Técnicas de regresión y análisis de datos

Informe nº: 17-0013-DEV-01
Fecha: 17 abril 2018

Extracto

En este documento se tratará de dar, de manera introductoria, una base sobre las distintas técnicas de regresión. Empezaremos explicando el concepto de estimador de máxima verosimilitud para luego ir derivando distintos problemas. Más adelante se explicará en qué consisten algunos modelos más avanzados.

Palabras clave: regresión, análisis, datos, Deming, chi-cuadrado, estimador máxima verosimilitud (MLE), regresión robusta, regresión simétrica.

Yingli Green Energy Europe, S.L.

R&D, Aftersales Service Center
Pol. Ind. Sur - Ctra. N-I km 32,1
E-28750 San Agustín del Guadalix
(Madrid) Spain

Elaborado por:

Ricardo Olivas González
Estudiante
Fecha: 17 abril 2018

Aprobado por:

José María Román
Laboratory Director
Fecha: 17 abril 2018

ÍNDICE

1 INTRODUCCIÓN

El análisis de regresión es un proceso estadístico para estimar la relación entre variables. Nace con Legendre y Gauss, los cuales utilizaron lo que hoy se denomina el método de mínimos cuadrados para determinar a partir de observaciones astronómicas, las órbitas de distintos cuerpos y así, predecir cuándo aparecerían en el cielo terrestre.

Las técnicas de análisis de datos han seguido evolucionando desde 1800 y actualmente es uno de los campos de mayor interés, ya que ante la gran cantidad de datos que surge a cada minuto, es necesario utilizar técnicas potentes de análisis. En este documento haremos una breve introducción a los métodos existentes más comunes.

2 ESTIMADOR DE MÁXIMA VEROSIMILITUD

Dado un conjunto de medidas (y_i, x_i) suponemos que no son exactas y contienen un error respecto del valor estimado que consideramos exacto. La intención es determinar los valores exactos de forma que se minimicen esos errores o distancias con un método de regresión (como en el caso de mínimos cuadrados). Esta minimización puede verse como la maximización de la distribución de probabilidad del error, que obtiene el estimador de máxima verosimilitud (maximun likehood estimation).

El método de máxima verosimilitud es un método objetivo para encontrar buenos estimadores puntuales del error. Sea una función que nos da la probabilidad de obtener los resultados medidos según unos parámetros, el estimador de máxima verosimilitud corresponde a los valores de parámetros que maximiza la probabilidad de obtener los errores existentes en el conjunto de medidas (y_i, x_i). [1]

Es importante señalar, que si tenemos un error sistemático, no podremos corregirlo por procesos numéricos o promedios.

En esta sección introduciremos el método de resolución que aplicaremos en cada método de regresión.

1) Definiremos una probabilidad o distribución P del error que puede tomar cualquier forma (Gaussiana, Poisson…) para cada par de datos (y_i, x_i). Habitualmente tendrá la forma de un multiplicatorio, puesto que asumimos que tenemos n variables independientes distribuidas según la misma distribución de probabilidad. La más generalizada es usar la distribución normal porque suele ser la distribución a la que converge la suma de un gran número de pequeñas desviaciones aleatorias (teorema central del límite), pero la realidad no siempre es esta.
2) Tomamos el logaritmo de P y cambiamos su signo, así tenemos una función más fácil de tratar y en lugar de maximizarla la minimizaremos (por ello el cambio de signo, realmente no es necesario cambiarlo). $-\log P = f(\{y_i, x_i\}; a_0 \dots a_M)$.
3) Derivamos $f(\{y_i, x_i\}; a_0 \dots a_M)$ respecto a sus parámetros y minimizamos. (En ocasiones derivar puede ser complicado y se utilizarán métodos numéricos).
4) Llegamos a un sistema de ecuaciones, resolvemos y hallamos los parámetros.
5) Por último, mediante propagación de errores hallaremos la incertidumbre de los parámetros estimados. Si los datos son independientes, cada uno contribuye con su propia incertidumbre: $a_k = a_k(y_i, x_i)$, y por lo tanto la incertidumbre del parámetro a_k se expresa como $(\delta a_k)^2 = \sum_{i=1}^{N} (\delta y_i)^2 \left(\frac{\partial a_k}{\partial y_i}\right)^2$.

3 MÍNIMOS CUADRADOS

Empezamos con el caso básico de mínimos cuadrados, donde se maximiza una distribución normal (con errores independientes y desviaciones estándar iguales para cada punto de la “y”, la “x” se supone libre de error).

Por lo tanto, suponiendo que cada dato y_i tiene un error aleatorio e independiente, y con distribución normal alrededor del modelo “real” *Y(x)*, y con la misma desviación estándar σ para todos los puntos, la probabilidad de todo el conjunto de datos es el producto de las probabilidades de cada punto [2]:

$$P = \prod_{i=1}^{N}\left\{exp\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i; a_0 \dots a_M)}{\sigma}\right)^2\right]\cdot\frac{1}{\sigma\sqrt{2\pi}}\right\} \quad (1)$$

$$P = \exp\left(\sum_{i=1}^{N}\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i; a_0 \dots a_M)}{\sigma}\right)^2\right]\right)\cdot\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N$$

$$-logP = \sum_{i=1}^{N}\left[\frac{[y_i - Y(x_i; a_0 \dots a_M)]^2}{2\sigma^2}\right] - N\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) \quad (2)$$

de forma que maximizar (1) es equivalente a minimizar el negativo de su logaritmo (2), llegando a la minimización conocida:

$$f(\{y_i, x_i\}; a_1 \dots a_M) \sim \sum_{i=1}^{N}[y_i - Y(x_i; a_0 \dots a_M)]^2 \quad (3)$$

Y minimizamos (3) respecto los parámetros a_M

3.1 Caso lineal $Y = a_0 + a_1 x$

Para el caso $Y = a_0 + a_1 x$, minimizamos respecto a_0 y a_1 y hallamos:

$$\left.\begin{aligned} 0 = \frac{\partial f}{\partial a_0} &= -2\sum[y_i - a_0 - a_1 x_i] \\ 0 = \frac{\partial f}{\partial a_1} &= -2\sum[y_i - a_0 - a_1 x_i]\cdot x_i \end{aligned}\right\} \quad (4)$$

Resolviendo el siguiente sistema de ecuaciones se obtienen los coeficientes,

$$\left.\begin{aligned} a_0 N + a_1\sum x_i &= \sum y_i \\ a_0\sum x_i + a_1\sum {x_i}^2 &= \sum x_i y_i \end{aligned}\right\}$$

$$a_0 = \frac{\sum {x_i}^2 \sum y_i - \sum x_i \sum x_i y_i}{N\sum {x_i}^2 - \left(\sum x_i\right)^2}$$

$$a_1 = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum {x_i}^2 - \left(\sum x_i\right)^2} = \frac{\frac{1}{N}\sum x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{N}\sum {x_i}^2 - \bar{x}^2} = \frac{\frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}\sum(x_i - \bar{x})^2} \quad (5)$$

Una vez calculados los parámetros nos preguntamos por la incertidumbre de los mismos ($\delta y = \sigma$).

$$\left.\begin{aligned} \frac{\partial a_0}{\partial y_i} &= \frac{\sum {x_i}^2 - \sum x_i \cdot x_i}{N\sum {x_i}^2 - \left(\sum x_i\right)^2} \\ \frac{\partial a_1}{\partial y_i} &= \frac{N x_i - \sum x_i}{N\sum {x_i}^2 - \left(\sum x_i\right)^2} \end{aligned}\right\} \quad (7)$$

Elevando (7) al cuadrado y aplicando el sumatorio $(\sum(\sum x_i) = N\sum x_i)$ se simplifican los términos y se llega a:

$$\left.\begin{aligned}\delta a_0^2 &= \frac{\sum {x_i}^2}{\Delta'}\,\sigma^2 \\ \delta a_1^2 &= \frac{N}{\Delta'}\,\sigma^2\end{aligned}\right\} \quad (8)$$
$$\Delta' = N\sum {x_i}^2 - \left(\sum x_i\right)^2$$

Estos términos σ_i (en propagación de errores se suele utilizar δ), son errores en la medida, ya sea por el equipo experimental o por desviación típica.

