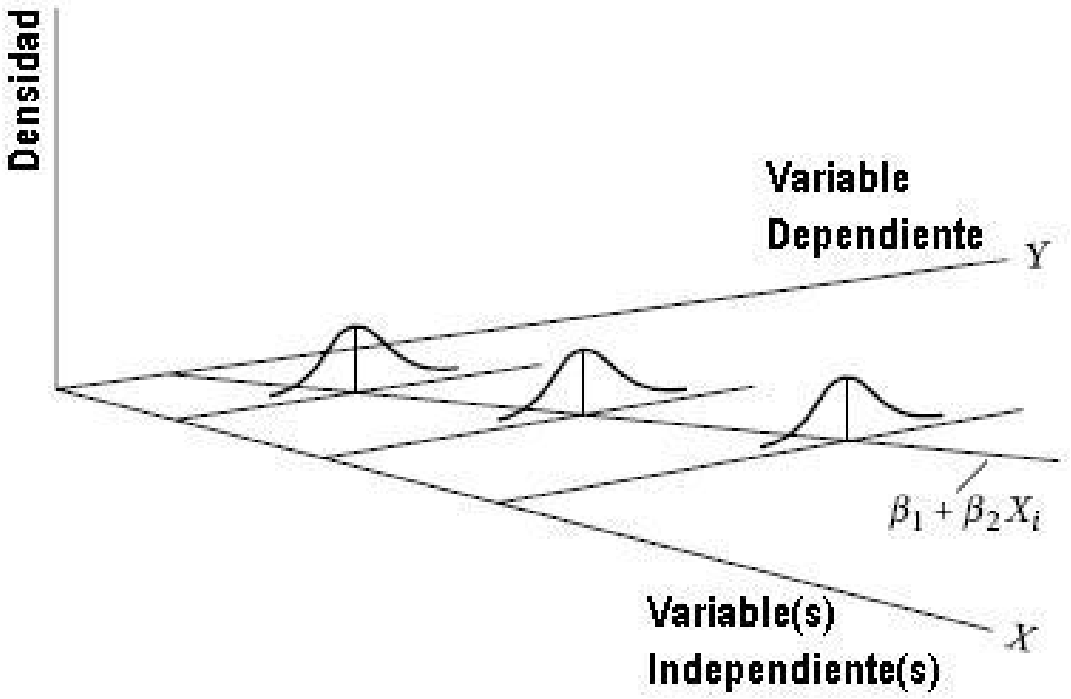
- Devore, Jay L.; *Probabilidad y Estadística para Ingeniería y Ciencias*. International Thomson Editores. México. ISBN-10: 9706864571.
- Walpole, Ronald E.; Raymond H.; Myers, Sharon L.; *Probabilidad y Estadística para Ingenieros*. Pretice-Hall Hispanoamericana, S.A. México. ISBN-10: 9701702646.
- Canavos, George C.; *Probabilidad y Estadística. Aplicaciones y Métodos.* McGraw-Hill. México. ISBN-10: 9684518560.

1.11 Enlaces externos

- Cálculo de regresiones lineales en línea. (en inglés)
- ZunZun.com Ajuste de curvas y superficies en línea. (en inglés)
- xuru.org Herramientas de regresión lineal en línea. (en inglés)
- Simulación de la recta de regresion de una variable bidimensional continua con R (lenguaje de programación)

Capítulo 2

Homocedasticidad



Distribución Homocedástica.

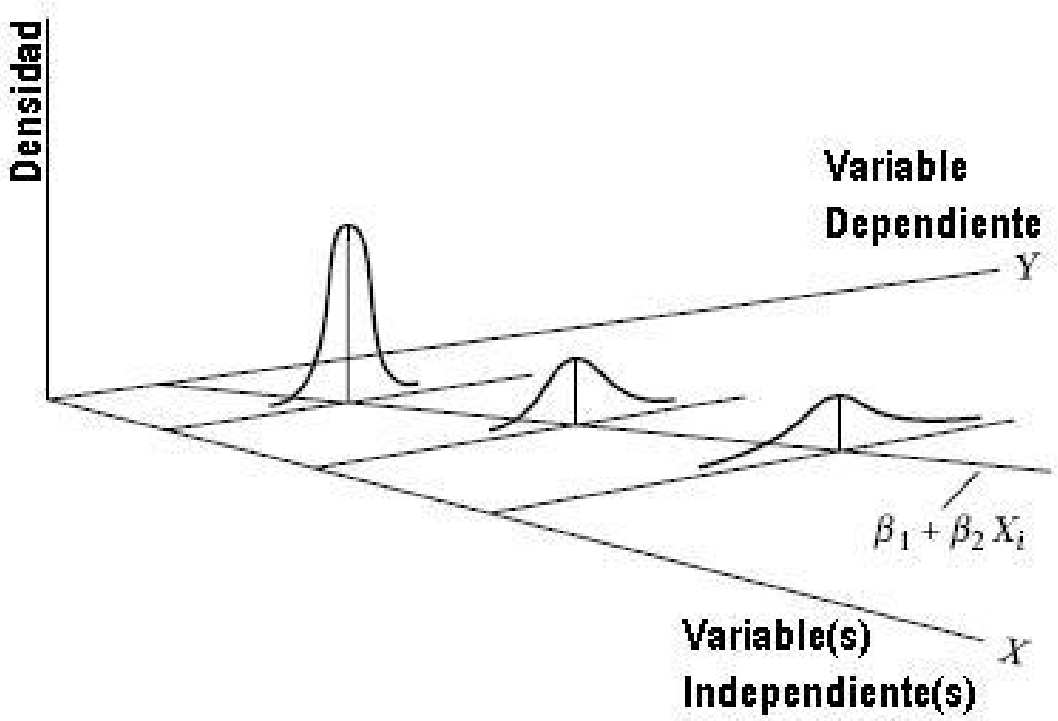
En estadísticas se dice que un modelo predictivo presenta **homocedasticidad** cuando la varianza del error de la variable endógena se mantiene a lo largo de las observaciones. En otras palabras, la varianza de los errores es constante.

Un modelo estadístico relaciona el valor de una variable a predecir con el de otras. Si el modelo es insesgado, el valor predicho es la media de la variable a predecir. En cualquier caso, el modelo da una idea del valor que tomará la variable a predecir.

Por simplificar el análisis, si se supone que la variable a predecir es escalar, aquí definida como η , y que se explica mediante un conjunto de variables que

$$\varepsilon = \eta - m(\boldsymbol{\xi})$$

Este error es una variable aleatoria: tomará un valor distinto cada vez que se ejecute el modelo. Se habla de **homocedasticidad** si el error cometido por el modelo tiene siempre la misma varianza. En particular, si el modelo es homocedástico, el valor de las variables explicativas, $\boldsymbol{\xi}$, no afectará a la varianza del error.



Distribución Heterocedástica.

La **homocedasticidad** es una propiedad fundamental del modelo de regresión lineal general y está dentro de sus supuestos clásicos básicos.

Formalizando, se dice que existe homocedasticidad cuando la varianza de los errores estocásticos de la regresión es la misma para cada observación *i* (de 1 a *n* observaciones), es decir:

$$E(\mu_i^2) = \sigma_\mu^2 \qquad \forall\, i = R$$

donde σ_μ^2 es un escalar constante para todo *i*. Lo que significaría que habría una distribución de probabilidad de idéntica amplitud para cada variable aleatoria.

Esta cualidad es necesaria, según el Teorema de Gauss-Márkov, para que en un modelo los coeficientes estimados sean los mejores o eficientes, lineales e insesgados.

Cuando no se cumple esta situación, se dice que existe heterocedasticidad, que es cuando la varianza de cada término de perturbación (u_i) no es un número constante σ^2 .

Este fenómeno suele ser muy común en datos de Corte Transversal y también se presenta, menos frecuentemente, en series de tiempo.

Si se regresiona un modelo a través de Mínimos Cuadrados Ordinarios con presencia de heterocedasticidad, los coeficientes siguen siendo lineales e insesgados pero ya no poseen mínima varianza (eficiencia).

2.1 Causas frecuentes de ausencia de homocedasticidad

2.1.1 Variables independientes que posean un gran recorrido con respecto a su propia media

Esto generalmente ocurre cuando se ha dispuesto arbitrariamente el orden de las observaciones, generando, casualmente que existan observaciones con grandes valores en una determinada variable explicativa y lo mismo con valores pequeños de esta misma variable.

2.1.2 Omisión de variables importantes dentro del modelo a estimar

Obviamente, si se omite una variable de relevancia en la especificación, tal variable quedará parcialmente recogida dentro de las perturbaciones aleatorias, introduciendo en estas su propia variación, que no será necesariamente fija.

2.1.3 Cambio de estructura

El hecho de que se produzca un cambio en la estructura determina un mal ajuste de los parámetros al conjunto de los datos muestrales. Y este no tiene porque influir del mismo modo en todo el recorrido de la muestra, pudiendo producir cuantías de desajuste del modelo diferentes y, por lo tanto, varianza no constante

2.1.4 Utilizar variables no relativizadas

Cuando existen observaciones dentro de una variable en concreto, y que poseen un valor mayor a las otras variables explicativas, puede originar valores del error diferentes. Esta situación es similar a la explicada al principio pero con la salvedad que en este caso se compara con las otras variables (inclusive con la dependiente) y no con respecto a su media.

2.2 Estimar en presencia de heterocedasticidad

2.2.1 Cálculo incorrecto de las varianza y parámetros ineficientes

La mayor varianza por empleo de MCO en presencia de heterocedasticidad puede producir un incremento de más de 10 veces en la varianza estimada del parámetro constante.

2.2.2 Invalidación de los contrastes de significancia

Ya que se aceptaría la hipótesis nula de los contrastes de significancia más veces de las reales. Generalmente resulta que ciertas variables podrían resultar no ser significativas cuando lo son realmente.

2.3 Referencias

- Damodar N. Gujarati. “Econometría”;
- Jorge Dresder Cid. “Nociones de Econometría Intermedia”
- Novales, A. “Econometría”

Capítulo 3

Regresión logística

En estadística, la **regresión logística** es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística.

La regresión logística es usada extensamente en las ciencias médicas y sociales. Otros nombres para regresión logística usados en varias áreas de aplicación incluyen **modelo logístico**, **modelo logit**, y **clasificador de máxima entropía**.

3.1 Introducción

La regresión logística analiza datos distribuidos binomialmente de la forma

$$Y_i \ \sim B(p_i, n_i), \text{ para } i = 1, \ldots, m,$$

donde los números de ensayos Bernoulli *ni* son conocidos y las probabilidades de éxito *pi* son desconocidas. Un ejemplo de esta distribución es el porcentaje de semillas (*pi*) que germinan después de que *ni* son plantadas.

El modelo es entonces obtenido a base de lo que cada ensayo (valor de *i*) y el conjunto de variables explicativas/independientes puedan informar acerca de la probabilidad final. Estas variables explicativas pueden pensarse como un vector *Xi* *k*-dimensional y el modelo toma entonces la forma

$$p_i = \mathrm{E}\left(\frac{Y_i}{n_i}\middle| X_i\right).$$

Los logits de las probabilidades binomiales desconocidas (*i.e.*, los logaritmos de la razón de momios) son modeladas como una función lineal de los *Xi*.

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

Note que un elemento particular de *Xi* puede ser ajustado a 1 para todo *i* obteniéndose una constante independiente en el modelo. Los parámetros desconocidos β_j son usualmente estimados a través de máxima verosimilitud.

La interpretación de los estimados del parámetro βj es como los efectos aditivos en el logaritmo de la razón de momios para una unidad de cambio en la *j*ésima variable explicativa. En el caso de una variable explicativa dicotómica, por ejemplo género, e^β es la estimación del odds ratio de tener el resultado para, por decir algo, hombres comparados con mujeres.

El modelo tiene una formulación equivalente dada por

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i})}}$$

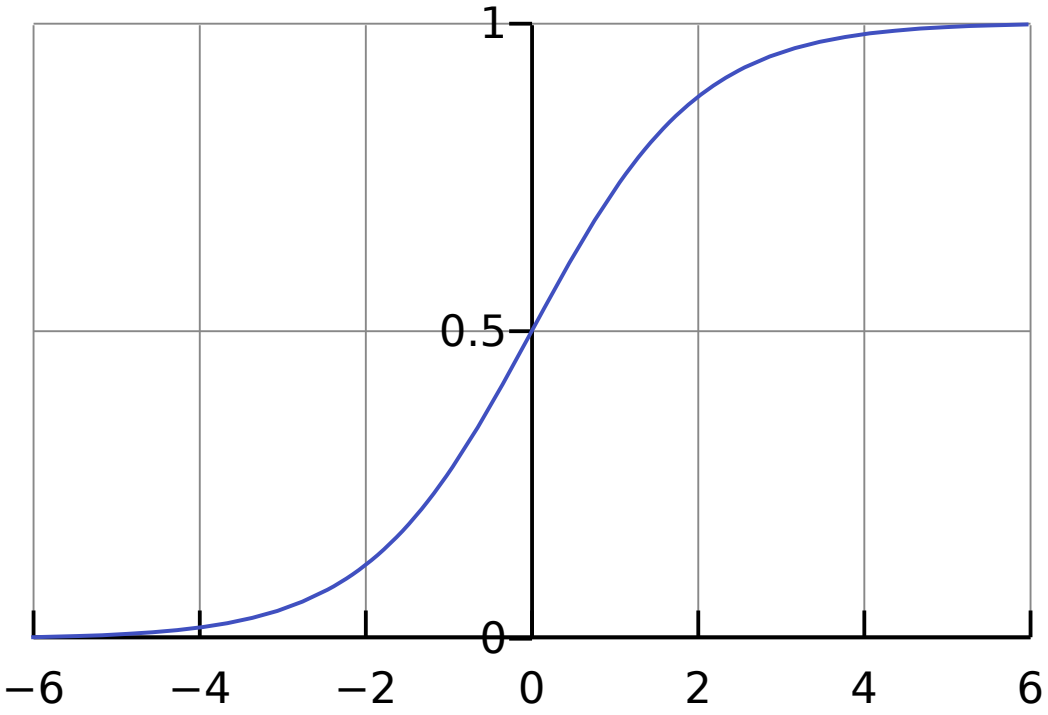
Esta forma funcional es comúnmente identificada como un "perceptrón" de una capa simple or red neuronal artificial de una sola capa. Una red neuronal de una sola capa calcula una salida continua en lugar de una función definida a trozos. La derivada de *pi* con respecto a $X = x_1...xk$ es calculada de la forma general:

$$y = \frac{1}{1+e^{-f(X)}}$$

donde $f(X)$ es una función analítica en X. Con esta escogencia, la red de capa simple es idéntica al modelo de regresión logística. Esta función tiene una derivada continua, la cual permite ser usada en propagación hacia atrás. Esta función también es preferida pues su derivada es fácilmente calculable:

$$y' = y(1-y)\frac{\mathrm{d}f}{\mathrm{d}X}$$

3.2 Implementación práctica



Función logística con $\beta_0 + \beta_1 x + e$ en el eje horizontal y $\pi(x)$ en el eje vertical.

La regresión logística unidimensional puede usarse para tratar de correlacionar la probabilidad de una variable cualitativa binaria (asumiremos que puede tomar los valores reales "0" y "1") con una variable escalar x. La idea es que la regresión logística aproxime la probabilidad de obtener "0" (no ocurre cierto suceso) o "1" (ocurre el suceso) con el valor de la variable explicativa x. En esas condiciones, la probabilidad aproximada del suceso se aproximará mediante una función logística del tipo:[1]

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta_1 x)}+1} = \frac{1}{e^{-(\beta_0+\beta_1 x)}+1},$$

que puede reducirse al cálculo de una regresión lineal para la función logit de la probabilidad:

$$g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x,$$

o una regresión exponencial:

$\frac{\pi(x)}{1-\pi(x)} = e^{(\beta_0+\beta_1 x)}$.

El grafo de la función logística se muestra en la figura que encabeza esta sección, la variable independiente es la combinación lineal $\beta_0 + \beta_1 x$ y la variable dependiente es la probabilidad estimada $\pi(x)$. Si se realiza la regresión lineal, la forma de la probabilidad estimada puede ser fácilmente recuperada a partir de los coeficientes calculados:[1]

Para hacer la regresión deben tomarse los valores X_i de las observaciones ordenados de mayor a menor y formar la siguiente tabla:

Donde ε*i* es "0" o "1" según el caso y además:

$$0 \leq \pi(X_i) = \frac{\sum_{k=1}^{i} \varepsilon_k}{i} \leq 1, \qquad g(X_i) = \ln\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = \beta_0 + \beta_1 X_i$$

El el cálculo de *g* pueden aparecer problemas al principio del intervalo si π(*Xj*) = 0 para algunos valores de *j*.

3.3 Ejemplo

Sea *p*(*x*) la probabilidad de éxito cuando el valor de la variable predictora es *x*. Entonces sea

$$p(x) = \frac{1}{1 + e^{-(B_0+B_1 x)}} = \frac{e^{B_0+B_1 x}}{1 + e^{B_0+B_1 x}}.$$

Después de algún álgebra se prueba que

$$\frac{p(x)}{1-p(x)} = e^{B_0+B_1 x},$$

donde $\frac{p(x)}{1-p(x)}$ son los odds en favor de éxito.

Si tomamos un valor de ejemplo, digamos *p*(50) = 2/3, entonces

$$\frac{p(50)}{1-p(50)} = \frac{\frac{2}{3}}{1-\frac{2}{3}} = 2.$$

Cuando *x* = 50, un éxito es dos veces tan probable como una falla. Es decir, se puede decir simplemente que los odds son 2 a 1.

3.4 Extensiones

Algunas extensiones del modelo existen para tratar variables dependientes multicategóricas y/o ordinales, tales como la regresión politómica. La clasificación en varias clases por regresión logística es conocida como regresión logística multinomial. Una extensión del modelo logístico para ajustar conjuntos de variables independientes es el campo aleatorio condicional.

3.5 Véase también

- Red neuronal artificial
- Minería de datos
- Modelos de regresión múltiple postulados y no postulados
- Análisis lineal discriminante

- Perceptrón
- Modelo probit
- Análisis de regla de variables
- Modelo de Jarrow-Turnbull

3.6 Referencias

[1] Hosmer, David W.; Lemeshow, Stanley (2000). *Applied Logistic Regression* (2nd edición). Wiley. ISBN 0-471-35632-8.

3.6.1 Bibliografía

- Agresti, Alan. (2002). *Categorical Data Analysis*. New York: Wiley-Interscience. ISBN 0-471-36093-7.
- Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press. ISBN 0-674-00560-0.
- Balakrishnan, N. (1991). *Handbook of the Logistic Distribution*. Marcel Dekker, Inc. ISBN 978-0824785871.
- Green, William H. (2003). *Econometric Analysis, fifth edition*. Prentice Hall. ISBN 0-13-066189-9.
- Hosmer, David W.; Stanley Lemeshow (2000). *Applied Logistic Regression, 2nd ed*. New York; Chichester, Wiley. ISBN 0-471-35632-8.

3.6.2 Enlaces externos

- Web-based logistic regression calculator
- A highly optimized Maximum Entropy modeling package
- MALLET Java library, includes a trainer for logistic models
- La Regresión Logística. Páginas de Bioestadística de la Sociedad Española de Hipertensión Arterial

Capítulo 4

Modelos de regresión múltiple postulados y no postulados

En estadística un **modelo de regresión múltiple no postulado** es uno de los métodos de regresión lineal.

