CALIBRADO Y REGRESIÓN

- **Introducción**
- **Límite de detección y cantidades relacionadas**
- **Método de los mínimos cuadrados**
  - Estimación de los parámetros
  - Validación del modelo
  - Heterocedasticidad y su solución
- **Intervalos de confianza**
  - De los parámetros por separado
  - De los parámetros conjuntamente
- **Predicciones hechas sobre la línea ajustada**
  - Predicción de nuevas respuestas
  - Interpolación de x a partir de y
- **Detección de outliers**
- **Otras posibilidades**
  - Regresión inversa
  - Método de adiciones/sustracciones patrón
  - Comparación de pendientes
  - Intersección de dos rectas de regresión
  - Validación de métodos
  - Regresión a través de un punto fijo
  - Linearización de funciones curvas
- **Regresión vs. correlación**

CALIBRADO Y REGRESIÓN

Introducción

Relación entre variables asociadas entre sí.
- Análisis Instrumental en que la Respuesta(s) está relacionada con la Concentración de analito(s)
- Construcción de modelos para predecir una variable (output) en función de una o varias entradas (inputs)

La relación se estudia mediante un **Análisis de regresión** y consiste en construir una **función matemática** que puede ser utilizada para **predecir** una variable a partir de las otras, o bien para **interpolar**.

Técnicas de regresión Modelo I: Dependencia de una variable aleatoria (variable dependiente o respuesta) en función de una variable controlada por el experimentador (variable independiente o de predicción).

Técnicas de regresión Modelo II: Dependencia entre variables cuando todas están sujetas a error

Desarrollo:
- **Regresión lineal simple:** Ajuste lineal de una respuesta a una sola variable independiente
- **Regresión múltiple**: Ajuste lineal de una (o varias) respuesta(s) a varias variables independientes

Simultáneamente iremos presentando las principales aplicaciones analíticas

MÉTODO DE LOS MÍNIMOS CUADRADOS

ESTIMACIÓN DE LOS PARÁMETROS

- Suponemos que existe una relación verdadera entre η (respuesta) y una variable independiente **x**

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

β_0 y β_1 : parámetros del modelo (ordenada en el origen y pendiente) **relacionados linealmente con** η

- En la realidad, únicamente se tiene acceso a un valor experimental y_i, sujeto a error:

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

β_0 y β_1 desconocidos. Se estiman mediante b_0 y b_1 a partir de la información de las medidas

- La línea estimada se denomina línea de **mínimos cuadrados** si se minimiza la suma de los cuadrados de los residuales

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

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

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

- Se deriva R frente a b_0 y b_1 y se iguala a cero, obteniendo las ecuaciones normales

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

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

MÉTODO DE LOS MÍNIMOS CUADRADOS

ESTIMACIÓN DE LOS PARÁMETROS

- Las estimaciones de los parámetros son:

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

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

- La **varianza residual**

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

es la varianza **no explicada** por la línea de regresión. A veces se la representa como $s_{y/x}^2$.

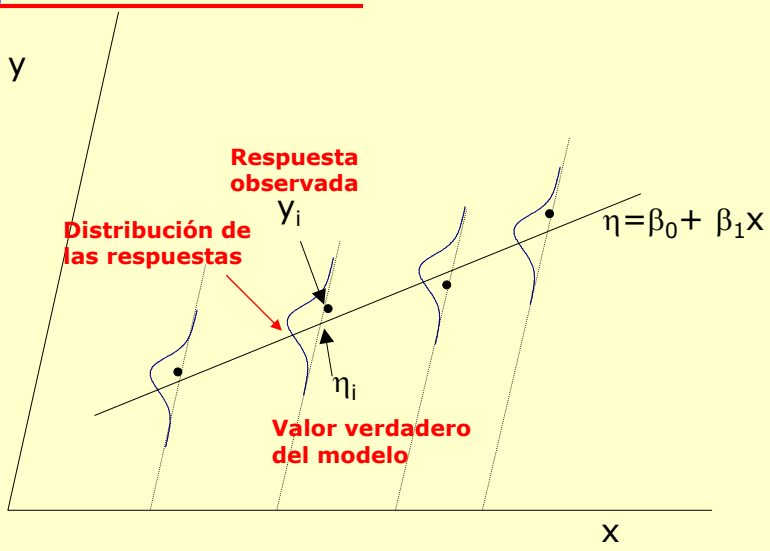
- Si el modelo es **correcto** s_e^2 es una **estimación de la varianza de las medidas** σ^2 **o puro error experimental.**

Condición fundamental de validez del modelo

MÉTODO DE LOS MÍNIMOS CUADRADOS

- **Se hacen las siguientes suposiciones implícitas**

  - Todos los residuales vienen de una población **normalmente distribuida con media cero**
  - Los e_i son **independientes**
  - Todos los e_i tienen la misma varianza σ^2. Condición de la **homocedasticidad**. Se comprueba mediante medidas replicadas y en caso negativo obliga a la utilización de pesos estadísticos. En calibrado eso significa que la precisión de las respuestas es independiente de la concentración.



MÉTODO DE LOS MÍNIMOS CUADRADOS

ETAPAS DEL PROCEDIMIENTO

- Selección de un modelo **Línea recta**
- Establecimiento del diseño experimental **Cómo se eligen los puntos**
- Estimación de los parámetros del modelo **Cálculo de b_0 y b_1**
- Validación del modelo **$N(0,\sigma^2)$, f.d.a, ¿r^2?**
- Determinación de los intervalos de confianza **s_{b0}, s_{b1}**
- Utilización del modelo para predecir o interpolar $\hat{y} \pm ts_{\hat{y}}$; $\hat{x} \pm ts_{\hat{x}}$

VALIDACIÓN DEL MODELO

No suele haber bastantes datos

Análisis de los residuales

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

- Se trata de verificar la hipótesis de que los residuales son **N(0,σ²)** mediante una prueba χ^2, rankit o un gráfico de normalidad

- **Gráfico de residuales**: 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

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

EXCEL

VALIDACIÓN DEL MODELO

Análisis de Varianza (ANOVA)

- Se trata de demostrar que la relación que liga las variables es la correcta: no hay falta de ajuste

- Se compara la **varianza debida a la presunta falta de ajuste con la varianza experimental σ²**, estimada a través de medidas replicadas de la variable observada

Si el modelo es correcto $s_{y/x}^2$ estimará σ^2

- P.e. En una línea de calibrado alguno de los puntos **se mide (observa) más de una vez**.

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		$n = \sum_i n_i = 14$		$\hat{y} = 0{,}0230 + 0{,}1181\,x$			$\bar{y} = 0{,}2001$	

- En este caso se tienen 14 puntos experimentales, correspondientes a k=7 situaciones diferentes .

VALIDACIÓN DEL MODELO

Análisis de Varianza (ANOVA)

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	$n = \sum_i n_i = 14$		$\hat{y} = 0{,}0230 + 0{,}1181\,x$			$\bar{y} = 0{,}2001$	

Desarrollo del ANOVA

Variación total de los datos **SS_T**: $SS_T = \sum_{i=1}^{k}\sum_{j=1}^{nk}(y_{ij} - \bar{y})^2$

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

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

$$SS_T = SS_{PE} + SS_{FDA} + SS_{REG}$$

$$SS_T = SS_{RES} + SS_{REG} \quad \leftarrow \textbf{SS}$$

$$(n-1) = (n-k) + (k-p) + (p-1) \quad \leftarrow \textbf{g.d.l.}$$

VALIDACIÓN DEL MODELO

Análisis de Varianza (ANOVA)

- Dividiendo SS por sus g.d.l. se obtienen MS: estimaciones de las varianzas del modelo:
  MS_{PE} es una estimación de σ^2
  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 que compara MS_{FDA}/MS_{PE} con el $F_{crít}$ con (k-2) y (n-k) g.d.l. Si **H_0 se rechaza el modelo no es adecuado**, mientras **si se retiene sí que lo es** y $MS_{RES} = s_e^2$ es 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		

- 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 g.d.l. para calcular MS_{PE}

Fuente	g.d.l	SS	MS	F
Regresión	1	0,19516	0,19516	1337,37
Residual	12	0,00175	0,000146	
Falta de ajuste	5	0,00169	0,000338	38,96
Puro error	7	0,00006	$8{,}68.10^{-6}$	
Total	13	0,19691		

HAY FALTA DE AJUSTE

F_{crit} con 5 y 7 g.d.l. y α = 0,05 vale 3,97 (p a posteriori = 0,00006)

VALIDACIÓN DEL MODELO

EXCEL

Análisis de Varianza (ANOVA)

- Existe otra comparación: **MS_{REG} con MS_{RES}** que es lo mismo que comprobar la hipótesis nula que $\beta_1 = 0$.

