CALIBRADO Y REGRESIÓN

INTRODUCCIÓN

Existen muchas situaciones en las que hay una relación entre dos variables asociadas. P.e. en Análisis Instrumental la respuesta está relacionada con la concentración de analito(s). La relación se estudia mediante un Análisis de regresión y consiste en una función matemática que puede ser utilizada para predecir una variable a a partir de las otras.

Se denominan técnicas de regresión Modelo I a las que estudian la dependencia de una variable aleatoria (variable dependiente o respuesta) en función de una variable controlada por el experimentador (variable independiente o de predicción). Se supone que la variable(s) independiente no está sujeta a error. La calibración es su principal aplicación analítica. Si ambas variables están sujetas a error se usan técnicas de regresión de Modelo II.

En esta Lección vamos a desarrollar los métodos de ajuste lineales de una respuesta a una sola variable independiente (regresión lineal simple presentando las principales aplicaciones analíticas.

MÉTODO DE LOS MÍNIMOS CUADRADOS

Estimación de los parámetros

En calibración, la y (variable independiente) es la respuesta del instrumento, mientras que la x es la concentración de los patrones, que debe estar libre de error. La verdadera relación matemática será:

$$\eta = \beta_0 + \beta_1 x$$

η es la respuesta verdadera, β_0 y β_1 los parámetros del modelo (abscisa y ordenada). Para cualquier concentración x_i, no podemos conocer η_i sino solo su valor experimental y_i, que al estar sujeto a error diferirá del verdadero:

$$y_i = \eta_i + \varepsilon_i \quad \text{ó} \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Los parámetros del modelo β_0 y β_1 son desconocidos, pero se pueden utilizar la información de las medidas para estimarlos mediante b_0 y b_1 respectivamente. Estas estimaciones se calculan de manera que los puntos calculados mediante esa línea estimada:

$$\hat{y} = b_0 + b_1 x$$

se ajusten lo más posible a los puntos experimentales.

La línea estimada se denomina línea de mínimos cuadrados si esa estimación se hace por el método de los mínimos cuadrados, en el cual se minimiza la suma de los cuadrados de los residuales:

$$e_i = y_i - \hat{y}_i$$

es decir se minimiza:

$$R = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - b_o - b_1 x_i)^2$$

Los cálculos se basan en derivar R frente a b_0 y b_1 e igualar a cero. De esa manera se obtienen las ecuaciones normales

$$\sum_i y_i - n b_0 - b_1 \sum_i x_i = 0$$

$$\sum_i x_i y_i - b_0 \sum_i x_i - b_1 \sum_i x_i^2 = 0$$

de donde se obtienen b_0 y b_1

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad ; \quad \bar{y} = (\sum_i y_i)/n \quad ; \quad \bar{x} = (\sum_i x_i)/n$$

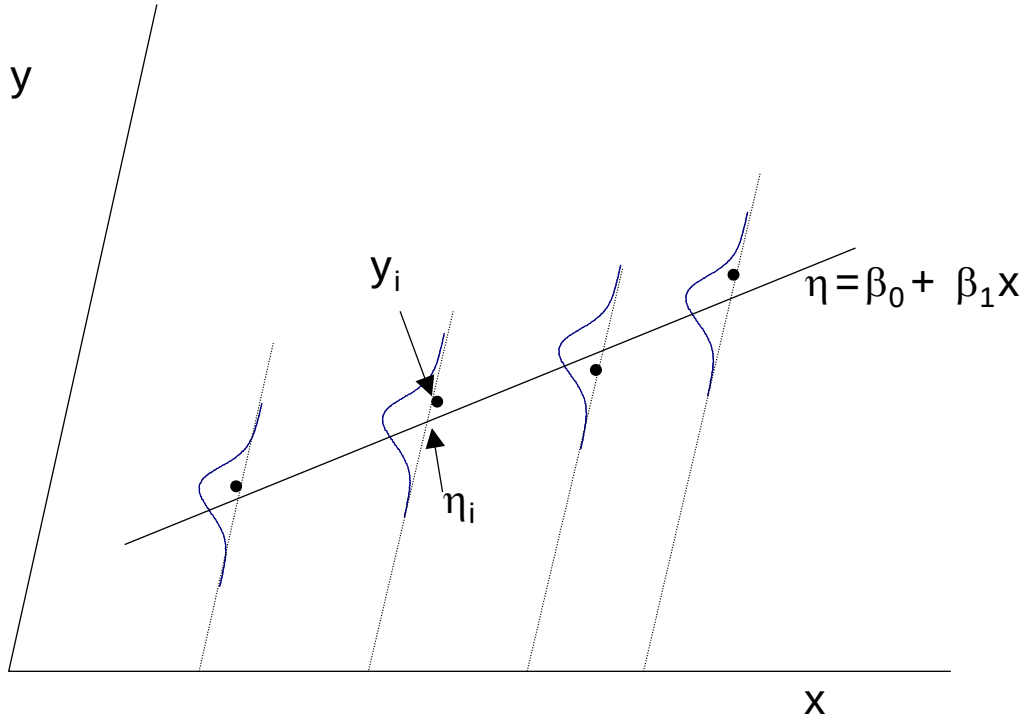
La varianza residual s_e^2, se determina mediante:

$$s_e^2 = \frac{\sum_i e_i^2}{n-2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$$

Este valor representa la **varianza que no puede ser explicada** por la línea de regresión. A veces se la representa como $s_{y/x}^2$. Si el modelo es correcto, $s_e^2 = s_{y/x}^2$ es una estimación de la varianza de las medidas σ^2 o puro error experimental.

El método de mínimos cuadrados hace las siguientes **suposiciones:**

1) Para cada valor de x_i los residuales vienen de una población normalmente distribuida, con media cero
2) Los e_i son independientes
3) Todos los e_i tienen la misma varianza σ^2 : La variable y siempre es observada con la misma precisión. En calibrado eso significa que la precisión de las respuestas es independiente de la concentración. Esta condición es la homocedasticidad. Se comprueba mediante medidas replicadas y en caso negativo obliga a la utilización de pesos estadísticos



Etapas de la regresión

1) Selección de un modelo
2) Establecimiento del diseño experimental
3) Estimación de los parámetros del modelo
4) Validación del modelo
5) Determinación de los intervalos de confianza

Debido a que nuestro principal interés es el calibrado, las etapas 1) y 2) no serán discutidas. La 1) porque hemos elegido una línea recta como modelo por ser el más adecuado ya que la Sensibilidad debe ser constante en todo el intervalo. El 2) sirve para ver como se reparte la variable x (concentración) sobre su dominio. Se tratará sobre la marcha. El 3), la estimación de los parámetros ya se ha hecho anteriormente, así que comenzaremos por el apartado 4)

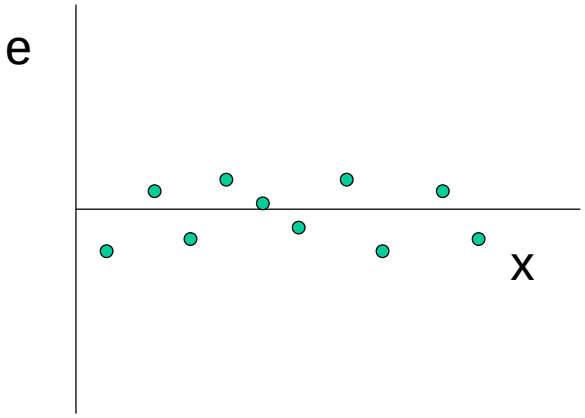
Validación del modelo

Debe verificarse a) que el modelo elegido describe adecuadamente la relación entre las dos variables (no hay falta de ajuste) y b) comprobar la suposición de que los residuales son normales con media cero y con varianza constante (N(0, σ^2).

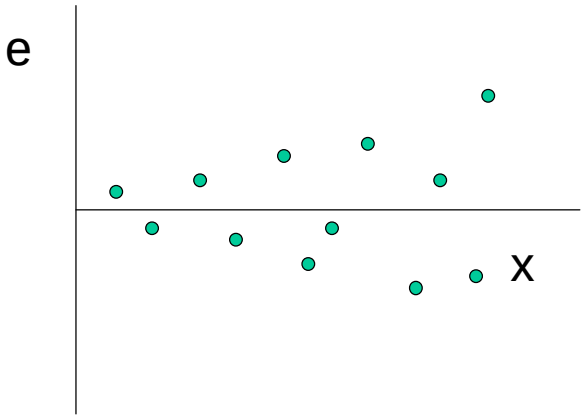
Análisis de los residuales

Los residuales $e_i = y_i - \hat{y}_i$ dan información valiosa acerca de las suposiciones hechas así como sobre la falta de ajuste. Su normalidad puede verificarse mediante una prueba χ^2 o un gráfico rankit, pero habitualmente no se dispone del número suficiente de medidas para cada x_i.

