Regressione lineare semplice

In chimica analitica strumentale la relazione esistente fra il segnale ottenuto e la concentrazione dell'analita che lo genera è solitamente lineare (almeno per un certo intervallo di concentrazione.

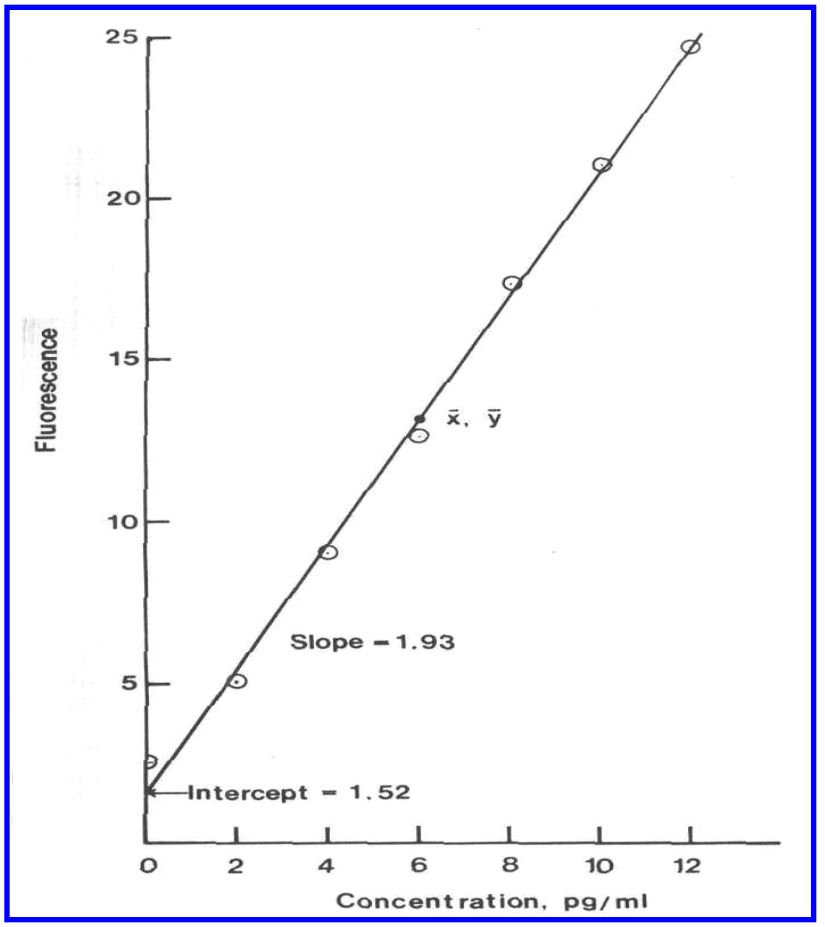
Detto Y il segnale, X la concentrazione ed ε l'errore associato alla misura di Y si può quindi scrivere l'equazione:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Dove β_0 e β_1 rappresentano i valori veri del modello che correla Y a X.

Dette b_0 e b_1, rispettivamente, le migliori stime di β_0 e β_1, ottenute da un metodo di interpolazione, i valori del segnale predetti dal modello sono dati da:

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



Metodo dei minimi quadrati

Uno dei metodi più comuni per ricavare le stime b_0 e b_1 è il metodo dei minimi quadrati.

Siano date n coppie di valori sperimentali:

$(x_1,y_1), (x_2, y_2), \ldots, (x_n, y_n)$;

per la generica coppia di valori (x_i,y_i) risulta:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ mentre $\hat{y}_i = b_0 + b_1 x_i$

L'approccio dei minimi quadrati ha per obiettivo determinare i valori di b_0 e b_1 tali che la somma degli scarti quadratici fra i valori di y_i e $\hat{y}_i$ sia minima.

I valori di b_0 e b_1 derivano quindi dalla soluzione delle equazioni:

$$\frac{\partial\left[\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2\right]}{\partial b_0} = 0 \qquad \frac{\partial\left[\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2\right]}{\partial b_1} = 0$$

Risolvendo il sistema di equazioni si ottiene:

$$b_1 = \frac{\sum_i^n X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_i^n X_i^2 - n\,\overline{X}^2}$$

equivalente a:

$$b_1 = \frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i^n (X_i - \overline{X})^2}$$

mentre:

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Considerando la relazione fra b_0 e b_1, l'equazione della regressione lineare ottenibile con il metodo dei minimi quadrati si può esprimere come:

$$\hat{y} = \overline{y} + b_1 (x - \overline{X})$$

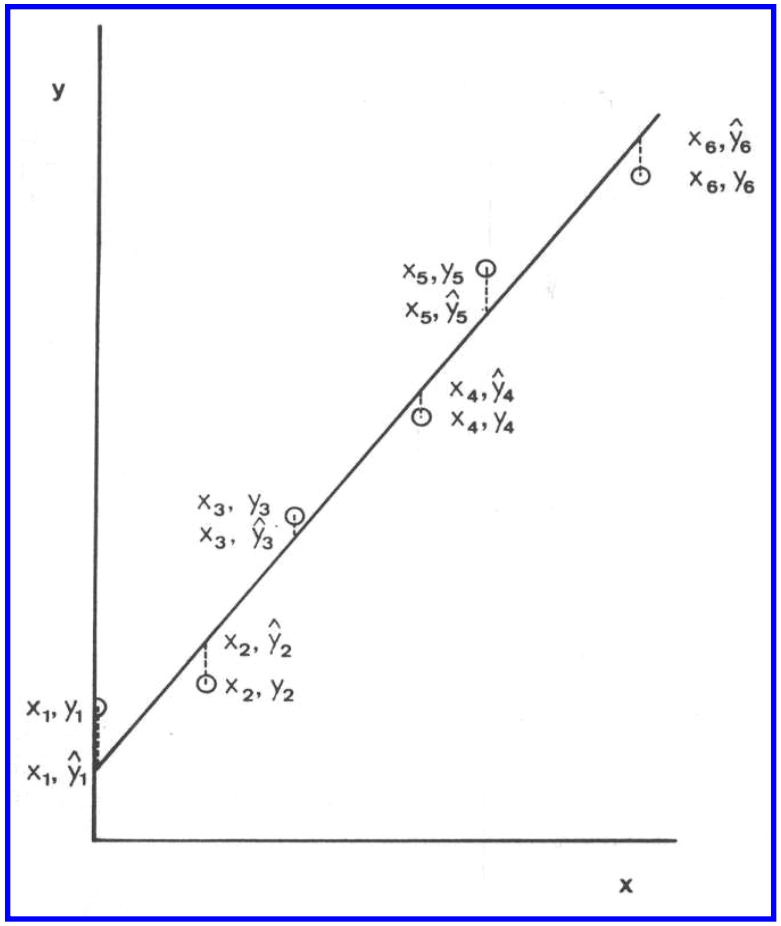
Nel caso del dato generico (x_i, y_i) l'equazione della regressione diventa:

$$\hat{y}_i = \overline{y} + b_1 (x_i - \overline{X})$$

sottraendo entrambi i membri dell'equazione dal valore y_i si ottiene:

$$y_i - \hat{y}_i = y_i - \overline{y} - b_1 (x_i - \overline{X})$$

il termine al primo membro viene definito residuo, ossia la differenza fra valore sperimentale della y e valore predetto dal modello a parità di concentrazione.



Sommando i residui su tutti i dati si ha:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)=\sum_{i=1}^{n}(y_i-\overline{Y})-b_1\sum_{i=1}^{n}(x_i-\overline{X})$$

Per definizione di media entrambe le sommatorie al secondo membro sono nulle, ad esempio risulta:

$$\sum_{i=1}^{n}(y_i-\overline{Y})=\sum_{i=1}^{n}y_i-n\overline{Y}=n\overline{Y}-n\overline{Y}=0$$

In definitiva: la sommatoria dei residui relativi ad una retta di regressione ottenuta con il metodo dei minimi quadrati è sempre nulla (il che significa che vi saranno sempre sia residui positivi che residui negativi, che alla fine si compenseranno).