Además de esto es interesante conocer las variancias y covariancias de las dos variables, así como su correlación (ej. Correlación de Pearson). [3]

4 Ajuste χ^2

Si queremos describir la dispersión de una población consideramos la varianza muestral. La distribución chi-cuadrado se introduce para este tipo de problemas.

Si x_i es una muestra de N variables aleatorias distribuidas normal e independientemente, con medias μ_i y varianzas ${\sigma_i}^2$, el estadístico

$$\chi^2 = \sum_{i=1}^{N}\left[\frac{x_i - \mu_i}{\sigma_i}\right]^2$$

es distribuido con función de densidad

$$f(x,N) = \frac{1}{2^{N/2}\cdot\Gamma(N/2)}\,x^{[N/2]-1}\cdot e^{-x/2}$$

Y con media N y varianza 2 N; ν grados de libertad. [21]

Ahora supongamos que cada par de datos (x_i,y_i) tiene su propia desviación estándar σ_i (nota: el error está en y_i). Entonces solo se añade el índice i a la minimización, ($P = \left(\frac{1}{\sigma_i\sqrt{2\pi}}\right)^N \exp\left(\sum_{i=1}^{N}\left[-\frac{1}{2}\left(\frac{y_i - Y(x_i)}{\sigma_i}\right)^2\right]\right)$):

$$f = \chi^2 = \sum_{i=1}^{N}\left(\frac{y_i - Y(x_i; a_0, \dots, a_M)}{\sigma\mathrm{i}}\right)^2 \quad (9)$$

N-M es el número de grados de libertad, queremos ajustar los $a_1, \dots, a_M$ parámetros que minimicen el valor de χ^2. Se puede calcular la bondad del ajuste mediante la función gamma. [2] Ver anexo (B).

Las ecuaciones que nos dan el chi cuadrado mínimo se obtienen derivando:

$$0 = \sum_{i=1}^{N}\left(\frac{y_i - Y}{{\sigma_i}^2}\right)\left(\frac{\partial Y(x_i; a_0, \dots, a_M)}{\partial a_k}\right) \qquad k = 0, \dots, M \quad (10)$$

4.1 Caso lineal $Y = a_0 + a_1 x$

$$\chi^2(a_0, a_1) = \sum_{i=1}^{N}\left(\frac{y_i - a_0 - a_1 x_i}{\sigma_i}\right)^2 \quad (11)$$

Ahora tenemos el mismo problema de mínimos cuadrados pero con un error individual, procedemos a resolverlo como en la ecuación (4). También podríamos agrupar la desviación individual en y_i y en $Y(x_i; a_1, \dots, a_M)$ en nuevas variables, calculando mínimos cuadrados normales y recuperando los parámetros al final del algoritmo.

$$\left.\begin{aligned} 0 = \frac{\partial \chi^2}{\partial a_0} = -2\sum \frac{[y_i - a_0 - a_1 x_i]}{{\sigma_i}^2} \\ 0 = \frac{\partial \chi^2}{\partial a_1} = -2\sum \frac{[y_i - a_0 - a_1 x_i]}{{\sigma_i}^2} x_i \end{aligned}\right\}$$

y llegamos a la solución:

$$\left.\begin{aligned} a_0 = \frac{S_{xx}S_y - S_x S_{xy}}{\Delta} \\ a_1 = \frac{SS_{xy} - S_x S_y}{\Delta} \end{aligned}\right\} \qquad (12)$$

con:

$$\Delta \equiv SS_{xx} - (S_x)^2$$

$$S \equiv \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2} \qquad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{{\sigma_i}^2} \qquad S_y \equiv \sum_{i=1}^{N} \frac{y_i}{{\sigma_i}^2}$$

$$S_{xx} \equiv \sum_{i=1}^{N} \frac{{x_i}^2}{{\sigma_i}^2} \qquad S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{{\sigma_i}^2}$$

Nota: Se pueden reescribir las soluciones en términos sencillos de varianzas y covarianzas teniendo en cuenta que tenemos dividiendo ${\sigma_i}^2$ dentro del sumatorio.

Podemos hallar también las incertidumbres de cada parámetro con la ecuación (6) (ahora sí tenemos σ_i, que además está presente en los parámetros):

$$\left.\begin{aligned} (\delta a_0)^2 = \frac{S_{xx}}{\Delta} \\ (\delta a_1)^2 = \frac{S}{\Delta} \end{aligned}\right\} \qquad (13)$$

También se puede hallar la covarianza, el coeficiente de correlación *r*, y la probabilidad Q para saber si es bueno el ajuste. También es interesante reescribir los parámetros para que no sean susceptibles a errores de redondeo a la hora de computar. [2]

Otra forma interesante de plantear el problema sería hacerlo de forma que se resuelva como en la sección 3.1. Podríamos separar la varianza (en *y*) de cada par de datos en una parte constante y otra particular $\sigma_i^2 = \sigma^2 \beta_i^2$. Entonces agrupando variables y renombrando:

$$\hat{y}_i = \frac{y_i}{\beta_i} \qquad \hat{x}_i = \frac{x_i}{\beta_i} \qquad \hat{a}_0 = \frac{1}{N}\sum_i \frac{a_0}{\beta_i^2}$$

Entonces podemos operar de igual forma que en el caso de mínimos cuadrados, esta vez derivando respecto a $\hat{a}_0$ y a_1, y al final retomando los valores a_k.

4.2 Línea recta con errores en ambas coordenadas

Si los datos experimentales están sujetos a error tanto en y como en x, el problema se complica. Podemos escribir la función mérito χ^2 (función que mide la concordancia entre los datos y el modelo, y va minimizando para unos parámetros elegidos):

$$\chi^2(a_0, a_1) = \sum_{i=1}^{N} \frac{(y_i - a_0 - a_1 x_i)}{{\sigma_{yi}}^2 + {a_1}^2 {\sigma_{xi}}^2} \quad (14)$$

La suma de varianzas ponderadas del denominador se puede entender como la varianza en la dirección del χ^2 más pequeño entre cada par de datos y la línea de pendiente a_1, o como la varianza de la combinación lineal de $y_i - a_0 - a_1 x_i$ y dos variables aleatorias y_i y x_i.

$$Var(y_i - a_0 - a_1 x_i) = Var(y_i) + {a_1}^2 Var(x_i) = {\sigma_{yi}}^2 + {a_1}^2 {\sigma_{xi}}^2 \equiv 1/w_i$$

Minimizar este caso ya no es tan sencillo, puesto que al tener el parámetro a_1 en el denominador, la derivada respecto de a_1 no es lineal. Con respecto al parámetro a_0 si es lineal, y podemos usar métodos numéricos (cómo el método de Brent) para minimizar una función general respecto de a_1, mientras que también a cada paso que está siendo minimizado respecto de a_0, cuya función lineal a minimizar sería ($\partial\chi^2/\partial a_0 = 0$):

$$0 = \left[\sum_i w_i (y_i - a_1 x_i)\right] / \sum_i w_i$$

Es más conveniente parametrizar la pendiente a_1 mediante el ángulo $\theta = arctan\, a_1$, ya que a_1 puede ir a infinito y χ^2 ser finito. Si tenemos una desviación pequeña en las y pero grande en las x, tendremos máximos en χ^2 cerca de la pendiente cero. A veces puede ocurrir que haya dos χ^2 mínimo, uno con pendiente negativa y otro positiva, pero sólo uno es correcto. Para elegir, escalaremos las y_i para tener una varianza igual que las x_i, entonces se hace un ajuste lineal con pesos derivados de la suma (escalada) ${\sigma_{yi}}^2 + {\sigma_{xi}}^2$.