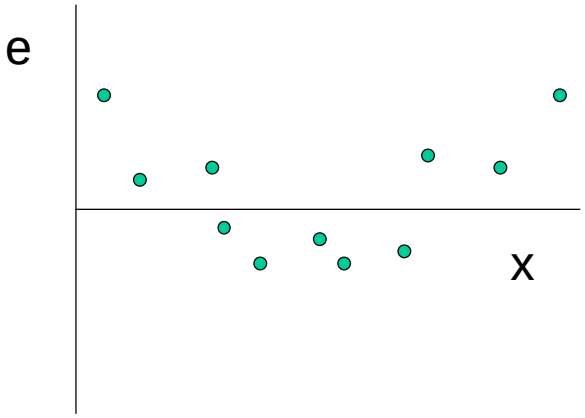
Lo más útil es un gráfico de residuales, o sea una representación de e_i en función de x_i. Su examen precisa experiencia previa, pero hay varios comportamientos típicos:



Los residuales aparecen esparcidos a lo largo del eje x de forma aleatoria y sin tendencias visibles sistemáticas. Esta es el comportamiento a esperar si el modelo es válido



No se cumple la condición de homocedasticidad, ya que aumenta la dispersión a lo largo del eje x



Existe un comportamiento anómalo que indica que el modelo subyacente no es una línea recta y que probablemente la adición de un término cuadrático mejoraría el ajuste

Análisis de varianza

Se pueden emplear las técnicas del **ANOVA**. Supongamos que se tiene una línea de calibrado pero con medidas **replicadas**, es decir cada punto de la línea de calibrado repite más de una vez. En el ejemplo siguiente se tienen 11 puntos experimentales, correspondientes a k=6 situaciones diferentes (solo se duplicaron 5 concentraciones)

.

C_i (mg/l)	0	0,5	1,0	1,5	2,0	2,5	3,0
y_{ij} (Absorb.)	0,0054	0,0823	0,1529	0,2129	0,2742	0,3133	0,3607
	0,0080	0,0842	0,1488	0,2064	0,2698	0,3179	0,3641
n_i	2	2	2	2	2	2	2
$\bar{y}_i$	0,0067	0,0833	0,1509	0,2097	0,2720	0,3156	0,3624
$\hat{y}_i$	0,023	0,082	0,141	0,200	0,259	0,318	0,377

k=7 $\quad \hat{y} = 0{,}0230 + 0{,}1181\,x \quad \bar{y} = 0{,}2001$

$n = \sum_i n_i = 14$

Se puede definir la variación total de los datos $\mathbf{SS_T}$ como:

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{nk}(y_{ij} - \bar{y})^2$$

La distancia de cada valor individual y_{ij} a la gran media (o centroide) $\bar{y}$ se puede dividir en partes:

$$(y_{ij} - \bar{y}) = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

siendo $\bar{y}_i$ la media de las determinaciones experimentales correspondiente a cada valor de x, e $\hat{y}_i$ el valor predicho por el modelo para ese valor de x.

Si se eleva al cuadrado la anterior expresión y se hace un sumatorio sobre i y j, los términos cruzados se anulan, y se demuestra que:

$$\sum_{i=1}^{k}\sum_{j=1}^{nk}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{nk}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^{k} n_i(\hat{y}_i - \bar{y})^2$$

$$SS_T = \underbrace{SS_{PE} + SS_{FDA}}_{SS_{RES}} + SS_{REG}$$

$$SS_T = SS_{RES} + SS_{REG}$$

$\mathbf{SS_{REG}}$ es la variación que es explicada por la línea de regresión
$\mathbf{SS_{RES}}$ es la variación residual (no es explicable por la regresión). Se divide en dos componentes
- $\mathbf{SS_{PE}}$ que mide la variación debida al puro error experimental
- $\mathbf{SS_{FDA}}$ que mide la variación debida a la falta de ajuste, es decir a la elección de un modelo de regresión equivocado (la línea recta en nuestro caso)

Los grados de libertad se reparten:
$\mathbf{SS_T = SS_{PE} + SS_{FDA} + SS_{REG}}$
n-1 = (n - k)+(k - 2) +1
En general el reparto es:
- A $\mathbf{SS_{REG}}$ le corresponden **p - 1** siendo p el nº de parámetros del modelo (p=2)
- A $\mathbf{SS_{FDA}}$ le corresponden **k - p**
- A $\mathbf{SS_{PE}}$ le corresponden el resto o sea **n - k**

Si dividimos las SS por sus g.d.l. se obtienen las MS que son estimaciones de las varianzas del modelo. $\mathbf{MS_{PE}}$ es una estimación de σ^2, y $\mathbf{MS_{FDA}}$ es también una estimación de σ^2 si el modelo está bien elegido.

El test de falta de ajuste es una prueba F de una cola que compara el cociente $\mathbf{MS_{FDA}/MS_{PE}}$ con el F crítico con (k-2) y (n-k) g.d.l. Si la H_0 se rechaza entonces el modelo no

es adecuado, mientras que si se retiene sí que lo es, y $\mathbf{MS_{RES}} = s_e^2$ se puede utilizar como un estimador de σ^2.

La tabla del ANOVA que resulta es:

Fuente	**g.d.l**	**SS**	**MS**	**F**
Regresión	p-1 = 1	SS_{REG}	$SS_{REG}/1$	MS_{REG}/MS_{RES}
Residual	n-p = n-2	SS_{RES}	$SS_{RES}/(n-2)$	
Falta de ajuste	k-p=k-2	SS_{FDA}	$SS_{FDA}/(k-2)$	MS_{FDA}/MS_{PE}
Puro error	n-k	SS_{PE}	$SS_{PE}/(n-k)$	
Total	n-1	SS_T		

Es importante darse cuanta que MS_{FDA} y MS_{PE} solo pueden ser comparadas cuando hay **replicación** de al menos una experiencia, pues si n = k (es decir solo hay una experiencia de cada situación experimental) no tenemos grados de libertad suficientes para calcular MS_{PE}.

Otra prueba F que puede hacerse es comparar $\mathbf{MS_{REG}}$ con $\mathbf{MS_{RES}}$ que es lo mismo que comprobar la hipótesis nula que $\beta_1 = 0$. Este prueba no tiene mucho sentido en calibrado químico, ya que éste se basa por definición en que la respuesta del instrumento cambia con la concentración de los patrones.

El cociente $\mathbf{SS_{REG}/SS_T}$ se denomina **coeficiente de determinación múltiple $\mathbf{R^2}$** y es la proporción de varianza total explicada por el modelo. Varía entre cero, que implica que x no tiene efecto sobre y, y uno que indica que x explica perfectamente y. Su raíz cuadrada es el coeficiente de correlación múltiple. Cuando el modelo es una recta (p=2), R^2 se convierte en $\mathbf{r^2}$ y se denomina **coeficiente de determinación**. Su raíz cuadrada, corregida por el signo de la pendiente, se llama coeficiente de correlación y su importancia ha sido **sobrestimada** (Ver Regresión y Correlación).

Aplicando el ANOVA a los datos de nuestro ejemplo:

Fuente	**g.d.l**	**SS**	**MS**	**F**
Regresión	1	0,19516	0,19516	1337,37
Residual	12	0,00175	0,0001459	
Falta de ajuste	5	0,00169	0,000338	38,96
Puro error	7	0,00006	$8{,}68.10^{-6}$	
Total	13	0,19691		

El valor F_{crit} con 5 y 7 g.d.l. y α = 0,05 vale 3,97 por lo que la falta de ajuste es significativa, y el modelo de la línea recta no describe de forma adecuada la relación entre x e y.

Obsérvese que la otra prueba F que puede realizarse comprueba que el modelo sí que explica una cantidad apreciable de la varianza. De hecho, el valor de r = $\sqrt{SS_{REG}/SS_T} = \sqrt{0{,}19516/0{,}19691} = 0{,}9955$, y el de r^2 = 0,9911 por lo que el modelo elegido (a pesar de ser erróneo) explica un 99,11 % de los datos.

Heterocedasticidad y su solución

En rigor, si no se cumple la condición de homocedasticidad, no se puede aplicar el método de mínimos cuadrados anterior. El problema se evita transformando las variables o empleando pesos en los mínimos cuadrados

Transformación

Depende de la forma en que la varianza de y, $s_{y_i}^2$ varía con y_i. Los dos casos más habituales son:

a) Varianza proporcional a y, la ecuación a ajustar es $\sqrt{y} = b_0 + b_1\sqrt{x}$

b) Varianza proporcional a y^2, es decir la desviación típica relativa es constante, se debe ajustar $\log y = b_0 + b_1 \log x$

Mínimos cuadrados con pesos

Se introduce el factor de peso, inversamente proporcional a la varianza

$$w_i = 1 / s_{y_i}^2$$

de manera que se da más importancia a las observaciones más precisas: la línea de regresión pasa más cerca de los puntos más precisos que de los más imprecisos. Los valores de los parámetros son:

$$b_1 = \frac{\sum_i w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_i w_i (x_i - \bar{x}_w)^2}$$

$$b_0 = \bar{y}_w - b_1 \bar{x}_w \quad ; \quad \bar{y}_w = (\sum_i w_i y_i) / \sum_i w_i \; ; \quad \bar{x}_w = (\sum_i w_i x_i) / \sum_i w_i$$

INCERTIDUMBRE E TERVALOS DE CONFIANZA

Los intervalos de confianza de los parámetros se dan con un 100(1-α)% de confianza y se valen $\pm t_{\alpha/2;n-2} \bullet s_{\text{parámetro}}$. La diferencia principal con otros intervalos es que ahora t se busca con n-2 g.d.l. (una línea recta tiene 2 parámetros)

