MODELS LINEALS GENERALITZATS. INTRODUCCIÓ (I)

Grau en Intel·ligència Artificial

Curs 2024-2025

Índex

1. Model lineal general. Necessitat d'ampliar-lo
   - Transformacions en models lineals
   - Models no lineals
   - Incompliment homoscedasticitat
   - Models amb variable discreta
2. Models lineals generalitzats
   - Definició de MLGz
   - Algunes famílies dels MLGz
3. Propietats de les famílies dels MLGz
   - Funció de variància
   - Funció link
   - Les tres parts de $\ell_i = \dfrac{\theta_i y - b\left(\theta_i\right)}{\Phi} + c\left(y, \phi\right)$
   - Ajust del model
4. Estimació del model
   - Quasiversemblança

Necessitat d'ampliar els Models lineals

Objectius de la modelització estadística

Descriptiu

Estudiar la relació entre una variable dependent (resposta) i una o més variables independents (predictors).

Predicció

Predir valors desconeguts d'una variable a partir de la informació d'altres variables.

Etiològic

Identificar Factors Rellevants i estimar l'efecte d'un canvi en les variables explicatives sobre la variable dependent

Model lineal general

Premisses

En el **model lineal general** o **model lineal** assumim:

- Dades independents
- Distribució normal amb paràmetres:
  - esperança (μ): $X\beta$
  - variància (σ^2): constant

Premisses en forma compacte

$Y|X = X\beta + \epsilon$ on $\epsilon \in N(0, \sigma)$ independents i σ no depèn de les X

Estimació

L'estimació es fa per mínims quadrats (`lm`) i coincideix amb màxima versemblança.

Tipus de models lineals

Hi ha diferents tipus de models lineals generals depenent de com són les variables predictores.

Tipus de models lineals

- **Regressió lineal, simple o múltiple**, quan les variables explicatives són covariables (*continues*).
- **Anàlisi de la variància**, quan les variables explicatives són factors (*categòriques*).
- **Anàlisi de la covariància**, quan les variables explicatives són covariables i factors.

Exemples

Aquesta diapositiva conté un resum dels exemples que veurem a continuació. En alguns, hem de recòrrer a un model lineal generalitzat per complir les premisses.

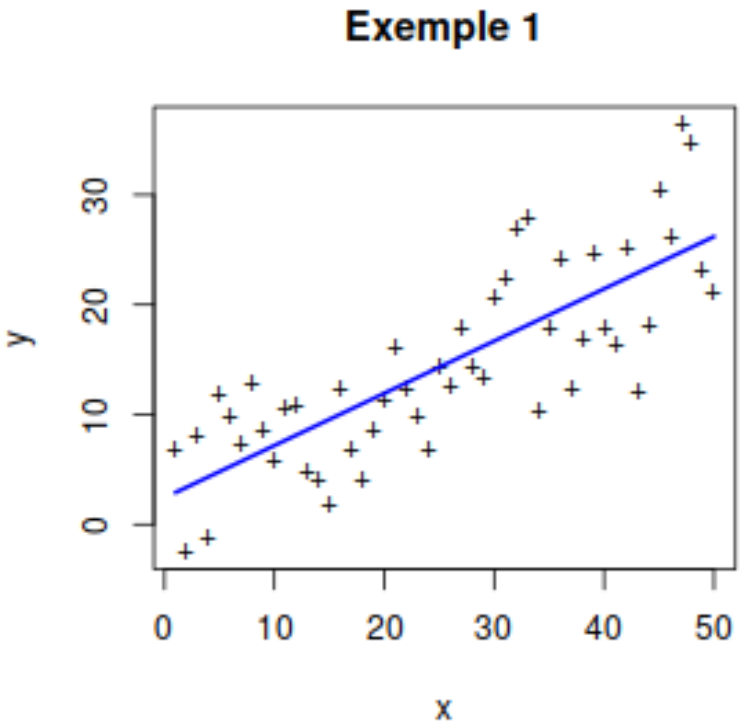
- Model lineal general aplicable
  1. Relació lineal
  2. Transformant resposta

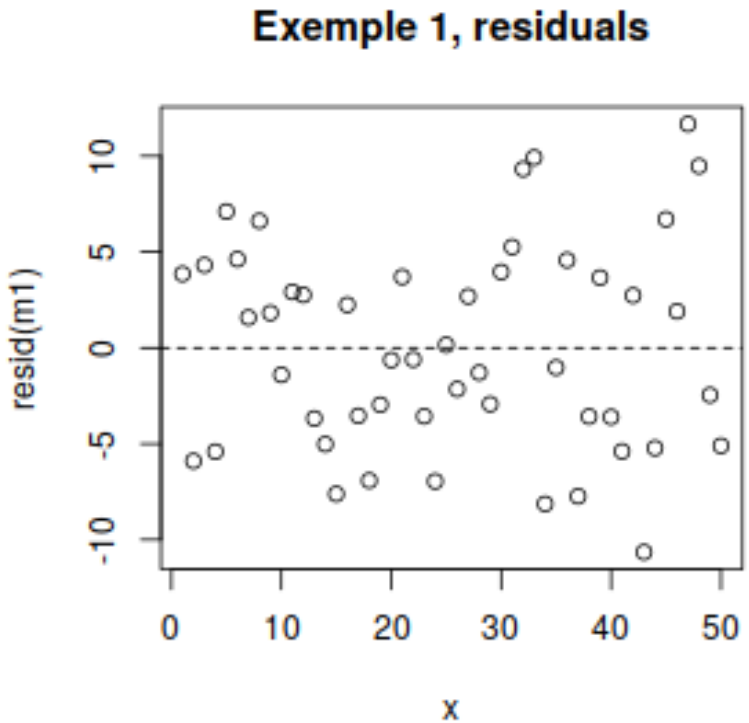
- Model lineal general NO aplicable
  3. Relació no lineal
  4. Incompliment homoscedasticitat
  5. Resposta discreta

Exemple 1: model lineal directe

Variable resposta contínua que compleix la tendència central i l'homoscedasticitat

Podem ajustar un model lineal general





Transformacions

En ocasions, si s'apliquen transformacions sobre la variable resposta podem arreglar l'incompliment de les premisses.
L'objectiu pot ser:

- Adaptar l'esperança per que quedi lineal
- Canviar la distribució de les dades així per a que la variància sigui constant
- Normalitzar la distribució de les dades experimentals

