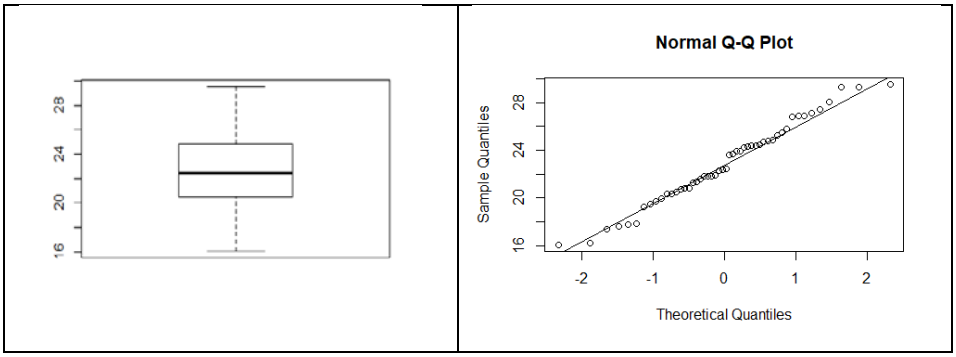
NOM:_____COGNOM:_____

(Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs)

Problema 1 (B4)

Un proveïdor vol estudiar la càrrega de la seva xarxa i per això registra el nombre d'usuaris (en milers de persones) en cinquanta llocs. A partir de l'estudi obté les següents dades:

$$\sum_{i=1}^{50} x_i = 1142'47 \qquad \sum_{i=1}^{50} x_i^2 = 26674'49$$



1. Feu una estimació puntual de l'esperança i de la desviació del nombre d'usuaris (1 punt)

   $\bar{x}$= 22'8494 i s_X= 3'4099

2. A partir dels gràfics argumenteu si podem suposar que les dades segueixen o no una distribució normal. (0'5 punts)

   A partir del boxplot i del qqnorm podem suposar que el nombre d'usuaris segueix una distribució normal.

   En estudis anteriors s'havia obtingut una mitjana de 22.000 usuaris.

En estudis anteriors s'havia obtingut una mitjana de 22.000 usuaris. El proveïdor vol posar a prova si hi ha hagut un augment dels usuaris amb un risc del 5%.

3. a) Indiqueu les hipòtesis, les premisses, la fórmula de l'estadístic i quina és la distribució d'aquest sota la hipòtesi nul·la (0'5 punts)

   $\begin{cases} H_0: \mu_X = 22 \\ H_1: \mu_X > 22 \end{cases}$, test unilateral perquè posem a prova si hi hagut un augment en els usuaris

   La premissa és X~ Normal . L'estadístic és $\hat{t} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$ i $\hat{t}$~ t_{49}

   b) Calculeu el valor de l'estadístic (0'5 punts)

   $\hat{t} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}} = 1'7614$

   c) Representeu gràficament el(s) punt(s) crític(s), les zones d'acceptació i de rebuig i el valor de l'estadístic (1 punt) [Per calcular els punts crítics fes la mitjana entre el valor amb els graus de llibertat immediatament més petit i el valor amb els graus de llibertat immediatament més gran]

   Com que n=50, els graus de llibertat són 49 que no estan a la taula. Per això busquem a les taules P(T < k1)= 0.95 (test unilateral) amb T~t_{40} i trobem que k1=1'684. I després P(T < k2)= 0.95 amb T~t_{60} i trobem k2=1'671

   Per trobar el punt crítics fem el que ens diu l'enunciat: (1'684+ 1'671)/2 i obtenim: 1'6775

   Els punts crítics és 1'6775. La zona d'acceptació és (-∞,1'6775). La zona de rebuig és (1'6775, +∞)

   Gràficament.

   d) A partir de l'estudi i del càlculs realitzats, interpreteu els resultats de la prova d'hipòtesi. (0'5 punts)
   El valor de l'estadístic 1'7614 pertany a la zona de rebuig, per tant podem concloure que tenim evidències que el nombre d'usuaris ha augmentat.

