TEMA 5: CALIBRATGE

MODELS DE CORRELACIÓ ENTRE VARIABLES

1. CONCEPTE DE CORRELACIÓ:

→ **Correlació**: variació sistemàtica d'una variable en funció d'una altra variable (**univariant**) o variables (**multivariant**) que pot ser descrita per un model de dependència.

La correlació entre parelles de variables es pot descriure amb diferents **models de dependència**, que es triaran segons el **coneixement** del problema (funcional, basat en relacions físiques, *vs.* **Correlacions empíriques**, que no es poden extrapolar més enllà les dades) i el tipus de model **matemàtic** (ex. Equacions linears *vs.* No linears) que millor descrigui la variació entre variables.

El fet que una variable estigui correlacionada amb una altra no vol dir que la variació d'una d'elles provoqui una variació en l'altra (**correlació ≠ causalitat**).

2. TIPUS DE MODELS DE DEPENDÈNCIA: Selecció de models univariants

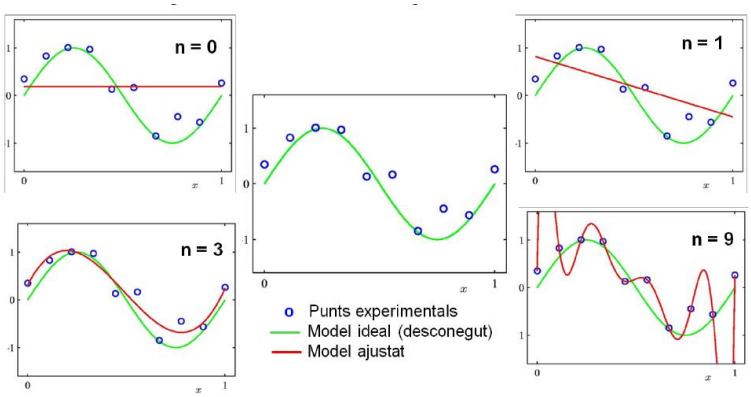
Segons el coneixement previ

- **Funcionals** (fisicoquímics). Aprofiten coneixement **científic** bàsic. Són equacions amb paràmetres que tenen sentit físic. Es poden extrapolar.
  - Ex: cinètica de primer ordre. $A \xrightarrow{k} B \qquad [A]=[A]_0 e^{-kt}$
- **Empírics**. Hi ha una funció matemàtica que s'adapta a les dades estudiades. No s'extrapolen. Només funcionen a l'interval de dades que fem servir. Per un interval de dades sí que hi ha correlació però a un domini experimental no hi ha correlació.

Segons la forma matemàtica

$y = a + b_1 x$	**Models lineals de primer ordre i de segon ordre fins n ordres**
$y = a + b_1 x + b_2 x_2 + b_3 x_3$	Són polinomis
$y = b_1 \exp(b_2 x)$	**Model no lineal**

Ajust amb models polinòmics: Cal triar el model més simple que descriu el comportament i variabilitat natural de les nostres dades (**principi de parsimònia)**



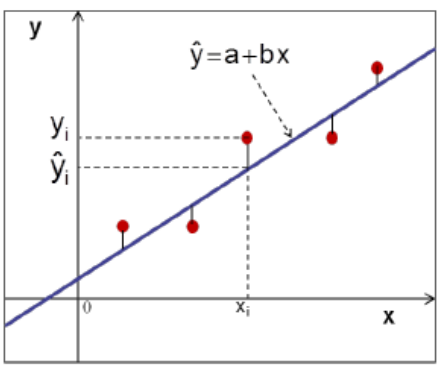
Hem de trobar un model com n=3 que l'error no és molt gran i la correlació de variables existeix.

El valor predit pot ser una mica diferent al valor real, cal que el model tingui sentit químic i que sigui lo més simple possible.

SLOPE: Pendent
INTERCEPT: Ordenada de l'origen

MODEL LINEAL UNIVARIANT:

- Moltes correlacions entre dues variables químiques poden ser descrites mitjançant un model lineal d'ordre 1 univariant (al menys, localment). Barret = valor predit



Ex: calibratge analític permet relacionar la resposta d'un sistema de mesura a la concentració (o quantitat) d'un anàlit. Per a una concentració donada, l'interval de l'anàlit, el model primer ordre lineal univariant és adequat per derivar la funció de calibratge (models multivariants es poden fer servir si considerem altres anàlits i/o interferències).