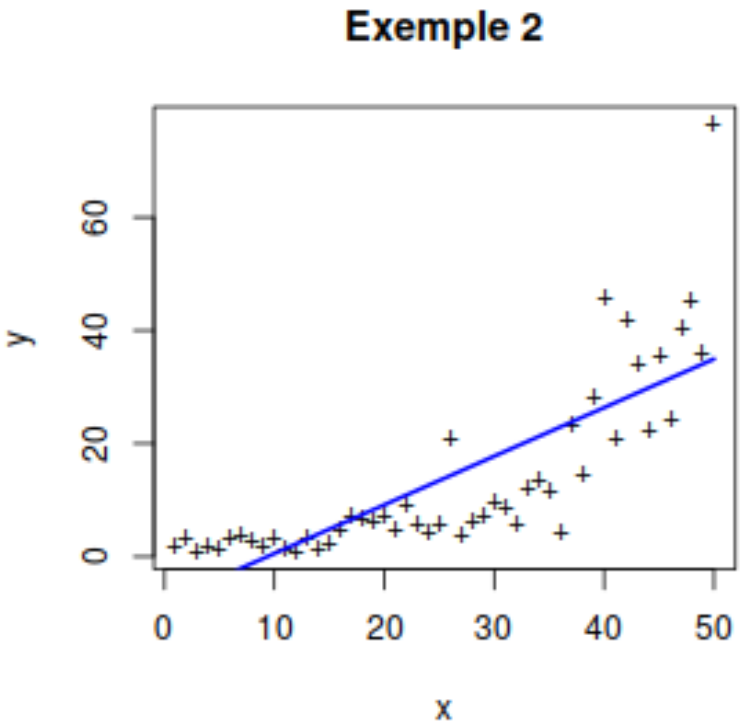
Transformació logarítmica

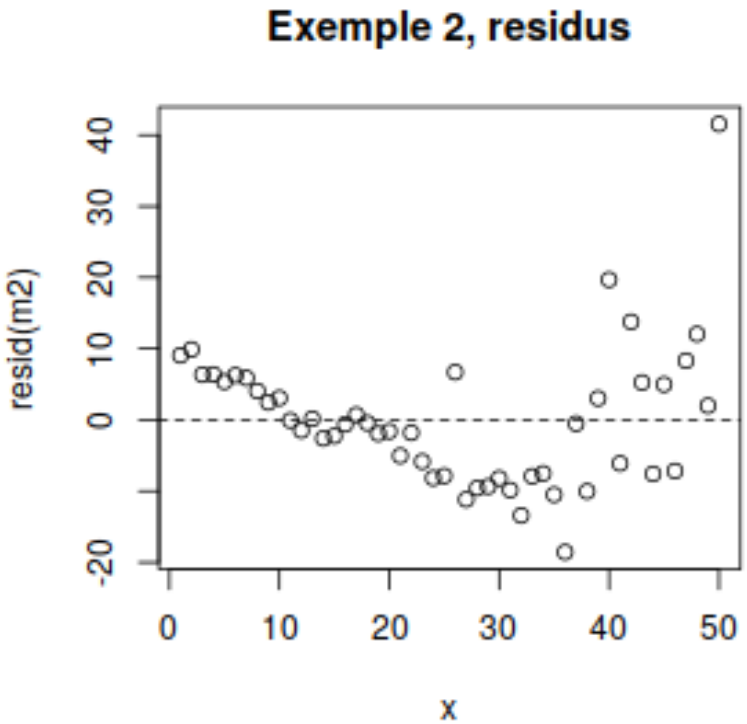
La **transformació logarítmica** sols funcionar prou bé en diverses situacions.

Exemple 2: abans de transformar la variable resposta

Situació amb variable resposta contínua que no compleix assumpcions, però es pot linealitzar.