Coefficiente di correlazione nella regressione lineare

Se si applica la definizione generale di coefficiente di correlazione al caso della regressione lineare si ottiene la relazione:

$$r = \frac{C(x,y)}{\sqrt{V(x)V(y)}} = \frac{\sum_{i=1}^{n}\{(x_i - \bar{x})(y_i - \bar{y})\}}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

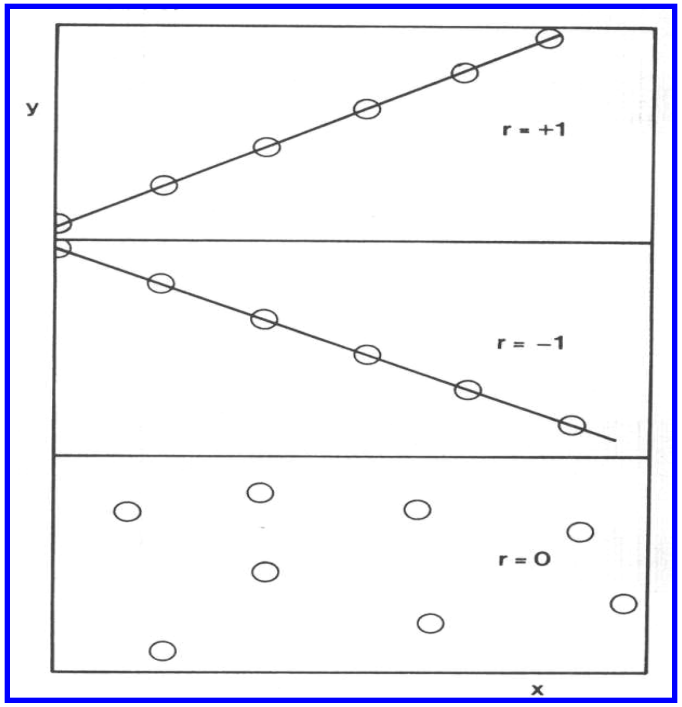
Un'attenta analisi dell'equazione mostra che $-1 \leq r \leq 1$.

$r = 1 \Rightarrow$ perfetta correlazione fra x e y

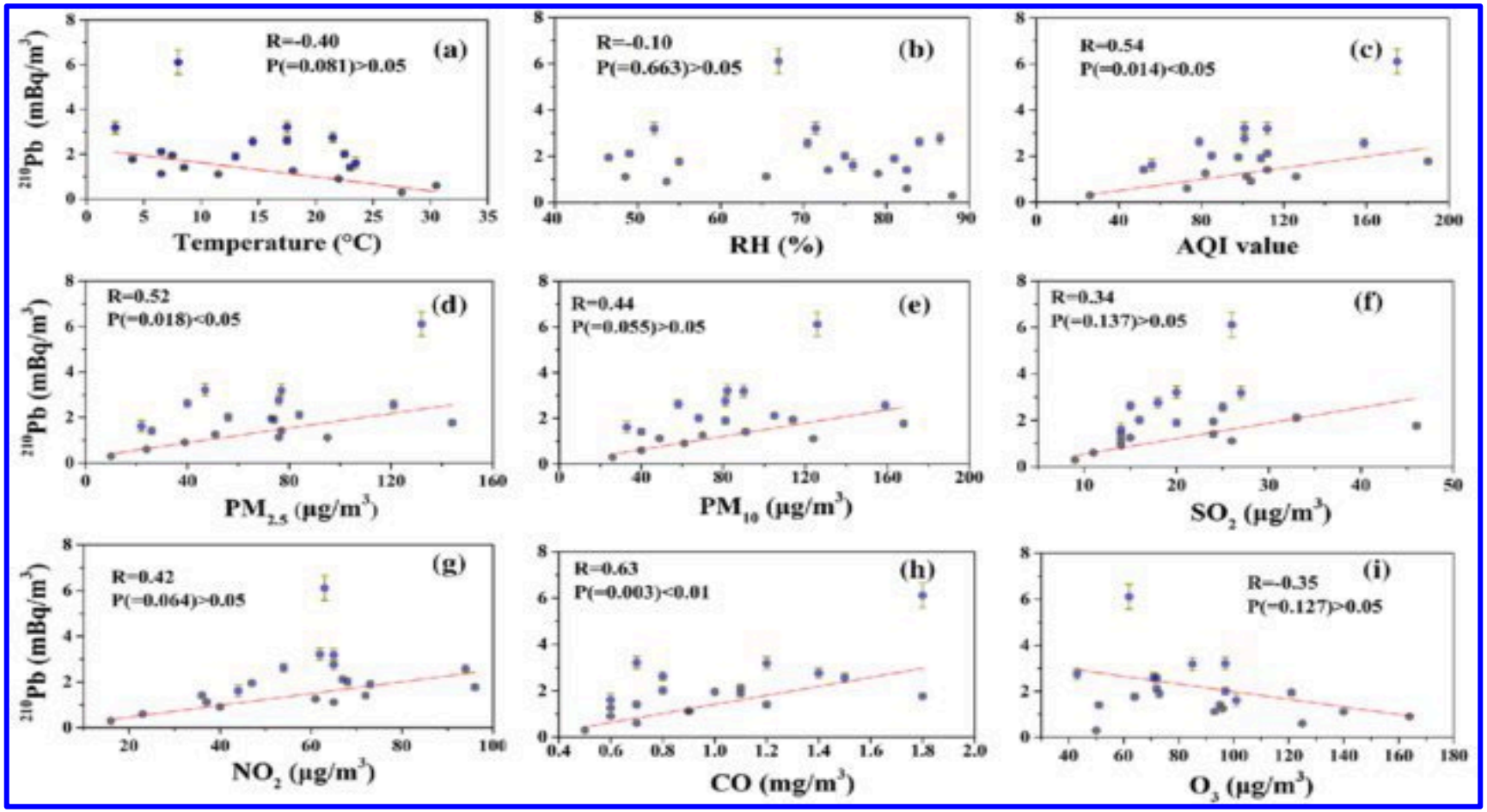
$r = -1 \Rightarrow$ perfetta correlazione negativa fra x e y

$r = 0 \Rightarrow$ nessuna correlazione lineare fra x e y

Le rette di taratura in chimica analitica forniscono frequentemente valori di r superiori a 0.99, mentre già valori inferiori a 0.90 sono raramente ritenuti accettabili.

