14.Model Lineal

Estadística

Grau en Matemàtiques

Josep A. Sanchez

Dept. Estadística i I.O.(UPC)

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Algunes situacions reals

Estem interessats en determinar la **producció d'insulina** en el teixit pancreàtic a diferents nivells de **concentració de glucosa**

Concentració								
1	1.53	1.61	3.75	2.89	3.26	2.83	2.86	2.59
2	3.15	3.96	3.59	1.89	1.45	3.49	1.56	2.44
3	3.89	4.80	3.69	5.70	5.62	5.79	4.75	5.33
4	8.18	5.64	7.36	5.33	8.82	5.26	8.75	7.10
5	5.86	5.46	5.69	6.49	7.81	9.03	7.49	8.98

Qüestió: La producció d'insulina depèn de la concentració de glucosa?

Algunes situacions reals

Volem estudiar la concentració de Cadmi en fetge, ronyó i pancreas en diferent peixos

Dos grups de peixos en funció del tipus d'aigua:

1. Concentració normal de Cadmi
2. Alta concentració de Cadmi

Fish Id.	Group	Liver	Kidney	Pancreas
1	A	0.38	0.09	0.65
2	B	0.14	0.36	0.9
3	B	0.18	0.29	0.34
4	A	0.24	0.19	0.43

Qüestions:

- Hi ha diferències en la concentració de Cadmi entre els dos grups?
- S'absorbeix el cadmi de forma similar en els diferents òrgans?

Algunes situacions reals

Un sistema de bases de dades permet fer cerques indicant un rang de dies anteriors. Es vol desenvolupar una fórmula que permeti predir el temps que triga la cerca en funció del nombre de dies indicats.

Observation Id.	Time	Days
1	0.65	1
2	0.79	2
3	1.36	4
4	2.26	8
5	3.59	16
6	5.39	25

Qüestió: És possible predir el temps de cerca quan coneixem el nombre de dies que s'indica?

Objectiu

En totes aquestes situacions, interessa descriure el comportament d'una variable aleatòria Y, coneguda com **variable dependent** en funció d'altres variables conegudes com a **variables independents, explicatives o predictores**. Quan aquestes variables X són categòriques, es coneixen com a **factors**, i si són numèriques, **covariables**.

Si les dades provenen d'un disseny experimental, les variables explicatives representen les **condicions experimentals**

Principi de Parsimonia

És important tenir en compte:

Principi de Parsimonia: Entre possibles models que ajusten de forma adequada les dades, el que tingui menys paràmetres serà preferible (the Cambridge Dictionary of Statistics, B.S. Everitt)

Per tant, serà especialment important seleccionar les variables a introduir en el model

G. Box:" Essentially, all models are wrong, but some are useful".

Clssificació de models basada en l'objectiu d'ús

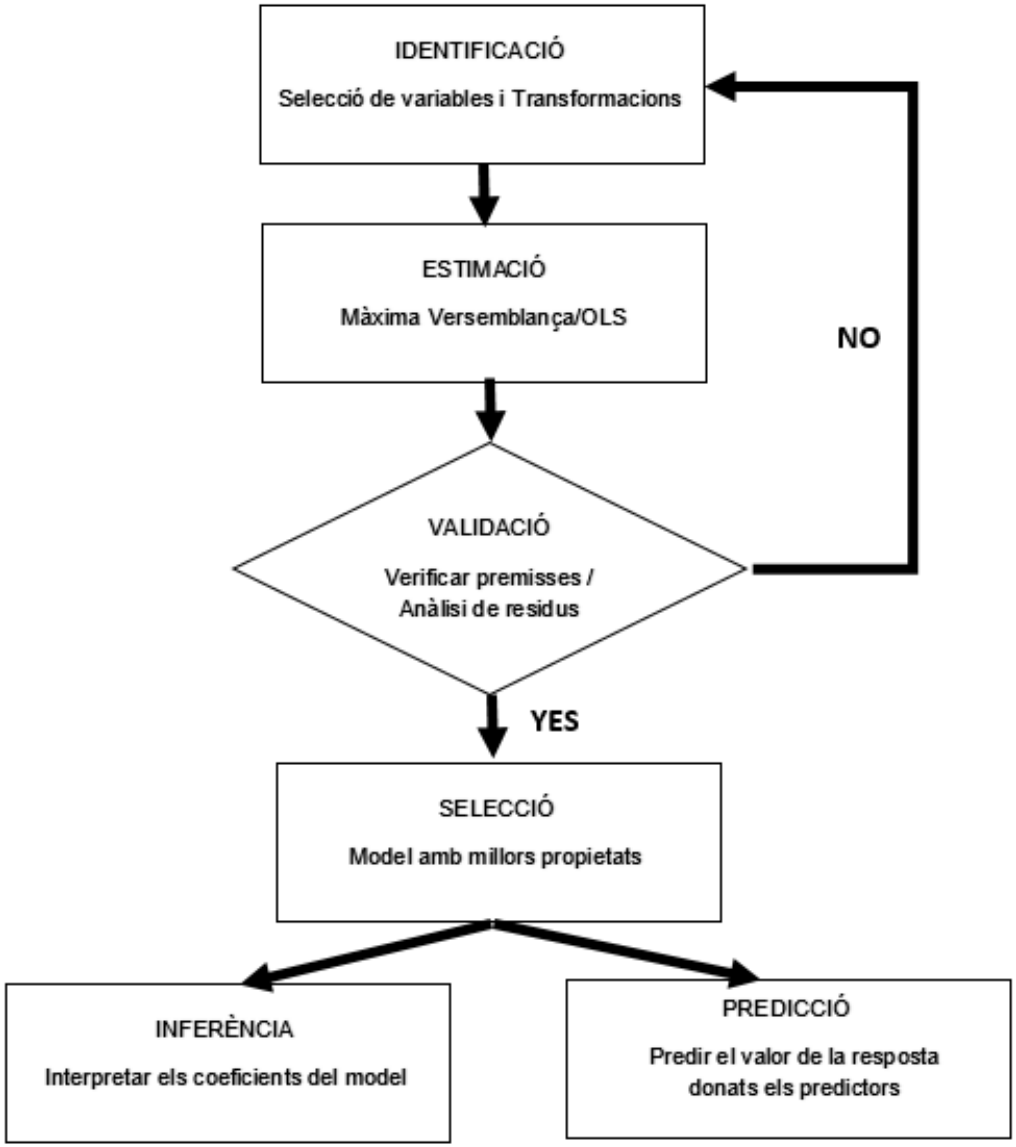
Podem distingir entre dos tipus de aspectes relacionats amb els models que depenen de l'objectiu final:

- **Interés Explicatiu**: Volem descobrir quines variables tenen influència en Y. No volem moltes variables en el model i les que hi siguin les volem el més independents possibles
- **Interés Predictiu**: Volem predir el valor de Y amb precisió quan coneixem els valors de les variables explicatives. Normalment són models amb moltes variables i amb variància residual molt petita

Fases de la Modelització Estadística

0) Dissenyar un experiment i obtenir les dades
1) Realitzar un anàlisi exploratori de les dades amb l'objectiu de:
   a) Localitzar observacions extremes o que difereixen molt de la resta
   b) Establir si té sentit la dependència funcional lineal entre X i Y
   c) Determinar si les variables explicatives estan correlacionades o no
2) Estimar els paràmetres del model
3) Verificar si es compleixen les assumpcions del model
4) Realitzar inferència a partir dels paràmetres del model
5) Explicar i interpretar el model obtingut
6) Predir noves observacions

Modelització estadística



Exemple: Regressió Lineal Simple

Un sistema de bases de dades permet fer cerques indicant un rang de dies anteriors. Es vol desenvolupar una fórmula que permeti predir el temps que triga la cerca en funció del nombre de dies indicats.