Para calcular los errores estándar, se puede hacer la expansión de Taylor (numéricamente) o mediante las proyecciones en los ejes a_0 y a_1 de los "límites de las regiones de confianza" y encontrar raíces (numéricamente) para $\Delta\chi^2 = 1$

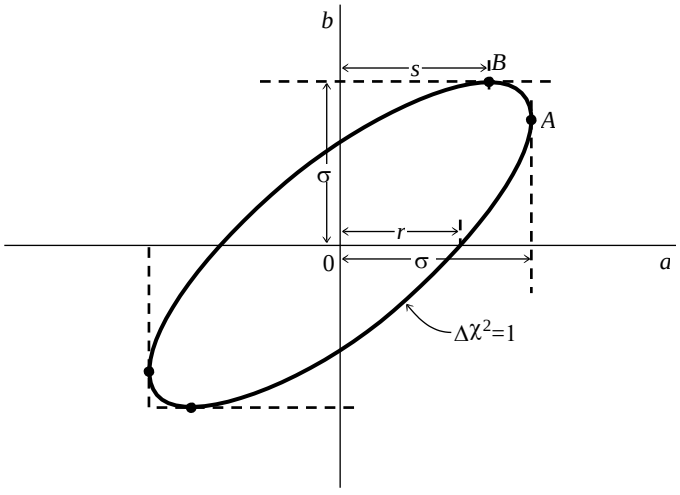

Figura 1: Errores estándar para los parámetros a_0 y a_1. El punto B se puede encontrar variando la pendiente a_1 mientras se minimiza a_0 simultáneamente. Esto da el error estándar σa_1 , y el valor s. El error estándar σa_0 se puede encontrar con la relación geométrica ${\sigma_{a_0}}^2 = s^2 + r^2$ [2]

El inconveniente de este método es que no hay simetría, los errores están relacionados en la función a minimizar, y se minimizan los dos a la vez, comprobando en la iteración que uno se mantiene mínimo mientras se minimiza el otro. Esto nos llevará a buscar un método simétrico (ortogonal, Deming…).

4.3 *"Mínimos cuadrados" lineales generales*

Extendemos lo aprendido a cualquier combinación lineal de parámetros:

$$y(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_M x^M$$

Pero ahora también vamos a generalizar a que las funciones de x puedan ser cualquier función y definiremos $X_k(x)$, con k=0,…,M, como funciones base. Estas funciones pueden ser no lineales, la linealidad sólo la queremos en la combinación con sus parámetros a_k.

$$Y(x) = \sum_{k=0}^{M} a_k X_k(x) \quad (15)$$

Se define entonces la función (vista en sección 4.1 para la recta):

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \sum_{k=0}^{M} a_k X_k(x_i)}{\sigma_i} \right)^2 \quad (16)$$

Trabajaremos entonces con matrices:

$$A_{ij} = \frac{X_j(x_i)}{\sigma_i} = \begin{pmatrix} \frac{X_0(x_1)}{\sigma_1} & \frac{X_1(x_1)}{\sigma_1} & \cdots & \frac{X_M(x_1)}{\sigma_1} \\ \frac{X_0(x_2)}{\sigma_2} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{X_0(x_N)}{\sigma_N} & \cdots & \cdots & \frac{X_M(x_N)}{\sigma_N} \end{pmatrix} \quad (17)$$

A_{ij} matriz Nx(M+1) con M funciones base y N datos. Normalmente tenemos que N≥(M+1) (más puntos que parámetros a resolver).

Definimos también los vectores:

$b_i = \frac{y_i}{\sigma_i}$, (Dimensión N) ($\vec{b}$, vector fila)

$\vec{a} = a_0, \ldots, a_M$, (Vector *columna* dimensión M+1, cuyos componentes son los parámetros a ajustar) (18)

La ecuación (16) se reescribiría así: $\chi^2 = \vec{b} - A_{ij}\vec{a}$

Veamos modos de resolver esto.

4.3.1 Solución por Ecuaciones Normales

La resolución típica consiste en usar ecuaciones normales. Para minimizar χ^2 derivamos respecto a_k e igualamos a cero:

$$0 = -2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - \sum_{j=0}^{M} a_j X_j(x_i) \right] X_k(x_i) \qquad k = 1, \ldots, M \quad (19)$$

Tenemos M ecuaciones (una por cada parámetro) que contienen sumatorios de la M ecuaciones base y los N datos. Intercambiando el orden de los sumandos, reescribimos (19):

$$\sum_{j=0}^{M} \alpha_{kj} a_j = \beta_k \quad (20)$$

donde,

$\alpha_{kj} = \sum_{i=1}^{N} \frac{X_j(x_i) X_k(x_i)}{{\sigma_i}^2}$ (Matriz (M+1)x(M+1)), es equivalente a $[\alpha] = \boldsymbol{A^T \cdot A}$

$\beta_k = \sum_{i=1}^{N} \frac{y_i X_k(x_i)}{{\sigma_i}^2}$ (Vector de longitud M+1), es equivalente a $[\beta] = \boldsymbol{A^T \cdot b}$

Las ecuaciones normales en forma matricial se escriben entonces como:

$$[\alpha] \cdot \boldsymbol{a} = [\beta] \Leftrightarrow (\boldsymbol{A^T \cdot A}) \cdot \boldsymbol{a} = \boldsymbol{A^T \cdot b} \quad (21)$$

Notar que nuestras funciones $Y(x_i)$ vienen definidas por $\boldsymbol{A \cdot a = b}$.

$(\boldsymbol{A^T \cdot A})$ es una matriz normal (conmuta con su traspuesta conjugada) y es simétrica y definida positiva, lo que nos permite emplear la factorización de Cholesky [4]. De aquí se puede despejar $\boldsymbol{a}$.

