MODELS LINEALS GENERALITZATS. INTRODUCCIÓ (II)

Grau en Intel·ligència Artificial

Curs 2024-2025

Índex

VALIDACIÓ NUMÈRICA

Models

Tipus de models

A part del model estimat, definim els models següents

- **Model nul**: quan s'exclouen les variables explicatives, deixant només el terme independent (si n'hi ha)
- **Model complet, o saturat**: hi ha tants paràmetres com dades, s'obté un ajust perfecte, els valors predits són les mateixes $\boldsymbol{y}$, anomenarem $\tilde{\theta}_i = q\,(y_i)$.

R

- **Model nul** $\rightarrow$ `glm(y ~ 1 ,...)`
- **Model saturat amb factors** $\rightarrow$ `glm(y ~` $F_1 * \cdots * F_p$ `,...)`

Contrast d'hipòtesi

Contrast d'hipòtesi

El nostre objectiu és trobar un estadístic que ens digui si el nostre model explica quelcom de la variable resposta.

Volem fer el següent **contrast d'hipòtesis**:

$$\begin{cases} H_0 : \text{El model ajusta bé les dades} \\ H_1 : \text{El model no ajusta bé les dades} \end{cases}$$

Estadístics

Veurem que hi ha 2 estadístics que podem emprar per comprovar la validesa del model:

- Estadístic de la **Deviància**
- Estadístic de **Pearson generalitzat**

Deviància: Estadístic escalat

Deviància escalada

És l'estadístic de contrast del test de raó de versemblança per comparar el nostre model amb el model saturat:

$$D^s\left(\boldsymbol{y}, \hat{\mu}, \Phi\right) = 2\left(\ell\left(\boldsymbol{y}, \phi; \boldsymbol{y}\right) - \ell\left(\hat{\mu}, \phi; \boldsymbol{y}\right)\right) =$$
$$= 2\sum_{i=1}^{N}\left(\frac{\tilde{\theta}_i y_i - b\left(\tilde{\theta}_i\right)}{\Phi} - \frac{\hat{\theta}_i y_i - b\left(\hat{\theta}_i\right)}{\Phi}\right)$$

asimptòticament la seva distribució és χ^2_{N-K}

Deviància: Estadístic no escalat

Deviància, o deviància no escalada

És la deviància escalada multiplicada per Φ.

$$D(\mathbf{y}, \hat{\mu}) = \Phi D^s(\mathbf{y}, \hat{\mu}, \Phi) =$$
$$= 2\sum_{i=1}^{N}\left(\left(\tilde{\theta}_i y_i - b\left(\tilde{\theta}_i\right)\right) - \left(\hat{\theta}_i y_i - b\left(\hat{\theta}_i\right)\right)\right)$$

No depèn del paràmetre de dispersió Φ

Deviància: Notes

Notes

- La deviància del nostre model s'anomena ***Residual deviance***, com més petita millor l'ajust.
- La del model nul s'anomena ***Null deviance***, és el valor màxim que pot obtenir la deviància.
- $\frac{\textit{Null deviance} - \textit{Residual deviance}}{\textit{Null deviance}} = 1 - \frac{\textit{Residual deviance}}{\textit{Null deviance}} \in [0, 1]$ es pot considerar un equivalent del coeficient de determinació R^2

Exercici: Comproveu les deviàncies de les següents famílies

Família	**Deviància** $D(\boldsymbol{y}, \hat{\mu})$
Normal	$\sum_{i=1}^{N} (y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^{N} \left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right)$
Binomial *m fix*	$2\sum_{i=1}^{N} \left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m - y_i)\log\left(\frac{m-y_i}{m-\hat{\mu}_i}\right)\right)$
Gamma	$2\sum_{i=1}^{N} \left(-\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i}{\hat{\mu}_i} - 1\right)$
Inversa Gaussiana	$\sum_{i=1}^{N} \frac{(y_i-\hat{\mu}_i)^2}{y_i\hat{\mu}_i^2}$

Estadístic de Pearson generalitzat I

Estadístic de Pearson

L'estadístic de Pearson generalitzat és una altra mesura de la bondat d'ajust del model, es defineix:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - \hat{\mu}_i\right)^2}{V\left(\hat{\mu}_i\right)}$$

Estadístic de Pearson generalitzat II

Distribució

Tant la distribució de la deviància escalada com la del estadístic de Pearson dividit per Φ asimptòticament són χ^2_{N-K}:

- $D^s\left(\boldsymbol{y}, \hat{\mu}, \Phi\right) = \frac{D(\boldsymbol{y}, \hat{\mu}, \Phi)}{\phi} \sim \chi^2_{N-K}$
- $\frac{\chi^2}{\Phi} \sim \chi^2_{N-K}$

En el cas de la normal la distribució és exacta, i coincideixen ambdues estimacions.