De los parámetros por separado

Las estimaciones de las desviaciones típicas respectivas son:

$$s_{b_o} = s_e \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}$$

$$s_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

y los intervalos son (con el 95 %)

Para $\beta_0 = b_0 \pm t_{0,025;n-2} . s_{b_0}$

Para $\beta_1 = b_1 \pm t_{0,025;n-2} . s_{b_1}$

Si cualquiera de los dos intervalos incluye a cero, se puede decir que dicho parámetro es igual a cero (Cf. Test de hipótesis)

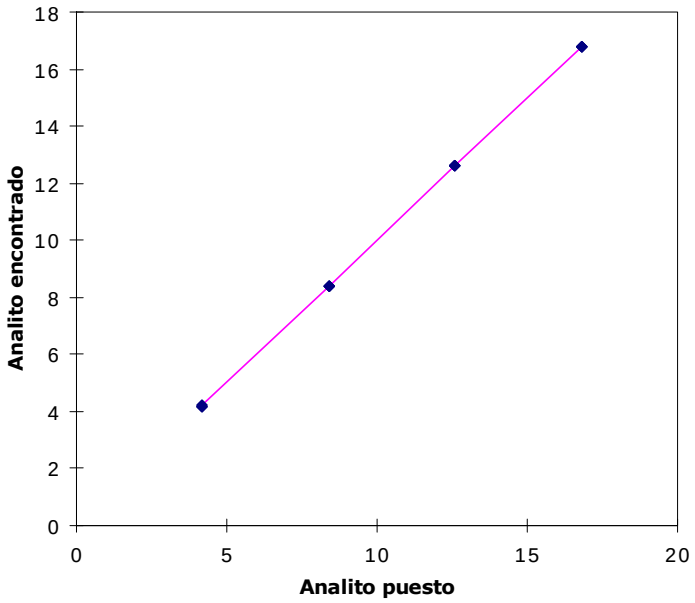
Una aplicación importante en Química Analítica es en la puesta a punto de un método. Este se valida analizando blancos dopados (spiked) con concentraciones diferentes de analito. Si se representa la concentración encontrada frente a la añadida, se debería obtener una línea de ordenada en el origen cero y pendiente uno. Si al calcular la regresión, los intervalos de confianza de b_0 y b_1 incluyen a esos valores, el método queda validado. Si la ordenada en el origen es mayor de cero existe un error sistemático constante (o una corrección incorrecta del blanco). Si la pendiente difiere e uno, existe un error sistemático proporcional que suele ser debido a la matriz.

Ejemplo

Para saber si un procedimiento analítico está libre de bias se llevan a cabo una serie de experiencias de recuperación (recovery) en las que se analizan muestras diferentes que contienen cantidades perfectamente conocidas del analito de interés. Determine si el procedimiento está o no libre de bias (con un nivel de significación α = 0,05)

Analito puesto	4,2	4,2	4,2	8,4	8,4	8,4	12,6	12,6	12,6	16,8	16,8	16,8
Analito encontrado	4,25	4,14	4,18	8,39	8,42	8,42	12,62	12,6	12,59	16,8	16,77	16,77

Al representar el analito encontrado frente al analito puesto y ajustar por mínimos cuadrados, se obtienen la Figura y resultados siguientes:



Parámetro	Valor	Desvi. típica	t	Intervalo	
Ordenada en el origen	0,005	0,02155226	2,228	-0,043	0,053
Pendiente	0,99912698	0,00187376	2,228	0,995	1,003

Como el intervalo de confianza de la ordenada en el origen incluye a cero, se puede mantener la H_0 de que dicha ordenada vale cero, luego no existen bias absolutos. El intervalo de confianza de la pendiente incluye a uno, luego se debe mantener la H_0 de que dicha pendiente vale uno, luego no existen bias relativos:

En conclusión, el método ensayado está libre de bias tanto absolutos como relativos.

De los parámetros en conjunto

Debido a la forma de obtenerlos, los valores de b_0 y b_1 no son independientes entre sí. Cuando se desea comprobar simultáneamente la hipótesis de que $\beta_0 = 0$ y $\beta_1 = 1$, debe utilizarse un test de hipótesis conjunto (joint hypothesis test), o bien construir una región conjunta de confianza que tenga en cuenta la correlación existente entre las estimaciones b_0 y b_1.

Esa región tiene forma de elipse y su fórmula es:

$$(\beta_0 - b_0)^2 + 2\bar{x}(\beta_0 - b_0)(\beta_1 - b_1) + (\sum x_i^2 / n)(\beta_1 - b_1)^2 = 2F_{\alpha;2,n-2}s_e^2 / n$$

La hipótesis conjunta implica calcular un valor de F

$$F = \frac{(\beta_0 - b_0) + 2\bar{x}(\beta_0 - b_0)(\beta_1 - b_1) + (\sum_i x_i^2 / n)((\beta_1 - b_1)^2}{2 s_e^2 / n}$$

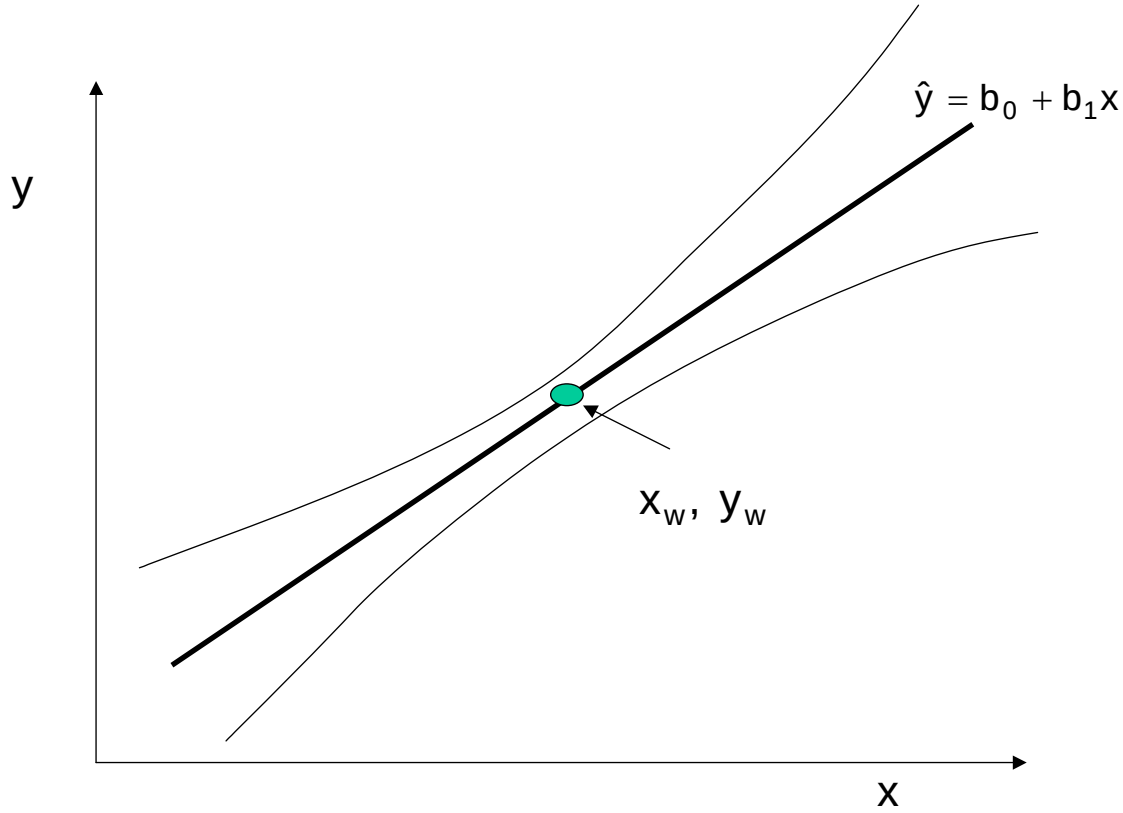
que se compara con F crítico con 2 y n-2 g.d.l. y el nivel de significación requerido.

De la respuesta predicha

Si $x = x_0$, se puede calcular $\hat{y}_0 = b_0 + b_1 x_0$. Ese valor predicho tiene un intervalo de confianza que es:

$$\hat{y}_0 \pm \sqrt{2F_{0,05;2,n-2}}\ s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

La representación gráfica es hiperbólica y la región dentro de las dos ramas se denomina banda de confianza de Working-Hotelling



El centroide $(\bar{x},\bar{y})$ (o en general (x_w,y_w)) es importante, ya que la línea de regresión pasa por él. Se observa que el intervalo de confianza es mínimo cuando $x_0 = \bar{x}$: la mayor precisión se obtiene en el centro de la línea de regresión. El intervalo también disminuye cuando $\sum(x_i-\bar{x})^2$ es muy grande, lo que obligaría en calibrado a utilizar patrones de muy baja y muy alta concentración lo cual no es aconsejable, ya que no se puede estimar la linealidad. El número de puntos también influye, pues t, 1/n y $\sum(x_i-\bar{x})^2$ dependen de n.