La ecuación (21) se puede resolver mediante descomposición LU, factorización de Cholesky o eliminación de Gauss Jordan. Si se quiere hallar la matriz de covarianzas *[C]* (que a continuación explicaremos) es más conveniente usar Gauss Jordan. Si sólo queremos hallar los parámetros podemos usar LU para ahorrar en el álgebra. En teoría Cholesky es el método más eficiente, pero en la práctica resulta que usa el tiempo de computación haciendo bucles sobre los datos para formar las ecuaciones, y Gauss-Jordan es más adecuado. [2]

La matriz inversa $C_{jk} \equiv [\alpha]_{kj}{}^{-1}$ está relaciona con las posibles incertidumbres de los parámetros **a**, para estimar estas incertidumbres consideramos:

$$a_j = \sum_{k=0}^{M} [\alpha]_{jk}{}^{-1} \beta_k = \sum_{k=0}^{M} C_{jk} \left[\sum_{i=1}^{N} \frac{y_i X_k(x_i)}{{\sigma_i}^2} \right] \quad (22)$$

y las varianza asociadas con los estimados a_j, se calculan mediante propagación como en la ecuación (6):

$$(\delta a_j)^2 = \sum_{i=1}^{N} {\sigma_i}^2 \left(\frac{\partial a_j}{\partial y_i} \right)^2 \quad (23)$$

Como C_{jk} es independiente de y_i,

$$\frac{\partial a_j}{\partial y_i} = \sum_{k=0}^{M} C_{jk} X_k(x_i) / {\sigma_i}^2 \quad (24)$$

Sustituyendo en (23):

$$(\delta a_j)^2 = \sum_{k=0}^{M} \sum_{l=0}^{M} C_{jk} C_{jl} \left[\sum_{i=1}^{N} \frac{X_k(x_i) X_l(x_i)}{{\sigma_i}^2} \right] \quad (25)$$

En (25) se han agrupado los sumatorios en k y l, y se han intercambiado el orden de los sumatorios, para así ver que tenemos entre corchetes la matriz $[\alpha]_{lk}$. Como es la matriz inversa de *[C]*, al multiplicar nos queda una matriz C_{jk} por la matriz unidad y quedan los términos:

$$\sigma^2(a_j) = C_{jj}$$

Es decir, los elementos diagonales de la matriz *[C]* son las varianzas (cuadrado de las incertidumbres) de los parámetros ajustados **a**. Además se puede probar que los elementos C_{jk} son las covarianzas entre a_j y a_k.

Prueba [5]
Sea $\boldsymbol{C}$ una matriz de entradas constantes y $\boldsymbol{V} = \boldsymbol{CU} \rightarrow Cov(\boldsymbol{V}) = \boldsymbol{C}Cov(\boldsymbol{U})\boldsymbol{C^T}$ (se puede demostrar), entonces buscamos la matriz de covarianza de $\boldsymbol{a}$. De la ecuación (21),

$$\boldsymbol{a} = (\boldsymbol{A^T} \cdot \boldsymbol{A})^{-1} \cdot \boldsymbol{A^T} \cdot \boldsymbol{b}$$

Definimos:

$$\boldsymbol{C} = (\boldsymbol{A^T} \cdot \boldsymbol{A})^{-1} \cdot \boldsymbol{A^T} \rightarrow \boldsymbol{C^T} = \boldsymbol{A} \cdot (\boldsymbol{A^T} \cdot \boldsymbol{A})^{-1}$$
$$\boldsymbol{U} = \boldsymbol{b}$$

La covarianza del vector $\boldsymbol{y}$ es justamente la varianza ${\sigma_i}^2$ veces la matriz identidad N-dimensional, esto es ${\sigma_i}^2\mathbb{I}$, porque las observaciones son independientes. Entonces la covarianza del vector $\boldsymbol{b}$ será la matriz identidad $\mathbb{I}$.

$$Cov(\boldsymbol{a}) = \boldsymbol{C}Cov(\boldsymbol{b})\boldsymbol{C^T} = (\boldsymbol{A^T} \cdot \boldsymbol{A})^{-1} \cdot \boldsymbol{A^T} \cdot \mathbb{I} \cdot \boldsymbol{A} \cdot (\boldsymbol{A^T} \cdot \boldsymbol{A})^{-1} = (\boldsymbol{A^T} \cdot \boldsymbol{A})^{-1} = [C]$$

4.3.2 Solución mediante Singular Value Decomposition

A veces podemos tener errores computacionales de redondeo, o ecuaciones normales muy próximas a ser singulares (matriz singular es aquella con determinante igual a cero). Esto a veces ocurre porque dos o más funciones base son similares y tendremos más de una combinación posible de estas funciones base para ajustar los parámetros. Tendremos problemas tanto sobredeterminados (más datos que parámetros) cómo indeterminados (combinación de parámetros ambigua).

Para estos casos se usa la descomposición en valores singulares que es una forma de factorizar una matriz real o compleja. Así cuando aparezca un elemento cero mientras calculamos las soluciones de las ecuaciones lineal, en lugar de no obtener solución (con Gauss pasaría) calcularemos la solución que minimiza los mínimos cuadrados.

La descomposición en valores singulares de una matriz *A* es la factorización de *A* en el producto de tres matrices $A = UDU^T$, donde las columnas de U y U^T son ortonormales y la matriz *D* es diagonal. [11]

5 AJUSTE SIMÉTRICO EN ERROR

Buscamos ahora un método simétrico, en el que podamos intercambiar la variable *x* y la variable *y* dependiendo de lo que deseemos en ese momento. *Deming* pensó en el problema de ajustar datos con error tanto en *x* como en *y*, pero no como en la sección 4.2 en la que la varianza de ambas estaba relacionada y se minimizaba *b* mientras manteníamos *a* mínimo. Deming propuso una minimización simultánea e independiente, que nos devuelva el caso de mínimos cuadrados normal cuando no halla error en una de las variables.

En este caso, la probabilidad que buscamos maximizar (*MLE*) será:

$$P = \prod_i^N (2\pi {\sigma_x}^2)^{-\frac{1}{2}} \exp\left(-\frac{(x_i - X_i(\ \ ))^2}{2{\sigma_x}^2}\right)(2\pi {\sigma_y}^2)^{-\frac{1}{2}} \exp\left(-\frac{(y_i - Y_i(\ \ ))^2}{2{\sigma_y}^2}\right) \quad (26)$$

Donde $Y_i(\ \ )$ y $X_i(\ \ )$ son los estimados. Para resolver el problema se suele tomar $Y_i(X_i) = a_0 + a_1 X_i(\ \ )$, y así se va resolviendo el problema para los parámetros a_k. Esto conlleva un problema, la pérdida de simetría, ya que un estimado depende del otro. Para solucionarlo podemos parametrizarlos, de forma que el cálculo de uno sea independiente del otro, hay métodos (Theil-Sen regression, Passing-Bablock regression).

5.1 Regresión de Deming

Partimos de la base anterior y definimos nuestros estimados $Y_i(\ \ )$ y $X_i(\ \ )$

$$x_i = X_i + e_{xi}$$
$$y_i = Y_i + e_{yi} = a_0 + a_1 X_i + e_{yi} \quad (27)$$

Donde e_{xi} y e_{yi} son los errores en cada variable y la relación de sus varianzas se asume conocida:

$$\lambda = \frac{{\sigma_y}^2}{{\sigma_x}^2} \quad (28)$$

Entonces podemos reescribir nuestro estimador (26):

$$P = \prod_i^N (2\pi {\sigma_x}^2)^{-\frac{1}{2}} \exp\left(-\frac{(x_i - X_i)^2}{2{\sigma_x}^2}\right)(2\pi\lambda {\sigma_x}^2)^{-\frac{1}{2}} \exp\left(-\frac{(y_i - a_0 - a_1 X_i)^2}{2\lambda {\sigma_x}^2}\right)$$

Tomando el logaritmo (esta vez no aplico signo negativo, buscaré maximizar):

$$f = log\ P = -\frac{n}{2}\log(4\pi^2 {\sigma_x}^4 \lambda) - \frac{\sum_i^n (x_i - X_i)^2}{2{\sigma_x}^2} - \frac{\sum_i^n (y_i - a_0 - a_1 X_i)^2}{2\lambda {\sigma_x}^2} \quad (29)$$