4.1 Modelo

Un modelo relaciona una o varias variables que hay que explicar Y a unas variables explicativas X, por una relación funcional Y = F (X)

- Un modelo físico es un modelo explicativo sostenido por una teoría.
- Un modelo estadístico, al contrario, es un modelo empírico nacido de datos disponibles, sin conocimientos a priori sobre los mecanismos en juego. Podemos sin embargo integrar en eso ecuaciones físicas (en el momento del pretratamiento de datos).

Disponemos de *n* de observaciones (*i* = 1,..., *n*) de *p* variables. La ecuación de regresión se escribe:

$y_i = a_o + a_1 x_{i,1} + \cdots + a_p x_{i,p} + \epsilon_i \qquad i = 1 \cdots n$

donde

- εi es el error del modelo;
- a_0, a_1, ..., *ap* son los coeficientes del modelo que hay que estimar.

El cálculo de los coeficientes *aj* y del error del modelo, a partir de las observaciones, es un problema bien dominado (ver Regresión lineal).

Más delicado es la elección de las variables que entran en este modelo. Puede ser postulado o no postulado.

4.2 Modelo postulado

Sólo los coeficientes del modelo precedente de regresión *son dirigidos por los datos*, la estructura polinómica del modelo es impuesta por el utilizador (según su peritaje del problema), que postula a priori:

- El tipo de modelo: lineal o polinómico, y el grado del polinomio,
- las variables que entrarán en el modelo.

Ejemplo de modelo polinómico con dos variables explicativas: $y_i = a_o + a_1 x_{i,1} + a_2 x_{i,2} + a_3 x_{i,1} x_{i,2} + a_4 x_{i,1}^2 + a_5 x_{i,1}^2 + \epsilon_i \qquad i = 1 \cdots n$

01

4.3 El problema de la selección de las variables explicativas

Cuando el número de variables explicativas es grande, puede hacerse que ciertas variables sean correlacionadas. En este caso hay que eliminar los doblones. El software utiliza para hacerlo métodos de selección paso a paso (ascendientes, descendentes o mixtos).

Sin embargo la calidad del modelo final repone en gran parte en la elección de las variables, y del grado del polinomio.

4.4 Modelo no postulado

El modelo *no postulado* es al contrario totalmente *dirigido por los datos* , tanto su estructura matemática como sus coeficientes. La selección de las variables explicativas no pide conocimiento a priori sobre el modelo: se efectúa entre un conjunto muy grande de variables, comprendiendo:

- **Variables explicativas simples**: A, B, C, (propuestas por los expertos del campo considerado y cuyo número p puede ser superior a n
- **Interacciones** o **acoplamiento** de estas variables, por ejemplo « A*B » (producido cruzado sobre variables centradas reducidas), pero también « interacciones lógicas » tal « A y B », « A o B », « A y B medios », « A si B es fuerte »,« A si B es medio », « A si B es débil », etc.;
- **Funciones de estas variables'**: por ejemplo cos (A) o cualquier función sinusoidal amortiguada o ampliada, función periódica no sinusoidal, efecto de umbral, etc.

La selección se produce *antes* del cálculo de los coeficientes de la regresión según el principio siguiente:

> Buscamos el factor o la *interacción* o la función mejor correlada a la respuesta. Habiéndolo encontrado, buscamos el factor o la interacción mejor correlada al *residuo* no explicado por la correlación precedente; etc. Este método pretende no contar dos veces la misma influencia, cuando los factores son correlados, y a ordenarlos por importancia decreciente.

La lista *por orden de importancia decreciente* encontrada y clasificada, no puede contar más términos que desconocidas (n). Si se guarda sólo un término en el modelo, deberá ser la primera de la lista. Si se guarda dos, serán ambos primeros, etc.

En efecto ya que cada uno de los términos de la lista *explica* el residuo no explicado por los precedentes, los últimos explican posiblemente sólo el *ruido*. ¿ Cuál criterio de parada escoger?

El número de términos conservados en el modelo puede ser, por ejemplo, el que minimiza el error estándar de predicción SEP (Standard error of Prediction), o el que maximiza el F de Fisher. Este número de término puede también ser escogido por el utilizador a partir de consideraciones físicas.

> **Ejemplo**: suponemos que el conjunto de las « variables explicativas » candidatas es {A,B,C,D,E,F,G}, y que el modelo obtenido es :
>
> Y = constante + a.A + b.(« E et G ») + c.(« D y F medios »)
>
> Observamos que:
>
> * las variables B y C, no pertinentes, no figuran en el modelo
>
> * la variable A apareció como término simple
>
> * las variables E y G de una parte, y D y F, por otra parte, aparecen sólo como « interacciones lógicas ».

Este modelo « *parsimonioso* », es decir conteniendo pocos términos (aquí tres), contrata 5 variables, y estará pegado mejor a la realidad física que un modelo polinómico. En efecto la conjunción « E y G » que significa « E y G fuertes simultáneamente » es encontrado más a menudo en la realidad física (ejemplo: la catálisis en química) que un término polinómico de tipo E.G.

4.5 Descomposición armónica

Un modelo no postulado será también eficaz en la descomposición armónica de las series.

En efecto, el principio se aplica también bien en caso de muestreo irregular (donde los métodos de tipo media móvil, ARIMA o Box y Jenkins son hechos caer en falta) que en los casos no estacionarios (donde Análisis armónico no se aplica). Permite descubrir y desenredar las interferencias de ciclos diversos y estacionalidad con roturas de tendencias en *escalón*, en *V*, *roturas logísticas*, motivos periódicos, y acontecimientos accidentales tales como picos aislados o *pedazos de ondas*.

4.6 Referencias

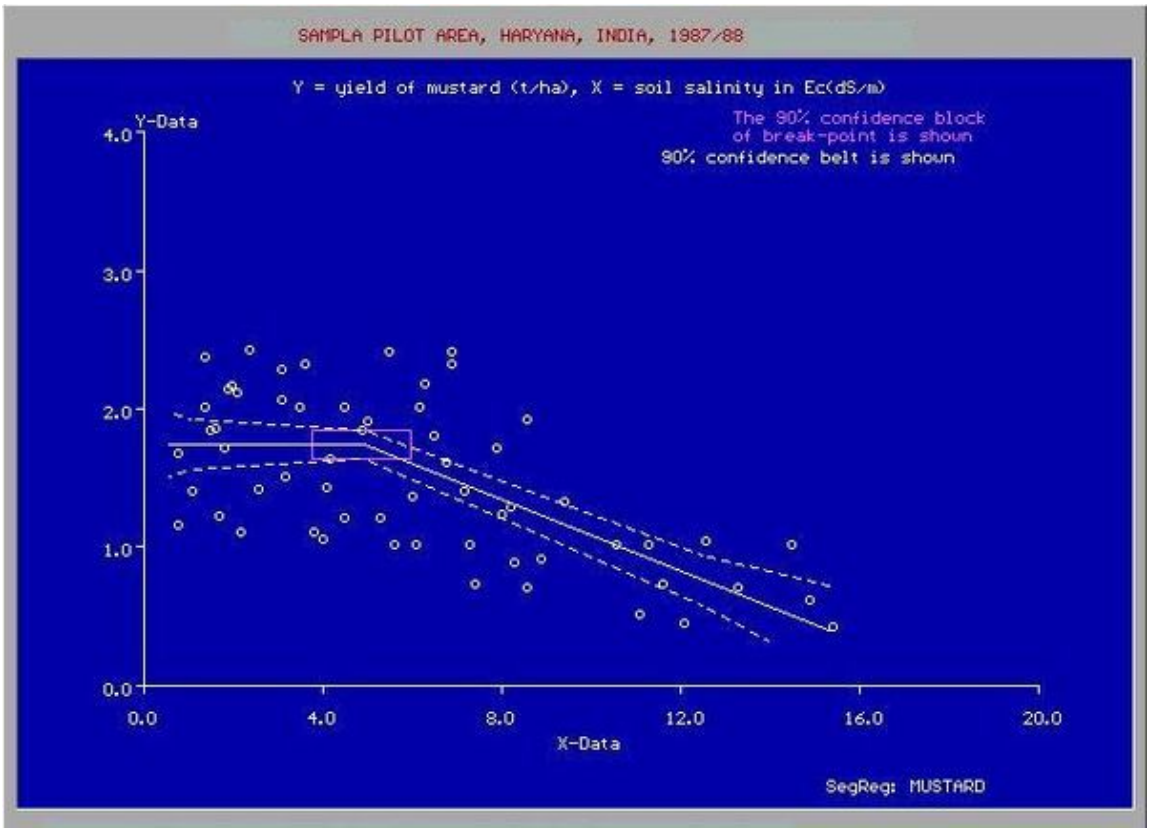
Lesty M. (1999) Une nouvelle approche dans le choix des régresseurs de la régression multiple en présence d'interactions et de colinéarités. La revue de Modulad, n°22, janvier 1999, pp. 41-77

Lesty M. (2002) La recherche des harmoniques, une nouvelle fonction du logiciel CORICO. La revue de Modulad, n°29, juin 2002, pp. 39-77

Capítulo 5

Regresión segmentada

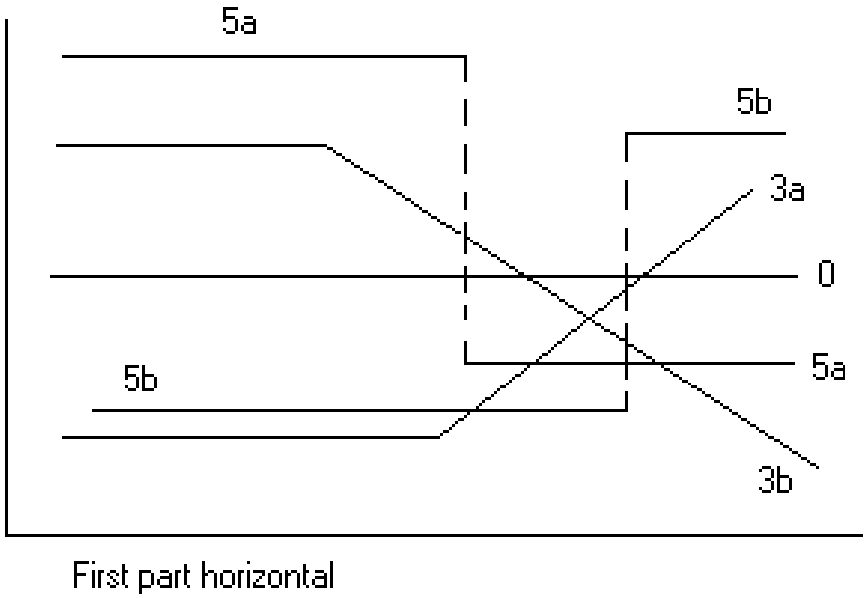
Regresión segmentada o *regresión por pedazos* es un método en el análisis de regresión en que el variable independiente es particionada en intervalos ajustando en cada intervalo una línea o curva a los datos. La regresión segmentada se puede aplicar también a la regresión con múltiples variables independientes particionando todas estas.



Regresión segmentada lineal, tipo 3

La regresión segmentada es util cuando el variable dependiente muestra una reacción abruptamente diferente a la variable independiente en los varios segmentos. En este caso el límite entre los segmentos se llama *punto de quiebra*.

Regresión segmentada lineal es la regresión segmentada en que la relación entre el variable dependiente e independiente dentro de los segmentos se obtiene por regresión lineal.



1.er miembro horizontal

5.1 Regresión segmentada lineal, 2 segmentos

Regresión segmentada lineal en dos segmentos separados por un punto de quiebra puede ser útil para cuantificar un cambio abrupto en la función de reacción de un factor de interés a la variación de otro factor influencial. El punto de quiebra se interpreta como un valor *seguro*, *crítico* o *umbral* cuando efectos (no) deseados suceden a uno de los dos lados.
El punto de quiebra puede ser un factor importante para la toma de decisiones de manejo.[1]

El análisis de la regresión segmentada se basa en la presencia de un juego de datos (**y, x**), donde **y** es el variable dependiente y **x** el variable independiente, es decir que el valor de **x** influye el valor de **y**.

El método de los mínimos cuadrados aplicado separadamente a cada segmento, por lo cual las dos líneas de regresión se ajustan a los datos tan cerca como posible minimizando la *suma de los cuadrados de las diferencias* (SCD) entre el valor observado (**y**) y valor calculado por regresión (*Yr*) de la variable dependiente, resulta en las ecuaciones siguientes:

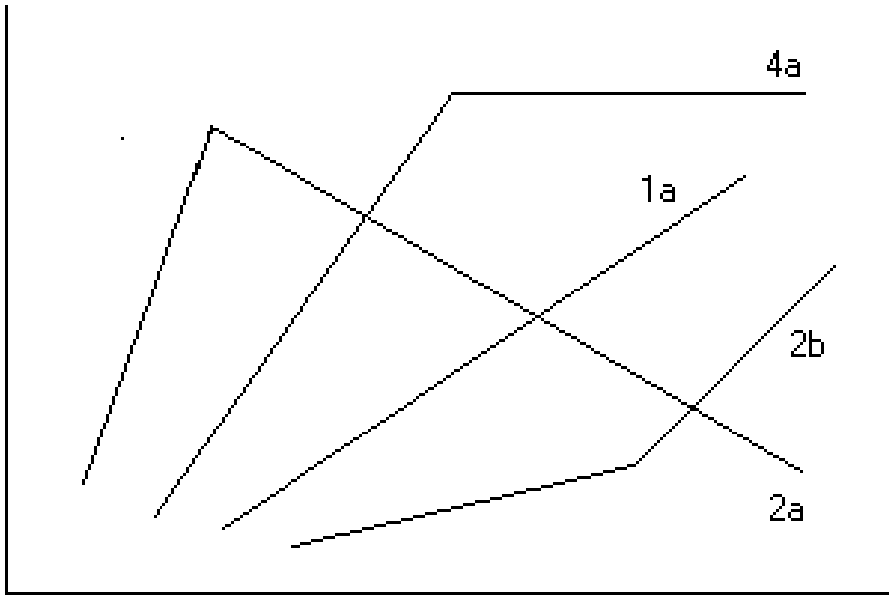
- $Yr = A_1 . \mathbf{x} + K_1$ para $\mathbf{x} <$ PQ (punto de quiebra)
- $Yr = A_2 . \mathbf{x} + K_2$ para $\mathbf{x} >$ PQ (punto de quiebra)

donde:

> *Yr* es el valor esperado (pronosticado) de **y** para un cierto valor de **x**
>
> A_1 y A_2 son los *coeficientes de regresión* indicando la inclinación de las líneas en los segmentos respectivos
>
> K_1 and K_2 son los *constantes de regresión* en los segmentos respectivos indicando los valores de *Yr* cuando $\mathbf{x} = 0$

Los datos pueden mostrar diferentes tipos de tendencia,[2] véase las figuras.

El método también rinde dos coeficientes de correlación:



First part sloping upward

1.er miembro inclinado hacia arriba

- $(R_1)^2$ = 1 – suma { $(\mathbf{y} - Yr)^2$ } / suma { $(\mathbf{y} - Ya1)^2$ } para **x** < PQ (punto de quiebra)
- $(R_2)^2$ = 1 – suma { $(\mathbf{y} - Yr)^2$ } / suma { $(\mathbf{y} - Ya2)^2$ } para **x** > PQ (punto de quiebra)

donde

suma { $(\mathbf{y} - Yr)^2$ } es la suma de cuadrados de las diferencias (SCD) minimizado por segmento
Ya1 e *Ya2* son los valores promedios de **y** en los segmentos respectivos

Cuando no se detecta un punto de quiebra, hay que volver a una regresión sin punto de quiebra.

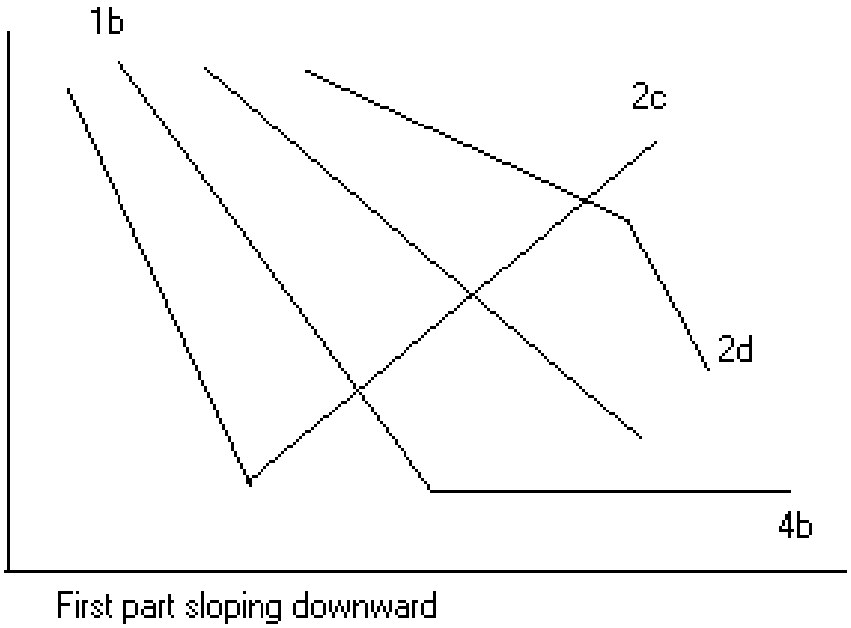
5.2 Ejemplo

Para la figura azul arriba, que da la relación entre la cosecha de mostaza (colza) en t/ha y la salinidad del suelo (**x** = *Ss*) expresada en conductividad eléctrica (EC en dS/m) de la solución del suelo,[3] se desprende que:

- *PQ* = 4.93, A_1 = 0, K_1 = 1.74, A_2 = –0.129, K_2 = 2.38, $(R_1)^2$ = 0.0035 (no significante), $(R_2)^2$ = 0.395 (significante) y:
- *Yr* = 1.74 t/ha para *Ss* < 4.93 (punto de quiebra)
- *Yr* = –0.129 *Ss* + 2.38 t/ha para *Ss* > 4.93 (punto de quiebra)

indicando que una salinidad del suelo < 4.93 dS/m es segura y una salinidad del suelo > 4.93 reduce la cosecha @ 0.129 tonelada/ha por unidad de aumento de salinidad de suelo.

La figura también muestra intervalos de confianza e inseguridad.