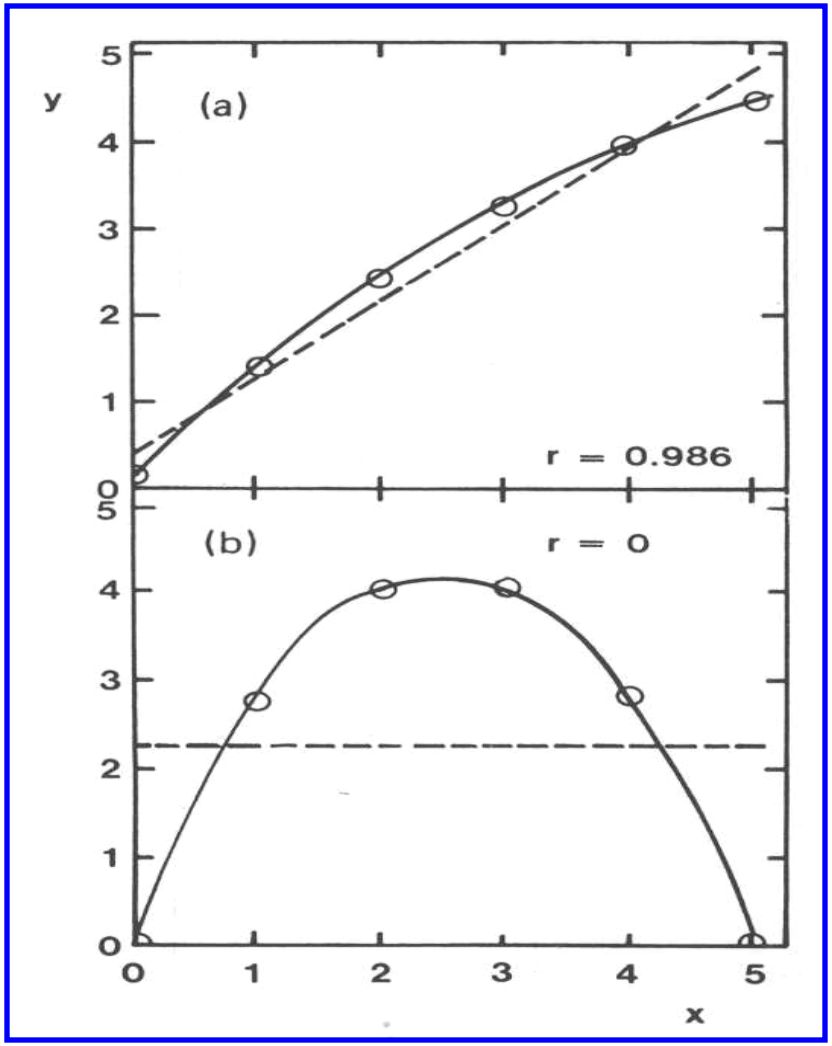


In altri contesti di ricerca, tuttavia, il coefficiente di correlazione può essere decisamente più basso rispetto ad 1 (in valore assoluto). Ciò accade spesso, ad esempio, quando si interpolano coppie di valori riferiti a grandezze di natura ambientale sulle quali incidono altre cause di variabilità oltre a quella che tende a farle aumentare o diminuire congiuntamente:



Bq = becquerel = attività di un radionuclide che ha un decadimento al secondo
AQI = Air Quality Index; RH = Room Humidity

Occorre fare attenzione al fatto che spesso valori apparentemente accettabili di r in realtà derivano da dati che non sono correlati linearmente (caso a):



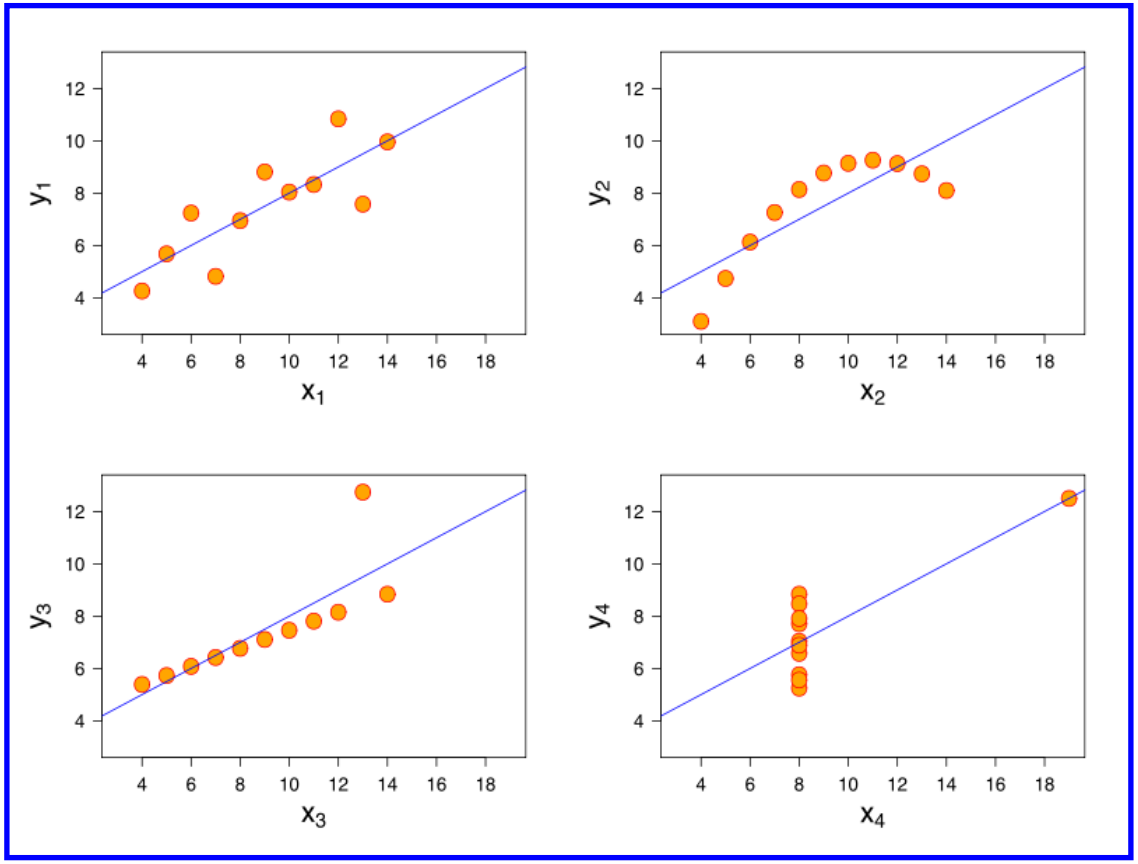
D'altra parte un valore di r praticamente nullo (caso b) significa solo che x e y non sono linearmente correlate ma questo non significa che esse non siano correlate da una funzione di ordine superiore.

L'analisi della distribuzione dei residui lungo l'intervallo di variazione di x può aiutare a comprendere la situazione.

Il quartetto di Anscombe

Nel 1973 lo statistico F.J. Anscombe pubblicò un articolo in cui mostrava come quattro set di dati molto diversi fra loro conducessero alla stessa retta di regressione, con r = 0.816:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



E' quindi opportuno sempre rappresentare graficamente i dati prima di avanzare ipotesi sulla bontà della regressione lineare.

Precisione della regressione lineare

I dati indispensabili per stabilire la precisione della regressione lineare sono le incertezze che caratterizzano la pendenza e l'intercetta della retta di regressione.

Un parametro fondamentale per la loro valutazione è:

$$s_{y/x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

ossia la deviazione standard sui residui della regressione.

Si può dimostrare che se il modello di regressione è valido questa grandezza è uno stimatore corretto della varianza associata ai dati sperimentali sottoposti alla regressione.

Le deviazioni standard sulla pendenza e sull'intercetta dipendono anche da $s_{y/x}$:

$$s_{b_1} = \frac{s_{y/x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad s_{b_0} = s_{y/x}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Per l'espressione corretta dei valori di b_0 e b_1 occorre considerare il loro intervallo di fiducia e quindi usare:

- ✓ la distribuzione t di Student ad n-2 gradi di libertà e ad un certo livello di significatività α
- ✓ i valori delle rispettive deviazioni standard:

$b_0 \pm t_{n-2,1-\alpha/2} \times s_{b_0}$ $\qquad$ $b_1 \pm t_{n-2,1-\alpha/2} \times s_{b_1}$

tipicamente si adotta $\alpha = 0.05$ (95% di fiducia / 5% di significatività).

E' importante notare che in questo caso le deviazioni standard su b_0 e b_1 contengono già il contributo del termine contenente i gradi di libertà, perché presente al denominatore di $s_{y/x}$, da cui esse dipendono linearmente.

Calcolo di una concentrazione a partire dal responso analitico

Dopo aver calcolato i parametri della retta di regressione è possibile risalire al valore di concentrazione x_0 associato ad un segnale analitico y_0:

$$x_0 = (y_0 - b_0)/b_1$$

La deviazione standard associata a tale valore è data dall'equazione:

$$s_{x_0} = \frac{s_{y/x}}{b_1}\left[\frac{1}{m} + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^{1/2}$$

nella quale compare anche il numero m degli eventuali replicati effettuati sul campione di cui si vuole valutare la concentrazione x_0, mentre n rappresenta il numero complessivo dei dati impiegati per la regressione lineare.