Ahora tenemos en cuenta el detalle de que *x* está sujeto a error, por lo que también tendremos que maximizar la probabilidad de su estimado (como Y_i depende de X_i no es necesario derivar respecto Y_i). Escribimos directamente las soluciones, la derivación se puede comprobar en el documento de Anders Christian Jensen [6]:

$$\frac{\partial f}{\partial X_i} = 0 \dashrightarrow X_i = \frac{\lambda x_i + a_1(y_i - a_0)}{\lambda + {a_1}^2} = x_i + \frac{a_1(y_i - a_0 - a_1 x_i)}{\lambda + {a_1}^2} \quad (30)$$

$$\frac{\partial f}{\partial a_0} = 0 \dashrightarrow a_0 = \frac{1}{N}\sum(y_i - a_1 x_i) = \bar{y} - \bar{x} a_1 \quad (31)$$

$$\frac{\partial f}{\partial a_1} = 0 \dashrightarrow a_1 = \frac{-(S_{yy} - \lambda S_{xx}) \pm \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}}}{-2S_{xy}} \quad (32)$$

Donde se han agrupado los términos

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - N\bar{y}^2$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - N\bar{x}^2$$

$$S_{xy} = \sum (y_i - \bar{y})(x_i - \bar{x}) = \sum x_i y_i - N\bar{x}\bar{y}$$

Así pues hemos llegado a las soluciones para el caso en el que tenemos en cuenta que en la variable independiente puede haber error. Notar que aunque en el estimador aparece ${\sigma_x}^2$ pero a la hora de calcular el máximo (igualando a cero la derivada) este desaparece para nuestros valores estimados de X_i y los parámetros. También es importante definir el denominado caso *ortogonal* en el que $\lambda = 1$ (varianzas iguales para ambas variables).

Vemos en la ecuación (32) que tenemos dos soluciones. Debido al hecho de que $S_{yy} - \lambda S_{xx} \leq \sqrt{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}}$ y que el signo de la pendiente debe ser igual al signo de S_{xy} ya que esta es la covarianza, y nos indica si la variable dependiente crece o decrece con la independiente (signo positivo o negativo), y la pendiente nos indica lo mismo.

Para calcular la incertdumbre de los parámetros procedemos a realizar la propagación de errores habitual.

$$(\delta a_k)^2 = \sum_{i=1}^{N} \left[{\sigma_y}^2 \left(\frac{\partial a_k}{\partial y_i}\right)^2 + {\sigma_x}^2 \left(\frac{\partial a_k}{\partial x_i}\right)^2 \right]$$

Es posible llegar a la misma solución de otra forma distinta, tal y cómo Glaister realiza [7]. Partiendo de los mínimos cuadrados ordinarios, podemos añadir el término de la variable independiente y minimizar la distancia (que es al fin y al cabo maximizar la probabilidad del estimador como hemos visto).

$$d^2 = \sum_{i=1}^{N} [(x_i - X_i)^2 - (y_i - Y_i)^2]$$

entonces definiendo la línea que pasa por los puntos estimados $Y_i = a_1 X_i + a_0$, y la línea que une los puntos obtenidos con los estimados $\frac{y_i - Y_i}{x_i - X_i}$. El producto de ambas pendientes será $\frac{y_i - Y_i}{x_i - X_i} \cdot a_1 = -\lambda = \frac{{\sigma_y}^2}{{\sigma_x}^2}$

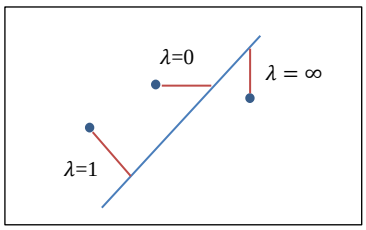

Figura 2: Esquema de cómo se realizan las regresiones para los distintos casos, ortogonal $(\boldsymbol{\lambda = 1})$***, con error sólo en*** $\mathbf{x}(\boldsymbol{\lambda = 0})$ ***y con error sólo en*** $\mathbf{y}(\boldsymbol{\lambda = \infty})$**.**

A partir de aquí Glaister construye la distancia a minimizar de forma que solo dependa de los parámetros a_0 y a_1 y se llega a los mismos resultados, ver [7].

Para finalizar esta sección, vamos a comprobar que efectivamente, tomando límites podemos recuperar la solución de mínimos cuadrados ordinarios, tanto para *x* como para *y*. Agrupando términos de la ecuación (32) y tomando límites:

$$\lim_{\lambda\to\infty}(-p + \sqrt{p^2 + \lambda}) = \lim_{\lambda\to\infty} \frac{\lambda}{p + \sqrt{p^2 + \lambda}} = \lim_{\lambda\to\infty} \frac{1}{\frac{p}{\lambda} + \sqrt{p^2/\lambda^2 + 1/\lambda}} = \lim_{\lambda\to\infty} \frac{1}{2p/\lambda}$$

Donde se ha tomado $p = \frac{(S_{yy} - \lambda S_{xx})}{-2S_{xy}}$, se llega entonces a:

$$\lim_{\lambda\to\infty} a_1 = \frac{S_{xy}}{S_{xx}} \quad (33)$$

Que concuerda con el caso de limites cuadrados ordinarios (sin error en x).

Tomando el límite $\lambda \to 0$

$$\lim_{\lambda\to 0}(-p + \sqrt{p^2 + \lambda}) = \frac{S_{yy}}{S_{xy}} \quad (34)$$

Que es el caso de mínimos cuadrados sin error en y *($X_i = b_0 + b_1 y_i$).*

Sabiendo que en el límite sin error en *x* a_1 es finito, podemos tomar el límite $\lambda \to \infty$ para X_i

$$X_i = x_i$$

Queda demostrado.

5.2 Caso general

En esta sección se intenta aclarar un poco si se podría llevar a cabo un caso más general, con la formulación de la sección 5.1.

Lo primero que podemos pensar en generalizar es el error. Tal y como se explicó en el caso 4Ajuste χ^2, se intenta asignar a cada par de datos una desviación estándar propia, en lugar de tomarla constante.

$$\sigma_{xi}^2 = {\sigma_x}^2 \alpha_i^2 \to \widehat{(x_\iota - X_\iota)} = \frac{x_i - X_i}{\alpha_i}$$

$$\sigma_{yi}^2 = {\sigma_y}^2 \beta_i^2 \to \widehat{(y_\iota - Y_\iota)} = \frac{y_i - Y_i}{\beta_i}$$

Esta sería una opción resolveríamos como en el caso Deming, y al final recuperaríamos el caso particular con error distinto en cada par de datos.

