Rev. Acad. Canar. Cienc., XVI (Núms. 1-2), 113-127 (2004) (publicado en julio de 2005)

CALIBRACIÓN POR REGRESIÓN LINEAL CON ERRORES ADICIONALES

Miguel Sánchez García*, María Inés Sobrón Fernández & Pedro Cuesta Álvaro*****

* Facultad de Medicina. Departamento de Estadística e Investigación Operativa

** Facultad de Matemáticas. Departamento de Estadística e Investigación Operativa

*** Centro de Proceso de Datos

Universidad Complutense. 28040 Madrid

ABSTRACT

Calibrating is meassuring an object as a function of other meassure of this very object. Though cheap and easy to obtain, however, this last meassure is not accurate. The true meassure can be obtained with great accuracy but it takes time and effort. In this work we synthesized controlled calibration and random calibration and also we make a comparative study of classical and reverse calibration methods. A synthesis of the multivariant calibration introduced by Brown (1982) is carried out. The most substantial paragraphs are these numbered 5 and 7, devoted respectively to the creation of a universal methodology to estimate uncertainity in univariant calibration and to the numerical comparisson of several calibration techniques.

KEY WORDS:.Linear Model. Calibration. Multivariate Calibration.

RESUMEN

La calibración consiste en estimar la medida de un objeto en función de otra medida del objeto que, aunque rápida y barata de obtener, es imprecisa. La verdadera medida se puede obtener con mucha precisión, pero cuesta mucho tiempo y recursos obtenerla. En el artículo se realiza una síntesis de la calibración controlada y de la calibración aleatoria, así como también se hace un estudio comparativo de los métodos de calibración clásico e inverso. Se resume la calibración multivariable, que fue introducida por Brown (1982). Las partes más esenciales son los epígrafes 5 y 7, que se dedican, respectivamente, a la creación de una metodología universal para estimar la incertidumbre en calibración univariante y a comparar numéricamente varias técnicas de calibración.

PALABRAS CLAVES: Modelo Lineal. Calibración. Calibración Multivariable.

1 INTRODUCCIÓN.

Un problema de calibración consiste, en el caso más elemental, en realizar inferencias sobre el valor de un vector numérico desconocido X^*; de dimensión $p\times1$; en función de los valores de un vector respuesta Y^* de dimensión $q\times1$. La relación entre Y y X se calibra, o estima, con datos experimentales (Y_i, X_i); i=1,2,…,n; siendo X_i e Y_i vectores de dimensión $p\times1$ y $q\times1$; respectivamente. La calibración es la operación inversa del método más usual de predecir el valor de Y^* en función del valor de X^*.

Se produce una asimetría entre X e Y, cuando se dispone de técnicas mediante las cuales se pueden medir con precisión los valores de X y X^*; mientras que los valores de Y e Y^* sólo se pueden medir por técnicas más imprecisas. En este caso la calibración recibe el nombre de Calibración Controlada; y el conjunto de datos (Y_i, X_i); i =1,2,....,n; recibe el nombre de conjunto de calibración. En la calibración controlada se fijan los valores de la variable X, mediante un diseño óptimo, convenientemente formulado.

Cuando los datos (Y_i, X_i); i=1,2,....,n; se miden con errores significativos, y ambas variables sólo se pueden modelizar de forma aleatoria; entonces la calibración se llama aleatoria. En este supuesto para predecir X^* en función de Y^*; se puede utilizar el modelo de regresión de X sobre Y. Este modelo no es lícito utilizarle en el caso de la calibración controlada.

La calibración se puede realizar también entre las medidas obtenidas con dos o más aparatos de medida, cuando no se dispone de las medidas estándar de varias Unidades Experimentales para el aparato que mide la variable X.

Un ejemplo de calibración se produce al medir la concentración de un cierto enzima en plasma sanguíneo. Dicha concentración se puede medir mediante un procedimiento de laboratorio que es tedioso y costoso, o se puede medir mediante un autoanalizador que es un procedimiento barato y rápido. Un diseño para calibrar el autonalizador puede consistir en seleccionar doce Unidades Experimentales, cada una de las cuales está formada por un cierto volumen de sangre, que tiene una concentración determinada de enzima. Cada Unidad experimental se divide en cinco Unidades Experimentales, del mismo volumen. En dos de estas cinco Unidades experimentales se mide la concentración con el método tradicional y en las otras tres con el autoanalizador. Con los datos así obtenidos, se halla la relación entre la medida exacta, realizada con el método tradicional, y las medidas obtenidas con el autoanalizador.

En general, la calibración tiene por objeto estimar el verdadero valor de la medida de un objeto, item, caso o Unidad Experimental; en función de la medida del objeto observada sobre la escala de un aparato. En calibración por regresión la verdadera medida, o medida real del objeto, es la variable independiente; mientras que la medida del objeto observada sobre el aparato es la variable dependiente. Por Calibración de un aparato de medida se entiende un proceso mediante el cual el calibrador decide si el aparato mide, o no, correctamente. Si el calibrador decide que el aparato mide incorrectamente, debe dar una orden de parar el proceso de medida. La calibración es una técnica indispensable en el control de calidad de los procesos productivos, pero también en cualquier proceso de medida, como pueden ser los laboratorios de Análisis Clínico; en el proceso de llenado de bombonas de oxigeno o de gas; y,en general, en cualquier proceso de medida.

El artículo consta de 6 apartados además de la introducción. En el epígrafe 2, se expone el modelo lineal de calibración con errores y se realizan los cálculos necesarios para la determinación del verdadero valor x de la medida del objeto. En el apartado 3 se desarrollan las técnicas clásicas del modelo de regresión lineal con hipótesis sobre los errores. En el epígrafe 4 se explican las relaciones entre las técnicas de calibracion clásica e inversa. En el epígrafe 5 se expone una técnica general para la estimación de la incertidumbre de las estimaciones de las medidas que se calibran. En el apartado 6, se da un breve resúmen de la calibración multivariable. Se finaliza, en el 7, con el análisis de dos modelos prácticos, uno de calibración univariante y el otro multivariante.