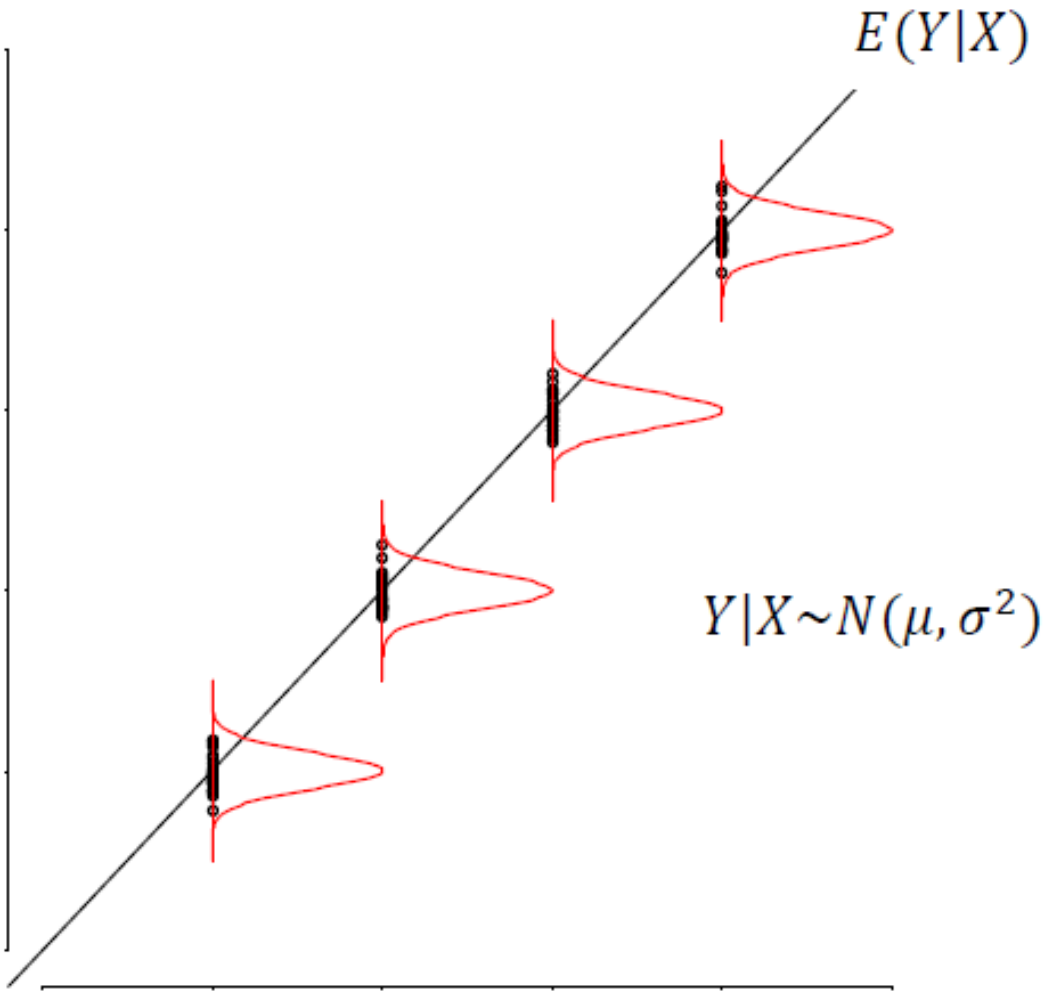
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \cdots, n$$

en forma matricial:

$$\begin{pmatrix} 0.65 \\ 0.79 \\ 1.36 \\ 2.26 \\ 3.59 \\ 5.39 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 8 \\ 1 & 16 \\ 1 & 25 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{pmatrix}$$

Exemple: Regressió Lineal Simple

La distribució de la resposta Y condicionada al valor de X segueix una distribució Normal. El seu valor esperat és una funció lineal de la variable X



Definició de Model Lineal

Objectiu: Explicar el comportament d'una variable aleatòria Y com a funció de $X_1, X_2, \cdots, X_{p-1}$.

Donat $n \in \mathbb{Z}^+$, $\forall i \in \{1, 2, \cdots, n\}$ sigui Y_i la variable relacionada con Y quan $X_1 = x_{1,i}, X_2 = x_{2,i}, \cdots, X_{p-1} = x_{p-1,i}$, on $x_{k,i} \in \mathbb{R}, \forall i, k$.

Definició:

$$\forall i \quad , Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \epsilon_i = \mu_i + \epsilon_i$$

Hipòtesis:

- $\forall i \in \{1, 2, \cdots, n\}$, $\epsilon_i \sim \mathrm{N}(0, \sigma_i^2)$;
- $\forall i \in \{1, 2, \cdots, n\}$, $\sigma_i^2 = \sigma^2$ (homoscedasticity);
- $\forall i, j \in \{1, 2, \cdots, n\}\, i \neq j$, ϵ_i indep.of ϵ_j.
- X valors fixats o, si són aleatoris, independents dels errors

β_0 es coneix com **intercept** i es considera que $x_{0,i} = 1 \qquad \forall i = 1 \cdots n,$.

Formulació del Model Lineal

En forma matricial,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & x_{3,1} & \cdots & x_{p-1,1} \\ 1 & x_{1,2} & x_{2,2} & x_{3,2} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} & x_{3,n} & \cdots & x_{p-1,n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Formulació del Model Lineal

Definint,

Vector de respostes:

$$Y_{n\times 1} = (Y_1, Y_2, \cdots, Y_n)'$$

Matriu de disseny:

$$X_{n\times p} = (x_{ij})$$

Vector de coeficients:

$$\beta_{p\times 1} = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$$

Errors:

$$\epsilon_{n\times 1} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$$

el model s'escriu com:

$$Y = X\ \beta + \epsilon \Longleftrightarrow \mu = E(Y|X) = X\beta$$

$$Y|X \sim N(X\beta, \sigma^2 I_n)$$

on I_n representa la matriu identitat de dimensió n.