Si queremos en lugar de una línea recta, tener una función polinómica, el caso se complica y la resolución algebraica puede alargarse mucho. No se resolverá explícitamente pero una forma sencilla de comprobar que también sería válido, es resolver cual sería el error en caso de $y_i = Y_i + e_{yi} = a_0 + a_1 X_i + a_2 X_i^2 + \cdots a_k X_i^k + e_{yi}$, con $X_i = x_i - e_{xi}$

El esquema a seguir sería el siguiente:

$$f = -logP \sim \frac{1}{2}\sum (y_i - a_0 + a_1(x_i - e_{xi}) - a_2(x_i - e_{xi})^2 - \cdots a_k(x_i - e_{xi})^k)^2 - \lambda {e_{xi}}^2$$

donde se ha sustituido $x_i - X_i = e_{xi}$

Al derivar respecto de e_{xi} vamos a obtener términos sin el error de x multiplicando, términos con hasta ${e_{xi}}^{(2k-1)}$ multiplicando, pero sólo un término con λ. Se llega a algo como:

$$\frac{\partial f}{\partial e_{xi}} = 0 \rightarrow e_{xi} \sim \left(\frac{1}{\lambda} \left(\frac{(y_i - \cdots - a_k (x_i)^k (a_1 + \cdots + c a_k x_{k-1}}{(1 - \frac{\mathcal{O}}{\lambda})} \right) \right)^{1/(2k-1)}$$

Entonces cuando $\lambda \rightarrow \infty$, $e_{xi} = 0$, sin embargo, no hemos resuelto el valor de cada parámetro a_k, que cómo vemos requiere un esfuerzo considerable, y también dependerá de λ. Aunque parece que a priori, si volvemos al caso en el que no hay varianza en x, efectivamente encontramos que el error e_{xi} es nulo.

5.3 Errors-in-variable models

También denominados modelos con errores de medida. No es una técnica de regresión sino un conjunto de ellas (el método de Deming estaría dentro de estos modelos). Además, existen modelos EIV (errors in variable) múltiples, es decir que tienen en cuenta más de 2 variables (ej.: *x,y,z*). Un ejemplo sencillo de estos métodos es el de Functional Relationship. Es un tipo de regresión EIV múltiple en la que se considera que una de las variables es determinista, es decir se considera que es la realidad. Esta variable real se le suele denominar T: [8][9]

$$\begin{gathered} X = T + e_x \\ Y = a_0 + a_1 T + e_y \\ Z = b_0 + b_1 T + e_z \end{gathered} \quad (35)$$

y la solución de los parámetros es:

$$\begin{gathered} a_1 = S_{yz}/S_{xz} \\ b_1 = S_{yz}/S_{xy} \\ a_0 = \bar{y} - a_1 \bar{x} \\ b_0 = \bar{z} - b_1 \bar{x} \end{gathered}$$

donde las S_{ij} son las covarianzas.

El modelo estructural es distinto, asumiría que las T_i son muestras aleatorias de una variable con media μ y varianza σ^2.

Este método requeriría un estudio más amplio ya que es útil para modelos no lineales con varias variables y está constituido por diversas técnicas.

5.3.1 Total Least Squares

O mínimos cuadrados totales, es un tipo de modelo de regresión en errores en las variables, que se puede interpretar como una generalización del modelo de Deming. En referencias dejo biografía recomendada [14] [15]. Se resuelve minimizando un funcional con restricciones (multiplicadores de Lagrange), como se puede ver rápidamente en [16] y [17].

6 MODELOS NO LINEALES

Pequeña introducción a cómo procederíamos si los parámetros a_k no formaran una combinación lineal. El procedimiento será computacional, minimizando χ^2 en cada iteración.

6.1 *Minimización de χ^2 [2]*

Se procede a minimizar χ^2 para determinar los parámetros, pero ahora se hará de manera iterativa. Dados unos valores prueba para los parámetros se desarrolla un proceso para mejorar la solución de prueba. Pararemos cuando χ^2 deja de decrecer.

Cerca del mínimo esperamos que χ^2 se aproxime a una función cuadrática de la forma:

$$\chi^2(\boldsymbol{a}) \approx \gamma - \boldsymbol{d}\cdot\boldsymbol{a} + \frac{1}{2}\boldsymbol{a}\cdot\boldsymbol{D}\cdot\boldsymbol{a}$$

d= (M+1) vector, **D**= matriz (M+1)x(M+1).

Si esta aproximación es buena, podemos hallar los parámetros **a** mínimos:

$$\boldsymbol{a}_{min} = \boldsymbol{a}_{current} + \boldsymbol{D^{-1}}\cdot[-\nabla\chi^2(\boldsymbol{a}_{current})] \quad (36)$$

Si la aproximación es un poco pobre, podemos dar un paso gradiente abajo, como en el método steepest descent [10]:

$$\boldsymbol{a}_{min} = \boldsymbol{a}_{current} - constant \times \nabla\chi^2(\boldsymbol{a}_{current}) \quad (37)$$

Con la constante suficientemente pequeña para no "agotar" la dirección cuesta abajo (esto requiere del estudio de numerical recipes y los métodos computacionales que usa).

Para usar estas dos fórmulas necesitamos ser capaces de computar el gradiente y conocer la matriz **D**, que es la matriz de derivadas segundas de χ^2 en cualquier **a** (matriz Hessiana).

6.1.1 Modelo. Cálculo del gradiente y la Hessiana

El modelo a ajustar es $Y = Y(x;\vec{a})$ y $\chi^2 = \sum_{i=1}^{N}\left(\frac{y_i - Y(x;\vec{a})}{\sigma_i}\right)^2$

El gradiente de χ^2 respecto a sus parámetros a_k, que será cero en su mínimo:

$$\frac{\partial\chi^2}{\partial a_k} = -2\sum_{i=1}^{N}\frac{[y_i - Y(x_i;\vec{a})]}{{\sigma_i}^2}\frac{\partial Y(x_i;\vec{a})}{\partial a_k} \qquad k = 0,1,\ldots M \quad (38)$$

Tomando más derivada parciales

$$\frac{\partial^2\chi^2}{\partial a_k\partial a_l} = 2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\frac{\partial Y(x_i;\vec{a})}{\partial a_l}\frac{\partial Y(x_i;\vec{a})}{\partial a_k} - [y_i - Y(x_i;\vec{a})]\frac{\partial^2 Y(x_i;\vec{a})}{\partial a_k\partial a_l}\right] \quad (39)$$

Definimos β_k y α_{kl} para quitarnos el factor 2:

$$\beta_k \equiv -\frac{1}{2}\frac{\partial\chi^2}{\partial a_k}; \qquad \alpha_{kl} \equiv \frac{1}{2}\frac{\partial^2\chi^2}{\partial a_k\partial a_l}$$

haciendo $[\alpha] = \frac{1}{2}\boldsymbol{D}$ en la ecuación (36), la reescribimos como el conjunto de ecuaciones lineales

$$\sum_{l=1}^{M} a_{kl}\,\delta a_l = \beta_k \quad (40)$$

Este conjunto se resuelve para los incrementos $\delta a_l = a_{lmin} - a_{lcurrent}$, que añadido a la aproximación actual, nos da la siguiente aproximación. En el contexto de mínimos cuadrados, a la matriz $[\alpha]$, igual a la mitad de la matriz Hessiana, se la suele denominar matriz de curvatura.

La ecuación steepest descent (37) , se transforma en:

$$\delta a_l = const \times \beta_k \quad (41)$$

Hay que darse cuenta de que las componentes a_{kl} de la matriz Hessiana dependen de las primeras y segundas derivadas respecto a sus parámetros. Nosotros ignoraremos las segundas derivadas. El término $\partial^2 Y(x_i;\vec{a}) \big/ \partial a_k \partial a_l$ se puede quitar cuando es cero o despreciable respecto al término de la primera derivada. Además el término $[y_i - Y(x_i;\vec{a})]$ puede tomar ambos signos y en general no está correlacionado con el modelo, por lo que los términos de la segunda derivada tienden a cancelarse sumados a todo *i*.

$$a_{kl} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial Y(x_i;\vec{a})}{\partial a_l} \frac{\partial Y(x_i;\vec{a})}{\partial a_k} \right] \quad (42)$$

La condición en el χ^2 mínimo es $\beta = 0$ para todo *k*. Independientemente de la definición de $[\alpha]$.

