Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Models estadístics i ciència de dades Conceptes

Bloc D – Probabilitat i Estadística

2024

Panoràmica

Bloc A. Les bases de la **Probabilitat** i les **Variables Aleatòries**

Bloc B. **Models probabilístics parametritzats** de Variables Aleatòries

Bloc C. **Estimació dels paràmetres** desconeguts (puntual i per IC)

(en mostres aleatòries i usant les bases de la inferència estadística)

Bloc D. Informació a partir de **models estadístics parametritzats**

- Usant eines (R) per fer l'explotació i anàlisis de les dades
- Fent èmfasi en la interpretació i la capacitat predictiva del model
- Introduint models (no parametritzats) vinculats a Ciència de Dades (machine learning, ...)

Índex

Estudis estadístics

Eines estadístiques

De IC a PH i NP. P-value

MODELS ESTADÍSTICS.

Notació i casos. Funcions en R. Avaluació

(a) Model sobre un valor concret d'un paràmetre
(**μ** d'una mostra o **μ_D** de la diferència en mostres aparellades)

(b) Model comparant el paràmetre **μ** en mostres independents

(c) Model LINEAL simple i múltiple
(coeficients d'una recta com a paràmetres **β_0 β_1** ...)

\+ Altres tècniques (ciència de dades: visualització multidimensional, clustering, machine learning, data mining,...)

Estudis estadístics

Un estudi estadístic treu **informació** de les **dades** analitzant la relació entre variables.

L'estudi ha de descriure (**Estadística Descriptiva**) les dades. Totes.

Com hem vist al bloc C, cal establir:

- l'obtenció de les dades — Només una m.a. justifica mesures d'incertesa com SE o ICs
- el paper de cada variable — Resposta **Y**, causa assignable **X**, i covariables **Z**
- el tipus d'estudi — Experimental o observacional
- el disseny — Aparellat, grups independents, ...

Per poder incrementar el coneixement, un estudi **científic** ha de ser **reproduïble**

> Si no es poden replicar els resultats, no és una investigació científica. Una investigació fracassa si no pot ser reproduïda

La incapacitat de reproduir un experiment és un problema <u>científic</u> i <u>social</u>:

> Les investigacions metodològicament pobres són un malbaratament de recursos

En el Bloc D veurem **eines** i **models** estadístics per representar relacions entre variables

Eines estadístiques. IC, PH i NP

- La inferència estadística permet inferir o estimar característiques de la població (paràmetres) a partir de les observacions d'una mostra, per quantificar i documentar unes possibles conclusions.

- Un **interval de confiança (IC)** permet estimar un paràmetre informant dels seus valors versemblants.

- Les proves de significació (**proves d'hipòtesis o PH**) plantegen avaluar l'evidència en contra d'un valor concret d'interès del paràmetre (hipòtesis) a partir de les dades. És una metodologia conservadora que planteja una hipòtesi nul·la (H_0) assignant un valor del paràmetre com a punt de partida i estudia si les dades proporcionen proves en contra seva. *No refutar H_0 no vol dir haver demostrat que H_0 és certa*.

- Els **contrastos de Neyman-Pearson (NP)** serveixen per **prendre decisions** acotant riscos a través de dues hipòtesis. L'avantatge sobre les PH és que permeten afitar els errors al decidir per una de les dues opcions confrontades.

A l'annex d'aquest bloc D trobareu més informació sobre PH i sobre els tipus d'errors al contrastar dues hipòtesis

A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més detalls de les proves de significació i contrast de dues hipòtesis al capítol 4

Eines estadístiques i p-valor (*p-value*)

- Un IC indica, amb un cert risc, els **valors versemblants d'un paràmetre** d'acord amb l'evidència aportada per les dades.

- El **P-valor** és la probabilitat d'obtenir per atzar un resultat més "extrem" que el de la mostra observada; una PH estableix (hipotèticament) un valor concret a un **paràmetre** clau.

  Quan més petit sigui el P-valor, menys versemblant és el valor assignat al paràmetre que s'estudia.

  - (és molt rar trobar situacions similars o més estranyes que la mostra observada: porta a dubte respecte el valor establert)
  - (és difícil justificar que les diferències entre la mostra observada i el valor contrastat es deuen només a l'atzar)

  Un P-valor que no sigui petit no aporta evidència per assegurar que el valor real del paràmetre sigui l'establert.

  Fal·làcia habitual: com el P-valor es gran, el valor hipotètic del paràmetre deu ser bo.

A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més detalls en el capítol 4

R usualment indica els **estimadors** ("Estimate") complementats amb el seu **error estàndard** ("Std.Error"), i l'**estadístic** (per ex. "t" o "t value") i el **P-valor** associats a un possible valor concret del paràmetre (per ex. "Pr(>|t|)" o "p-value")

MODELS ESTADÍSTICS. Notació

Un model estadístic explica la variabilitat d'una resposta (Y) separant una part determinista (una **fórmula amb paràmetre/s θ**) i una part aleatòria ($\boldsymbol{\varepsilon}$, amb una distribució adequada):

$$\boldsymbol{Y_i = f(\theta) + \varepsilon_i}$$

$\boldsymbol{f(\theta)}$ és la part determinista que explica el valor de la resposta Y en funció dels paràmetres **θ** dels quals obtindrem estimacions puntuals i intervals de confiança. Per exemple:

- el valor del paràmetre de tendència central (mitjana) en una mostra de la variable de resposta Y
- la diferència de mitjanes en dues mostres (aparellades o independents) de la variable de resposta Y
- l'equació d'una recta que relaciona dues variables Y i X en una mostra

$\boldsymbol{\varepsilon_i}$ (soroll, error, residu,...) representa la part individual no recollida pel model determinista

- No informa sobre la relació entre les variables, i és diferent per a cada observació
- Un model és millor si té ε petites. Interessa V(ε) petita, i amb esperança 0
- Molt habitualment es modela amb una distribució Normal: $\varepsilon \sim N(0, \sigma)$ on interessa σ mínima

Per a una observació *i*, a partir del model i de les característiques de la observació, podem obtenir:
- una **predicció**, $\widehat{Y}_i = f(\hat{\theta})$, aplicant l'estimació de la part determinista del model
- un **error** o **residu** (**e**$_i$) fent la diferència entre l'observat Y_i i la predicció del model ($\widehat{Y}_i$)

MODELS ESTADÍSTICS.

a) Model sobre el valor del **paràmetre μ** (predicció d'un cas *i*)

$\mathbf{Y_i = \mu + \varepsilon_i}$

Objectiu: estimació puntual i per IC de μ ($\hat{\mu} = \bar{y}$) $(\hat{Y}_i = \hat{\mu} = \bar{y})$

Cas particular: **mostres aparellades**, estudi de la diferència de paràmetres μ_1, μ_2

$\mathbf{D_i = \mu_D + \varepsilon_i}$ D = Y1 – Y2 (la diferència com a variable de resposta)

Objectiu: estimació puntual i per IC de μ_D ($\hat{\mu}_D = \bar{d}$) $(\hat{D}_i = \hat{\mu}_D = \bar{d})$

b) Model comparant el **paràmetre μ en *k* mostres independents** (*k≥2* grups)

$\mathbf{Y_i = \mu + \vartheta_j + \varepsilon_i}$

Objectiu: estimació puntual i per IC de cada μ_j ($\hat{\mu}_j = \bar{y} + \hat{\vartheta}_j$) $(\hat{Y}_{ij} = \hat{\mu}_j = \bar{y} + \hat{\vartheta}_j)$

El model contempla μ com a mitjana de referència i ϑ_j com a canvi de la mitjana del grup *j*

c) Model **lineal simple o múltiple**

$\mathbf{Y_i = \beta_0 + \beta_1 X1_i + \varepsilon_i}$ $(\hat{Y}_i = \hat{\beta}_0 + \cdots + \hat{\beta}_k Xk_i)$

$\mathbf{Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \cdots + \beta_k Xk_i + \varepsilon_i}$ $= b_0 + b_k Xk_i)$

Objectiu: estimació puntual i per IC dels coeficients β_i ($\hat{\beta}_0 = b_0$, $\hat{\beta}_1 = b_{1,...}$) de l'equació d'una recta

Determinen una relació entre la resposta i les variables explicatives a través d'una relació lineal

+ Altres tècniques (ciència de dades: visualització multidimensional, clustering, machine learning,...)