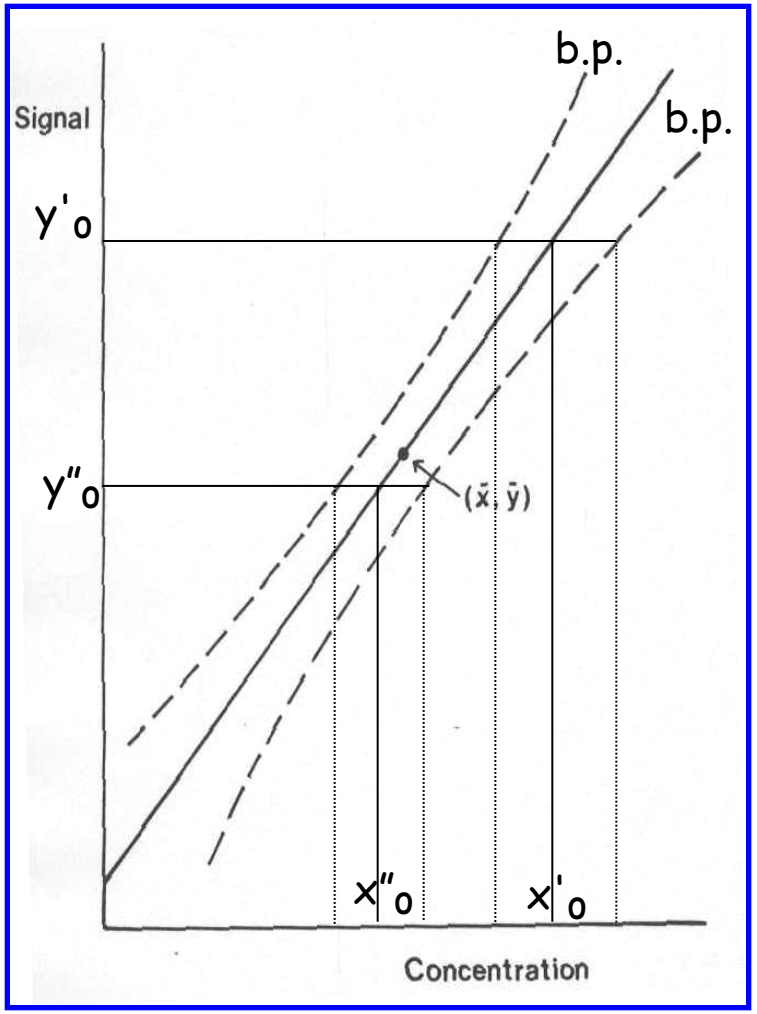
Anche in questo caso occorre calcolare l'intervallo di fiducia sul valore di x_0 usando la distribuzione t di Student, con gradi di libertà n+m-3:

$$x_0 \pm t_{n+m-3,1-\alpha/2} \times s_{x_0}$$

E' interessante evidenziare che l'intervallo di fiducia sulla concentrazione estrapolata può essere determinato graficamente, con buona approssimazione, usando le cosiddette bande di predizione (b.p.), come mostrato in figura.

Esse esprimono l'escursione stimata per il segnale y (entro un certo livello di fiducia) in corrispondenza di un dato valore di concentrazione e si avvicinano al massimo alla retta di interpolazione in corrispondenza del punto avente come coordinate le medie dei valori di y e di x.

A parità di tutte le altre condizioni l'intervallo di fiducia su x_0 tende quindi ad aumentare quando più i valori di y_0 utilizzati si allontanano dalla media $\bar{y}$:



Metodo dell'aggiunta standard

Quando è necessario misurare la concentrazione di un analita in una matrice molto complessa la calibrazione in solvente puro (calibrazione esterna) può portare a risultati di bassa accuratezza a causa dell'effetto matrice.

L'effetto matrice è quello che gli altri componenti della matrice possono esercitare sul responso di un analita, rendendolo significativamente diverso (per eccesso o per difetto) rispetto a quello ottenuto in solvente puro, a parità di concentrazione.

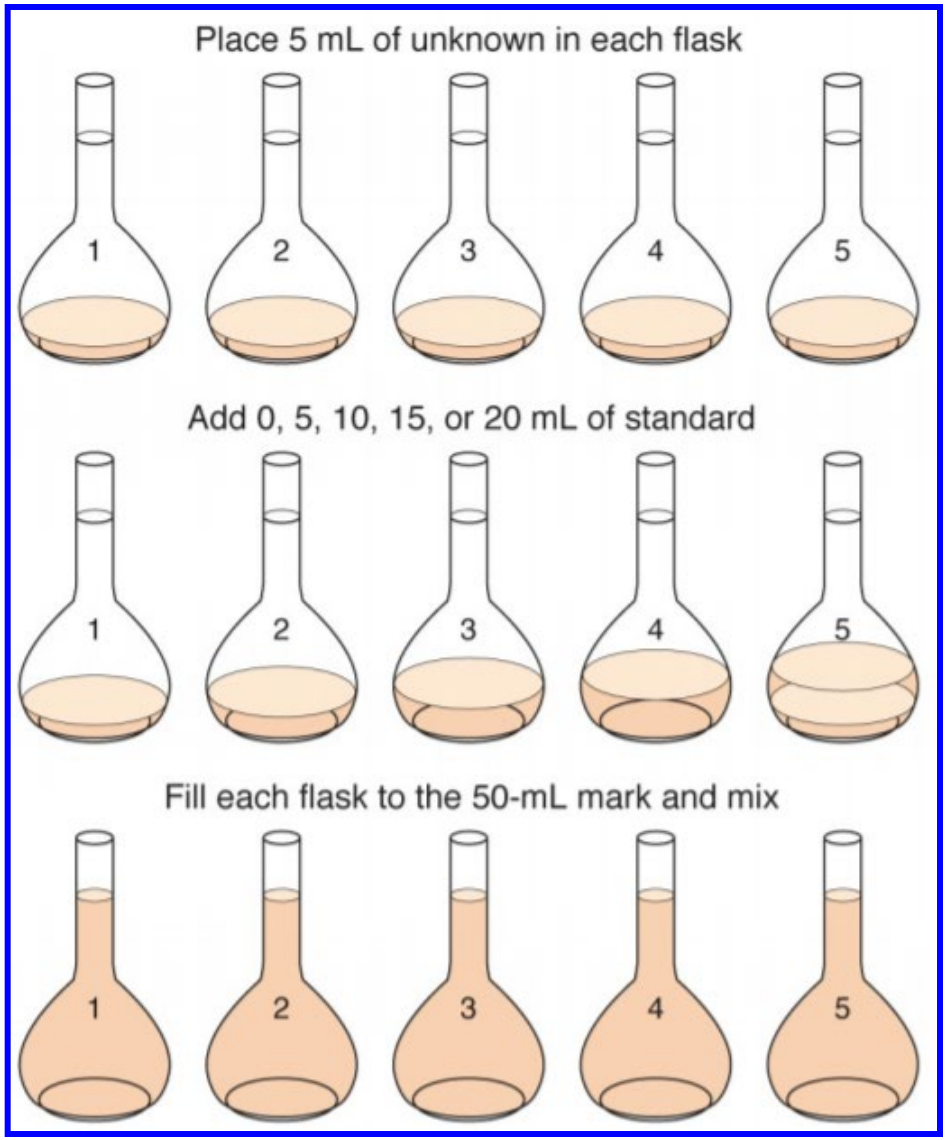
In presenza di tale effetto si può procedere con il cosiddetto metodo dell'aggiunta standard.

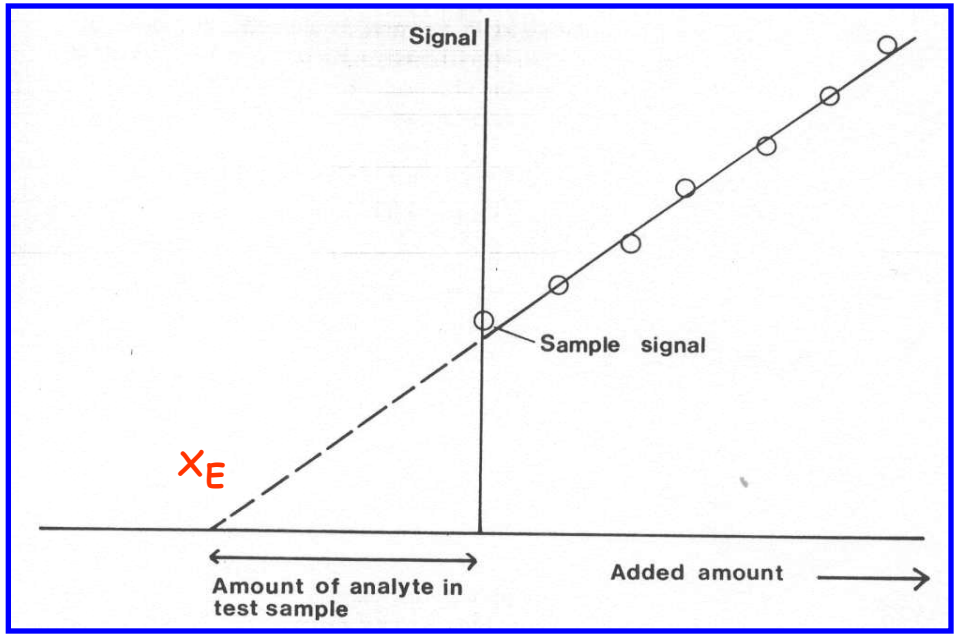
Con tale metodo si misura il responso dell'analita direttamente in matrice, attraverso l'aggiunta a questa di opportuni volumi di uno standard concentrato dell'analita.