Formulació del Model Lineal

Donat que,

$$Y|X \sim N(X\beta, \sigma^2 I_n)$$

tenim que,

$$E((Y - X\beta)(Y - X\beta)') = E\left[\begin{pmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & \epsilon_1\epsilon_3 & \cdots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & \epsilon_2\epsilon_3 & \cdots & \epsilon_2\epsilon_n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & \epsilon_n\epsilon_3 & \cdots & \epsilon_n^2 \end{pmatrix}\right] = \sigma^2 \cdot I_n$$

Formulació del Model Lineal

Observació: Les variables $X_1, X_2, \cdots, X_{p-1}$ poden ser funció d'un altre conjunt de variables. Poden existir $\{Z_1, Z_2, \cdots, Z_m\}$, $m \in \mathbb{N}$, tal que

$$X_i = g_i(Z_1, Z_2, \cdots, Z_m) \qquad \forall i \in \{1, 2, \cdots, p-1\}$$

Exemples:

- $p = 3,\ m = 1;\quad X_1 = Z_1 \qquad X_2 = Z_1^2$

$$\forall i \qquad Y_i = \beta_0 + \beta_1 z_{1,i} + \beta_2 z_{2,i}^2$$

- $p = 3,\ m = 3;\quad X_1 = e^{Z_1} Z_2 \qquad X_2 = Z_3 - Z_2$

$$\forall i \quad Y_i = \beta_0 + \beta_1 e^{z_{1,i}}\, z_{2,i} + \beta_2 (z_{3,i} - z_{2,i})$$

Exemples de Models Lineals

Els models que es fan servir en l'**Anàlisi de la Variància (ANOVA)** són Models Lineals amb variables explicatives de tipus factor (categòriques)

Exemple: Volem comparar la pressió sanguínea (Y) en dos tipus d'individus, uns que han pres una medicació (tractament) i d'altres que no (control).

$$Y_{ij} = \mu_i + \epsilon_{ij}, \ \ \forall i \in \{1,2\}, \ \forall j \in \{1,2,\cdots,n_i\}$$

en forma matricial,

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

Exemples de Models Lineals

Els models coneguts com **models de regressió lineal** són també un cas particular de Model Lineal. En aquest cas, les variables explicatives són numèriques i no categòriques.

Exemple: Volem estudiar el nivell d'un agent químic (Y) en una planta en funció de la quantitat d'aquest agent present en la terra (X).

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \cdots, n$$

en formulació matricial,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Exemples de Models Lineals

El model conegut com **Anàlis de la Covariància (ANCOVA)** és un model lineal on els coeficients de regressió per a una variable numèrica poden canviar segons els nivells d'una variable categòrica

Exemple: Volem estudiar els nivells d'un fàrmac en sang (Y) com a funció de la dosi (X_1). A més, s'ha de tenir en compte el gènere de l'individu (X_2), ja que es sospita que l'efecte pot ser diferent segons el gènere.

$$Y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \epsilon_{ij}, \qquad i \in \{1, 2\}, j \in \{1, 2, \cdots, n_i\}$$

En forma matricial,

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ 1 & x_{12} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ 0 & 0 & 1 & x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{02} \\ \beta_{12} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

Exemples de Models No-lineals

Exemple: Volem estudiar la producció de llet en vaques com a funció del nombre de dies des del part. Si x_i és el número de dies després del part i Y_i indica la producció de llet (en litres), un possible model seria:

$$Y_i = \exp(\beta_0 + \beta_1 x_i + \beta_2 \log(x_i)) + \epsilon_i \qquad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

Exemple: (model jeràrquic) Volem estudiar la qualitat d'una matèria prima suministrada per a diferent proveidors. Per això, de cada proveidor seleccionem aleatòriament un conjunt de b enviaments i de cada un d'ells obtenim n observacions. Un possible model seria:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{k(ij)}$$

on $i = 1, \cdots a, \quad j = 1, \cdots, b, \quad k = 1, \cdots, n \qquad \epsilon_{k(ij)} \sim \mathrm{N}(0, \sigma^2)$

El model NUL

El **model NUL** només conté un paràmetre (el model més simple):

$$Y_i = \beta_0 + \epsilon_i, \qquad i = 1, \cdots, n$$

en forma matricial,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} \beta_0 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

És l'equivalent a l'estudi d'una mostra d'una variable aleatòria

β_0 es denota com *intercept* (ordenada a l'origen). Habitualment considerarem models amb *intercept* ja que contenen el model NUL com a submodel

Estimació de Paràmetres (Mínims quadrats)

Sigui $y = (y_1, y_2, \cdots y_n)'$ una realització de Y i $\hat{\beta}$ una estimació de β

1) **Mínims Quadrats Ordinaris (OLS)** minimitza:

$$S(\beta) = ||y - \hat{y}||_2^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{j,i}\right)^2$$

on $\hat{y} = \hat{\mu} = X\hat{\beta}$.
Solució: $\hat{\beta} = (X'X)^{-1}X'y$, si $X'X$ no és una matriu singular

2) **Mínims Quadrats Ponderats (WLS)** minimitza:

$$S(\beta) = \sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} w_i\left(y_i - \sum_{j=1}^{p}\beta_j x_{j,i}\right)^2$$

on $w_i^{-1} = V(Y_i)$.

3) **Mínims Quadrats Ponderats amb observacions correlacionades** minimitza:

$$S(\beta) = (y - X\beta)'V^{-1}(y - X\beta)$$

on $V = V(Y)$.
Solució: $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$, si $X'V^{-1}X$ no és una matriu singular

Observació: No es requereix cap distribució per Y

Estimació de Paràmetres (Màxima Versemblança)

L'estimador de β maximitza:

$$L(\beta; y, x) = (\sqrt{2\pi}\sigma)^{-n} exp\left(-\sum_{i=1}^{n} \frac{(y_i - \sum_{j=1}^{p} \beta_j x_{j,i})^2}{2\sigma^2}\right)$$

que és equivalent a

$$l(\beta; y, x) = -n\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(y_i - \sum_{j=1}^{p} \beta_j x_{j,i})^2}{2\sigma^2} =$$

$$= -n\log(\sqrt{2\pi}\sigma) - \frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}$$

Definim les eqüacions de l'Score en aquest cas:

$$U_j = \frac{\partial l}{\partial \beta_j} = \frac{1}{\sigma^2}\left(X'(Y - X\beta)\right)_j \qquad \forall j$$

El vector $U = (U_1, U_2, \cdots U_p)^t$ és el vector dels Scores

$$U_j = 0 \qquad \forall j \Longleftrightarrow X'Y = X'X\beta \Longleftrightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

Si el rang de $X'X$ és igual a p, existeix una única solució $\hat{\beta}$, que és U.M.V.U.E.