El proveïdor realitza un segon estudi per estudiar la dispersió del nombre d'usuaris connectats. En aquest cas registra el nombre d'usuaris (en milers de persones) en trenta llocs, obtenint les següents dades que segueixen una distribució normal:

$$\sum_{i=1}^{30} x_i = 553'3 \qquad \sum_{i=1}^{30} x_i^2 = 10565'65$$

4. a) Calculeu l'interval de confiança per la desviació poblacional amb una confiança del 95% (1 punt)

La premissa és X~ Normal

s^2 = 12.4467, n= 30. Amb les taules trobem que: $\chi^2_{29,1-\frac{\alpha}{2}}$ = 45'722 i $\chi^2_{29,\frac{\alpha}{2}}$ = 16'047

$\left(\frac{s^2\cdot(n-1)}{\chi^2_{29,1-\frac{\alpha}{2}}}, \frac{s^2\cdot(n-1)}{\chi^2_{29,\frac{\alpha}{2}}}\right)$ = (7'895, 22'493). I per tant, per la desviació poblacional: (2'810, 4'743)

b) Interpreteu el resultat anterior (0'5 punts)

Amb un 95% de confiança el valor de la desviació del nombre d'usuaris connectats en milers de persones està entre 2'810 i 4'743.

S'estudia si per nombre d'usuaris s'assoleix l'amortització de la xarxa, per fer-ho es mira la variable dicotòmica d'amortitzat o no amortitzat, obtenint que la xarxa està amortitzada en 37 llocs dels 50 que hem avaluat.

5. Poseu a prova si el valor esperat de la proporció d'amortitzats és del 80% o no. Amb un risc del 5%:

a) Indiqueu l'estimació puntual de la proporció d'amortitzats (0'5 punts)

P = #A / 50 =37/50

b) Indiqueu les hipòtesis, premisses, la fórmula de l'estadístic i dir quina distribució segueix sota la hipòtesis nul·la (1 punt)

$\begin{cases} H_0: \pi_A = 0'8 \\ H_1: \pi_A \neq 0'8 \end{cases}$, test bilateral

Premisses: $(1-\pi_0)n \geq 5$ i $\pi_0 \cdot n \geq 5$. Tenim que: 0'2 · 50 = 10 ≥ 5 i 0'8 · 50 = 40 ≥ 5

Fórmula de l'estadístic: $\hat{z} = \frac{p-p_0}{\sqrt{\frac{p_0\cdot(1-p_0)}{n}}}$, $\hat{z}$ ~ N(0,1)

c) Calculeu del valor de l'estadístic: (0'5 punts)

$\hat{z} = \frac{p-p_0}{\sqrt{\frac{p_0\cdot(1-p_0)}{n}}} = \frac{0'74-0'8}{\sqrt{\frac{0'8\cdot0'2}{50}}}$ = -1'0607

d) Realitzeu la representació gràfica de l'estadístic amb el/s punt/s crític/s i les zones d'acceptació i rebuig (0'5 punts)

Punts crítics -1.96 i 1.96
Zona rebuig: (-∞,-1'96) ∪ (1'96, +∞)
Zona acceptació: (-1'96, 1'96)
Representació gràfica

e) Calculeu l'interval de confiança pel valor esperat de la proporció de punts amortitzats i interpreteu-lo (1 punt)

$(P - Z_{1-\alpha/2} \cdot \sqrt{\frac{P\cdot(1-P)}{n}}, P + Z_{1-\alpha/2} \cdot \sqrt{\frac{P\cdot(1-P)}{n}})$ = (0'6184, 0'8615)

Amb un 95% de confiança el valor de la proporció d'amortització de la xarxa està entre 0'6184 i 0'8615

f) En funció dels dos apartats anteriors, a quina conclusió arribeu sobre la prova d'hipòtesi. Interpreteu els resultats. (1 punt)

El valor de l'estadístic -1'07 està dins de la regió d'acceptació (apartat d) i el valor 0'8 pertany a l'interval de confiança (apartat e). Per tant, ambdós resultats ens indiquen que no hi ha evidència per rebutjar que la proporció d'amortització de la xarxa sigui de 0'8.