Per exemple: PCA [Y1, Y2, ... Yk] ---- *transformació geomètrica* ----> $[\Psi 1, \Psi 2 \ldots]$ $[\ldots \Psi k]$

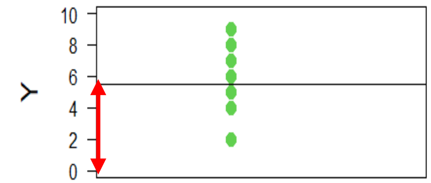
MODELS ESTADÍSTICS. Funció lm() de R

(lm de "*linear model*" però només el cas c) és el que anomenarem estrictament model lineal)

En els casos anteriors de models, la funció lm() en R que li correspon és:

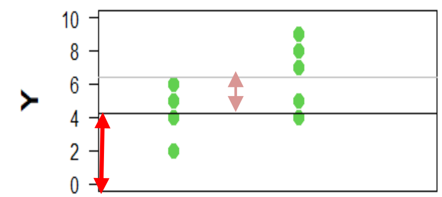
a) $Y_i = \mu + \varepsilon_i$ `lm(Y~1)`

$D_i = \mu_D + \varepsilon_i$ `lm(D~1)`



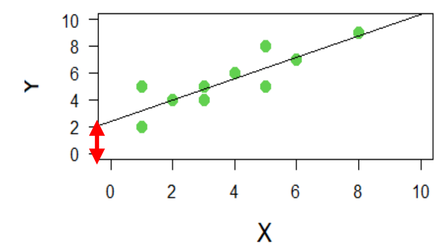
b) $Y_i = \mu + \vartheta_k + \varepsilon_i$ `lm(Y~G)`

`(G columna separant k grups)`



c) $Y_i = \beta_0+\beta_1 X_i + \varepsilon_i$ `lm(Y~X)`

`(o lm(Y~X1+X2+...+Xk))`



En R, els resultat de les **estimacions puntuals dels paràmetres** són:

- **"Coefficients"** on trobarem **"Estimate"** per als diferents paràmetres:
  - $\hat{\mu}$ o $\hat{\mu}_D$ o $\hat{\beta}_0$ com a valors base o de referència de la resposta Y ( )
  - $\widehat{\vartheta}_j$ com a canvi en les mitjanes entre grups ( )
  - $\hat{\beta}_1$ com a pendent de l'equació de la recta

També, per a cada estimació puntual, podem trobar:

- **"Std Error"** és l'error estàndard de l'estimador que permet el càlcul de IC
- **"p-value"** per avaluar la versemblança de valors concrets del paràmetre (per defecte el valor 0)

MODELS ESTADÍSTICS. Avaluació

Avaluarem els models en dos aspectes:

- **Validació de les premisses** per comprovar i assegurar que té sentit aplicar el model

  (en principi, les comunes a tots els models seran **mostra aleatòria** i **normalitat** de la resposta)

- **Anàlisi dels resultats** que pot implicar:
  - Indicadors per valorar la "qualitat" dels resultats i, si escau, la **capacitat predictiva del model**
  - **Interpretabilitat** dels resultats

L'avaluació ha de permetre detectar si un model és adequat, o si pot presentar problemes com per exemple de **sobre-ajustament** o **overfitting** (més detalls al final d'aquest bloc D)

En els següents apartats es presentaran cadascun dels models indicats prèviament, amb les seves particularitats i concretant la **validació de les premisses** i l'**anàlisi dels resultats**

Abans de presentar els models, veurem el cas d'haver de transformar les dades, per exemple amb logaritmes per aconseguir complir la premissa de normalitat de la resposta. Veurem com desfer la transformació en les resultats

MODELS ESTADÍSTICS. Transformacions

En qualsevol dels models indicats, a vegades cal una **transformació** (molt sovint **logarítmica**) de la variable resposta o de les variables explicatives per complir les premisses del model (per exemple la premissa de Normalitat)

Després cal **desfer* la transformació a les estimacions** (puntuals o per IC) obtingudes:

- En el cas de models pel paràmetre μ, si la normalitat es compleix per ln(Y), llavors el model serà

  ln(Y) = Y' = μ' + ε amb estimacions $\bar{y}'$ i IC = [inf',sup']

  i desfent logaritmes obtenim l'estimador puntual i l'IC buscat: **$\exp(\bar{y}')$** i **[exp(inf') , exp(sup')]**

  - En mostres aparellades, no feu logaritmes de D=Y1-Y2 si D pot ser negativa. Quan la diferència augmenta proporcionalment a Y, proveu la diferència de logaritmes o logaritme del rati Y1/Y2.

    ln(Y1) - ln(Y2) = ln(Y1/Y2) = Y'' = μ'' + ε amb estimacions $\bar{y}''$ i IC = [inf'',sup'']

    I desfent logaritmes obtenim estimadors del valor esperat del rati Y1/Y2: **$\exp(\bar{y}'')$** i **[exp(inf'') , exp(sup'')]**

- En el cas del model lineal simple o múltiple, s'ajusta una recta amb les variables transformades i després es poden desfer les transformacions a les prediccions

* Al desfer una transformació logarítmica, ja no fem un IC de μ, sinó de μ' (μ és **mitjana aritmètica** i μ' és **mitjana geomètrica**)

Models estadístics
Model (a) i funcions en R
(model sobre el valor del paràmetre μ)

Model per estimar μ

Notació del model: $Y_i = \mu + \varepsilon_i$

Funciones de R: `lm(Y~1)` i `summary(lm(Y~1))`

(la constant **1** a la dreta de **~** és la sintaxi que usa R per indicar que només volem estimar el paràmetre **μ**)

R proporciona:

- l'estimació puntual de μ ($\hat{\mu} = \bar{y}$). R ho indica com a **"Estimate"** de l'**intercept**
- l'estimació de l'error tipus (se) assumint m.a de l'estimador anterior

  R ho indica com a **"Std. Error"** (*se*) de l'**intercept** ($\bar{y}$), que permetria construir un IC per a μ

$$IC(1-\alpha\%, \mu) = \bar{y} \pm t_{n-1,1-\alpha/2} \cdot se$$

- el valor de l'estadístic (**"t value"**) i el **p-value** (**"Pr(>|t|)"**) per avaluar la versemblança d'un possible valor concret del paràmetre (per defecte el valor 0)
- la desviació residual (**"Residual Standard error"**) o desviació de la part aleatòria que no recull el model

Model per estimar μ

Validació de les premisses

Les **premisses** són: **mostra aleatòria** i **normalitat de la resposta**

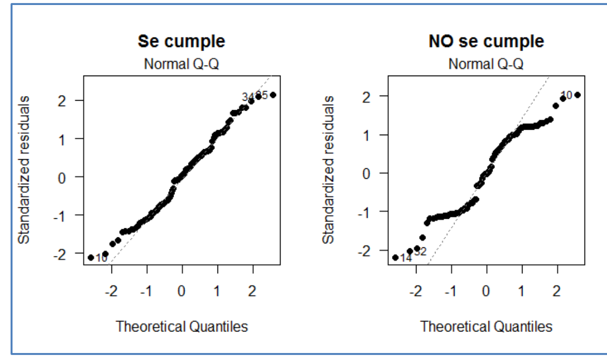
- La premissa de **mostra aleatòria** (m.a.) **no es pot verificar**. Depèn de que el disseny de recollida de dades s'hagi realitzat de forma correcta. Es podria únicament verificar la independència respecte a l'ordre de recollida de les dades.
- La premissa de **normalitat** l'avaluarem gràficament amb un **QQ-plot de Normalitat**

Aquest gràfic representa els quantils empírics de les dades enfront dels quantils teòrics

Si els punts queden (aproximadament) sobre la recta tenim dades (evidència) a favor normalitat

En R es construeix amb `qqnorm(Y) qqline(Y)`

(més informació a l'annex del bloc B)



Anàlisi dels resultats

El resultat bàsic és l'estimació puntual (i per IC) de la mitjana poblacional

La desviació residual és l'indicador de la variabilitat que el model no recull

Model per estimar μ. Exemple

Ex: mostra de 9 valors positius i negatius (per exemple errors en unes mesures)

```
> X = c(-4,-2,-1, 0, 0, 4, 8, 8, 9) # mean=2.4 sd=4.9
```

Solució per Bloc C (IC)

```
> t.test(X)
```

```
t = 1.496, df = 8, p-value = 0.173
  alternative hypothesis: true mean is not equal to 0
  95 percent confidence interval:
   -1.323423  6.212312
  sample estimates: mean of x  2.444444
```

Solució per Bloc D (model estadístic)

```
> (mod <- lm(Y~1))
> summary(mod)
```

```
Residuals:
   Min      1Q Median      3Q     Max
-6.444 -3.444 -2.444  5.556  6.556

Coefficients: Estimate  Std.Error   t value   Pr(>|t|)
(Intercept)     2.444       1.634      1.496     0.173
Residual standard error: 4.902 on 8 degrees of freedom
```

IC(μ, 95%) = **2.444** $\pm$ $t_{8,0.975}$ **1.634** = **[-1.32 , 6.21]**

L'estadístic: (**2.444**-0)/1.634 = **1.496** amb **p-value** P(|t_8|>1.496)= pt(-1.496,8)+(1-pt(1.496,8)) = **0.173**

Les premisses que s'assumeixen són mostra aleatòria i Normalitat de X

- Els resultats: **estimació puntual** de 2.44 unitats per a la mitjana de la variable «error en les mesures». Amb **confiança** del 95% la mitjana poblacional estaria entre -1.32 i 6.21 (rang de valors versemblants).

  p-valor associat a PH μ=0 dóna 0.173 (gran), per tant no seria una hipòtesi *sense cap ni peus,* però no és demostra res