Distribució dels estimadors (Màxima Versemblança)

Com l'estimador de β és lineal en Y, la distribució d'aquest estimador també serà Normal i:

$$E(\hat{\beta}|X) = E((X'X)^{-1}X'Y|X) = (X'X)^{-1}X'E(Y|X)$$

$$= (X'X)^{-1}X'X\beta = \beta$$

$$V(\hat{\beta}|X) = V((X'X)^{-1}X'Y|X) = \Big((X'X)^{-1}X'\Big)V(Y|X)\Big((X'X)^{-1}X'\Big)' =$$

$$= \Big((X'X)^{-1}X'\Big)V(Y|X)\Big((X(X'X)^{-1}\Big)$$

i com $V(Y|X) = E\Big((Y - X\beta)(Y - X\beta)'|X\Big) = E(\epsilon\epsilon') = V(\epsilon) = \sigma^2 I_n$

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Per tant, $\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$

Estimació de Paràmetres (Màxima Versemblança)

La matriu $\mathcal{J} = E(UU')$ és la **matriu d'informació de Fisher** del model pel paràmetre β

Sota la assumpció de Normalitat:

$$\begin{aligned} \mathcal{J} = E(UU') &= E\Big(\frac{1}{\sigma^2}X'(Y - X\beta)(Y - X\beta)'X\frac{1}{\sigma^2}\Big) \\ &= \frac{1}{\sigma^2}X'E\Big((Y - X\beta)(Y - X\beta)'\Big)X\frac{1}{\sigma^2} \\ &= \frac{1}{\sigma^2}X'X. \end{aligned}$$

Prediccions i residus

El vector de valors predits és

$$\hat{Y} = X\hat{\beta} = (\hat{Y}_1, \hat{Y}_2, \cdots \hat{Y}_n)'$$

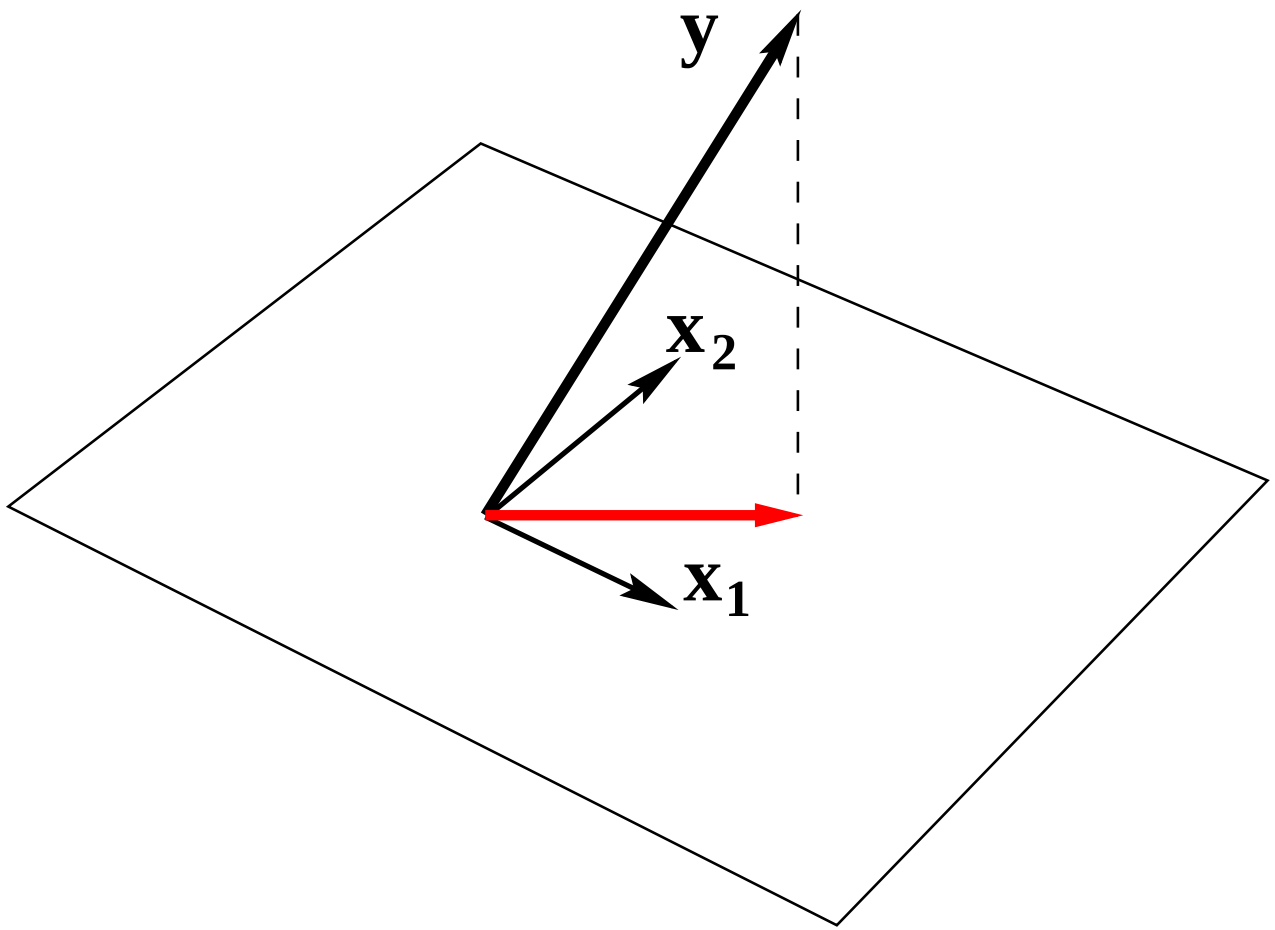
Si $(y_1, y_2, \cdots y_n)'$ és una realització del vector Y, el residu d'una observació y_i és

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

$e = Y - X\hat{\beta}$ es ortogonal a las columnes de la matriu X

$$\begin{aligned} X'e &= X'(Y - X\hat{\beta}) = X'(Y - X(X'X)^{-1}X'Y) \\ &= X'Y - X'X(X'X)^{-1}X'Y = X'Y - X'Y = 0 \end{aligned}$$

Interpretació geomètrica del vector de residus



Estimació de la variància residual

La funció de log-versemblança coma funció de σ^2 és:

$$l(\sigma^2;\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{2\,\sigma^2}$$

diferenciant i igualant a zero:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 = 0 \Longleftrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

L'estimador de màxima versemblança de σ^2 és:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \frac{RSS}{n}$$

on **RSS** és la Suma de Quadrats Residual (*Residual Sum of Squares*).

Estimació de la variància residual

Assumint que $p = rang(X'X)$, es verifica que:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \sim \chi^2_{n-p}$$

i per tant, l'estimador de màxima versemblança té biaix. Fent servir el mètode dels moments podem corregir el biaix de l'estimador: Definim

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{n}{n-p}\hat{\sigma}^2 = \frac{RSS}{n-p}$$

i tindrem

$$E\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) = n - p \quad \Rightarrow \quad E(S^2) = \sigma^2$$

Aquest estimador també es coneix com Error Quadràtic Mig (*Mean Square Error*)

Observació: p is the number of parameters that have been estimated (p-1 variables explicatives+ *intercept*).

Alguns casos simples

- Model Nul: $y_i = \beta_0 + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \qquad i = 1, \cdots n.$

$$\hat{\beta}_0 = \overline{y} \qquad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- Regressió Lineal Simple:
  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \quad i = 1, \cdots, n$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2} = \frac{Cov(X,Y)}{V(X)} = r_{XY} \frac{S_Y}{S_X}$$

Observacions:

- $(\overline{x}, \overline{y})$ pertany a la recta de regressió
- El coeficient de correlació (r_{XY}) és una mesura de relació lineal entre X and Y.