PREDICCIONES HECHAS SOBRE LA LÍNEA AJUSTADA

El objetivo último del análisis de regresión es utilizar la línea para predecir, bien valores de y (y de su incertidumbre asociada) a partir de un x, o como en el calibrado analítico, el predecir el valor de la x (y de su incertidumbre asociada) a partir de una y. Los correspondientes intervalos de confianza se denominan intervalos de regresión

Predicción de nuevas respuestas

Además de la incertidumbre de la regresión, estimada mediante $s_e^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$ tenemos la incertidumbre de la observación, que si el modelo es correcto, también es estimada por s_e^2. Ambas varianzas se pueden sumar, pues son independientes por lo que la varianza de una respuesta predicha a partir de un x_0 es

$$s_{\hat{y}_0}^2 = s_e^2 + s_e^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) = s_e^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

$$s_{\hat{y}_0} = s_e\sqrt{\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

y si se quiere estimar una media medida m veces, la $s_{\hat{y}_0}^2$ vale

$$s^2_{\hat{y}_0} = \frac{s^2_e}{m} + s^2_e\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right) = s^2_e\left(\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)$$

$$s_{\hat{y}_0} = s_e\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

y el intervalo del 95 % de confianza es $\hat{y}_0 \pm t_{0,025;n-2}\, s_{\hat{y}_0}$ aunque algunos textos indican n+m-3 g.d.l.

Interpolación de x a partir de la respuesta

Es la aplicación más importante en calibrado. A partir de m medidas hechas sobre una muestra $\bar{y}_s$ se predice $\hat{x}_s = \frac{\bar{y}_s - b_0}{b_1}$. La precisión de esa estimación depende no solo de la fiabilidad del modelo (a través de b_0 y b_1) sino también de la precisión de la Respuesta observada $\bar{y}_s$ que, si el modelo está bien elegido y explica correctamente la situación experimental planteada, puede asimilarse a s^2_e (homogeneidad de varianzas). No obstante, la determinación de la desviación típica o incertidumbre asociada es complicada y se acepta la siguiente **aproximación**

$$s_{\hat{x}_0} = \frac{s_e}{b_1}\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}_s - \bar{y})^2}{b_1^2\sum(x_i - \bar{x})^2}}$$

El intervalo de confianza es $\hat{x}_s \pm t_{0,025;n-2}\, s_{\hat{x}_0}$ aunque algunos textos utilizan n+m-3 g.d.l.

Los mismos comentarios hechos más arriba acerca de cómo estrechar el intervalo de confianza son ahora aplicables:

1) Cuanto mayores m y n mejor (efecto sobre t, 1/n, 1/m y $\sum(x_i - \bar{x})^2$.
2) El menor valor de $s_{\hat{x}_0}$ se obtiene cuando $\bar{y}_s = \bar{y}$ (o sea en la mitad de la línea de calibrado)
3) El intervalo también disminuye cuando $\sum(x_i - \bar{x})^2$ es muy grande, lo que obliga a utilizar patrones de muy baja y muy alta concentración lo cual no es aconsejable

Si la varianza de la medida de la muestra, s^2_s, es diferente de la varianza de los patrones de calibrado, la fórmula es ligeramente diferente:

$$s_{\hat{x}_0} = \frac{1}{b_1}\sqrt{\frac{s^2_s}{m} + s^2_e\left(\frac{1}{n} + \frac{(\bar{y}_s - \bar{y})^2}{b_1^2\sum(x_i - \bar{x})^2}\right)}$$

En el caso de datos heterocedásticos las fórmulas cambian:

$$s_e = \sqrt{\frac{\sum w_i(y_i - \hat{y}_i)^2}{n-2}}$$

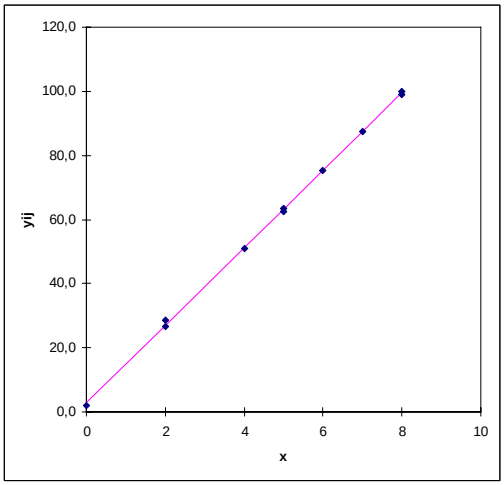
$$s_{\hat{x}_0} = \frac{s_e}{b_1}\sqrt{\frac{1}{w_s\, m} + \frac{1}{\sum w_i} + \frac{(\bar{y}_s - \bar{y}_w)^2\sum w_i}{b_1^2(\sum w_i\sum w_i x_i^2 - (\sum w_i x_i)^2)}}$$

Ejemplo

Se prepara una línea de calibrado a partir de los siguientes datos:

Concentración (mg/l)	0	2	4	5	6	7	8
Respuesta	2,0	28,7	51,1	63,3	75,4	87,6	99,0
		26,7		62,5			99,8

1) Determine los parámetros del calibrado (Utilice la macro de EXCEL).



Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,9997607
Coeficiente de determinación R^2	**0,9995214**
R^2 ajustado	0,9994616
Error típico s_e	**0,7598695**
Observaciones	10

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	9646,7498	9646,7498	16707,174	1,435E-14
Residuos	8	4,6192133	0,5774017		
Total	9	9651,369			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95,0%*
Intercepción	**2,8310136**	0,5007019	5,6540901	0,0004791	**1,6763922**	**3,985635**
x	**12,080635**	0,0934627	129,25623	1,435E-14	**11,86511**	**12,296161**

Se observa que el modelo explica prácticamente toda la información de la línea de calibrado

2) Calcule la incertidumbre con la que se obtiene la concentración de una muestra que se determina por triplicado y cuyas respuestas son:34,3 ; 37,5 y 36,4
La fórmula a utilizar es:

$$s_{\hat{x}_0} = \frac{s_e}{b_1}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(\bar{y}_s-\bar{y})^2}{b_1^2\sum(x_i-\bar{x})^2}}$$

La macro de EXCEL da directamente s_e y b1, pero el resto de los parámetros deben calcularse.

En este caso valen: $\bar{y}=59{,}6\,;\bar{y}_s=36{,}07;m=3;n=10;\sum(x_i-\bar{x})^2=66{,}1$

Por tanto, la incertidumbre vale $s_{\hat{x}_0}=0{,}0440654$.

Si el valor de $\hat{x}_0$ no es directamente el resultado de la determinación analítica sino que se obtienen mediante una operación algebraica en la que intervienen, además de $\hat{x}_0$ otras magnitudes (masa o volumen iniciales de muestra, factores de dilución, volúmenes de enrase….), la incertidumbre del resultado final se calculará teniendo en cuenta las incertidumbres de dichas magnitudes según las leyes de propagación de errores.

(Nota: repita los cálculos sin asumir que $s_s^2 = s_e^2$)

LÍMITE DE DETECCIÓN Y CONCEPTOS RELACIONADOS

Según la IUPAC es la concentración x_L o cantidad q_L derivada de la medida más pequeña y_L que puede ser detectadas con una certeza razonable con un procedimiento analítico dado.

$$y_L = \bar{y}_{bl} + k\, s_{bl}$$

$$x_L = k\, s_{bl} / S$$

donde S es la pendiente de la línea de calibrado (sensibilidad). La IUPAC recomendaba (de acuerdo con Kaiser) que k = 3.

En términos generales, el límite de detección es la concentración que origina una señal instrumental significativamente diferente de la del blanco. Hay diferentes versiones, pero todas incluso la de la IUPAC solo protegen contra errores α o de tipo I, pero no contra errores β o de tipo II. Hoy día incluso la IUPAC reconoce 3 versiones, que en rigor deben definirse en el dominio de la señal:

- *Límite de decisión*: señal a partir de la cual se puede decidir a posteriori si el resultado obtenido indica o no detección
- *Límite de detección* señal a partir de la cual se puede confiar a priori que el procedimiento analítico permitirá detectar
- *Límite de determinación o cuantificación* señal a partir de la cual un procedimiento analítico es capaz de dar un resultado con la suficiente precisión

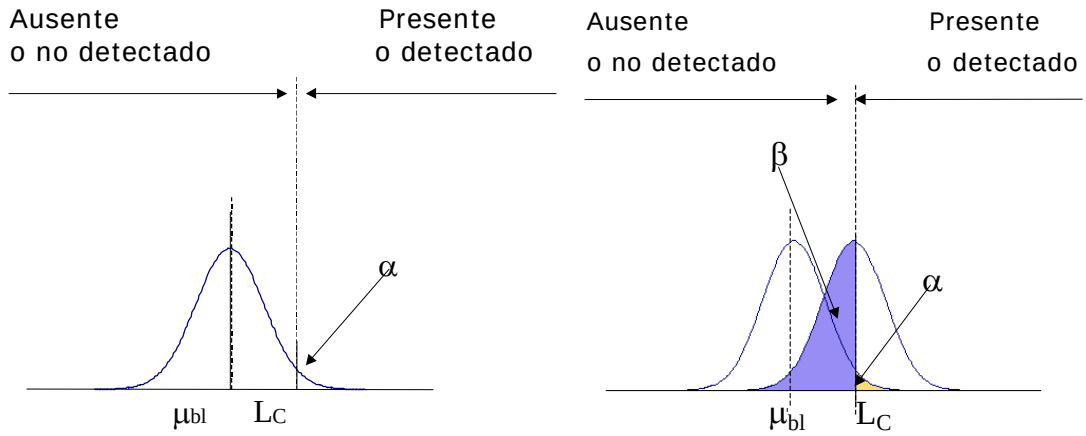
Límite de decisión CCα (Decisión 2002/657/CE

$$L_C = \mu_{bl} + k_C\, \sigma_{bl}$$

El valor de k_C depende del nivel de significación α. Este límite protege contra errores de tipo I (falsos positivos): Hay solo un α % de probabilidades de que una señal mayor o igual que L_C pertenezca al blanco, luego se puede concluir con una probabilidad (1-α) que el componente ha sido detectado. En estas condiciones el error β o de tipo II puede llegar a ser del 50%

Si k_C= 3 se tiene la primitiva versión de la IUPAC (límite de Kaiser). La IUPAC y la ISO prefieren hoy día un α del 5% lo cual se corresponde con k_C = 1,645

Este límite ha sido propuesto para tomar una decisión a posteriori, es decir después de que se ha medido la respuesta, sobre la presencia de un componente. Se le puede definir como nivel crítico o límite de decisión por encima del cual una señal puede ser reconocida como detectada.