2 MODELO DE CALIBRACIÓN POR REGRESIÓN LINEAL CON ERRORES.

En un ejemplo de calibración, la variable X mide la concentración real del gas en la mezcla. La variable Y es la lectura de la concentración medida por el aparato. Calibrar es determinar la variable concentración X, en función de la variable Y.

El modelo de calibración lineal sin error se formula por la siguiente ecuación

$$y_0 = a_0+b_0x_0 \qquad (2\text{-}1)$$

siendo x_0 la verdadera medida del objeto ó patrón, (en el ejemplo, concentración del gas); y_0 la medida del objeto observada sobre el aparato de medida, b_0 la pendiente de la recta o cambio en la unidad de medida y a_0 la ordenada en el origen. Cuando el modelo de regresión es lineal y no hay errores, la calibración se realizaría por la ecuación (2-2):

$$x_0 = (y_0-a_0)/b_0 \qquad (2\text{-}2)$$

En este supuesto, se determina sin error la verdadera medida del patrón x_0.

El modelo de regresión lineal, con error, para calibración; tiene por ecuación:

$$y = a_0+b_0x_0 + \varepsilon_3 \qquad (2\text{-}1\text{-}1)$$

Si se conocieran todos los elementos de este modelo, el problema de calibración con el modelo de error, que consiste en determinar el valor de x_0 en función del valor conocido de y, es

$$x_0 = (y-a_0)/b_0 + (\varepsilon_3/b_0) \qquad (2\text{-}2\text{-}1)$$

El problema verdadero se presenta porque no se pueden conocer con exactitud, a_0, b_0 y ε_3. En este caso, la fórmula (2-2-1) no se puede evaluar.

Cuando en el modelo de calibración lineal no se cometen errores en las medidas x e y; pero sí en la determinación o estimación de los parámetros a_0 y b_0; así como en el modelo; la ecuación de la Regresión Lineal se formula por la ecuación (2-3):

$$y(1) = (a_0+\varepsilon_1) + (b_0+\varepsilon_2)x_0 + \varepsilon_3 = y_0 + \varepsilon_1 + \varepsilon_2x_0 + \varepsilon_3 \qquad (2\text{-}3)$$

siendo ε_1 el error que se comete en la determinación o estimación del parámetro a_0; ε_2 el error similar para el parámetro b_0 y ε_3 el error que mide la desviación del modelo real respecto del modelo lineal. En este supuesto, llamando $a=a_0+\varepsilon_1$ y $b=b_0+\varepsilon_2$; la calibración exacta se realizaría por la ecuación (2-4):

$$x_0 = (y(1)-a_0-\varepsilon_1-\varepsilon_3)/(b_0+\varepsilon_2)= (y(1)-a-\varepsilon_3)/b = (y_0-a_0)/b_0 \qquad (2\text{-}4)$$

en cuyo caso los errores impedirían determinar con exactitud el verdadero valor x_0 del patrón, teniendo que ser estimado.

Teniendo en cuenta la ecuación (2-4), se tiene:

$$0 = (b-\varepsilon_2)(y(1)-a-\varepsilon_3)-b(y_0-a_0) = ((b-\varepsilon_2)(y(1)-a-\varepsilon_3) - b(y(1)-\varepsilon_1-\varepsilon_2x_0-\varepsilon 3-a+\varepsilon_1)) =$$

$$= b(\varepsilon_2x_0) - \varepsilon_2(y(1)-a-\varepsilon_3) \qquad (2\text{-}4\text{-}1)$$

Despejando en (2-4-1) x_0, se obtiene la igualdad:

$$x_0 = ((y(1)-a)/b) + (\varepsilon_3/b) \qquad (2\text{-}4\text{-}2)$$

En la fórmula (2-4-2) el único valor desconocido es el error ε_3; mientras que en la fórmula (2-2-1) eran desconocidos ε_3, b_0 y a_0.

Cuando en el modelo de calibración lineal se cometen errores tanto en las medidas x_0 e y_0; como en la determinación o estimación por a y b de los parámetros a_0 y b_0; así como en el propio modelo; la ecuación del Modelo de Regresión Lineal es de la forma

$$(y(2)+\varepsilon_4) = (a_0+\varepsilon_1) + (b_0+\varepsilon_2)(x_0+\varepsilon_5) + \varepsilon_3 \qquad (2\text{-}5)$$

siendo, como antes, ε_1 el error que se comete en la determinación o estimación del parámetro a_0; ε_2 el error similar para el parámetro b_0, ε_3 es el error que mide la desviación del modelo real respecto del modelo lineal, ε_5 es el error que se comete al medir el verdadero valor del patrón x_0 y ε_4 es el error que se comete al medir u observar la verdadera medida del aparato y_0. Llamando $x=x_0+\varepsilon_5$, $a=a_0+\varepsilon_1$; y $b=b_0+\varepsilon_2$; la calibración exacta se realizar por la ecuación:

$$x_0 = ((y(2)-a-\varepsilon_3+\varepsilon_4)/(b_0+\varepsilon_2)) - \varepsilon_5 = ((y(2)-a-\varepsilon_3+\varepsilon_4)/b) - \varepsilon_5 = (y_0-a_0)/b_0 \qquad (2\text{-}6)$$

Teniendo en cuenta que:

$$(y(2)+\varepsilon_4) = (a_0+\varepsilon_1) +(b_0+\varepsilon_2)(x_0+\varepsilon_5)+\varepsilon_3=y_0+\varepsilon_1+\varepsilon_2(x_0+\varepsilon_5)+ b_0\,\varepsilon_5+\varepsilon_3$$

se verifica que

$$y_0 = y(2)+\varepsilon_4 - (\varepsilon_1+\varepsilon_2(x_0+\varepsilon_5) + b_0\,\varepsilon_5+\varepsilon_3)$$