- La part residual que el model no recull s'indica amb la **desviació dels residus** (4.902), o desviació tipus de la diferència en la mostra entre el valor observat i el valor de mitjana estimada

Model per estimar μ. Cas de μ_D en mostres aparellades

Notació del model: $D_i = \mu_D + \varepsilon_i$ (amb D=Y1-Y2)

Funcions de R: `lm(D~1)` i `summary(lm(D~1))`

(la constant **1** a la dreta de **~** és la sintaxi que usa R per indicar que només volem estimar el paràmetre **μ**)

R proporciona: (com el cas anterior per μ d'una mostra)

- l'estimació puntual de μ ($\hat{\mu}_D = \bar{d}$). R ho indica com a **"Estimate"** de l'**intercept**
- l'estimació de l'error tipus (se) assumint m.a. de l'estimador anterior

  R ho indica com a **"Std. Error"** (*se*) de l'**intercept** ($\bar{d}$), que permetria construir l'IC per a μ_D

$$IC(1-\alpha\%, \mu_D) = \bar{d} \pm t_{n-1,1-\alpha/2} \cdot se$$

- El valor de l'estadístic (**"t value"**) i el **p-value** (**"Pr(>|t|)"**) per avaluar la versemblança d'un possible valor concret del paràmetre (per defecte el valor 0)
- la desviació residual (**"Residual Standard error"**) o desviació de la part aleatòria que no recull el model

Anàlisi dels resultats

El resultat bàsic és l'estimació puntual (i per IC) de la mitjana poblacional de la diferència

En aquest cas, és interessant avaluar la versemblança del valor 0 per a la diferència de mitjanes, indicant que no es diferencien i que representen mostres aparellades amb comportament mitjà equivalent.

La desviació residual és l'indicador de la part residual que el model no recull

Model per estimar μ. Cas de μ_D en mostres aparellades

Validació de les premisses

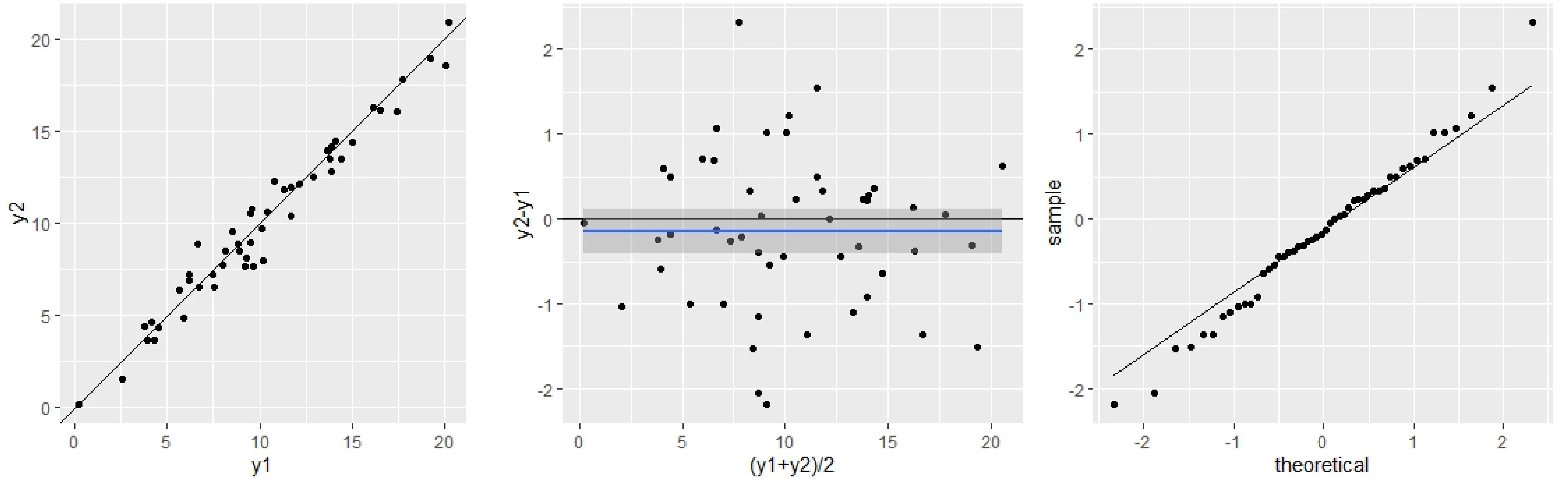
Les **premisses** són: **mostra aleatòria** i **normalitat de la diferència**
i **diferència constant entre parelles (efecte additiu)**

- La premissa de **mostra aleatòria** (m.a.) **no es pot verificar** (com en el cas anterior) (únicament verificar la independència respecte a l'ordre de recollida de les dades)
- La premissa de **normalitat** l'avaluarem gràficament (com en el cas anterior, amb QQ-plot)
- En una comparació de mitjanes s'assumeix un **efecte additiu** (un algorisme redueix el temps d'execució en 2 segons) però a vegades es té un **efecte multiplicatiu** (un algorisme redueix a la meitat els temps d'execució)
  - El **gràfic de Bland-Altman** (introduït al final del bloc C) representa les diferències de les respostes per cada individu en funció de les seves mitjanes. Permet estudiar si hi ha o no un efecte additiu o multiplicatiu, i decidir si convindria una transformació a les dades
  - **Veurem 3 situacions especials:**
    Cas 1: sense efecte lineal (additiu)
    Cas 2: amb efecte multiplicatiu
    Cas 3: amb efecte multiplicatiu i transformació logarítmica: $(\ln Y1, \ln Y2)$

Model per estimar μ. Cas de μ_D en mostres aparellades

Validació de les premisses. Cas 1: sense efecte lineal (additiu)

Exemple de 50 observacions aparellades (D = Y1-Y2)



Cas amb diferència de mitjanes puntual estimada de -0.146

L'IC per la diferència de mitjanes és $IC_{95\%}(\mu_{Y2}-\mu_{Y1})$ = [-0.40 a 0.11]

És a dir, $Y_2 = Y_1 - 0.146$ en mitjana, o bé $Y_2 = Y_1 + [-0.40 \text{ a } 0.11]$

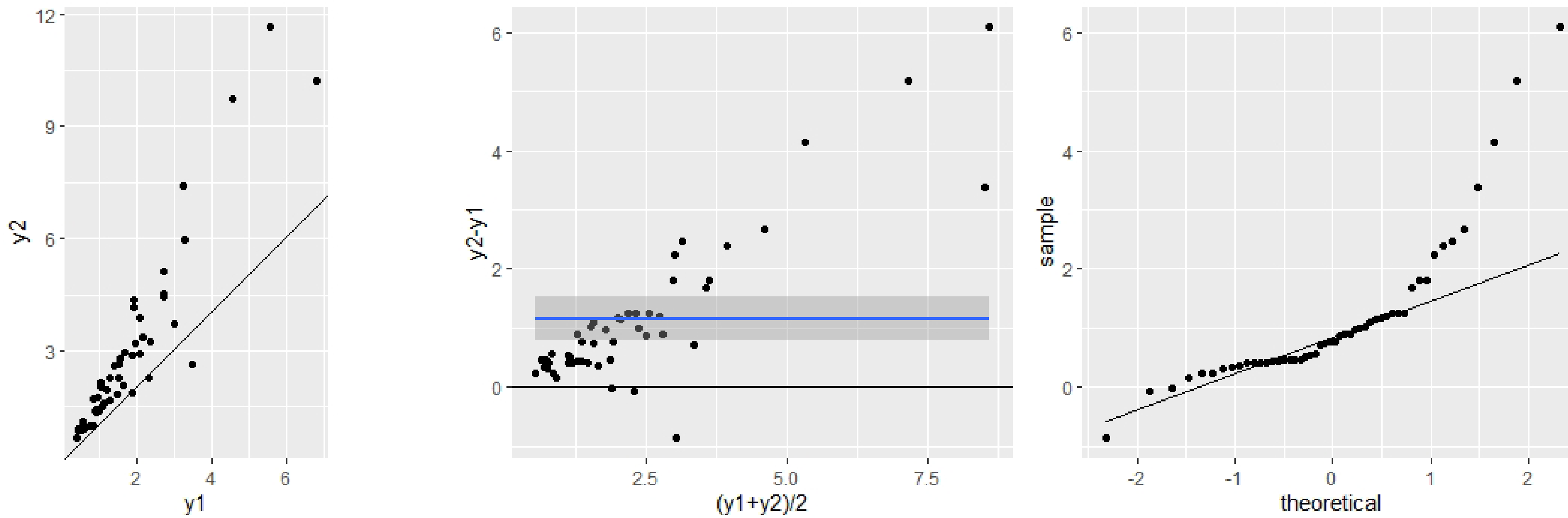
Per tant, **no hi ha evidència de què ambdues mitjanes siguin diferents** (el valor 0 per a la diferència de mitjanes és versemblant)

Es pot assumir Normalitat de la variable diferència ja que tots els quantils observats s'ajusten força bé als quantils teòrics de la Normal.