Alguns casos simples

Regressió Lineal Simple:

Desviació estàndard dels estimadors dels paràmetres:

$$S_{\hat{\beta}_0} = \hat{\sigma}\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\right]^{1/2}$$

$$S_{\hat{\beta}_1} = \hat{\sigma}\,\frac{1}{[\sum_{i=1}^{n} x_i^2 - n\,\overline{x}^2]^{1/2}}$$

Interpretació dels paràmetres

Suposem el següent model:

$$\forall i \qquad Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \epsilon_i = \mu_i + \epsilon_i$$

Sigui μ_i el valor esperat de la resposta sota les condicions $X_i = (1, x_{1,i}, x_{2,i}, \cdots, x_{j,i}, \cdots, x_{p-1,i})$ i μ_i^* sota les condicions $X_i^* = (1, x_{1,i}, x_{2,i}, \cdots, x_{j,i} + 1, \cdots, x_{p-1,i})$. Llavors,

$$\mu_i^* - \mu_i = \hat{\beta}_j$$

Per tant,

- $\hat{\beta}_j$ és el canvi esperat en la resposta si augmentem la covariant $x_{j,i}$ en una unitat i deixem la resta igual
- En el cas en que $j = 0$ i per tant es tracta de l'estimació de l'*intercept* ($\hat{\beta}_0$) la interpretació fa referència al valor esperat de la resposta quan totes les covariant valen zero

La desviació estàndard residual S és l'error associat a les prediccions. Per exemple, 95% de les prediccions tindran un residu en:

$$(-t_{n-p,1-\alpha/2} S, t_{n-p,1-\alpha/2} S) \simeq (-1.96S, 1.96S)$$

Inferència pels paràmetres del model

Com $\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$ cada component verifica:

$$\hat{\beta}_i|X \sim N(\beta_i, \sigma^2[(X'X)^{-1}]_{ii})$$

Per tant, donat un valor $a \in \mathbb{R}$, podem resoldre el test:

$$\begin{cases} H_0 : \beta_i = a \\ H_1 : \beta_i \neq a \end{cases}$$

al nivell de significació α calculant l'estadístic

$$t = \frac{\hat{\beta}_i - a}{S\sqrt{[(X'X)^{-1}]_{ii}}}$$

i rebutjant la H_0 si

$$\left| \frac{\hat{\beta}_i - a}{S\sqrt{[(X'X)^{-1}]_{ii}}} \right| \geq t_{n-p,1-\alpha/2}$$

Inferència pels paràmetres del model

En particular, és interessant fer un **test de significació** dels paràmetres:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

En cas que no rebutgem H_0, implica que la covariant X_i no té influència significativa en Y

Interval de confiança pels paràmetres:

$$\hat{\beta}_i \pm t_{n-p,\alpha/2} S \sqrt{[(X^t X)^{-1}]_{ii}}$$

Observació: Si l'interval conté el valor 0, la covariable corresponent no és estadísticament significativa.

Taula de l'ANOVA de la regressió

Considerem

$$SS_{Total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad SS_{Model} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad SS_{Residual} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Es compleix que

$$SS_{Total} = SS_{Model} + SS_{Residual}$$

Demostració:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = SS_{Model} + SS_{Residual} + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

Cal veure que $\sum_{i=1}^{n} e_i(\hat{Y}_i - \bar{Y}) = 0$, o vectorialment $e'(X\beta - \mathbf{1}\bar{y}) = 0$. Però el vector e es ortogonal a les columnes de X (inclosa la columna d'uns associada a l'*intercept*):

$$e'X = (Y - \hat{Y})'X = Y'(I - X(X'X)^{-1}X')X = Y'(X - X) = 0$$

Taula de l'ANOVA de la regressió

La taula de l'ANOVA és:

Font	SS	Graus Llib.	*MSS*	F
Model	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$p-1$	$SS_{Model}/(p-1)$	$F = \frac{SS_{Model}/(p-1)}{SS_{Residual}/(n-p)}$
Residuals	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-p$	$SS_{Residual}/(n-p)$	
Total	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$		

Observació: $S^2 = RSS/(n-p)$

Mesura de bondat d'ajust

Coeficient de determinació: $R^2 = \frac{SS_{Model}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}}$

S'interpreta com la proporció de variabilitat observada en Y que és explicada pel model

Tests per comparar models

Sigui el model lineal $Y = X\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2 I)$

Es vol realitzar un test d'hipòtesi per comparar dos models, on un d'ells és una versió restringida de l'altre:

$$\begin{cases} H_0 : \text{Model restringit} \\ H_1 : \text{Model complet} \end{cases}$$

Aquesta situcació implica una hipòtesi lineal de rang q sobre els parèmetres i es resol amb el test lineal general de la F.

Si H_0 és certa, llavors:

$$\hat{F} = \frac{(SS_{H_0} - SS_{Residual})/q}{SS_{Residual}/(n-p)} \approx F_{q,n-p}$$

on SS_{H_0} i $SS_{Residual}$ són respectivament les sumes de quadrats residuals del model restringit i del complet

Tests per comparar models

Si apliquem el test de la raó de versemblança generalitzada:

Pel model restringit, $\tilde{\sigma}^2_{H_0} = \frac{SS_{H_0}}{n}$ i $L(\tilde{\sigma}^2_{H_0}) = \frac{e^{-n/2}}{(2\pi\tilde{\sigma}^2_{H_0})^{n/2}}$

Pel model complet $\hat{\sigma}^2_{MV} = \frac{SS_{Residual}}{n}$ i $L(\hat{\sigma}^2_{MV}) = \frac{e^{-n/2}}{(2\pi\hat{\sigma}^2_{MV})^{n/2}}$

Per tant, l'estadístic del test de raó de versemblança és:

$$\Lambda = \left(\frac{\hat{\sigma}^2_{MV}}{\tilde{\sigma}^2_{H_0}}\right)^{n/2} = \left(\frac{SS_{Residual}}{SS_{H_0}}\right)^{n/2}$$

i per tant, l'estadístic $\hat{F}$ és una funció decreixent de l'estadístic Λ

$$\hat{F} = \frac{n-p}{q}\left(\Lambda^{-n/2} - 1\right)$$

Test Omnibus

Si β_0 correspon a l'*intercept*, el test **omnibus** es defineix com:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_{p-1} = 0 \\ H_1 : \exists i \quad \beta_i \neq 0 \end{cases}$$

Observació: Aquest test compara el model actual amb el model Nul

El procediment de resolució del test és:

$$\text{Si } H_0 \text{ és cert} \qquad F = \frac{SS_{Model}/(p-1)}{SS_{Residual}/(n-p)} \sim F_{p-1,n-p}$$

Per tant,

$$H_0 \text{ es rebutja si} \qquad F \geq F_{1-\alpha,p-1,n-p}$$

on $SS_{Model} = SS_{Total} - SS_{Residual}$

Observation: El test Omnibús no és equivalent a testar la significació de cada coeficient per separat

Distribució de les prediccions

Si definim el **vector de valors predits** com $\hat{Y} = X\hat{\beta}$, la seva distribució serà:

$$\hat{Y}|X \sim N(X\beta, \sigma^2 X(X'X)^{-1}X')$$

deguta a que és una funció lineal de variables aleatòries Normals i

$$E(\hat{Y}|X) = XE(\hat{\beta}|X) = X\beta$$

$$\begin{aligned} E((\hat{Y} - X\beta)(\hat{Y} - X\beta)'|X) &= XE((\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X)X' \\ &= X\sigma^2(X'X)^{-1}X' \\ &= \sigma^2 X(X'X)^{-1}X' \end{aligned}$$

La matriu $H = X(X'X)^{-1}X'$ s'anomena **matriu de projecció** (*hat matrix*) degut a que $\hat{Y} = X(X'X)^{-1}X'Y$

Interval de Predicció (PI)

Sigui $X_0 = (1, x_{1,0}, x_{2,0}, \cdots, x_{p-1,0})'$ unes certes condicions experimentals.