No compleix ni la tendència central ni l'homoscedasticitat

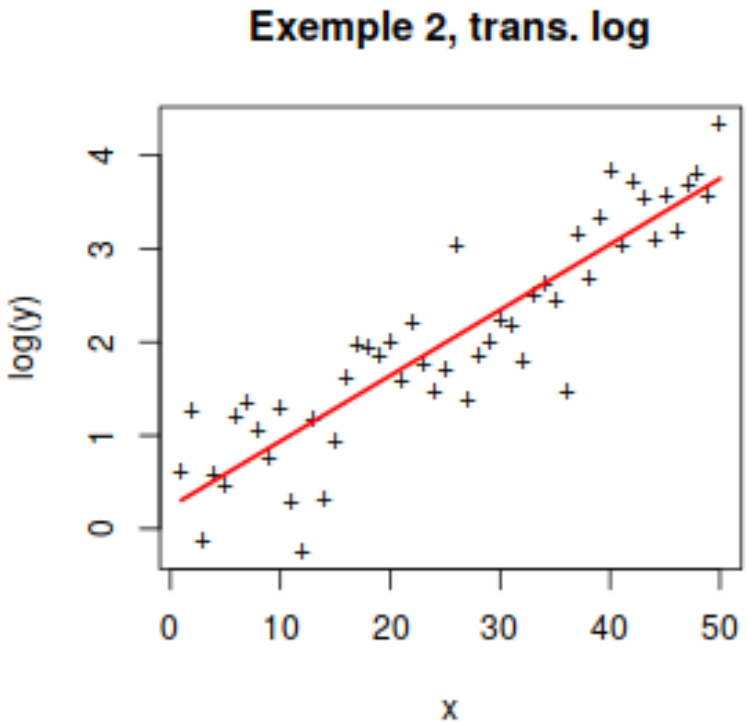


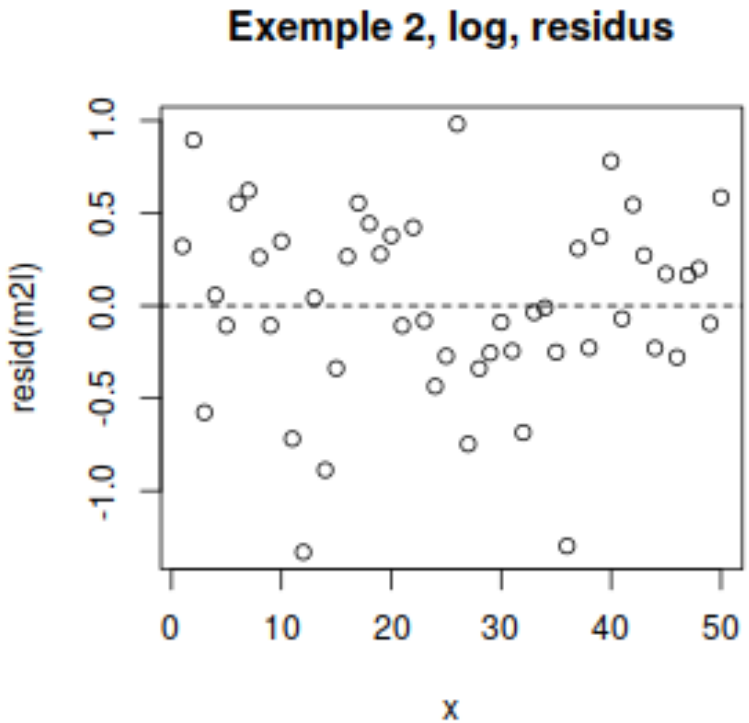


Exemple 2: transformat logarítmicament

L'exemple transformant la y pel $\log(y)$

Amb $\log(y)$ és una recta de regressió, compleix les condicions

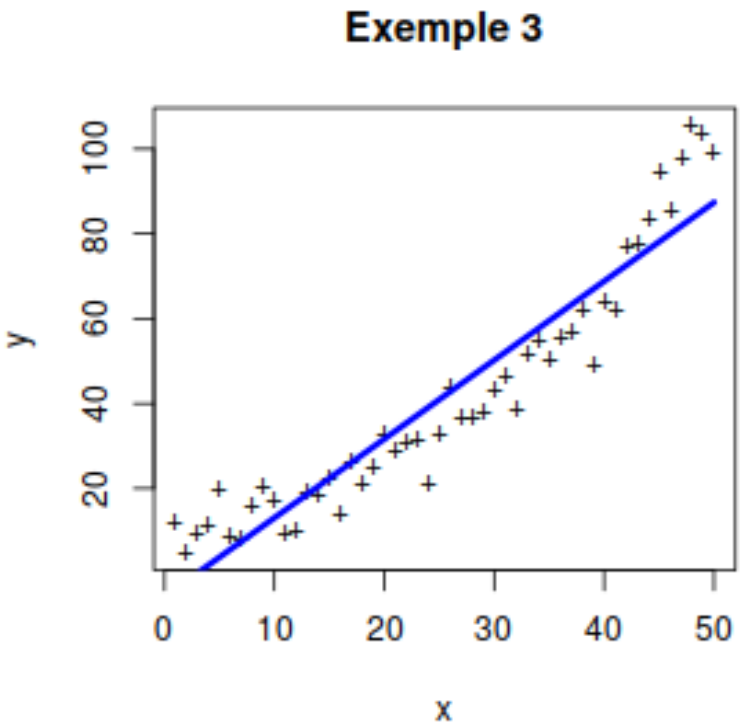


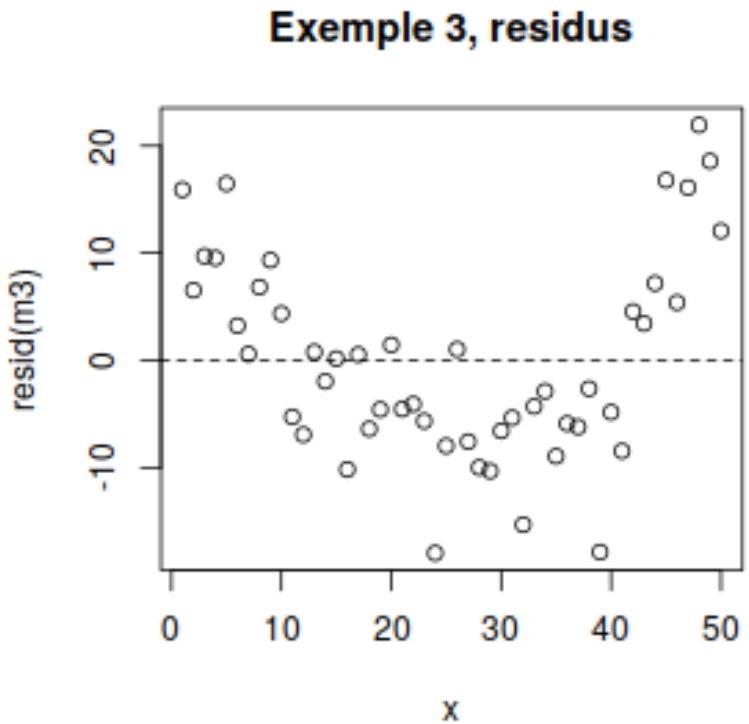


Exemple 3: model normal però no lineal

No és una recta de regressió, no compleix la tendència central

Sí que compleix la homoscedasticitat

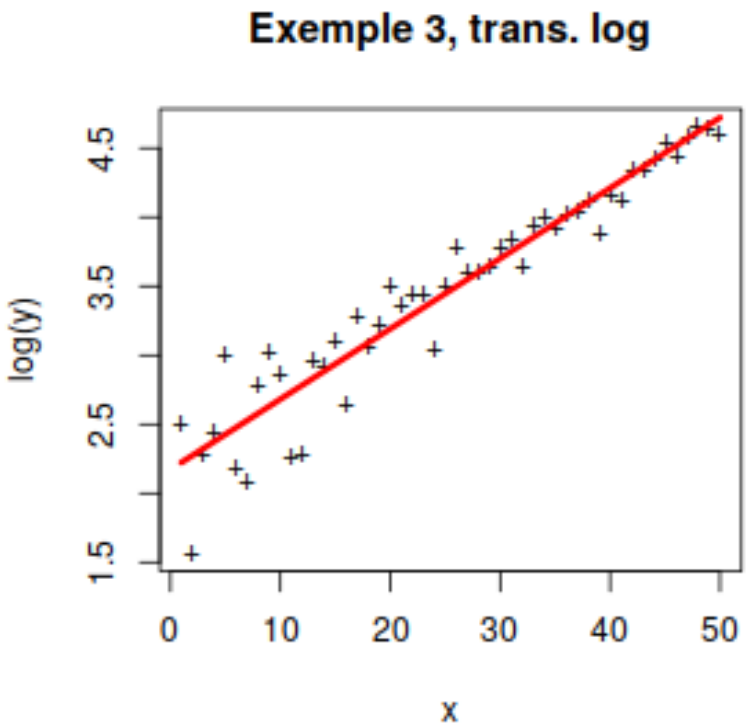