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}	$SS_{RES}/(n-2)$	
Total	n-1	SS_T		

y en este caso resulta

Fuente	g.d.l	SS	MS	F
Regresión	1	0,19516	0,19516	1337,6
Residual	12	0,000175	0,0001459	
Total	13	0,19691		

F_{crit} con 1 y 12 g.d.l. y $\alpha = 0,05$ vale 4,74 (p a posteriori 7.10^{-8})

La prueba dice que el modelo sí que explica una cantidad apreciable de la varianza.

- El cociente SS_{REG}/SS_T es el **coeficiente de determinación múltiple R^2** y es la proporción de varianza o información 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 **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.**

VALIDACIÓN DEL MODELO

Heterocedasticidad

Si no se cumple la condición de **homocedasticidad,** no se puede aplicar el método de mínimos cuadrados anterior.

2) 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:

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

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

DETERMINACIÓN DE LOS INTERVALOS DE CONFIANZA

De los parámetros individuales

- Los **intervalos de confianza de un parámetro** con un 100(1-α)% de confianza son .

$$\pm t_{\alpha/2;n-2} \bullet s_{parámetro}$$

- Las estimación de las **s** respectivas son:

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

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

EXCEL

$$\beta_0 = b_0 \pm t_{0,025;n-2} \cdot s_{b0}$$
$$\beta_1 = b_1 \pm t0_{,025;n-2} \cdot s_{b1}$$

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

De los parámetros conjuntamente

- Debido a la forma de obtenerlos, los valores de b_0 y b_1 **no son independientes** entre sí. Si se desea comprobar simultáneamente la hipótesis de que $\beta_0 = 0$ y $\beta_1 = 1$, se usa un test de hipótesis conjunto (**joint hypothesis test**), o construir la 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

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

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

DETERMINACIÓN DE LOS INTERVALOS DE CONFIANZA

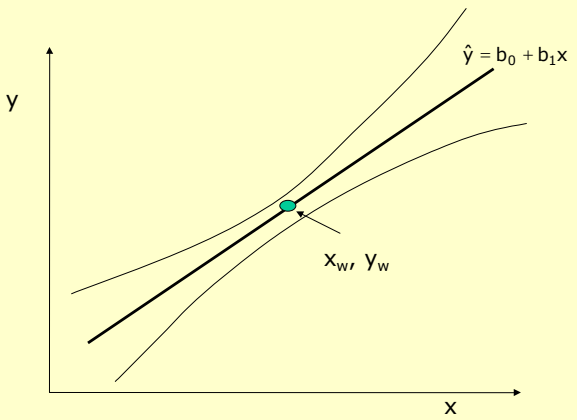
Intervalo de confianza de la línea de regresión

- 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** es importante, ya que la línea de regresión pasa por él.

- Se observa que el intervalo de confianza disminuye cuando x_0 coincide con el centroide: **la mayor precisión se obtiene en el centro de la línea de regresión**.

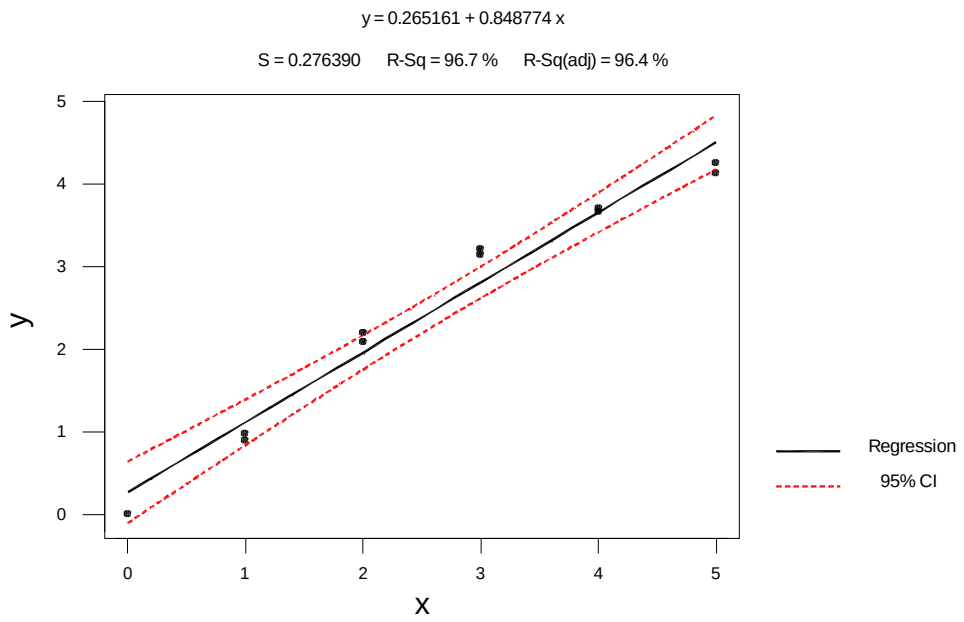
- El intervalo también disminuye cuando

$$\Sigma(x_i - \bar{x})^2$$

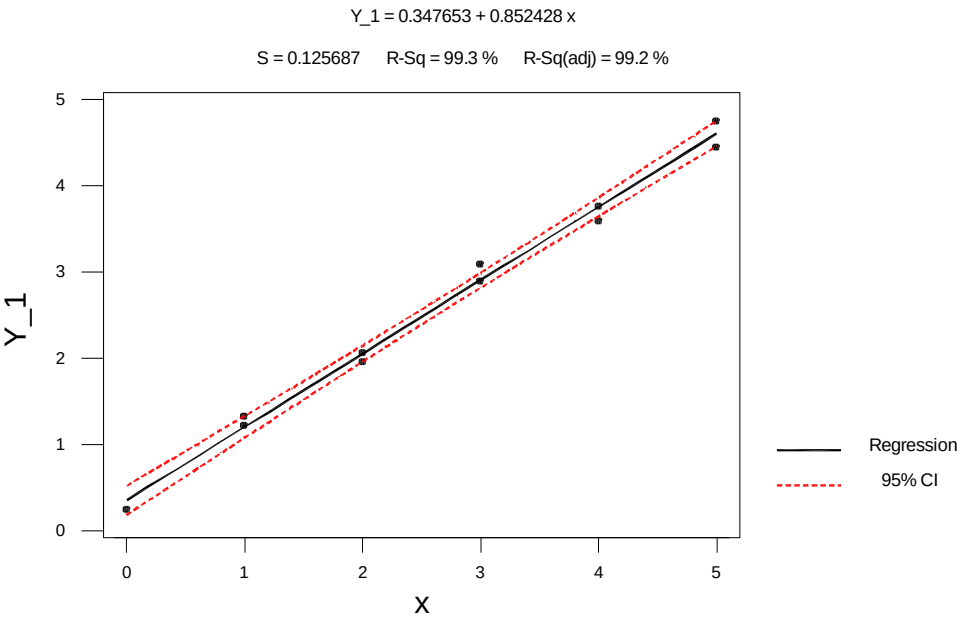
es muy grande, lo que obligaría en calibrado a utilizar patrones de muy baja y muy alta concentración lo cual **no es aconsejable**.