6.1.2 Método de Levenberg-Marquardt

Consiste en ir variando suavemente entre los extremos del método de la Hessiana-Inversa y el método steepest descent. Primero consideramos la constante del método steepest descent, ¿qué orden de magnitud tiene? ¿Y su escala? No hay información sobre ello en el gradiente, ya que sólo nos dice la pendiente, no como de lejos se extiende esa pendiente. La primera idea/percepción de Marquard es que las componentes de la Hessiana, aunque no son útiles de una forma precisa, nos da información de la escala del problema.

La cantidad χ^2 no tiene dimensiones. $[\beta_k]$ tendrá dimensiones de $[a_k]^{-1}$ (cada componente de β_k puede tener diferentes dimensiones, cm^{-1}…). La constante de proporcionalidad entre β_k y δa_k tendrá entonces dimensiones de $[a_k]^2$. En las componentes de [α] solo hay una cantidad con esas dimensiones: $1/\alpha_{kk}$, esa debe ser la escala de la constante. Si es demasiado grande dividimos por una constante λ. Reemplazaremos la ecuación (41) por

$$\delta a_l = \frac{1}{\lambda \alpha_{ll}} \beta_l \qquad \text{ó} \qquad \lambda \alpha_{ll} \delta a_l = \beta_l \quad (43)$$

La segunda idea es que las ecuaciones anteriores y la ecuación $\sum_{l=1}^{M} \alpha_{kl} \delta a_l = \beta_k$ pueden combinarse si definimos una matriz nueva α' de la siguiente forma

$$\alpha'_{jj} \equiv \alpha_{jj}(1 + \lambda)$$
$$\alpha'_{jk} \equiv \alpha_{jk} \quad (j \neq k) \quad (44)$$

Y reemplazar (40) y (43) por

$$\sum_{l=1}^{M} \alpha'_{kl} \delta a_l = \beta_k \quad (45)$$

si λ es muy grande es diagolmente dominante y es igual que $\lambda \alpha_{ll} \delta a_l = \beta_l$, si λ es cercana a cero, se llega al método Hessiano.

Dada una suposición inicial para el conjunto de parámetros **a** ajustados, el método de Marquard es el siguiente:

- Calcular (numéricamente) $\chi^2(\vec{a})$
- Elegir un valor modesto para λ ($\lambda = 0{,}001$)
- **(*)** Resolver las ecuaciones lineales (45) para δa_l y evaluar $\chi^2(\vec{a} + \delta\vec{a})$
- Si $\chi^2(\vec{a} + \delta\vec{a}) \geq \chi^2(\vec{a})$ incrementamos λ un factor 10 y volvemos a **(*)**

- Si $\chi^2(\vec{a}+\delta\vec{a}) < \chi^2(\vec{a})$, disminuimos λ un factor 10, actualizamos la solución prueba $\vec{a} \leftarrow \vec{a}+\delta\vec{a}$, y volvemos a **(*)**

Falta la condición para parar la iteración. Pararemos de iterar la primera o segunda ocasión en que χ^2 diminuya una cantidad despreciable <<1 (p.e <0,01).

Una vez hemos llegado a un mínimo aceptable, hacemos $\lambda = 0$ y computamos la matriz $[C] \equiv [\alpha]^{-1}$. Que es la matriz de covarianza estimada de los errores estándar de los parámetros estimados $\vec{a}$.

7 MÉTODOS DE REGRESIÓN ROBUSTA

Son métodos de regresión diseñados para no ser excesivamente afectados si los supuestos (ej, suposiciones de distribución normal de mínimos cuadrados) son violados por el proceso de generación de datos. Puede ocurrir por cualquier pequeña desviación (departures) de todos los puntos, o por grandes desviaciones de un pequeño número de puntos. La última interpretación, relacionada con el concepto de valores atípicos (outliers) suele tener más peso y produce más diferencias.

Los estadistas han desarrollado varios tipos de estimadores estadísticos robustos. Se pueden agrupar (en general), en las siguientes categorías:

- M-estimadores: Vienen de los argumentos de máxima verosimilitud (máximum-likelihood). Suelen ser los más relevantes para la estimación de parámetros (model-fitting). En la página 696 de [2] se ve con más detalle, consiste en elegir otra distribución que no sea la normal (tendrá una "cola" más larga, por lo que será menos sensible a valores atípicos), y derivar como hemos estado haciendo a lo largo del documento.
- L-estimadores: "combinaciones lineales de *estadísticos de orden*". Se aplican a estimaciones del valor central y la tendencia central, aunque pueden ocasionalmente aplicarse para estimar parámetros. 2 estimadores-L típicos son la mediana y *Tukey's trimean* (que se define como la media ponderada del primer, segundo y tercer cuartil en una distribución, con pesos ¼, ½ y ¼ respectivamente).
- R-estimadores están basados en pruebas de rango. Por ejemplo, la igualdad o desigualdad de dos distribuciones puede estimarse con el test de Wilcoxon calculando el rango medio de una distribución en una muestra combinada de ambas distribuciones.

Hay muchos métodos de regresión robusta, como puede ser Theil–Sen estimator (en estadística no paramétrica).

7.1 Modelo lineal generalizado

Las regresiones lineares ordinarias predicen el valor esperado de una cantidad desconocida como una combinación lineal de unos observables. Implica una variación constante de la variable respuesta (esperada) para un cambio en los valores observado. Esto no funciona siempre. Por ejemplo, si predecimos que con 10º C menos, 1000 personas menos irán a la playa, no tenemos en cuenta que la playa puede ser grande o pequeña, deberíamos hallar una proporción (un 50% menos). Cuando hallamos una proporción, hablamos de exponential-response models (log-linear model). También hay modelos en los que se evalúa la tendencia de que una persona vaya o no a la playa sin recurrir a probabilidades fijas (log-odds or logistic model).
En el modelo lineal generalizado, cada salida Y de las variables dependientes se asume que es generada a partir de una distribución particular perteneciente a la familia exponencial (incluye distribución normal, binomial, Poisson, gamma…). (La familia exponencial requiere estudio, tienen unas propiedades algebraicas y estadísticas comunes).

La media **μ** de la distribución depende de las variables independientes **X** a través de la fórmula

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{X\beta})$$

$E(\boldsymbol{Y})$ es el valor esperado de **Y**; $\boldsymbol{X\beta}$ es el «predictor lineal», una combinación lineal de parámetros desconocidos $\boldsymbol{\beta}$; g es la función de enlace.

Con esta notación, la varianza es típicamente una función **V** de la media:

$$Var(\boldsymbol{Y}) = V(\boldsymbol{\mu}) = V(g^{-1}(\boldsymbol{X\beta}))$$

Es conveniente si **V** proviene de una distribución en la familia exponencial, pero podría simplemente ser que la varianza es una función del valor ajustado.

Los parámetros desconocidos $\boldsymbol{\beta}$ son generalmente estimados por máxima verosimilitud, máxima cuasi-verosimilitud [12], o técnicas de inferencia bayesiana [13].

8 ESTADÍSTICA MULTIVARIANTE

Los métodos estadísticos multivariantes y el análisis multivariante son herramientas estadísticas que estudian el comportamiento de tres o más variables al mismo tiempo. Se usan principalmente para buscar las variables menos representativas para poder eliminarlas, simplificando así modelos estadísticos en los que el número de variables sea un problema y para comprender la relación entre varios grupos de variables.