Estadístic de Pearson generalitzat III

Exemples d'estadístics de Pearson generalitzats

Família	Estadístic de Pearson gen. χ^2
Normal	$\sum_{i=1}^{N} (y_i - \hat{\mu}_i)^2$
Poisson	$\sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$
Binomial m_i *fixes*	$\sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(1 - \hat{\mu}_i/m_i)}$
Gamma	$\sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2}$
Inversa Gaussiana	$\sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^3}$

Estimació de Φ

Estimació de Φ

- En els casos en que Φ sigui desconeguda l'haurem d'estimar, $\hat{\Phi}$
- $\hat{\Phi}$ tindrà (asimptòticament) una distribució $\boldsymbol{\Phi} \cdot \frac{\chi^2_\nu}{\nu}$ on els graus de llibertat (ν) són:
  $nombre\ d'observacions - nombre\ de\ paràmetres$.

Variància de y_x

Per estimar $Var\left(y_x\right) = \Phi V\left(\mu_x\right)$ necessitem Φ, que sabem que és constant. Tenim dos casos:

- Φ és una constant coneguda (Binomial, Poisson,...). Llavors, $\widehat{Var}\left(y_x\right) = \boldsymbol{\Phi} V\left(\hat{y}_x\right)$
- Φ és una constant desconeguda (Normal, Gamma...) i s'ha d'estimar: $\widehat{Var}\left(y_x\right) = \hat{\boldsymbol{\Phi}} V\left(\hat{y}_x\right)$

Estimació de Φ

Hi ha dificultats pràctiques per estimar Φ per màxima versemblança, però un cop coneixem les $\hat{\beta}$ el podem estimar pel mètode dels moments, tenim dues opcions:

- A partir de la deviància$= D(\boldsymbol{y}, \hat{\mu})$:
  com que $\frac{D(\boldsymbol{y},\hat{\mu})}{\Phi} \sim \chi^2_{N-K} \Rightarrow E\left[\frac{D(\boldsymbol{y},\hat{\mu})}{\Phi}\right] = N - K$ i

$$\hat{\Phi}_{deviance} = \frac{D(\boldsymbol{y}, \hat{\mu})}{N - K}$$

- A partir de l'estadístic de Pearson generalitzat$= \chi^2$:
  com que $\frac{\chi^2}{\Phi} \sim \chi^2_{N-K} \Rightarrow E\left[\frac{\chi^2}{\Phi}\right] = N - K$ i

$$\hat{\Phi}_{pearson} = \frac{\chi^2}{N - K}$$

Validació gràfica del model

Residus: 3 tipus

- **Resposta (Response)**: $r_i = y_i - \hat{\mu}_i$, és intuïtiu però poc útil pels mlg. La seva variància depèn de $V(\hat{\mu}_i)$.
- **Pearson (Pearson)**: $r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$, també és intuïtiu, és com el de resposta però homogeneïtzant les seves variàncies.
  Compleix que $\sum_{i=1}^N \left(r_i^P\right)^2 = \chi^2 = \textit{estadístic de Pearson}$.
- **Deviància (Deviance)**: $r_i^D = sign\left(y_i - \hat{\mu}_i\right) d_i$ on $d_i = \sqrt{2\left(\left(\tilde{\theta}_i y - b\left(\tilde{\theta}_i\right)\right) - \left(\hat{\theta}_i y - b\left(\hat{\theta}_i\right)\right)\right)}$, és poc intuïtiu, s'obté de la deviància (que ve de la màxima versemblança), té les variàncies homogeneïtzades.
  Compleix que $\sum_{i=1}^N \left(r_i^D\right)^2 = \sum_{i=1}^N d_i^2 = D\left(\mathbf{y}, \hat{\mu}_i\right) = \textit{estadístic deviància}$.

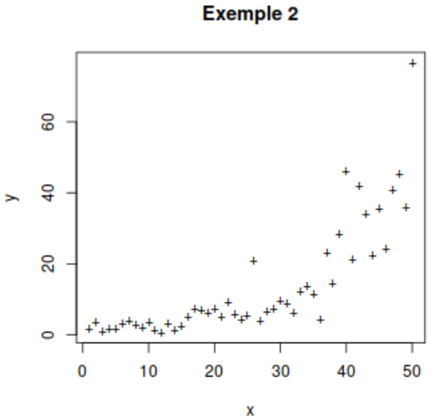
*La família normal té $V(\mu) = 1 \Rightarrow$ **els dos tipus de residus coincideixen** (és l'única).*

Eina gràfica per analitzar la validesa del model

Eina gràfica: ***Residus de Pearson*** vs ***valors predits***

1. **Aquests residus han d'anar al voltant de zero**, sense patrons, ja que $E\left[r_{pearson,i}\right] = 0$. Si no es compleix és que no hem escollit bé:
   - El predictor lineal, $\eta = X\beta$, o
   - La funció link $g\left(\mu\right) = \eta$.
2. **La variància d'aquests residus és constant**. Per tant, a la gràfica, la seva dispersió no ha de créixer, ni decréixer, en funció dels valors predits.
   - Si la 1a no es compleix, aquesta ja no importa.
   - Si aquesta no es compleix, és que no hem escollit bé la funció de variància, $V\left(\mu\right)$, és a dir, no hem escollit bé la família.

Exemple 2 de la 1a secció, com a MLGz



1. Predictor lineal: $\eta = \alpha + \beta x$
2. Funció link: sembla que les prediccions han de seguir la funció $\mu = e^{\alpha+\beta x}$, per tant el link és la seva inversa $g(\mu) = \log(\mu)$
3. Família: De les dades veiem que $Var(y_x) \propto \mu_x^2$ per tant ha de ser la família Gamma.