- El **número de puntos** también influye

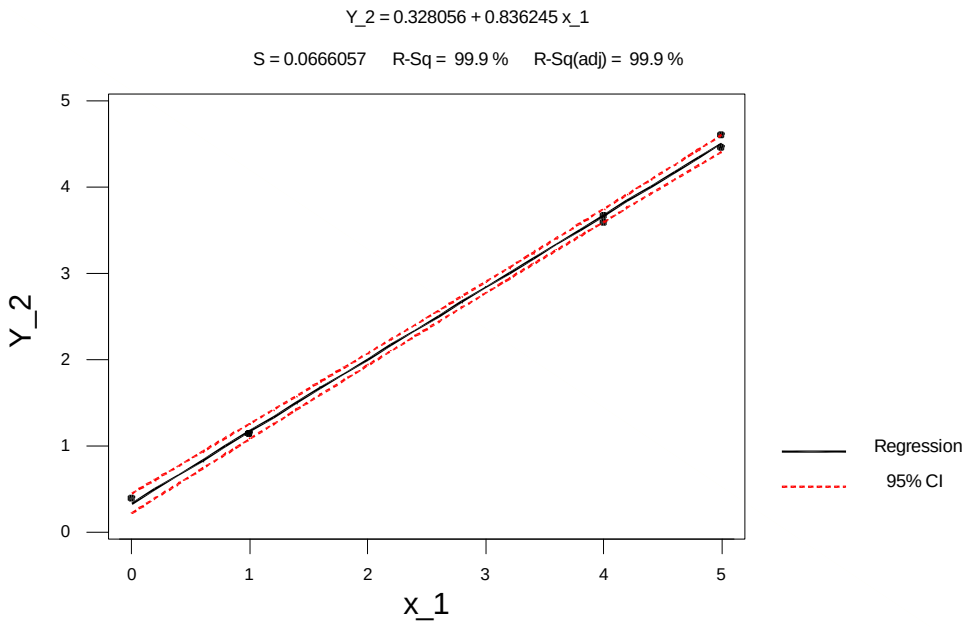
ZONA DE CONFIANZA DE LA LÍNEA DE REGRESIÓN



ZONA DE CONFIANZA DE LA LÍNEA DE REGRESIÓN



ZONA DE CONFIANZA DE LA LÍNEA DE REGRESIÓN



PREDICCIONES SOBRE LA LÍNEA AJUSTADA

Predicción de nuevas respuestas

$$\sigma^2_{total} = \sigma^2_{regresión} + \sigma^2_{observ.}$$

$$s^2_{total} = s^2_{regresión} + s^2_{observ.}$$

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

$$s^2_{observción} = s^2_e \quad \text{(modelo bien elegido)}$$

- Si la predicción se hace a partir de un valor de x

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

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

- Si se quiere predecir la media de una respuesta, medida **m** veces, la varianza vale

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

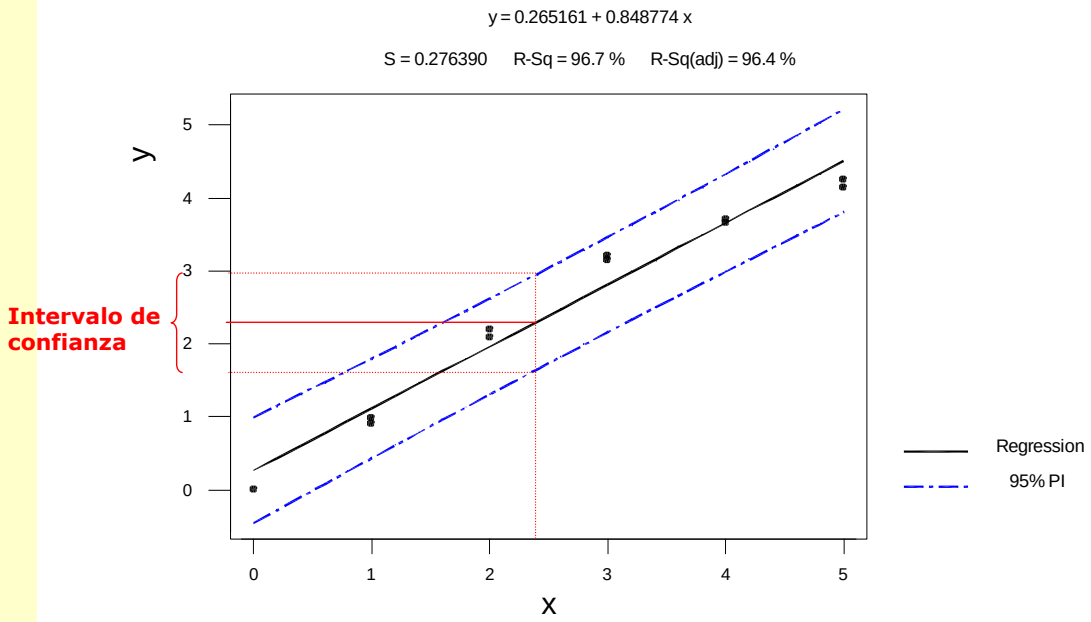
$$s_{\hat{y}_0} = s_e\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

- El **intervalo de confianza** es:

$$\hat{y}_0 \pm t_{0,025;n-2}\, s_{\hat{y}_0}$$

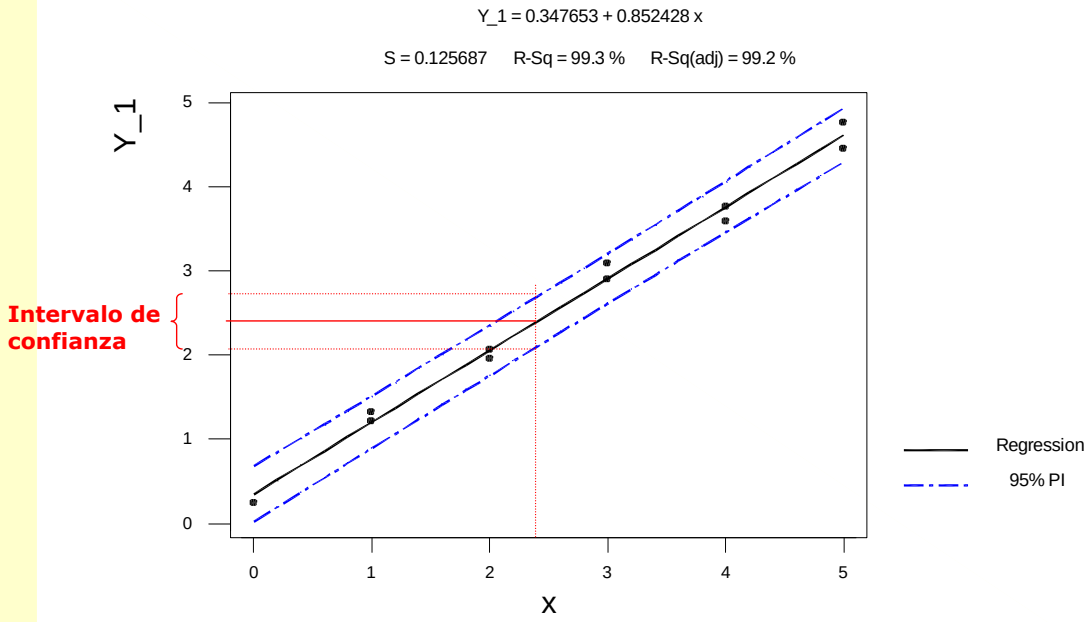
aunque algunos textos indican n+m-3 g.d.l

PREDICCIÓN DE NUEVAS RESPUESTAS



Intervalo de confianza

PREDICCIÓN DE NUEVAS RESPUESTAS



Intervalo de confianza

PREDICCIONES SOBRE LA LÍNEA AJUSTADA

Predicción de x a partir de y (interpolación)

- Con **m** observaciones de la respuesta, se obtiene $\bar{y}_s$ y se **predice** $\hat{x}_s = \dfrac{\bar{y}_s - b_0}{b_1}$
- La varianza de la predicción deberá tener en cuenta la **varianza de la observación y la varianza del ajuste**. Su determinación es complicada, y se acepta la siguiente aproximación:

1) Si la varianza de la observación es la misma que la de las observaciones realizadas durante el calibrado (s_e^2)

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

2) Si la varianza de la observación de la muestra (s_s^2) es diferente de la de los patrones de calibrado:

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

3) En el caso de datos heterocedásticos las fórmulas cambian:

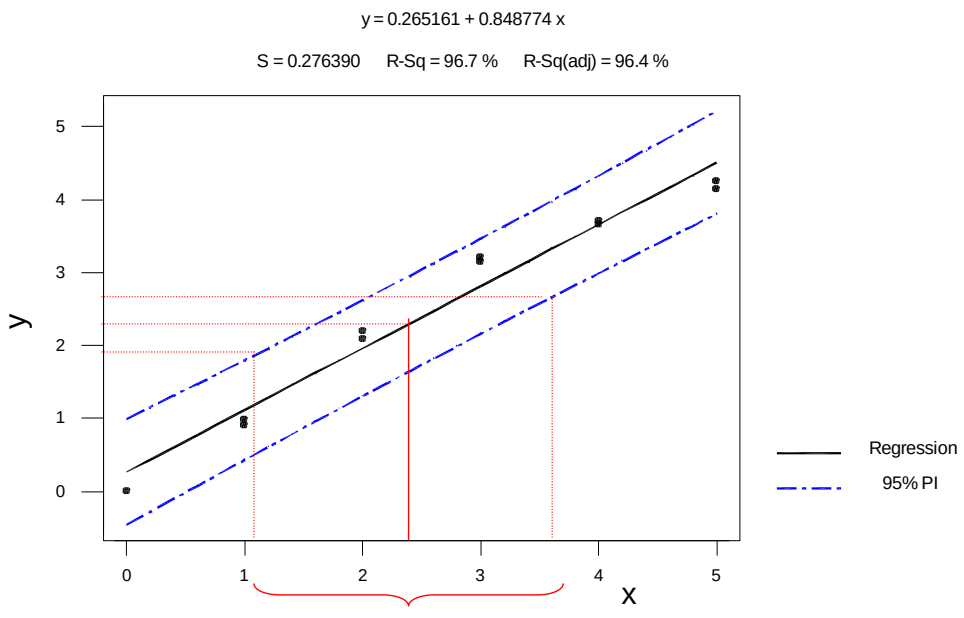
$$s_e = \sqrt{\frac{\sum w_i (y_i - \hat{y}_i)^2}{n-2}}$$

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

- El **intervalo de confianza** es $\hat{x}_s \pm t_{0,025;n-2}\, s_{\hat{x}_0}$

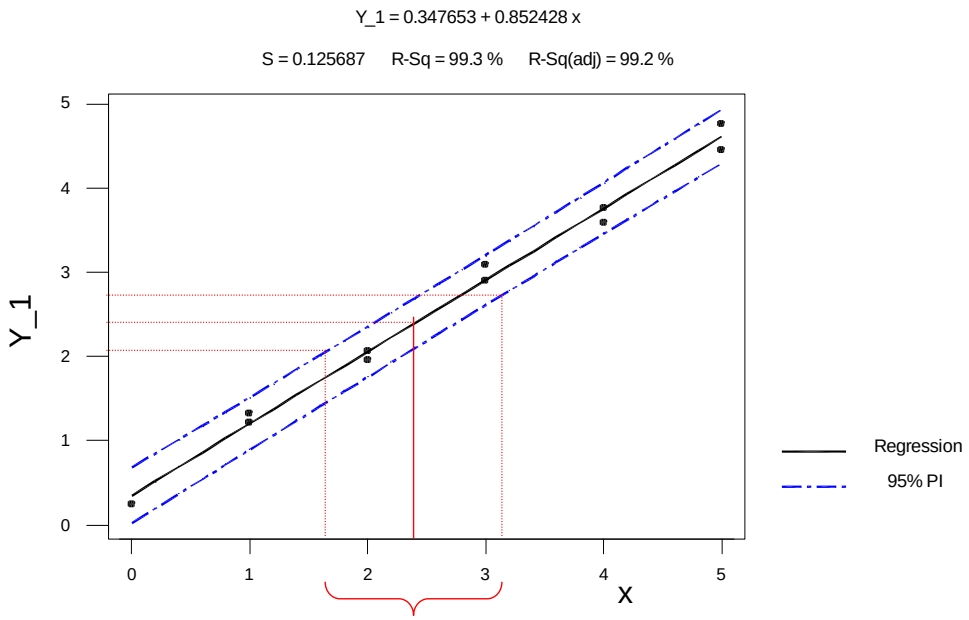
aunque algunos textos utilizan n+m-3 g.d.l.

PREDICCIÓN DE NUEVAS RESPUESTAS



Intervalo de confianza

PREDICCIÓN DE NUEVAS RESPUESTAS



Intervalo de confianza

PREDICCIONES SOBRE LA LÍNEA AJUSTADA

Como estrechar el intervalo de confianza

$$\hat{x}_s \pm t_{0,025;n-2}\, s_{\hat{x}_0}$$

$$\hat{x}_s = \frac{\bar{y}_s - b_0}{b_1}$$

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

1) Cuanto **mayores m** y **n** mejor (efecto sobre t, 1/n, 1/m)

2) El menor valor de $s_{\hat{x}_0}$ se obtiene cuando la respuesta a interpolar coincide con el **centroide**: o sea en la mitad de la línea de calibrado

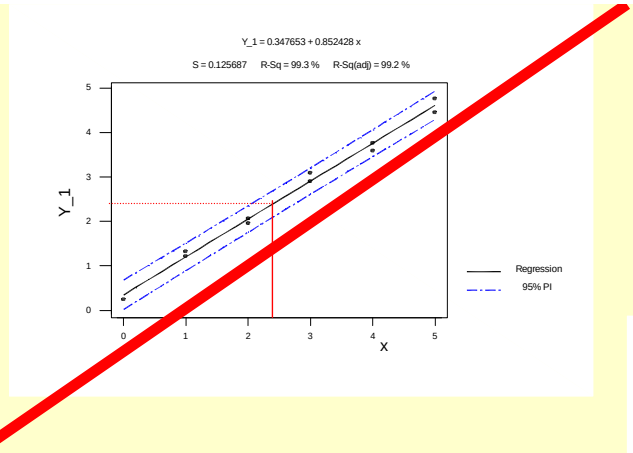
3) El intervalo también disminuye cuando el término

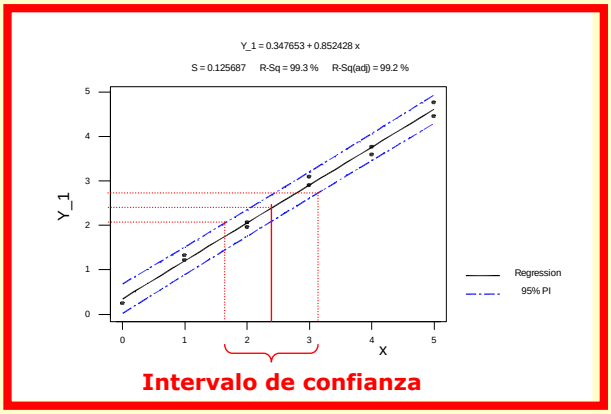
$$\sum (x_i - \bar{x})^2$$

es muy **grande**, lo que obliga a utilizar patrones de muy baja y muy alta concentración lo cual no es aconsejable

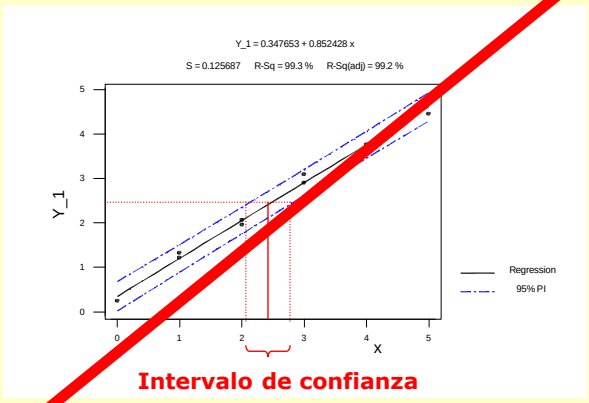
4) Cuanto mayor es b_1 (Sensibilidad según la IUPC) menor es s_{x0}

5) Cuanto menor es s_e (mejor el ajuste) menor es s_{x0}





Intervalo de confianza



Intervalo de confianza

EXCEL

LÍMITE DE DETECCIÓN Y CANTIDADES RELACIONADAS

- Habitualmente, dentro del campo de la Química Analítica, es la concentración x_L o cantidad q_L derivada de la respuesta más pequeña y_L que puede ser detectada, con una **certeza razonable**, mediante un procedimiento analítico dado.

