NOM:_____

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

Problema 1 (B4). Per comparar M i N, els dos proveïdors de pel·lícules més populars, hem calculat la diferència D entre els seus temps de baixada de 120 films escollits a l'atzar de la llista de la CSMS. Encara que estava planificat inicialment comparar els temps en tots els casos conjuntament, els resultats mostren 2 grups clarament diferenciats: 100 casos amb valors al voltant de la igualtat (Empat E: n_E=100, mitjana$_E$=0, sd$_E$=4); i 20 casos amb clara avantatge per N (G=Guanya N, n_G=20, mitjana$_G$=40, sd$_G$=12). Dins de cada grup, D ~ N. *(Totes valen igual)*

Dades auxiliars:	> qt(0.975,99) [1] 1.984217 > qt(0.95,19) [1] 1.729133 > pt(8,19) [1] 0.9999999	> qchisq(0.05,20) [1] 10.85081 > qchisq(0.95,20) [1] 31.41043 > qchisq(0.025,20) [1] 9.590777	> qchisq(0.975,20) [1] 34.16961 > qchisq(0.05,19) [1] 10.11701 > qchisq(0.95,19) [1] 30.14353	> qchisq(0.025,19) [1] 8.906516 > qchisq(0.975,19) [1] 32.85233
Usage	pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE) qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)			

En general, errors de detall (taules, càlcul, ...) resten ½ punt.
Respostes de qualitat (preguntes 5, 7, 9 i 10) donen punts extra.

1) Estimeu per interval (99%) i interpreteu la proporció de casos en que guanya N

p=20/120 ≈ 0.167
IC (π, 99%) = p + c(–1, 1)*2.576*sqrt(0.167*0.833/120) ≈ 0.167 ± 0.088 ≈ **(0.079, 0.254)**
Amb una confiança del 99%, aquesta proporció es algun valor entre el 8% i el 25%.

[O bé: IC (π, 99%) = p + c(–1, 1)*2.576*sqrt(0.5*0.5/120) ≈ ≈ 0.167 ± 0.118 ≈ (0.049, 0.284)]
[No demanat: les condicions de mostra gran es compleixen ja que 0.049*120≈5.89 > 5]

Estimeu per interval (95%) el valor de la esperança de D en els casos 'E' d'empat.
2) Estadístic, distribució, premisses,

$$\hat{t} = \frac{(\bar{d} - \mu_D)}{s_D / \sqrt{n}} \to t_{n-1}$$

σ desconeguda i D ~ N

[O bé $\hat{t} = \frac{(\bar{d} - \mu_D)}{s_D / \sqrt{n}} \to z$ σ desconeguda i n≥100]

3) càlcul, interpretació..

IC (μ, 0.95) = $\bar{d}$ ± t $_{n-1,1-\alpha/2}$ $s_D/\sqrt{n}$ = 0 ± 1.984· 4/√100 ≈ 0 ± 0.794 ≈ **(-0.794, 0.794)**
Amb una confiança del 95%, E(D) és algun valor entre -0.8 i +0.8.

[O bé: IC (μ, 0.95) = $\bar{d}$ ± $Z_{1-\alpha/2}$ $s_D/\sqrt{n}$ = 0 ± 1.96· 4/√100 ≈ 0 ± 0.784 ≈ (-0.784, 0.784)]

4) Estimeu per interval (90%) el valor poblacional de la s_d en els casos 'G' on guanya N. Estadístic, distribució, premisses,

$(n-1)*S^2/\sigma^2 \sim \chi^2_{n-1}$ σ desconeguda i D ~ N

$$IC(\sigma^2, 0.90) = \left(\frac{s^2(n-1)}{\chi^2_{19,0.95}}, \frac{s^2(n-1)}{\chi^2_{19,0.05}} \right) \approx \left(\frac{144*19}{30.143}, \frac{144*19}{10.117} \right) \approx (90.77 , 270.43)$$

IC(σ,0.9) ≈ (9.5 , 16.4)

Si només σ^2, ½ punt

5) Podeu demostrar (amb alfa=0.05 unilateral) que és cert que N guanya en els casos G?