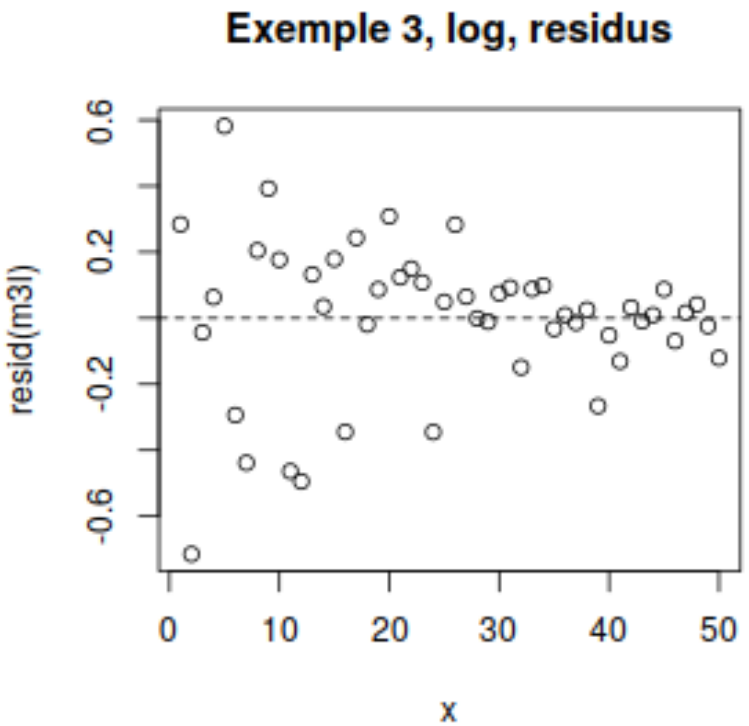


Exemple 3: model normal transformat logarítmicament

L'exemple transformant la y pel $\log(y)$

Amb $\log(y)$ no és una recta de regressió, no compleix l'homoscedasticitat



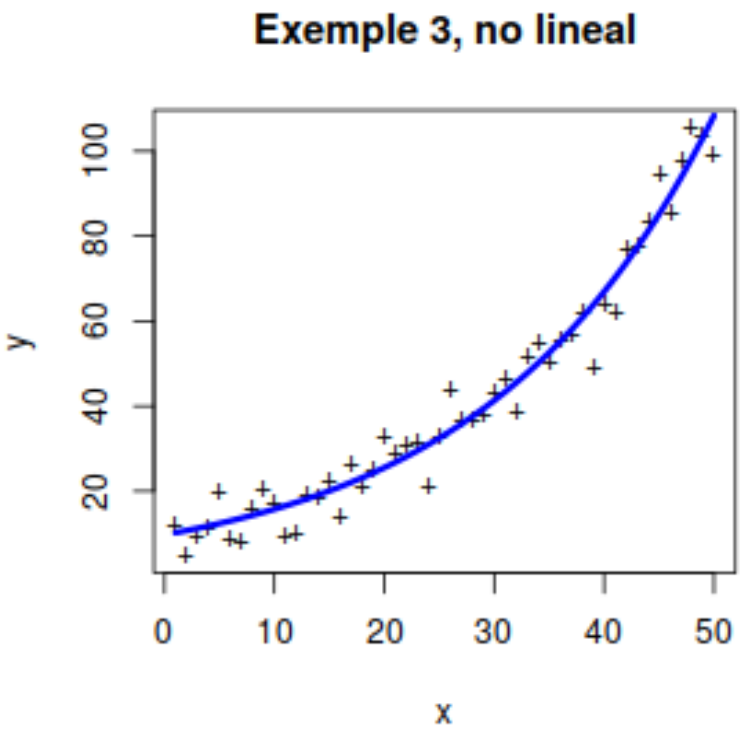


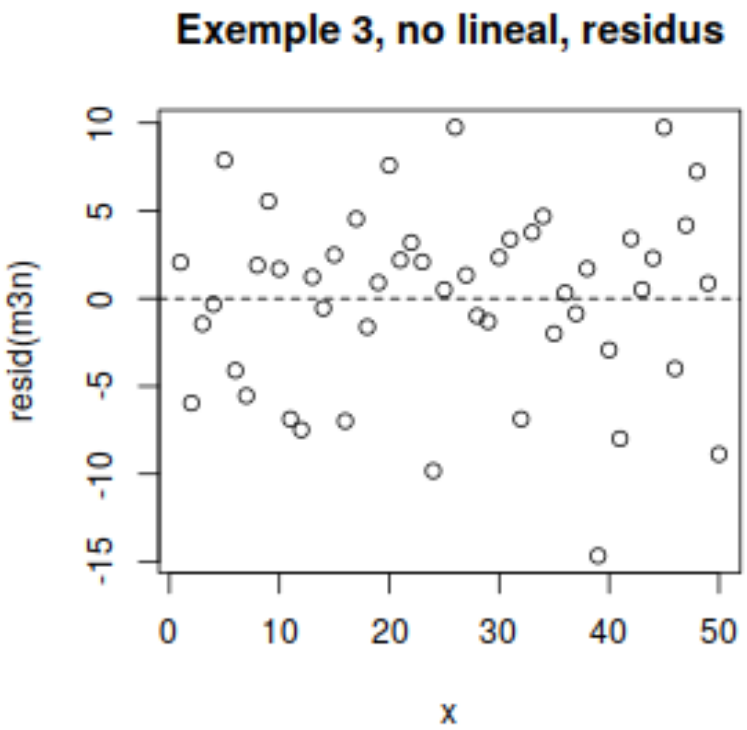
Exemple 3: model NO lineal

L'exemple com a model no lineal,

`nls(y3~exp(a+b*x),start=list(a=2,b=0.5))`

Compleix les condicions, tendència central i homoscedasticitat





Observació

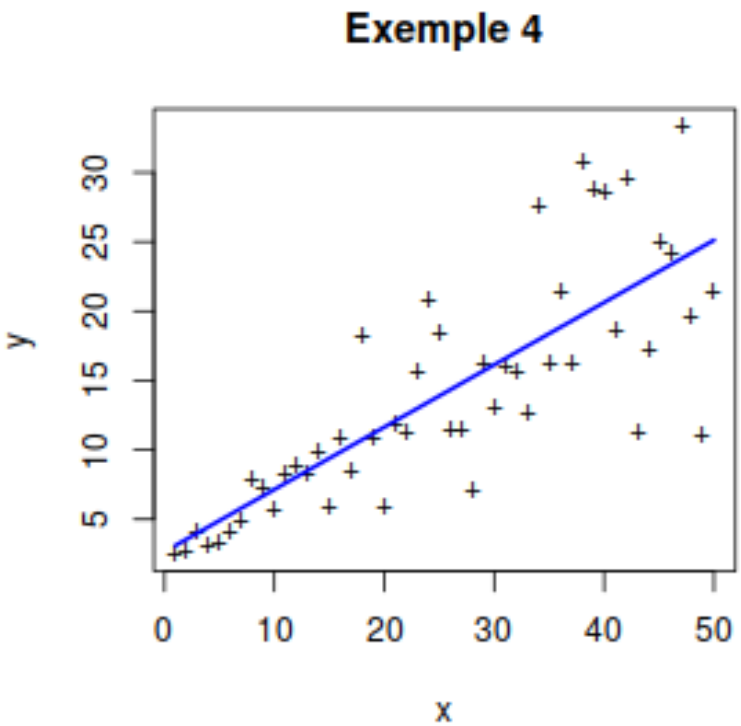
No és equivalent a fer el logaritme de la variable resposta perquè