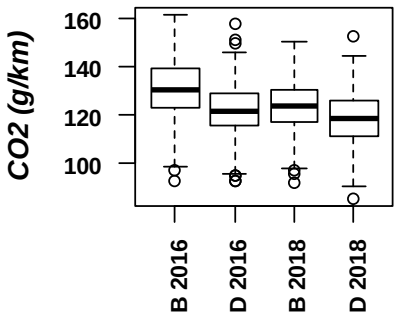
NOM:_____COGNOMS:_____

(*Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs. Totes les preguntes valen igual*)

Problema 2 (B5)

En els darrers anys, s'estan impulsant mesures per reduir la contaminació dels cotxes amb mesures com la restricció d'entrada a les grans ciutats de cotxes dièsel o l'encariment de determinants combustibles. Per avaluar l'impacte d'aquestes mesures, s'ha fet un estudi de les emissions de CO_2 dels vehicles de benzina i dièsel abans (2016) i després (2018) de l'entrada en vigor d'aquestes mesures en una mostra de 100 cotxes cada any (no han entrat en el mostreig ni els vehicles híbrids ni elèctrics). El resum de la descriptiva la tens a la següent taula i figura.

				Emissions CO_2 (g/km)		
Any	**Tipus**	**N**	**n**	**Mitjana**	**Desviació típica**	**Desviació típica global**
2016	Benzina (B)	100	40	130	12	13
	Dièsel (D)		60	122	10	
2018	Benzina (B)	100	52	124	10	11
	Dièsel (D)		48	118	10	



1. Fes un contrast d'hipòtesi (α=0.05) per determinar si la proporció poblacional de vehicles dièsel ha decrescut de 2016 a 2018.

Contrast d'hipòtesi (0.5 punts)	**Càlcul de l'estadístic i la seva distribució sota H_0 (1 punt)**
Especifica si és unilateral o bilateral $\begin{cases} H_0: \pi_{2016} = \pi_{2018} \\ H_1: \pi_{2016} > \pi_{2018} \end{cases}$ unilateral	$p_{2016} = \frac{60}{100} = 0.60 \quad p_{2018} = \frac{48}{100} = 0.48 \quad p = \frac{100 \cdot 60 + 100 \cdot 48}{200} = 0.54$ $Z = \frac{p_{2016} - p_{2018}}{\sqrt{\frac{p \cdot (1-p)}{n} + \frac{p \cdot (1-p)}{n}}} \sim N(0,1)$ sota H_0 $Z = \frac{0.60 - 0.48}{\sqrt{\frac{2 \cdot 0.54 \cdot 0.46}{100}}} = 1.70$
Punt crític (α=0.05) i p-valor (1 punt)	**Conclusió (0.5 punts)**
Punt crític = 1.645 (taules) P-valor = 0.046 (taules)	Com que l'estadístic està a la regió de rebuig d'H_0 i el p-valor és inferior a 0.05, rebutgem H_0 i assumim que la proporció de vehicles dièsel ha decrescut de 2016 a 2018.

2. L'altre pregunta que és fa l'associació és si el consum mitjà de CO_2 a l'any 2018 és inferior en els vehicles dièsel que en els de benzina. Assumint normalitat i variàncies poblacionals iguals, fes un contrast d'hipòtesi (α=0.05) per discernir-ho.

Contrast d'hipòtesi (0.5 punts)	**Càlcul de l'estadístic i distribució sota H_0 (1 punts)**
Especifica si és unilateral o bilateral $\begin{cases} H_0: \mu_D = \mu_B \\ H_1: \mu_D < \mu_B \end{cases}$ Unilateral	$s^2_{pooled} = \frac{(n_B - 1) \cdot S_B^2 + (n_D - 1) \cdot S_D^2}{(n_B + n_D - 2)} = \frac{51 \cdot 10^2 + 47 \cdot 10^2}{52 + 48 - 2} = 100$ $t = \frac{\bar{y}_B - \bar{y}_D}{\sqrt{\frac{S^2_{pooled}}{n_B} + \frac{S^2_{pooled}}{n_D}}} \sim t_{98}$ sota H_0 $t = \frac{124 - 118}{\sqrt{\frac{100}{52} + \frac{100}{48}}} = 3.00$

Punt crític (α=0.05) i fites pel p-valor † (1 punt)	**Conclusió (0.5 punts)**
† (una fita superior i una inferior) Punt crític = 1.66 (taules interpolant entre 1.658 i 1.671) P-valor = 0.0013 Emprant taules: 0.001 < p-valor < 0.0025	Com que l'estadístic està a la regió de rebuig d'H_0 i el p-valor és inferior a 0.05, rebutgem H_0 i es conclou que el consum de CO_2 dels vehicles dièsel és inferior als de vehicles de benzina.