Exemple 2 de la 1a secció, com a MLGz

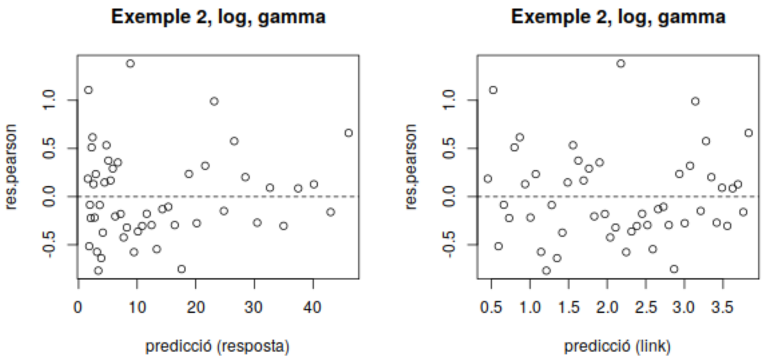
```
Call: glm(formula = y2 ~ x, family = Gamma(link = log))
Deviance Residuals:
     Min        1Q    Median        3Q       Max
 -1.1799   -0.3390   -0.1233    0.2189    1.0136

Coefficients:  Estimate Std. Error t value Pr(>|t|)
   (Intercept) 0.387464   0.135526   2.859   0.00627 **
   x           0.068855   0.004625  14.886   < 2e-16 ***
(Dispersion parameter for Gamma family taken to be 0.2227708)

Null deviance: 60.122  on 49  degrees of freedom Residual
     deviance: 11.401  on 48  degrees of freedom
AIC: 278.8
Number of Fisher Scoring iterations: 5
```

Exemple 2 de la 1a secció, com a MLGz

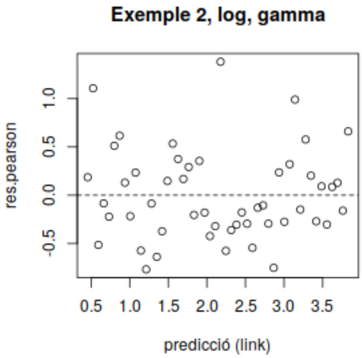
Comparació gràfica $r_{pearson}$ vs predicció,
(És preferible tipus tipus link $\hat{\eta}_x$ a tipus resposta $\hat{\mu}_x$)

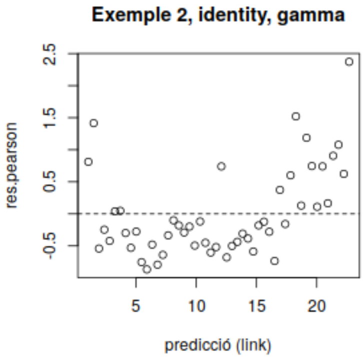


Exemple 2 de la 1a secció, com a MLGz

Comparació gràfica: models amb el link diferent.

Amb el link identitat no es compleix $E\left[r_{perason}\right] = 0$.

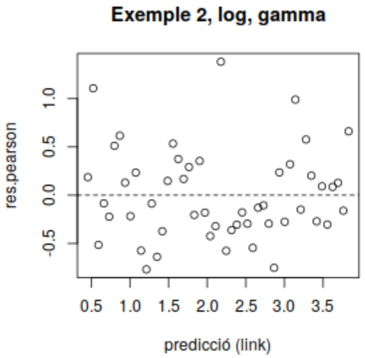




Exemple 2 de la 1a secció, com a MLGz

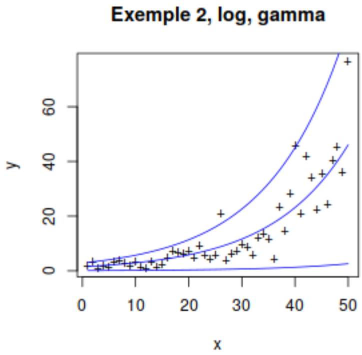
Comparació gràfica: models amb diferent família.

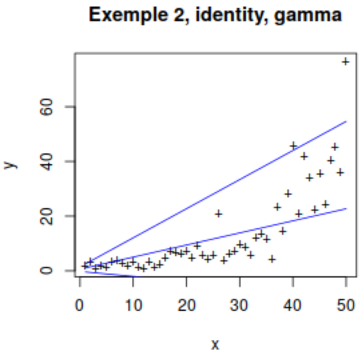
Amb la família normal no es compleix $Var\left[r_{perason}\right] = ctnt$.



Exemple 2 de la 1a secció, com a MLGz

Com que són models amb una sola variable regressora, també ajuda a veure l'efecte del link i/o la família que hem escollit, la gràfica de la *"pseudo banda de predicció"*: $\hat{\mu}_x \pm 2S_x$.