1.^{er} miembro inclinado hacia abajo

5.3 Procedimiento de pruebas

Las siguientes *pruebas estadísticas* se emplean para determinar el tipo de tendencia:

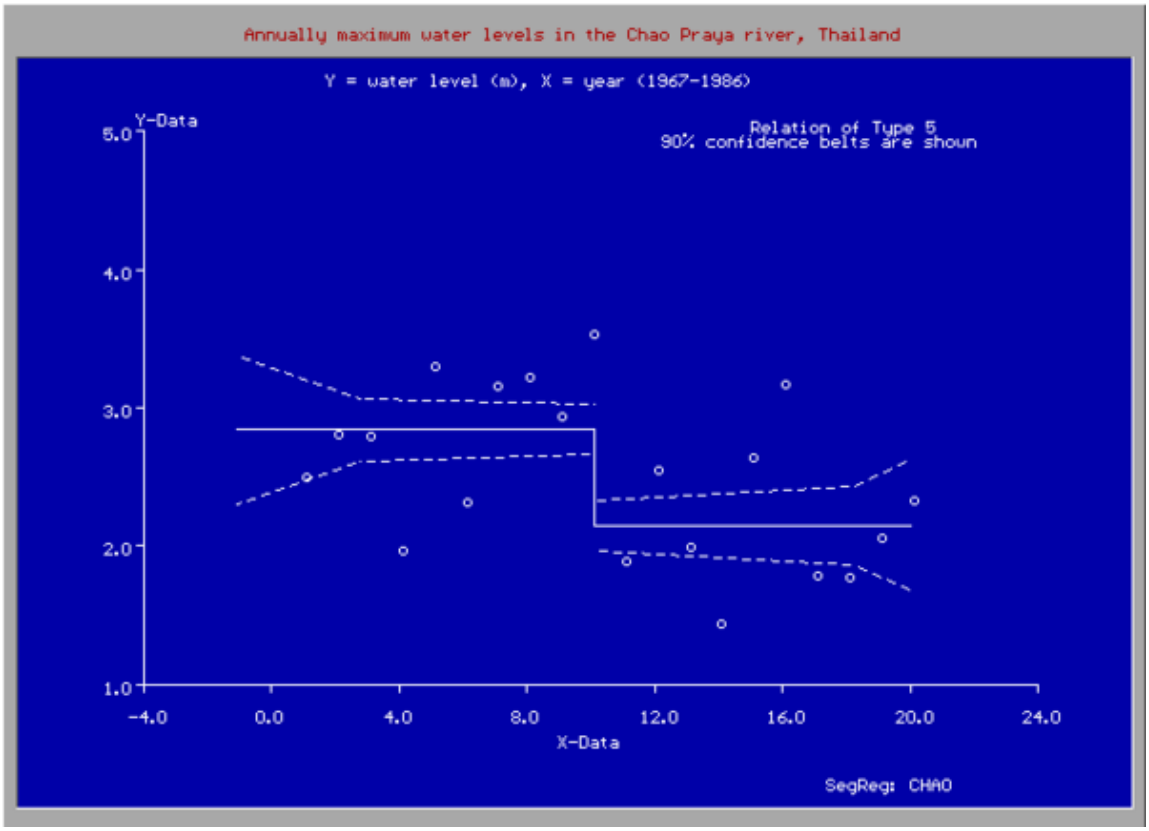
1. Significatividad estadística del punto de quiebra (PQ) expresando PQ como una función de los coeficientes de regresión A_1 y A_2, los promedios Y_1 e Y_2 de los datos **y**, y los promedios X_1 y X_2 de los datos **x** (al lado izquierdo y derecho de PQ respectivamente), utilizando la leyes de propagación de errores en adiciones y multiplicaciones para la computación del error estándar (ES) de PQ, seguido por la prueba t de Student
2. Significatividad estadística de A_1 y A_2 aplicando la prueba t de Student y el error estándar ES de A_1 y A_2
3. Significatividad estadística de la diferencia de A_1 y A_2 aplicando la prueba t de Student y el error estándar ES de la diferencia
4. Significatividad estadística de de la diferencia de Y_1 e Y_2 aplicando la prueba t de Student y el error estándar ES de la diferencia

Adicionalmente se emplea de coeficiente de correlación de todos los datos (*Ra*), el *coeficiente de determinación* (o *coeficiente de explicación*), intervalos de confianza de las funciones (líneas) de regresión, y un análisis de la varianza (ANOVA).[4]

El *coeficiente de determinación* de todos los datos (*Cd*), lo cual se debe maximizar bajo las condiciones especificados arriba en *pruebas estadísticas*, se defina como:

- $Cd = 1 - \text{suma } \{ (\mathbf{y} - Yr)^2 \} / \text{suma } \{ (\mathbf{y} - Ya)^2 \}$

donde *Yr* es el valor esperado (pronosticado) de **y** de acuerdo a las ecuaciones de regresión previas, y *Ya* es el promedio de todo los valores **y**. El coeficiente *Cd* puede variar entre 0 (ninguna explicación de la regresión segmentada) y 1



Ejemplo de una serie temporal de descargas de un río, tipo 5

(perfecta explicación).
En una regresión lineal pura, sin segmentación, los valores de *Cd* y Ra^2 son iguales. En la regresión segmentada, *Cd* debe ser significativamente mayor que Ra^2 para justificar la segmentación.

La optimización del punto de quiebra PQ se alcanza probando una serie de puntos tentativos y seleccionando el punto que tiene el coeficiente *Cd* máximo.

5.4 Referencias

[1] *Frequency and Regression Analysis*. Chapter 6 in: H.P.Ritzema (ed., 1994), *Drainage Principles and Applications*, Publ. 16, pp. 175-224, International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. ISBN 90 70754 3 39. Bajar de: , bajo no. 13, o directamente como PDF:

[2] *Drainage research in farmers' fields: analysis of data*. Part of project "Liquid Gold" of the International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. Bajar como PDF:

[3] R.J.Oosterbaan, D.P.Sharma, K.N.Singh and K.V.G.K.Rao, 1990, *Crop production and soil salinity: evaluation of field data from India by segmented linear regression*. In: Proceedings of the Symposium on Land Drainage for Salinity Control in Arid and Semi-Arid Regions, February 25th to March 2nd, 1990, Cairo, Egypt, Vol. 3, Session V, p. 373 - 383

[4] *Statistical significance of segmented linear regression with break-point using variance analysis and F-tests*. Bajar de: , bajo. no. 13, o directamente como PDF:

5.5 Enlaces externos

- SegReg, programa libre para regresión segmentada lineal con 2 variables independientes.

Capítulo 6

Econometría

La **econometría** (del griego οἰκονόμος *oiko-nomos*, 'regla para la administración doméstica' y μετρία *metría*, 'relativo a la medida') es la rama de la economía que hace un uso extensivo de modelos matemáticos y estadísticos así como de la programación lineal y la teoría de juegos para analizar, interpretar y hacer predicciones sobre sistemas económicos, prediciendo variables como el precio, las reacciones del mercado, el coste de producción, la tendencia de los negocios y las consecuencias de la política económica.

6.1 Introducción

La economía, perteneciente a las ciencias sociales, trata de explicar el funcionamiento del sistema económico en sus distintos aspectos, como producción, consumo, dinero, distribución del ingreso, etc. La herramienta más utilizada por los economistas es la construcción de modelos económicos teóricos y matemáticos que describan el comportamiento de los agentes económicos. Sin embargo, esos modelos deben contrastarse con los datos disponibles para saber si éstos tienen capacidad explicativa y predictiva, y poder en definitiva optar entre unas u otras opciones. La construcción de tales modelos es la finalidad de la econometría.

Los econometristas (economistas cuantitativos) han tratado de emular a las ciencias naturales (física, química) con mejor o peor resultado a través del tiempo. Hay que considerar que tratan con uno de los fenómenos más complejos que conocemos, el comportamiento de las personas. Actualmente, la econometría no necesariamente requiere o presupone una teoría económica subyacente al análisis econométrico. Más aún: la econometría moderna se precia de prescindir voluntariamente de la teoría económica por considerarla un obstáculo si se quiere realizar un análisis riguroso (ésta es, por ejemplo, la filosofía del método de Vector Autorregresivos - VAR).

En la elaboración de la econometría se unen la estadística y la investigación social y la teoría económica. El mayor problema con el que se enfrentan los económetras en su investigación es la escasez de datos, los sesgos que pueden presentar los datos existentes y la ausencia o insuficiencia de una teoría económica adecuada. Aún así, la econometría es la única aproximación científica al entendimiento de los fenómenos económicos.

6.1.1 Definiciones de econometría

Entre las **definiciones** de econometría que los economistas relevantes han formulado a lo largo de la historia, podemos destacar las siguientes:

- Ragnar Frisch (1930): 'La experiencia ha mostrado que cada uno de estos tres puntos de vista, el de la estadística, la teoría económica y las matemáticas, es necesario, pero por sí mismo no suficiente para una comprensión real de las relaciones cuantitativas de la vida económica moderna. Es la *unión* de los tres aspectos lo que constituye una herramienta de análisis potente. Es la unión lo que constituye la econometría".

- Paul Samuelson, Tjalling Koopmans y Richard Stone (1954): '... el análisis cuantitativo de fenómenos económicos actuales, basado en el desarrollo congruente de teoría y observaciones, y relacionado por métodos apropiados de inferencia.'

- Valavanis (1959): 'El objetivo de la econometría es expresar las teorías económicas bajo una forma matemática a fin de verificarlas por métodos estadísticos y medir el impacto de una variable sobre otra, así como predecir acontecimientos futuros y dar consejos de política económica ante resultados deseables.'

- A.G. Barbancho (1962): 'La econometría es la rama más operativa de la Ciencia económica, trata de representar numéricamente las relaciones económicas mediante una adecuada combinación de la Teoría económica matemática y la Estadística. De forma que las matemáticas, como lenguaje y forma de expresión simbólica e instrumento eficaz en el proceso deductivo, representan el medio unificador; y teoría económica, economía matemática o estadística económica serían consideraciones parciales de su contenido.'

- Klein (1962): 'El principal objetivo de la econometría es dar contenido empírico al razonamiento a priori de la economía.'

- Malinvaud (1966): '... aplicación de las matemáticas y método estadístico al estudio de fenómenos económicos.'

- Christ (1966): 'Producción de declaraciones de economía cuantitativa que explican el comportamiento de variables ya observadas, o predicen la conducta de variables aún no observadas.'

- Intriligator (1978): 'Rama de la economía que se ocupa de la estimación empírica de relaciones económicas.'

- G.C. Chow (1983): 'Arte y ciencia de usar métodos para la medida de relaciones económicas.'

En cualquier caso, la definición de economía es tan amplia que todas las anteriores son aceptables.

6.1.2 Descripción somera de la econometría

La econometría se ocupa de obtener, a partir de los valores reales de variables económicas y a través del análisis estadístico y matemático (mas no de la teoría económica, como si se usa en las ciencias naturales, como la física), los *valores* que tendrían los *parámetros* (en el caso concreto de la estimación paramétrica) de los modelos en los que esas variables económicas aparecieran, así como de *comprobar el grado de validez* de esos modelos, y ver en qué medida estos modelos pueden usarse para explicar la economía de un agente económico (como una empresa o un consumidor), o la de un agregado de agentes económicos, como podría ser un sector del mercado, o una zona de un país, o todo un país, o cualquier otra zona económica; su evolución en el tiempo (por ejemplo, decir si ha habido o no cambio estructural), poder predecir valores futuros de la variables, y sugerir medidas de política económica conforme a objetivos deseados (por ejemplo, para poder aplicar técnicas de optimización matemática para racionalizar el uso de recursos dentro de una empresa, o bien para decidir qué valores debería adoptar la política fiscal de un gobierno para conseguir ciertos niveles de recaudación impositiva).

Usualmente se usan técnicas estadísticas diversas para estudiar la economía, pero uno de los métodos más usados es el que se mostrará aquí.

6.1.3 Concepto de modelo econométrico

La econometría, igual que la economía, tiene como objetivo explicar una variable en función de otras. Esto implica que el punto de partida para el análisis econométrico es el modelo económico y este se transformará en modelo econométrico cuando se han añadido las especificaciones necesarias para su aplicación empírica. Es decir, cuando se han definido las variables (endógenas, exógenas) que explican y determinan el modelo, los parámetros estructurales que acompañan a las variables, las ecuaciones y su formulación en forma matemática, la perturbación aleatoria que explica la parte no sistemática del modelo, y los datos estadísticos.

A partir del modelo econométrico especificado, en una segunda etapa se procede a la estimación, fase estadística que asigna valores numéricos a los parámetros de las ecuaciones del modelo. Para ello se utilizan métodos estadísticos como pueden ser: Mínimos cuadrados ordinarios, Máxima verosimilitud, Mínimos cuadrados bietápicos, etc. Al recibir los parámetros el valor numérico definen el concepto de estructura que ha de tener valor estable en el tiempo especificado.

La tercera etapa en la elaboración del modelo es la verificación y contrastación, donde se someten los parámetros y la variable aleatoria a unos contrastes estadísticos para cuantificar en términos probabilísticos la validez del modelo estimado.

La cuarta etapa consiste en la aplicación del modelo conforme al objetivo del mismo. En general los modelos econométricos son útiles para:

1. Análisis estructural y entender como funciona la economía.
2. Predicción de los valores futuros de las variables económicas.
3. Simular con fines de planificación distintas posibilidades de las variables exógenas.
4. Simular con fines de control valores óptimos de variables instrumentales de política económica y de empresa.

6.2 Métodos de la econometría

6.2.1 El método de mínimos cuadrados (estimación MCO)

También se conoce como teoría de la regresión lineal, y estará más desarrollado en la parte estadística de la enciclopedia. No obstante, aquí se dará un resumen general sobre la aplicación del método de mínimos cuadrados.

Se parte de representar las relaciones entre una variable económica endógena y una o más variables exógenas de forma lineal, de la siguiente manera:

$$Y = a_1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n$$

"Y" es la variable endógena, cuyo valor es determinado por las exógenas, X_1 hasta X_n . Cuales son las variables elegidas depende de la teoría económica que se tenga en mente, y también de análisis estadísticos y económicos previos. El objetivo buscado sería obtener los valores de los parámetros desde a_1 hasta β_n . A menudo este modelo se suele completar añadiendo un término más a la suma, llamado término independiente, que es un parámetro más a buscar. Así:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n \ .$$

En el que β_0 es una constante, que también hay que averiguar. A veces resulta útil, por motivos estadísticos, suponer que siempre hay una constante en el modelo, y contrastar la hipótesis de si es distinta, o no, de cero para reescribirlo de acuerdo con ello.

Además, se supone que esta relación no es del todo determinista, esto es, existirá siempre un cierto grado de error aleatorio (en realidad, se entiende que encubre a todas aquellas variables y factores que no se hayan podido incluir en el modelo) que se suele representar añadiendo a la suma una letra representa una variable aleatoria.Así:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n + \mu$$

Se suele suponer que μ es una variable aleatoria normal, con media cero y varianza constante en todas las muestras (aunque sea desconocida).

Se toma una muestra estadística, que corresponda a observaciones de los valores que hayan tomado esas variables en distintos momentos del tiempo (o, dependiendo del tipo de modelo, los valores que hayan tomado en distintas áreas o zonas o agentes económicos a considerar).

Por ejemplo, en un determinado modelo podemos estar interesados en averiguar como la renta ha dependido de los niveles de precios, de empleo y de tipos de interés a *lo largo de los años* en *cierto país*, mientras que en otro podemos estar interesados en ver como, a *lo largo de un mismo año*, ha dependido la renta *de distintos países* de esas mismas variables. Por lo que tendríamos que observar, en el primer caso, la renta, niveles de empleo, precios y tipos de interés del año 1, lo mismo, pero del año 2, etcétera, para obtener la muestra a lo largo de varios años, mientras que en el segundo caso tendríamos que tener en cuenta los valores de cada uno de los países para obtener la muestra. Cada una de esas observaciones para cada año, o país, se llamaría observación muestral. Nótese que aún se podría hacer un análisis más ambicioso teniendo en cuenta *país y año*.

Una vez tomada la muestra, se aplica un método, que tiene su justificación matemática y estadística, llamado método de mínimos cuadrados. Este consiste en, básicamente, minimizar la suma de los errores (elevados al cuadrado) que se tendrían, suponiendo distintos valores posibles para los parámetros, al estimar los valores de la variable endógena a partir de los de las variables exógenas en cada una de las observaciones muestrales, usando el modelo propuesto, y comparar esos valores con los que realmente tomó la variable endógena. Los parámetros que lograran ese mínimo, el de las suma de los errores cuadráticos, se acepta que son los que estamos buscando, de acuerdo con criterios estadísticos.

También, este método nos proporcionará información (en forma de ciertos valores estadísticos adicionales, que se obtienen además de los parámetros) para ver en qué medida los valores de los parámetros que hemos obtenido resultan fiables, por ejemplo, para hacer **contrastes de hipótesis**, esto es, ver **si ciertas suposiciones que se habían hecho acerca del modelo resultan, o no, ciertas**. Se puede usar también esta información adicional para comprobar si se pueden prescindir de algunas de esas variables, para ver si es posible que los valores de los parámetros hayan cambiado con el tiempo (o si los valores de los parámetros son diferentes en una zona económica de los de otra, por ejemplo), o para ver en qué grado son válidas predicciones acerca del futuro valor de la variable endógena si se supone que las variables exógenas adoptarán nuevos valores.

6.2.2 Problemas del método de los mínimos cuadrados

El método de los mínimos cuadrados tiene toda una serie de problemas, cuya solución, en muchas ocasiones aproximada, ha estado ocupando el trabajo de los investigadores en el campo de la econometría.

De entrada, el método presupone que la relación entre las variables es lineal y está bien especificada. Para los casos de no linealidad se recurre, bien a métodos para obtener una relación lineal que sea equivalente, bien a aproximaciones lineales, o bien a métodos de optimización que absorban la relación no lineal para obtener también unos valores de los parámetros que minimicen el error cuadrático.

Otro supuesto del modelo es el de normalidad de los errores del modelo, que es importante de cara a los contrastes de hipótesis con muestras pequeñas. No obstante, en muestras grandes el teorema del límite central justifica el suponer una distribución normal para el estimador de mínimos cuadrados.

No obstante, el problema se complica considerablemente, sobre todo a la hora de hacer contrastes de hipótesis, si se cree que la varianza de los errores del modelo cambia con el tiempo. Es el fenómeno conocido como heterocedasticidad (el fenómeno contrario es la homocedasticidad). Este fenómeno se puede detectar con ciertas técnicas estadísticas. Para resolverlo hay que usar métodos que intenten estimar el cambiante valor de la varianza y usar lo obtenido para corregir los valores de la muestra. Esto nos llevaría al método conocido como mínimos cuadrados generalizados. Una versión más complicada de este problema es cuando se supone que, además, no solo cambia la varianza del error sino que también los errores de distintos periodos están correlacionados, lo que se llama autocorrelación. También hay métodos para detectar este problema y para corregirlo en cierta medida modificando los valores de la muestra, que también son parte del método de los mínimos cuadrados generalizados.

Otro problema que se da es el de la multicolinealidad, que generalmente sucede cuando alguna de las variables exógenas en realidad depende, también de forma estadística, de otra variable exógena del mismo modelo considerado, lo que introduce un sesgo en la información aportada a la variable endógena y puede hacer que el método de mínimos cuadrados no se pueda aplicar correctamente. Generalmente la solución suele ser averiguar qué variables están causando la multicolinealidad y reescribir el modelo de acuerdo con ello.

También hay que tener en cuenta que en ciertos modelos puede haber relaciones dinámicas, esto es, que una variable exógena dependa, además, de los valores que ella misma y/u otras variables tomaron en tiempos anteriores. Para resolver estos problemas se estudian lo que se llama modelos de series temporales.

6.3 Software econométrico

Entre los programas más empleados se encuentran SAS, Stata, RATS, TSP, SPSS, Limdep y WinBugs. Para más detalles, pueden verse las siguientes referencias.