- Cal tenir una sèrie de punts experimentals (x_i, y_i) per a ajustar el model. Mirar si hi ha una correlació lineal entre y i x (en l'interval de concentració estudiat):
- Es vol trobar els paràmetres del model (a,b) que proporcionin els residuals menors.

$y = \alpha + \beta x$ → $\hat{y} = a + bx$

Model escollit → Model ajustat

$e_i = y_i - \hat{y}_i$ **Residual:** diferència entre el valor experimental i l'ajustat pel model

OBTENCIÓ DEL MODEL. MÈTODES DE MÍNIMS QUADRATS

- Troben els paràmetres del model (a,b) minimitzant la suma de quadrats dels residuals entre els punts experimentals i els ajustats pel model.

$$ss_R = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - b x_i)^2$$

Tipus de mètodes de mínims quadrats:

→ **MÍNIMS QUADRATS ORDINARIS (OLS): COMÚ**

- ✓ L'error només és als valors de **y**.
- ✓ Els residuals són similars en tots els punts (**homocedàstic**). Residual: és independent valor x!!
- ✓ Tots els punts conten igual en l'ajust.

→ **Mínims quadrats sospesats (WLS): ponderats!!**

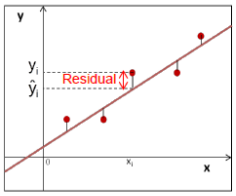
- ✓ L'error només és als valors de **y**.
- ✓ Els residuals varien segons els punts (**heteroscedàstic**).
- ✓ Es dóna més pes a l'ajust als punts que tenen menys error.

RESIDUAL = y_i – ^y_i

→ **MÍNIMS QUADRATS TOTALS (TLS):**

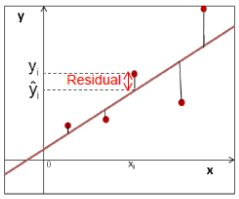
- ✓ L'error experimental es troba als valors de **x** i els valors de y.

OLS



$$SS_R = \sum_i (y_i - \hat{y}_i)^2$$

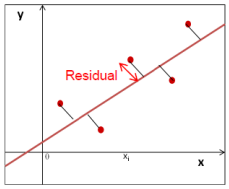
WLS



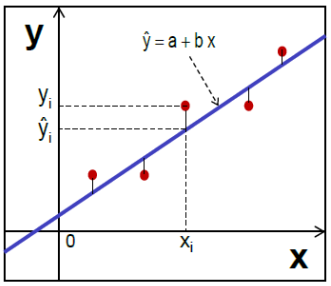
$$SS_R = \sum_i \frac{(y_i - \hat{y}_i)^2}{s_i^2}$$

(s_i ,desviació estàndard del punt i)

TLS



REGRESSIÓ LINEAL PER MÍNIMS QUADRATS (OLS)



$$ss_R = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - b x_i)^2$$

$$\frac{\partial ss_R}{\partial a} = \frac{\partial ss_R}{\partial b} = 0$$

Els paràmetres **a** i **b** fan mínima la suma de quadrats dels residuals

ERROR:

$$s^2_{y/x} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$$ **Variància dels residuals**

$$a = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2}$$

$$b = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2}$$

n nombre de parelles de punts

a → $s^2_a = s^2_{y/x}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)$ Variància associada a l'ordenada

b → $s^2_b = \frac{s^2_{y/x}}{\sum_i (x_i - \bar{x})^2}$ Variància associada al pendent

n nombre de parelles de punts

No *és una línia per culpa dels residuals.*

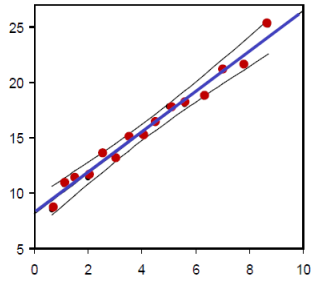
Intervals de confiança dels paràmetres

a (ordenada) $a \pm t(\alpha, n-2)\, s_a$ $\quad y = \alpha + \beta x = a \pm t s_a + (b \pm t s_b)x$

b (pendent) $b \pm t(\alpha, n-2)\, s_b$

Banda de regressió

Indica la incertesa associada a la recta del model de regressió



UNIVERSITAT DE BARCELONA

Banda de regressió

Per a un valor de x_i donat:

$$y = a + bx \pm t\, s_{y/x}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