Sustituyendo este valor en la ecuación (2-6) se obtiene

$$(b-\varepsilon_2)(y(2)-a-\varepsilon_3+\varepsilon_4-b\varepsilon_5) = b(y(2)+\varepsilon_4-(\varepsilon_1+\varepsilon_2(x_0+\varepsilon_5) +(b-\varepsilon_2)\,\varepsilon_5+\varepsilon_3)-a+\varepsilon_1) \qquad (2\text{-}6\text{-}1)$$

Simplificando en la ecuación (2-6-1) se obtiene:

$$-\varepsilon_2(y(2)-a-\varepsilon_3+\varepsilon_4) = -b\varepsilon_2(x_0+\varepsilon_5)$$

Despejando se obtiene:

$$x_0 = ((y(2)-a-\varepsilon_3+\varepsilon_4)/b) - \varepsilon_5 \qquad (2\text{-}7)$$

Simulando en los modelos (2-4-2) y (2-7) se puede obtener información útil para una mejor aproximación en la estimación del verdadero valor del patrón x_0.

3 MODELO LINEAL DE CALIBRACIÓN CON HIPÓTESIS.

Hasta aquí no hemos hablado de hipótesis. En el modelo (2-1-1) podríamos poner a_0 y b_0 como quisiéramos, con tal de que b_0 fuera distinto de cero, para que conociendo ε_3. se pudiera hallar x_0 por la fórmula (2-2-1); afirmación que no deja de ser una utopía. Como consecuencia se necesitan hacer hipótesis sobre el modelo lineal (2-2-1). Las hipótesis que se suelen hacer sobre este modelo, son las siguientes:

(H1) El error ε_3 es una variable aleatoria que se distribuye como una normal de media cero y desviación típica σ desconocida. Como consecuencia, la media ó esperanza matemática de y es

$$E(y/x)= a_0+b_0x \tag{3-1}$$

(H2) Para cada medida de y, y_j, que se haga con valores iguales ó distintos de x; la variable error $\varepsilon_j(x)$; tiene como desviación típica ó incertidumbre una cantidad σ; que es independiente de x, y de la medida j-ésima. El modelo (2-2-1) queda como:

$$y_j=a_0+b_0x+\varepsilon_j(x) \tag{3-2}$$

La hipótesis segunda afirma que la varianza de los errores es: $V(\varepsilon_j(x))=\sigma^2$; que es independiente de la concentración.

(H3) Para distintas medidas y_j de un mismo objeto, las variables error $\varepsilon_j(x)$ son incorreladas, que en variables normales equivale a que sean independientes.

Si se cumplen las hipótesis anteriores, y se realizan n experiencias con diferentes valores de x, (concentraciones), y con los datos resultantes se estiman los parámetros del modelo (2-2-1) por mínimos cuadrados, que son también estimadores de máxima verosimilitud cuando se cumplen las hipótesis H1, H2 y H3; se obtienen los estimadores a y b por las fórmulas

$$b = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{Cov(y,x)}{Var(x)} = \frac{S_{xy}}{S_{xx}} \tag{3-3}$$

y

$$a = \bar{y} - b\,\bar{x} \tag{3-4}$$

Las siguientes conclusiones se encuentran demostradas en muchos autores. Los resultados se han tomado de Sánchez y otros (6).

(a) El estimador mínimo cuadrático b de b_0, se distribuye, cuando se cumplen las hipótesis H1, H2 y H3, como una variable normal de media b_0 y varianza ó cuadrado de la incertidumbre igual a

$$(\text{INCER(b)})^2 = \hat{s}^2(b) = \frac{\sigma^2}{nVar(x)} \tag{3-5}$$

donde

$$Var(x) = S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3-6}$$

(b) El estimador mínimo cuadrático a de a_0, se distribuye, cuando se cumplen las hipótesis H1, H2 y H3, como una variable normal de media a_0 y varianza ó cuadrado de la incertidumbre igual a

$$(\text{INCER}(a_0))^2 = \hat{s}^2(a_0) = \sigma^2\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nVar(x)}} \tag{3-7}$$

(c) La varianza del error de la regresión es

$$E(\sum_{i=1}^{n}(y_i - a - bx)^2) = E(\sum_{i=1}^{n}\hat{\varepsilon}_i^2) = (n-2)\sigma^2 \tag{3-8}$$

(d) El estimador de la media de las observaciones $\bar{y}(x_0)$ en un punto x_0; se distribuye, cuando se cumplen las hipótesis H1, H2 y H3, como una variable normal de media $a_0+b_0x_0$ y varianza ó cuadrado de la incertidumbre igual a

$$(\text{INCER}(\bar{y}(x_0))^2 = \hat{s}^2(\bar{y}(x_0)) = \sigma^2\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nVar(x)}} \tag{3-9}$$

(e) El estimador de la observación $y(x_0)$ en un punto x_0; se distribuye; cuando se cumplen las hipótesis H1, H2 y H3, como una variable normal de media $a_0+b_0x_0$ y varianza ó cuadrado de la incertidumbre igual a

$$(\text{INCER}(y(x_0))^2 = \hat{s}^2(y(x_0)) = \sigma^2\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nVar(x)}} \tag{3-10}$$

(f) El estimador de la varianza del error es

$$\hat{s}^2 = \frac{n}{n-2}S_{yy}(1 - \hat{r}^2) \tag{3-11}$$

donde S_{yy} es la varianza empírica de la variable dependiente y; mientras que $\hat{r}^2$ es el coeficiente de correlación estimado o empírico.