3. Fes una estimació de l'interval del 90% de confiança per a la variació del consum mitjà de CO_2 de 2016 a 2018 per Km i cotxe. **Pista**: *Per trobar les mitjanes respectives de les emissions dels anys 2016 i 2018 has de fer la mitjana ponderada dels cotxes de benzina i dièsel per cada any* (2.5 punts)

Primer, es calcula la mitjana de les emissions al 2016 i 2018

$$\bar{y}_{2016} = \frac{130 \cdot 40 + 122 \cdot 60}{100} = 125.2$$

$$\bar{y}_{2018} = \frac{124 \cdot 52 + 118 \cdot 48}{100} = 121.12$$

A continuació, es calcula el IC90% per la diferència de mitjanes:

$$s^2_{pooled} = \frac{(n_{2018} - 1) \cdot S^2_{2018} + (n_{2016} - 1) \cdot S^2_{2016}}{(n_{2018} + n_{2016} - 2)} = \frac{11^2 + 13^2}{2} = 145$$

$$IC(90\%, \mu_{2018} - \mu_{2016}) = (\bar{y}_{2018} - \bar{y}_{2016}) \mp z_{0.95} \cdot s_{pool} \cdot \sqrt{\frac{1}{n_{2018}} + \frac{1}{n_{2016}}} = (121.12 - 125.2) \mp 1.65 \cdot 12.04 \cdot \sqrt{\frac{2}{100}}$$

$$= [-6.88, -1.28]$$

4. Contesta de forma argumentada però el més breument possible a les 3 qüestions següents. (0.5 punts cada qüestió)

a. Creus que la mesura adoptada de restringir l'entrada de vehicles dièsel ha tingut impacte en la reducció d'emissions de CO_2 per part dels vehicles? Per què?

El IC90% no conté el zero i, per tant, amb una confiança del 90% podem dir que les emissions de CO_2 en els 2 períodes no són les mateixes i, de fet, han decrescut. No obstant aquest descens no es pot atribuir a la mesura de permetre l'entrada a menys cotxes dièsel ja que aquests tenen unes emissions inferiors en CO_2. El canvi és degut a les més baixes emissions de tots els tipus de vehicles (benzina i dièsel) en el 2018.

b. Amb la informació que disposes, creus que les premisses de Normalitat i Variància constant assumides en el apartat 2 són raonables? Per què?

Sí. Veient els boxplots de l'enunciat existeix una simetria en les 2 distribucions que no descarta que siguin dades que segueixen una distribució Normal. A més, visualment, s'observa que mitjana i mediana són molt similars. En quant a la variància constant, l'estimació puntual de la desviació és la mateixa en les dues mostres i l'amplada dels boxplots és molt similar.

c. Per què les desviacions típiques globals tant a l'any 2016 com 2018 són més grans que les desviacions típiques de cada tipus de cotxe per separat?

Perquè la desviació global prové de dues distribucions que probablement tenen mitjanes diferents i per tant, al contemplar tots els tipus de cotxes, tindrà una major variabilitat.

NOM:_____COGNOM:_____

(***Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs***)

Problema 3 (B6)

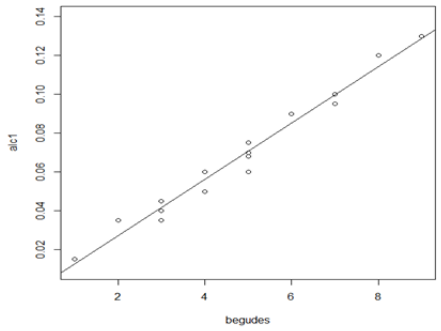
Volem comparar l'efecte de dos tipus de beguda en la quantitat d'alcohol en sang mitjançant una regressió lineal depenent de la quantitat de beguda. El nombre d'unitats de beguda ingerida per 16 persones pren els valors 1,2,3,3,3,4,4,5,5,5,5,6,7,7,8,9. Les respostes "alc1" i "alc2" són, respectivament, els valors d'alcohol en sang després d'ingerir un dels dos tipus de beguda; es mesuren en dos grups independents de 16 persones cadascun on cada individu té assignat el nombre i tipus de beguda.