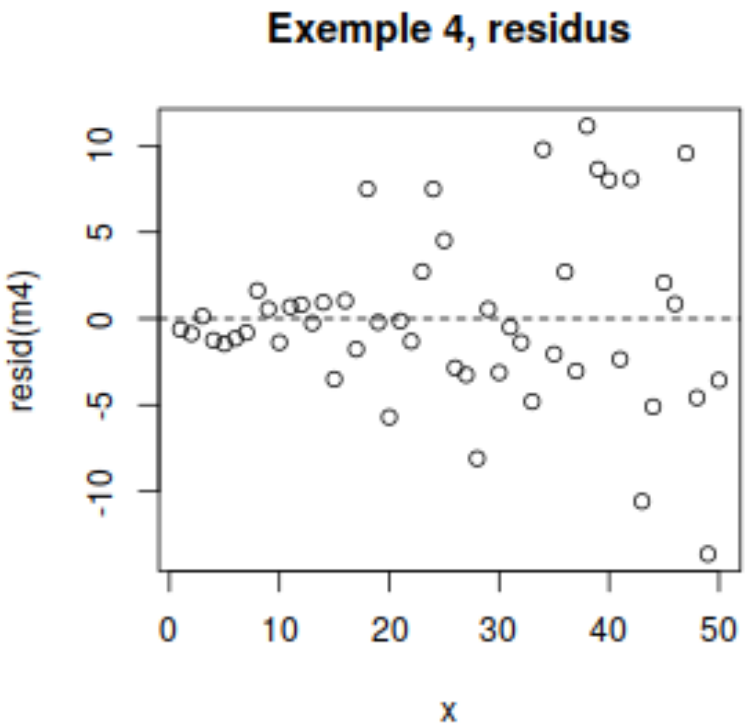
$log(E[Y_3]) \neq E[log(Y_3)]$

Exemple 4: no es compleix la homoscedasticitat

Variable resposta contínua que no compleix homoscedasticitat

Podria ser una recta de regressió però no es compleix la homoscedasticitat

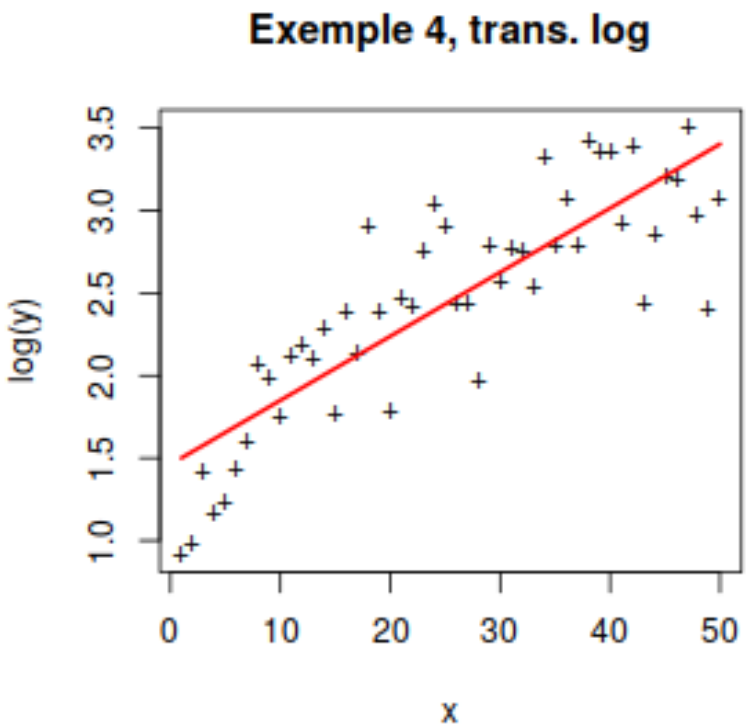


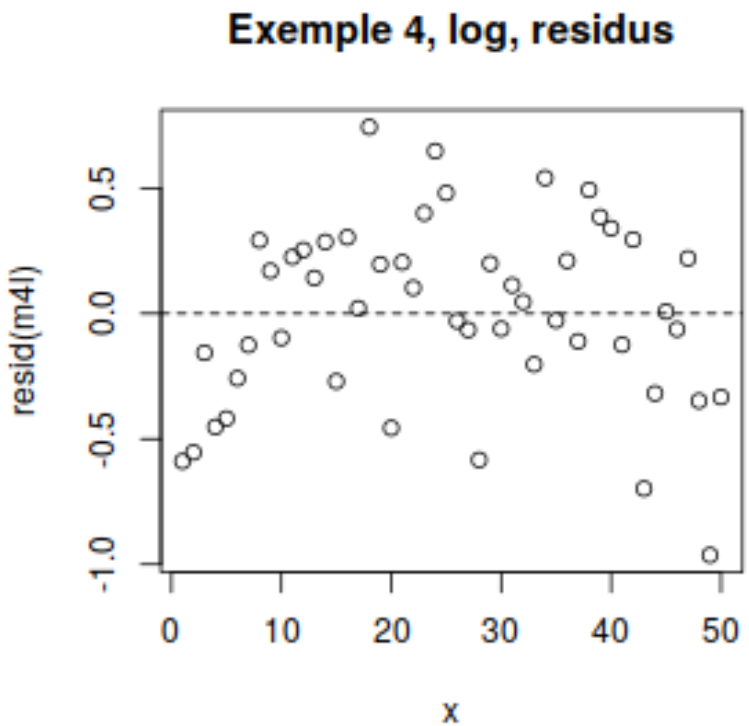


Exemple 4: transformat logarítmicament

Si transformem l'exemple anterior per obtenir homoscedasticitat

Ara el problema no és la variància, si no els valors predits: la recta no descriu la concavitat de les dades $\longleftrightarrow$ els residus no són aleatòriament al voltant de 0

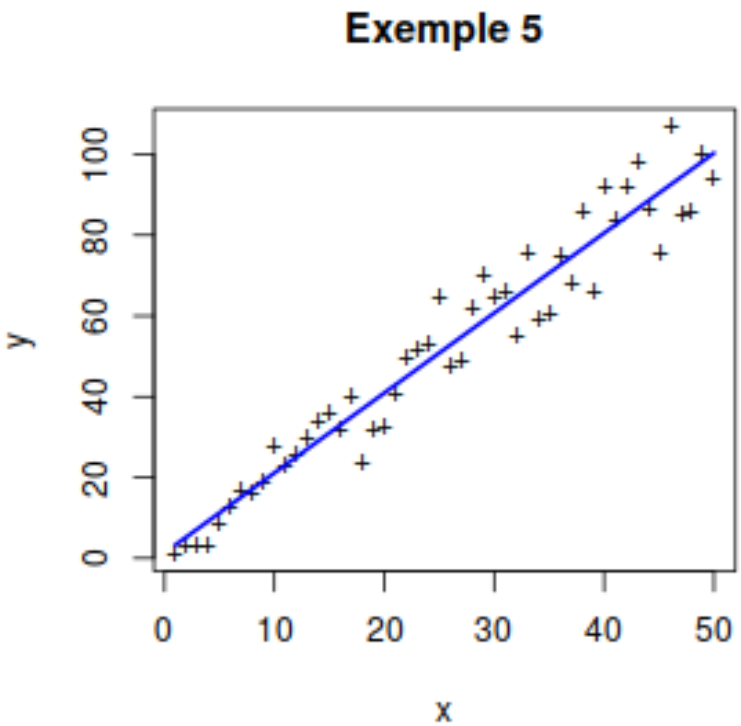


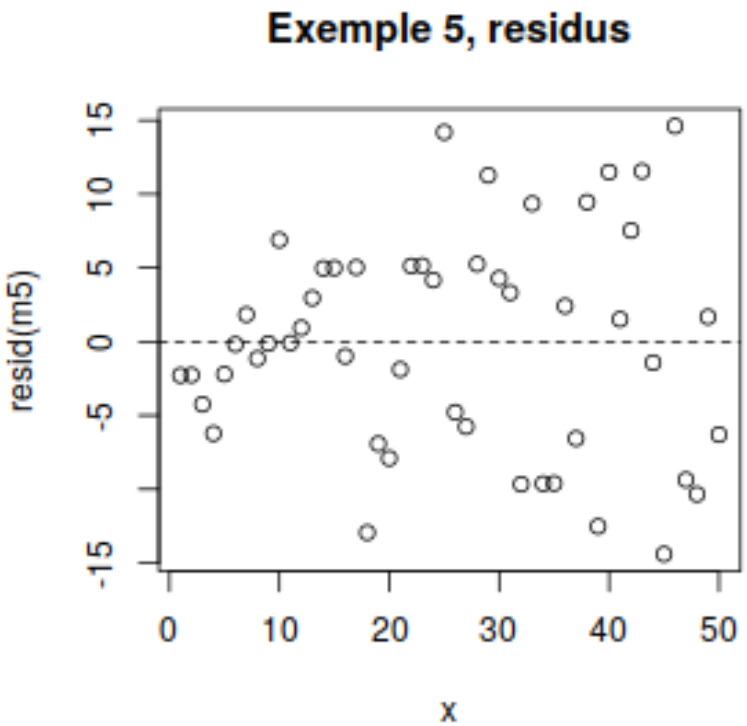


Exemple 5: resposta variable discreta (Poisson)