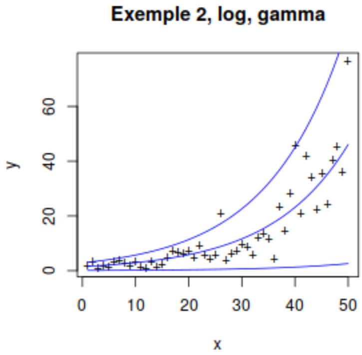


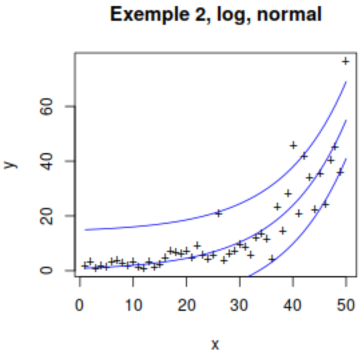


Exemple 2 de la 1a secció, com a MLGz

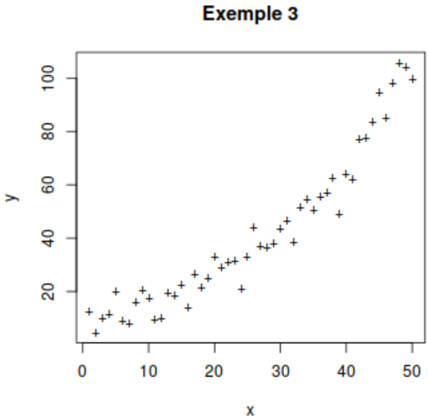
Pseudo banda de predicció: $\hat{\mu}_x \pm 2S_x$.

Ara canviem la família, no el link.





Exemple 3 de la 1a secció



1. Predictor lineal: $\eta = \alpha + \beta x$
2. Funció link: sembla que les prediccions han de seguir la funció $\mu = e^{\alpha+\beta x}$, per tant el link és la seva inversa $g(\mu) = \log(\mu)$
3. Família: La variància sembla constant $\Rightarrow$ Normal

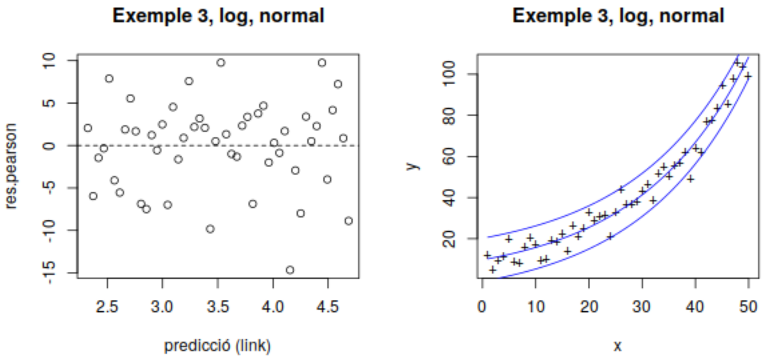
Exemple 3 de la 1a secció

```
Call: glm(formula = y3 ~ x, family = gaussian(link = log))
Deviance Residuals:
     Min        1Q    Median       3Q       Max
  -14.670    -2.697     0.882    3.011     9.759
Coefficients:    Estimate Std. Error t value Pr(>|t|)
    (Intercept) 2.275433   0.065654   34.66   <2e-16 ***
    x           0.048195   0.001579   30.53   <2e-16 ***
(Dispersion parameter for gaussian family taken to be 27.48464)
    Null deviance: 40348.2  on 49  degrees of freedom
Residual deviance:  1319.3  on 48  degrees of freedom
AIC: 311.53
Number of Fisher Scoring iterations: 4
```

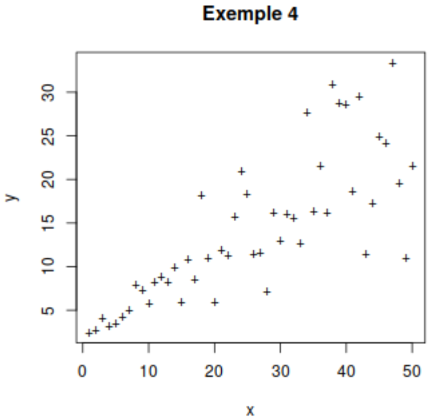
Exemple 3 de la 1a secció

gràfica $r_{pearson}$ vs predicció (link) i

"Pseudo banda de predicció": $\hat{\mu}_x \pm 2S_x$.



Exemple 4 de la 1a secció



1. Predictor lineal: $\eta = \alpha + \beta x$
2. Funció link: sembla que les prediccions han de seguir la funció $\mu = \alpha + \beta x$, per tant el link és la identitat $g(\mu) = \mu$
3. Família: De les dades veiem que $Var(y_x) \propto \mu_x^2$ per tant ha de ser la família Gamma.