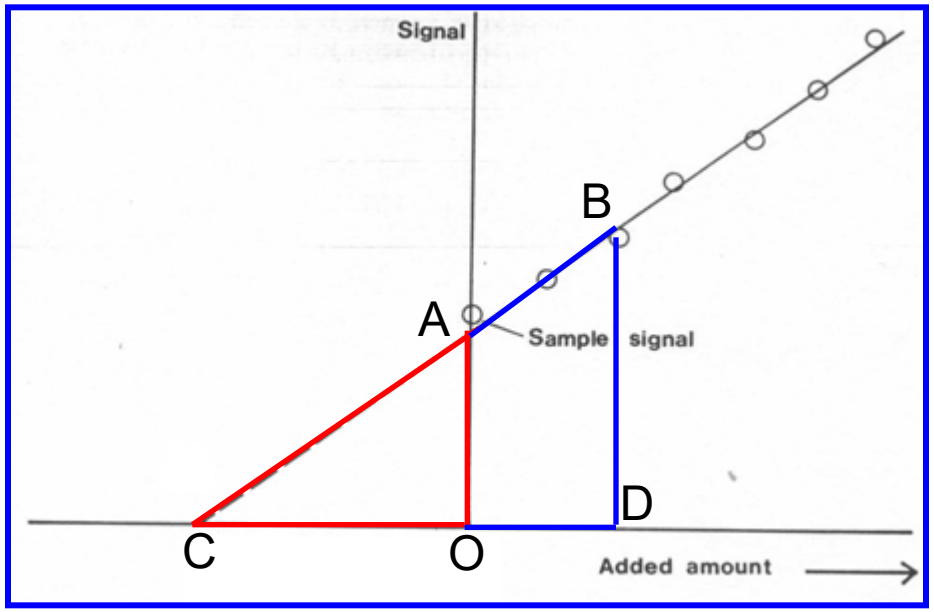
Di fatto si esegue una calibrazione in matrice ma, contestualmente, si può estrapolare il valore della concentrazione incognita dell'analita presente inizialmente nella matrice, attraverso un'opportuna interpolazione dei dati mediante la regressione lineare.

Si procede come segue:

- ✓ si prelevano più aliquote della soluzione da analizzare e le si suddividono in altrettanti contenitori;
- ✓ a ciascuna aliquota, tranne una (la 1) , si aggiunge un volume noto di una soluzione standard concentrata di analita, in modo da ottenere una concentrazione nota di analita aggiunto (che sarà 0 nel primo caso);
- ✓ si porta a volume nei vari casi, si analizza ciascuna delle soluzioni, compresa quella senza aggiunta di analita, e si riporta in grafico il segnale in funzione della concentrazione aggiunta.







L'intercetta sull'asse x (x_E) della retta di interpolazione del segnale in funzione della concentrazione aggiunta fornisce la concentrazione di campione presente nella soluzione da analizzare purché il responso del bianco non sia diverso da zero.

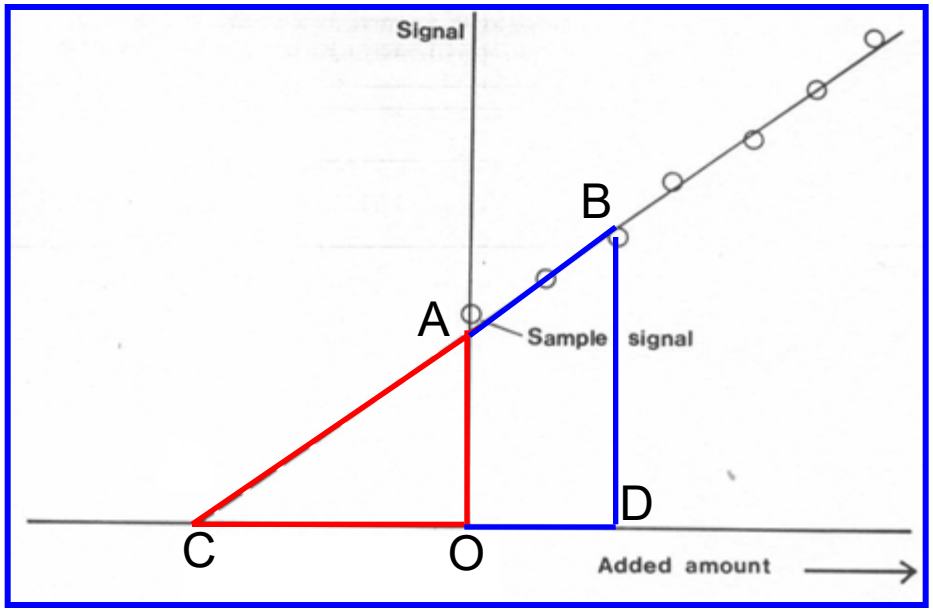
Ciò è dimostrabile considerando la similitudine di triangoli rettangoli aventi in comune un angolo acuto, nel caso specifico i triangoli *ACO* e *BCD*.

Risulta:

AO = segnale dovuto alla concentrazione incognita x_E

BD = segnale dovuto alla concentrazione incognita e a quella aggiunta, $x_E + x_{agg}$

OD = x_{agg}



Dette b_0 e b_1 l'intercetta e la pendenza della retta di regressione ottenuta sui responsi misurati, si ha:

$AO = b_0 + b_1 * x_E$
$BD = b_0 + b_1*(x_E + x_{agg})$

In virtù della similitudine fra i due triangoli rettangoli possiamo scrivere la seguente proporzione:

$AO : BD = CO : (CO+OD)$ ossia:

$[b_0 + b_1 * x_E] : [b_0 + b_1*(x_E + x_{agg})] = CO : [CO + x_{agg}]$ e quindi:

$b_0 CO + b_0 x_{agg} + b1 x_E CO + b_1 x_E x_{agg} = b_0 CO + b_1 x_E CO + b_1 x_{agg} CO$

Se risulta $b_0 \approx 0$ (ossia, se l'intervallo di fiducia di b_0 include il valore zero), come nell'ipotesi iniziale, l'equazione diventa $x_E = CO$, come volevasi dimostrare.

Anche in questo caso è possibile valutare l'errore sulla concentrazione x_E.