- EViews
- Gauss

- Gretl
- Microfit
- R
- Limdep
- SPSS
- Stata

6.4 Lecturas recomendadas

- Jeffrey M. Wooldridge: "Introductory Econometrics: A Modern Approach, 4e"
- Alfonso Novales: "Econometría"
- J.M. Caridad y Ocerín: "Econometría: modelos econométricos y series temporales"
- Damodar M. Gujaratí: "Econometría"
- Pindyck: "Modelos y pronósticos"
- Guisán, María del Carmen: "Econometría"
- Alfonso G. Barbancho: "Fundamentos y posibilidades de la Econometría"

6.5 Véase también

- economía
- economía social
- microeconomía
- serie temporal

6.6 Referencia

6.6.1 Bibliografía

- *Handbook of Econometrics* Elsevier. Links to volume chapter-preview links:
  Zvi Griliches and Michael D. Intriligator, ed. (1983). v. 1; (1984),v. 2; (1986), description, v. 3; (1994), description, v. 4
  Robert F. Engle and Daniel L. McFadden, ed. (2001).Description, v. 5
  James J. Heckman and Edward E. Leamer, ed. (2007). Description, v. 6A & v. 6B
- *Handbook of Statistics*, v. 11, *Econometrics* (1993), Elsevier. Links to first-page chapter previews.
- *International Encyclopedia of the Social & Behavioral Sciences* (2001), Statistics, "Econometrics and Time Series," links to first-page previews of 21 articles.
- Angrist, Joshua & Pischke, Jörn-Steffen (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics*], 24(2), , pp. 3–30. Abstract.*
- Eatwell, John, *et al.*, eds. (1990). *Econometrics: The New Palgrave*. Article-preview links (from *The New Palgrave: A Dictionary of Economics, 1987).*
- Geweke, John; Horowitz, Joel; Pesaran, Hashem (2008). «Econometrics». En Durlauf, Steven N.; Blume, Lawrence E. *The New Palgrave Dictionary of Economics* (Palgrave Macmillan). doi:10.1057/9780230226203.0425.

- Greene, William H. (2012, 7th ed.) *Econometric Analysis*, Prentice Hall.
- Hayashi, Fumio. (2000) *Econometrics*, Princeton University Press. ISBN 0-691-01018-8 Description and contents links.
- Hamilton, James D. (1994) *Time Series Analysis*, Princeton University Press. Description and preview.
- Hughes Hallett, Andrew J (1989). «Econometrics and the Theory of Economic Policy: The Tinbergen-Theil Contributions 40 Years On». *Oxford Economic Papers* **41** (1): 189–214.
- Kelejian, Harry H., and Wallace E. Oates (1989, 3rd ed.) *Introduction to Econometrics*.
- Kennedy, Peter (2003). *A guide to econometrics*. Cambridge, Mass: MIT Press. ISBN 978-0-262-61183-1.
- Russell Davidson and James G. MacKinnon (2004). *Econometric Theory and Methods*. New York: Oxford University Press. Description.
- Mills, Terence C., and Kerry Patterson, ed. *Palgrave Handbook of Econometrics*:

  (2007) v. 1: *Econometric Theory*v. 1. Links to description and contents.

  (2009) v. 2, *Applied Econometrics*. Palgrave Macmillan. ISBN 978-1-4039-1799-7 Links to description and contents.

- Pearl, Judea (2009, 2nd ed.). *Causality: Models, Reasoning and Inference*, Cambridge University Press, Description, TOC, and preview, ch. 1-10 and ch. 11. 5 economics-journal reviews, including Kevin D. Hoover, *Economics Journal*.
- Pindyck, Robert S., and Daniel L. Rubinfeld (1998, 4th ed.). *Econometric Methods and Economic Forecasts*, McGraw-Hill.
- Santos Silva, J.M.C. and Tenreyro, Silvana (2006), "The Log of Gravity," The Review of Economics and Statistics, 88(4), pp. 641–658. <http://www.mitpressjournals.org/doi/pdfplus/10.1162/rest.88.4.641>
- Studenmund, A.H. (2011, 6th ed.). *Using Econometrics: A Practical Guide*. Contents (chapter-preview) links.
- Wooldridge, Jeffrey (2003). *Introductory Econometrics: A Modern Approach*. Mason: Thomson South-Western. ISBN 0-324-11364-1 Chapter-preview links in brief and detail.

6.6.2 Enlaces externos

- Economía Social
- Diapositivas Libro wooldridge
- Asociación de Econometría Aplicada

Capítulo 7

Mínimos cuadrados

Mínimos cuadrados es una técnica de análisis numérico enmarcada dentro de la optimización matemática, en la que, dados un conjunto de pares ordenados: variable independiente, variable dependiente, y una familia de funciones, se intenta encontrar la función continua, dentro de dicha familia, que mejor se aproxime a los datos (un "mejor ajuste"), de acuerdo con el criterio de *mínimo error cuadrático*.

En su forma más simple, intenta minimizar la suma de cuadrados de las diferencias *en las ordenadas* (llamadas *residuos*) entre los puntos generados por la función elegida y los correspondientes valores en los datos. Específicamente, se llama *mínimos cuadrados promedio* (LMS) cuando el número de datos medidos es 1 y se usa el método de descenso por gradiente para minimizar el residuo cuadrado. Se puede demostrar que LMS minimiza el residuo cuadrado esperado, con el mínimo de operaciones (por iteración), pero requiere un gran número de iteraciones para converger.

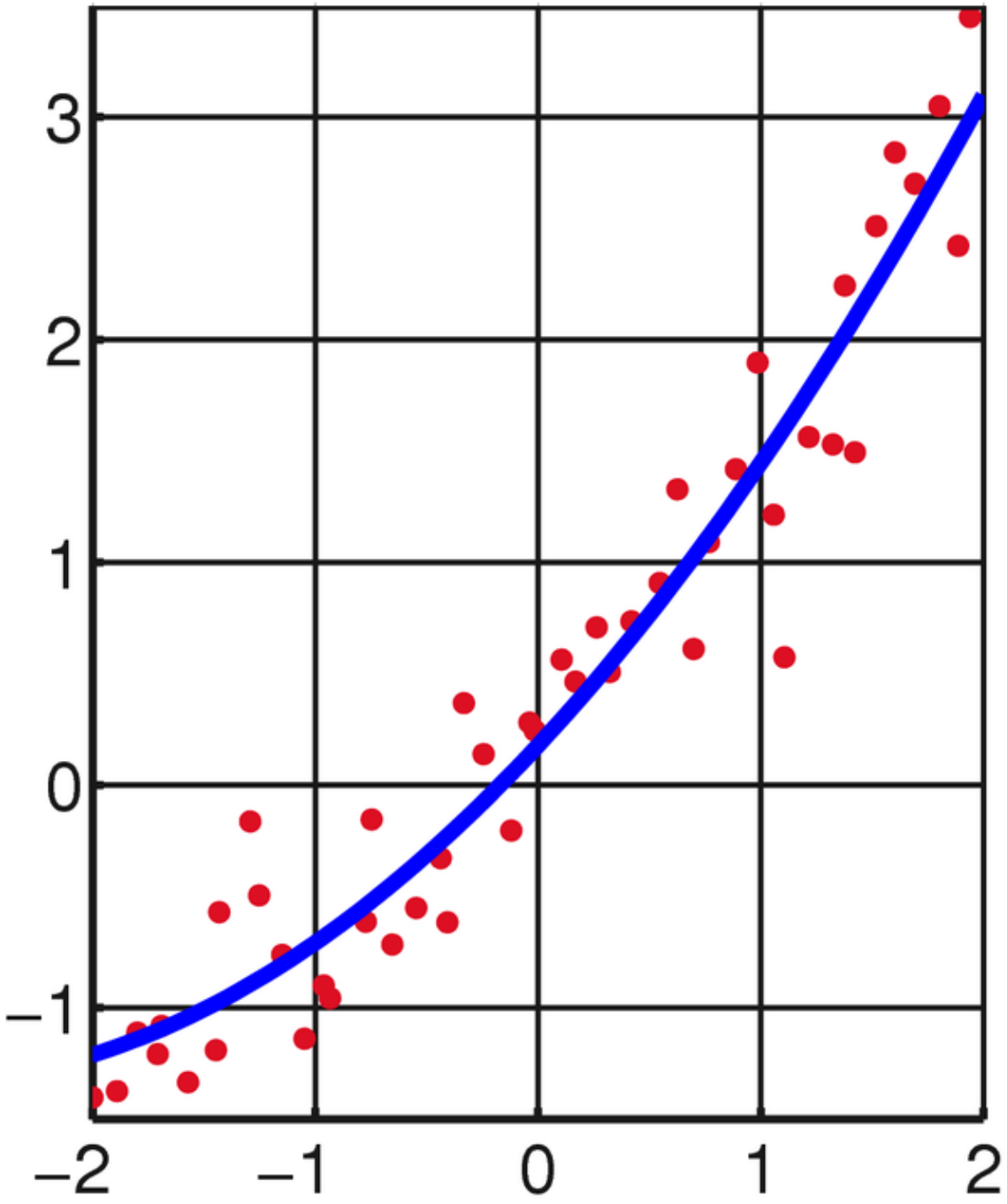
Desde un punto de vista estadístico, un requisito implícito para que funcione el método de mínimos cuadrados es que los errores de cada medida estén distribuidos de forma aleatoria. El teorema de Gauss-Márkov prueba que los estimadores mínimos cuadráticos carecen de sesgo y que el muestreo de datos no tiene que ajustarse, por ejemplo, a una distribución normal. También es importante que los datos a procesar estén bien escogidos, para que permitan visibilidad en las variables que han de ser resueltas (para dar más peso a un dato en particular, véase mínimos cuadrados ponderados).

La técnica de mínimos cuadrados se usa comúnmente en el ajuste de curvas. Muchos otros problemas de optimización pueden expresarse también en forma de mínimos cuadrados, minimizando la energía o maximizando la entropía.

7.1 Historia

El día de Año Nuevo de 1801, el astrónomo italiano Giuseppe Piazzi descubrió el planeta enano Ceres. Fue capaz de seguir su órbita durante 40 días. Durante el curso de ese año, muchos científicos intentaron estimar su trayectoria con base en las observaciones de Piazzi (resolver las ecuaciones no lineales de Kepler de movimiento es muy difícil). La mayoría de las evaluaciones fueron inútiles; el único cálculo suficientemente preciso para permitir a Franz Xaver von Zach, astrónomo alemán, reencontrar a Ceres al final del año fue el de Carl Friedrich Gauss, por entonces un joven de 24 años (los fundamentos de su enfoque ya los había planteado en 1795, cuando aún tenía 18 años). Sin embargo, su método de mínimos cuadrados no se publicó sino hasta 1809, y apareció en el segundo volumen de su trabajo sobre mecánica celeste, *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*. El francés Adrien-Marie Legendre desarrolló el mismo método de forma independiente en 1805.

En 1829, Gauss fue capaz de establecer la razón del éxito maravilloso de este procedimiento: simplemente, el método de mínimos cuadrados es óptimo en muchos aspectos. El argumento concreto se conoce como teorema de Gauss-Márkov.



El resultado del ajuste de un conjunto de datos a una función cuadrática.

7.2 Formulación formal del problema bidimensional

Sea $\{(x_k, y_k)\}_{k=1}^{n}$ un conjunto de n puntos en el plano real, y sea $\{f_j(x)\}_{j=1}^{m}$ una base de m funciones linealmente independiente en un espacio de funciones. Queremos encontrar una función $f(x)$ que sea combinación lineal de las funciones base, de modo que $f(x_k) \approx y_k$, esto es:

$$f(x) = \sum_{j=1}^{m} c_j f_j(x)$$

Karl Friedrich Gauss

Por tanto, se trata de hallar los m coeficientes c_j que hagan que la función aproximante $f(x)$dé la mejor aproximación para los puntos dados (x_k, y_k). El criterio de "mejor aproximación" puede variar, pero en general se basa en aquél que minimice una "acumulación" del error individual (en cada punto) sobre el conjunto total. En primer lugar, el error (con signo positivo o negativo) de la función $f(x)$en un solo punto, (x_k, y_k) , se define como:

$$e_k = y_k - f(x_k)$$

pero se intenta medir y minimizar el error en todo el conjunto de la aproximación, $\{(x_k, y_k)\}_{k=1}^{n}$. En matemáticas, existen diversas formas de definir el error, sobre todo cuando éste se refiere a un conjunto de puntos (y no sólo a uno), a una función, etc. Dicho error (el error "total" sobre el conjunto de puntos considerado) suele definirse con alguna de las siguientes fórmulas:

Error Máximo: $E_\infty(f) = max(|e_k|)$
Error Medio: $E_m(f) = \frac{\sum_{k=1}^{n} |e_k|}{n}$
Error cuadrático medio: $E_{cm}(f) = \sqrt{\frac{\sum_{k=1}^{n} (e_k)^2}{n}}$

La aproximación por mínimos cuadrados se basa en la minimización del error cuadrático medio o, equivalentemente, en la minimización del radicando de dicho error, el llamado error cuadrático, definido como:

$$E_c(f) = \frac{\sum_{k=1}^{n} (e_k)^2}{n}$$

Para alcanzar este objetivo, se utiliza el hecho que la función *f* debe poder describirse como una combinación lineal de una base de funciones. Los coeficientes de la combinación lineal serán los parámetros que queremos determinar. Por ejemplo, supongamos que *f* es una función cuadrática, lo que quiere decir que es una combinación lineal, $f(x) = ax^2 + bx + c$, de las funciones $f_1(x) = x^2$, $f_2(x) = x$ y $f_3(x) = 1$ (m=3 en este caso), y que se pretende determinar los valores de los coeficientes: a, b, c, de modo que minimicen la suma (*S*) de los cuadrados de los residuos:

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - ax_i^2 - bx_i - c)^2$$

Esto explica el nombre de *mínimos cuadrados*. A las funciones que multiplican a los coeficientes buscados, que en este caso son: x^2 , x y 1 , se les conoce con el nombre de funciones base de la aproximación, y pueden ser funciones cualesquiera. Para ese caso general se deduce a continuación la fórmula de la mejor aproximación discreta (i.e. para un conjunto finito de puntos), lineal y según el criterio del error cuadrático medio, que es la llamada *aproximación lineal por mínimos cuadrados*. Es posible generar otro tipo de aproximaciones, si se toman los errores máximo o medio, por ejemplo, pero la dificultad que entraña operar con ellos, debido al valor absoluto de su expresión, hace que sean difíciles de tratar y casi no se usen.

7.3 Solución del problema de los mínimos cuadrados

La aproximación mínimo cuadrática consiste en minimizar el error cuadrático mencionado más arriba, y tiene solución general cuando se trata de un problema de aproximación lineal (lineal en sus coeficientes c_j) cualesquiera que sean las funciones base: $f_j(x)$ antes mencionadas. Por lineal se entiende que la aproximación buscada se expresa como una combinación lineal de dichas funciones base. Para hallar esta expresión se puede seguir un camino analítico, expuesto abajo, mediante el cálculo multivariable, consistente en optimizar los coeficientes c_j ; o bien, alternativamente, seguir un camino geométrico con el uso de el álgebra lineal, como se explica más abajo, en la llamada deducción geométrica. Para los Modelos estáticos uniecuacionales, el método de mínimos cuadrados no ha sido superado, a pesar de diversos intentos para ello, desde principios del Siglo XIX. Se puede demostrar que, en su género, es el que proporciona la mejor aproximación.

7.3.1 Deducción analítica de la aproximación discreta mínimo cuadrática lineal

Sea $\{(x_k, y_k)\}_{k=1}^{n}$ un conjunto de n pares con abscisas distintas, y sea $\{f_j(x)\}_{j=1}^{m}$ un conjunto de m funciones linealmente independientes (en un espacio vectorial de funciones), que se llamarán funciones base. Se desea encontrar una función $f(x)$ de dicho espacio, o sea, combinación lineal de las funciones base, tomando por ello la forma:
$f(x) = c_1 f_1(x) + c_2 f_2(x) + ... + c_m f_m(x) = \sum_{j=1}^{m} c_j f_j(x)$.
Ello equivale por tanto a hallar los m coeficientes: $\{c_j(x)\}_{j=1}^{m}$. En concreto, se desea que tal función $f(x)$ sea la mejor aproximación a los n pares $(x_k, y_k)_1^n$ empleando, como criterio de "mejor", el criterio del mínimo error cuadrático medio de la función $f(x)$ con respecto a los puntos $(x_k, y_k)_1^n$.

El error cuadrático medio será para tal caso:
$E_{cm} = \sqrt{\frac{\sum_{k=1}^{n}(e_k)^2}{n}} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - f(x_k))^2} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \sum_{j=1}^{m} c_j f_j(x_k))^2}$

Minimizar el error cuadrático medio es equivalente a minimizar el error cuadrático, definido como el radicando del error cuadrático medio, esto es:
$E_c = \sum_{k=1}^{n}(y_k - \sum_{j=1}^{m} c_j f_j(x_k))^2$

Así, los c_j que minimizan E_{cm} también minimizan E_c , y podrán ser calculados derivando e igualando a cero este último:
$\frac{\partial E_c}{\partial c_i} = \sum_{k=1}^{n} 2(y_k - \sum_{j=1}^{m} c_j f_j(x_k))(-f_i(x_k)) = 0$ Siendo i=1,2, . . .,m

Se obtiene un sistema de m ecuaciones con m incógnitas, que recibe el nombre de *"Ecuaciones Normales de Gauss"*.
Operando con ellas:
$\sum_{k=1}^{n}(\sum_{j=1}^{m} c_j f_j(x_k))f_i(x_k) = \sum_{k=1}^{n} y_k f_i(x_k)$, para i=1,2, . . .,m
$\sum_{j=1}^{m}(\sum_{k=1}^{n} f_i(x_k)f_j(x_k))c_j = \sum_{k=1}^{n} y_k f_i(x_k)$, para i=1,2, . . .,m
Si se desarrolla la suma, se visualiza la ecuación "i-ésima" del sistema de m ecuaciones normales: $(\sum_{k=1}^{n} f_i(x_k)f_1(x_k))c_1 + (\sum_{k=1}^{n} f_i(x_k)f_2(x_k))c_2 + ... + (\sum_{k=1}^{n} f_i(x_k)f_m(x_k))c_m = \sum_{k=1}^{n} y_k f_i(x_k)$, para cada i=1,2, . . .,m
Lo cual, en forma matricial, se expresa como:

$$\begin{bmatrix} (f_1,f_1)_d & (f_1,f_2)_d & ... & (f_1,f_m)_d \\ (f_2,f_1)_d & (f_2,f_2)_d & ... & (f_2,f_m)_d \\ ... & ... & ... & ... \\ (f_m,f_1)_d & (f_m,f_2)_d & ... & (f_m,f_m)_d \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ ... \\ c_m \end{bmatrix} = \begin{bmatrix} (f_1,y)_d \\ (f_2,y)_d \\ ... \\ (f_m,y)_d \end{bmatrix}$$

Siendo $(a,b)_d$ el producto escalar discreto, definido para dos funciones dadas h(x) y g(x) como:
$(h(x),g(x))_d = \sum_{k=1}^{n} h(x_k)g(x_k)$,
y para una función h(x) y vector cualquiera u, como:
$(h(x),u)_d = \sum_{k=1}^{n} h(x_k)u_k$

La resolución de dicho sistema permite obtener, para cualquier base de funciones derivables localmente, la función f(x) que sea mejor aproximación mínimo cuadrática al conjunto de puntos antes mencionado. La solución es óptima –esto es, proporciona la mejor aproximación siguiendo el criterio de mínimo error cuadrático–, puesto que se obtiene al optimizar el problema.

Corolario

Si se tratara de hallar el conjunto de coeficientes $\{c_j\}$ tal que $f(x)$ pase exactamente por todos los pares $\{(x_k,y_k)\}_{j=1}^{n}$, esto es, tales que $f(x)$ *interpole* a $\{(x_k,y_k)\}_{j=1}^{n}$, entonces tendría que cumplirse que:
$\sum_{j=1}^{m} c_j f_j(x_k) = y_k$
Que en forma matricial se expresa como:

$$\begin{bmatrix} f_1(x_1) & f_2(x_1) & ... & f_m(x_1) \\ f_1(x_2) & f_2(x_2) & ... & f_m(x_2) \\ ... & ... & ... & ... \\ f_1(x_n) & f_2(x_n) & ... & f_m(x_n) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ ... \\ c_m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} = A \cdot c = b$$

Esto establece un sistema de n ecuaciones y m incógnitas, y como en general n>m, quedaría sobredeterminado: no tendría siempre una solución general. Por tanto, la aproximación tratará en realidad de hallar el vector c que mejor aproxime $A \cdot c = b$.