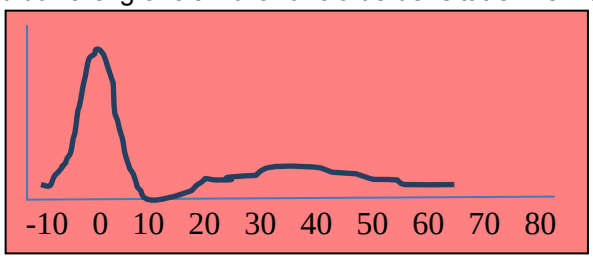
$$\hat{t} = (40-0)/(12/\sqrt{20}) \approx 14.907$$

Sí, ja que l'estadístic **'t' = 14.907 >1,729133 = $t_{19,0.95}$**

[També: De fet, les dades auxiliars mostren que amb 19 gdl, la prob acumulada de la t per al valor 8 ja és major del 0.999999. Per tant, al ser unilateral, P<0.0001]

També: $IC_{BIL}90\%$: $40 \pm 1.73 \cdot 12/\sqrt{20} \approx [35.4;44.6]$ o, millor, **$IC_{UNI}95\% \approx [35.4; \infty]$**

6) Creieu que el model Normal podria ser una bona aproximació per la distribució conjunta dels 120 casos? (Per ajudar-te, dibuixa el gràfic amb la funció de densitat en l'eix d'ordenades i el valors de D en el d'abscisses):



No, encara que dins els grups sí. La distribució conjunta té dos 'pics' i la N no podria ser una bona aproximació.

7) Quant valen els estadístics mitjana i variància del 120 casos junts?

E: $\Sigma d = 0*100 = 0$; $\Sigma d^2 = 16*99 + 0\text{^}2/100 = 1584$

G: $\Sigma d = 40*20 = 800$; $\Sigma d^2 = 144*19 + 800\text{^}2/20 = 34736$

Tots: $\Sigma d = 0+800 = 800$; $\Sigma d^2 = 1584 + 34736 = 36320$

$\bar{d} = 800/120 \approx$ **6.667** $S^2 = (36320 - 800^2/120)/119 \approx$ **260.4**; $S \approx$ **16.14**

Atenció: la 'pooled' és incorrecta ja que estima la dispersió respecte a les mitjanes de cada subgrup (0 i 40), en lloc de respecte a la mitjana global (6.67).

8) Posa a prova la hipòtesi d'igualtat (amb alfa=0.05 bilateral) en els 120 casos conjuntament. [Si no has contestat la pregunta anterior, utilitza mitjana=5 i sd=15.]. Estadístic, premisses, càlcul i interpretació.

$$\hat{t} = \frac{(\bar{d} - \mu_D)}{s_D/\sqrt{n}} \rightarrow z$$

σ desconeguda i n≥100

$$\hat{t} = (6.667-0)/(16.14/\sqrt{120}) \approx 6.667/1.473 \approx 4.526$$

$$O: \hat{t} = (5-0)/(15/\sqrt{120}) \approx 3.65$$

Sí que rebutgem la hipòtesi d'igualtat, ja que l'estadístic 't' = 4.526 >1,96 = $Z_{0.975}$

[O bé: IC (μ, 0.95) = $\bar{d} \pm Z_{1-\alpha/2}\ s_D/\sqrt{n} = 6.667 \pm 1.96 \cdot 16.14/\sqrt{120} \approx$
$\approx 6.667 \pm 1.96*1.473 \approx 6.667 \pm 2.887 \approx [3.779, 9.554]$

També: $5 \pm 1.96 \cdot 15/\sqrt{120} \approx [2.3, 7.7]$

9) Faci una interpretació global

D'acord amb la hipòtesi prèvia, hem demostrat (punt 8) que N és més ràpid. Ara bé, els resultats apunten a que aquest avantatge no és comú en totes les execucions: sembla que es concentra en una cada 6 execucions (IC99% de 8 a 25%) a on hi ha el guany.

10) Vostè treballa per N i vol fer públics els resultats. Però el revisor de una revista informàtica ha dit que els subgrups E i G han estat suggerits pels resultats i per tant algunes hipòtesis no eren prèvies a les dades. A més a més, la seva filiació deixa clar cert conflicte de interessos. Segons el revisor aquests arguments fan que les troballes no tinguin cap valor. Redacti una resposta al revisor.