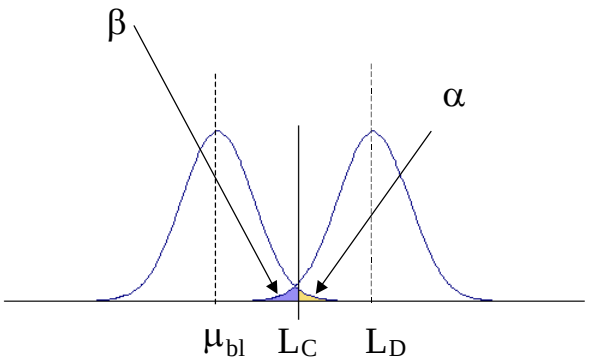
Límite de detección o Capacidad de detección CCβ (Decisión 2002/657/CE)

Para reducir el error de tipo II o β (falsos negativos) no queda más remedio que separar más las distribuciones del blanco y el analito.

$$L_D = L_C + k_D\, \sigma_{bl} = \mu_{bl} + k'_D\, \sigma_{bl} \quad \text{siendo} \quad k'_D = k_C + k_D$$

Para una muestra que contenga analito, y tenga una concentración que origine una respuesta L_D, solo el β% de las medidas estarán por debajo de L_C y serán asignadas al blanco. Por lo tanto, dado L_C, L_D protege contra falsos negativos.

Anteriormente, se admitía $k_C = k_D = 3$ con lo que $\alpha = \beta = 0{,}13$ % y además $L_D = \mu_{bl} + 6\ \sigma_{bl}$. Hoy día, la IUPAC y la ISO proponen $\alpha = \beta = 5$ %, por lo que el valor de k'_D es ahora de 2 * 1,645 = 3,29. Este límite puede ser usado a priori



Límite de cuantificación

Es el nivel al cual la precisión de la medida será satisfactoria para la determinación cuantitativa. Es decir es la concentración que se puede determinar con una desviación típica relativa (DTR) máxima fijada y una adecuada exactitud.

$$L_Q = \mu_{bl} + k_Q\ \sigma_{bl}$$

Si DTR es del 5%, $k_Q = 20$. La IUPAC propone un valor de 10, con lo que la DTR es del 10%. En la definición se supone que la σ en el límite de cuantificación es igual que σ_{bl} , pero eso debe comprobarse y suele utilizarse el valor de σ al nivel que se espera para L_Q. La utilidad de este límite de cuantificación está siendo cuestionada en la actualidad.

Estimación del Límite de detección

Es fundamental el conocer la estimación adecuada de μ_{bl} y σ_{bl}, y hay diferentes posibilidades: Un *blanco de reactivos o de disolvente* puede dar lugar a resultados demasiado optimistas. Debe emplearse un **blanco analítico** que contiene todos los reactivos y se analiza de la misma manera que las muestras. El blanco ideal sería un **blanco de matriz** que duplicaría exactamente la muestra excepto por el analito., aunque en ocasiones de utiliza una muestra real con una concentración de analito cercana a la esperada para el blanco. Deben hacerse al menos 10 réplicas del blanco en diferentes condiciones de precisión intermedia (días, operadores, equipamiento).

Si la determinación se hace con pocas réplicas los valores de k_C, k_D y k_Q deben sustituirse por valores de t con n-1 g.d.l. Otra posibilidad es estimar σ_B a partir de $s_{y/x}$ de una línea de calibrado.

1) **Si la corrección del blanco no es parte del proceso analítico**, L_D se calcula simplemente a partir de las expresiones anteriores:

$$L_C = \mu_{bl} + k_D\ \sigma_{bl} = \bar{y} + k_D s_{bl}$$

$$L_D = \mu_{bl} + k'_D\ \sigma_{bl} = \bar{y} + k'_D\ s_{bl}$$

siendo $\bar{y}$ y s_{bl} estimaciones experimentales de la respuesta del blanco y de su desviación típica. El factor k'_D debe ser al menos 2*1,65 =3,3 para que los errores α y β sean del 5%.

Para convertir L_D en unidades de concentración, se utiliza la pendiente de la línea de calibrado:

$$x_C = \frac{(L_D - \bar{y})}{b_1}$$

2) **Si la corrección del blanco es parte del proceso analítico**, hay dos posibilidades:
2.1) Si la respuesta medida **se corrige previamente** de la del blanco, es decir con cada muestra (o grupo de muestras) se analiza un blanco, y la respuesta de cada muestra (o de todas las del grupo) es corregida del blanco de forma individual.
La decisión se hace comparando la señal neta con cero.

$y_N = y_{bruta} - y_{bl}$ por lo que
$\sigma^2_N = \sigma^2_{bruta} + \sigma^2_{bl} = 2\sigma^2_{bl}$; $s^2_N = s^2_{bruta} + s^2_{bl} = 2s^2_{bl}$

Si la muestra no contiene analito, $y_{bruta} = y_{bl}$, y su diferencia será $N(0, 2\sigma^2_{bl})$. Por lo tanto los límites de decisión y detección para señales corregidas del blanco serán:

$$L_C = k_C \sigma_0 = k_C \sqrt{2}\, s_{bl}$$

$$L_D = k'_D\, \sigma_0 = k'_D \sqrt{2}\, s_{bl}$$

2.2) Si se miden N réplicas de la muestra, de las que se *resta la media de n determinaciones del blanco* que se llevan a cabo de forma separada, ahora

$y_N = y_{bruta} - \overline{y}_{bl}$ por lo que

$\sigma^2_N = \sigma^2_{bruta} + \sigma^2_{bl}$ /n y si σ es independiente de la concentración $\sigma_0 = \sqrt{(1/N) + (1/n)}\, \sigma_{bl}$

$$L_C = k_C \sigma_0 = k_C \sqrt{(1/N) + (1/n)}\, s_{bl}$$

$$L_D = k'_D\, \sigma_0 = k'_D \sqrt{(1/N) + (1/n)}\, s_{bl}$$

Para convertir L_D en unidades de concentración, se utiliza la pendiente de la línea de calibrado:

$$x_C = \frac{L_D}{b_1}$$

En caso de métodos que implican la medida de picos con líneas base muy ruidosas se ha introducido el concepto de Límite de detección del Método: mínima concentración de sustancia que puede ser identificada, medida y reportada con un 99% de confianza como conteniendo una concentración de analito mayor que cero y se determina del análisis de una muestra en una matriz dada conteniendo el analito. Es confuso pues depende del método, del instrumento y del operador.

Límites de concentración a partir de la línea de calibrado

Ya que la línea de calibrado es solo una estimación de la verdadera línea de regresión, debe tenerse en cuenta su incertidumbre. Para ello se siguen estos pasos:

1) La ecuación que da el intervalo de confianza para la media de m respuestas a un valor particular de $x = x_0$ es:

$$p + q\,x_0 \pm t_{n-2}\, s_{y/x} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\Sigma\left(x_i - \overline{x}\right)^2}}$$

2) Se calcula y_C que es el límite superior de confianza con una cola para la media de m respuestas cuando la concentración de analito es cero

$$y_C = p + t_{\alpha,n-2}\, s_{y/x} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{\left(\overline{x}\right)^2}{\Sigma\left(x_i - \overline{x}\right)^2}}$$

3) Se calcula x_D, lo cual puede hacerse de tres posibles formas:

3.1) Como la intersección de la línea $y = y_C$ con la curva describiendo el límite inferior de confianza $y = y_L$ siendo

$$y_L = p + q\,x_D \;\; - t_{\beta,n-2}\, s_{y/x} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{\left(x_D - \overline{x}\right)^2}{\Sigma\left(x_i - \overline{x}\right)^2}}$$

que es un método engorroso

3.2) Por iteración. El valor de x_D se define como el valor más pequeño de x que origina un valor para y_L mayor o igual que y_C.