alc1 <- c(0.015,0.035,0.04,0.045,0.035,0.06,0.05,0.07,0.06,0.068,0.075,0.09,0.095,0.10,0.12,0.13)

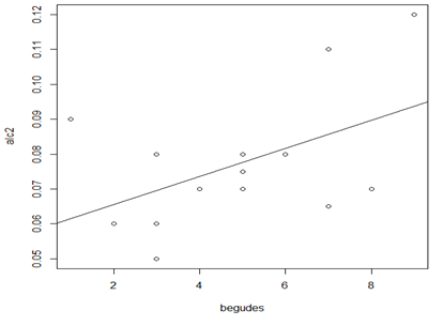
alc2 <- c(0.09,0.06,0.05,0.06,0.08,0.07,0.07,0.08,0.07,0.075,0.08,0.08,0.11,0.065,0.07,0.12)

Alguns dels resultats són:

mean(alc1)=0.068 sd(alc1)=0.032 mean(begudes)=4.81 sd(begudes)=2.20 mean(alc2)=0.077 sd(alc2)=0.018



cov(begudes,alc1) = 0.07



cov(begudes,alc2) = 0.02

```
lm(formula = alc1 ~ begudes)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0016920  0.0033004  -0.513    0.616
begudes      0.0144814  ?x+??++??   23.088 1.52e-12 ***
Residual standard error: 0.005338 on 14 degrees of freedom
Multiple R-squared:  0.9744
```

Pel cas de l'alcohol segons el nombre de begudes tipus 1:

- indiqueu la recta ajustada i interpreteu què n'indiquen els coeficients (1 punt)

alc1 = -0.0016920 + 0.0144814 * begudes

-0.0016920 és la ordenada a l'origen i indica el nivell d'alcohol en sang amb 0 begudes seguint el model de la recta ajustada

0.0144814 és el pendent i indica que amb una beguda més l'alcohol en sang augmenta uns 0.0145

- a les proves d'hipòtesis per posar a prova si la ordenada a l'origen i el pendent es poden considerar nul·les o no, indiqueu les hipòtesis, resultats, interpretació i conclusió (2 punts)

H_0: $\beta_0=0$ amb estadístic -0.513 i p_value 0.616

H_1: $\beta_0\neq 0$

Com p_value és més gran que un risc de 0.05 (o estadístic=-0.513 dins zona d'acceptació de punt crítics $\pm t_{14,0.975}=2.145$) res s'oposa a acceptar H_0: la ordenada a l'orígen és 0 (a 0 begudes l'alcohol en sang és 0)

H_0: $\beta_1=0$ amb estadístic 23.088 i p_value pràcticament 0

H_1: $\beta_1\neq 0$

Com p_value és més petit que un risc de 0.05 (o estadístic 23.088 fora zona d'acceptació de punt crítics $\pm t_{14,0.975}=2.145$) no és raonable acceptar H_0. I concloure que el pendent és diferent de 0, per tant hi ha relació positiva entre una beguda més i l'alcohol en sang

- calculeu i interpreteu un interval de confiança al 95% pel pendent (1 punt)

$b_1 \pm t_{14,0.975}\, S_{b1}$

$= 0.0145 \pm 2.145\ (0.0145/23.088) = 0.0145 \pm 0.0013 = [\ 0.0132\ ,\ 0.0158\ ]$

Amb un 95% de confiança el valor del pendent (és a dir de l'augment d'alcohol en sang per cada unitat més de beguda de tipus 1) estarà entre 0.0132 i 0.0158

Pel cas de l'alcohol segons el nombre de begudes tipus 2:
- calculeu la recta ajustada (1 punt)

$b_1 = 0.02 / 2.2^2 = 0.004$

$b_0 = 0.077 – (0.004*4.81) = 0.058$ per tant alc2 = 0.058 + 0.004 * begudes

- calculeu el coeficient de correlació, el de determinació i la desviació residual. Comparant-los amb els del cas anterior comenteu què indiquen (2 punts)