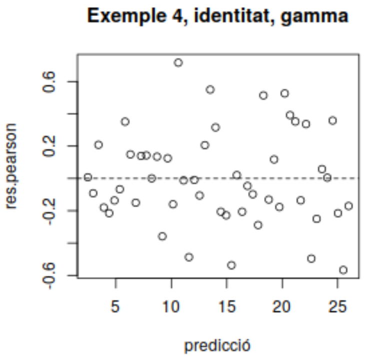
Exemple 4 de la 1a secció

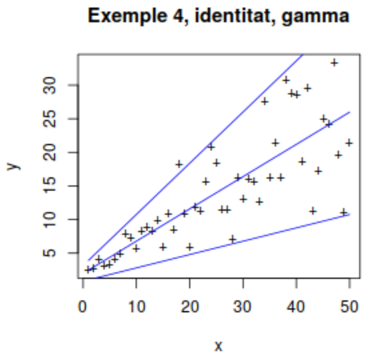
```
glm(formula = y4 ~ x, family = Gamma(link = identity))
Deviance Residuals:
      Min          1Q      Median          3Q         Max
 -0.73495    -0.19121    -0.02964     0.14083     0.59370
Coefficients:  Estimate Std. Error t value Pr(>|t|)
   (Intercept)  1.99004    0.41773   4.764 1.79e-05 ***
   x            0.47991    0.03223  14.892  < 2e-16 ***
(Dispersion parameter for Gamma family taken to be 0.08598135)
    Null deviance: 19.5132  on 49  degrees of freedom
Residual deviance:  4.3873  on 48  degrees of freedom
AIC: 272.17
Number of Fisher Scoring iterations: 5
```

Exemple 4 de la 1a secció

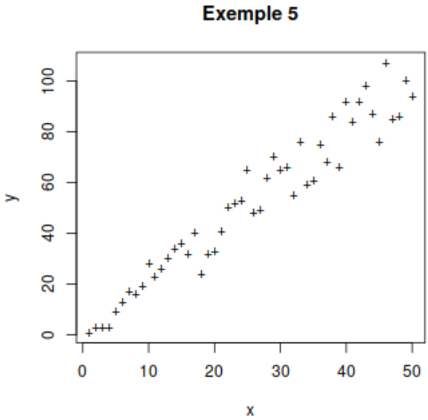
gràfica $r_{pearson}$ vs predicció (link) i

"Pseudo banda de predicció": $\hat{\mu}_x \pm 2S_x$.





Exemple 5 de la 1a secció, Poisson



1. Predictor lineal: $\eta = \alpha + \beta x$
2. Funció link: sembla que les prediccions han de seguir la funció $\mu = \alpha + \beta x$, per tant el link és la identitat $g(\mu) = \mu$
3. Família: Sabem que és Poisson

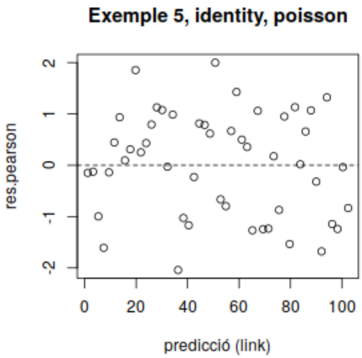
Exemple 5 de la 1a secció, Poisson

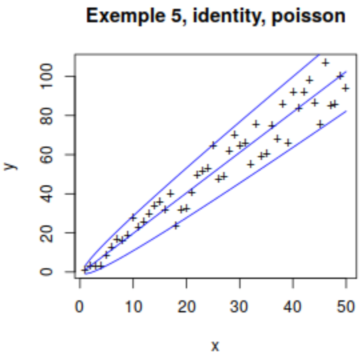
```
Call: glm(formula=y5~x, family=poisson(link=identity))
Deviance Residuals:
     Min        1Q    Median       3Q      Max
  -2.1766   -0.8760   0.1338   0.7920   1.9138
Coefficients:  Estimate Std. Error z value Pr(>|z|)
   (Intercept) -0.90758    0.77021  -1.178    0.239
   x            2.06696    0.05061  40.841   <2e-16 ***
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 986.466  on 49  degrees of freedom
Residual deviance:  50.644  on 48  degrees of freedom
AIC: 328.71
Number of Fisher Scoring iterations: 4
```

Exemple 5 de la 1a secció, Poisson

gràfica $r_{pearson}$ vs predicció (link) i

"Pseudo banda de predicció": $\hat{\mu}_x \pm 2S_x$.



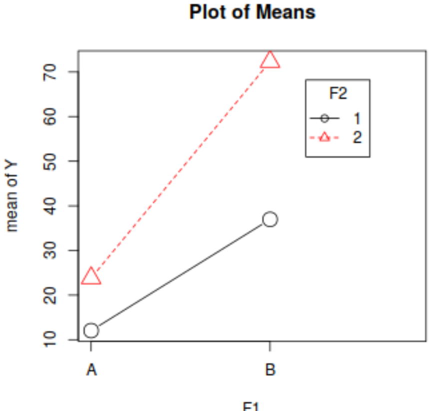


Exemple de dos factors

S'ha de considerar les interaccions: **Canvia l'efecte d'un factor sobre la resposta segons el nivell d'un altre factor?**

Taula de mitjanes

		Y	
		F2	
F1		1	2
A	N	20.00	20.00
	mitjana	12.04	23.83
B	N	20.00	20.00
	mitjana	36.97	72.32

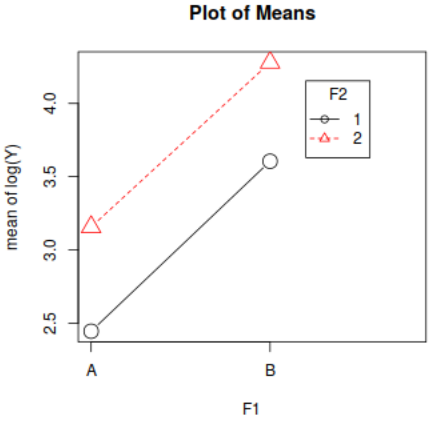


Exemple de dos factors

Que hi hagi interacció o no depen de l'escala.