Es més estreta:

- ✓ quan n augmenta
- ✓ prop del centre del model $(\bar{x}, \bar{y})$

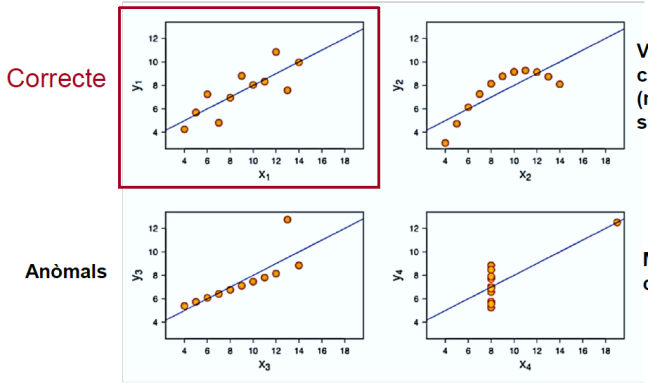
DOF: N-2 (dos coeficients, amb 2 punts tinc recta)

$$s^2_{y/x} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$$ (variance of the residuals)

$$s^2_a = s^2_{y/x}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)$$ (variance associated with the intercept)

$$s^2_b = \frac{s^2_{y/x}}{\sum_i (x_i - \bar{x})^2}$$ (variance associated with the slope)

INSPECCIÓ VISUAL DE LA GRÀFICA X/Y:

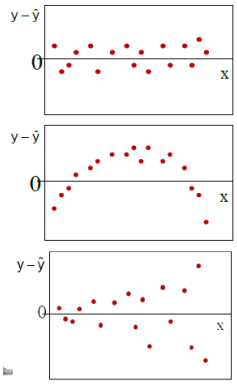


Hem d'interpolar a la zona central perquè és on menys error té.

El correcte: la correlació entre variables pot ser bona pq residuals són homocedàstics no hem de buscar r^2 molt lineals.

Pels anòmals no puc aplicar els tests que sabem pq no són rèpliques són una població de dades.

Estudi dels residuals (ei = yi – ŷi) en funció de x. INSPECCIÓ RESIDUALS:



Model lineal correcte.
- ✓ Residuals amb patró aleatori
- ✓ Residuals similars (homocedàstics)

Model lineal incorrecte: model massa senzill
- ✓ Residuals amb patró sistemàtic → test estadístic (chi-quadrada) per saber la distribució
- ✓ **Cal un model més complex.**

Model lineal incorrecte.
- ✓ Residuals diferents (heterocedàstics).
- ✓ **Us de WLS:** WLS per a punts amb menor residual

SIGNIFICACIÓ DE LA CORRELACIÓ I DELS PARÀMETRES DEL MODEL:

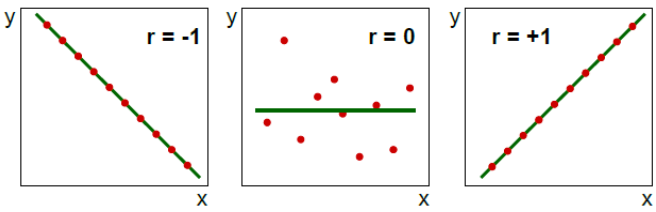
Covariància Relació entre la variació de les parelles de punts (x,y). Cov >> 0 o << 0 . Bona relació lineal x/y.

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

$$r = \frac{\text{cov}(x,y)}{s_x s_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Coeficient de correlació, r Relació entre la variació de les parelles de punts (x,y). Covariància estandarditzada. Varia entre -1 i 1.

Coeficient de determinació, r^2: variància de y descrita pel model lineal. Varia entre 0 i 1.
cov: Com canvis de la y els puc explicar com canvis de la x. R2: explica el % de variació de la y es deu a la x R: cov normalitzada a s (x) i s(y). Variabilitat regressió. 1-r2: altra variabilitat.