3.3) De forma aproximada (AOAC)

$$x_D = x_C + t_{\beta,n-2}\, s_{y/x} / p \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{\left(2x_C - \bar{x}\right)^2}{\Sigma\left(x_i - \bar{x}\right)^2}}$$

siendo $x_C = (y_C - q)/p$

A partir de x_D puede calcularse también $y_D = p + q\, x_D$

Límites de concentración aproximados

Si el modelo de regresión está bien elegido (no hay falta de ajuste) entonces s_e^2 puede tomarse como una buena estimación de σ^2. Si hacemos la suposición adicional de que $s_{bl}^2 = s_e^2$, entonces:

$$L_C = \mu_{bl} + k_C\, \sigma_{bl} = b_0 + k_c s_e$$
$$L_c = b_0 + b_1 x_C$$

Por lo tanto

$$x_C = \frac{k_C s_e}{b_1}$$

Expresiones análogas se derivan para x_D y x_Q

$$x_D = \frac{k_D s_e}{b_1}$$
$$x_Q = \frac{k_Q s_e}{b_1}$$

Los valores de k_C, k_D y k_Q recomendados en estos momentos son respectivamente: 1,65, 3,29 y 10.

Sensibilidad

La IUPAC define la sensibilidad como *la pendiente* de la línea de calibrado. A veces se emplea erróneamente como sinónimo de límite de detección.

La pendiente por sí misma no indica nada, ya que no basta con conocerla para saber si dos concentraciones pueden ser discriminadas entre sí. Se necesita también conocer la desviación típica de esa pendiente.

Se ha propuesto

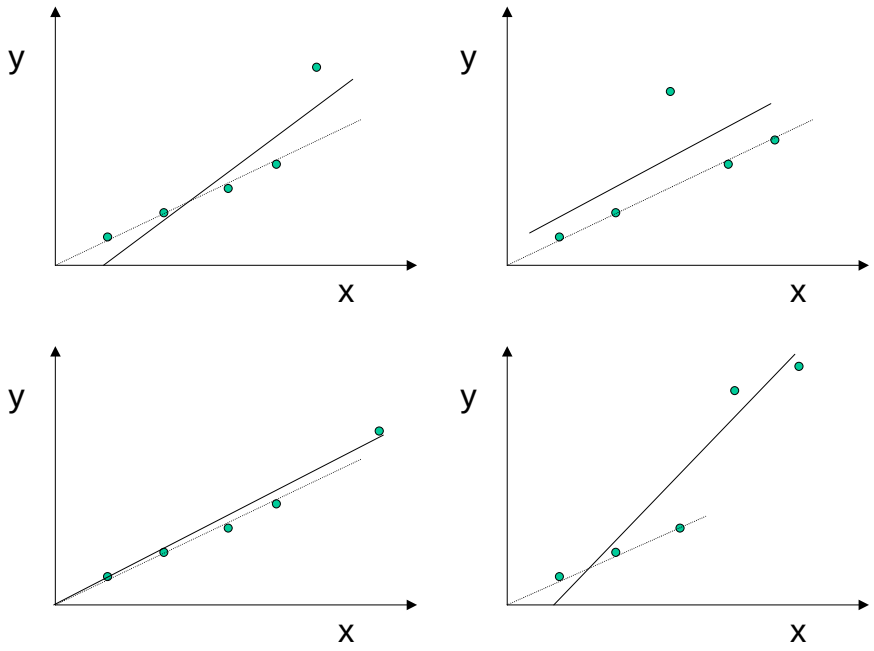
$$d = (t_{1-\alpha/2} + t_{1-\beta})\, s \sqrt{2}\, (1/p)$$

siendo los valores t para α=0,05 (2 colas) y β=0,05 (1 cola) para el número de g.d.l. con el que se determinó s (desviación típica de la señal)

DETECCIÓN DE OUTLIERS

Los "outliers" o elementos espurios son observaciones atípicas, que en el caso de los mínimos cuadrados pueden llegar a tener gran efecto sobre las estimaciones de los parámetros. Hay dos tipos principales: **"outliers" de la regresión**, que no son representativos del ajuste y **"outliers" de nivelación** ("leverage") que son puntos extremos que tienen una gran influencia sobre los parámetros y que pueden ser buenos o malos. En el caso de la regresión a una línea recta pueden detectarse mediante una representación gráfica, pero en el caso de la regresión múltiple su visualización es más difícil.

Para la detección de los de **regresión** existen varios criterios. El más sencillo es el valor absoluto del residual estandarizado: $\left|e_i / s_e\right|$. Si dicho valor es mayor de 2 o 3 el punto se rechaza. Este criterio se basa en la distribución normal de los residuales y la probabilidad de que un residual sea superior a 2 o 3 veces la s_e. Por desgracia no suele ser concluyente pues los "outliers" "atraen" la línea de regresión.



El segundo criterio es la distancia de Cook al cuadrado (Cook's square distance)

$$CD^2_{(i)} = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}^{(i)}_j)^2}{p s^2_e}$$

donde $\hat{y}_j$ son los valores predichos con la línea de regresión calculada con todos los puntos, $\hat{y}^{(i)}_j$ son los valores predichos con la línea de regresión calculada eliminando el sospechoso; p es el número de parámetros (2 para una recta) y s^2_e la varianza residual del ajuste con todos los puntos. Este criterio es más sensible a los "outliers" de regresión que el anterior. Cuanto mayor es $CD^2_{(i)}$, mayor es la probabilidad de que el "outlier" no sea representativo de la regresión. El valor de corte suele ser 1.

Para los "outliers" de **nivelación** la cuestión es más complicada, y existen otros criterios aunque la $CD^2_{(i)}$ suele dar buenos resultados. Siempre es conveniente indicar el criterio que sirvió para eliminar los puntos sospechosos.

OTRAS POSIBILIDADES

Regresión inversa

Algunos instrumentos analíticos comerciales incorporan la regresión inversa ya que se facilita el cálculo de la concentración. Hay mucha controversia sobre el método, puesto que se ajusta la variable x, libre de error, a la variable y, sujeta a error: $\hat{x} = b_0 + b_1 y$

Método de adiciones patrón

Cuando hay efecto de matriz y no se puede preparar una línea de calibrado en la que aquélla se duplique exactamente, no se puede aplicar el método del calibrado lineal. Una solución es el método de adiciones patrón (MAP o SAM). En este método se añaden cantidades conocidas del analito que se quiere determinar a alícuotas de la muestra desconocida (o a la misma alícuota si el método de análisis no es destructivo). Se representa la respuesta obtenida frente a cantidad de analito añadido y se aplica el método de regresión a una línea recta. La

cantidad de analito presente en la muestra, x_s, se estima extrapolando a ordenada cero la línea. En ausencia de errores sistemáticos absolutos, $\hat{x}_s = b_0 / b_1$

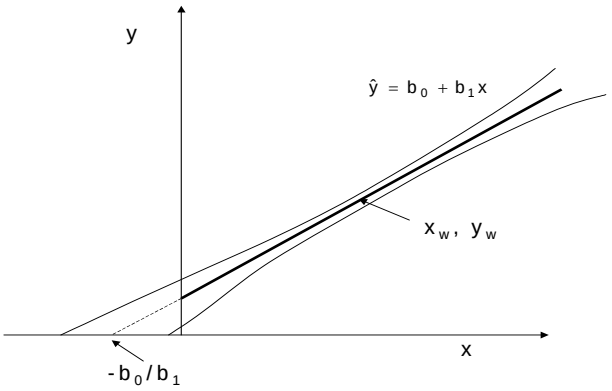
$$V_{eq} = - b_0 / b_1$$

En el caso del M.A.P. siempre debe hacerse una determinación en blanco, ya que el modelo que liga la Respuesta con la Concentración es ahora R = k C. Ese volumen equivalente, $(V_{eq})_{blanco}$ debe sustraerse del correspondiente a la muestra. De esa forma, el volumen equivalente real de la muestra $(V_{eq})_{muestra}$ será:

$$(V_{eq})_{muestra} = V_{eq} - (V_{eq})_{blanco}$$

Dicho resultado se ha obtenido a partir de dos ajustes por mínimos cuadrados (que no son exactos), por lo que su varianza será:

$$s^2_{(V_{eq})_{muestra}} = s^2_{V_{eq}} + s^2_{(V_{eq})_{blanco}}$$



Los dos componentes de la varianza se estiman a partir de la siguiente expresión general:

$$s^2_{V_{eq}} = \frac{s_e^2}{b_1^2}\left(\frac{1}{n} + \frac{(\bar{R})^2}{b_1^2 \sum (V_i - \bar{V})^2}\right)$$

en la que se deben sustituir los parámetros correspondientes a las dos determinaciones por el M.A.P. de la muestra y del blanco.

V_{eq} se convierte en concentración multiplicando por el factor C_p/V_0 y la desviación típica, o incertidumbre, del resultado final se calculará mediante las leyes de propagación de errores.

Para datos heterocedásticos:

$$s_{\hat{x}_0} = \frac{s_e}{b_1}\sqrt{\frac{1}{\sum w_i} + \frac{\bar{y}_w^2}{b_1^2(\sum w_i x_i^2 - \sum w_i \bar{x}_w^2)}}$$

donde todos los datos tienen el significado habitual.

Un problema del MAP es que al **extrapola**r se trabaja en una zona en la cual la imprecisión es muy grande. Una ventaja es su uso para detectar efectos de matriz que originen un error sistemático relativo. Para ello se compara la pendiente de un MAP con la de una línea de calibrado acuosa. Si la matriz no interfiere, ambas líneas deben tener la misma pendiente.