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

siendo S es la pendiente de la línea de calibrado (**Sensibilidad**).

- En realidad debe definirse en el **dominio de la respuesta** o variable observada
- Límite de detección es aquella señal que difiere significativamente de la señal del blanco.
- La **IUPAC** reconoce 3 versiones:

Límite de decisión CCα (Decisión 2002/657/CE

- **Límite de decisión** (decisión de detección) a partir del cual se puede decidir a posteriori si el resultado obtenido indica o no detección del analito

Capacidad de detección CCβ (Decisión 2002/657/CE

- **Límite de detección** al cual se puede confiar a priori que el procedimiento analítico permita detectar el analito
- **Límite de determinación o cuantificación** a partir del cual un procedimiento analítico es capaz de dar un resultado con la suficiente precisión

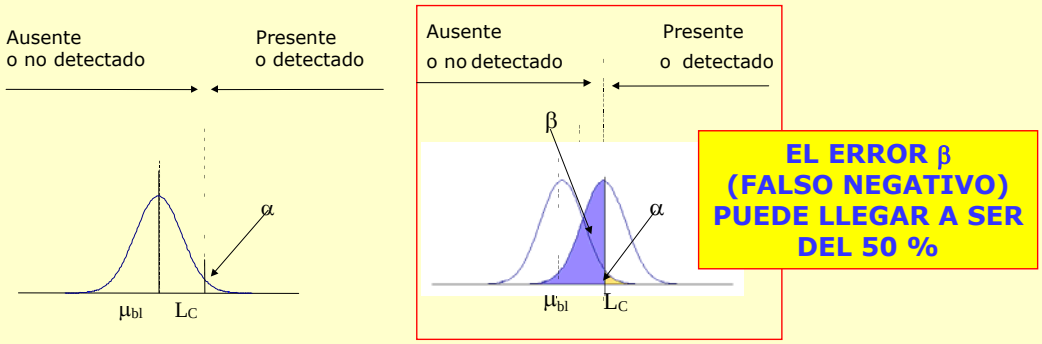
LÍMITE DE DETECCIÓN Y CANTIDADES RELACIONADAS

Límite de decisión (decisión de detección, valor crítico, Límite de decisión CCα)

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

k_c = 2,33 para α = 0,01

- La IUPAC y la ISO proponen **k_c = 1,645**, lo cual significa α = 0,05. Si k_C = 3, coincide numéricamente con la primitiva versión de la IUPAC (α=0,13%)
- Hay un α **% de probabilidades de que una señal mayor o igual que LC (CCα) pertenezca al blanco (FALSO POSITIVO)**, luego se puede concluir que con una elevada probabilidad (1-α = 0,95 ó 95 %) que el componente ha sido detectado



EL ERROR β (FALSO NEGATIVO) PUEDE LLEGAR A SER DEL 50 %

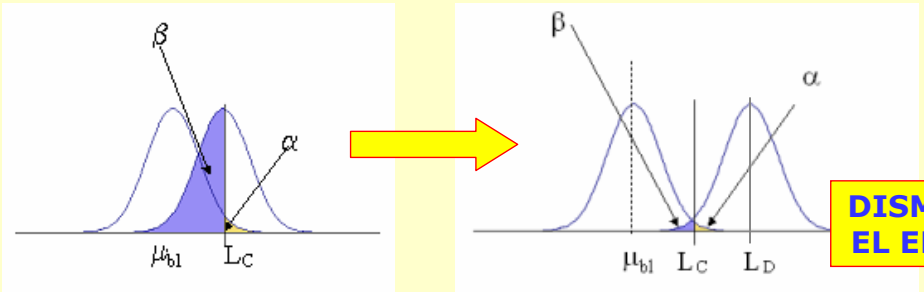
- Este límite ha sido propuesto para tomar una decisión **a posteriori**, es decir después de que se ha medido la respuesta, sobre la presencia de un componente.
- Se le puede definir como **nivel crítico o límite de decisión** por encima del cual una señal puede ser reconocida como detectada

LÍMITE DE DETECCIÓN Y CANTIDADES RELACIONADAS

Límite de detección (mínimo valor detectable, Capacidad de detección CCβ)

- Para reducir el error β no queda más remedio que **separar más las distribuciones** del blanco y el analito.

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



DISMINUYE EL ERROR β

- Para una muestra que **no contiene analito** (su verdadera concentración corresponde a una señal μ_{bl}), menos del α% de las medidas excederán a LC (CCα).
- Para una muestra **conteniendo analito** y con una concentración que origina una respuesta LD, solo el β% de las medidas estarán por debajo de LC y serán indistinguibles del blanco. Por lo tanto, dado LC (CCα), LD (CCβ) **protege contra falsos negativos**.
- La IUPAC propone **α=β=5 % y k'_D vale 3,29** (2 * 1,645). **LD (CCβ) = μ_{bl} + 3,29 σ_{bl}**
- Si k_C=k_D= 3, entonces α=β=0,13 % y LD (CCβ) = μ_{bl} + 6 σ_{bl}
- Este límite puede ser usado **a priori**

LÍMITE DE DETECCIÓN Y CANTIDADES RELACIONADAS

Límite de cuantificación

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

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

- Si DTR es del 5%, $k_Q = 20$.

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

$$D.T.R. = \frac{\sigma_{bl}}{(L_Q - \mu_{bl})} = \frac{1}{k_Q}$$

$$k_Q = \frac{1}{D.T.R.}$$

- La **IUPAC** propone una DTR del 10% por lo que el valor $k_Q = 1/0{,}1 = 10$
- 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

LÍMITE DE DETECCIÓN Y CANTIDADES RELACIONADAS

$$L_C \equiv CC\alpha = \mu_{bl} + k_C \sigma_{bl}$$

$$L_D \equiv CC\beta = \mu_{bl} + k'_D \sigma_{bl}$$

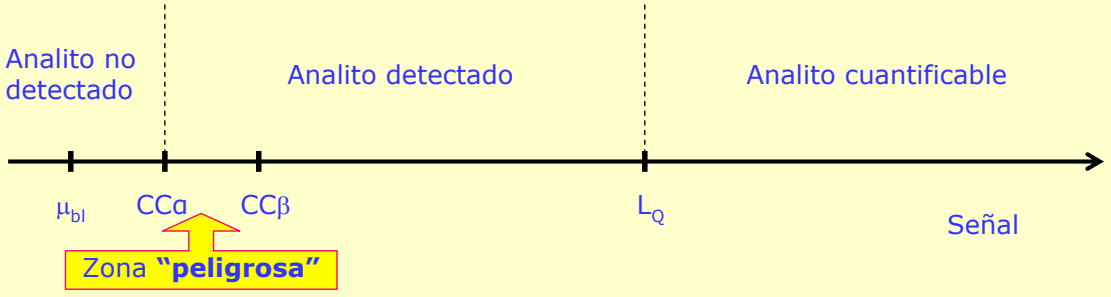
$$L_Q = \mu_{bl} + k_Q \sigma_{bl}$$

Si α y β valen 0,05 y la D.T.R. de L_Q es el 10%

$$L_C \equiv CC\alpha = \mu_{bl} + 1{,}65 \sigma_{bl}$$

$$L_D \equiv CC\beta = \mu_{bl} + 3{,}29 \sigma_{bl}$$

$$L_Q = \mu_{bl} + 10 \sigma_{bl}$$



DETERMINACIÓN DEL BLANCO

- Es fundamental el conocer la estimación adecuada de μ_{bl} y σ_{bl}.
- 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 es un blanco de matriz**