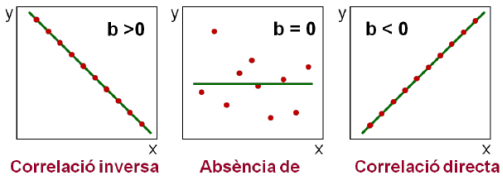
C. inversa Absència de c. C. directa

Test t sobre la significació del coeficient de correlació r: FER!!!

$$t_{calc} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

H_0: r = 0 (dades no correlacionades) $t_{calc} < t_{taules}$
H_1: r ≠ 0 (dades correlacionades) $t_{calc} > t_{taules}$
t_{taules}(α, 2 cues, n-2 gdl)

Quan la hipòtesi nul.la no es compleix, la correlació entre x i y és significativa



y = bx + a

Si r=0,5 no sé si hi ha correlació.
També es pot fer un test t sobre el pendent: <0 ó >0 difereix de 0: sí
Si b → 0 no hi ha correlació.
Puc fer el mateix test t sobre el pendent: b/Sb→ valor estadístic calculat per a r.
Test t de l'ordenada de l'origen per tal de simplificar el model a y=bx
Tcalc =a/Sa no el podem proposar si l'ordenada de l'origen és 0. Malament.

Test t sobre el pendent

$t_{calc} = b/S_b$
t_{tab} (α, 2 cues, n-2 gdl)

H_0: b = 0 (dades no correlacionades) $t_{calc} < t_{tab}$
H_1: b ≠ 0 (dades correlacionades) $t_{calc} > t_{tab}$

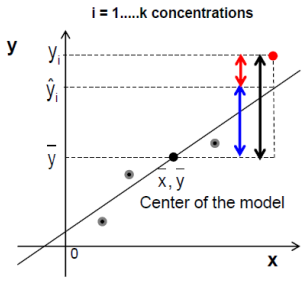
Test t sobre l'ordenada a l'origen

$t_{calc} = a/S_a$
t_{tab} (α, 2 cues, n-2 gdl)

H_0: a = 0 $t_{calc} < t_{tab}$
H_1: a ≠ 0 $t_{calc} > t_{tab}$

H_0 certa ⇒ model **y = bx**
H_0 falsa ⇒ model **y = bx + a**

ANOVA DE LA REGRESSIÓ



$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$SS_{overall}$ = $SS_{regression}$ + $SS_{residual}$

d.o.f. = number of coefficients - 1

d.o.f. = k – number of coefficients

H_o $s^2_{regression} = s^2_{residual}$

H_1 $s^2_{regression} > s^2_{residual}$ (good description of data variation by the model)

Estudia si la variància associada a la descripció de la y mitjançant el model lineal és superior a la variància residual.

També serveix per dir si hi ha correlació o no. Com de diferent és el valor experimental al valors associats al model.

Tinc n valors de y_i (experimentals) + valor central y (mitjana). Hi haurà correlació quan vario el valor de x. Tindré una línia recta si no hi ha correlació. Tinc 2 fonts de variabilitat: y són diferents per tema aleatori o perquè hi ha un model de regressió que les fa diferents a mida que augmenta la x.

Miro la distància entre el punt yi i la y mitjana. Si ssreg^2> s^2ale hi ha correlació.

Font de variació	Suma de quadrats	graus de llibertat	S^2	Fcalc
Regressió	$SS_{REG} = \sum_i (\hat{y}_i - \bar{y})^2$	k-1	S^2_{REG}	$\frac{S^2_{REG}}{S^2_R}$
Residual	$SS_R = \sum_i (y_i - \hat{y}_i)^2$	n-k	S^2_R	
Total	$SS_T = \sum_i (y_i - \bar{y})^2 = SS_{REG} + SS_R$	n-1	S^2_{TOT}	

n: nombre de parelles de punts k: nombre de paràmetres ajustats (a,b)

UNIVERSITAT DE BARCELONA

$r^2 = \frac{SS^2_{REGR}}{SS^2_{TOTAL}}$ % variància explicada pel model

CALIBRATGE UNIVARIANT ANALÍTIC. EXEMPLE DE MODEL LINEAL:

Calibratge univariant analític: model lineal que relaciona una resposta mesurada amb la concentració d'un anàlit amb la finalitat predictiva.

- ✓ Vàries estratègies es poden seguir per comprovar **com de bo** és el model més enllà de la validació estadística.
- ✓ Ex: la bondat del model es pot provar quan s'aplica a **mostres no utilitzades** en la construcció del model lineal (**mostres de validació**). Tot i que un model pot descriure correctament les dades utilitzades per a la seva construcció, pot donar resultats erronis (**exbiaixats**) quan s'aplica a mostres externes (per exemple, no s'ha considerat una variable rellevant, com també l'efecte de la composició de la matriu). El **biaix relacionat** i **l'error predictiu** es pot calcular.