Del modelo práctico: $y = a+bx+\hat{\varepsilon}$ se infiere que

$$x=((y-a)/b)+(\hat{\varepsilon}/b) \tag{3-12}$$

En modelo (2-2-1) son desconocidos (b_0, b_1, ε_3 y x_0); mientras que en el modelo (2-3) sólo son desconocidos (x_0 y ε_3)

La estimación práctica se realiza por

$$y(obs) = a+bx_0+\hat{\varepsilon} \Rightarrow x_0 = (y(obs)-a)/b) + (\hat{\varepsilon}/b) \tag{3-13}$$

Puesto que $(\hat{\varepsilon}/b)$ es desconocido, por ser desconocido $\hat{\varepsilon}$; tenemos que estimar x_0 por $\hat{x}$; donde $\hat{x} = ((y-a)/b)$. El cuadrado de la incertidumbre de $\hat{x}$ es

$$(\text{INCER}(\hat{x}))^2 = E(\hat{x}-x_0)^2 = E((y-a)/b-(y-a_0)/b_0-\varepsilon/b)^2 =$$

$$= E(\frac{y(obs)-a}{b} - (y-a_0)/b_0-(\varepsilon/b_0))^2 \tag{3-14}$$

Si aproximamos la ecuación anterior por el método DELTA se tiene

$$E(\hat{x}-x_0)^2 \approx E(\frac{y(obs)-y(est)}{b})^2 + E(a-a_0)^2/b^2 +$$

$$+ (\frac{y(obs)-a}{b})^2 \frac{1}{b^2} E(b-b_0)^2 + E(\frac{\hat{\varepsilon}}{b})^2 \qquad (3\text{-}15)$$

La fórmula (3-15) se utiliza para calcular intervalos de confianza para el verdadero valor x_0.

4 TÉCNICAS DE CALIBRACIÓN CLÁSICA E INVERSA.

En la industria farmacéutica antes de que se introduzca un producto en el mercado se necesita comprobar y demostrar que el producto cumple los límites especificados en la farmacopea sobre dosis, potencia, disolución, contenido uniforme y desintegración. El procedimiento para contrastar el medicamento conlleva la determinación de la variabilidad y la precisión de un ensayo. Por tanto es esencial la calibración de un instrumento para obtener medidas precisas sobre las unidades experimentales de una muestra del producto. Con frecuencia se utiliza un método de calibración lineal para determinar los resultados del ensayo, esto es, las concentraciónes del medicamentos en muestras, que se usan para contrastar la bondad del medicamento. La calibración lineal típica implica dos técnicas que se usan frecuentemente, la clásica y la inversa. En el método clásico, se preparan varias concentraciones x_i de las dosis del medicamento y con un aparato se miden los valores de la variable respuesta Y_i (Absorvancias). Con los datos (X_i,Y_i) se puede obtener una curva de calibración lineal ajustando un modelo de regresión lineal de Y sobre X. En un experimento se puede determinar la concentración x_0, reemplazando en la recta de regresión el valor de y por su Absorvancia. En el método inverso, se obtiene la recta de regresión de x sobre y. La concentración de un experimento dado se obtiene de forma similar. Las técnicas clásica e inversa se han usado y se siguen usando con frecuencia, siendo ampliamente aceptadas en el área de la calibración de un aparato de medida.

Para explicar analíticamente los procedimientos clásico e inverso, sea x_i la i-ésima preparación estándar conocida, (concentración) e y_i la correspondiente respuesta; es decir, la Absorvancia medida con el aparato, i=1,2,...,n.

El estimador clásico se obtiene del ajuste de la recta de regresión de y sobre x, cuya ecuación estimada es

$$y = a + bx + \hat{\varepsilon} \qquad (4\text{-}1)$$

El cálculo de x, denominado x_C; por el modelo clásico se obtiene despejando la x de la ecuación anterior, bajo el supuesto de que el error es cero. Llamando S_{xx} a la varianza de X y S_{xy} a la covarianza entre X e Y; se obtiene

$$x_C = \frac{(y_0 - a)}{b} = \bar{x} + \frac{(y_0 - \bar{y})S_{xx}}{S_{xy}} \qquad (4\text{-}2)$$

Alternativamente, utilizando la regresión inversa de X sobre Y se obtiene, llamando S_{yy} a la varianza empírica de la variable Y, la estimación inversa, dada por la fórmula

$$x_I = \bar{x} + \frac{(y_0 - \bar{y})S_{xy}}{S_{yy}} \qquad (4\text{-}3)$$

Para comparar la diferencia entre x_C y x_I se suele utilizar el cociente (x_I/x_C) o el cociente relativo $(x_I - \bar{x})/(x_C - \bar{x})$.

Teniendo en cuenta la definición del coeficiente de correlación R, se verifica:

$$(x_I - \bar{x})/(x_C - \bar{x}) = (S_{xy}^2/S_{xx}S_{yy}) = R^2$$

y $$x_I/x_C = 1+(1-R^2)(\bar{x}/x_C - 1) \quad (4\text{-}4)$$

De las relaciones anteriores se obtienen las siguientes consecuencias:

(C1) Los dos estimadores son iguales si y sólo sí $R^2=1$.

(C2) Puesto que $R^2 \leq 1$, el estimador inverso es siempre más próximo a $\bar{x}$ que el estimador clásico.

(C3) La distribución del cociente relativo es independiente del verdadero valor desconocido x_0.

El objetivo es calcular $p = P(1-((x_I - \bar{x})/(x_C - \bar{x}))>\delta)$

Llamando $SSR = S_{xy}^2/S_{xx}$ y $SSE = S_{yy} - SSR$; entonces $R^2/(1-R^2) = SSR/SSE$ y
$p = p(\theta) = P(1-((x_I - \bar{x})/(x_C - \bar{x}))>\delta) = P(F<(n-2)(1-\delta)/\delta)$;

donde la variable F tiene una distribución F de Snedecor no central con 1 y n-2 grados de libertad y el parámetro de no centralidad es $S_{xx}\theta$; siendo $\theta = b^2/\sigma^2$. Por observación directa se sabe que $p(\theta)$ es una función decreciente de θ.