Se puede demostrar que la matriz de coeficientes de las ecuaciones normales de Gauss coincide con $A^t \cdot A$, siendo A la matriz de coeficientes exactas, y como el término independiente de las ecuaciones normales de Gauss coincide con el vector $A^t \cdot b$, se tiene que los valores $\{c_j\}$ que mejor aproximan f(x) pueden calcularse como la solución al sistema:
$A^t \cdot A \cdot c = A^t \cdot b$
que es, precisamente, el sistema de las ecuaciones normales de Gauss.

7.3.2 Deducción geométrica de la aproximación discreta mínimo cuadrática lineal

La mejor aproximación deberá tender a interpolar la función de la que proviene el conjunto de pares (x_k, y_k) , esto es, deberá tender a pasar exactamente por todos los puntos. Eso supone que se debería cumplir que:

$$f(x_k) = y_k \quad \text{con} k = 1, 2, \ldots, n$$

Sustituyendo f(x) por su expresión como combinación lineal de una base de m funciones:

$$\sum_{j=1}^{m} c_j f_j(x_k) = y_k \quad \text{con} k = 1, \ldots, n$$

Esto es, se tendría que verificar exactamente un sistema de n ecuaciones y m incógnitas, pero como en general n>m, dicho sistema estaría sobredeterminado y, por tanto, sin solución general. De ahí surge la necesidad de aproximarlo.

Dicho sistema podría expresarse en forma matricial como:

$$\begin{bmatrix} f_1(x_1) & f_2(x_1) & ... & f_m(x_1) \\ f_1(x_2) & f_2(x_2) & ... & f_m(x_2) \\ ... & ... & ... & ... \\ f_1(x_n) & f_2(x_n) & ... & f_m(x_n) \end{bmatrix} \times \begin{bmatrix} c_1 \\ c_2 \\ ... \\ c_m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}$$

Esto es:

$$Ac = b$$

La aproximación trata de hallar el vector c aproximante que mejor aproxime el sistema $Ac = b$.

Con dicho vector c aproximante, es posible definir el vector residuo como:

$$r = b - Ac$$

De manera que el mínimo error cuadrático supone minimizar el residuo, definiendo su tamaño según la norma euclídea o usual del residuo, que equivale al error cuadrático:

$$||r||_2 = \sqrt{(r,r)_2} = \sqrt{r^t r} = \sqrt{\sum_{k=1}^{n} (r_k)^2}$$

siendo $(r, r)_2$ el producto interior o escalar del vector residuo sobre sí mismo.

Si atendemos al sistema $Ac = b$, entonces se ve claramente que al multiplicar A y c, lo que se realiza es una combinación lineal de las columnas de A:

$$Ac = \begin{bmatrix} A_1 & A_2 & ... & A_m \end{bmatrix} \times \begin{bmatrix} c_1 \\ c_2 \\ ... \\ c_m \end{bmatrix} = c_1 A_1 + c_2 A_2 + ... + c_m A_m$$

El problema de aproximación será hallar aquella combinación lineal de columnas de la matriz A lo más cercana posible al vector b. Se comprueba que el conjunto de las columnas de A generan un espacio vectorial o Span lineal: $span(A_1, A_2, ..., A_m)$, al que el vector b no tiene porqué pertenecer (si lo hiciera, el sistema A.c=b tendría solución).

Entonces, de los infinitos vectores del $span(A_1, A_2, ..., A_m)$ que son combinación lineal de los vectores de la base, se tratará de hallar el más cercano al vector b.

De entre todos ellos, el que cumple esto con respecto a la norma euclídea es la proyección ortogonal de b sobre $span(A_1, A_2, ..., A_m)$, y que por tanto hace que el tamaño del vector r, que será el vector que une los extremos de los vectores b y proyección ortogonal de b sobre el span, sea mínimo, esto es, que minimiza su norma euclídea.

Es inmediato ver que si el residuo une b con su proyección ortogonal, entonces es a su vez ortogonal al $span(A_1, A_2, ..., A_m)$, y a cada uno de los vectores de la base, esto es, ortogonal a cada columna de A.

La condición de minimización del residuo será:

$$r \perp span(A_1, A_2..., A_m)$$

Que es cierto si y sólo si:

$$r \perp A_j, \forall j \iff A_j \perp r, \forall j \iff (A_j, r)_2 = 0 = A_j^t r, \forall j = 1, 2, ..., m$$

A su vez, cada una de las m condiciones de perpendicularidad se pueden agrupar en una sola:

$$A^t r = 0$$

Sustituyendo el residuo por su expresión:

$$A^t(b - Ac) = 0 \iff A^t Ac = A^t b$$

Por tanto, la mejor aproximación mínimo cuadrada lineal para un conjunto de puntos discretos, sean cuales sean las funciones base, se obtiene al resolver el sistema cuadrado:

$$A^t Ac = A^t b$$

A esta ecuación se le llama *ecuación normal de Gauss*, y es válida para cualquier conjunto de funciones base. Si estas son la unidad y la función x, entonces la aproximación se llama regresión lineal.

7.4 Mínimos cuadrados y análisis de regresión

En el análisis de regresión, se sustituye la relación

$$f(x_i) \approx y_i$$

por

$$f(x_i) = y_i + \varepsilon_i,$$

siendo el término de perturbación ε una variable aleatoria con media cero. Obsérvese que estamos asumiendo que los valores *x* son exactos, y que todos los errores están en los valores *y*. De nuevo, distinguimos entre regresión lineal, en cuyo caso la función *f* es lineal para los parámetros a ser determinados (ej., $f(x) = ax^2 + bx + c$), y regresión no lineal. Como antes, la regresión lineal es mucho más sencilla que la no lineal. (Es tentador pensar que la razón del nombre *regresión lineal* es que la gráfica de la función $f(x) = ax + b$ es una línea. Ajustar una curva $f(x) = ax^2 + bx + c$, estimando *a*, *b* y *c* por mínimos cuadrados es un ejemplo de regresión *lineal* porque el vector de estimadores mínimos cuadráticos de *a*, *b* y *c* es una transformación lineal del vector cuyos componentes son $f(xi) + \varepsilon i$).

Los parámetros (*a*, *b* y *c* en el ejemplo anterior) se estiman con frecuencia mediante mínimos cuadrados: se toman aquellos valores que minimicen la suma *S*. El teorema de Gauss-Márkov establece que los estimadores mínimos cuadráticos son óptimos en el sentido de que son los estimadores lineales insesgados de menor varianza, y por tanto

de menor error cuadrático medio, si tomamos $f(x) = ax + b$ estando a y b por determinar y con los términos de perturbación ε independientes y distribuidos idénticamente (véase el artículo si desea una explicación más detallada y con condiciones menos restrictivas sobre los términos de perturbación).

La estimación de mínimos cuadrados para modelos lineales es notoria por su falta de robustez frente a valores atípicos (*outliers*). Si la distribución de los atípicos es asimétrica, los estimadores pueden estar sesgados. En presencia de cualquier valor atípico, los estimadores mínimos cuadráticos son ineficientes y pueden serlo en extremo. Si aparecen valores atípicos en los datos, son más apropiados los métodos de regresión robusta.

7.5 Referencias

- Abdi, H (2003). « (2003). Least-squares.». *M. Lewis-Beck, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage. pp. 792-795*.

7.6 Véase también

- Regresión isotónica
- Filtro de mínimos cuadrados promedio
- Estimación de mínimos cuadrados de coeficientes para regresión lineal
- Regresión lineal
- Mínimos cuadrados móviles
- Análisis de regresión
- Regresión robusta
- Valor eficaz
- Mínimos cuadrados totales
- Mínimos cuadrados ponderados
- Análisis de la varianza
- Ecuaciones normales del problema de cuadrados mínimos
- Algoritmo de Levenberg-Marquardt

7.7 Enlaces externos

En español:

- Regresión Lineal Simple
- Regresión Lineal y Cuadrática
- Regresión Polinomial
- Mínimos Cuadrados

En inglés:

- http://www.physics.csbsju.edu/stats/least_squares.html
- Zunzun.com - Ajuste de curvas y superficies en línea

- http://www.orbitals.com/self/least/least.htm
- Mínimos cuadrados en PlanetMath
- levmar implementación en C/C++ por cuadrados mínimos no lineales, GNU General Public License.
- SysLinea implementación en Pascal por cuadrados mínimos no lineales, GNU General Public License.
- lmfit otra implementación del algoritmo de Levenberg y Marquardt en C/C++, dominio público

Capítulo 8

Regularización de Tíjonov

La **Regularización de Tíjonov** es el método de regularización usado más comúnmente. En algunos campos, también se conoce como **regresión de arista**.

En su forma más simple, un sistema de ecuaciones lineales mal determinado:

$$A\mathbf{x} = \mathbf{b}\ ,$$

donde A es una matriz de dimensiones $m \times n$, x es un vector vertical con n celdas y b es otro vector vertical con m celdas, es reemplazado por el problema de encontrar un x que minimice

$$\|A\mathbf{x} - \mathbf{b}\|^2 + \alpha^2\|\mathbf{x}\|^2$$

dado un *factor de Tíjonov* $\alpha > 0$ elegido apropiadamente. La expresión $\|\cdot\|$ representa la norma euclídea. Su uso mejora el condicionamiento del problema, posibilitando su solución por métodos numéricos. Una solución explícita, denotada $\hat{x}$, es la siguiente:

$$\hat{x} = (A^T A + \alpha^2 I)^{-1} A^T \mathbf{b}$$

donde I es la matriz identidad $n \times n$. Para $\alpha = 0$, esto se reduce al método de mínimos cuadrados, siempre que $(\mathrm{A^T A})^{-1}$ exista.

8.1 Interpretación bayesiana

Aunque en principio la solución propuesta pueda parecer artificial, y de hecho el parámetro α tiene un carácter algo arbitrario, el proceso se puede justificar desde un punto de vista bayesiano. Nótese que para resolver cualquier problema indeterminado se deben introducir ciertas restricciones adicionales para establecer una solución estable. Estadísticamente se puede asumir que a priori sabemos que x es una variable aleatoria con una distribución normal multidimensional. Sin pérdida de generalidad, tomemos la media como 0 y asumamos que cada componente es independiente, con una desviación estándar σ_x . Los datos de b pueden tener ruido, que asumimos también independiente con media 0 y desviación estándar σ_b . Bajo estas condiciones, la regularización de Tíjonov es la solución más probable dados los datos conocidos y la distribución a priori de x , de acuerdo con el teorema de Bayes. Entonces, el parámetro de Tíjonov viene dado por $\alpha = \frac{\sigma_b}{\sigma_x}$ ...

8.2 Regularización de Tíjonov generalizada

Para distribuciones normales multivariadas de x y su error, se puede aplicar una transformación a las variables que reduce el problema al caso anterior. Equivalentemente, se puede minimizar

$$\|Ax - b\|_P^2 + \alpha^2\|x - x_0\|_Q^2$$

donde $\|x\|_P$ es la norma con peso $x^T Px$. En la interpretación bayesiana, P es la matriz de covarianza invertida b , x_0 es el valor esperado de x , y αQ es la matriz de covarianza invertida de x .

Esta expresión se puede resolver explícitamente mediante la fórmula

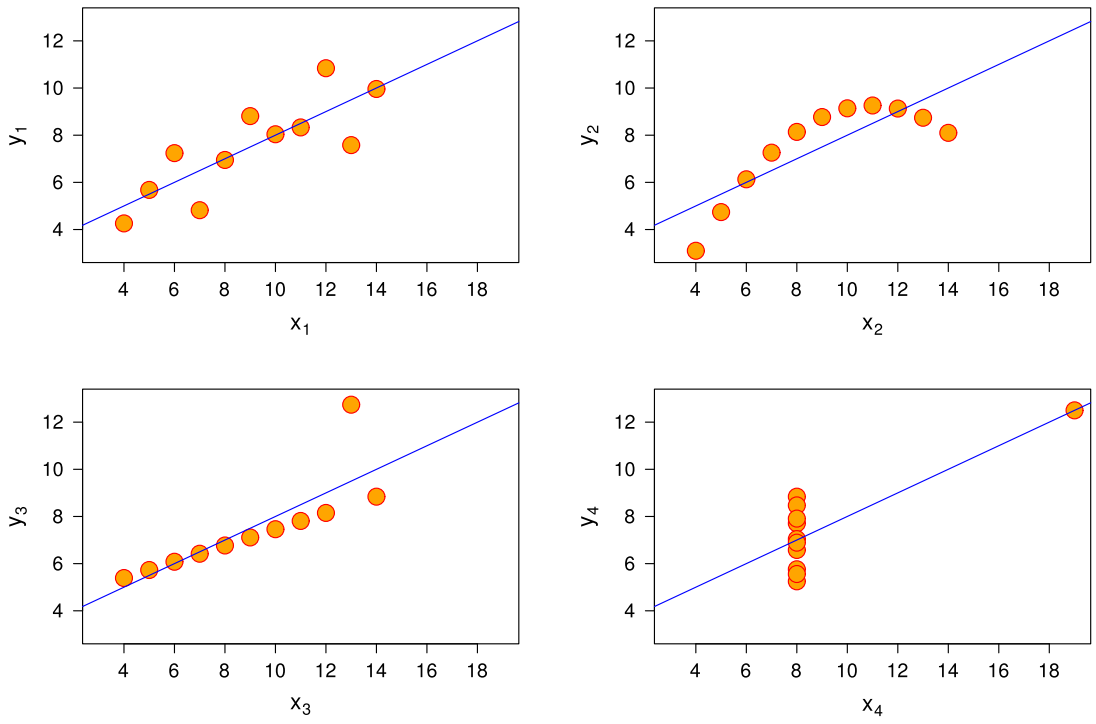
$$x_0 + (A^T PA + \alpha^2 Q)^{-1} A^T P(b - Ax_0).$$

8.3 Referencias

- Tikhonov AN, 1943, *On the stability of inverse problems*, Dokl. Akad. Nauk SSSR, 39, No. 5, 195-198
- Tikhonov AN, 1963, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math Dokl 4, 1035-1038 English translation of Dokl Akad Nauk SSSR 151, 1963, 501-504
- Tikhonov AN and Arsenin VA, 1977, *Solution of Ill-posed Problems*, Winston & Sons, Washington, ISBN 0-470-99124-0.
- Hansen, P.C., *Rank-deficient and Discrete ill-posed problems*, SIAM
- Hoerl AE, 1962, *Application of ridge analysis to regression problems*, Chemical Engineering Progress, 58, 54-59.
- Foster M, 1961, *An application of the Wiener-Kolmogorov smoothing theory to matrix inversion*, J. SIAM, 9, 387-392
- Phillips DL, 1962, *A technique for the numerical solution of certain integral equations of the first kind*, J Assoc Comput Mach, 9, 84-97
- Tarantola A, 2005, *Inverse Problem Theory* (free PDF version), Society for Industrial and Applied Mathematics, ISBN 0-89871-572-5
- Wahba, G, 1990, *spline Models for Observational Data*, SIAM

Capítulo 9

Cuarteto de Anscombe



El **cuarteto de Anscombe** comprende cuatro conjuntos de datos que tienen las mismas propiedades estadísticas, pero que evidentemente son distintas al inspeccionar sus gráficos respectivos.

Cada conjunto consiste de once puntos (x, y) y fueron construidos por el estadístico F. J. Anscombe. El cuarteto es una demostración de la importancia de mirar gráficamente un conjunto de datos antes de analizarlos.

Para los cuatro conjuntos de datos:

El primer gráfico (arriba a la izquierda) muestra lo que parece una relación lineal simple, correspondiente a dos variables correlacionadas cumpliendo con la suposición de normalidad. El segundo gráfico (arriba a la derecha) no está distribuido normalmente, aunque se observa relación entre los datos, esta no es lineal y el coeficiente de correlación de Pearson no es relevante. En la tercera gráfica (abajo a la izquierda) la distribución es lineal pero con una línea de regresión diferente de la que se sale el dato extremo que influye lo suficiente como para alterar la línea de regresión y disminuir el coeficiente de correlación de 1 a 0.816. Por último, la cuarta gráfica (abajo a la derecha) es un ejemplo de muestra en la que un valor atípico es suficiente para producir un coeficiente de correlación alto incluso cuando la relación entre las dos variables no es lineal.

Edward Tufte usó el cuarteto en la primera página del primer capítulo de su libro *The Visual Display of Quantitative*

Information, para enfatizar la importancia de *mirar* los datos antes de analizarlos.

9.1 Referencias externas

- F.J. Anscombe, “Graphs in Statistical Analysis,” American Statistician, 27 (febrero de 1973), 17-21.
- Departmento de Física, Universidad de Toronto
- Departmento de Computación, City University, Londres
- Ajuste de curvas, Central Queensland University, Australia

Capítulo 10

Modelo de valoración de activos financieros

El **modelo de valoración de activos financieros**, denominada en inglés **Capital asset pricing model** (**CAPM**) es un modelo introducido por Jack L. Treynor, William Sharpe, John Litner y Jan Mossin de forma independiente, basado en trabajos anteriores de Harry Markowitz sobre la diversificación y la Teoría Moderna de Portfolio. Sharpe, profesor de la Universidad de Stanford recibió el Premio Nobel de Economía (en conjunto con Harry Markowitz y Merton Miller, profesor de University of Chicago Booth School of Business) por su contribución al campo de la economía financiera.

10.1 Fórmula

El CAPM es un modelo para calcular el precio de un activo y pasivo o una cartera de inversiones. Para activos individuales, se hace uso de la recta *stock market line (SML)* (la cual simboliza el retorno esperado de todos los activos de un mercado como función del riesgo no diversificable) y su relación con el retorno esperado y el riesgo sistémico (beta), para mostrar cómo el mercado debe estimar el precio de un activo individual en relación a la clase a la que pertenece.

La línea SML permite calcular la proporción de recompensa-a-riesgo para cualquier activo en relación con el mercado general.

La relación de equilibrio que describe el CAPM es:

$$E(r_i) = r_f + \beta_{im}(E(r_m) - r_f)$$

donde:

- $E(r_i)$ es la tasa de rendimiento esperada de capital sobre el activo i .
- β_{im} es el *beta* (cantidad de riesgo con respecto al Portafolio de Mercado), o también

$$\beta_{im} = \frac{Cov(r_i, r_m)}{Var(r_m)}$$

- ($E(r_m) - r_f$) es el exceso de rentabilidad del portafolio de mercado.
- (r_m) Rendimiento del mercado.
- (r_f) Rendimiento de un activo libre de riesgo.

Es importante tener presente que se trata de un Beta no apalancado, es decir que se supone que una empresa no tiene deuda en su estructura de capital, por lo tanto no se incorpora el riesgo financiero, y en caso de querer incorporarlo, debemos determinar un Beta apalancado; por lo tanto el rendimiento esperado será más alto.

10.2 Precio de un activo

Una vez que el retorno esperado, $E(R_i)$, es calculado utilizando CAPM, los futuros flujos de caja que producirá ese activo pueden ser descontados a su valor actual neto utilizando esta tasa, para poder así determinar el precio adecuado del activo o título valor.

En teoría, un activo es apreciado correctamente cuando su precio observado es igual al valor calculado utilizando CAPM. Si el precio es mayor que la valuación obtenida, el activo está sobrevaluado, y viceversa.