$r = 0.02 /(2.2*0.018) = 0.5$ i per tant $R^2 = (r)^2 = 0.25$

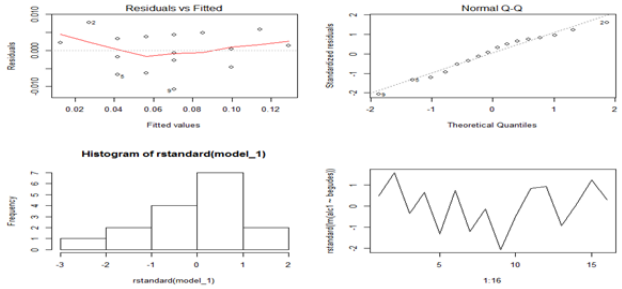
$s = sqrt(s^2) = sqrt(15*0.018^2*(1-0.25)/14) = sqrt(0.00026) = 0.016$

El coeficient de determinació d'un 25% és baix comparat amb el del cas anterior que era de més d'un 97% (0.9744) com a % d'explicació de la variabilitat d'alcohol explicat pel nombre de begudes. I també la correlació de 0.5 és molt inferior a l'anterior (sqrt(0.9744)=0.99) pràcticament al màxim de positiva. Indica que un (simple) model lineal pot explicar un 25% de la variabilitat.

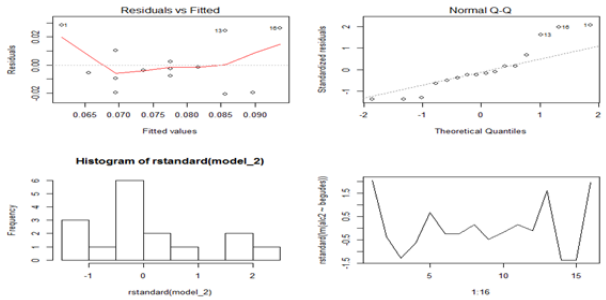
La desviació residual de 0.016 és molt superior a la del cas anterior (0.005338), per tant els punts s'ajusten molt més a una recta en el cas anterior que en aquest

- Compareu les premisses en els dos models (indiqueu els gràfics d'on es dedueixen) (2 punts)

alc1 ~ begudes :



alc2 ~ begudes



En general les premisses es compleixen més en el cas de alc1 que en el alc2. En els gràfics de la pàgina anterior ja es veu que la linealitat és millor en el cas de alc1, i en el gràfic Residuals.vs.fitted a alc 1 la línia vermella és més plana indicant millor ajustament dels punts a la recta ajustada

La normalitat en el cas alc1 tant al NormalQ-Q (punts ben alineats) com a l'histograma (forma de campana tot i que una mica aplanada per l'esquerra) es compleixen, mentre que per alc2 la normalitat falla sobretot a l'extrem superior estant menys alineats i l'histograma més cua per la dreta.

La homocedasticitat es compleix en el cas alc1 doncs no hi ha zones de més i menys variabilitat respecte la recta. Però en el cas alc 2 és més heterocedàstic doncs la variabilitat és més gran en els dos extrems. Es veu a gràfic Residuals.vs.fitted i a l'últim

La independència es compleix en el cas alc1 i també força en el cas alc2 doncs en els gràfics Residuals.vs.fitted i últim els punts no segueixen cap patró que permeti relacionar-los indicant alguna dependència entre les observacions.

- Feu una valoració global comparant els dos models de regressió indicant què ens diuen en cada cas sobre la relació entre nombre de begudes i alcohol en sang. (1 punt)

En el cas alc1 hi ha una bona relació lineal i bon ajust a la recta (amb correlació i coeficient de determinació alts, i desviació residual petita). Indica que hi ha una forta relació positiva complint les premisses, i que una beguda més del tipus 1 preveu un augment de l'alcohol en sang d'entre 0.0132 i 0.0158.

En el cas alc2 la relació lineal i per tant la recta ajustada no són bones (correlació i coeficient de determinació baixos, i desviació residual gran). Indica que no hi ha relació lineal entre una beguda més del tipus 2 i un augment de l'alcohol en sang. En aquest cas possiblement el pendent no és raonablement diferent de 0, indicant que la recta és plana tot i que el gràfic inicial por semblar que té pendent positiu degut a l'escalat de l'eix d'ordenades. Potser convindria estudiar un model més sofisticat.