Para más información consultar el curso de Análisis de Datos y Estadística Avanzada del master de Astrofísica UCM+UAM [18] y el Manual abreviado de Análisis Estadístico Multivariante (Jesús Montanero Fernández) [19]

8.1 Análisis componentes principales

Técnica utilizada para reducir la dimensionalidad de un conjunto de datos. Busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Esta convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales. [20]

9 CONCLUSIONES

Hemos realizado un estudio sobre los métodos de regresión vistos desde la base, es decir, desde los estimadores de máxima verosimilitud. A partir de este concepto se puede construir gran parte de la estadística de regresión, ya que de él se derivan los modelos lineales, no-lineales y robustos.

También hemos introducido métodos simétricos, sin embargo, vemos que se pierde parte de la simetría, y que no hemos propuesto un modelo "único" para un problema simétrico. Una posible forma es parametrizar la recta en función del ángulo del vector director de la recta, que puede ser una indicación para continuar y ampliar a casos más generales. Aun así, aplicando el método de Deming y MLE y con el ordenador, deberíamos ser capaces de estudiar (numéricamente) datos con errores en la variable dependiente e independiente.

Debemos ampliar la información presentada en este documento, incluyendo aspectos tales como bases estadísticas o de econometría, pruebas de hipótesis, grados de libertad, etc. que no hemos abordado. Sin embargo, sí es muy conveniente seguir indagando sobre modelos de errores en variables como el Total Least Squares, para poder generalizar un método lo más simétrico posible en n variables.

Hemos apuntado una información preliminar sobre modelos de regresión robusta, que en el marco del MLE puede tenerse en cuenta utilizado distribuciones más generales que la normal.

Para seguir avanzado en el tema de análisis de grandes cantidades de datos, estudiar de verdad la estadística multivariante es necesario, ya que nos permite reducir la dimensionalidad y poder manejarnos.

Finalmente, en los anexos hemos añadido una definición de la estadística Bayesiana, que se basa en el conocimiento del sistema que permite plantear hipótesis a priori sobre el mismo.

Otro aspecto no contemplado está relacionado con los grados de libertad en el sistema estadístico, que se plantea en otro anexo de los anexos.

REFERENCIAS

1. http://webs.ucm.es/info/Astrof/POPIA/asignaturas/ana_dat_est/tema01.pdf , Pag. 10
2. Numerical Recipes in Fortran 77, Chapter 15
3. http://webs.ucm.es/info/Astrof/POPIA/asignaturas/ana_dat_est/tema05.pdf
4. http://lya.fciencias.unam.mx/gfgf/pa20081/data/lecturas/sistemas/EcuNormal.html
5. Devore J.L., Berk K.N. Modern mathematical statistics with applications (2ed., Springer, 2011)(ISBN 9781461403906) Cap 12.8
6. Deming regression (MethComp package) May 2007, Anders Christian Jensen
7. Least Squares Revisited, P. Glaister. The Mathematical Gazette, Vol. 85, No. 502 (Mar., 2001), pp. 104-107
8. Tesis doctoral metodologías de calibración de bases de datos de reanálisis de clima marítimo, Antonio Tomás Sampedro. http://www.tesisenred.net/bitstream/handle/10803/10622/03de12.ATS.cap3.pdf?sequence=4
9. Validation of ocean wind and wave data using triple collocation, S. Caires and A. Sterl
10. Numerical Recipes in Fortran 77, Chapter 10.6
11. https://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage/book-chapter-4.pdf
12. https://es.wikipedia.org/wiki/Cuasi-verosimilitud
13. https://en.wikipedia.org/wiki/Bayesian_inference
14. Total Least Squares, Duke University, Fall 2017
15. https://www.mff.cuni.cz/veda/konference/wds/proc/pdf08/WDS08_115_m4_Pesta.pdf
16. Michael Krystek and Mathias Anton 2007 Meas. Sci. Technol. 18 3438
17. https://academic.oup.com/gji/article/190/2/1135/644394#31573090
18. http://webs.ucm.es/info/Astrof/POPIA/asignaturas/ana_dat_est/
19. http://matematicas.unex.es/~jmf/Archivos/Manual%20de%20Estad%C3%ADstica%20Multivariante.pdf
20. https://es.wikipedia.org/wiki/An%C3%A1lisis_de_componentes_principales
21. Statistics for physicists, B.R. Martin. Department of Physics, University College London, England, 1971

ANEXOS

A INFERENCIA BAYESIANA

"La evidencia sobre el verdadero estado del mundo se expresa en términos de grados de creencia o, más específicamente, las probabilidades bayesianas". Es otro tipo de estadística distinta a la frecuentista. En términos generales (hay que estudiar la matemática asociada) consiste en vincular la probabilidad de A dado B, con la probabilidad de B dado A. La estadística tradicional sólo admite probabilidades basadas en experimentos repetibles y que tengan una confirmación empírica mientras que los llamados estadísticos bayesianos permiten probabilidades subjetivas. La inferencia se entiende como un proceso de actualización de las medidas de credibilidad al conocerse nuevas evidencias.

Inferencia:
Dada una nueva evidencia, el teorema de Bayes ajusta las probabilidades de la misma de la siguiente manera:

$$P(H_0|E) = \frac{P(E|H_0)\,P(H_0)}{P(E)}$$

- H_0 representa una hipótesis, llamada hipótesis nula, que ha sido inferida antes de que la nueva evidencia E, resultara disponible.
- P (H_0) es la *probabilidad a priori* de H_0.
- $P(E|H_0)$ es la *probabilidad condicional* de que se cumpla la evidencia E si la hipótesis H_0 es verdadera. Se llama también la función de verosimilitud cuando se expresa como una función de E dado H_0.
- $P(E)$ es la *probabilidad marginal* de E: la probabilidad de observar la nueva evidencia E bajo todas las hipótesis mutuamente excluyentes. Se puede calcular como la suma del producto de todas las hipótesis mutuamente excluyentes por las correspondientes probabilidades condicionales: $\sum P(E|H_i)\ P(H_i)$
- $P(H_0|E)$ se llama la probabilidad a posteriori de H_0 dado E.

$\frac{P(E|H_0)}{P(E)}$ representa el impacto que la evidencia tiene en la creencia en la hipótesis. El teorema de Bayes mide cuánto la nueva evidencia es capaz de alterar la creencia en la hipótesis.

https://eva.fing.edu.uy/pluginfile.php/81075/mod_resource/content/1/ESTADISTICA%20BAYESIANA.pdf

B GRADOS DE LIBERTAD Y BONDAD DEL AJUSTE

"La geometría nos describe a los grados de libertad como espacios e hiperespacios de libertad a través de los cuales una medida de resumen puede moverse y tomar diferentes valores. El punto de vista algebraico los describe como el número de ecuaciones que se establecen usando los datos. Ambos puntos de vista están relacionados y ayudan a comprender con mayor profundidad el concepto de grados de libertad. Las aplicaciones de los grados de libertad están extendidas a través de toda la estadística, el cálculo de la desviación estándar y la prueba t-de Student son solo algunos ejemplos."

http://www.redalyc.org/pdf/2031/203129458002.pdf

La bondad de ajuste de un modelo estadístico describe lo bien que se ajusta un conjunto de observaciones. Las medidas de bondad en general resumen la discrepancia entre los valores observados y los que valores esperados en el modelo de estudio.

Para chi-cuadrado: http://maxwell.ucsc.edu/~drip/133/ch4.pdf