Calibratge univariant. Es mesura una resposta i es determina la concentració de l'anàlit.

- No hi ha d'haver interferents a les mostres, o
- La resposta mesurada ha de ser selectiva per a l'anàlit.

Calibratge multivariant. Permet determinar un o més anàlits simultàniament a partir d'una resposta multivariant.

- Pot haver interferents a les mostres.
- No cal selectivitat en la mesura.

Construcció d'un model de calibratge

1. **Selecció de mostres de calibratge (anàlit i matriu) representatives de les mostres desconegudes de predicció.**
   a. Patrons: mostres de calibratge de concentració ben coneguda (materials de referència, patrons de matriu sintètica, patrons d'anàlit pur)
   b. Han de reflectir l'interval natural de variabilitat de concentracions.
2. **Realització acurada de la mesura de resposta instrumental (replicats).**
3. **Rebuig de mostres anòmales**
   a. Inspecció visual
   b. Mètodes de regressió robusta (LMS, least median of squares regression).
4. **Construcció del model lineal de calibratge.**

Validació del model de calibratge: *la x és interpolada*

- **Validació matemàtica del model lineal.** Comprovació que l'ús d'un model lineal de primer ordre és adequat per a descriure la dependència entre resposta i concentració.
- **Validació analítica del model.** Comprovació del bon funcionament del model de calibratge aplicant-lo sobre mostres noves de validació de concentració perfectament coneguda.
  - ✓ **Càlcul de l'error en la predicció de les concentracions (RMSEP).**
  - ✓ **Càlcul del biaix en les prediccions.**

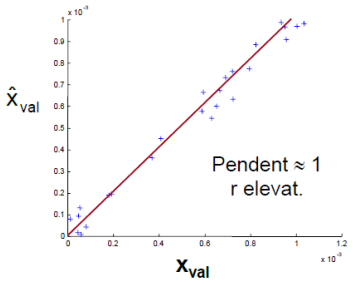
Model de calibratge

$$y_{cal} = bx_{cal} + a$$

Predicció de x (concentració) en mostres de validació

$$\hat{x}_{val} = \frac{y_{val} - a}{b}$$

Comparació dels valors predits, $\hat{x}_{val}$, amb els valors de referència, x_{val}



UNIVERSITAT DE BARCELONA

$$RMSEP = \sqrt{\frac{\sum_i (\hat{x}_{i,val} - x_{i,val})^2}{n}}$$

Ha de ser petit i reflectir l'error esperat.

$$biaix = \frac{\sum_i (\hat{x}_{i,val} - x_{i,val})}{n}$$

Ha de ser proper a zero. Biaix > 0 o < 0, error sistemàtic en el model

ESTRATÈGIES ANALÍTIQUES DE CALIBRATGE: solucions per a respostes sense sentit químic!

Calibratge extern: *x interpolada*

- ✓ No hi ha efecte (significatiu) de la matriu en la resposta, o bé la matriu pot ser reproduïda
- ✓ No hi ha errors sistemàtics que afectin els paràmetres de la funció de calibratge.

Patró intern

- ✓ Hi ha fluctuacions en la resposta que afecten per igual a l'anàlit i al patró intern. La resposta del model lineal (y) és el quocients de les respostes de l'anàlit i el patró intern i x és la relació de concentracions entre ánàlit i patró intern.

Addició d'estàndard *x extrapolada, punt on la resposta és 0.*

- ✓ Hi ha efecte significatiu de la matriu. S'afegeix una quantitat coneguda d'anàlit a la matriu de la mostra d'anàlisi Si hi ha efecte de matriu: hi ha una variable descontrolada que afecta en el model.
- ✓ El model es fa entre la resposta mesurada (y) i la quantitat d'anàlit afegida (x).
- ✓ La resposta analítica ha de ser nul·la en absència d'anàlit.
- ✓ O bé fem calibratge multivariant o simulem la matriu en els patrons.