Model per estimar μ. Cas de μ_D en mostres aparellades

Validació de les premisses. Cas 2: amb efecte multiplicatiu

Nou exemple de 50 observacions aparellades



La diferència de mitjanes estimada és 1.16 amb un $IC_{95\%}(\mu_{Y2}-\mu_{Y1})$ = [0.38 a 1.92]
Però aquest valor no ens informa bé, ja que **l'efecte no és constant**

Per a valors grans, l'efecte és més gran i té mes variabilitat

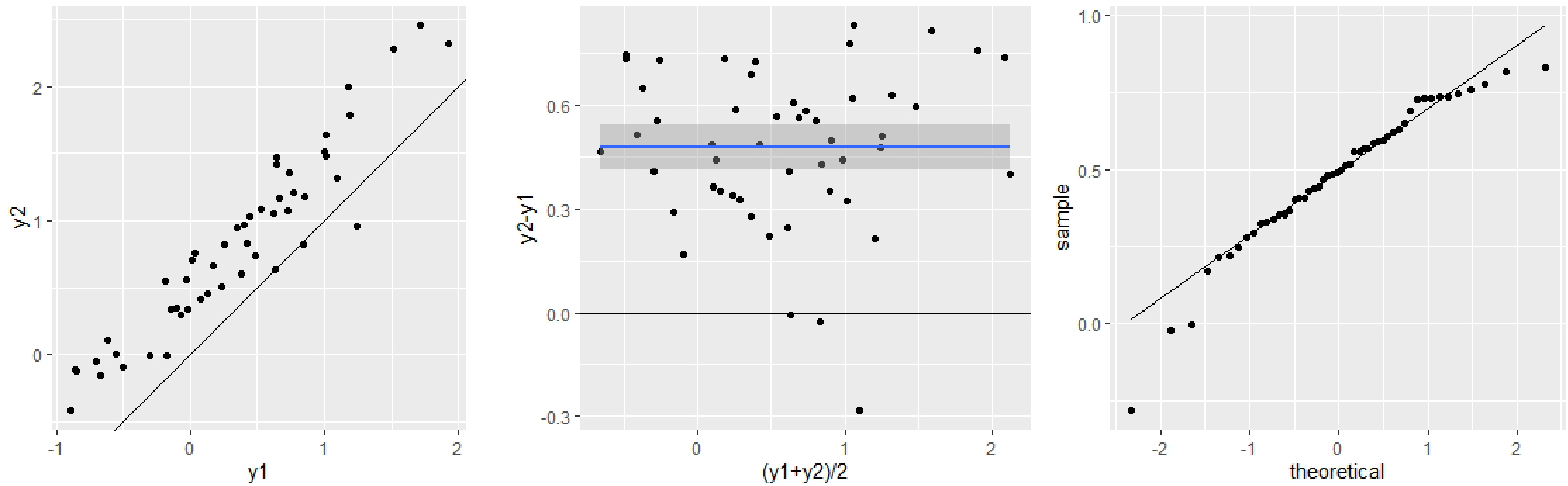
Una transformació sobre les variables que solucioni aquests problemes, farà la interpretació més fàcil → Provarem fent la transformació logarítmica (natural)

La distribució de les diferències NO és Normal

Model per estimar μ. Cas de μ_D en mostres aparellades

Validació de les premisses. Cas 3: amb efecte multiplicatiu i transformació log

Aplicar logaritmes a cada variable del exemple anterior de 50 observacions aparellades



La diferència mitjana estimada dels logaritmes és 0.48 amb $IC_{95\%}(\mu_{Y2'}-\mu_{Y1'})$ de 0.21 a 0.75

Si Y1,Y2 són les variables originals i Y1',Y2' són les log-transformades:

$$\left.\begin{array}{l} Y2' = \ln(Y2) \\ Y1' = \ln(Y1) \end{array}\right\} \rightarrow \quad Y2' = Y1' + 0.48 \quad (en\ mitjana)$$

$$\rightarrow \ln(Y2) = \ln(Y1) + 0.48 \ \rightarrow Y2 = e^{\ln(Y1)+0.48} \rightarrow Y2 = e^{\ln(Y1)} \cdot e^{0.48} = 1.62 \cdot Y1$$

Interpretació: Y2 és en mitjana 1.62 ($IC_{95\%}$ de 1.23 a 2.12) vegades més gran que Y1

(ja que exp(0.21)= $e^{0.21}$ =1.23 i exp(0.75)= $e^{0.75}$ =2.12)

Es pot assumir Normalitat de la variable diferència de logaritmes (= logaritme del quocient) ja que tots els quantils observats s'ajusten prou bé als quantils teòrics de la Normal.

Al desfer la transformació logarítmica, les estimacions obtingudes infra-estimen el valor real

Model per estimar μ. Cas de μ_D en mostres aparellades

Exemple de 50 observacions aparellades (D = Y2-Y1)

Solució per Bloc C (IC)
 	
`>t.test(Y2,Y1,paired=T)`
	`t = -1.1434 , df = 49, p-value = 0.2584`
	`alternative hypothesis: true mean is not equal to 0`
	`95 percent confidence interval:`
	`-0.4037   0.1109`
	`sample estimates: mean of x  -0.146`

Solució per Bloc D (model estadístic)
 	
`> (mod <- lm(D~1))`
`> summary(mod)`
	`Coefficients:`
	`             Estimate   Std. Error   t value   Pr(>|t|)`
	`(Intercept)   -0.146      0.128      -1.143     0.258`
	`Residual standard error: 0.9054 on 49 degrees of freedom`

IC(μ, 95%) = $-0.146 \pm t_{49,0.975} \cdot 0.128$ = **[-0.40 , 0.11]**

L'estadístic: (-**0.146**-0)/0.128 = -**1.143** amb **p-valor** $P(|t_{49}|>1.143)$= pt(-1.143,49)+(1-pt(1.143,49)) = **0.258**

Les premisses que s'assumeixen són mostra aleatòria i Normalitat de Y_2-Y_1 i diferència constant

- Els resultats indiquen una **estimació puntual** de -0.146 com a mitjana esperada de la diferència, i amb **confiança del 95%** la diferència estarà entre -0.40 i 0.11
- El p-valor "gran" indica que el valor 0 és versemblant, per tant no hi ha evidència per dubtar que la diferència mitjana sigui nul·la i que les mitjanes en les dues mostres siguin equivalents (l'atzar pot explicar les discrepàncies observades en les diferències)
- La part no recollida pel model és la **desviació dels residus** o "Residual Standard error" (0.9054)

Models estadístics
Model (b) i funcions de R
(model comparant el paràmetre μ en diverses mostres independents)

Model comparant el paràmetre μ en mostres independents

Notació del model: $Y_i = \mu + \vartheta_k + \varepsilon_i$

El model contempla com a paràmetres: μ com a mitjana de referència i ϑ_k com a canvi de la mitjana del grup *k* (per tant, atenció!, no contempla les diverses μ_k)

Funcions de R: `lm(Y~G)` i `summary(lm(Y~G))`

(G és una columna amb caràcters identificant els grups o factors; si és numèrica cal usar **as.factor(G)**)

R proporciona:

- l'estimació puntual de la **μ de referència** (R ho indica com a **"Estimate"** de l'**intercept**)
  i l'estimació del **canvi en les mitjanes** entre grups (indicat amb **"Estimate"** per a cada canvi de grup de G)
- l'estimació de l'error tipus (se) per a cada estimació anterior (indicat com a **"Std. Error"**). Permet calcular IC
- l'estadístic (**"t value"**) amb el **p-value** (**"Pr(>|t|)"**) per a cada estimació i que permet avaluar la versemblança d'un possible valor concret del paràmetre (per defecte el valor 0)
- la desviació residual (**"Residual Standard error"**) o desviació de la part aleatòria que no recull el model

Altres funcions en R (veure més endavant en un exemple) permeten obtenir resultats complementaris:

```
confint(lm(Y~G))              # Interval de confiança
library(emmeans)              # Llibreria per calcular mitjanes
emmeans(lm(Y~G),~G)           # Mitjana per cada grup
plot(emmeans(lm(Y~G),~G))     # Es requereix library(ggplot2)
pairs(emmeans(lm(Y~G),~G))    # Fa comparacions 2 a 2
```

Model comparant el paràmetre μ en mostres independents

Validació de les premisses

Les **premisses** són: **dues (*k*) mostres independents**, per disseny,

mostra aleatòria i **normalitat dins de cada grup**

i **homoscedasticitat entre grups** (variabilitat semblant entre els grups)

- La premissa de **mostra aleatòria** (m.a.) **no es pot verificar** (com en el casos anteriors) (bàsicament verificar la independència respecte a l'ordre de recollida de les dades)
- La premissa de **normalitat** l'avaluarem gràficament dins de cada grup (amb QQ-plot)

- La premissa d'**homoscedasticitat** fa referència a igualtat de variàncies (igualtat dels paràmetres de variància desconeguts)

  La comprovació d'homocedascititat es pot fer generalment de forma gràfica (boxplot múltiple, per obtenir una escala comuna).