El valor predit pel valor X_0 és $\hat{y}_0 = X_0'\hat{\beta}$.

Un **interval de predicció (PI)** al $(1-\alpha)\%$ per aquesta observació puntual és

$$\hat{y}_0 \pm t_{1-\alpha/2,n-p}\sqrt{S^2\left(1 + X_0'(X'X)^{-1}X_0\right)}$$

- En el cas particular de la regressió lineal simple, l'interval de predicció és:

$$IC_{1-\alpha}(y_0) = \hat{y}_0 \pm t_{1-\alpha/2,n-p}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(X_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}$$

Interval de Confiança per la resposta esperada (CI)

Sigui $X_0 = (1, x_{1,0}, x_{2,0}, \cdots, x_{p-1,0})'$ unes certes condicions experimentals.

El valor esperat de la resposta pel valor X_0 és $\mu_0 = E(Y|X_0) = X_0'\beta$.

En conseqüència, $\hat{y}_0$ és un estimador sense biaix de $E(Y|X_0)$ i

$$V(\hat{y}_0) = \sigma^2 X_0'(X'X)^{-1}X_0$$

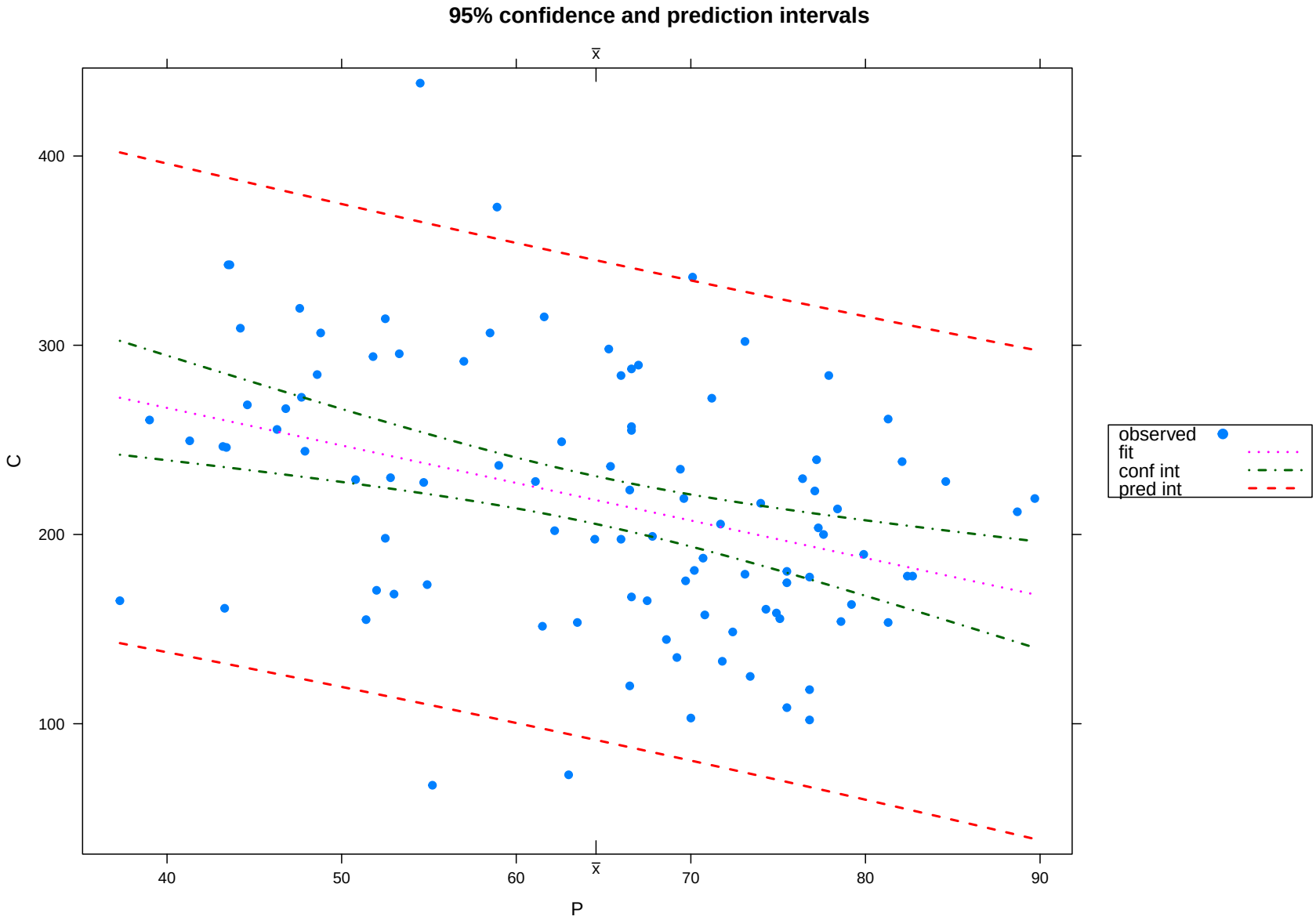
i per tant, un interval de confiança per a μ_0 is:

$$IC_{1-\alpha}(\mu_0) = \hat{y}_0 \pm t_{1-\alpha/2,n-p}\sqrt{S^2\, X_0'(X'X)^{-1}X_0}$$

- En el cas particular de la regressió lineal simple, l'interval de confiança per μ_0 és:

$$\hat{y}_0 \pm t_{1-\alpha/2,n-p}\sqrt{S^2\left(\frac{1}{n} + \frac{(X_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}$$

Intervals de predicció i confiança



Distribució del vector de residus

El residu i-ésim es defineix com $e_i = y_i - \hat{y}_i$.

El vector $e = (Y_1 - \hat{Y}_1, Y_2 - \hat{Y}_2, \cdots, Y_n - \hat{Y}_n)'$ correspon al vector de residus i verifica:

$$e|X \sim N(0, \sigma^2(I - X(X'X)^{-1}X'))$$

perquè és una combinació lineal de variables aleatòries Normals i

$$E(e|X) = E(Y - X\hat{\beta}|X) = X\beta - X\beta = 0$$

i com $E(YY'|X) = E(\hat{Y}\hat{Y}'|X) + E((Y - \hat{Y})(Y - \hat{Y})'|X)$

$V(e|X) = E((Y-\hat{Y})(Y-\hat{Y})'|X) = \sigma^2 I - \sigma^2 X(X'X)^{-1}X' = \sigma^2(I - X(X'X)^{-1}X').$

Residus Estandarditzats i Studentitzats

Una estimació de la variància de $(y_i - \hat{y}_i)$ es pot obtenir com:

$$S^2(1 - [X(X'X)^{-1}X']_{ii})$$

Denotant per $h_{ii} = [X(X'X)X']_{ii}$, els **residus estandarditzats** es defineixen com:

$$\textbf{Stand.Res} = \frac{y_i - \hat{y}_i}{S\sqrt{1 - h_{ii}}}$$

i el **residus studentitzats** com:

$$\textbf{Student.Res} = \frac{y_i - \hat{y}_i}{S_{(-i)}\sqrt{1 - h_{ii}}}$$

on $S^2_{(-i)}$ és la estimació de la variància residual quan es suprimeix l'observació i-ésima del model

El terme h_{ii} s'anomena el **coeficient d'apalancament o anclatge** de l'observació i-ésima (*leverage*)

Multicol · lienalitat

Definició: Terme usat en l'anàlisi de la regressió per indicar situacions on les varibales explicatives estan relacionades d'acord a una funció lineal... (the Cambridge Dictionary of Statistics, B.S. Everitt)

Quan dues variables predictores són linealment dependents, s'anomenen **col · linears**

La **Multicol · linealitat** implica que $det(X'X)$ és molt petit o igaul a zero en el cas extrem.