CALIBRATGE EXTERN: Predicció d'un valor de x a partir de y

Interval de confiança

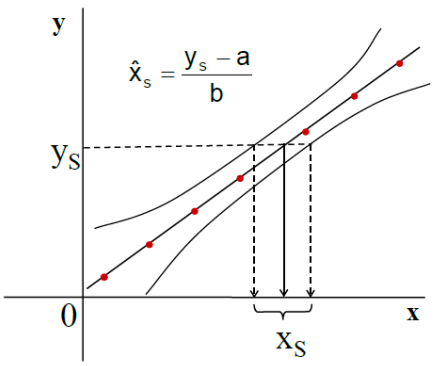
$$x_S = \hat{x}_S \pm t_{(\alpha, n-2)} \cdot s_{x_s}$$

n punts (n patrons), sense replicats

$$s_{x_s} = \frac{s_{y/x}}{b} \sqrt{1 + \frac{1}{n} + \frac{(y_S - \bar{y})^2}{b^2 \sum_i (x_i - \bar{x})^2}}$$

n punts, m replicats

$$s_{x_s} = \frac{s_{y/x}}{b} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_S - \bar{y})^2}{b^2 \sum_i (x_i - \bar{x})^2}}$$



ADDICIÓ ESTÀNDARD: variable composició de matriu afecta.

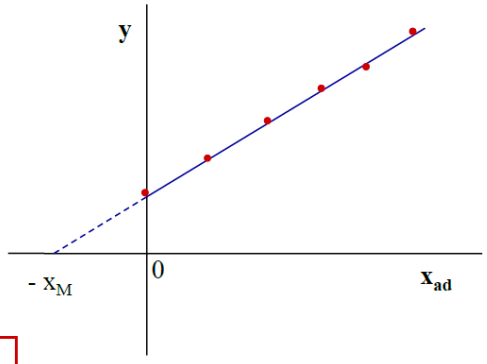
$$y = k\,(x_M + x_{ad}) = \underbrace{k\,x_M}_{a} + \underbrace{k}_{b}\,x_{ad}$$

$$y = 0 \rightarrow -x_{ad} = \frac{a}{b} = x_M$$

Interval de confiança

$$x_M = \hat{x}_M \pm t_{(\alpha, n-2)} \cdot s_{x_M}$$

$$s_{x_M} = \frac{s_{y/x}}{b} \sqrt{\frac{1}{n} + \frac{\bar{y}^2}{b^2 \sum_i (x_i - \bar{x})^2}}$$



PARÀMETRES DE QUALITAT

Límit de detecció (LOD): és la mínima concentració d'anàlit que proporciona un senyal (y_{LOD}) . Un instrument en absència de senyal pot donar resposta. PROBLEMA: a concentracions molt petites!. Significativament diferent de la mitjana del blanc (y_B), considerant una variació aleatòria de la resposta de la solució del blanc, (IUPAC: 20 blancs). Aquesta resposta és 0 o no. És cnt o no?... Comparem 2 distribucions, blanc sense anàlit i distribució al voltant de la resposta del blanc.

$$y_{LOD} = y_B + 3\,s_B$$

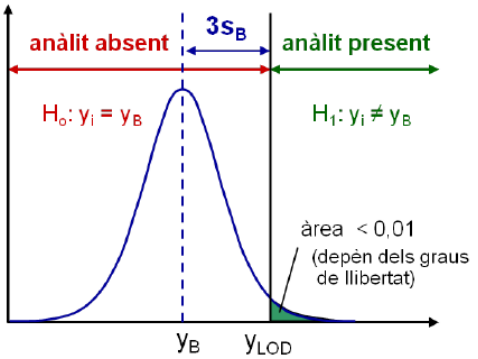
$$x_{LOD} = \frac{y_{LOD} - a}{b} = \frac{y_B + 3\,s_B - a}{b}$$

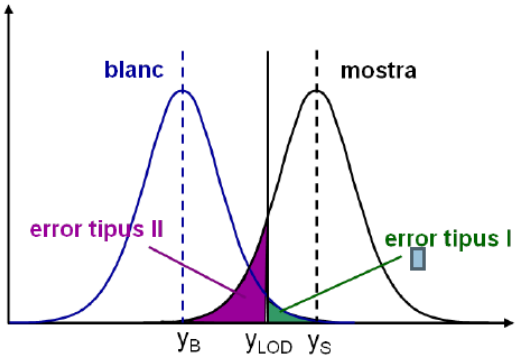
Aproximació $y_B \approx a$ $s_B \approx s_a$

$$y_{LOD} \approx a + 3\,s_a$$

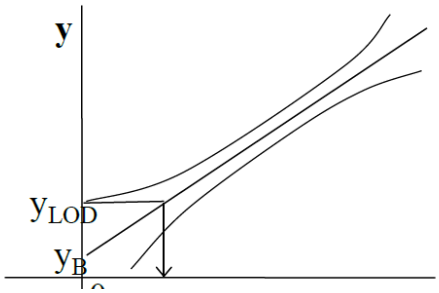
$$x_{LOD} = \frac{y_{LOD} - a}{b} \approx \frac{3 s_a}{b}$$