Amb $\log(Y)$ no es veu interacció

		log(Y) F2	
F1		1	2
A	N	20.000	20.000
	mitjana	2.445	3.156
B	N	20.000	20.000
	mitjana	3.605	4.279



Exemple de dos factors

$\mathtt{glm}(\mathtt{Y} \sim \mathtt{F1} * \mathtt{F2}, \mathtt{family} = \mathtt{gaussian}(\mathtt{link} = \textit{identity}))$

```
Analysis of Deviance Table (Type II tests)
Response: Y
Error estimate based on Pearson residuals
           Sum Sq Df F value    Pr(>F)
F1        26951.0  1 1716.26 < 2.2e-16 ***
F2        11110.4  1  707.52 < 2.2e-16 ***
F1:F2      2775.1  1  176.72 < 2.2e-16 ***
Residuals  1193.5 76
```

Exemple de dos factors

$\texttt{glm}(\texttt{Y} \sim \texttt{F1} * \texttt{F2}, \texttt{family} = \texttt{gaussian}(\texttt{link} = \textit{log}))$

```
Analysis of Deviance Table (Type II tests)
Response: Y
Error estimate based on Pearson residuals
            Sum Sq Df   F value Pr(>F)
F1         29725.9  1 1892.9661 <2e-16 ***
F2         13885.3  1  884.2250 <2e-16 ***
F1:F2          0.3  1    0.0184 0.8925
Residuals   1193.5 76
```

Exemple de dos factors

Estimacions de les μ_{ij}. Model factorial amb els dos links

Les estimacions no canvien emprem un link o un altre. Per tant, en presència només de factors, no importa la funció d'enllaç escollida per fer prediccions.

$\hat{\mu}_{ij}$ amb *link* = *identity*

id	F1	F2	emmean	SE
01	A	1	12.0400	0.886
02	B	1	36.9695	0.886
03	A	2	23.8300	0.886
04	B	2	72.3185	0.886

$\hat{\mu}_{ij}$ amb *link* = *log*

id	F1	F2	response	SE
01	A	1	12.0400	0.886
02	B	1	36.9695	0.886
03	A	2	23.8300	0.886
04	B	2	72.3185	0.886

Exemple de dos factors

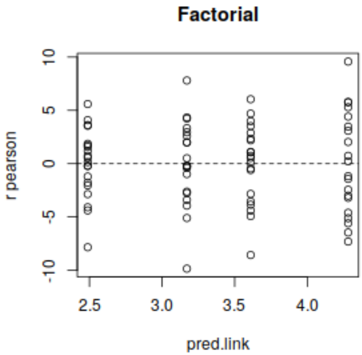
$\mathtt{glm(Y \sim F1 + F2, family = gaussian(link = \textit{log}))}$

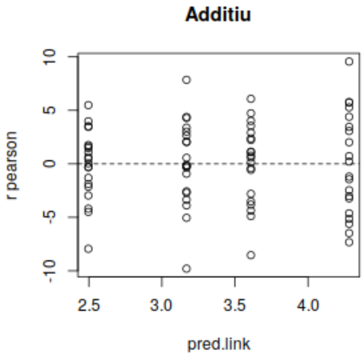
```
Analysis of Deviance Table (Type II tests)
Response: Y
Error estimate based on Pearson residuals
            Sum Sq Df F value    Pr(>F)
F1         29725.9  1 1917.41 < 2.2e-16 ***
F2         13885.3  1  895.64 < 2.2e-16 ***
Residuals   1193.7 77
```

Exemple de dos factors

gràfica $r_{pearson}$ vs predicció (link)

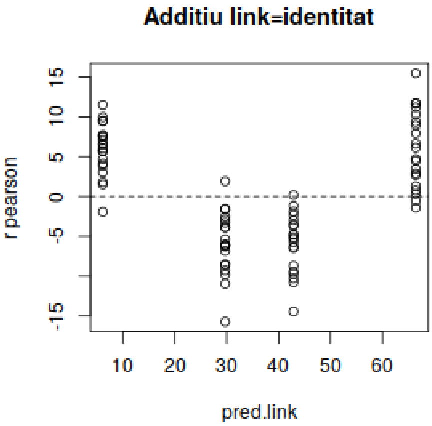
Amb interacció i sense



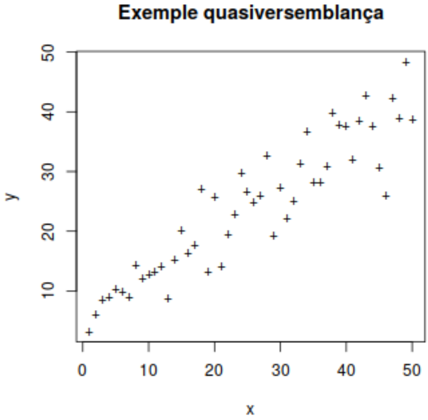


Exemple de dos factors

gràfica $r_{pearson}$ vs predicció (link). Sense interacció i link=identitat



Exemple quasiversemblança



1. Predictor lineal: $\eta = \alpha + \beta x$
2. Funció link: sembla que les prediccions han de seguir la funció $\mu = \alpha + \beta x$, per tant el link és la identitat $g(\mu) = \mu$
3. Família: De les dades veiem que la variància augmenta però $Var(y_x) \propto \mu_x$, com la Poisson, però són dades contínues, família desconeguda, utilitzarem quasiversemblança amb $V(\mu) = \mu$.