Por último, y cuando el método de análisis químico sea destructivo, no quedará más remedio que hacer adiciones sobre porciones (alícuotas) diferentes de la misma muestra.

Comparación de las pendientes de dos líneas de regresión

Dadas las pendientes de dos líneas de regresión b_{11} y b_{12}, se las puede comparar mediante un test t:

$$t = \frac{b_{11} - b_{12}}{\sqrt{s_{ep}^2\left(\frac{1}{\sum (x_{i1} - \bar{x})^2} + \frac{1}{\sum (x_{i2} - \bar{x})^2}\right)}}$$

$$s_{ep}^2 = \frac{(n_1 - 2)s_{e1}^2 + (n_2 - 2)s_{ep}^2}{n_1 + n_2 - 4}$$

donde el t crítico se busca con n_1+n_2-4 g.d.l. y el nivel de significación deseado.

Si las varianzas residuales de ambas líneas no son comparables, el t crítico se calcula previamente como:

$$t' = \frac{t_1 s_{b_{11}}^2 + t_2 s_{b_{12}}^2}{s_{b_{11}}^2 + s_{b_{12}}^2}$$

Este procedimiento de comparación es útil en la validación de métodos analíticos.

Intersección de dos líneas de regresión

En algunas valoraciones (fotométricas o conductimétricas) el punto final es la intersección de dos líneas rectas. Las ecuaciones y valores son los siguientes:

Línea 1: $y_1 = b_0 + b_1 x_1$ con n_1 puntos

Línea 2: $y_2 = b_0' + b_1' x_1$ con n_2 puntos

La intersección es:

$$\hat{x}_1 = \frac{(b_0 - b_0')}{(b_1' - b_1')} = \frac{\Delta b_0}{\Delta b_1}$$

Los límites del intervalo de confianza se obtienen con las raíces de la siguiente ecuación de segundo grado

$$\hat{x}_1^2((\Delta b_1)^2 - t^2 s^2_{\Delta b_1}) - 2\hat{x}_1(\Delta b_0 \Delta b_1 - t^2 s^2_{\Delta b_0 \Delta b_1}) + ((\Delta b_0)^2 - t^2 s^2_{\Delta b_0}) = 0$$

$$s^2_{\Delta b_1} = s^2_{ep}\left(1 \Big/ \sum (x_{i1} - \bar{x}_1)^2 + 1 \Big/ \sum (x_{i2} - \bar{x})^2\right)$$

$$s^2_{\Delta b_0} = s^2_{ep}\left(1/n_1 + 1/n_2 + \bar{x}_1^2 \Big/ \sum (x_{i1} - \bar{x}_1)^2 + \bar{x}_2^2 \Big/ \sum (x_{i2} - \bar{x})^2\right)$$

$$s_{\Delta b_0 \Delta b_1} = s^2_{ep}\left(\bar{x}_1 \Big/ \sum (x_{i1} - \bar{x}_1)^2 + \bar{x}_2 \Big/ \sum (x_{i2} - \bar{x})^2\right)$$

donde la s^2_{ep} es una varianza promediada calculada como en el caso de la comparación de dos líneas de regresión.

El valor de t es el tabulado con n_1+n_2-4 g.d.l.

Validación de métodos

1) Ensayos de recuperación

Se utilizan para validar un procedimiento analítico. Para ello, se analizan blancos dopados (spiked) con concentraciones diferentes de analito exactamente conocidas.

Se representa la concentración de analito encontrada frente a la añadida. Al hacer la regresión, se debería obtener una línea recta de ordenada en el origen cero y pendiente uno o lo que es lo mismo, los intervalos de confianza de b_0 y b_1 deberían incluir a cero y a uno, respectivamente.

Si la ordenada en el origen es mayor de cero existe un error sistemático (bias) constante (o una corrección incorrecta del blanco)

Si la pendiente difiere de uno, existe un error sistemático (bias) proporcional que suele ser debido a la matriz.

2) Comparación de métodos

Para comparar y validar un método frente a un método de referencia, se analizan blancos o muestras reales dopadas (spiked) con diferentes cantidades de analito. En este caso, no es estrictamente necesario el conocer exactamente la concentración de analito presente en cada caso.

Se representan las concentraciones encontradas mediante el método a validar, frente a las encontradas con el método de referencia. Al hacer la regresión, se debería obtener una línea recta de ordenada en el origen cero y pendiente uno, o bien los intervalos de confianza de b_0 y b_1 deben incluir a cero y a uno, respectivamente.

Si la ordenada en el origen es mayor de cero el método a validar introduce un error sistemático (bias) constante respecto del método de referencia.

Si la pendiente difiere de uno, el método a validar introduce un error sistemático (bias) proporcional respecto del método de referencia.

3) Efectos de matriz

Los efectos de matriz se detectan comparando la pendiente de una línea de calibrado hecha con patrones puros, y la pendiente de un M.A.P. realizado sobre una muestra real. Las concentraciones de analito utilizadas en ambos casos deberían ser comparables.

Si las pendientes son idénticas, o sus intervalos se solapan, o la prueba t dice que son comparables, no hay efecto de matriz: Las determinaciones analíticas pueden hacerse por calibrado lineal.
Si las pendientes son diferentes, hay efecto de matriz: Es necesario utilizar el M.A.P.

Regresión a través de un punto fijo

En ocasiones se fuerza a que la línea pase a través de un punto (x_0, y_0). Por tanto la ecuación de la recta debe cumplir:

$$y_0 = b_0 + b_1 x_0$$

por lo que el modelo debe cumplir

$$\hat{y} = y_0 + b_1(x - x_0)$$

Se tiene pues un modelo con un único parámetro b_1, y debe minimizarse la expresión

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - y_0 - b_1(x_i - x_0))^2$$

respecto de b_1, para el que se obtiene la siguiente expresión:

$$b_1 = \frac{\sum (x_i - x_0)(y_i - y_0)}{\sum (x_i - x_0)^2}$$

La varianza residual s_e^2 o $s_{y/x}^2$, que es una estimación de σ^2 cuando el modelo es correcto vale:

$$s_e^2 = \frac{\sum (e_i)^2}{n-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-1}$$

Si el punto fijo es el origen, $x_0 = 0$ e $y_0 = 0$, las ecuaciones se simplifican y tenemos

$$\hat{y} = b_1 x$$

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

La desviaciones típicas del modelo son:

- Para la pendiente $s_{b_1} = s_e \sqrt{\frac{1}{\sum x_i^2}}$
- Para la estimación de la respuesta a partir de un valor x_k

$$s_{\hat{y}_0} = s_e x_k / \sqrt{x_i^2}$$

- Para la estimación de la respuesta media a partir de un x_k

$$s_{\hat{y}_0} = s_e / \sqrt{1/m + x_k^2 / \sum x_i^2}$$

- Para la estimación de una x_s a partir de una respuesta y_s, media de m valores

$$s_{\hat{x}_0} = (s_e / b_1)\sqrt{1/m + y_s^2 / (b_1^2 \sum x_i^2)}$$

Este modelo solo debe ser utilizado cuando haya buenas razones a priori para ello, y eso no incluye aquellos casos en los que se haya demostrado que b_0 no es significativamente diferente de cero.

Linearización de funciones curvas

Cuando la relación entre las dos variables no puede ser representada por un línea recta, cabe la posibilidad de hacer una ajuste polinómico (ver Regresión Lineal Múltiple) o no lineal (ver Métodos de Ajuste No Lineal). Una alternativa es transformar una o ambas variables, de manera que se obtenga una relación más sencilla. Por ejemplo, la ecuación de Arrhenius, $k = A e^{(-E/RT)}$, la ecuación de decaimiento de la actividad de un radionúclido, $A_t = A_0 e^{-0{,}693\, t / t_{1/2}}$, la ecuación de Michaelis-Menten $v = \frac{v_{max}[S]}{K_m + [S]}$. Tomando logaritmos en

esas expresiones, se obtienen ecuaciones que pueden ser luego ajustadas por regresión a una línea recta

Ecuación	Ecuación linearizada	y	x
$k = A\,e^{(-E/RT)}$	$\ln k = \ln A - \frac{E}{RT}$	$\ln k$	$\frac{1}{T}$
$A_t = A_0\,e^{-0{,}693\,t/t_{1/2}}$	$\ln A_t = \ln A_0 - \frac{0{,}693\,t}{t_{1/2}}$	$\ln A_t$	t
$v = \frac{v_{max}[S]}{K_m + [S]}$	$\frac{1}{v} = \frac{1}{v_{max}} + \frac{K_m}{v_{max}} \cdot \frac{1}{[S]}$	$\frac{1}{v}$	$\frac{1}{[S]}$

El problema es que la condición de homocedasticidad, que suele cumplirse con los datos originales, no se mantiene con los datos transformados. Eso implica el que la regresión deba hacerse utilizando pesos estadísticos, de la forma indicada en la página 7. De forma general, si una variable y se transforma por medio de $y_i = f(y_i)$, la expresión para calcular los pesos es:

$$w_i = 1 \Big/ s^2_{f(y_i)}$$

$$s^2_{f(y_i)} = \left(\frac{d(f(y))}{dy}\right)^2 s^2_y$$

Por ejemplo, si se transforma A en ln A, se tiene

$$s^2_{\ln A} = \left(\frac{d(\ln A)}{dA}\right)^2 s^2_A = \left(\frac{1}{A^2}\right) s^2_A$$

$$w_i = 1 \Big/ s^2_{\ln A_i} = A_i^2 \Big/ s^2_{A_i}$$

y dado que s_A^2 es constante, se puede tomar $w_i = A_i^2$. Cálculos similares pueden hacerse para cualquier otra transformación, aunque no es usual hacerlos.