Interpretació estadística $y_{LOD} = y_B + t\,s_B$ ≈ 3 (bona aproximació)





També es pot estimar a partir de la banda de regressió:



Marco la frontera. Resposta mínima a partir de la qual jo puc detectar l'anàlit. Les condicions estadístiques per rebutjar o acceptar H0 les marca el LOD: la zona de rebuig és molt petita. T*3!!

Ylod = Ys + t Sb

La fluctuació del planc afecta que LOD sigui més alt del que caldria. No preocupa només la resposta del blanc sinó si fluctua molt. A nivell pràctic: $Y_{LOD} = Y_S + t \cdot S_B$ → $X_{LOD} = 3*S_A/b$

$$x_{LOD} = x_B + ts_{x_B} = 0 + 3\frac{s_{x/y}}{b}\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_B - \bar{y})^2}{b^2 \sum_i (x_i - \bar{x})^2}}$$

Sa la tindrem sempre però Sblanc no!

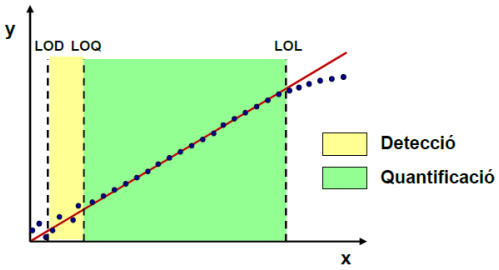
No universal a vegades no funciona. Les farem quan no tinguem informació del blanc. Hipòtesi que podem fer:

- significativament idèntica a l'ordenada de l'origen. Comprovar-ho test Fischer.
- Sinó, podem negligir-les
- La fluctuació del blanc ≈ fluctuació de l'ordenada de l'origen.

<u>Límit de quantificació (LOQ):</u> aproximació: $\mathbf{X_{LOQ} = 10*S_A/b}$

IUPAC approach

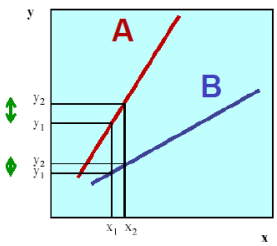
$$y_{LOQ} = y_B + 10\, s_B \qquad x_{LOQ} = \frac{y_{LOQ} - a}{b}$$



Confidence band approach

$$x_{LOQ} = x_B + ts_{x_B} = 0 + 10\frac{s_{x/y}}{b}\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_B - \bar{y})^2}{b^2 \sum_i (x_i - \bar{x})^2}}$$

<u>Interval de linealitat:</u> LOQ-LOL interval de concentracions en què la relació concentració/resposta està ben descrita <u>per un model lineal d'ordre 1.</u> Acaba en LOL. El LOD és extrapolació, no puc quantificar.



<u>Sensibilitat:</u> és la capacitat d'un mètode de distingir x (ex. Concentracions) similars: s = dy/dx =b pendent. Capacitat de discriminar 2 concentracions molt pròximes. 2 sistemes amb errors del pendent similars. Pendent més gran, més sensibilitat però si Sb és molt gran tindré LOQ horrible. SbA ≈ SbB

Smètode A > S mètode B

<u>ALTRES APLICACIONS DE MODELS LINEALS:</u> MULTIVARIANTS: SEMINARIS. Inventar-me variables addicionals. Una vegada les variables (Factors; x) que afecten una variable dependent (y) són identificades, pot servir per establir el (multivariant) model de la **relació y-x** i per predir la variació de la variable y quan canviem els valors de x en un rang donat.

$y = a + b_1x_1 + b_2x_2 + \ldots$
(multivariate linear model without interactions)
$y = a + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + \ldots$
(multivariate linear model with interactions)
$y = a + b_{11}x_1^2 + b_1x_1 + b_{22}x_2^2 + b_2x_2 + b_{12}x_1x_2 + \ldots$
(multivariate linear model with interactions and second-order terms)
(../..)