10.3 Retorno requerido para un activo específico

CAPM calcula la tasa de retorno apropiada y requerida para descontar los flujos de efectivo futuros que producirá un activo, dada la apreciación de riesgo que tiene ese activo. Betas mayores a 1 simbolizan que el activo tiene un riesgo mayor al promedio de todo el mercado; betas debajo de 1 indican un riesgo menor. Por lo tanto, un activo con un beta alto debe ser descontado a una mayor tasa, como medio para recompensar al inversor por asumir el riesgo que el activo acarrea. Esto se basa en el principio que dice que los inversores, entre más riesgosa sea la inversión, requieren mayores retornos.

Puesto que el beta refleja la sensibilidad específica al riesgo no diversificable del mercado, el mercado, como un todo, tiene un beta de 1. Puesto que es imposible calcular el retorno esperado de todo el mercado, usualmente se utilizan índices, tales como el S&P 500 o el Dow Jones.

10.4 Riesgo y diversificación

El riesgo dentro de un portafolio incluye el riesgo sistemático, conocido también como riesgo no diversificable. Este riesgo se refiere al riesgo al que están expuestos todos los activos en un mercado. Por el contrario, el riesgo diversificable es aquel intrínseco a cada activo individual. El riesgo diversificable se puede disminuir agregando activos al portafolio que se mitiguen unos a otros, o sea diversificando el portafolio. Sin embargo, el riesgo sistemático no puede ser disminuido.

Por lo tanto, un inversor racional no debería tomar ningún riesgo que sea diversificable, pues solamente el riesgo no diversificable es recompensado en el alcance de este modelo. Por lo tanto, la tasa de retorno requerida para un determinado activo, debe estar vinculada con la contribución que hace ese activo al riesgo general de un determinado portafolio.

10.5 Suposiciones de CAPM

El modelo asume varios aspectos sobre inversores y mercados:

1. Los individuos son adversos al riesgo, y maximizan la utilidad de su riqueza en el próximo período. Es un modelo plurianual.

2. Los individuos no pueden afectar los precios, y tienen expectativas homogéneas respecto a las varianzas-covarianzas y acerca de los retornos esperados de los activos.

3. El retorno de los activos, se distribuye de manera normal. Explicando el retorno con la esperanza matemática y el riesgo con la desviación estándar.

4. Existe un activo libre de riesgo, al cual los individuos pueden prestar y/o endeudarse en cantidades ilimitadas. El mercado de activos es perfecto. La información es gratis y está disponible en forma instantánea para todos los individuos.

5. La oferta de activos es fija.

10.6 Inconvenientes de CAPM

- El modelo no explica adecuadamente la variación en los retornos de los títulos valores. Estudios empíricos muestran que activos con bajos betas pueden ofrecer retornos más altos de los que el modelo sugiere.
- El modelo asume que todos los inversores tienen acceso a la misma información, y se ponen de acuerdo sobre el riesgo y el retorno esperado para todos los activos.
- El portafolio del mercado consiste de todos los activos en todos los mercados, donde cada activo es ponderado por su capitalización de mercado. Esto asume que los inversores no tienen preferencias entre mercados y activos, y que escogen activos solamente en función de su perfil de riesgo-retorno.

10.7 Referencias

- Fama, Eugene F. (1968). *Risk, Return and Equilibrium: Some Clarifying Comments*. Journal of Finance Vol. 23, No. 1, pp. 29-40.
- Fama, Eugene F. and Kenneth French (1992). *The Cross-Section of Expected Stock Returns*. Journal of Finance, June 1992, 427-466.
- Fischer; Jensen & Scholes (1972). *The Capital Asset Pricing Model: Some Empirical Tests*, pp. 79-121 in M. Jensen ed., Studies in the Theory of Capital Markets. New York: Praeger Publishers.
- French, Craig W. (2003). *The Treynor Capital Asset Pricing Model*, Journal of Investment Management, Vol. 1, No. 2, pp. 60-72. Available at http://www.joim.com/
- French, Craig W. (2002). *Jack Treynor's 'Toward a Theory of Market Value of Risky Assets'* (December). Available at http://ssrn.com/abstract=628187
- Lintner, John (1965). *The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets*, Review of Economics and Statistics, 47 (1), 13-37.
- Markowitz, Harry M. (1999). *The early history of portfolio theory: 1600-1960*, Financial Analysts Journal, Vol. 55, No. 4
- Meggison, Smart, Gitman. "Corporate Finance", segunda edición. Thomson South-Western.
- Mehrling, Perry (2005). Fischer Black and the Revolutionary Idea of Finance. Hoboken: John Wiley & Sons, Inc.
- Mossin, Jan. (1966). *Equilibrium in a Capital Asset Market*, Econometrica, Vol. 34, No. 4, pp. 768-783.
- Mullins, David W. (1982). *Does the capital asset pricing model work?*, Harvard Business Review, January-February 1982, 105-113.
- Quiroga Díaz, Javier Fernando (????). Elaboración del Wiki.
- Ross, Stephen A. (1977). *The Capital Asset Pricing Model (CAPM), Short-sale Restrictions and Related Issues*, Journal of Finance, 32 (177)
- Rubinstein, Mark (2006). A History of the Theory of Investments. Hoboken: John Wiley & Sons, Inc.
- Sharpe, William F. (1964). *Capital asset prices: A theory of market equilibrium under conditions of risk*, Journal of Finance, 19 (3), 425-442
- Stone, Bernell K. (1970) Risk, Return, and Equilibrium: A General Single-Period Theory of Asset Selection and Capital-Market Equilibrium. Cambridge: MIT Press.
- Tobin, James (1958). *Liquidity preference as behavior towards risk*, The Review of Economic Studies, 25
- Treynor, Jack L. (1961). *Market Value, Time, and Risk*. Unpublished manuscript.
- Treynor, Jack L. (1962). *Toward a Theory of Market Value of Risky Assets*. Unpublished manuscript. A final version was published in 1999, in Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics. Robert A. Korajczyk (editor) London: Risk Books, pp. 15-22.

Capítulo 11

Análisis armónico

En matemáticas, el **análisis armónico** o **análisis de Fourier** estudia la representación de funciones o señales como superposición de ondas "básicas" o armónicos.

Investiga y generaliza las nociones de series de Fourier y transformadas de Fourier. A lo largo de los siglos XIX y XX se ha convertido en una materia enorme con aplicaciones en campos diversos como el procesamiento de señales, la mecánica cuántica o la neurociencia.

11.1 Serie de Fourier

Las *series de Fourier* se utilizan para descomponer una función, señal u onda periódica como suma infinita o finita de funciones, señales u ondas armónicas o sinusoidales; es decir, una serie de Fourier es un tipo de *serie trigonométrica*.

11.2 Transformada de Fourier

La transformada clásica de Fourier en $\mathbf{R}^n$ aún es un área de investigación activa, sobre todo en la transformación de Fourier sobre objetos más generales, como las distribuciones temperadas. Por ejemplo, si imponemos algunos requerimientos sobre una distribución f, podemos intentar trasladarlos a términos de su transformada de Fourier. El Teorema de Paley-Wiener es un ejemplo de ello, que implica inmediatamente que si f es una distribución de soporte compacto (lo que incluye a las funciones de soporte compacto), entonces su transformada de Fourier no tiene nunca el soporte compacto. Esto es un tipo muy elemental de un principio de incertidumbre en términos del análisis armónico.

Las series de Fourier pueden ser estudiadas convenientemente en el contexto de los espacios de Hilbert, lo que nos da una conexión entre el análisis armónico y el análisis funcional.

11.3 Análisis armónico abstracto

Una de las ramas más modernas del análisis armónico, que tiene sus raíces a mediados del siglo XX, es el análisis sobre grupos topológicos. El ideal central que lo motiva es la de las varias transformadas de Fourier, que pueden ser generalizadas a una transformación de funciones definidas sobre grupos localmente compactos.

La teoría para los grupos localmente compactos abelianos se llama dualidad de Pontryagin, que se considera una proposición muy satisfactoria ya que explica las características envueltas en el análisis armónico. En su página se encuentra desarrollada en detalle.

El análisis armónico estudia las propiedades de tal dualidad y la transformada de Fourier; y pretende extender tales características a otros marcos, por ejemplo en el del caso de los grupos de Lie no abelianos.

Para grupos generales no abelianos localmente compactos, el análisis armónico está muy relacionado con la teoría unitaria de representación de grupos unitarios. Para grupos compactos, el Teorema de Peter-Weyl explica cómo se pueden conseguir armónicos extrayendo una representación irreducible de cada clase de equivalencia de representaciones. Esta elección de armónicos goza de algunas de las propiedades útiles de la transformada de Fourier clásica

de forma que lleva convoluciones a productos escalares, o por otra parte mostrando cierta comprensión sobre la estructura de grupo subyacente.

11.4 Referencias

- Elias Stein and Guido Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, 1971. ISBN 0-691-08078-X
- Elias Stein with Timothy S. Murphy, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, 1993.
- Elias Stein, *Topics in Harmonic Analysis Related to the Littlewood-Paley Theory*, Princeton University Press, 1970.
- Yitzhak Katznelson, *An introduction to harmonic analysis*, Third edition. Cambridge University Press, 2004. ISBN 0-521-83829-0; 0-521-54359-2
- Yurii I. Lyubich. *Introduction to the Theory of Banach Representations of Groups*. Translated from the 1985 Russian-language edition (Kharkov, Ukraine). Birkhäuser Verlag. 1988.

Capítulo 12

Teorema de Gauss-Márkov

En estadística, el **Teorema de Gauss-Márkov**, formulado por Carl Friedrich Gauss y Andréi Márkov, establece que en un modelo lineal general (MLG) en el que se establezcan los siguientes supuestos:

- Correcta especificación: el MLG ha de ser una combinación lineal de los parámetros (β) y no necesariamente de las variables: $Y = X\beta + u$
- Muestreo aleatorio simple: la muestra de observaciones del vector $(y_i,\ x_{2i},\ x_{3i},\ \ldots,\ x_{ki})$ es una muestra aleatoria simple y, por lo tanto, el vector $(y_i,\ X_i')$ es independiente del vector $(y_i,\ X_j')$
- Esperanza condicionada de las perturbaciones nula: $E(u_i|X_i') = 0$
- Correcta identificación: la matriz de regresoras (X) ha de tener rango completo: rg(X)=K<=N
- Homocedasticidad: $Var(U|X) = \sigma^2 I$

el estimador mínimo cuadrático ordinario (MCO) de B es el estimador lineal e insesgado óptimo (ELIO o BLUE: best linear unbiased estimator), es decir, el estimador MCO es el estimador eficiente dentro de la clase de estimadores lineales e insesgados.

Dicho teorema se basa en 10 supuestos, denominados, Supuestos de Gauss Márkov; que sirven como hipótesis a la demostración del mismo:

1. El modelo esta correctamente especificado.
2. Debe ser lineal en los parámetros.
3. El valor de la media condicional es cero.
4. Hay homocedasticidad.
5. No existe correlación entre las perturbaciones.
6. La covarianza entre u_i y x_i es cero.
7. El número de observaciones es mayor que el de parámetros.
8. Existe variabilidad entre los x .
9. No hay multicolinealidad perfecta.
10. Las x son no estocásticas, es decir, son fijas en muestras repetidas.

12.1 Enlaces externos

- Demostración del teorema

Capítulo 13

Análisis de la regresión

En estadística, el análisis de la regresión es un proceso estadístico para la estimación de relaciones entre variables. Incluye muchas técnicas para el modelado y análisis de diversas variables, cuando la atención se centra en la relación entre una variable dependiente y una o más variables independientes. Más específicamente, el análisis de regresión ayuda a entender cómo el valor típico de la variable dependiente cambia cuando cualquiera de las variables independientes es variada, mientras que se mantienen las otras variables independientes fijas. Más comúnmente, el análisis de regresión estima la esperanza condicional de la variable dependiente dadas las variables independientes - es decir, el valor promedio de la variable dependiente cuando se fijan las variables independientes. Con menor frecuencia, la atención se centra en un cuantil, u otro parámetro de localización de la distribución condicional de la variable dependiente dadas las variables independientes. En todos los casos, el objetivo es la estimación de una función de las variables independientes llamada la **función de regresión**. En el análisis de regresión, también es de interés para caracterizar la variación de la variable dependiente en torno a la función de regresión que puede ser descrito por una distribución de probabilidad.

El análisis de regresión es ampliamente utilizado para la predicción y previsión, donde su uso tiene superposición sustancial en el campo de aprendizaje automático. El análisis de regresión se utiliza también para comprender que cuales de las variables independientes están relacionadas con la variable dependiente, y explorar las formas de estas relaciones. En circunstancias limitadas, el análisis de regresión puede utilizarse para inferir relaciones causales entre las variables independientes y dependientes. Sin embargo, esto puede llevar a ilusiones o falsas relaciones, por lo que se recomienda precaución,[1] por ejemplo, la correlación no implica causalidad.

Se han desarrollado muchas técnicas para llevar a cabo análisis de regresión. Métodos familiares tales como regresión lineal y ordinaria de mínimos cuadrados de regresión son paramétrica, en que la función de regresión se define en términos de un número finito de desconocidos parámetros que se estiman a partir de los datos. regresión no paramétrica se refiere a las técnicas que permiten que la función de regresión mienta en un conjunto específico de funciones, que puede ser de dimensión infinita.

El desempeño de los métodos de análisis de regresión en la práctica depende de la forma del proceso de generación de datos, y cómo se relaciona con el método de regresión que se utiliza. Dado que la forma verdadera del proceso de generación de datos generalmente no se conoce, el análisis de regresión depende a menudo hasta cierto punto de hacer suposiciones acerca de este proceso. Estos supuestos son a veces comprobable si una cantidad suficiente de datos está disponible. Los modelos de regresión para la predicciamente, aunque pueden no funcionar de manera óptima. Sin embargo, en muchas aplicaciones, sobre todo con pequeños efectos o las cuestiones de causalidad sobre la base de los datos de observación, métodos de regresión pueden dar resultados engañosos.[2][3]

13.1 Historia

La primera forma de regresión fue el método de mínimos cuadrados, que fue publicado por Legendre en 1805,[4] y por Gauss en 1809.[5] Legendre y Gauss tanto aplicaron el método para el problema de determinar, a partir de observaciones astronómicas, la órbitas de los cuerpos sobre el Sol (la mayoría de los cometas, sino también más tarde el entonces recién descubiertos planetas menores). Gauss publicó un desarrollo posterior de la teoría de los mínimos cuadrados en 1821,[6] incluyendo una versión del teorema de Gauss-Markov.

El término “regresión” fue acuñado por Francis Galton en el siglo XIX para describir un fenómeno biológico. El fe-

nómeno fue que las alturas de los descendientes de ancestros altos tienden a regresar hacia abajo, hacia un promedio normal (un fenómeno conocido como regresión hacia la media).[7][8] Para Galton, la regresión sólo tenía este significado biológico,,[9][10] pero su trabajo se extendió más tarde por Udny Yule y Karl Pearson a un contexto estadístico más general.[11][12] En la obra de Yule y Pearson, la distribución conjunta de la respuesta y las variables explicativas se supone que es de Gauss. Esta suposición se vio debilitada por RA Fisher en sus obras de 1922 y 1925.[13][14][15] Fisher supone que la distribución condicional de la variable de respuesta es de Gauss, pero la distribución conjunta no es necesario. A este respecto, la asunción de Fisher está más cerca de la formulación de Gauss de 1821.

En los años 1950 y 1960, los economistas utilizan calculadoras electromecánicas para calcular regresiones. Antes de 1970, a veces tardaba hasta 24 horas para recibir el resultado de una regresión.[16]

Los métodos de regresión siguen siendo un área de investigación activa. En las últimas décadas, los nuevos métodos han sido desarrollados para la regresión robusta, la regresión que implica respuestas correlacionadas, tales como series de tiempo y las curvas de crecimiento, regresión en la que los predictores o variables de respuesta son curvas, imágenes, gráficos y otros objetos de datos complejos, los métodos de regresión Aceptar varios tipos de datos faltantes, la regresión no paramétrica, bayesianos métodos de regresión, regresión en el que las variables de predicción se miden con error, regresión con más variables predictoras que las observaciones y la inferencia causal con la regresión.

13.2 Modelos de regresión

Regresión lineal

- **Regresión lineal simple**

Dadas dos variables (Y: variable dependiente; X: independiente) se trata de encontrar una función simple (lineal) de X que nos permita aproximar Y mediante: $\hat{Y} = a + bX$

a (ordenada en el origen, constante)

b (pendiente de la recta)

A la cantidad $e=Y-\hat{Y}$ se le denomina residuo o error residual.

Así, en el ejemplo de Pearson: $\hat{Y} = 85\text{ cm} + 0{,}5X$

Donde $\hat{Y}$ es la altura predicha del hijo y X la altura del padre: En media, el hijo gana 0,5 cm por cada cm del padre.

- **Regresión lineal múltiple**

13.2.1 Regresión no lineal

- Regresión segmentada

13.3 Véase también

- Contraste de hipótesis

13.4 Referencias

[1] Armstrong, J. Scott (2012). «Illusions in Regression Analysis». *International Journal of Forecasting (forthcoming)* **28** (3): 689. doi:10.1016/j.ijforecast.2012.02.001.

[2] David A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press (2005)

[3] R. Dennis Cook; Sanford Weisberg Criticism and Influence Analysis in Regression, *Sociological Methodology*, Vol. 13. (1982), pp. 313–361

[4] A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*, Firmin Didot, Paris, 1805. "Sur la Méthode des moindres quarrés" appears as an appendix.

[5] C.F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*. (1809)

[6] C.F. Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae*. (1821/1823)

[7] Mogull, Robert G. (2004). *Second-Semester Applied Statistics*. Kendall/Hunt Publishing Company. p. 59. ISBN 0-7575-1181-3.

[8] Galton, Francis (1989). «Kinship and Correlation (reprinted 1989)». *Statistical Science* (Institute of Mathematical Statistics) **4** (2): 80–86. doi:10.1214/ss/1177012581. JSTOR 2245330.

[9] Francis Galton. "Typical laws of heredity", Nature 15 (1877), 492–495, 512–514, 532–533. *(Galton uses the term "reversion" in this paper, which discusses the size of peas.)*

[10] Francis Galton. Presidential address, Section H, Anthropology. (1885) *(Galton uses the term "regression" in this paper, which discusses the height of humans.)*

[11] Yule, G. Udny (1897). «On the Theory of Correlation». *Journal of the Royal Statistical Society* (Blackwell Publishing) **60** (4): 812–54. doi:10.2307/2979746. JSTOR 2979746.

[12] Pearson, Karl; Yule, G.U.; Blanchard, Norman; Lee,Alice (1903). «The Law of Ancestral Heredity». *Biometrika* (Biometrika Trust) **2** (2): 211–236. doi:10.1093/biomet/2.2.211. JSTOR 2331683.

[13] Fisher, R.A. (1922). «The goodness of fit of regression formulae, and the distribution of regression coefficients». *Journal of the Royal Statistical Society* (Blackwell Publishing) **85** (4): 597–612. doi:10.2307/2341124. JSTOR 2341124.

[14] Ronald A. Fisher (1954). *Statistical Methods for Research Workers* (Twelfth edición). Edinburgh: Oliver and Boyd. ISBN 0-05-002170-2.

[15] Aldrich, John (2005). «Fisher and Regression». *Statistical Science* **20** (4): 401–417. doi:10.1214/088342305000000331. JSTOR 20061201.