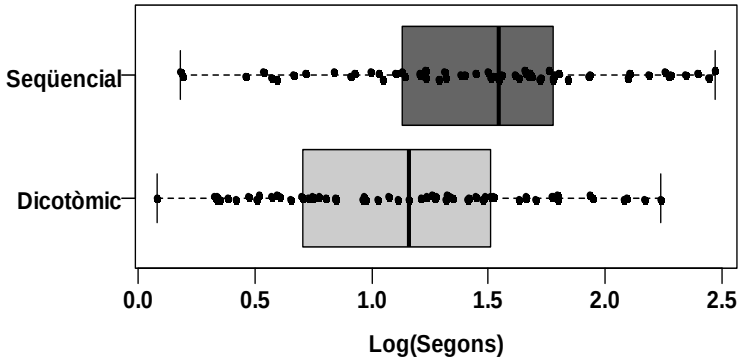
a) És cert, per poder ser contrastades, les hipòtesis han de ser prèvies a les dades: com l'anàlisi fins al punt 5 ha estat suggerit per les dades, han de interpretar-se com a temptatius o suggeridors d'hipòtesis.
   És mol important però, que els resultats son consistents amb l'anàlisi (hipòtesis) pre-planificat del punt 8 amb totes les dades.
b) Conflicte d'interessos: Quelcom semblant a "Jo no he pogut interferir en els resultats, ja que les dades i la seva obtenció han estat **auditades per l'inspector** extern XXX (el seu informe està a disposició). L'objectiu i els mètodes varen ser enviats prèviament a XXX, qui garanteix que l'anàlisi ha seguit el protocol quan així ho diu."

NOM:_____

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

Problema 2 (B5). Es vol comparar la velocitat de dos algoritmes de cerca, *Seqüencial* i *Dicotòmic*, en vectors prèviament endreçats. Es generen 122 vectors endreçats i s'envien per fer una cerca a un dels dos algoritmes de forma aleatòria. Com que el temps de la cerca té una distribució asimètrica amb cua cap a la dreta, treballem amb els logaritmes dels temps. Siguin *X* el logaritme del temps que triga el algorisme *Seqüencial* i *Y* el del *Dicotòmic*. Els estadístics descriptius per a cada mostra estan a la següent taula. A més a més, es mostra el boxplot dels logaritmes dels temps:

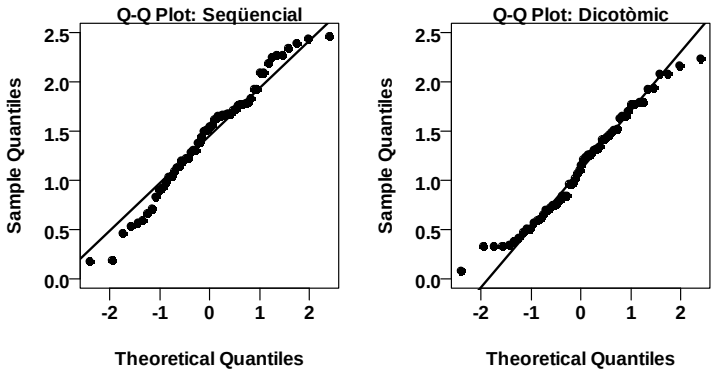
	N	Mitjana	Desv. est	Mediana	Mínim	Màxim
Seqüencial	61	1.47	0.56	1.55	0.18	2.47
Dicotòmic	61	1.14	0.54	1.16	0.08	2.24



1) (1 punt). Es tracta de dues mostres independents o aparellades? Raoneu la resposta.

Independents, ja que per disseny no hi ha cap connexió entre l'observació ièssima de S i la ièssima de D; per la qual cosa no hi haurà relació entre S i D

2) (1 punt). Veient els següents Q-Q plots, sembla raonable suposar que les variables X i Y segueixen una distribució normal? Raoneu la resposta.



Sí, totes dues distribucions estan prou a prop de la línia recta que indica igualtat entre els percentils observats i els que s'haurien d'observar en cas d'una Normal.

3) (2 punts). Estudiem primer la variabilitat dels (logaritmes dels) temps d'ambdós algoritmes. Es pot suposar que són iguals? Responeu aquesta pregunta plantejant i resolent la hipòtesi adient (amb α = 0.1) i explicant quines són les premisses que cal fer per realitzar aquesta prova.