La formula è leggermente diversa da quella adottata per x_0:

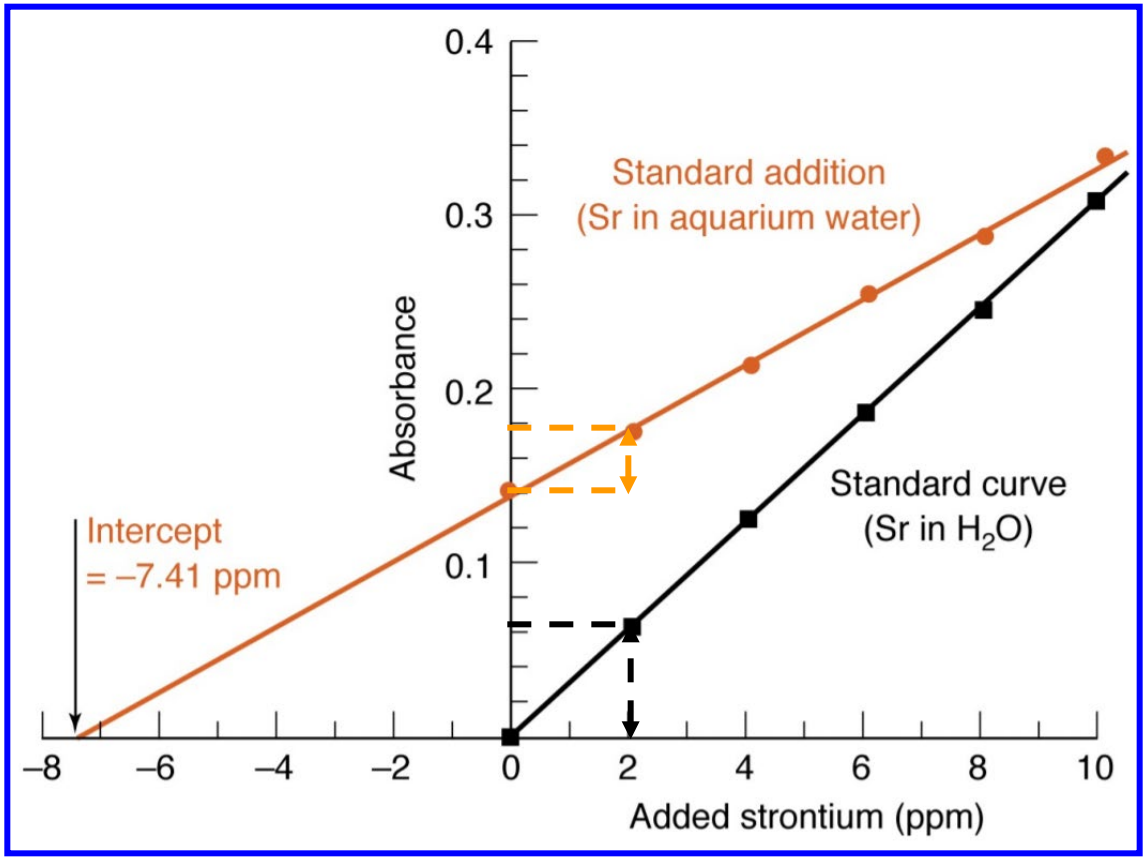
$$s_{x_E} = \frac{s_{y/x}}{b_1}\left[\frac{1}{n} + \frac{\bar{y}^2}{b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^{1/2}$$

n rappresenta il numero di dati complessivamente utilizzati per realizzare la retta di regressione.

Come per x_0 il calcolo dell'intervallo di fiducia per il valore di x_E al livello di significatività α implica l'uso della distribuzione t di Student:

$$x_E \pm t_{n-2,1-\alpha/2} \times s_{x_E}$$

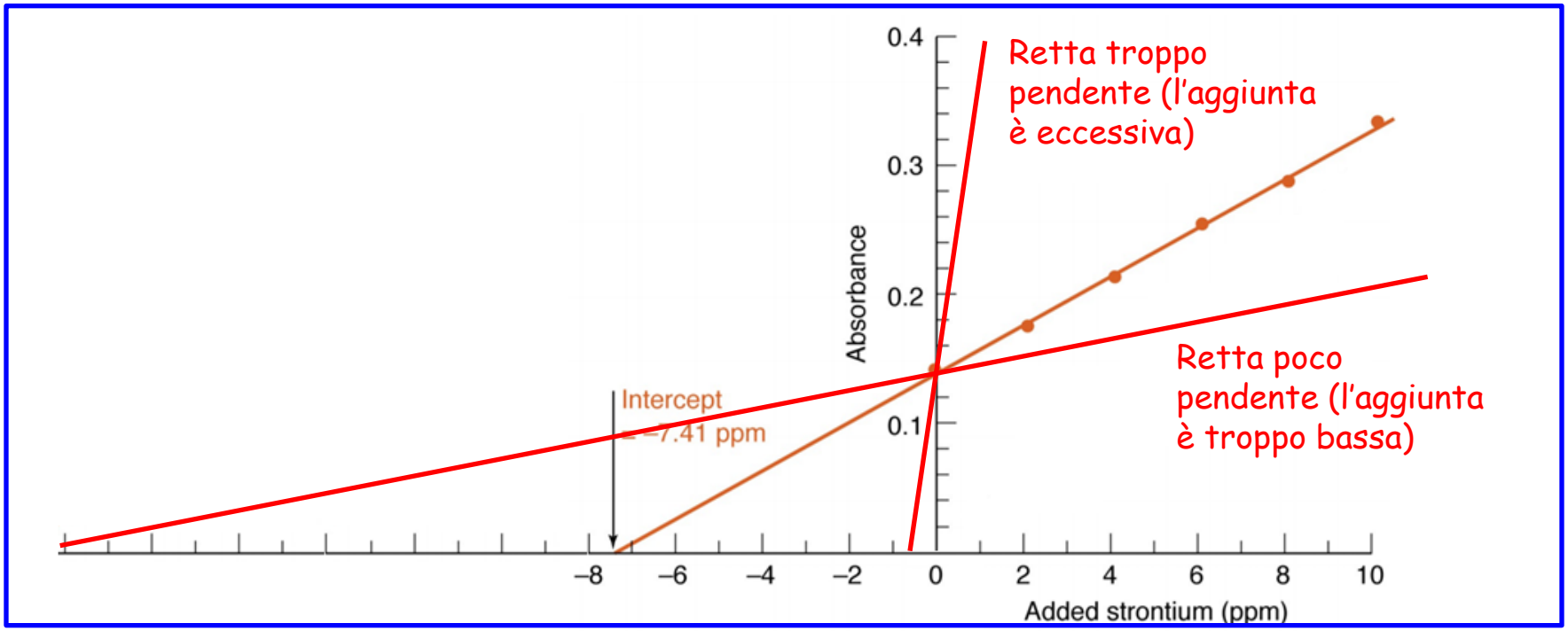
Esempio di applicazione del metodo dell'aggiunta standard all'analisi dello stronzio presente nell'acqua di un acquario mediante spettroscopia atomica di assorbimento:



Si noti che in acqua pura 2 ppm di Sr determinano un'assorbanza di circa 0.06 unità, mentre nell'acqua di acquario essa è inferiore a 0.04 unità, quindi l'effetto matrice si traduce in un decremento di responso di circa il 33%.

Criterio per la determinazione delle aggiunte di concentrazione ottimali

Nell'esempio precedente la prima aggiunta di stronzio alla matrice corrispondeva a 2 ppm, a fronte di una concentrazione dell'elemento nella matrice pari a 7.41 ppm, dunque era paragonabile a quest'ultima.
Laddove l'aggiunta sia eccessiva/bassa rispetto alla concentrazione iniziale in matrice, la retta di aggiunta standard che ne deriva è molto/poco pendente, come mostrato in figura:



L'intercetta sul semiasse negativo diventa quindi o molto vicina allo zero o molto grande, rispettivamente. In entrambi i casi l'errore sulla concentrazione ricavata è rilevante. E' dunque opportuno effettuare una sola aggiunta inizialmente e, se essa non è ottimale, rifarla con un valore diverso.

Uso della regressione lineare per il confronto di metodi analitici

La regressione lineare con il metodo dei minimi quadrati può essere impiegata in modo molto efficace quando due metodi analitici vanno confrontati in un ampio intervallo di concentrazioni.