1) 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

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

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:

$$L_C = k_C\sigma_0 = k_C\sqrt{2}\,\sigma_{bl}$$
$$L_D = k'_D\,\sigma_0 = k'_D\sqrt{2}\,\sigma_{bl} \quad \text{siendo} \quad k'_D = k_C + k_D$$

2) Si se determina n veces **un único blanco** de forma separada, la corrección se hace restando a todas las muestras la media de ese blanco

$$y_N = y_{bruta} - \bar{y}_{bl}$$
$$\sigma^2_N = \sigma^2_{bruta} + \sigma^2_{bl}/n$$

y si σ es independiente de la concentración

$$\sigma_0 = \left(\sqrt{1 + (1/n)}\right)\sigma_{bl}$$
$$L_C = k_C\sigma_0 = k_C\sqrt{1 + (1/n)}\,\sigma_{bl}$$
$$L_D = k'_D\,\sigma_0 = k'_D\sqrt{1 + (1/n)}\,\sigma_{bl} \quad \text{siendo} \quad k'_D = k_C + k_D$$

- Si se usan **pocas réplicas,** k_C, k_D y k_Q deben sustituirse por valores de t con n-1 g.d.l.

LÍMITES DE CONCENTRACIÓN

- Los límites en función de la respuesta pueden transformarse en límites de **concentración** a partir de la pendiente de la línea de calibrado:

$$X_{LC} = \frac{k_C\,\sigma_e}{b_1} \qquad y \qquad X_{LD} = \frac{k'_D\,\sigma_e}{b_1}$$

- En esas fórmulas se supone que el **blanco es bien conocido** ya que su variabilidad no se tiene en cuenta. Si este no es el caso, las fórmulas varían ligeramente

$$x_c = \sqrt{1 + (1/n)}\,k_c\sigma_{bl}/b_1$$
$$x_D = \sqrt{1 + (1/n)}\,k'_D\sigma_{bl}/b_1$$

- Si se ha llevado a cabo una línea de calibrado y **el modelo está bien elegido**, se puede asimilar σ_{bl} con σ_e, y μ_{bl} con b_0 de manera que:

$$L_C = \mu_{bl} + k_C\sigma_{bl} = b_0 + k_C\sigma_e$$
$$L_C = b_0 + b_1X_{LC}$$
$$X_{LC} = \frac{k_C\,\sigma_e}{b_1} = \frac{1{,}65\,\sigma_e}{b_1}$$

$$L_D = \mu_{bl} + k'_D\sigma_{bl} = b_0 + k'_D\sigma_e$$
$$L_D = b_0 + b_1X_{LD}$$
$$X_{LD} = \frac{k'_D\sigma_e}{b_1} = \frac{3{,}29\,\sigma_e}{b_1}$$

LÍMITES DE CONCENTRACIÓN A PARTIR DE LA LÍNEA DE CALIBRADO

La línea de calibrado es solo una **estimación de la verdadera línea de regresión**, debe tenerse en cuenta su incertidumbre.

1) 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 x_0 es cero

$$b_0 + b_1 x_0 \pm t_{n-2}\, s_{y/x} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = b_0 + t_{n-2}\, s_{y/x} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

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

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

$$y_L = b_0 + b_1 x_D - t_{\beta,n-2}\, s_{y/x} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_D - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

que es un método engorroso

2.2) 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.

2.3) De forma aproximada (AOAC)

$$x_D = x_C + t_{\beta,n-2}\, s_{y/x} / b_1 \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(2x_C - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

siendo $x_C = (y_C - b_0)/b_1$

3) A **partir de x_D** puede calcularse también $y_D = b_0 + b_1 x_D$

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 $d = (t_{1-\alpha/2} + t_{1-\beta})\, s \sqrt{2}\, (1/b_1)$

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 (ESPURIOS)

- Son observaciones atípicas (**espurias**) que no son representativas del ajuste. Se detectan

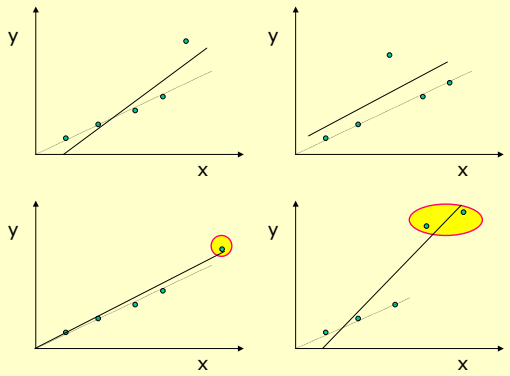
  1) Si el **valor absoluto** del residual estandarizado: $|e_i / s_e| \geq 2$ ó 3, el punto se rechaza.

  2) Calculando la distancia de Cook al cuadrado (**Cook's square distance**)

  El **valor de corte suele ser 1**

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

nº de parámetros



- Existen además los "outliers" de **nivelación** ("leverage") que son puntos extremos que tienen una gran influencia sobre los parámetros

APLICACIONES DE INTERÉS ANALÍTICO

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 de calibrado lineal.

- Una solución es el **método de adiciones patrón (MAP o SAM)**: 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) y se observa la variación producida en la Respuesta.

- Sobre V_0 mL de problema de concentración C_0 desconocida, se hacen adiciones iguales V_p de un patrón de concentración C_p. La **concentración resultante** es en cada momento:

$$C_i = (V_0C_0 + V_pC_p)/(V_0 + V_p)$$

- Si se mide la Respuesta, relacionada **linealmente** con la concentración del componente buscado a través de una relación del tipo: $R = k\,C$, ésta valdrá en cada momento:

$$R_i = k\,C_i = k\,(V_0C_0 + V_pC_p)/(V_0 + V_p)$$

$$Q_i = R_i\,(V_0+V_p) = k\,(V_0C_0+V_pC_p)$$

  siendo Q la **señal corregida por la dilución** (por tanto proporcional a la cantidad de componente).

- Si se **representa Q_i frente a V_p** se debe obtener una **recta**, cuya extrapolación a cero origina un Volumen negativo $\mathbf{V_{eq}}$, a partir del cual se puede conocer la concentración C_0:

$$Q_i = k\,(V_0C_0 + V_{eq}C_p) = 0;\ (V_0C_0 + V_{eq}C_p) = 0$$

$$C_0 = -\,(V_{eq}\,C_p)/V_0$$

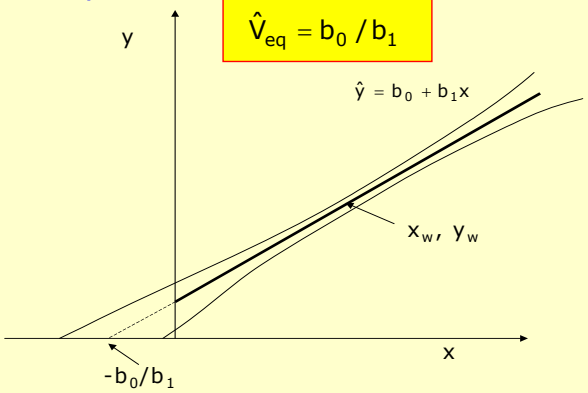
- El **valor absoluto de V_{eq} (Volumen equivalente)** es el volumen de la disolución patrón añadida que contiene la misma cantidad de analito que la muestra.