[16] Rodney Ramcharan. Regressions: Why Are Economists Obessessed with Them? March 2006. Accessed 2011-12-03.

13.5 Enlaces externos

- Francis Galton. "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute*, 15:246-263 (1886).
- A non-mathematical explanation of regression toward the mean.
- A simulation of regression toward the mean.
- Amanda Wachsmuth, Leland Wilkinson, Gerard E. Dallal. Galton's Bend: An Undiscovered Nonlinearity in Galton's Family Stature Regression Data and a Likely Explanation Based on Pearson and Lee's Stature Data

Capítulo 14

Regresión robusta

En Estadística robusta, una **regresión robusta** es una forma de análisis de la regresión diseñada para eludir algunas limitaciones tradicionales de los métodos paramétricos y no paramétricos. El análisis de regresión busca encontrar la relación entre una o más variables independientes y una variable dependiente. Algunos métodos utilizados de regresión, como mínimos cuadrados ordinarios, tienen propiedades favorables si sus suposiciones subyacentes se cumplen para los datos estudiados, pero pueden dar resultados engañosos si esas suposiciones no son ciertas; se dice que mínimos cuadrados ordinarios no es robusto a violaciones de los supuestos. Los métodos de regresión robusta están diseñados para no ser excesivamente afectados por violaciones de los supuestos por el proceso de generación de datos subyacente.

En particular, las estimaciones con los mínimos cuadrados son altamente no robustos a los valores atípicos. Si bien no existe una definición exacta de un valor atípico o de una observación atípica, los valores atípicos son observaciones que no siguen el patrón de las otras observaciones. Esto no es normalmente un problema si el valor atípico es simplemente una observación extrema extraída de la cola de una distribución normal, pero si los resultados atípicos de error de medición no normal o alguna otra violación de supuestos ordinarios de mínimos cuadrados estándar, entonces se compromete la validez de los resultados de la regresión si se utiliza una técnica de regresión no-robusta.

14.1 Aplicaciones

14.1.1 Errores heteroscedásticos

Un caso en el que la estimación robusta se debe considerar es cuando hay una fuerte sospecha de heterocedasticidad. En el modelo homoscedástico, se asume que la varianza del término de error es constante para todos los valores de x. Heteroscedasticidad permite la variación que dependerá de x, que es más preciso para muchos escenarios reales. Por ejemplo, la variación del gasto suele ser mayor para las personas con ingresos más altos que para las personas con ingresos más bajos. Los paquetes de software normalmente por defecto a una modelo homoscedástica, a pesar de que este modelo puede ser menos precisa que un modelo heteroscedastic. Un enfoque simple (Tofallis, 2008) es la aplicación de mínimos cuadrados a los errores porcentuales ya que esto reduce la influencia de los valores más grandes de la variable dependiente en comparación con los mínimos cuadrados ordinarios.

14.1.2 La presencia de valores atípicos

Otra situación común en la que se utiliza estimación robusta se produce cuando los datos contienen valores atípicos. En presencia de valores atípicos que no provienen de un mismo proceso de generación de datos que el resto de los datos, la estimación por mínimos cuadrados es ineficaz y puede estar sesgada. Debido a que las predicciones con mínimos cuadrados son arrastradas hacia los valores atípicos, y debido a que la varianza de las estimaciones se inflan artificialmente, el resultado es que los valores atípicos se pueden enmascarar. (En muchas situaciones, como algunas zonas de la geoestadística y estadísticas médicas, son precisamente los valores atípicos los que son de interés.)

Aunque a veces se afirma que los mínimos cuadrados (o métodos estadísticos clásicos en general) son robustos, solo son robustos en el sentido de que el tipo I tasa de error no aumenta bajo violaciones del modelo. De hecho, el tipo I tasa de error tiende a ser más bajo que el nivel nominal cuando los valores atípicos están presentes, y con frecuencia

hay un dramático incremento en la tasa de error de tipo II. La reducción de la tasa de error de tipo I ha sido etiquetado como el conservadurismo de los métodos clásicos. Otras etiquetas pueden incluir la ineficacia o inadmisibilidad.

14.2 Historia e impopularidad de la regresión robusta

A pesar de su rendimiento superior sobre la estimación de mínimos cuadrados, en muchos casos, aún no se utilizan ampliamente métodos robustos para la regresión. Hay varias razones que pueden ayudar a explicar su impopularidad (Hampel et al. 1986, 2005). Una posible razón es que hay varios métodos que compiten y el campo empezó con muchas salidas en falso. Además, el cálculo de las estimaciones robustas es mucho más intensiva computacionalmente que la estimación por mínimos cuadrados. Sin embargo, en los últimos años esta objeción se ha vuelto menos relevante dado que la potencia de cálculo ha aumentado considerablemente. Otra razón de la poca utilización de la regresión robusta puede ser que algunos paquetes populares de software estadístico no aplicaron los métodos (Stromberg, 2004). La creencia de muchos estadísticos de que los métodos clásicos son robustos puede ser otra razón.

Aunque la adopción de métodos robustos han sido lenta, las materias de estadística convencionales y los libros de texto modernos a menudo incluyen la discusión de estos métodos (por ejemplo, los libros de Seber y Lee, y Faraway). Además, los paquetes de software estadísticos modernos, como R, Stata y S-PLUS incluyen una funcionalidad considerable para la estimación robusta (véase, por ejemplo, los libros de Venables y Ripley, y por Maronna et al.).

14.3 Los métodos de regresión robusta

14.3.1 Alternativas a los mínimos cuadrados

Los métodos más simples de estimación de parámetros en un modelo de regresión que son menos sensibles a los valores atípicos que las estimaciones de mínimos cuadrados, es el uso de Mínimas desviaciones absolutas. Incluso entonces, los valores extremos graves aún puede tener un impacto considerable en el modelo, motivando la investigación sobre enfoques aún más robustos.

En 1973, Peter J. Huber presentó los modelos de regresión M-estimación. La M enlas siglas de M-estimación son por "Tipo de máxima verosimilitud". El método es robusto a los valores atípicos en la variable de respuesta, pero resultó no ser resistente a los valores atípicos en las variables explicativas (puntos de influencia). De hecho, cuando hay valores extremos en las variables explicativas, el método no tiene ninguna ventaja sobre los mínimos cuadrados.

En la década de 1980, se propusieron varias alternativas al M-estimación como intentos de superar la falta de resistencia. Mínimos cuadrados recortados (LTS) es una alternativa viable y es actualmente (2007) en la opción preferida de Rousseeuw y Ryan (1997, 2008). El Theil-Sen estimador tiene un punto de ruptura inferior LTS pero es estadísticamente eficiente y popular. Otra solución propuesta fue S-estimación. Este método encuentra una línea (plano o hiperplano) que minimiza una estimación robusta de la escala (de la que el método obtiene el S en su nombre) de los residuos. Este método es altamente resistente a los puntos de influencia, y es robusto a los valores atípicos en la respuesta. Sin embargo, se encontró también que este método es ineficaz.

14.3.2 Alternativas paramétricas

Otro enfoque para la estimación robusta de modelos de regresión es reemplazar la distribución normal con una distribución de cola pesada. Una distribución t con entre 4 y 6 grados de libertad se considera que es una buena elección en diferentes situaciones prácticas. La regresión bayesiana robusta, siendo totalmente paramétrica se basa en gran medida de estas distribuciones.

Bajo el supuesto de residuos t-distribuidos, la distribución es una localización escala. Es decir, $x \leftarrow (x - \mu)/\sigma$. Los grados de libertad de la distribución t son a veces llamados el parámetro de curtosis. Lange, Little y Taylor (1989) discuten este modelo en cierta profundidad desde un punto de vista no Bayesiano.[1] Una estudio que toma en cuenta lo bayesiano aparece en Gelman et al. (2003).[2]

Un enfoque paramétrico alternativa es suponer que los residuos siguen una mezcla de distribuciones normales, en particular, una distribución normal contaminada en la que la mayoría de las observaciones son de una distribución normal especificada, pero una pequeña proporción son de una distribución normal con mucho mayor varianza. Eso es,

los residuos tienen probabilidad $1-\varepsilon$ de venir de una distribución normal con varianza σ^2 , En donde ε es pequeño, y la probabilidad ε de venir de una distribución normal con varianza $c\sigma^2$ para algunos $c > 1$

$$e_i \sim (1-\varepsilon)N(0,\sigma^2) + \varepsilon N(0, c\sigma^2).$$

Típicamente, $\varepsilon < 0.1$. Esto a veces se llama el ε Modelo de la contaminación.

Enfoques paramétricos tienen la ventaja de que la teoría de probabilidad proporciona un 'fuera de la plataforma' enfoque a la inferencia (aunque para los modelos de mezcla tales como la ε -Contaminación modelo, no pudo aplicarse las condiciones usuales de regularidad), y que es posible construir modelos de simulación a partir del ajuste. Sin embargo, estos modelos paramétricos todavía asumen que el modelo subyacente es literalmente cierto. Como tales, no tienen en cuenta las distribuciones residuales sesgadas o precisiones observación finitos.

14.4 Referencias

[1] Lange, K. L.; R. J. A. Little and J. M. G. Taylor (1989). «Robust statistical modeling using the *t*-distribution». *Journal of the American Statistical Association* **84** (408): 881–896. doi:10.2307/2290063. JSTOR 2290063.

[2] Gelman, A.; J. B. Carlin, H. S. Stern and D. B. Rubin (2003). Bayesian Data Analysis (Second ed.). Chapman & Hall/CRC.

14.5 Bibliografía adicional

- Andersen, R. (2008). *Modern Methods for Robust Regression*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-152.
- Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.
- Breiman, L. (2001). «Statistical Modeling: the Two Cultures». *Statistical Science* **16** (3): 199–231. doi:10.1214/ss/1009213725. JSTOR 2676681.
- Faraway, J. J. (2004). *Linear Models with R*. Chapman & Hall/CRC.
- Draper, David (1988). «Rank-Based Robust Analysis of Linear Models. I. Exposition and Review». *Statistical Science* **3** (2): 239–257. doi:10.1214/ss/1177012915. JSTOR 2245578.
- McKean, Joseph W. (2004). «Robust Analysis of Linear Models». *Statistical Science* **19** (4): 562–570. doi:10.1214/088342304000000 JSTOR 4144426.
- Gelman, A.; J. B. Carlin, H. S. Stern and D. B. Rubin (2003). *Bayesian Data Analysis (Second Edition)*. Chapman & Hall/CRC.
- Hampel, F. R.; E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel (1986, 2005). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.
- Lange, K. L.; R. J. A. Little and J. M. G. Taylor (1989). «Robust statistical modeling using the *t*-distribution». *Journal of the American Statistical Association* **84** (408): 881–896. doi:10.2307/2290063. JSTOR 2290063.
- Maronna, R.; D. Martin and V. Yohai (2006). *Robust Statistics: Theory and Methods*. Wiley.
- Radchenko S.G. (2005). *Robust methods for statistical models estimation: Monograph. (on russian language)*. Kiev: PP «Sanspariel» ISBN 966-96574-0-7. p. 504.
- Rousseeuw, P. J.; A. M. Leroy (1986, 2003). *Robust Regression and Outlier Detection*. Wiley.
- Ryan, T. P. (1997, 2008). *Modern Regression Methods*. Wiley.
- Seber, G. A. F.; A. J. Lee (2003). *Linear Regression Analysis (Second Edition)*. Wiley.

- Stromberg, A. J. (2004). «Why write statistical software? The case of robust statistical methods». *Journal of Statistical Software*.
- Strutz, Tilo (2010). *Data Fitting and Uncertainty - A practical introduction to weighted least squares and beyond*. Vieweg+Teubner. ISBN 978-3-8348-1022-9.
- Tofallis, Chris (2008). «Least Squares Percentage Regression». *Journal of Modern Applied Statistical Methods* **7**: 526–534.
- Venables, W. N.; B. D. Ripley (2002). *Modern Applied Statistics with S*. Springer.

Capítulo 15

Valor eficaz

En electricidad y electrónica, en corriente alterna, el valor cuadrático medio (en inglés *root mean square*, abreviado RMS o rms), de una corriente variable es denominado **valor eficaz**. Se define como el valor de una corriente rigurosamente constante (corriente continua) que al circular por una determinada resistencia óhmica pura produce los mismos efectos caloríficos (*igual potencia disipada*) que dicha corriente variable (corriente alterna). De esa forma una corriente eficaz es capaz de producir el mismo trabajo que su valor en corriente directa o continua. Como se podrá observar derivado de las ecuaciones siguientes, el valor eficaz es independiente de la frecuencia o periodo de la señal.

Al ser la intensidad de esta corriente variable una función continua **i(t)** se puede calcular:

$$I_{ef} = \sqrt{\frac{1}{T}\int_{t_0}^{t_0+T} i^2(t)\,dt}$$

donde:

T

Esta expresión es válida para cualquier forma de onda, sea ésta sinusoidal o no, siendo por tanto aplicable a señales de radiofrecuencia y de audio o vídeo.

En el caso de una corriente alterna sinusoidal (como lo es, con bastante aproximación, la de la red eléctrica) con una amplitud máxima o de pico $\mathbf{I_{max}}$, el valor eficaz $\mathbf{I_{ef}}$ es:

$$I_{ef} = \frac{I_{max}}{\sqrt{2}}$$

En el caso de una señal triangular con una amplitud máxima $\mathbf{I_{max}}$, el valor eficaz $\mathbf{I_{ef}}$ es:

$$I_{ef} = \frac{I_{max}}{\sqrt{3}}$$

Para una señal cuadrada es:

$$I_{ef} = I_{max}$$

Para el cálculo de potencias eficaces $\mathbf{P_{ef}}$ por ser proporcional con el cuadrado de la amplitud de la tensión eléctrica, para el caso de señales sinusoidales se tiene:

$$P_{ef} = \frac{P_{max}}{2}$$

Del mismo modo para señales triangulares:

$$P_{ef} = \frac{P_{max}}{3}$$

Es común el uso del valor eficaz para voltajes también y su definición es equivalente:

$$V_{ef} = \sqrt{\frac{1}{T}\int_{t_0}^{t_0+T} v^2(t)\,dt}$$

Valor eficaz de una señal de corriente o voltaje con offset

En ocasiones una señal de corriente o voltaje posee un componente de continua, que se le suele llamar *offset*, que implica un desplazamiento hacia arriba o hacia abajo de la forma

$$f(t) + a$$

donde a puede ser positivo o negativo, positivo si se desplaza hacia arriba y negativo si se desplaza hacia abajo.

Su valor efectivo en caso de ser senoidal será:

$$V_{ef} = \sqrt{\frac{V_{max}^2}{2} + a^2}$$

en caso de ser triangular:

$$V_{ef} = \sqrt{\frac{V_{max}^2}{3} + a^2}$$

en caso de ser cuadrada:

$$V_{ef} = \sqrt{V_{max}^2 + a^2}$$

Capítulo 16

Análisis de la varianza

En estadística, el **análisis de la varianza** (ANOVA, **AN**alysis **O**f **VA**riance, según terminología inglesa) es una colección de modelos estadísticos y sus procedimientos asociados, en el cual la varianza está particionada en ciertos componentes debidos a diferentes variables explicativas.

Las técnicas iniciales del análisis de varianza fueron desarrolladas por el estadístico y genetista R. A. Fisher en los años 1920 y 1930 y es algunas veces conocido como "Anova de Fisher" o "análisis de varianza de Fisher", debido al uso de la distribución F de Fisher como parte del contraste de hipótesis.

16.1 Introducción

El análisis de la varianza parte de los conceptos de regresión lineal. Un análisis de la varianza permite determinar si diferentes tratamientos muestran diferencias significativas o por el contrario puede suponerse que sus medias poblacionales no difieren. El análisis de la varianza permite superar las limitaciones de hacer contrastes bilaterales por parejas (que son un mal método para determinar si un conjunto de variables con $n > 2$ difieren entre sí. El primer concepto fundamental es que todo valor observado puede expresarse mediante la siguiente función:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

Donde *yij* sería el valor observado (variable dependiente), y τ_i es el efecto del tratamiento *i*.

μ

τ_i

ϵ_{ij}

Por tanto, a la función de pronóstico la podemos llamar "media del tratamiento i":

$$y_i = \mu + \tau_i$$

Podemos resumir que las puntuaciones observadas equivalen a las puntuaciones esperadas, más el error aleatorio ($y_{ij} = y_i + e_{ij}$). A partir de esa idea, se puede operar:

1. Restamos a ambos lados de la ecuación (para mantener la igualdad) la media de la variable dependiente:

$$y_{ij} - \overline{y} = y_i + e_{ij} - \overline{y}$$

1. Operando se llega finalmente a que:

$$\sum_i \sum_j (y_{ij} - \overline{y})^2 = n \sum_i (y_i - \overline{y}_i)^2 + \sum_i \sum_j (y_{ij} - \overline{y}_i)^2$$

Esta eccuación se reescribe frecuentemente como:

$$SS_{total} = SS_{fact} + SS_{error}$$

de un factor, que es el caso más sencillo, la idea básica del análisis de la varianza es comparar la variación total de un conjunto de muestras y descomponerla como:

$$SS_{total} = SS_{fact} + SS_{int}$$

Donde:

SS_{fact}

SS_{int}

En el caso de que la diferencia debida al factor o tratamiento no sean estadísticamente significativa puede probarse que las varianzas muestrales son iguales:

$$\hat{s}_{fact} = \frac{SS_{fact}}{a-1}, \qquad \hat{s}_{int} = \frac{SS_{int}}{a(b-1)}$$

Donde:

a

b

Así lo que un simple test a partir de la F de Snedecor puede decidir si el factor o tratamiento es estadísticamente significativo.

16.1.1 Visión general

Existen tres clases conceptuales de estos modelos:

1. El Modelo de efectos fijos asume que los datos provienen de poblaciones normales las cuales podrían diferir únicamente en sus medias. (Modelo 1)
2. El Modelo de efectos aleatorios asume que los datos describen una jerarquía de diferentes poblaciones cuyas diferencias quedan restringidas por la jerarquía. Ejemplo: El experimentador ha aprendido y ha considerado en el experimento sólo tres de muchos más métodos posibles, el método de enseñanza es un factor aleatorio en el experimento. (Modelo 2)
3. El Modelo de efectos mixtos describen situaciones que éste puede tomar. Ejemplo: Si el método de enseñanza es analizado como un factor que puede influir donde están presentes ambos tipos de factores: fijos y aleatorios. (Modelo 3)

16.1.2 Supuestos previos

El ANOVA parte de algunos supuestos o hipótesis que han de cumplirse:

- La variable dependiente debe medirse al menos a nivel de intervalo.
- Independencia de las observaciones.
- La distribución de los residuales debe ser normal.
- Homocedasticidad: homogeneidad de las varianzas.

La técnica fundamental consiste en la separación de la suma de cuadrados (SS, 'sum of squares') en componentes relativos a los factores contemplados en el modelo. Como ejemplo, mostramos el modelo para un ANOVA simplificado con un tipo de factores en diferentes niveles. (Si los niveles son cuantitativos y los efectos son lineales, puede resultar apropiado un análisis de regresión lineal)

$$SS_{\text{Total}} = SS_{\text{Error}} + SS_{\text{Factores}}$$

El número de grados de libertad (gl) puede separarse de forma similar y corresponde con la forma en que la distribución chi-cuadrado (χ^2 o Ji-cuadrada) describe la suma de cuadrados asociada.

$$gl_{\text{Total}} = gl_{\text{Error}} + gl_{\text{Factores}}$$

16.2 Tipos de modelo

16.2.1 Modelo I: Efectos fijos

El modelo de *efectos fijos* de análisis de la varianza se aplica a situaciones en las que el experimentador ha sometido al grupo o material analizado a varios factores, cada uno de los cuales le afecta sólo a la media, permaneciendo la "variable respuesta" con una distribución normal.