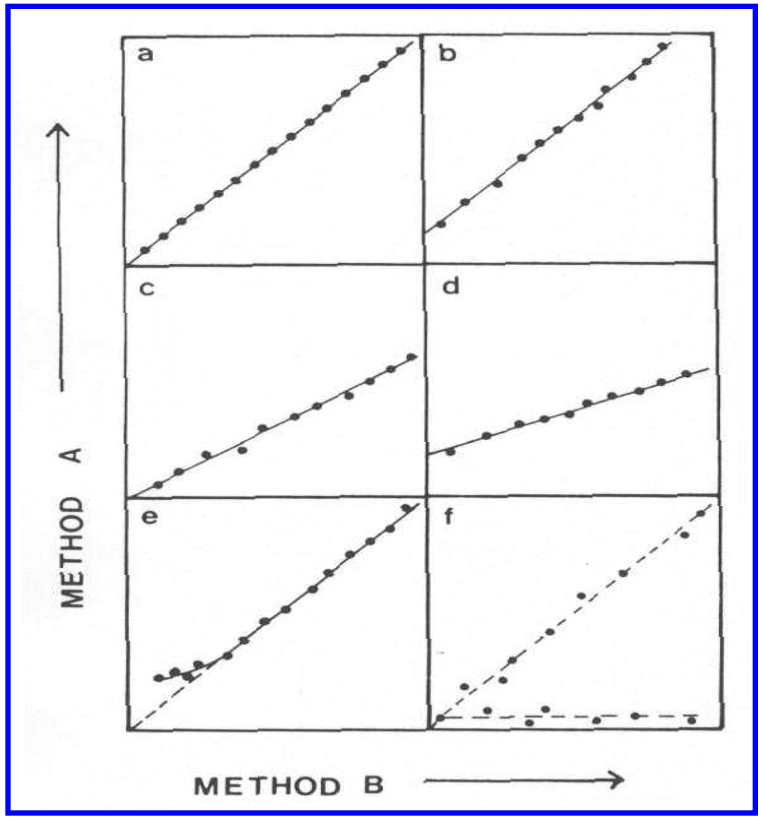
In tal caso si effettuano n coppie di misure con i due metodi e si riportano in grafico i valori ottenuti da uno dei due metodi (A) contro quelli derivanti dall'altro (B). Sono possibili 6 casi:

a) $b_0 = 0$, $b_1 = 1$, $r \approx 1$:

è il caso ideale, in cui i due metodi forniscono risultati identici;

b) $b_0 \neq 0$, $b_1 = 1$, $r \approx 1$:

il metodo A fornisce un risultato costantemente spostato verso valori maggiori rispetto al metodo B



c) $b_0 = 0$, $b_1 \neq 1$, $r \approx 1$:

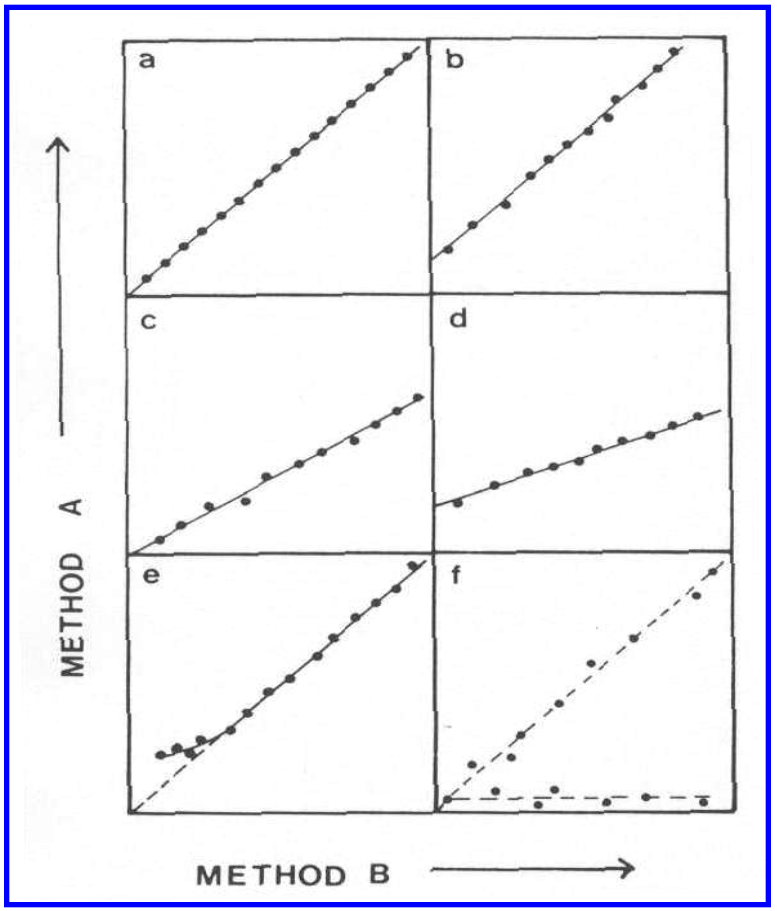
il metodo B fornisce responsi proporzionalmente più elevati

d) $b_0 \neq 0$, $b_1 \neq 1$, $r \approx 1$:

è una combinazione dei casi b e c

e) esiste una deviazione dalla linearità per uno dei due metodi in un certo intervallo di concentrazione

f) i campioni contengono quantità variabili di due specie dell'analita, una delle quali non è rivelata affatto dal metodo A, mentre il metodo B è sensibile ad entrambe.

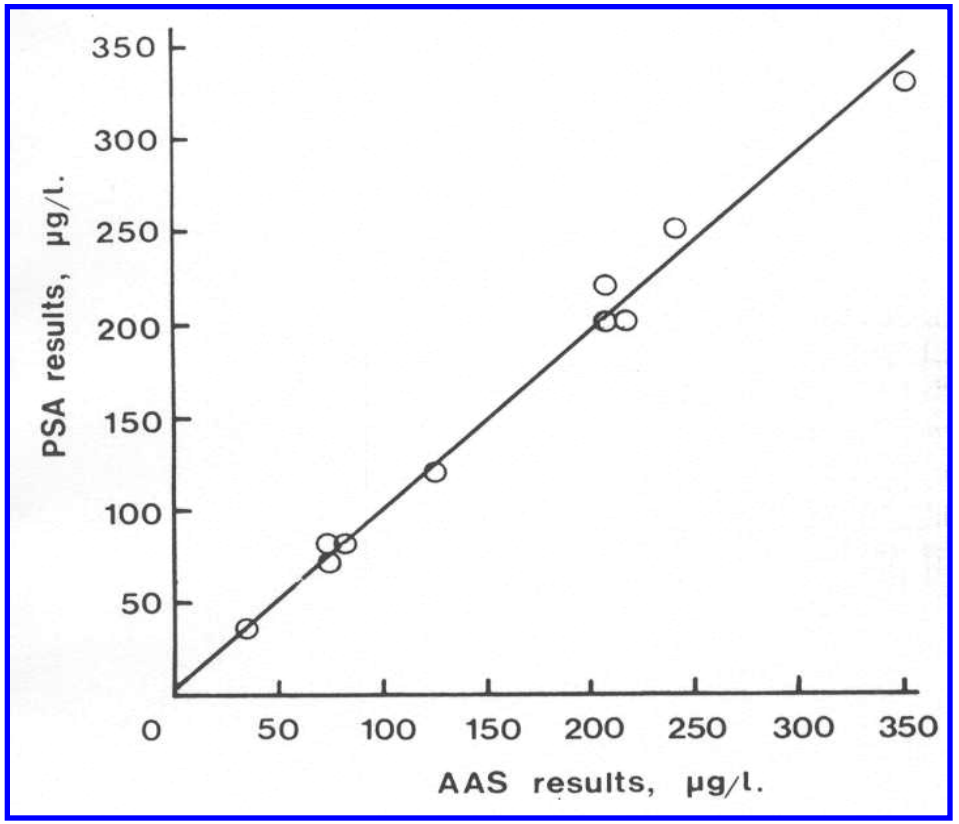


In generale il confronto fra i due metodi si effettua calcolando gli intervalli di fiducia per la pendenza e l'intercetta della retta di regressione e verificando che i valori 1 e 0, rispettivamente, siano compresi in tali intervalli.

Un esempio numerico

Si considerino i dati relativi alla concentrazione (µg/L) di piombo in 10 succhi di frutta ottenuti con un'analisi di stripping potenziometrica (PSA) e con la spettroscopia atomica di assorbimento (AAS):

Campione	AAS	PSA
1	35	35
2	75	70
3	75	80
4	80	80
5	125	120
6	205	200
7	205	220
8	215	200
9	240	250
10	350	330



La retta di regressione ottenuta ha i parametri:

$b_0 = 3.87$ $b_1 = 0.963$ $r = 0.9945$

$s_{y/x} = 10.56$ $s_{b_0} = 6.64$ $s_{b_1} = 0.0357$

Considerando un livello di fiducia del 95% risulta $t_{8,0.975} = 2.31$ e quindi:

$b_0 = 4 \pm 15$ $b_1 = 0.96 \pm 0.08$

I risultati dei due metodi PSA e AAS si possono dunque ritenere non significativamente diversi al 95 % di fiducia.