En general los procedimientos de estimación clásico e inverso no son intercambiables. Otro procedimiento intermedio entre los dos anteriores es el de estimar la recta de regresión de tal forma que se minimicen los cuadrados de los errores medidos como la distancia ortogonal de los puntos a la recta.

5 GENERALIZACIÓN DE LA MODELIZACIÓN CON ERRORES.

Si se conociera la función del cambio de escala f, y no se produjera ningún error en el proceso de medida, la relación entre la cantidad exacta ó medida del patrón x_0 y la medida real u_0 proporcionada por el aparato sería $u_0=f(x_0)$.

Si el patrón tuviera un error $\eta(x_0)$, y no hubiera error en las medidas, se obtendría otra cantidad z_0, relacionada con el patrón x_0 por la ecuación

$$z_0=f(x_0+\eta(x_0))=u_0+\varepsilon_1(\eta(x_0)) \quad (5\text{-}1)$$

Si al medir se cometen, además, otros tipos de error se obtiene:

$$y_0=z_0+\varepsilon_2+\varepsilon_3+\ldots+\varepsilon_k = f(x_0+\eta(x_0))+\varepsilon_2+\varepsilon_3+\ldots+\varepsilon_k =$$
$$= u_0+\varepsilon_1(\eta(x_0))+\varepsilon_2+\varepsilon_3+\ldots+\varepsilon_k = u_0+\varepsilon_1(\eta(x_0)) + \varepsilon_2^* \quad (5\text{-}2)$$

donde $\varepsilon_2^* = \varepsilon_2+\varepsilon_3+\ldots+\varepsilon_k$

El objetivo es hallar con precisión x_0, la cantidad, proporción ó concentración exacta del gas de interés, oxígeno, en la botella. Puesto que la función f es la función de un cambio de escala, debe ser biunívoca. Por tanto si fuera conocida f, la cantidad x_0 se calcularía por la fórmula

$$x_0 = f^{-1}(y_0-(\varepsilon_2+\varepsilon_3+\ldots+\varepsilon_k)) - (\eta(x_0) = f^{-1}(y_0-\varepsilon_2^*) - (\eta(x_0) \quad (5\text{-}3)$$

Esta es la fórmula clave para calcular ó estimar los errores.

En la practica f suele ser desconocida y se suele sustituir por una función sencilla g, que en muchas ocasiones es lineal. En tal caso, conocida g, los errores se calcularían de la fórmula

$$x_0 = f^{-1}(y_0-\varepsilon_2^*) - (\eta(x_0) = g^{-1}(Y_0-\varepsilon_2^*) - (\eta(x_0) + f^{-1}(y_0-\varepsilon_2^*) - g^{-1}(y_0-\varepsilon_2^*) \quad (5\text{-}4)$$

Si además g es desconocida, entonces debe ser estimada por h; y, en este caso, la fórmula clave para el cálculo de los errores sería

$$x_0 = h^{-1}(y_0-\varepsilon_2^*) - (\eta(x_0) + f^{-1}(y_0-\varepsilon_2^*) - g^{-1}(y_0-\varepsilon_2^*) + g^{-1}(y_0-\varepsilon_2^*) - h^{-1}(y_0-\varepsilon_2^*) \quad (5\text{-}5)$$

La fórmula anterior es clave para el cálculo de errores en calibración.

6 CALIBRACIÓN MULTIVARIABLE.

Cuando se miden más de una variable respuesta y (o) más de una variable control x, el problema que consiste en determinar los valores de x en función de los valores observados de y se denomina calibración multivariable.

Por ejemplo, la calibración se necesita cuando un método de medida precisa; pero que es largo, caro o laborioso se quiere remplazar por una técnica de medida que aunque menos precisa, es más rápida de obtener, más barata o ambas cosas a la vez. Esto sucede en las industrias químicas o de alimentación, donde se quiere estimar la composición de un determinado producto. El método preciso tradicional conlleva química húmeda y se sustituye por una técnica de medida de las absorvancias cerca del infrarrojo. En general se dispone de un conjunto de n Unidades Experimentales, el conjunto de calibración, sobre el que se miden las absorvancias en q diferentes frecuencias de luz infrarroja, conjuntamente con determinaciones precisas de las proporciones de las sustancias químicas presentes.

Usando los datos de entrenamiento se halla un predictor de las proporciones de las componentes químicas presentes en otras n* Unidades Experimentales diferentes, predicción que se debe realizar con el predictor evaluado en las observancias observadas de luz infrarroja. Brown (1982) distingue dos casos:

(C1) Las proporciones de sustancias químicas en el conjunto de calibración son estrictamente medidas y controladas y forman la muestra de un Diseño de Experimentos Óptimo. Este caso se denomina Calibración Controlada; y

(C2) La proporción de componentes químicas sólo se puede modelizar de forma aleatoria, tanto en el conjunto de calibración, como en el muestral. En este supuesto, la calibración se denomina aleatoria.

En calibración controlada es más apropiado hacer la regresión de las absorvancias sobre las proporciones de distintos productos en el objeto alimenticio o químico, e invertir la relación para hacer la predicción de la calibración; mientras que en calibración aleatoria puede que sea más apropiada realizar la regresión de las proporciones sobre las absorvancias. Brown (1989) ha demostrado que estas distinciones son borrosas en la práctica; ya que en muchos problemas espectroscópicos reales sucede que el número de Unidades Experimentales controladas es muy pequeño cuando se le compara con el número de variables. En estos casos, la calibración controlada por mínimos cuadrados generalizados y la calibración aleatoria por mínimos cuadrados son procedimientos idénticamente indeterminados

Se denota por y_{ij} la Absorvancia en la j-ésima frecuencia infrarroja en la Unidad Experimental i-ésima y por x_{ik} la proporción de la k-ésima componente del producto químico

en la i-ésima muestra. Si se denota por Y la matriz de absorvancias, de dimensión n×q y por X la matriz de proporciones de las componentes; el modelo de regresión lineal multivariante tradicional para la regresión de las absorvancias sobre las proporciones de componentes es

$$Y=XB+E \qquad (6\text{-}1)$$

donde B, de dimensión p×q es la matriz de los coeficientes de regresión que son desconocidos y $E = (\varepsilon_1;\varepsilon_2;\ldots,\varepsilon_j;\ldots;\varepsilon_n)^T$ es la matriz de errores. Cada ε_j es un vector de dimensión q; de tal forma que $E(\varepsilon_j)=0$, $E(\varepsilon_j,\varepsilon_j^T)=\Sigma$; para j=1,2,...,n. y $E(\varepsilon_i,\varepsilon_j^T)=0$; siendo una matriz de dimensión q×q.