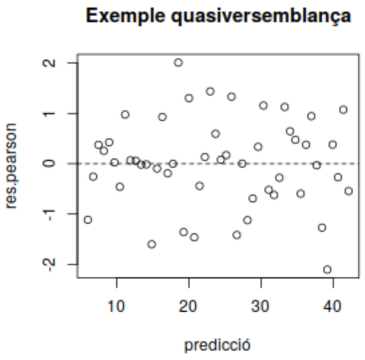
Exemple quasiversemblança

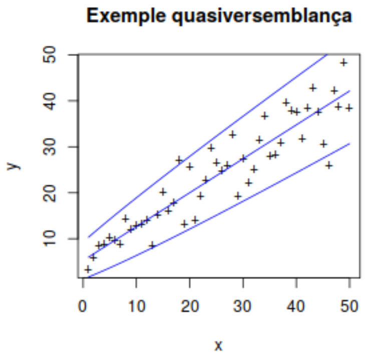
```
glm(formula=yq~x,family=quasi(link=identity,variance="mu"))
Deviance Residuals:
     Min        1Q    Median        3Q       Max
 -2.2437   -0.5135    0.0132    0.4549    1.8768
Coefficients:    Estimate Std. Error t value Pr(>|t|)
    (Intercept)   5.25941    0.85830   6.128  1.6e-07 ***
    x             0.73786    0.03819  19.323  < 2e-16 ***
(Dispersion parameter for quasi family taken to be 0.7799602)
    Null deviance: 284.295  on 49  degrees of freedom
Residual deviance:  38.335  on 48  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 5
```

Exemple quasiversemblança

gràfica $r_{pearson}$ vs predicció (link) i

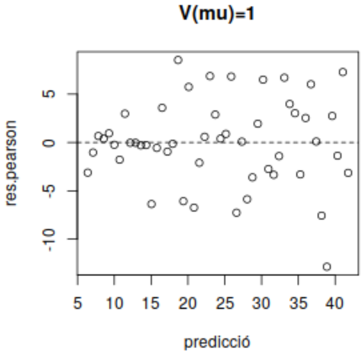
"Pseudo banda de predicció": $\hat{\mu}_x \pm 2S_x$.

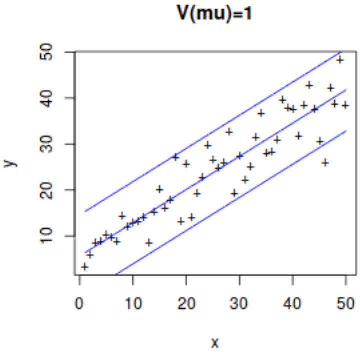




Exemple quasiversemblança

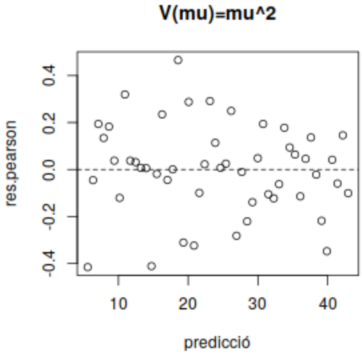
$V(\mu) = 1 \Leftrightarrow$ *Normal*: gràfica $r_{pearson}$ vs predicció (link) i *"Pseudo banda de predicció": $\hat{\mu}_x \pm 2S_x$. No compleix.*

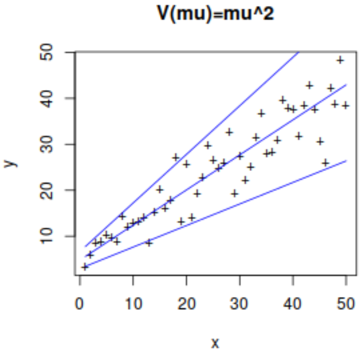




Exemple quasiversemblança

$V(\mu) = \mu^2 \Leftrightarrow$ *Gamma*: gràfica $r_{pearson}$ vs predicció (link) i *"Pseudo banda de predicció"*: $\hat{\mu}_x \pm 2S_x$. *No compleix.*





Prediccions

Prediccions

Interval de predicció de y_x

Un **interval de predicció** és un rang de valors dins del qual es preveu que es trobi una nova observació de la variable dependent, donat un conjunt de valors dels predictors.

Nota: A diferència dels intervals de confiança, per l'estimació de la mitjana de la resposta, un interval de predicció té en compte tant la variabilitat de la mitjana com la variabilitat inherent de la nova observació.