APLICACIONES DE INTERÉS ANALÍTICO

Método de adiciones patrón

- Si se **ajusta Q_i al V_p mediante una regresión** según $Q_i = b_0 + b_1 V_p$, V_{eq} se estima

$$\hat{V}_{eq} = b_0 / b_1$$



- Siempre debe hacerse una **determinación en blanco**, cuyo volumen equivalente debe sustraerse del correspondiente a la muestra (y cuya **varianza se suma** a la de V_{eq})
- Si la **dilución producida es pequeña o inexistente**, no es precisa corregir la respuesta por el volumen, y se representa directamente R_i vs. V_p
- La **incertidumbre** del V_{eq} el error se calcula:

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

- Un problema del MAP es que al **extrapolar** se trabaja en una zona en la cual la imprecisión es muy grande.

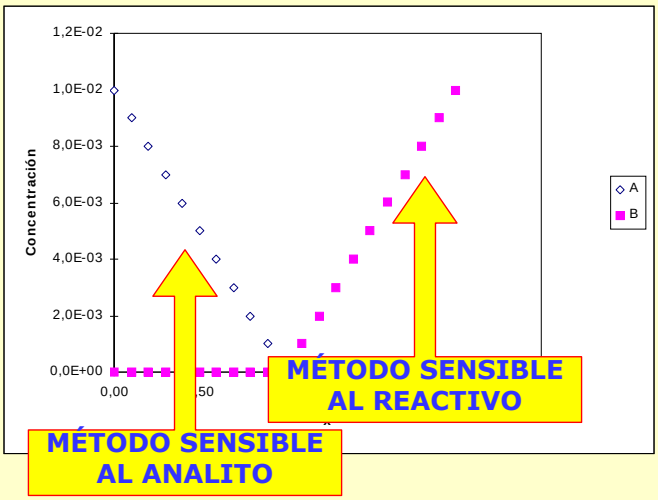
APLICACIONES DE INTERÉS ANALÍTICO

Método de sustracciones patrón

- Se determina el analito adicionando a la muestra cantidades conocidas de un reactivo patrón que **reaccione cuantitativamente** con él.
- De esta manera se origina una **disminución lineal de la cantidad de analito**, por lo que se puede seguir la respuesta obtenida en función de la cantidad de reactivo añadido y se puede aplicar el método de regresión a una línea recta.
- La **cantidad estequiométrica de reactivo** que reacciona con el analito presente en la muestra, x_{eq}, se estima extrapolando a ordenada cero la línea. En ausencia de errores sistemáticos absolutos,

$$\hat{x}_{eq} = -b_0 / b_1$$

- En ocasiones, **la respuesta medida es función del reactivo añadido**, por lo que dicha respuesta comenzará a **aumentar después** del punto estequiométrico. También en este caso, la cantidad x_{eq} se estima extrapolando a cero la línea recta obtenida.



APLICACIONES DE INTERÉS ANALÍTICO

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 una prueba de significación (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:

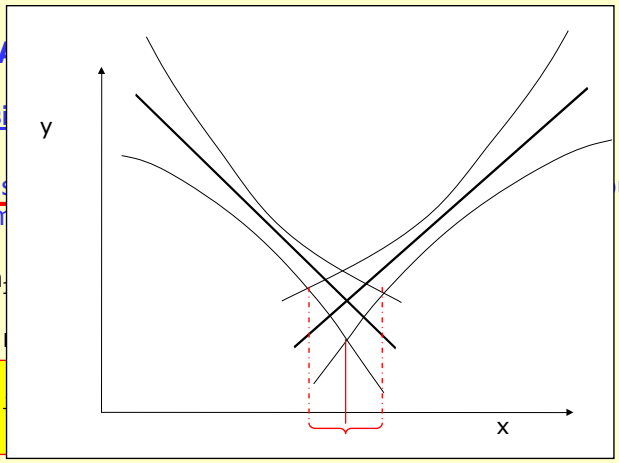
$$t' = \frac{t_1 s_{b_{11}}^2 + t_2 s_{b_{12}}^2}{s_{b_{11}}^2 + s_{b_{12}}^2}$$

- La prueba puede hacerse de **forma rápida**: si los intervalos de confianza de b_{11} y b_{12} no se solapan las pendientes difieren significativamente, y viceversa.

COMPARACIÓN DE LAS SENSIBILIDADES DE DOS PROCEDIMIENTOS

APLICACIONES DE INTERÉS ANA

Intersección de dos líneas de regresi



- En ocasiones, como en las valoraciones unto de **corte de dos líneas rectas**, que m

Línea 1: $y_1 = b_0 + b_1 x_1$ con n

Línea 2: $y_2 = b_0' + b_1' x_1$ con

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

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

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

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

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

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

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

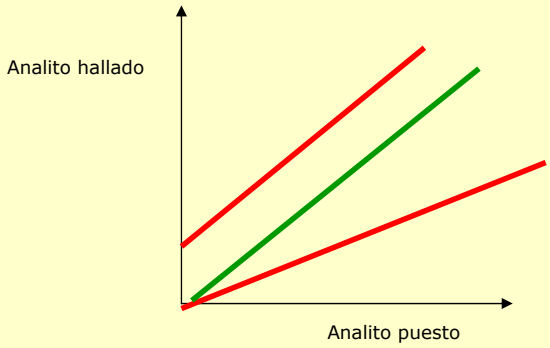
- El **valor de t** es el tabulado con n_1+n_2-4 g.d.l.

APLICACIONES DE INTERÉS ANALÍTICO

Ensayos de recuperación

ALTERNATIVA A LA PRUEBA DE SIGNIFCACIÓN

- Sirven para **validar** un procedimiento analítico
- Se analizan **blancos dopados o fortalecidos (spiked)** con concentraciones diferentes de analito **exactamente** conocidas.
- Se representa la **concentración encontrada frente a la añadida**



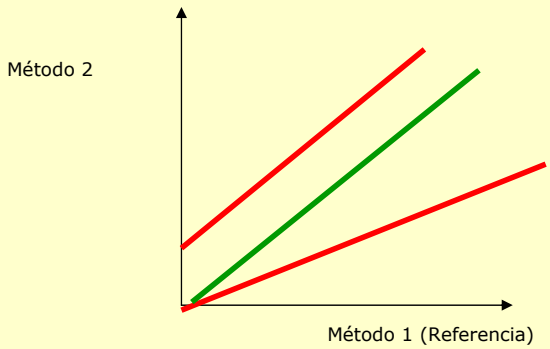
- Al hacer la regresión, se debería obtener una línea recta de ordenada en el origen cero y pendiente uno, o bien los **intervalos de confianza de b_0 y b_1 deben incluir a cero y a uno,** respectivamente.
- Si la ordenada en el origen es mayor de cero existe un error sistemático (bias, sesgo) constante (o una corrección incorrecta del blanco)
- Si la pendiente difiere de uno, existe un error sistemático (bias, sesgo) proporcional que suele ser debido a la matriz.

APLICACIONES DE INTERÉS ANALÍTICO

Comparación de métodos

ALTERNATIVA A LA PRUEBA DE SIGNIFCACIÓN

- Sirve para comparar **dos métodos o para validar** uno si el otro está validado
- Se analiza la misma muestra mediante los **dos procedimientos**.
- Se representa la **concentración (cantidad) hallada por un método frente a la del otro.** En abcisas se representan los resultados del método validado si alguno lo fuere

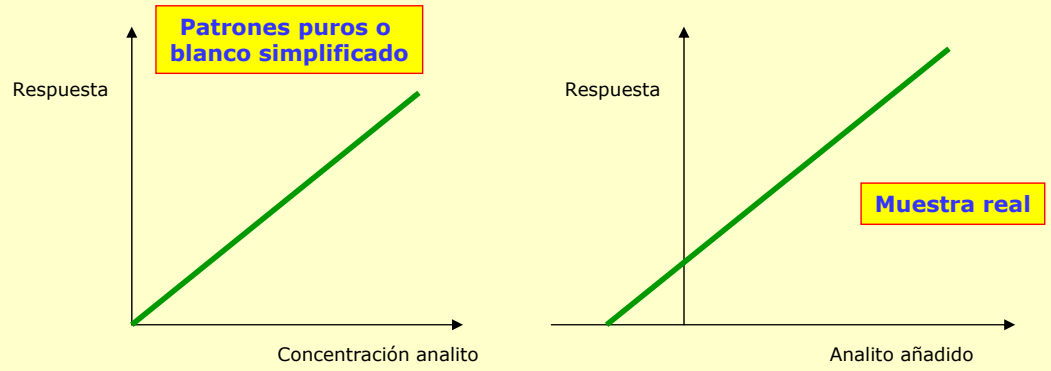


- Al hacer la regresión, se debería obtener una línea recta de ordenada en el origen cero y pendiente uno, o bien los intervalos de **confianza de b_0 y b_1 deben incluir a cero y a uno**, respectivamente.
- Si la ordenada en el origen es mayor de cero hay una diferencia sistemática (bias, sesgo) constante (o un bias o sesgo sistemático del Método 2)
- Si la pendiente difiere de uno, existe una diferencia sistemática (bias, sesgo) proporcional