Model comparant el paràmetre μ en mostres independents

R^2: Indicador de variabilitat explicada pels grups

- En aquest model $\mathbf{Y_i = \mu} + \vartheta_k + \varepsilon_i$ podem **descomposar la variabilitat**

La variabilitat **total** de la resposta té una part **explicada** (entre els grups o factors) i una part no explicada (o **residual)**

En particular, els diferents tipus de variabilitat (*sumes de quadrats*) es defineixen d'aquesta manera:

Variabilitat total (T): $SS_T = \sum_j \sum_i \left(y_{i,j} - \bar{y}\right)^2$

Variabilitat residual (R) o dins dels grups: $SS_R = \sum_j \sum_i \left(y_{i,j} - \bar{y}_j\right)^2$

Variabilitat entre (E) els grups: $SS_E = \sum_j n_j \left(\bar{y}_j - \bar{y}\right)^2$

- S'anomena **coeficient de determinació** o $\mathbf{R^2}$ al rati entre la variabilitat explicada i la total

$$R^2 = \frac{variabilitat\ explicada}{variabilitat\ total} = \frac{\text{"}suma\ de\ diferències\ entre\ grups, al\ quadrat\text{"}}{\text{"}suma\ de\ diferències\ totals, al\ quadrat\text{"}} = \frac{SS_E}{SS_T}$$

No requereix m.a.
ni altres premisses

Com més gran és el valor de R^2, millor representa el model la relació entre les variables:

R^2 és màxim (1=100%) si la relació és perfectament determinista, la part residual és 0, les mitjanes de cada grup són diferents i dins cada grup no hi ha variació, tota la variació és entre grups
R^2 és mínim (0) si el factor no determina res de la variació de Y, les mitjanes de cada grup són idèntiques, no hi ha variació entre grups, tota la variació és dins els grups

- R es refereix a aquest indicador com **"Multiple R-squared"** (i afegeix una variant, "Adjusted R-squared")

A la referència a bibliografia (Estadística per a enginyers informàtics) hi ha més detalls al capítol 6.6. I també veure *app*

Model comparant el paràmetre μ en mostres independents

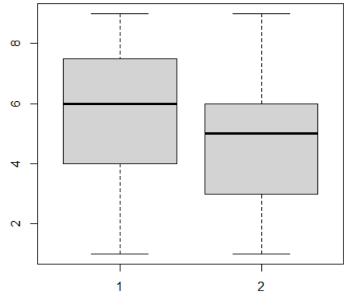
Exemple de 2 mostres independents per comparar μ_1 i μ_2

Solució per Bloc C (IC de la diferència)

```
Y1 <-c(1,2,3,5,6,6,7,7,8,8,9)         # mean=5.6  sd=2.62
Y2 <-c(1,1,3,4,5,5,6,7,9)             # mean=4.5  sd=2.65
Y <- c(1,2,3,5,6,6,7,7,8,8,9,1,1,3…)  # 2 mostres juntes
G <- c(1,1,1,1,1,1,1,1,1,1,1,2,2,2…)  # a lm() cal usar as.factor(G)
# mean(Y1)-mean(Y2) -> 1.080808       # o bé 5.636364- 4.555556 = 1.080808
```

Estudiem la normalitat de Y1 I Y2 amb qqnorm(Y1) i qqline(Y1) i amb qqnorm(Y2) i qqline(Y2)

boxplot(X1,X2)




```
t.test(Y1,Y2,var.equal=T)  # o t.test(Y~as.factor(G),var.equal=T)
```

```
t = 0.91335, df = 18, p-value = 0.3731
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   -1.405312  3.566928
  sample estimates: mean of x mean of y  5.636364  4.555556
```

Les premisses que s'assumeixen són mostra aleatòria, Normalitat de Y1 i Y2, i homoscedasticitat (variabilitat semblant en els dos grups comprovada gràficament al boxplot)

- Els resultats indiquen que encara que l'**estimació puntual** de la diferència de mitjanes és de 1.08, el rati senyal/soroll és 0.91 (amb un p-valor "gran") i per tant la comparació no arriba a cap conclusió "qualitativa" (*alguna μ és superior?*)
- Les dades són compatibles amb una diferència de mitjanes poblacionals des de -1.40 fins a 3.57 amb una **confiança del 95%.**
- La igualtat de mitjanes no es pot descartar (però no hi ha cap evidència a favor)

Model comparant el paràmetre μ en mostres independents

Exemple de 2 mostres independents per comparar μ_1 i μ_2

Solució per Bloc D (model estadístic amb IC de la μ de referència i de la diferència)

els estimadors principals ara són una de les mitjanes 5.6364 com a referència, i el canvi -1.08 (5.636 - 1.08 → 4.55)

```
lm(Y~as.factor(G))
summary(lm(Y~as.factor(G)))
```

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   5.6364      0.7938      7.100     1.28e-06 ***
G2           -1.0808      1.1833     -0.913     0.373

Residual standard error: 2.633 on 18 degrees of freedom
Multiple R-squared:  0.04429,   Adjusted R-squared:  -0.008803
F-statistic: 0.8342 on 1 and 18 DF,  p-value: 0.3731
```

IC (intercept) $(5.6364 \pm t_{18,0.975}\, 0.7938 = [3.97, 7.30]$ amb estadístic: $(5.6364-0) / 0.7938 = 7.1$ (p-valor “petit”)

IC ϑ_k (G2) $(-1.0808 \pm t_{18,0.975}\, 1.1833 = [-3.567, 1.405]$ amb estadístic: $(-1.0808-0) / 1.1833 = -0.913$ (p-valor “gran”)

- En aquests resultats també veiem que l’**estimació puntual** de la diferència de mitjanes és de -1.08
  - El rati senyal/soroll del terme constant (*intercept*) és 7.1 (amb un p-valor “petit”) i per tant la mitjana de la categoria de referència (*G1*) no és versemblant que sigui zero. I el rati senyal/soroll de la diferència de mitjanes és -0.9 (amb un p-valor “gran”) i per tant **la comparació entre G1 i G2 no és concloent** (el signe de la diferència podria ser + o -)
- La part residual que el model no recull és **2.633** (“Residual Standard error”, estimació conjunta de la desviació estàndard intragrup, o desviació de la diferència en la mostra entre l’observat i la predicció).
- **R^2** és 0.04429: els **grups només expliquen** un 4.4% de la variabilitat total

Model comparant el paràmetre μ en mostres independents

Exemple de 2 mostres independents per comparar μ_1 i μ_2

Solució per Bloc D (model estadístic amb resultats de funcions complementàries)

```
confint(lm(Y~as.factor(G)))
```

	2.5 %	97.5 %
(Intercept)	3.968623	7.304104
G2	-3.566928	1.405312

(IC dels paràmetres μ_1 i ϑ_2 del model, també calculats abans)

```
emmeans(lm(Y~as.factor(G),~G))
```

G	emmean	SE	df	lower.CL	upper.CL
1	5.64	0.794	18	3.97	7.3
2	4.56	0.878	18	2.71	6.4

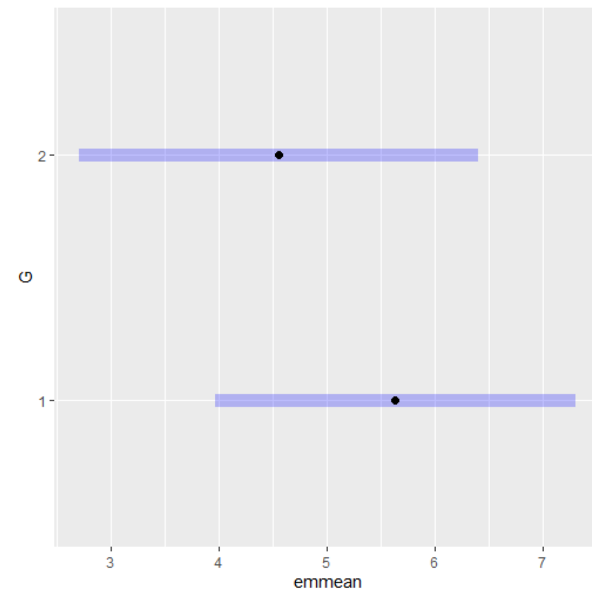
(Estimació i SE permeten calcular IC per la μ_1 i μ_2 dels dos grups)

```
pairs(emmeans(lm(Y~as.factor(G),~G))
```

contrast	estimate	SE	df	t.ratio	p.value
G1 - G2	1.08	1.18	18	0.913	0.3731

(Estimació i SE permeten calcular IC del canvi ϑ_2 entre els dos grups)