REGRESIÓN vs. CORRELACIÓN

La correlación sirve para estudiar la asociación entre dos variables aleatorias: no hay variable dependiente ni independiente. Esa asociación se cuantifica mediante la covarianza y el coeficiente de correlación.

Si tenemos dos parámetros determinados sobre n muestras: $y_{11},\ldots,y_{1n}$ y por otro lado $y_{21},\ldots,y_{2n}$ con sus respectivas medias $\bar{y}_1$ e $\bar{y}_2$. Se define la covarianza como:

$$\mathrm{cov}(y_1, y_2) = \frac{1}{n-1}\sum (y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)$$

Este valor es una estimación de la verdadera covarianza y puede variar entre $-\infty$ cuando las variables están asociadas negativamente (una disminuye cuando la otra aumenta) y $+\infty$ cuando la asociación es positiva.

Su magnitud depende de la escala de medida, por lo que utiliza más el coeficiente de correlación producto-momento o **coeficiente de correlación de Pearson**:

$$r(y_1, y_2) = \frac{\mathrm{cov}(y_1, y_2)}{s_{y_1} s_{y_2}} = \frac{\sum (y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)}{\sqrt{\sum (y_{1i} - \bar{y}_1)^2 \sum (y_{2i} - \bar{y}_2)^2}}$$

$r(y_1,y_2)$ estima el verdadero $\rho(y_1,y_2)$ y vale entre –1 y +1 dependiendo del tipo de asociación: Los valores –1 y +1 indican una perfecta relación linear entre ambas variables. Un coeficiente de correlación que no es significativamente diferente de cero indica que las variables no están

correlacionadas. Esto no quiere decir que no haya relación entre ellas, sino que esa relación no es lineal.

Se puede comprobar la H_0 de que $\rho = 0$. Para ello se siguen los pasos habituales:

1) H_0: $\rho = 0$; $H_1 \neq 0$
2) $\alpha = 0{,}05$ (95% de confianza)
3) $t_{cal} = \dfrac{r\sqrt{n-2}}{1-r^2}$
4) t_{crit} se busca en una tabla de dos colas con n-2 g.d.l.
5) Si $t_{cal} < t_{crit}$ H_0 se mantiene. Si $t_{cal} > t_{crit}$ H_0 se rechaza y se acepta H_1

Este test solo permite comprobar si $\rho=0$, pues la distribución de r es asimétrica y no normal. Para calcular los intervalos de confianza de un valor r dado, se calcula primero una nueva variable z:

$$z = 0{,}5\ln[(1+r)/(1-r)]$$

que está normalmente distribuida con una $\sigma = \sqrt{1/(n-3)}$. Por ello, el intervalo de confianza de z es $z \pm 1{,}96 \cdot 1/\sqrt{(n-3)}$, e invirtiendo la transformación, se obtiene el intervalo de confianza de r, que no es simétrico alrededor de r.

Aunque sirven para diferentes propósitos, hay evidentes relaciones matemáticas entre el **coeficiente de correlación de Pearson r(x,y) y los coeficientes de la regresión b_0 y b_1 de x sobre y**. Resumiendo:

$$b_1 = r(x,y)\frac{s_y}{s_x}$$

$$s_x = \sqrt{\frac{(x_i - \bar{x})^2}{n-1}} \quad \text{y} \quad s_y = \sqrt{\frac{(y_i - \bar{y})^2}{n-1}}$$

Si no hay correlación entre x e y, no existe una regresión lineal significativa entre x e y. Por tanto el test de hipótesis de que $\rho=0$ da idénticos resultados al de $\beta_1=0$ y ambos son equivalentes.

Si se eleva r al cuadrado, se obtiene finalmente $r^2 = \dfrac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \dfrac{SS_{REG}}{SS_T}$.

En regresión, el coeficiente de correlación elevado al cuadrado, r^2 se denomina **coeficiente de determinación** y expresa la **proporción de variación total que es explicada por la regresión**. Si r=-1 o r=+1 todas las observaciones se ajustan perfectamente a una línea recta y por tanto toda la variación en y puede explicarse en términos de la línea de regresión ($r^2=1$). Si por el contrario r=0, no hay regresión entre x e y por lo que la regresión no explica nada de la variación de y. Además en ese caso $b_1=0$ es decir la línea de regresión es paralela al eje x.

La **utilidad real de r ha sido sobrestimada** en muchas ocasiones debido a que es un parámetro proporcionado rutinariamente por todas las calculadoras y paquetes estadísticos. Por tanto:

1) Lo verdaderamente útil no es r sino **r^2** que expresa la proporción de variación explicada por la regresión.
2) En el caso concreto del calibrado analítico que se realiza en análisis instrumental, siempre deben obtenerse valores de r próximos a la unidad (*"con muchos nueves"*), de manera que también **r^2 sea próximo a uno**. Esto es lógico ya que en análisis instrumental siempre se trabaja con respuestas (y) que dependen linealmente de la concentración (x), por lo que si r^2 no es próximo a la unidad es que *"algo va mal"*.
3) En preferible la realización de la prueba **$F = MS_{FDA}/MS_{PE}$** indicada en la validación del modelo, pues eso permite conocer si la varianza no explicada es superior o no a la puramente experimental. Eso implica el hacer medidas replicadas.

BIBLIOGRAFÍA

Massart D.L., Vandeginste, B.G.M., Buydens L.M.C., De Jong S., Lewi P.J. and Smeyers-Verbeke J., *Handbook of Chemometrics and Qualimetrics*. Elsevier, Amsterdam, 1997

EJERCICIOS PROPUESTOS

1.-A partir de los datos siguientes correspondientes a una línea de calibrado, demostrar si el modelo elegido presenta o no falta de ajuste

C	0	5	10	10	20	20	30	40	40	50
R	3,6	13,5	24,6	22,9	44,1	43,3	63,9	84,0	82,5	105,3

2.- Se determinan varias muestras problema, obteniendo las respuestas que se indican a continuación. Determine en todos los casos la concentración del problema y su incertidumbre (asuma que las varianzas de la muestra y los patrones son homogéneas)
Muestra 1: 49,4; 49,6; 49,2
Muestra 2: 91,5; 93,5; 90,0; 94,2
Muestra 3: 3,8; 3,6; 3,6; 4;0

3.- Se tiene la siguiente línea de calibrado correspondiente a la determinación fluorimétrica de Ácido acetilsalicílico (AAS)

C	0	2,0	4,0	6,0	8,0	10,0
R	3	803	1586	2350	3090	3802

Calcule todos los parámetros de la línea de calibrado y determine la incertidumbre del resultado obtenido al medir una muestra en los siguientes casos (sin sumir que las varianzas de los patrones y la muestra son homogéneas):
Muestra 1: 1540; 1524 ; 1531
Muestra 2: 500
Muestra 3: 340; 355; 334; 347
Muestra 4: 3507 ; 3516; 3513; 3538; 3501; 3523; 3512

4.- Para saber si un procedimiento analítico está libre de bias se llevan a cabo una serie de experiencias de recuperación (recovery) en las que se analizan muestras diferentes que contienen cantidades perfectamente conocidas del analito de interés. A partir de los resultados demuestre si el procedimiento está o no libre de bias (con un nivel de significación α = 0,05) y si fuera necesaria de qué tipo de bias se trata. (En todos los casos, la cantidad de analito puesto es la que aparece en la columna de la izquierda))

Analito	**Analito hallado**		
Puesto	**Caso 1**	**Caso 2**	**Caso 3**
4,20	4,35	4,52	4,34
4,20	4,35	4,47	4,40
4,20	4,37	4,54	4,40
8,40	8,57	8,86	8,74
8,40	8,59	8,83	8,70
8,40	8,59	8,79	8,67
12,60	12,74	13,08	12,98
12,60	12,75	13,15	13,01
12,60	12,76	13,09	13,03
16,80	17,02	17,38	17,35
16,80	17,01	17,39	17,35
16,80	16,98	17,47	17,39

5.- Demuestre si hay efecto de matriz en la determinación de cinc por AAS, al trabajar en medio ácido acético, clorhidrato de hidroxilamina, acetato amónico, cloruro magnésico y acetato sódico. La concentración de cinc está en mg/l.

	Absorbancia					
Concentración	**Agua**	**HAcO**	**$NH_2OH.HCl$**	**NH_4AcO**	**$MgCl_2$**	**NaAcO**
0,0	0,0029	-0,0007	0,0047	0,0084	-0,0038	0,0048
0,5	0,0766	0,0727	0,0585	0,0629	0,0862	0,0826
1,0	0,1545	0,1384	0,1187	0,1172	0,1642	0,1646
1,5	0,2176	0,2071	0,1723	0,1771	0,2234	0,2476
2,0	0,2740	0,2590	0,2187	0,2343	0,2946	0,3280