Este modelo se supone cuando el investigador se interesa únicamente por los niveles del factor presentes en el experimento, por lo que cualquier variación observada en las puntuaciones se deberá al error experimental.

16.2.2 Modelo II: Efectos aleatorios (componentes de varianza)

Los modelos de *efectos aleatorios* se usan para describir situaciones en que ocurren diferencias incomparables en el material o grupo experimental. El ejemplo más simple es el de estimar la media desconocida de una población compuesta de individuos diferentes y en el que esas diferencias se mezclan con los errores del instrumento de medición.

Este modelo se supone cuando el investigador está interesado en una población de niveles, teóricamente infinitos, del factor de estudio, de los que únicamente una muestra al azar (*t* niveles) están presentes en el experimento.

16.3 Grados de libertad

Los grados de libertad pueden descomponerse al igual que la suma de cuadrados. Así, GLtotal = GLentre + GLdentro. Los GLentre se calculan como: a - 1, donde a es el número de tratamientos o niveles del factor. Los GLdentro se calculan como N - a, donde N es el número total de observaciones o valores de la variable medida (la variable respuesta).

16.4 Pruebas de significación

El análisis de varianza lleva a la realización de pruebas de significación estadística, usando la denominada distribución F de Snedecor.

16.5 Tablas ANOVA

Una vez que se han calculado las sumas de cuadrados, las medias cuadráticas, los grados de libertad y la F, se procede a elaborar una tabla que reuna la información, denominada "Tabla de Análisis de varianza o ANOVA", que adopta la siguiente forma:

$= +$

16.6 Bibliografía

- M.R. Spiegel; J. Schiller; R. A. Srinivasan (2007). «9. Análisis de la varianza». *Probabilidad y Estadística* [*Schaum's Outline of Theory and Problems of Probability and Statistics*]. Schaum (2ª edición). México D.F.: McGraw-Hill. pp. 335–371. ISBN 978-970-10-4231-1.

- F. J. Tejedor Tejedor (1999). *Análisis de varianza*. Schaum. Madrid: La Muralla S.A. ISBN 84-7635-388-X.

16.7 Enlaces externos

- Del análisis de varianza clásico al ANOVA robusto.

16.8 Texto e imágenes de origen, colaboradores y licencias

16.8.1 Texto

- **Regresión lineal** *Fuente:* https://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal?oldid=82160733 *Colaboradores:* Joseaperez, JorgeGG, Riviera, Elwikipedista, Tano4595, Felipealvarez, Magister Mathematicae, Alhen, BOT-Superzerocool, Vitamine, Gaeddal, GermanX, Banfield, BOTpolicia, CEM-bot, Daniel De Leon Martinez, Laura Fiorucci, Marianov, Roberpl, Davius, Antur, Gafotas, Ggenellina, Ingenioso Hidalgo, Thijs!bot, Alvaro qc, Xabier, Diego D E, Yeza, Gusgus, JAnDbot, Kved, Rjgalindo, TXiKiBoT, Juan renombrado, Hlnodovic, Dhcp, Marvelshine, Alefisico, Icchu, Snakeeater, VolkovBot, Technopat, Matdrodes, Muro Bot, PaintBot, Drinibot, Pacomegia, Correogsk, Tirithel, Dnu72, HUB, Antón Francho, Carro e, Botito777, Alexbot, Juan Mayordomo, Raulshc, Hucknall, UA31, Chomolungma, AVBOT, MastiBot, HanPritcher, NjardarBot, Diegusjaimes, DrFO.Tn.Bot~eswiki, Andreasmperu, Luckas-bot, Madmaxsrx, Jcoronelf, El Quinche, FariBOT, Sergiportero, Mcapdevila, SuperBraulio13, Jkbw, Botarel, BOTirithel, TiriBOT, Blinski, TorQue Astur, Rouxfederico, PatruBOT, AldanaN, Ivanpares, EmausBot, Sergio Andres Segovia, ConPermiso, Gecime, Alfonso Aguilar, Fbport, MerlIwBot, Acratta, Johnbot, Elvisor, Maria Antonia Aguilar C., Balles2601 y Anónimos: 118
- **Homocedasticidad** *Fuente:* https://es.wikipedia.org/wiki/Homocedasticidad?oldid=82257667 *Colaboradores:* Tano4595, Zam, CEM-bot, Booksboy, Isha, JAnDbot, Lasai, Hlnodovic, Muro Bot, Eduardosalg, Ccelis, Juan Mayordomo, CayoMarcio, Mcapdevila, Rubinbot, Botarel, Tomgc, PatruBOT, Nachosan, EmausBot, SUPUL SINAC, Grillitus, MerlIwBot, KLBot2 y Anónimos: 18
- **Regresión logística** *Fuente:* https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica?oldid=74604986 *Colaboradores:* Joseaperez, GermanX, Dlc, Davius, Julian Mendez, TXiKiBoT, VolkovBot, Synthebot, Lmmolinero, PaintBot, Trujilloleonardo, Loveless, Bigsus-bot, BOTarate, Juan Mayordomo, MastiBot, Luckas-bot, El Quinche, Xqbot, Ivanpares, EmausBot, KLBot2 y Anónimos: 3
- **Modelos de regresión múltiple postulados y no postulados** *Fuente:* https://es.wikipedia.org/wiki/Modelos_de_regresi%C3%B3n_m%C3%BAltiple_postulados_y_no_postulados?oldid=70365872 *Colaboradores:* Sabbut, Superzerocool, GermanX, CEM-bot, Juan Mayordomo, El Quinche, KLBot2 y Anónimos: 5
- **Regresión segmentada** *Fuente:* https://es.wikipedia.org/wiki/Regresi%C3%B3n_segmentada?oldid=76915843 *Colaboradores:* BOT-Superzerocool, CEM-bot, Mr. Moonlight, CommonsDelinker, Juan Mayordomo, LucienBOT, Luckas-bot, ConPermiso, Grillitus, KLBot2 y Elvisor
- **Econometría** *Fuente:* https://es.wikipedia.org/wiki/Econometr%C3%ADa?oldid=83391389 *Colaboradores:* JorgeGG, SpeedyGonzalez, Wesisnay, Robbot, Elwikipedista, Tano4595, ManuP, Ictlogist, Petronas, Rembiapo pohyiete (bot), Maltusnet, Kokoo, Orgullobot~eswiki, RobotQuistnix, Yrbot, FlaBot, Maleiva, BOTijo, YurikBot, Suntalzel, KnightRider, C-3POrao, Eskimbot, Maldoror, Vicfaria, Nihilo, Lmendo, CEM-bot, Damifb, Davius, Rastrojo, Booksboy, Ggenellina, Thijs!bot, Artemiorguez, Eccgs, TXiKiBoT, Josedomingoh, Netito777, Raspatan, Fixertool, Idioma-bot, Lnumb, Pólux, Almendro, Lnegro, VolkovBot, Nicoguaro, AlleborgoBot, Muro Bot, YonaBot, SieBot, Anual, BOTarate, Pacomegia, Correogsk, Xqno, Guille46, Jloc25, PixelBot, Arhendt, Açipni-Lovrij, Kintaro, Diegusjaimes, Arjuno3, Luckas-bot, El Quinche, Ptbotgourou, SuperBraulio13, Xqbot, AstaBOTh15, Yosicogito, Jerowiki, TjBot, Juanda1234567, AHG2010, EmausBot, Bachi 2805, HRoestBot, WikitanvirBot, CocuBot, JABO, AvocatoBot, MetroBot, John plaut, DanielithoMoya, Helmy oved, Addbot y Anónimos: 81
- **Mínimos cuadrados** *Fuente:* https://es.wikipedia.org/wiki/M%C3%ADnimos_cuadrados?oldid=79074853 *Colaboradores:* Riviera, Tano4595, Balderai, Soulreaper, Heimy, GermanX, Tigerfenix, BOTpolicia, CEM-bot, Daniel De Leon Martinez, Alexav8, Davius, Antur, Jjafjjaf, FrancoGG, Drackangel, Thijs!bot, Alvaro qc, Escarbot, JAnDbot, Rafa3040, Muro de Aguas, TXiKiBoT, Hlnodovic, Felipebm, Urdangaray, Technopat, Cesar.romero.avello, KronT, Matdrodes, CJMax, Muro Bot, BotMultichill, SieBot, PaintBot, Loveless, Fdefreitas6, BOTarate, Pacomegia, Correogsk, DorganBot, Javierito92, WitoGLX, Juan Mayordomo, Hucknall, SergioN, AVBOT, HanPritcher, Dstanizzo, Pasmargo, EdBever, Luckas-bot, Jaromero, Amirobot, Ptbotgourou, ArthurBot, MartinDM, SuperBraulio13, Xqbot, Jkbw, Joticajulian, Rubinbot, Botarel, AstaBOTh15, J1wu, Ivanpares, Afrasiab, EmausBot, ZéroBot, J. A. Gélvez, WikitanvirBot, Palissy, Deachp, Pedro Jara V, MetroBot, Acratta, Jxwx, Addbot, Tburroni y Anónimos: 56
- **Regularización de Tíjonov** *Fuente:* https://es.wikipedia.org/wiki/Regularizaci%C3%B3n_de_T%C3%ADjonov?oldid=64617183 *Colaboradores:* Riviera, AlfonsoERomero, Davius, Hlnodovic, VolkovBot, PaintBot, Bigsus-bot, STBot~eswiki, Jvelasco85, Juan Mayordomo, DumZiBoT, MystBot, DiegoFb, SassoBot, EmausBot, ZéroBot, KLBot2 y Anónimos: 1
- **Cuarteto de Anscombe** *Fuente:* https://es.wikipedia.org/wiki/Cuarteto_de_Anscombe?oldid=73925906 *Colaboradores:* Juanjo Bazan, Davius, Thijs!bot, TXiKiBoT, Ignacioerrico, Rovnet, Urdangaray, Daniel Ajoy, Alecs.bot, Alexbot, Juan Mayordomo, VanBot, UA31, Nocturnogatuno, Luckas-bot, Xqbot, PatruBOT, KamikazeBot, EmausBot, ZéroBot, Addbot y Anónimos: 3
- **Modelo de valoración de activos financieros** *Fuente:* https://es.wikipedia.org/wiki/Modelo_de_valoraci%C3%B3n_de_activos_financieros?oldid=82587441 *Colaboradores:* Jynus, Sms, Jsanchezes, JohnWest~eswiki, Dianai, ManuP, Petronas, Airunp, RobotQuistnix, Chobot, Floydian~eswiki, Yrbot, FlaBot, BOTijo, YurikBot, KnightRider, CEM-bot, Davius, Dajuliani, Thijs!bot, Rrmsjp, JAnDbot, Mandrake33, Muro de Aguas, Karla guzman, VolkovBot, Technopat, Murdockcrc, BotMultichill, PaintBot, Anual, Javierito92, Alecs.bot, MastiBot, Mariocastrogama, Diegusjaimes, Andreasmperu, ArthurBot, Xqbot, Jkbw, D'ohBot, RedBot, KamikazeBot, Perico1992, ZéroBot, ChuispastonBot, Raul196756, Addbot y Anónimos: 47
- **Análisis armónico** *Fuente:* https://es.wikipedia.org/wiki/An%C3%A1lisis_arm%C3%B3nico?oldid=82462566 *Colaboradores:* Ivn, Ricardos, Sms, Tostadora, Elwikipedista, Petronas, Rembiapo pohyiete (bot), Boogiepazzo, Orgullobot~eswiki, BOT-Superzerocool, Davidsevilla, FlaBot, YurikBot, Cheveri, Chlewbot, CEM-bot, Roberpl, Davius, Flobo, Thijs!bot, Matdrodes, Seraphita~eswiki, Muro Bot, Tommy Boy, Juan Mayordomo, CarsracBot, El Quinche, Jkbw, MerlIwBot, Addbot y Anónimos: 13
- **Teorema de Gauss-Márkov** *Fuente:* https://es.wikipedia.org/wiki/Teorema_de_Gauss-M%C3%A1rkov?oldid=78316660 *Colaboradores:* Stoni, Dianai, Gengiskanhg, Hispa, Orgullobot~eswiki, YurikBot, DarkDante, Alfredobi, Thijs!bot, IrwinSantos, Botones, Cgb, TXiKiBoT, Hlnodovic, VolkovBot, Urdangaray, Muro Bot, PaintBot, Loveless, Farisori, Eduardosalg, Petruss, Juan Mayordomo, Carlos Rogério Santana, AVBOT, Chzelada, Ptbotgourou, Hampcky, KLBot2, YFdyh-bot y Anónimos: 13
- **Análisis de la regresión** *Fuente:* https://es.wikipedia.org/wiki/An%C3%A1lisis_de_la_regresi%C3%B3n?oldid=82826409 *Colaboradores:* LP, Amadís, BOT-Superzerocool, GermanX, Lauranrg, Botones, Matdrodes, Muro Bot, Gerakibot, ???? robot, SrDonPatrón, Juan Mayordomo, UA31, AVBOT, MarcoAurelio, Ezarate, Cépey, EdBever, KamikazeBot, Ripchip Bot, GrouchoBot, Ivanpares, Wikiléptico, EmausBot, ConPermiso, Grillitus, ChuispastonBot, Antonorsi, MerlIwBot, KLBot2, Acratta, JYBot, Ralgisbot, Ihtizon, Arisdas y Anónimos: 30

- **Regresión robusta** *Fuente:* https://es.wikipedia.org/wiki/Regresi%C3%B3n_robusta?oldid=81750944 *Colaboradores:* CEM-bot, Ivanpares, Grillitus, Danielglezschez17, MetroBot y Invadibot
- **Valor eficaz** *Fuente:* https://es.wikipedia.org/wiki/Valor_eficaz?oldid=76883838 *Colaboradores:* PACO, Franjesus, Felipealvarez, LeonardoRob0t, Xuankar, Rembiapo pohyiete (bot), Roo72, RobotQuistnix, Chobot, Mabuimo, Yrbot, Echani, Robespierre, George McFinnigan, CEM-bot, Jorgelrm, Davius, Rastrojo, Thijs!bot, JAnDbot, TXiKiBoT, Nolaiz, Humberto, Rei-bot, Technopat, Patxistein, Muro Bot, Charly2807, Sureda, Fede Threepwood, JaviMad, StarBOT, Juan Mayordomo, Raulshc, Sebarex, AVBOT, David0811, MastiBot, Davidgutierrezalvarez, Madalberta, Luckas-bot, Jvdura, Mcapdevila, SuperBraulio13, Gusbelluwiki, Jerowiki, PatruBOT, MerlIwBot, Travelour, Helmy oved, Wikilario, Addbot y Anónimos: 55
- **Análisis de la varianza** *Fuente:* https://es.wikipedia.org/wiki/An%C3%A1lisis_de_la_varianza?oldid=80123228 *Colaboradores:* JorgeGG, Juan Manuel, Rembiapo pohyiete (bot), LP, RobotQuistnix, Yrbot, BOT-Superzerocool, BOTijo, JAGT, KnightRider, Jjgibaja, Chlewbot, Paintman, Futbolero, CEM-bot, Davius, Resped, Thijs!bot, Lauranrg, Botones, JAnDbot, Mion, Alfambra, Rafa3040, TXiKiBoT, Macalla, VolkovBot, The Bear That Wasn't, Matdrodes, Muro Bot, SieBot, Trujilloleonardo, Loveless, Bigsus-bot, PixelBot, Alecs.bot, PetrohsW, Alexbot, Juan Mayordomo, AVBOT, Diegusjaimes, InflaBOT, Luckas-bot, Wikisilki, Nallimbot, Ptbotgourou, Deemonita, SuperBraulio13, Jkbw, Botarel, Nikolin rio, Gorigori, D'ohBot, TiriBOT, MondalorBot, TobeBot, Rorduna, Ripchip Bot, Humbefa, CentroBabbage, GrouchoBot, EmausBot, Jcaraballo, WikitanvirBot, Acarabal, KLBot2, Elvisor, Phaliel y Anónimos: 70

16.8.2 Imágenes

- **Archivo:Anscombe.svg** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/b/b6/Anscombe.svg *Licencia:* GPL *Colaboradores:* ? *Artista original:* ?
- **Archivo:Artículo_bueno.svg** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/e/e5/Art%C3%ADculo_bueno.svg *Licencia:* Public domain *Colaboradores:* Circle taken from Image:Symbol support vote.svg *Artista original:* Paintman y Chabacano
- **Archivo:CHAO.png** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/6/69/CHAO.png *Licencia:* Attribution *Colaboradores:* Trabajo propio *Artista original:* The original uploader was R.J.Oosterbaan de Wikipedia en inglés
- **Archivo:Carl_Friedrich_Gauss.jpg** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/9/9b/Carl_Friedrich_Gauss.jpg *Licencia:* Public domain *Colaboradores:* Gauß-Gesellschaft Göttingen e.V. (Foto: A. Wittmann). *Artista original:* Gottlieb Biermann
A. Wittmann (photo)
- **Archivo:Heterocedastico.JPG** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/6/68/Heterocedastico.JPG *Licencia:* Public domain *Colaboradores:* ? *Artista original:* ?
- **Archivo:Homocedastico.JPG** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/b/bf/Homocedastico.JPG *Licencia:* Public domain *Colaboradores:* ? *Artista original:* ?
- **Archivo:Linear_least_squares2.png** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/9/94/Linear_least_squares2.png *Licencia:* Public domain *Colaboradores:* self-made with MATLAB, tweaked in Inkscape. *Artista original:* Oleg Alexandrov
- **Archivo:Linear_regression.svg** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/3/3a/Linear_regression.svg *Licencia:* Public domain *Colaboradores:* Trabajo propio *Artista original:* Sewaqu
- **Archivo:Logistic-curve.svg** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/8/88/Logistic-curve.svg *Licencia:* Public domain *Colaboradores:* Created from scratch with gnuplot *Artista original:* Qef (talk)
- **Archivo:MUSTARD.JPG** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/c/c9/Segmented_linear_regression_graph_showing_yield_of_mustard_plants_vs_soil_salinity_in_Haryana%2C_India%2C_1987%E2%80%931988.jpg *Licencia:* Public domain *Colaboradores:* Trabajo propio; transferred from nl.wikipedia. *Artista original:* R.Oosterbaan (R.J. Oosterbaan) at nl.wikipedia.
- **Archivo:SegReg1.gif** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/6/68/SegReg1.gif *Licencia:* Attribution *Colaboradores:* Transferred from en.wikipedia *Artista original:* R.J.Oosterbaan at en.wikipedia
- **Archivo:SegReg2.gif** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/8/8a/SegReg2.gif *Licencia:* Attribution *Colaboradores:* Transferred from en.wikipedia *Artista original:* R.J.Oosterbaan at en.wikipedia
- **Archivo:SegReg3.gif** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/4/45/SegReg3.gif *Licencia:* Attribution *Colaboradores:* Transferred from en.wikipedia *Artista original:* R.J.Oosterbaan at en.wikipedia
- **Archivo:Translation_arrow.svg** *Fuente:* https://upload.wikimedia.org/wikipedia/commons/2/2a/Translation_arrow.svg *Licencia:* CC-BY-SA-3.0 *Colaboradores:* Este gráfico vectorial fue creado con Inkscape. *Artista original:* Jesse Burgheimer

16.8.3 Licencia de contenido

- Creative Commons Attribution-Share Alike 3.0