Variable resposta Poisson

Podria ser una recta de regressió però no es compleix la homoscedasticitat





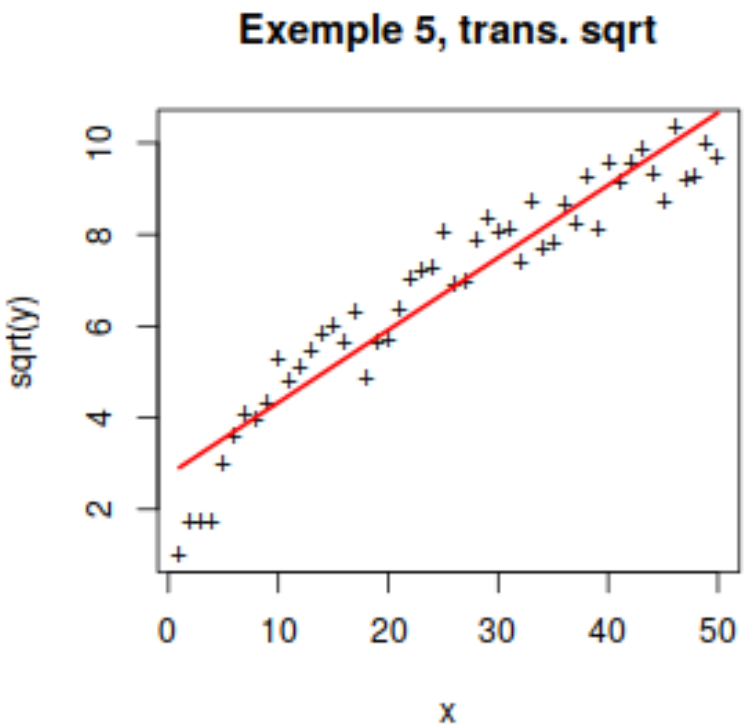
Models de variable discreta (transformacions)

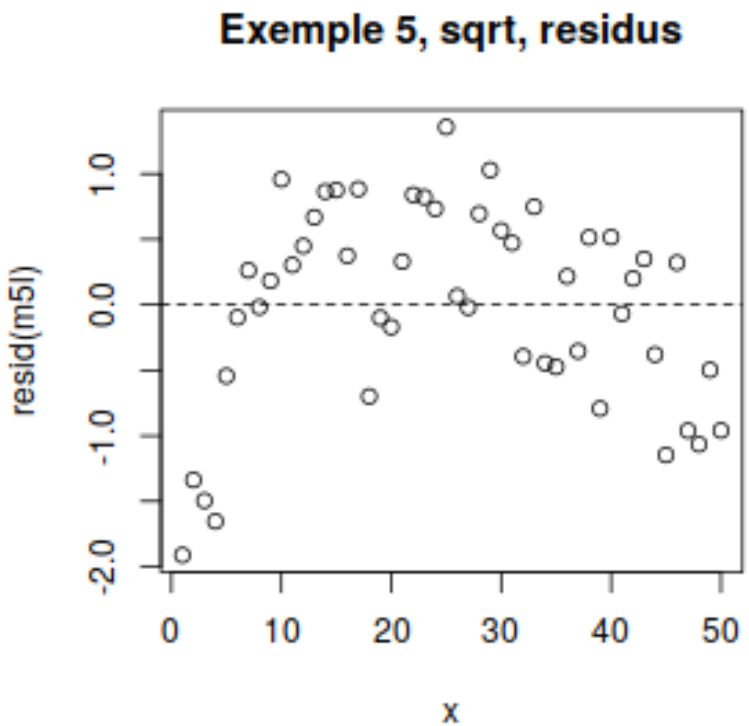
Casos en que la variable resposta no és contínua

- $Y \sim$ **Binomial**
  - En alguns casos pot funcionar la transformació arcsin $\sqrt{\frac{y}{N}}$, o similars.
- $Y \sim$ **Poisson**
  - En alguns casos pot funcionar la transformació $\sqrt{y}$, o similars.

Models de variable discreta

Si transformem, $\sqrt{y}$, l'exemple anterior per obtenir homoscedasticitat, el problema no és la variància, si no els valors predits, la recta no descriu la concavitat de les dades, i més clar, els residus no són aleatòriament al voltant de 0





Limitacions

Limitacions dels models lineals

- Només admet la distribució **normal**.
- La **variància** ha de ser **constant**.
- L'esperança ha de ser $X\beta$, **lineal** respecte les variables explicatives.

Introducció als Models lineals Generalitzats

Models Lineals Generalitzats

Com en els models lineal també tenim:

- $Y = (Y_1, \ldots, Y_N)^t$
- $X = (X_{ij})$
- $\beta = (\beta_1, \ldots, \beta_K)^t$.

Components dels Models Lineals Generalitzats (I)

Els **MLGz** tenen una component **Determinista** i una **Aleatòria**:

Component **Determinista**

Aquesta component conté:

- **Predictor lineal**. $\eta_i = (X_{i,1}, \ldots, X_{i,K})\,\beta$, en global $\boldsymbol{\eta = X\beta}$. No és necessari que sigui la μ.
- **Funció d'enllaç (link)**. Funció bijectiva que relaciona el valor esperat μ_i amb el predictor lineal, η_i:
  - $g(\mu_i) = \eta_i \Longleftrightarrow \boldsymbol{g(\mu) = \eta = X\beta}$.
  - $\mu_i = g^{-1}(\eta_i) \Longleftrightarrow \boldsymbol{\mu = g^{-1}(\eta) = g^{-1}(X\beta)}$.

Components dels Models Lineals Generalitzats (II)

Component **Aleatòria**

La funció de densitat de la variable resposta ha de pertànyer a la *Familía Exponencial* (no ha de ser necessàriament Normal). És a dir, ha de ser de la forma:

$$f_{Y_i}(y;\theta_i,\phi) = e^{\frac{\theta_i y - b(\theta_i)}{a(\phi)} + c(y,\phi)}$$

.

- Les Y_i han de ser independents.
- θ_i és el paràmetre canònic i és funció de μ_i, $\theta_i = \theta(\mu_i)$.
- Anomenarem $\Phi = a(\phi)$ paràmetre de dispersió i a $\sqrt{\Phi}$ s'anomena paràmetre de escala, per totes les Y_i tenen el mateix valor.