APLICACIONES DE INTERÉS ANALÍTICO

Detección de efectos de matriz **¿Hay que aplicar el M.A.P.?**

- Los efectos de matriz se buscan **comparando las pendientes** de una **línea de calibrado** hecha con patrones puros, y la de un M.A.P. realizado sobre una **muestra real**
- Las **concentraciones** de analito en ambos casos deben 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 pueden hacerse por calibrado lineal.
- Si las **pendientes son diferentes, hay efecto de matriz**: Es necesario utilizar el M.A.P.



APLICACIONES DE INTERÉS ANALÍTICO

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:

$$y_0 = b_0 + b_1 x_0 \quad \Rightarrow \quad b_0 = y_0 - b_1 x_0$$

y el modelo también debe cumplir

$$\hat{y} = b_0 + b_1 x = (y_0 - b_1 x_0) + b_1 x = y_0 + b_1 (x - x_0)$$

- 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

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

APLICACIONES DE INTERÉS ANALÍTICO

Regresión a través de un punto fijo

- 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 sólo debe ser utilizado cuando haya **buenas razones a priori** para ello.

APLICACIONES DE INTERÉS ANALÍTICO

Linearización de funciones curvas

- Cuando la relación entre las dos variables **no puede ser representada por una línea recta**, cabe la posibilidad de hacer una ajuste polinómico (**Regresión Lineal Múltiple**) o un ajuste no lineal. Una alternativa es **transformar una o ambas variables**, de manera que se obtenga una relación más sencilla.

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

- 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**.

APLICACIONES DE INTERÉS ANALÍTICO

Linearización de funciones curvas

- De **forma general, si una variable y se transforma** por medio de $y_i = f(y_i)$, la expresión para calcular los pesos es:

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

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

- Por ejemplo, si se transforma **x en ln x**, se tiene

$$s^2_{\ln x} = \left(\frac{d(\ln x)}{dx}\right)^2 s^2_x = \left(\frac{1}{x^2}\right) s^2_x$$

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

- Este es el caso de la **Absorbancia y la Transmitancia**. A = -log T

$$s^2_A = \left(\frac{d(-\log T)}{dT}\right)^2 s^2_T = \left(\frac{2{,}30^2}{T^2}\right) s^2_T$$

$$w_i = 1 / s^2_{A_i} = T_i^2 / s^2_{T_i}$$

El factor $2{,}30^2$ es cte. y no se necesita

CORRELACIÓN Y REGRESIÓN

- 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**.
- Sean **2 parámetros determinados sobre n muestras**: $y_{11},\ldots,y_{1n}$ e $y_{21},\ldots,y_{2n}$ con sus respectivas medias $\bar{y}_1$ e $\bar{y}_2$. Se define la covarianza como:

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

- Varía entre **-∞** cuando las variables están asociadas **negativamente** (una disminuye cuando la otra aumenta) y **+∞** cuando la asociación **es positiva**.
- **Coeficiente de correlación producto-momento** o coeficiente de correlación de **Pearson**:

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

  Varía entre **–1 y +1** dependiendo del tipo de asociación. Valores de –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 significa que no hay relación entre ellas, sino que esta **relación no es lineal**.

CORRELACIÓN vs. REGRESIÓN

- **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 r^2

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SS_{REG}}{SS_T}$$

coeficiente de determinación: 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**. Lo verdaderamente útil **no es r sino r^2** que expresa la proporción de variación explicada por la regresión

REGRESIÓN CON AMBAS VARIABLES SUJETAS A ERROR

- Hasta ahora se ha **supuesto que sólo la respuesta estaba sujeta a error** y que la variable independiente, x, se conocía exactamente (Regresión Modelo I).

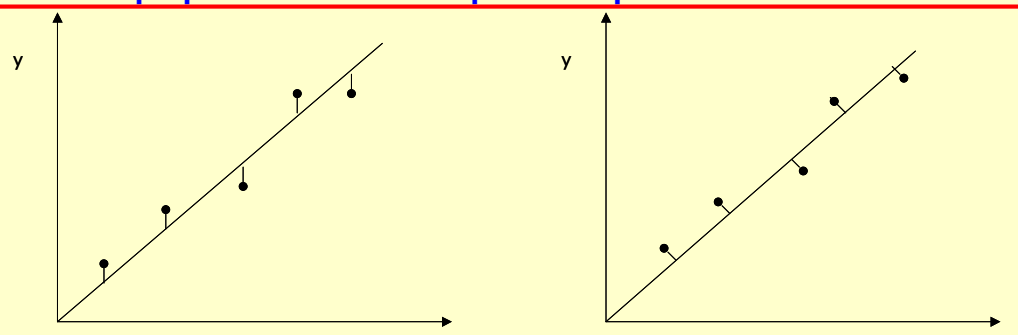
- Hay situaciones en las que **no es asumible que x esté libre de errores** (p.e. comparación de resultados de dos métodos, o patrones preparados con gran incertidumbre). Si η_i es el verdadero valor de $\mathbf{y_i}$ y ξ_i el verdadero valor de $\mathbf{x}_i$, entonces

$$y_i = \eta_i + \varepsilon_i$$
$$x_i = \xi_i + \delta_i$$

- El modelo que describe la relación lineal será $\eta_i = \beta_0 + \beta_1\xi_i$ y por tanto

$$y_i = \beta_0 + \beta_1(x_i - \delta_i) + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_i + (\varepsilon_i - \beta_1\delta_i)$$

- Si se supone que $\sigma_\varepsilon^2 = \sigma_\delta^2$, para obtener estimaciones insesgadas de los coeficientes de la regresión, lo que debe minimizarse es d_i^2 es decir la **suma de los cuadrados de las distancias perpendiculares de los puntos experimentales a la línea de regresión**.



REGRESIÓN CON AMBAS VARIABLES SUJETAS A ERROR

Comparación de la regresión por mínimos cuadrados y ortogonal

- El método recibe el nombre de regresión de la distancia ortogonal (**orthogonal distance regresión ORD**), y los valores de los coeficientes se obtienen mediante las siguientes fórmulas:

$$b_1 = \frac{s_y^2 - s_x^2 + \sqrt{(s_x^2 - s_y^2)^2 + 4(\text{cov}(y,x))^2}}{2\,\text{cov}(y,x)}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

- siendo s_y^2 y s_x^2 las varianzas respectivas de las variables y x, y cov(y,x) la covarianza de y y x calculadas con

$$s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

$$\text{cov}(y,x) = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

- Existe una **relación aproximada** entre los valores de b_1 obtenidos por mínimos cuadrados (MC) y por regresión ortogonal (ORD):

$$b_1(\text{ORD}) = \frac{b_1(\text{MC})}{(1 - \frac{s_{ex}^2}{s_x^2})}$$

- siendo s_{ex}^2 la varianza de un valor de x **individual** (es decir se observa un mismo valor de x varias veces) y s_x^2 la varianza de la variable x, que solo depende del intervalo estudiado de x y de su distribución. Si el cociente **s_{ex}^2/s_x^2 es mayor de 0,2** se pueden esperar errores significativos en la estimación de b_1 por mínimos cuadrados.