El modelo análogo para la calibración es

$$Y^*=X^*B+E^* \qquad (6\text{-}2);$$

que satisface las mismas hipótesis que el modelo (5-1). La estimación de X^* por los mínimos cuadrados generalizados se obtiene por la ecuación

$$\hat{X}^* = Y^* \hat{\Sigma}^{-1} \hat{B}^T (\hat{B} \hat{\Sigma}^{-1} \hat{B}^T)^{-1} \qquad (6\text{-}3)$$

que se obtiene invirtiendo la estimación de los mínimos cuadrados generalizados aplicada al modelo (5-1).

Un modelo alternativo es

$$X=YD+F \qquad (6\text{-}4)$$

para el conjunto de calibración controlada; siendo $X=YD+F^*$ para la verdadera calibración. La matriz de errores se ajusta al modelo precedente.

Con este modelo se verifica:

$$X\hat{X} = Y^* \hat{D} \quad \text{donde} \quad \hat{D} = (Y^T Y)^{-1} Y^T X \qquad (6\text{-}5)$$

Estas estimaciones son independientes de la matriz de covarianza de los errores.

En Denham y Brown(1993) los modelos previos se simplifican suponiendo:

(1) que los errores están relacionados por un modelo autorregresivo, en cuyo caso se reduce su dimensión.

(2) Se utilizan Splines para la estimación, en el supuesto que tanto las absorvancias Y como la matriz de coeficientes de regresión son funciones continuas conocidas, salvo algún parámetro, de las frecuencias.

(3) Se usan técnicas de pseudo inversas, (Moore Penrose) de mínima longitud.

En Tormod Naes (1985) se restringe la variabilidad de los errores, cambiando el modelo de errores por la ecuación E=Pt+e; siendo incorrelados los errores de la matriz e; así como los errores f; y P una matriz numérica de pocas columnas.

En Brown (1982), además de comparar los resultados de la regresión de Y sobre X, con los de X sobre Y, y concluir que en la mayoría de los ejemplos prácticos ambos modelos son muy indeterminados por falta de información se analizan los datos por métodos bayesianos, despues de introducir probabilidades a priori sobre el espacio de parámetros.

De todos los artículos comentados se concluye que, a pesar de la simplicidad del modelo lineal, en la mayoría de los casos reales faltan datos para estimar los parámetros de dicho modelo. La falta de datos es exagerada cuando se quiere contrastar la validez del modelo lineal. Otra dificultad que suele aparecer en la calibración multivariable es la falta de biunicidad entre los valores de las variables X e Y; hecho que no sucedía en el caso univariante.

7 MODELOS PRÁCTICOS DE CALIBRACIÓN.

En primer lugar se desarrolla un modelo univariante sobre la concentración de un cierto gas en una bombona, formada por mezcla de gases. Para ello, se tomó una muestra de 45 Unidades Experimentales en las que se midió, con precisión, la concentración del gas; así como las absorvancias a una determinada frecuencia. Los datos se recogen en la tabla 7.1; de la forma siguiente:

En las filas números 1, 5, 9, 13 y 17; se recogen las medidas de las concentraciones de las 45 Unidades Experimentales; en las filas 2, 6, 10, 14 y 18 se exponen las predicciones con las medias de 49 parámetros estimados con 49 regresiones obtenidas con la selección aleatoria del 75 % de datos; en las filas 3, 7, 11, 15 y 19 se presentan las predicciones realizadas con los parámetros de la regresión obtenidos con los 45 datos; y finalmente en las filas números 4, 8, 12, 16 y 20; se escriben las absorvancias. Se observa que no hay diferencias prácticas entre las predicciones realizadas con ambos métodos.

Se supone que las absorvancias se medían con un aparato que estaba bien calibrado. Por experiencias previas se sabe que el modelo de regresión lineal se ajustaba bien. Las predicciones se realizan por la fórmula de la predicción media.
En la tabla 7-2 se recogen las concentraciones y los intervalos de confianza estimados para estas concentraciones. Las concentraciones son las de las filas 1, 4, 7, 10 y 13; los extremos inferiores de los intervalos se escriben en las filas 2, 5, 8, 11 y 14; y los extremos superiores en las filas 3, 6, 9, 12 y 15.

TABLA 7.1. TABLA DE CONCENTRACIONES, PREDICCIONES Y ABSORVANCIAS.