```
plot(emmeans(lm(Y~as.factor(G)),~G))
```



- En aquests resultats veiem, en diversos formats, els mateixos IC presentats en els resultats anteriors
- El gràfic representa els dos IC de les respectives μ_1 i μ_2 amb l'estimació puntual com un punt negre i els IC en blau. Així es poden comparar i veure si tenen solapament (és versemblant que vinguin d'un mateix valor de μ) o no (no seria versemblant que μ_1 i μ_2 coincideixin)

Un disseny balancejat (és a dir, si els k grups tenen igual nombre d'observacions) és més eficient en el sentit que proporciona un SE més baix que un no balancejat. Un SE menor proporciona IC més precisos.

Models estadístics
Model (c) i funcions en R
(model lineal simple i múltiple)

Model lineal simple ... i múltiple

Notació del model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (O, si fos múltiple: $Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \cdots + \beta_k Xk_i + \varepsilon_i$)

El model contempla com a paràmetres els coeficients β_i (de l'equació d'una recta)

Els coeficients de la recta estimada en el cas **simple** $Y = b_0 + b_1 X$ es calculen per mínims quadrats obtenint

$b_1 = r_{XY} \cdot \frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2}$ $b_0 = \bar{Y} - b_1 \bar{X}$ i variància residual $s^2 = \frac{\sum(e_i^2)}{n-2} = \frac{(n-1)S_Y^2(1-r^2)}{n-2}$ (r_{XY} és la correlació i S_{XY} la covariància)

En el cas del model lineal simple b_0 és la constant o ordenada a l'origen (**intercept**), i b_1 el pendent indicant el canvi en la resposta associat a un increment en una unitat en la variable explicativa X . App.

Les variables X_i poden ser tant quantitatives com qualitatives, i les anomenem també predictores

Funcions de R: `lm(Y~X)` i `summary(lm(Y~X))`

(o bé pel cas múltiple: `lm(Y~X1+X2+...)` i `summary(lm(Y~X1+X2+...))`)

R proporciona per a tots els coeficients de la recta ajustada obtenim

- l'estimació dels paràmetres (**"Estimate"**) incloent la b_0 (o *intercept*) com a resposta base per valors nuls o de referència de les X, i les b_i com a pendent o efecte de grup
- l'estimació de l'error tipus per a cada estimació anterior (**"Std. Error"**). Permet calcular IC
- estadístic senyal/soroll (**"t value"**) amb el **p-value** (**"Pr(>|t|)"**) per a cada estimació i que permet avaluar la versemblança d'un possible valor concret del paràmetre (per defecte el valor 0)
- la variància o desviació residual (**"Residual Standard error"**) que no recull el model

Model lineal simple

R^2: Indicador de capacitat predictiva

- En el model lineal simple i múltiple (tal com hem vist en el model de comparar mitjanes en mostres independents) podem **descomposar la variabilitat**

La variabilitat **total** de la resposta Y té una part **explicada** (variacions en algun predictor X suposa canvis en la Y) i una part no explicada, restant o **residual**

En particular, així es fa la descomposició de la variabilitat

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

$$SQ_{Total} = SQ_{Residual} + SQ_{Explicada}$$

- Anomenem **coeficient de determinació** o **R^2** al rati entre la variabilitat explicada pel model i la variabilitat total de la resposta

$$R^2 = \frac{variabilitat\ explicada}{variabilitat\ total} = \frac{SQ_E}{SQ_T}$$

No requereix m.a.
ni altres premisses

Com més gran és el valor de R^2, millor representa el model la relació entre les variables

R^2 és màxim (1 = 100%) si la relació és perfectament determinista, la part residual és zero
R^2 és mínim (0) si el model no determina res de la variació de Y que prové de la part aleatòria i no de les X

En regressió lineal simple, $R^2 = (r_{XY})^2$, és a dir, equival al quadrat del coeficient de correlació lineal
R^2 és indicador de "bondat" de l'ajust o **capacitat predictiva** i r_{XY} ho és de l'associació de les variables

- R es refereix a aquest indicador com **"Multiple R-squared"** (i afegeix una variant, "Adjusted R-squared")

A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més detalls al capítol 6.6

Model lineal simple i múltiple. Exemple

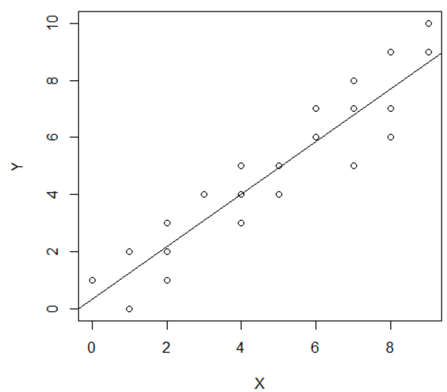
Exemple de mostra de dues variables amb relació lineal

```
Y <-c(1,0,1,3,4,5,4,6,2,6,5,2,9,7,7,5,6,9,3,6,8,7,10,5,4,6)
X <-c(0,1,2,2,3,4,4,6,1,8,7,2,9,6,8,5,6,8,4,6,7,7,9,7,5,6)
cor(X,Y) # = 0.9239073
lm(Y~X)
summary(lm(Y~X))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.7710 -0.8719  0.1483  0.8054  1.3903
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.33581    0.44460   0.755    0.457
X            0.91931    0.07771  11.830 1.68e-11 ***

Residual standard error: 1.015 on 24 degrees of freedom
Multiple R-squared:  0.8536,    Adjusted R-squared:  0.8475
F-statistic: 139.9 on 1 and 24 DF,  p-value: 1.68e-11
```