Interval de predicció de y_x en els MLGz

En els models lineals és fàcil obtenir-los ja que sabem que:

$$y_x - \hat{y}_x \sim N\left(0, \sigma_e^2 + \sigma_{\hat{y}_x}^2\right)$$

En els MLGz es complica una mica, però tenim 3 opcions.

Maneres de calcular un interval de predicció (I)

1. **Pseudo-interval**. És intuïtiu i fàcil de calcular, però no és rigorós ja que assumeix distribució Normal:

$$IP_{95\%}\left(\mu_x\right) = \left(\hat{\mu}_x \pm 2\sqrt{\Phi V\left(\hat{\mu}_x\right)}\right) \approx \left(\hat{\mu}_x \pm 2\sqrt{\hat{\Phi} V\left(\hat{\mu}_x\right)}\right)$$

2. **Asimptòtic**. Es considera que $\beta = \hat{\beta}$, $\Phi = \hat{\Phi}$ i que $F_{y_x}\left(s\right)$ és la funció de distribució de y_x que dona la família del MLGz. Llavors:

$$IP_{1-\alpha}\left(y_x\right) = \left(F_{y_x}^{-1}\left(\frac{\alpha}{2}\right), F_{y_x}^{-1}\left(1-\frac{\alpha}{2}\right)\right)$$

on $F_{y_x}^{-1}$ és la funció inversa de F_{y_x}.

Maneres de calcular un interval de predicció (II)

3. **Bootstrap paramètric**. Consisteix en generar un nombre alt de valors de $y_{x,j}$ en tres passos:

    1. Simulem el valor de Φ:
        - Si Φ és desconegut, generem el seu valor a partir de $\Phi_j = \hat{\Phi}\left(\frac{\chi^2_\nu}{\nu}\right)^{-1}$
        - Si Φ és conegut, sempre $\Phi_j = \Phi$.
    2. Generem el valor de β_j a partir de $N\left(\hat{\beta}, \hat{\Sigma}_{\hat{\beta}}\right)$.
    3. Generem el valor de $y_{x,j}$ a partir F_x que és la distribució de la família amb els paràmetres $\mu_{x,j} = g^{-1}\left(x\beta_j\right)$ i $\sigma^2_{\mu_{x,j}} = \Phi_j V\left(\mu_{x,j}\right)$.

    Un cop tenim el vector de les $(y_{x,j})$ en calculem els percentils $\frac{\alpha}{2}$ i $1-\frac{\alpha}{2}$, que conformaran l'**interval de predicció**.

Tests Raó de versemblances

Test Raó de versemblança I

Test pels models lineals generalitzats

Tant el test omnibus com els tests dels efectes es basen en la comparació de dos models **encaixats** amb el test de raó de versemblança.

Test Raó de versemblança II

Test de raó de versemblança entre ***mlg*** encaixats, $m_1 \subset m_2$

Models: m_1 amb K_1 paràmetres, m_2 en té $K_2\,(> K_1)$ i N dades.

El test de raó de versemblança en general es contrasta amb:

$$\Lambda = 2\,(\ell_2 - \ell_1) \sim \chi^2_{K_2 - K_1}.$$

En els MLGz, s'empra el següent estadístic:

$$\Lambda = \frac{2(\ell_2 - \ell_1)}{\Phi} = \frac{deviance_1 - deviance_2}{\Phi} = \frac{dif.dev}{\Phi}$$

Hi ha dues possibilitats (veure diapositives següents)

Test Raó de versemblança III

1) Si Φ és coneguda:

Com el cal de la Binomial, Poisson,... aleshores:

$$\Lambda = \frac{dif.dev}{\Phi} \sim \chi^2_{K_2 - K_1}$$

i per tant l'hem de contrastar amb una $\chi^2_{K_2 - K_1}$.

R

```
anova(m1,m2,test="LR")
Anova(mod,test.statistic="LRT")
```

`m1`, `m2` són dos models encaixats i `mod` és el model més complert.

Test Raó de versemblança IV

2) Si Φ no és coneguda

Com és el cas de la Normal o la Gamma. Aleshores:

$$\Lambda = \frac{dif.dev}{\hat{\Phi}}, \text{ on } dif.dev \sim \chi^2_{K_2-K_1} \text{ i } \hat{\Phi} \sim \frac{\chi^2_{N-K}}{N-K}$$

Φ l'estimem amb $\hat{\Phi}$, que pot ser la de *Pearson* o la de *Deviance*:

$$F = \frac{\Lambda}{K_2-K_1} \sim \frac{\left(\chi^2_{K_2-K_1}\right)/(K_2-K_1)}{\left(\chi^2_{N-K}\right)/(N-K)} \sim F\left(K_2-K_1, N-K\right)$$

R

```
anova(m1 ,m2,test="F")
Anova(mod,test.statistic="F")
```

Podem especificar el paràmetre error.estimate="pearson" (default) o "deviance"

Take-home messages

- L'**estadístic de Pearson** i de la **Deviància** serveixen per fer la validació numèrica del model.
- La validació gràfica la podem fer emprant el **gràfic de residus de Pearson vs. valors predits**
- Hi ha **3 mètodes per fer prediccions**: pseudo-interval, assimptòtic, bootstrap.
- Per comparar models encaixats, fem servir el **test de raó de versemblances**.