F1	30.40	41.80	44.10	42.70	38.70	45.40	39.10	36.40	40.70
F2	31.32	40.92	44.10	42.11	38.07	45.16	38.41	36.66	40.68
F3	31.33	40.93	44.12	42.12	38.08	45.17	38.41	36.77	40.69
F4	1.1398	1.6022	1.7321	1.6618	1.4597	1.9243	1.4674	1.4142	1.5203
F5	39.90	41.60	43.10	43.30	35.90	40.70	37.60	36.30	48.70
F6	39.18	43.26	44.51	43.57	37.35	39.31	37.29	36.91	48.48
F7	39.18	43.27	44.51	43.57	37.36	39.32	37.30	36.91	48.48
F8	1.4750	1.5692	1.7115	1.7145	1.3636	1.5116	1.4205	1.4013	2.0132
F9	35.90	44.80	39.60	41.60	36.00	40.40	37.10	48.20	48.90
F10	35.95	44.32	38.35	40.38	36.72	38.45	37.22	48.72	49.95
F11	35.95	44.32	38.36	40.39	36.73	38.46	37.23	48.73	49.95
F12	1.3534	1.7721	1.4735	1.5752	1.3922	1.4787	1.4172	1.9922	2.0535
F13	40.80	44.80	45.20	31.60	42.30	44.50	39.40	43.30	46.80
F14	40.23	44.66	45.98	33.15	44.00	44.51	38.31	43.20	48.12
F15	40.23	44.67	45.98	33.16	44.00	44.51	38.32	43.20	48.12
F16	1.5675	1.7892	1.9050	1.2834	1.2760	1.7614	1.4715	1.716	1.962
F17	38.60	43.60	41.80	43.00	43.30	46.30	41.60	40.70	42.70
F18	37.33	43.30	41.11	42.38	44.83	47.31	40.88	40.00	42.46
F19	37.34	43.31	41.11	42.38	44.83	47.32	40.89	40.00	42.46
F20	1.4255	1.7212	1.6114	1.6750	1.2976	1.9518	1.6003	1.5557	1.279

TABLA 7.2. TABLA DE CONCENTRACIONES, PREDICCIONES Y ABSORVANCIAS.

F1	30.40	41.80	44.10	42.70	38.70	45.40	39.10	36.40	40.70
F2	30.27	40.10	43.30	41.31	37.21	44.31	37.56	35.88	39.87
F3	32.75	41.83	44.98	42.93	39.07	46.06	39.38	37.84	41.53
F4	39.90	44.70	43.10	43.30	35.90	40.70	37.60	36.30	48.70
F5	38.35	42.45	43.69	42.76	36.48	38.48	36.42	36.03	47.48
F6	40.10	44.10	45.39	44.41	38.39	40.22	38.33	37.97	49.52
F7	35.90	44.80	39.60	41.60	36.00	40.40	37.10	48.20	48.90
F8	35.04	43.50	37.50	39.56	35.84	37.61	36.35	47.70	48.86
F9	37.07	45.19	39.33	41.23	37.80	39.42	38.27	49.77	51.04
F10	40.80	44.80	45.20	31.60	42.30	44.50	39.40	43.30	46.80
F11	39.41	43.83	45.10	32.17	43.18	43.68	37.46	42.39	47.13
F12	41.09	45.54	46.91	34.46	44.86	45.38	39.29	44.04	49.14
F13	38.60	43.60	41.80	43.00	43.30	46.30	41.60	40.70	42.70
F14	36.46	42.49	40.29	41.57	44.00	46.37	40.07	39.17	41.65
F15	38.37	44.14	41.94	43.20	45.72	48.31	41.72	40.87	43.285

Se observa que hay 15 concentraciones que caen fuera del correspondiente intervalo de confianza; hecho que prueba que la regresión lineal no se ajusta bien a los datos; ya que los intervalos de confianza se construyeron con un nivel de significación de 0.01; por la fórmula habitual.

Teniendo en cuenta estas observaciones se formuló el siguiente modelo

$$y = a + bx + c(Pun-750)/1499 + \varepsilon \qquad (7\text{-II})$$

Se aplicó el modelo de regresión lineal a 1499 ejemplos, donde los datos se obtuvieron por simulación del conjunto de las 45 Unidades Experimentales iniciales. Se eligió una probabilidad de 0.75 para la pertenencia a cada ejemplo de cada Unidad Experimental; y con cada ejemplo se estimaron los valores de las 45 concentraciones. El valor de la función Pun(j) es el número de ejemplos en que la estimación por el modelo de la concentración es inferior a la verdadera concentración. Los 45 valores de Pun(j) aparecen en las filas 2,6,10,14 y 18 de la tabla 7. En las filas 1,5,9,13 y 17 se colocan las verdaderas concentraciones. En las filas 3,7,11,15 y 19 aparecen los valores de los extremos inferiores de los intervalos de confianza y, finalmente en las filas 4,8,12,16 y 20 los extremos superiores de los citados intervalos de confianza. El valor estimado de c es de 0.08. Si se examinan los intervalos de confianza de la tabla 7.3, se observa que tan sólo la concentración de una Unidad Experimental no pertenece al intervalo de confianza; aunque estos intervalos de confianza están calculados para un nivel de significación de 0.05. Teniendo en cuenta lo dicho se concluye que el model (7-II) es adecuado para la calibración.

TABLA 7.3. TABLA DE CONCENTRACIONES Y PREDICCIONES.

F1	30.40	31.60	35.90	35.90	36.00	36.30	36.40	37.10	37.60
F2	0	0	0	603	0	0	1	219	1461
F3	29.50	31.40	35.73	34.87	35.08	35.27	35.12	35.81	37.06
F4	32.02	33.74	37.67	36.87	37.08	37.25	37.11	37.74	38.89
F5	38.60	38.70	39.10	39.40	39.60	39.90	40.40	40.70	40.70
F6	1499	1499	1499	1499	1499	1499	1499	856	1499
F7	37.13	37.89	38.23	38.13	38.17	39.02	38.28	39.93	39.15
F8	38.96	39.67	39.98	39.89	39.93	40.71	40.02	41.58	40.83
F9	40.70	40.80	41.60	41.60	41.60	41.80	41.80	42.30	42.70
F10	1499	1499	0	1499	1499	1499	1499	0	1499
F11	39.84	40.07	41,72	40.23	40.74	40.78	40.96	42.46	41.97
F12	41.50	41.72	43.38	41.88	42.39	42.42	42.61	44.13	43.63
F13	42.70	43.00	43.10	43.30	43.30	43.30	43.60	44.10	44.50
F14	1499	1499	0	1	1326	0	1495	671	771
F15	42.31	42.23	42.96	42.03	42.89	43.28	43.15	43.20	43.66
F16	43.98	43.91	44.66	43.69	44.59	44.99	44.86	44.91	45.40
F17	44.80	44.80	45.20	45.40	46.30	46.80	48.20	48.70	48.90
F18	1499	1364	0	1458	0	0	6	1352	0
F19	44.14	44.35	44.39	44.92	45.66	46.42	46.99	47.95	48.15
F20	45.91	46.14	46.18	46.76	47.58	48.42	49.06	50.12	50.33