Si existe Multicol · linealitat

- La interpretació del model és més complicada
- La variància del estimadors $\hat{\beta}_i$ és gran
- La matriu $X'X$ pot ser singular, i les equacions normals donen infinites solucions per $\hat{\beta}$

Observació: Inclús si hi ha Multicol · linealitat, les prediccions són correctes si el model és correcte

Diagnòstic de la Multicol · linealitat

Para detectar la multicol · linealitat és apropiat realitzar un diagrama de punts múltiple de totes les variables predictores (`pairs`)

Correlació entre parelles de predictors

$$r_{x_1x_2} = \frac{\sum_i x_{1i}x_{2i} - n\overline{x}_1\,\overline{x}_2}{\left[\sum_i x_{1i}^2 - n\overline{x}_1^2\right]^{1/2}\left[\sum_i x_{2i}^2 - n\overline{x}_2^2\right]^{1/2}}$$

Observació: Si un terme de correlació és relativament alt, una de les variables implicades s'ha d'eliminar del model i tornar a fer l'ajust

Diagnòstic de la Multicol · linealitat (VIF)

És convenient calculat el **Factor d'Inflament de la variància** (*Variance Inflation Factor (VIF)*) de cada variable, definit com

$$VIF(X_j) = \frac{1}{1 - R^2(X_j)}$$

on $R^2(X_j)$ és la R^2 obtinguda en la regressió que té com a variable resposta la variable X_j respecte a la resta de predictores

VIF	Conclusió
VIF=1	no correlacionades
$1 < VIF < 5$	moderadament correlacionades
$VIF > 5$	altament correlacionades

Observació: $1 - R^2(x_j)$ és coneix com a **tolerància**

Diagnòstic de la Multicol · linealitat (VIF)

El terme VIF prové de la següent propietat:

$$V(\hat{\beta}_j) = \frac{1}{1 - R^2(X_i)} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2} = VIF(X_j) \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}$$

Leverage

Donada una observació Y_i, su factor d'apalancament (*leverage*) es defineix com l'element [ii] de la matriu de projecció H:

$$h_{ii} = [X(X'X)^{-1}X']_{ii}$$

És una mesura de la distància entre $(x_{1,i}, x_{2,i}, \cdots, x_{p-1,i})$ i el centroid $(\overline{x_1}, \overline{x_2}, \cdots, \overline{x_p})$.

El *leverage* verifica que $1/n \leq h_{ii} \leq 1 \qquad \forall i$.

Observacions que:

$$h_{ii} > 3p/n \qquad \text{or} \qquad h_{ii} > 0,99$$

tenen un *leverage* elevat i són considerades potencials dades influents

Observacions Influents (Distància de Cook)

Una **observació influent** és una observació que afecta de forma significativa els valors dels coeficients de la regressió.

La **distància de Cook** es pot calcular com:

$$D_i = \frac{\sum_{i=1}^{n}(\hat{y}_j - \hat{y}_{j(-i)})^2}{pS^2} \qquad i = 1, \cdots, n$$

on $\hat{y}_{j(-i)}$ fa referència a la predicció per a la observació j-ésima pel model sense la observació i-ésima. Una expressió alternativa és:

$$D_i = \frac{1}{p}\frac{h_{ii}}{(1-h_{ii})^2} \cdot \frac{e_i^2}{S^2} \qquad i = 1, \cdots, n$$

i pot ser elevada si la observació té un *leverage* alt o si té un residu estandarditzat elevat

Observacions Influents (Distància de Cook)

Observacions:

- Les dades extremes (*Outliers*) són observacions influents
- Si el *leverage* i el residu d'una observació són alts, la observació serà influent

Observacions Influents (Distància de Cook)

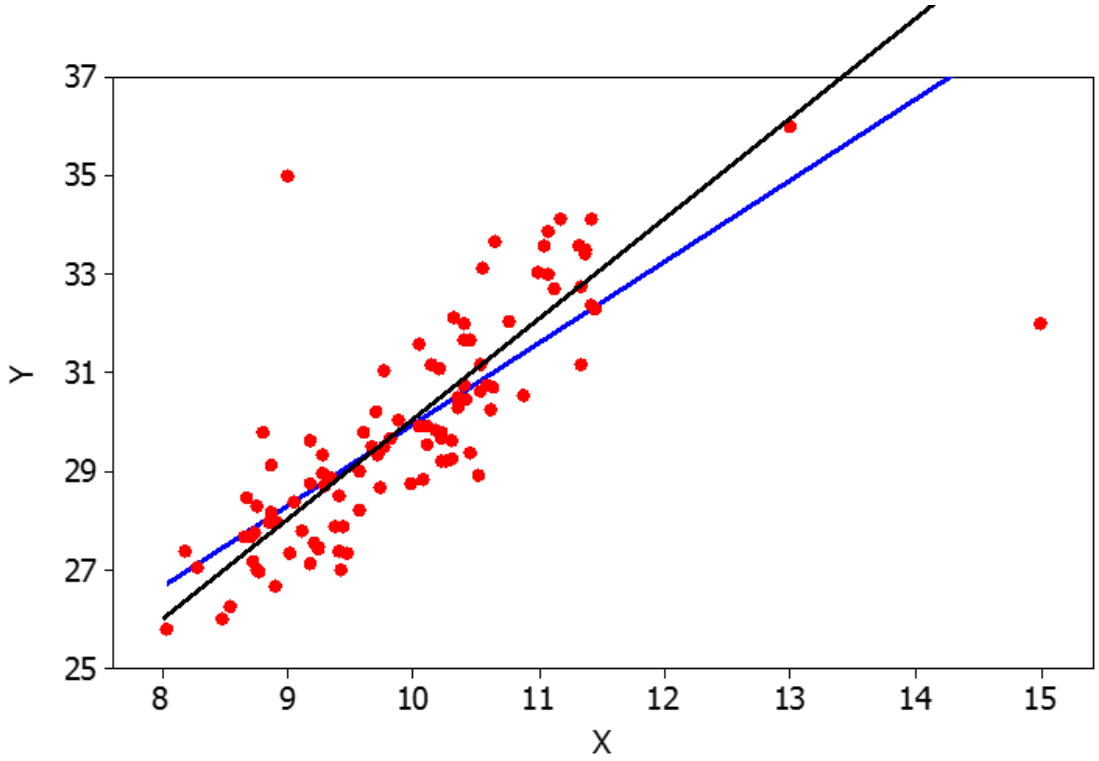


Figure 1: Dues línies de regressió: amb totes les observacions (blau), sense el punt influent (negre)