```
plot(X,Y)
abline(lm(Y~X))
```



- El model recull un **85.36 % de la variabilitat total** de la resposta. La resta és **residual** amb desviació tipus 1.015
- L'**ordenada a l'origen** (**o terme independent**) és 0.33581 amb IC 0.33581 $\pm$ qt(0.975,24)*0.4446 → **[-0.58,1.25]**
- El **pendent** (**o terme lineal**) és 0.91931 amb IC 0.91931 $\pm$ qt(0.975,24)*0.07771 → **[0.76,1.08]**
  Per a cada unitat més a X, el pendent indica 0.919 unitats més a Y (amb **95% de confiança** entre 0.76 i 1.08)
- 0 no és un valor versemblant per al pendent (no és a l'interval). Però 1 si és versemblant: +1 a X implica +1 a Y
  0 sí és un valor versemblant per a l'ordenada a l'origen (és a l'interval)
  No ho podem confirmar, però una possible recta Y = 0 + 1 X + ε suggereix considerable concordança entre X i Y.

Model lineal simple

Validació de les premisses

Les premisses són:

- **linealitat** (la forma del núvol de punts s'ajusta a una recta o a un pla en el cas múltiple)
- **mostra aleatòria** (implica independència dels residus)
- **normalitat dels residus**
- **homoscedasticitat dels residus** (variabilitat homogènia dels residus)

La linealitat fa referència a la part determinista, mentre que les altres fan referència a la part aleatòria o residual (per això el que es valida és la independència, normalitat i homoscedasticitat dels residus):

$$Y_i = \underbrace{\beta_0 + \beta_1 X1_i + \cdots \beta_k Xk_i}_{\text{Linealitat}} + \epsilon_i \rightarrow \epsilon_i \sim \underbrace{N}_{\text{Normalitat}}(\underbrace{0, \sigma}_{\text{Homoscedasticitat}}) \qquad \epsilon_i, \epsilon_j \ \underbrace{ind.}_{\text{Independència}} \forall i, j$$

Ara veurem aquestes premisses en gràfics on podrem comprovar-les

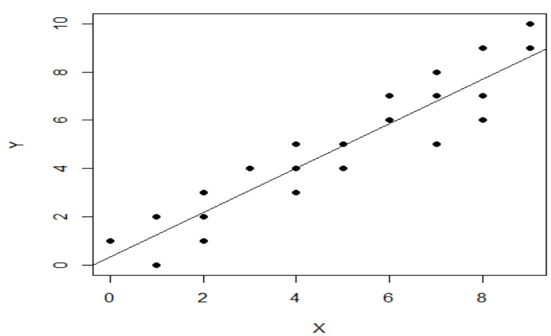
App. A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més detalls al capítol 7

Model lineal simple

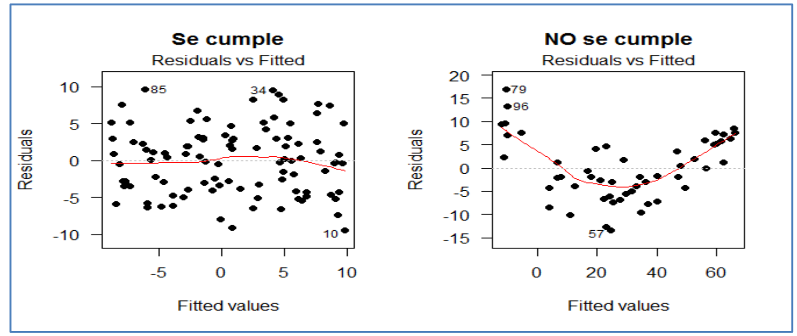
Validació de les premisses

Les **premisses** del model lineal i els gràfics on estudiar-les són:

- **Linealitat** (Y i X s'ajusten a una recta, pel cas simple, o a un pla o hiperplà, pel cas múltiple)

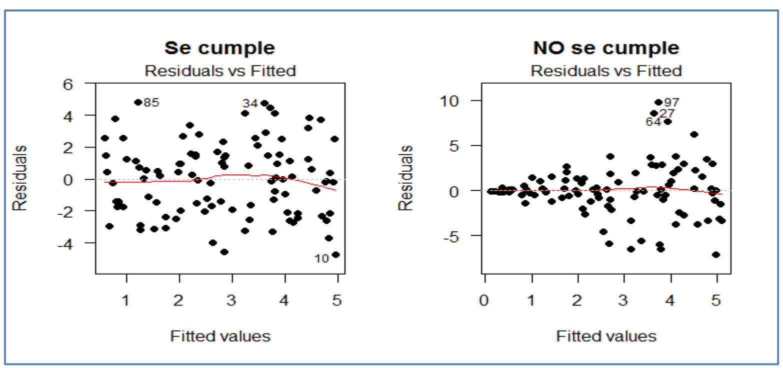


Plot Y i X on veure si s'ajusta a una recta (en aquest cas sí)



Gràfics de residus enfront les prediccions on veure si estan per sota i per sobre del 0 uniformement. Esquerra compleix linealitat; i dreta no

- **Homoscedasticitat** (ε$\sim N(0, \boldsymbol{\sigma})$ o variabilitat constant)

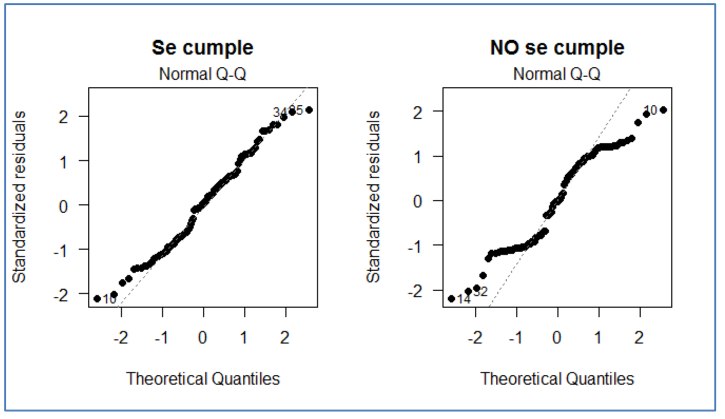


Gràfic dels residus enfront les prediccions on veure que es distancien del zero de la mateixa forma, sense zones amb més i menys dispersió.
En aquests gràfics:
Esquerra compleix homoscedasticitat; i dreta no

Model lineal simple. Validació

Validació de les premisses

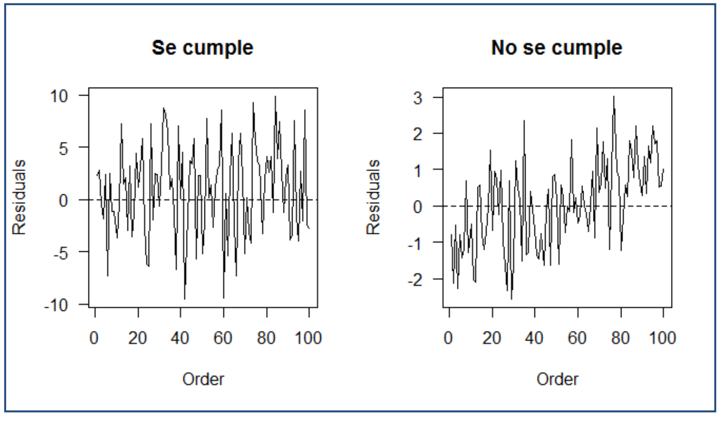
- **Normalitat** ($\varepsilon \sim N(0, \sigma)$ residus amb distribució Normal)



qqnorm i qqline(on veure si els residus s'ajusten al model Normal

En aquests gràfics un compleix i l'altre no

- **Independència** (mostra aleatòria i no dependència entre observacions)



La independència es garanteix amb un bon disseny i recollida de dades
Però observant els residus enfront l'ordre de recollida es pot veure si hi ha alguna dependència. S'espera no trobar cap patró específic

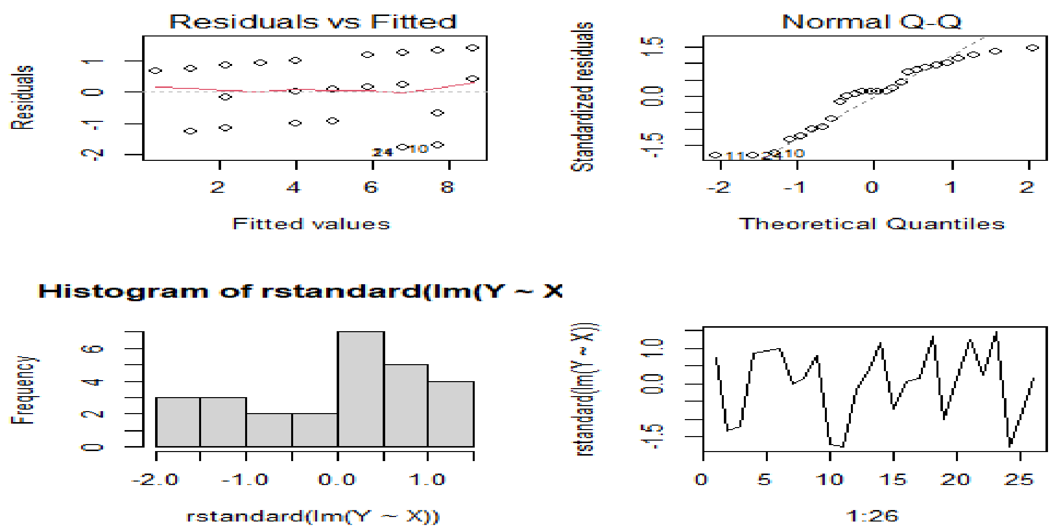
En aquest cas un compleix i l'altre no, perquè hi ha una tendència creixent dels residus al llarg del temps

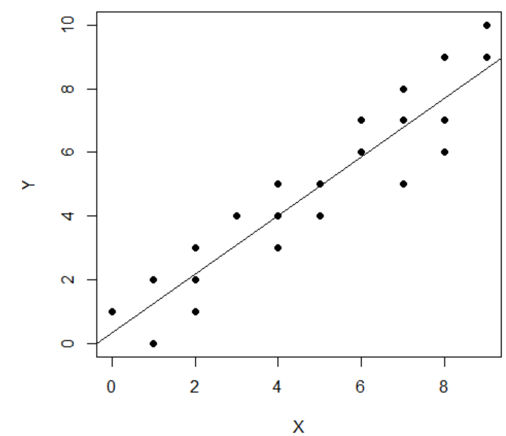
Model lineal simple i múltiple

Exemple de mostra de dues variables amb relació lineal

```
Y <-c(1,0,1,3,4,5,4,6,2,6,5,2,9,7,7,5,6,9,3,6,8,7,10,5,4,6)
X <-c(0,1,2,2,3,4,4,6,1,8,7,2,9,6,8,5,6,8,4,6,7,7,9,7,5,6)
cor(X,Y) # = 0.9239073
lm(Y~X)
par(mfrow=c(2,2))
plot(lm(Y~X),c(2,1)) # o bé plot(model,c(2,3))
hist(rstandard(lm(Y~X)))
plot(rstandard(lm(Y~X)),type="l")
```

```
plot(X,Y)
abline(lm(Y~X))
```





Totes les **premisses** semblen acceptables, tot i que al QQ-plot es pot sospitar que la Normalitat falla (a la vista dels extrems, no hi ha *cues llargues*)

Model lineal simple

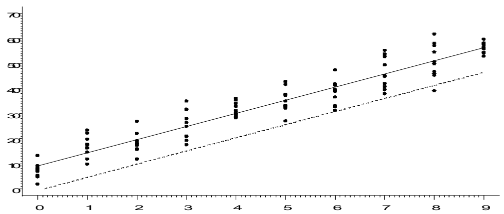
Anàlisi dels resultats (Funcions R per fer prediccions)

- A partir d'una predicció puntual (aplicant la fórmula de l'estimació de la part determinista del model) es pot calcular un interval a la predicció amb dos tipus d'enfocament:
  - **Valor individual**. Estimar el **valor** de la resposta per a **una** observació amb uns valors concrets de les variables predictores [Ex: Quin és el retard *previsible* pel vol BCN-ROM de Vueling de les 8:00?]
    `predict(…, interval = "prediction")`
  - **Valor esperat**. Estimar la **mitjana** de la resposta en **totes** les observacions amb uns valors concrets de les variables predictores [Ex: Quin és el retard *esperat* per als vols BCN-ROM de Vueling que surten a les 8:00?]
    `predict(…, interval = "confidence")`

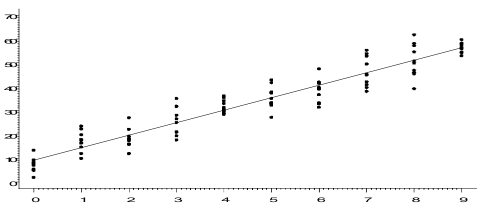
En ambdós casos, **l'estimació puntual de la predicció és la mateixa, però no la seva incertesa.**

En el cas del valor individual tenim un rang més ampli de valors plausibles:

(predicció sobre valors individuals)



(predicció sobre valors esperats)



- R proporciona la funció predict():