EJEMPLO DE CALIBRACIÓN MULTIVARIABLE

En la tabla 7-4 aparecen los valores de cuatro variables. En las filas 1,5, y 9 aparecen los porcentajes de agua de deteminadas sustancias; en las filas 2,6 y 10 aparecen los porcentajes de proteina de las mismas sustancias; en las filas 3,7 y 11 aparecen las reflectancias para una determinada longitud de onda y, finalmente; en las filas 4,8 y 12 aparecen las reflectancias para otra longitud de onda.

Para el análisis de estos datos, se dividió cada variable en nueve intervalos, asignando a los valores de cada variable perteneciente a cada uno de los intervalos números enteros correlativos del 1 al 9. Con esta transformación de las variables se obtuvo la tabla de contingencia múltiple 7-5.

TABLA 7-4. TABLA DE PORCENTAJES Y REFLECTANCIAS.

F1	9.00	9.94	9.92	9.06	10.02	10.06	9.95	9.12	9.88
F2	10.13	9.16	9.09	10.64	9.27	9.22	9.23	10.21	9.12
F3	361	361	362	362	362	363	363	360	362
F4	108	107	109	110	111	113	113	106	110
F5	9.03	9.81	9.90	9.89	10.12	10.04	10.00	9.16	9.98
F6	10.21	10.99	9.13	9.15	9.03	9.03	9.02	10.16	9.02
F7	360	351	353	352	354	353	352	363	363
F8	108	95	91	92	96	98	93	113	111
F9	9.88	9.94	10.12	10.11	9.99	9.87	10.3	10.1	
F10	9.06	10.85	9.04	9.01	9.12	9.18	9.10	9.06	
F11	364	350	363	364	364	362	353	351	
F12	111	90	112	113	113	109	99	100	

TABLA 7-5. TABLA DE CONTINGENCIA MÚLTIPLE.

% AGUA	1	1	1	1	2	6	7	7	7	7	7	7	7	8	8	8	8	9
% PROTEI	6	6	6	8	6	9	1	1	1	1	1	9	1	1	1	1	2	1
REFLECT1	7	7	8	8	9	1	2	2	8	8	9	1	1	2	3	9	8	2
REFLECT2	7	8	8	8	9	2	1	2	7	8	9	1	4	4	3	9	9	4
Frecuencia	1	1	1	1	1	1	2	1	1	3	4	1	1	1	1	3	1	1

En la tabla 7-6 se escriben los valores mínimo, máximo y longitud del intervalo correspondiente acada uno de los valores de las variables discretas:

TABLA 7-6. TABLA DE EQUIVALENCIAS.

% AGUA	Minimo	Máxim	INC	% PROTEIN	Mínimo	Máxim	INC
	9	10.3	0.14		9.01	10.99	0.22
Reflectancia 1	Minimo	Máxim	INC	Reflectanc 2	Mínimo	Máxim	INC
	350	364	1.56		90	113	2.56

Si se examina la tabla 7-5 se observa que las variables 3 y 4 de reflectancias para dos frecuencias diferentes son fuertemente dependientes, por lo que sólo una de ellas es suficiente para predecir el % de agua y el % de proteina. Esta predección se refleja en la tabla 7-7.

TABLA 7-7. PREDICCIÓN.

REFLECTANCIA 1	1	1	1	2	2	2	3	7	8	8	8	8	9	9	9
% AGUA	6	7	6	7	8	9	1	1	1	1	7	8	2	7	8
% PROTEINA	9	9	1	1	1	1	8	6	6	8	1	2	6	1	1
FRECUENCIA	1	1	1	3	1	1	1	2	1	1	4	1	1	4	3

Si se examina la tabla 7-7 se observa que el % de Proteina se predice mejor que el % de Agua; aunque la reflectancia no es un predictor preciso de ninguna de las otras dos variables.

8 BIBLIOGRAFIA.

(1) P.J. Brown (1982) *"Multivariate Calibration"* Jour. Royal Statis. Soc. B. Vol 44, N° 3 pp 287-321.

(2) Charles Hagwood (1992) *" The calibration Problem as an ill-posed inverse problem" Jour. Of Statistical Planning and Inference"* Vol 31 pp 179-185.

(3) M.C.Denham and P.J.Brown (1993) *"Calibratión with many Variables"* Appl. Statis. Vol 42 No 3 pp 515-528

(4) Phillip J. Brown and Rolf Sundberg (1987) *"Confident and Conflict in Multivariate Calibratión"*. Jour. Royal Statis. Soc. B. Vol 49, N° 1 pp 46-57.

(5) Robert W. Mee and Keith R. Eberhardt. (1996). *"A Comparison of UncertaintyCriteria for Calibratión"*. Technometrics, Vol 38, No 3, pp 221-237

(6) Sánchez G, M; Frutos C, G y Cuesta A., P.(1996) "Estadística y Matemáticas Aplicadas" Síntesis.

(7) Sheing Chung Chow and Jun Shao (1990) *" On the Difference Between the Classical and Inverse Methods of Calibration"* Appl. Statis. Vol 39 No 2 pp 219-228

(8) Tormod Naes (1985) *"Multivariate Calibration When the Error Covariance Matrix is Structured"* Technometrics , Vol 27, No 3, pp 301-311.

(9) Wei –Yin Loh (1987) *" Calibrating Confidence Coefficients"* JASA.Vol 82, N° 397 pp 155-170.