Mesures d'adequació del model a les dades (R^2)

Coeficient de correlació múltiple o de determinació:

$$R^2 = \frac{SS_{Model}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

$$R^2 = c' R_{xx} c$$

on $c = (r_{x_1 y}, r_{x_2 y}, \cdots r_{x_p y})$ i $R_{xx} = (r_{x_i x_j})_{i,j}$ essent $r_{a,b}$ la correlació lineal dels vectors a i b

1) $R^2 \in [0, 1]$
2) quan més gran millor ajust del model
3) És una mesura de la correlació entre els valors observats i els predits pel model
4) Si les columnes d'X estan incorrelades, llavors $R_{xx} = I_p$ i aleshores

$$R^2 = c' \cdot c = ||c||_2^2$$

Mesures d'adequació del model a les dades (R^2 – *ajustada*)

És evident que si aumentem el nombre de variables en el model, la R^2 creix.

Per tal de penalitzar models que tinguin més covariables i poder comparar models amb diferent nombre de predictors, es calcula el **coeficient de determinació ajustat** (R^2-ajustat)

$$R^2_{adj} = 1 - \frac{SS_{Residual}/(n-p)}{SS_{Total}/(n-1)} = 1 - \frac{S^2}{S^2_Y} = 1 - (1-R^2)\frac{n-1}{n-p}$$

Observació: Si aumenta p, R^2_{adj} decreix amb relació a R^2

Validació del model: Anàlisi de residus

Verificando les assumpcions del model:

- Diagrama de punts (*scatterplot*) de y vs x: S'ha d'observar la linealitat
- Diagrama de punts de e_i vs $\hat{y}_i$, No s'ha d'observar tendències ni canvis en la variància.
- qq-plot per e_i, s'ha de validar la normalitat dels residus d'aord a la línia de referència
- Algunes vegades s'obté el plot e_i vs order per detectar una possible dependència.

Resum

- Estimació de paràmetres: $\hat{\beta} = (X'X)^{-1}X'y$
- Prediccions: $\hat{y}_i = (X\hat{\beta})_i$.
- Suma de quadrats residual: $SS_{Residual} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- Estimació de la variància residual: $S^2 = \hat{\sigma}^2 = MSE = \frac{SS_{Residual}}{n-p}$
- Desviació estàndard de $\hat{\beta}$: $S_{\hat{\beta}_j} = S \cdot \sqrt{c_{jj}}$, essent c_{jj} els elements diagonals de $(X'X)^{-1}$
- Interval de confiança pel paràmetre: $\hat{\beta}_i \pm t_{1-\alpha/2,n-p} S_{\hat{\beta}_j}$
- Coef. de determinació:

$$R^2 = \frac{SS_{Model}}{SS_{Total}} = \frac{SS_{Total} - SS_{Residual}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

Errors freqüents

- No verificar si la relació és lineal
- Basar-se en els resultats automatitzats sense verificació visual
- No tenir en compte que les estimacions de paràmetres depenen de les unitats de les variables predictores i de la resposta
- Confondre el coeficient de determinació i el de correlació
- Fer servir variables explicatives altament correlacionades
- Fer servir la regressió per predir més eenllà del rang de les variables predictores mesurades (extrapolació)

Errors freqüents

- Fer servir masses variables predictores (sobreajust,*overfitting*).
- Mesurar només un subconjunt petit del rang total de les variables explicatives
- Confondre correlació amb causalitat: dues variables poden estar altament correlacionades, però cap de elles controla l'altre

Errors freqüents: Extrapolació

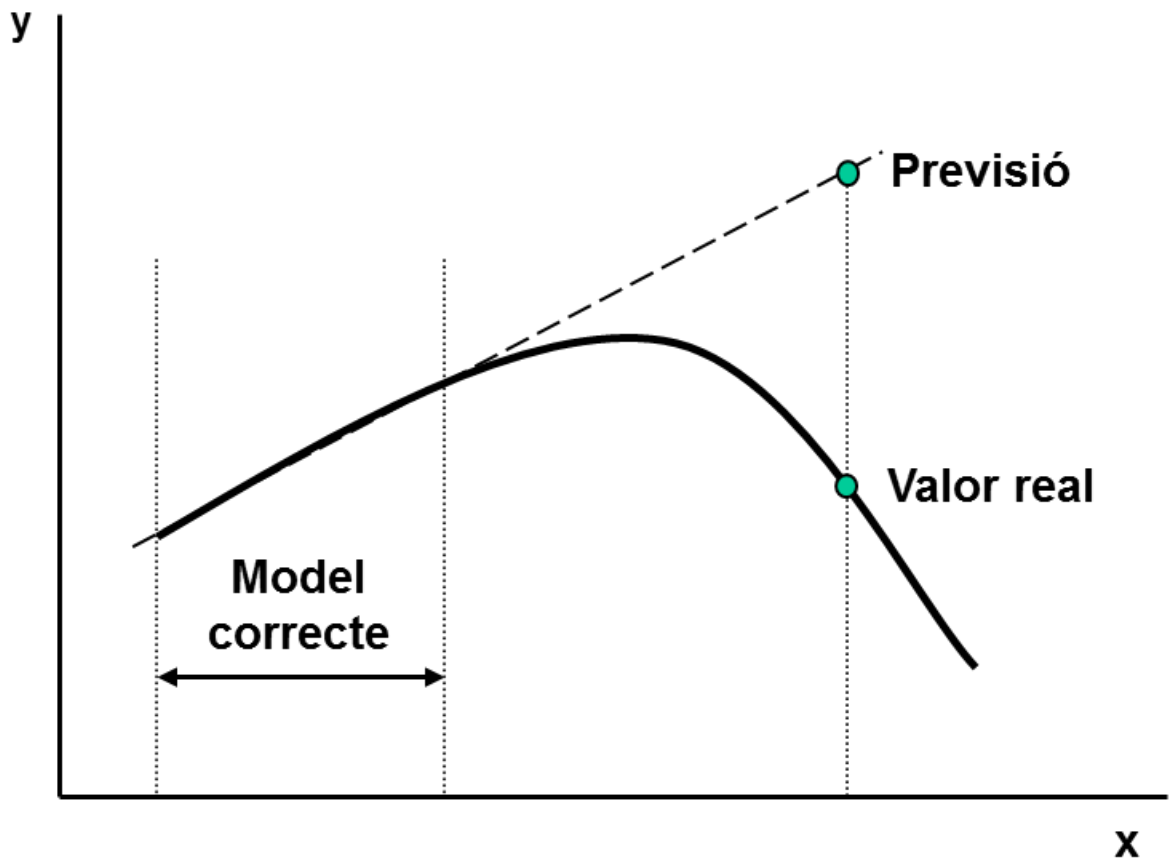


Figure 2: Extrapolació en general no és aconsellable