Comentaris sobre Models Lineals Generalitzats (II)

Comentaris sobre els **MLGz**'s

- El model lineal general és un cas particular dels MLGz's on:
  - Funció de densitat $\rightarrow Y_i \sim N$
  - Funció link $\rightarrow g(\mu_i) = \mu_i$
- La matriu X del predictor lineal $\eta = X\beta$, es construeix igual que en els models lineals.
- La variància no cal que sigui constant, només ho serà per la família normal. Com canvia la variància en funció de l'esperança depèn de la família.
- La funció link l'escollim en funció de com és l'esperança del model que volem descriure.
- Estimarem els paràmetres per màxima versemblança, amb les propietats que això comporta.

Comparativa MLG vs. MLGz

- MLG: Model Lineal General
- MLGz: Model Lineal Generalitzat

	MLG	MLGz
Dependència	$y's$ independents	$y's$ independents
Familia	$y_i \sim N(\mu_i, \sigma^2)$	$y_i \sim$ familia exponencial
Esperança	$\mu_i = X\beta$	$g(\mu_i) = X\beta$
Estimació	Mínims quadrats	Màxima versemblança

Algunes famílies dels MLGz: $\log\left(f_{Y_i}\left(y\right)\right)=\frac{\theta_i y-b(\theta_i)}{a(\phi)}+c\left(y,\phi\right)$

Normal, família gaussiana:

- Funció de densitat: $f_{Y_i}\left(y;\mu_i,\sigma\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{y-\mu_i}{\sigma}\right)^2}$
- $\ell_i=\log f_{Y_i}\left(y;\mu_i,\sigma\right)=-\frac{y^2}{2\sigma^2}+\frac{\mu_i y}{\sigma^2}-\frac{\mu_i^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)=$
$=\frac{\mu_i y-\mu_i^2/2}{\sigma^2}-\frac{y^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)$

Agafant: $\boldsymbol{\Phi=a\left(\phi\right)=\phi=\sigma^2}$ i $\boldsymbol{\theta_i=\mu_i}$

queda $\ell_i=\frac{\theta_i y-\frac{\theta_i^2}{2}}{a(\phi)}-\frac{y^2}{2a(\phi)}-\frac{1}{2}\log\left(2\pi\phi\right)\Rightarrow$

$$\boldsymbol{b\left(\theta_i\right)=\frac{\theta_i^2}{2}},\ \boldsymbol{\Phi=a\left(\phi\right)=\phi}\text{ i }\boldsymbol{c\left(y,\phi\right)}=-\frac{y^2}{2\phi}-\frac{1}{2}\log\left(2\pi\phi\right)$$

$$\ell_i=\frac{\theta_i y-b\left(\theta_i\right)}{a\left(\phi\right)}+c\left(y,\phi\right)$$

Algunes famílies dels MLGz: $\log\left(f_{Y_i}\left(y\right)\right) = \frac{\theta_i y - b(\theta_i)}{a(\phi)} + c\left(y, \phi\right)$

Poisson:

- $f_{Y_i}\left(y; \lambda_i\right) = e^{-\lambda_i} \frac{\lambda_i^y}{y!}$
- $\ell_i = \log f_{Y_i}\left(y; \lambda_i\right) = y \log \lambda_i - \lambda_i - \log y!$

Agafant: $\boldsymbol{\phi = \Phi = a\left(\phi\right) = 1}$ i $\boldsymbol{\theta_i = \log \lambda_i}$

queda $\ell_i = \frac{y\theta_i - e^{\theta_i}}{1} - \log y! \Rightarrow$

$\boldsymbol{b\left(\theta_i\right) = e^{\theta_i}}$ i $\boldsymbol{c\left(y, \phi\right)} = -\log y! \Rightarrow$

$$\ell_i = \frac{\theta_i y - b\left(\theta_i\right)}{a\left(\phi\right)} + c\left(y, \phi\right)$$

Algunes famílies dels MLGz: $\log\left(f_{Y_i}(y)\right) = \frac{\theta_i y - b(\theta_i)}{a(\phi)} + c(y,\phi)$

Exercici: Altres famílies en que es pot escriure $\ell_i = \frac{\theta_i y - b(\theta_i)}{a(\phi)} + c(y,\phi)$. Comproveu la taula següent:

Família	$f_{Y_i}(y)$	θ_i	ϕ	Φ	$b(\theta_i)$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu_i}{\sigma}\right)^2}$	μ_i	σ^2	ϕ	$\frac{\theta_i^2}{2}$
Poisson	$e^{-\lambda_i}\frac{\lambda_i^y}{y!}$	$\log\lambda_i$	1	1	e^{θ_i}
Binomial *N fix*	$\binom{N}{y} p_i^y \left(1-p_i^{N-y}\right)$	$\log\frac{p_i}{1-p_i}$	1	1	$N\log\left(1+e^{\theta_i}\right)$
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta_i y}$	$\theta_i = -\frac{\beta_i}{\phi}$	α	ϕ^{-1}	$-\log(-\theta_i)$
Inv.Gaussiana	$\sqrt{\frac{\lambda}{2\pi y^3}}\, e^{-\frac{\lambda(y-\nu_i)^2}{2y\nu_i^2}}$	$\frac{-1}{2\nu_i^2}$	λ	ϕ^{-1}	$-\sqrt{-2\theta_i}$

Propietats de les famílies dels GLM

Propietats dels MLGz

A partir de l'expressió com una familia exponencial, es pot deduir l'esperança i la variància de la variable resposta, així com la funció de variància, que ens **indica com canvia la variància en funció del valor esperat**.

Esperança

$$\mu_i = E\left[Y_i|\left(\theta_i, \phi\right)\right] = \boldsymbol{b'\left(\theta_i\right)} \Longleftrightarrow \theta_i = b'^{-1}\left(\mu_i\right) = q\left(\mu_i\right)$$

Variància

$$Var\left(Y_i|\left(\theta_i, \phi\right)\right) = \boldsymbol{\Phi b''\left(\theta_i\right)}$$

Funció de variància

$$\boldsymbol{V\left(\mu_i\right) = b''\left(b'^{-1}\left(\mu_i\right)\right)}$$

Comproveu la taula següent pel càlcul de la funció de variància:

Família	$b(\theta_i)$	$\mu_i = b'(\theta_i)$	$\theta_i = q(\mu_i)$	$b''(\theta_i)$	$V(\mu_i)$
Normal	$\frac{\theta_i^2}{2}$	θ_i	μ_i	1	1
Poisson	e^{θ_i}	e^{θ_i}	$\log \mu_i$	e^{θ_i}	μ_i
Binomial *N fix*	$N \log\left(1 + e^{\theta_i}\right)$	$\frac{Ne^{\theta_i}}{1+e^{\theta_i}}$	$\log\left(\frac{\mu_i}{N-\mu_i}\right)$	$\frac{Ne^{\theta_i}}{\left(1+e^{\theta_i}\right)^2}$	$\mu_i\left(1 - \frac{\mu_i}{N}\right)$
			$\log\left(\frac{p_i}{1-p_i}\right)$		$Np_i(1-p_i)$
Gamma	$-\log(-\theta_i)$	$-\frac{1}{\theta_i}$	$-\frac{1}{\mu_i}$	$\frac{1}{\theta_i^2}$	μ_i^2
Inversa Gaussiana	$-\sqrt{-2\theta_i}$	$\frac{1}{\sqrt{-2\theta_i}}$	$-\frac{1}{2\mu_i^2}$	$\frac{1}{\sqrt{-8\theta_i^3}}$	μ_i^3