```
predict(lm( ))  # predicció puntual a les obs
new <- data.frame([nom_columna]=[valor on fer pred],...)
predict(lm( ),new ) # pred puntual a nova obs
predict(lm( ),new,interval=...) # pred per interval 95%
predict(lm( ),new,interval=...,level=...)
```

Indicant només el model, fa prediccions puntuals a les pròpies observacions. Però és pot indicar:
- valors de noves observacions (new) per predir
- el tipus d'interval a partir de la predicció puntual
- el nivell de confiança per l'interval

A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més detalls al capítol 7.2

Model lineal múltiple. Exemple

Amb variables explicatives quantitatives i qualitatives

Un model estadístic pot combinar els models (b) i (c), però la inclusió de diferents variables explicatives suposa dificultats addicionals; fonamentalment, si no són independents entre elles.

El forat de gènere (*gender pay gap*) es refereix a la diferència de sou existent entre un treballador home i una treballadora dona. Una empresa de consultoria estudia en una mostra de 30 homes i 20 dones les relacions entre salari amb experiència, edat i sexe

Y: salari anual en milers eur (és la variable resposta)

G: gènere ("m", home, "w", dona)

Xp: nombre d'anys de vida laboral

Age: edat

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	44.0387	10.9254	4.031	0.000207
xp	1.1575	0.5638	2.053	0.045789
age	-0.3088	0.4683	-0.659	0.512878
Gw	-5.2430	0.8572	-6.116	1.94e-07

```
Residual standard error: 2.663 on 46 degr of freedom
Multiple R-squared: 0.7647, Adjusted R-squared:  0.7494
```

- A "Estimate" hi ha l'**estimació puntual** del pendent per a les quantitatives, o de l'efecte de grup per a les qualitatives
- La constant ("*intercept"*) ara es refereix a l'estimació de la resposta mitjana per a les categories de referència (aquí "home"), i per al valor 0 de les variables quantitatives (encara que *edat=0* no és interpretable)
- Els "estimates" de les variables suposen el canvi (amb un increment 1 si és una quantitativa) respecte els valors de referència, *suposant que les altres covariables no canvien* (difícil si estan correlacionades)
- En general, no podem aplicar dissenys experimentals per a obtenir les dades independents.

Model lineal múltiple. Exemple

Amb variables explicatives quantitatives i qualitatives

El forat de gènere (*gender pay gap*) es refereix a la diferència de sou existent entre un treballador home i una treballadora dona. Una empresa de consultoria estudia en una mostra de 30 homes i 20 dones les relacions entre salari amb experiència, edat i sexe

Y: salari anual en milers eur
(és la variable resposta)

G: gènere ("m", home, "w", dona)

Xp: nombre d'anys de vida laboral

Age: edat

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	44.0387	10.9254	4.031	0.000207
xp	1.1575	0.5638	2.053	0.045789
age	-0.3088	0.4683	-0.659	0.512878
Gw	-5.2430	0.8572	-6.116	1.94e-07

Residual standard error: 2.663 on 46 degr of freedom
Multiple R-squared: 0.7647, Adjusted R-squared: 0.7494

- La constant (*"intercept")*, l'estimació de la resposta mitjana per a les categories de referència (aquí "home"), i per al valor 0 de les variables quantitatives: una base de **44039 €**, i un standard error de **10925€**
- Els coeficients ajustats per a cada variable (amb les altres fixades) s'interpreten com:
  - cada any més de vida laboral "representa"* un increment de salari de **1157.5 €**
  - cada any més d'edat, **308.8 €** menys Noteu el gran SE (468 €) de l'estimació, una conseqüència de la alta correlació entre edat i experiència (i que el fa impossible de valorar-ho)
  - si es tracta d'una dona, el salari mitjà es redueix en **5243 €**
  - no tenim les dades però si el gènere no fos independent de l'experiència les dues estimacions no serien separables, s'estarien *confonent*.
  - El model serveix per predir *globalment*, no es pot trobar una contribució *particular* per cada variable explicativa.

* "representa" no en el sentit de "causa" sinó en el sentit de que "s'associa amb"

Altres tècniques
(amb funcions en R)

Altres tècniques

Two types of models (L. Breiman, Two cultures. 2001)

Models for Explanation (theory driven).

- Models for understanding the true generative mechanism of data (this implies that we look for the causal relationships of the response).
- Imply interpretable and parsimonious models. This leads using parametric (statistical) models. Error follows a probability distribution.
- We are interested in significance. We focus on global measures of fit and significance of coefficients (we use p.values).
- Prediction may allow forecasting the future in presence of change (or intervention policies).
- Modeling needs expert domain.

Models for Prediction (data driven).

- Models are considered mere algorithms.
- The model is considered a black box (complexity is not a problem).
- Correlated predictors are enough.
- We are just interested in the accuracy of predictions (significance is uninteresting).
- We focus on the Generalization Error of the model. Error should be made minimal (we don't care on the residual analysis).
- Predictions fail in presence of change.
- Accuracy and interpretability are in conflict.

Descriptive Analytics
discovering patterns in data

- Profiling
- Principal Component Analyisis
- Correspodence Analyisis
- Clustering
- Association Rules
- ...

Predictive Analytics
Setting models for prediction

- Multiple Regression
- Logistic Regression (GLM)
- Neural networks
- Decision trees
- Knn neighbors
- PCR & PLSR
- Discriminant Analysis
- SVM
- Deep Learning

Altres tècniques

Forces tècniques de mineria de dades, ciència de dades o intel·ligència artificial presenten algunes característiques que les distingeixen de les tècniques estadístiques més clàssiques:

- habitualment són observacionals i no experimentals
- acostumen a usar dades massives i sense un disseny previ, que fa dubtar de l'aleatorietat
- usen indicadors de capacitat predictiva específics de cada tècnica (tot i que molts cops són indicadors de % de variabilitat explicada com R^2 en estadística més clàssica)
- majoritàriament impliquen models no paramètrics i molt gràfics
- moltes tècniques són algorismes iteratius més que de recerca d'un model. Són computacionalment exigents però cada vegada més poden treballar amb quantitats enormes de dades

Funcions de R:

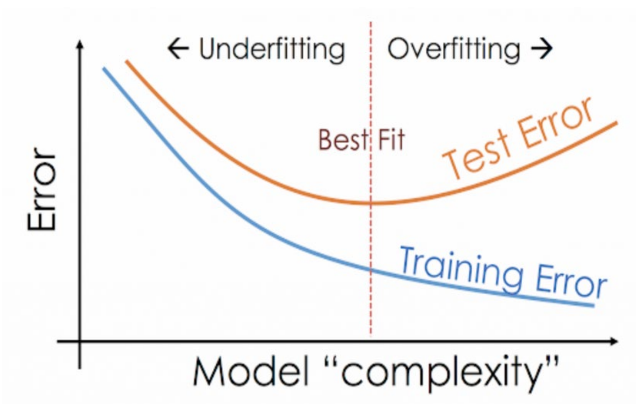
- Tècniques de visualització multidimensional **`PCA()`**
- Tècniques d'agrupament o clustering **`HCPC()`** **`kmeans()`**

(al final del fitxer d'exercicis d'aquest bloc D hi ha explicacions i alguns exemples d'aquestes tècniques)

Hi ha moltes altres tècniques (**`randomForest(),neuralnet(),...`**) amb moltes variants. Cal, però, en totes elles controlar la capacitat predictiva i possibles problemes de sobre-ajustament que veurem a la següent transparència

Capacitat predictiva i sobre-ajustament

- La disponibilitat cada cop més habitual de grans quantitats de dades (on aplicar tècniques de ciència de dades però també model lineal, sobretot múltiple) permet avaluar la **capacitat predictiva** dividint la mostra de dades en un conjunt d'**entrenament** i un conjunt de **test** (per ex separant aleatòriament un 70% i 30% respectivament)
  - El conjunt d'**entrenament** s'utilitza per **ajustar** el model
  - El conjunt de **test** es reserva per **avaluar** el model en dades diferents a les d'entrenament (l'objectiu és avaluar el model en unes dades independents de les usades per crear-lo)
  - La capacitat predictiva es pot avaluar mesurant un "**error**" (minimitzant d'alguna manera les diferències entre els valors observats i els predits)
- El **sobre-ajustament (overfitting)** en un model es produeix quan el model s'ajusta molt a les dades d'entrenament, però té un rendiment deficient en predir noves dades:
  - Un model amb sobre-ajustament predirà de forma deficient observacions futures
  - Habitualment passa en models excessivament complexos que no generalitzen bé més enllà de les dades d'entrenament. En el cas del model lineal múltiple pot passar quan s'inclouen masses variables predictores que poden no tenir una relació real amb la variable de resposta



Models més complexos poden anar ajustant més bé (menys error) la mostra d'entrenament i també la de test. Però pot arribar un punt on afegir complexitat al model pot ajustar bé la d'entrenament però després fallen en la test (i per tant en l'aplicació futura).
El millor model serà el que no passa a ser sobre-ajustat per les dades d'entrenament

En les il·lustracions la relació estadística és real, però la interpretació causal no