Accorgimenti nell'applicazione del metodo dei minimi quadrati al confronto di due metodi analitici

Una delle assunzioni fondamentali per l'applicazione del metodo dei minimi quadrati è che l'errore che insiste sulla grandezza usata come x sia trascurabile.

Nel caso del confronto di metodi questa assunzione può non essere vera, perché anche i valori delle x derivano da misure sperimentali, tuttavia test pratici mostrano che l'approccio può essere applicato con successo purché:

- ✓ si riportino sull'asse x i dati derivanti dal metodo più preciso dei due (aspetto da valutare preliminarmente con un F-test sulle varianze)
- ✓ si usi un numero ragionevole di dati (almeno una decina) per il confronto, considerando che il calcolo degli intervalli di fiducia della pendenza e dell'intercetta si basa su n-2 gradi di libertà.
- ✓ i punti coprano nel modo più uniforme possibile l'intervallo di concentrazione d'interesse per il confronto.

Un'altra assunzione fondamentale del metodo dei minimi quadrati è che l'errore sulla grandezza y NON cambi con la concentrazione (omoschedasticità).

Questa condizione, valutabile con una serie di test sulle varianze osservate replicando la misura più volte a ciascuna delle concentrazioni coinvolte nella calibrazione, può non essere vera se si esplorano molti ordini di grandezza di concentrazione.

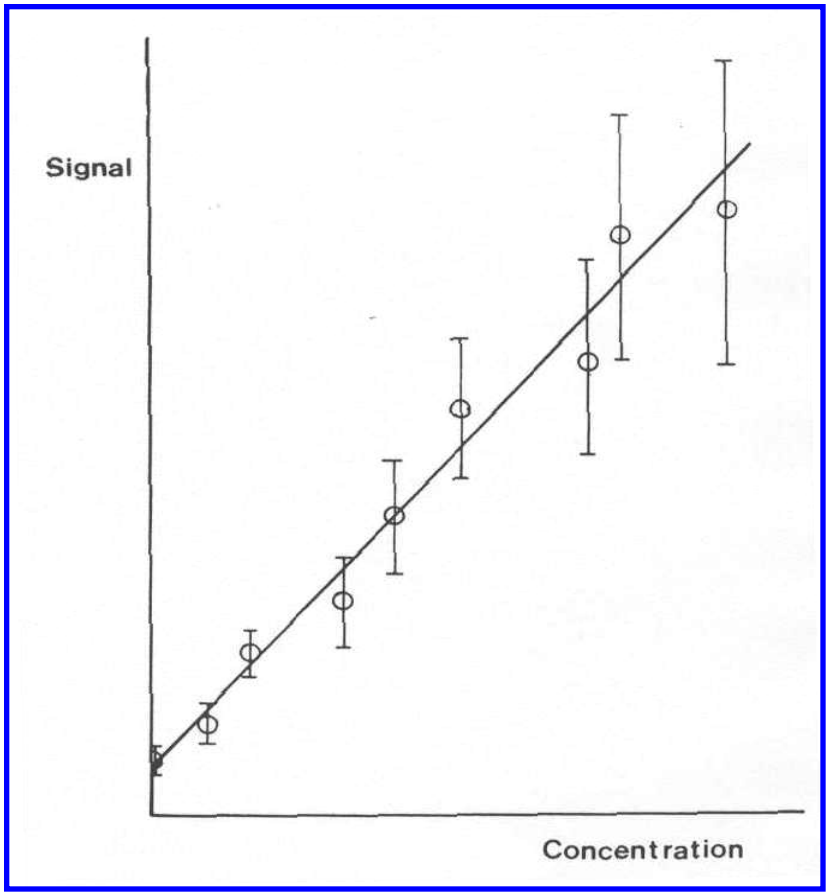
In tal caso occorre far uso della regressione lineare pesata.

Regressione lineare pesata

La regressione lineare pesata va introdotta quando è evidente che l'errore che incide sui valori del segnale NON è indipendente dalla concentrazione (ossia cresce, o diminuisce, in modo significativo con essa).

Si supponga di:

- ✓ aver effettuato misure con un metodo analitico su nove campioni a diversa concentrazione
- ✓ di aver replicato un congruo numero di replicati per ciascuno di essi, ottenendo così un intervallo di fiducia del segnale per ogni concentrazione
- ✓ che l'intervallo di fiducia del responso cresca significativamente con la concentrazione:



Poiché è più importante che la retta di regressione passi più vicino ai punti con un'incertezza inferiore è opportuno dare ad essi un peso statistico maggiore rispetto agli altri.

Ciò è possibile usando come peso un valore inversamente proporzionale alla varianza s_i caratteristica per i vari punti. Di solito si usa come peso il valore (weight):

$$w_i = s_i^{-2} / \left(\sum_{i=1}^{n} s_i^{-2} / n \right)$$

per comodità i pesi w_i sono scalati in modo che la loro somma sia n, il numero dei punti della retta.

Noti i vari w_i si calcolano le medie pesate dei valori di x e y:

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i / n \qquad \bar{y}_w = \sum_{i=1}^{n} w_i y_i / n$$

In analogia con quanto accade con la regressione lineare non pesata si possono poi calcolare i parametri della regressione, b_0 e b_1:

$$b_1 = \frac{\sum_{i=1}^{n} w_i x_i y_i - n\bar{x}_w \bar{y}_w}{\sum_{i=1}^{n} w_i x_i^2 - n\bar{x}_w^2}$$

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

In questo caso le formule per il calcolo dell'incertezza su un valore di concentrazione x_0 determinato a partire dal corrispondente segnale y_0 sono più complesse. La deviazione standard sui residui è data dall'equazione:

$$s_{(y/x)w} = \sqrt{\frac{\sum_{i=1}^{n} w_i y_i^2 - n\bar{y}_w^2 - b_1^2\left(\sum_{i=1}^{n} w_i x_i^2 - n\bar{x}_w^2\right)}{n-2}}$$

La deviazione standard sul valore di concentrazione estrapolato è data da:

$$(s_{x_0})_w = \frac{s_{(y/x)w}}{b_1}\left[\frac{1}{w_0} + \frac{1}{n} + \frac{(y_0 - \bar{y}_w)^2}{b_1^2\left(\sum_{i=1}^{n} w_i x_i^2 - n\bar{x}_w^2\right)}\right]^{1/2}$$

dove w_0 è il peso associato alla determinazione y_0.

La valutazione di casi reali mostra che molto spesso l'uso della regressione lineare pesata al posto di quella convenzionale non modifica tanto il valore di x_0 ma rende molto più realistico l'intervallo di fiducia associato ad esso.

Un esempio numerico

Si supponga di aver misurato l'assorbanza di sei soluzioni standard, effettuando per ciascuna più replicati ed ottenendo i valori:

Concentr., µg/L	0	2	4	6	8	10
Assorbanza	0.009	0.158	0.301	0.472	0.577	0.739
Deviazione Stand.	0.001	0.004	0.010	0.013	0.017	0.022

Se procediamo nel calcolo dei parametri della regressione con i due metodi otteniamo i valori:

regressione convenzionale $b_0 = 0.0133$ $b_1 = 0.0725$

regressione pesata $b_{0_w} = 0.0091$ $b_{1_w} = 0.0738$

La differenza fra i valori appare trascurabile, come confermerebbe la valutazione dei relativi intervalli di fiducia.

La differenza fra i due metodi appare però evidente se si valuta l'intervallo di fiducia sui valori di concentrazione corrispondenti ad assorbanze 0.1 e 0.6:

Assorbanza	Intervallo di fiducia	
	regr. conv.	regr. pesata
0.1	1.2 ± 0.6	1.23 ± 0.12
0.6	8.1 ± 0.6	8.0 ± 0.7

La stima dell'intervallo di fiducia mediante la regressione lineare pesata è molto più realistica alle basse concentrazioni, ossia in prossimità del "baricentro" della retta.

Ciò si riflette nella diversa forma delle bande di predizione rispetto al caso della regressione lineare convenzionale.