En totes les famílies : $\boldsymbol{Var\left(Y_i \middle| (\theta_i, \phi)\right) = \Phi V(\mu_i)}$

Link canònic

Definició

En tots els MLGz, anomenarem **canònic** al link que compleix:

$$q\left(g^{-1}\left(\eta_i\right)\right) = \theta_i = \eta_i \leftrightarrow g^{-1}\left(\eta_i\right) = \mu_i = b'\left(\eta_i\right) \rightarrow$$

$$\eta_i = q\left(\mu_i\right) \Rightarrow g\left(\mu_i\right) = q\left(\mu_i\right)$$

Avantatges del link canònic:

No sempre podem escollir el link canònic, ja que el link l'escollim en funció dels valors esperats del model, però en el cas que el puguem utilitzar:

- L'estimador màxim versemblant $\hat{\beta}$ serà més fàcil de calcular.
- Facilitarà la interpretació del model

Comproveu la taula pel càlcul del link canònic.

Família	$b(\theta_i)$	$\mu_i = b'(\theta_i)$	$\theta_i = q(\mu_i)$	*Link canònic*
Normal	$\frac{\theta_i^2}{2}$	θ_i	μ_i	$\boldsymbol{\mu_i}$
Poisson	e^{θ_i}	e^{θ_i}	$\log \mu_i$	$\boldsymbol{\log \mu_i}$
Binomial *N fix*	$N \log\left(1 + e^{\theta_i}\right)$	$\frac{Ne^{\theta_i}}{1+e^{\theta_i}}$	$\log\left(\frac{\mu_i}{N-\mu_i}\right)$	$\boldsymbol{\log\left(\frac{\mu_i}{N-\mu_i}\right)}$
			$\log\left(\frac{p_i}{1-p_i}\right)$	$\boldsymbol{\log\left(\frac{p_i}{1-p_i}\right)}$
Gamma	$-\log(-\theta_i)$	$-\frac{1}{\theta_i}$	$\frac{1}{\mu_i}$	$\boldsymbol{\mu_i^{-1}}$
Inversa Gaussiana	$-\sqrt{-2\theta_i}$	$\frac{1}{\sqrt{-2\theta_i}}$	$-\frac{1}{2}\frac{1}{\mu_i^2}$	$\boldsymbol{\mu_i^{-2}}$

Plantejament del model, necessitem determinar:

1. **El predictor lineal**, $\eta = X\beta$, de forma anàloga als models lineals.
2. **La funció link**, $g(\mu) = \eta$, segons els casos:
   - Quan hi ha variables explicatives contínues està estretament relacionat amb la funció de regressió que volem modelar.
   - Quan el model no hi ha covariables contínues:
     1. Si és d'un sol factor, no importa la funció link que utilitzem.
     2. Si hi ha diversos factors, la significació de les interaccions pot canviar segons el link $\Rightarrow$ afecta a les simplificacions del model.
     3. En el cas factorial les estimacions $\hat{\mu}_x$ i $\hat{\Phi}$ no depenen del link.
3. **Família de distribucions** (només se'n necessita la funció de variància).
   - Coneixem la família teòricament, o bé, de les dades en veiem la funció de variància que tenen i això determina la distribució.

Estimació del model

Log-versemblança

La log-versemblança per a un conjunt de n observacions és:

$$\ell(\beta) = \sum_{i=1}^{n} \log f(y_i \mid \mathbf{x}_i, \beta),$$

on $f(y_i \mid \mathbf{x}_i, \beta)$ és la funció de densitat o de probabilitat de la variable de resposta Y_i, condicionada a les covariables $\mathbf{x}_i$.

Exemple

En el cas d'un model Poisson, la log-versemblança es pot escriure:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log(\mu_i) - \mu_i \right],$$

on $\mu_i = \exp(\mathbf{x}_i^\top \beta)$.

Estimació per Màxima Versemblança en MLGz

L'**estimació per màxima versemblança (MLE)** consisteix a trobar els valors dels paràmetres β que maximitzen la log-versemblança $\ell(\beta)$:

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta).$$

Passos clau

1. Escriure la log-versemblança $\ell(\beta)$ per la distribució específica de la variable de resposta Y_i.
2. Trobar les β_i resolent el sistema $\frac{\partial \ell(\beta)}{\partial \beta} = 0$.

En molts casos, no hi ha una solució analítica tancada, i es requereixen **mètodes numèrics** com l'algoritme iteratiu de *Fisher scoring* o *Newton-Raphson*.

Ajust del model

Estimarem els paràmetres del model β i Φ en dos passos:

Passos

1. $\hat{\beta}$, estimació per màxima versemblança de β.
   - Per estimar β no es necessita Φ, només la funció de variància.
2. Si Φ és desconegut, obtimdrem $\hat{\Phi}$ fent estimació pel mètode dels moments de Φ, quan ja coneixem $\hat{\beta}$. [*Més endavant, veurem com fer l'estimació*]

Propietats assimptòtiques de l'estimador de β

- **Normal**: $\hat{\beta} \sim \boldsymbol{N(\mu, \sigma)}$
- **NO esbiaixat**: $\boldsymbol{E}\left[\hat{\beta}\right] = \beta$

Quasiversemblança

Models quasiversemblants

- En l'estimació de les $\beta's$, de la família de distribucions només es necessita la **funció de variància**.
- Per tant, si no coneixem la família però sí $V(\mu) \Rightarrow$ podem calcular $\hat{\beta}$ per estimació **quasiversemblant**.

Propietats

- Al no tenir la la distribució, no podem calcular certes coses com $\Pr(Y)$, intervals de predicció, AIC, BIC
- És un mètode no paramètric (com mínims quadrats) que té bones propietats asimptòtiques, com els estimadors màxim-versemblants.

Estimació en Models Quasiversemblants

No es coneix la distribució exacta de la variable de resposta, però es té informació sobre la seva mitjana i variància:

$$\mathbb{E}(Y_i) = \mu_i, \quad \text{Var}(Y_i) = \phi V(\mu_i),$$

- La funció de **quasi-versemblança** és:

$$\ell_q(\mu_i, y_i) = \int \frac{y_i - t}{\phi V(t)}\, dt$$

- Els paràmetres β s'estimen resolent:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \beta} = 0$$

- El paràmetre de dispersió ϕ s'estima posteriorment a partir dels residus.

Take-home messages

- Les diferències fonamentals dels MLGz amb els MLG són:
  - **Funció link** per perfilar la relació funcional entre el predictor lineal i l'esperança.
  - Permeten **distribucions de la familia exponencial** per considerar variàncies no constants.
  - Estimació per màxima versemblança en **MLGz**
- La **funció de Variància** $V(\mu_i)$ determina com canvia la dispersió en funció del valor esperat. Depén de la familia escollida.
- El **paràmetre de dispersió** Φ és comú a totes les dades i, juntament amb la funció de variància, determina la variabilitat de la resposta.