Premisses: mostres aleatòria, independents i variables provinents d'una distribució Normal

$$\begin{cases} H_0: \sigma_S = \sigma_D \\ H_1: \sigma_S \neq \sigma_D \end{cases}$$

$$F = \frac{S^2_{Major}}{S^2_{Menor}} = \frac{0.56^2}{0.54^2} = 1.075 < F_{60,60,0.95} = 1.53$$

No hi ha res que s'oposi a que les variàncies poblacionals siguin iguals

4) (3 punts). Per saber si es pot suposar que hi ha diferències entre ambdós algoritmes, es vol plantejar una prova d'hipòtesi d'igualtat de mitjanes: $H_0: \mu_S = \mu_D$ vs. $H_1: \mu_S \neq \mu_D$ (nivell de significació del 5%)

a) Calcula la variància conjunta (pooled variance) i el valor de l'estadístic (Desenvolupa els càlculs).

$$S^2 = \frac{(n_D - 1)\cdot S_D^2 + (n_S - 1)\cdot S_S^2}{n_D + n_S - 2} = \frac{(61-1)\cdot 0.54^2 + (61-1)\cdot 0.56^2}{61+61-2} = 0.3026$$

$$t = \frac{\bar{x}_S - \bar{x}_D}{S\sqrt{1/n_S + 1/n_D}} = \frac{1.47 - 1.14}{\sqrt{0.3026 \cdot \left(1/61 + 1/61\right)}} = 3.313$$

b) Digues quin és el punt crític amb un 5% de significació i treu conclusions sobre la prova d'hipòtesi.

$t = 3.313 > t_{120,0.975} = 1.98$. Rebutgem la hipòtesi nul·la de que les mitjanes siguin iguals

c) Calcula l'interval de confiança del 90% per a la diferència de mitjanes dels logaritmes i interpreta'l.

$$IC(\mu_S - \mu_D, 90\%) = (\bar{x}_S - \bar{x}_D) \mp t_{120,0.95}\cdot S\sqrt{\frac{1}{n_S}+\frac{1}{n_D}} = 0.33 \mp 1.658 \cdot 0.550 \cdot \sqrt{\frac{1}{61}+\frac{1}{61}} = [0.1648, 0.4951]$$

Amb un 90% de confiança, la diferència de mitjanes dels logaritmes poblacionals es troba entre 0.16 i 0.50. Per tant, amb un 90% de confiança, el quocient entre les mitjanes dels temps poblacionals està entre 1.18 ($e^{0.1648}$) i 1.64 ($e^{0.4951}$). L'algoritme dicotòmic és entre un 18 i un 64% més ràpid.

5) (3 punts). A l'hora d'implementar els algoritmes s'han programat de manera que si l'execució triga més de 5 segons s'avorta l'execució i dóna un missatge d'error. Volem comparar si la probabilitat d'error en ambdós algoritmes és pot considerar la mateixa. La taula de contingència de la dreta resumeix els resultats obtinguts:

	Temps < 5 s.	Temps > 5 s.
Seqüencial	32	29
Dicotòmic	47	14

(Aquest problema es podia haver resolt amb una comparació de dues proporcions o amb una prova de chi-quadrat, arribant en tots 2 casos a la mateixa conclusió. Aquí és mostra la resolució pel cas de la chi-quadrat)

a) Planteja la prova d'hipòtesi i especifica si és unilateral o bilateral. Quina és l'expressió de l'estadístic? Quina és la distribució de l'estadístic i sota quines premisses?

$H_0: P(E|S) = P(E|D)$ vs. $H_1: P(E|S) \neq P(E|D)$. És una prova bilateral

$$\chi^2 = \sum_{i=1}^{4}\frac{(f_i - e_i)^2}{e_i}$$

La distribució de l'estadístic és una $\chi^2_{(f-1)(c-1)} = \chi^2_1$ sota la premissa de mostra prou gran (eij > 5 $\forall$i,j) i m.a.s independent

b) Quant val l'estadístic? (Desenvolupa els càlculs)

e_{ij}		
	$\frac{61\cdot 79}{122} = 39.5$	$\frac{61\cdot 43}{122} = 21.5$
	$\frac{61\cdot 79}{122} = 39.5$	$\frac{61\cdot 43}{122} = 21.5$

$$\chi^2 = \frac{(32-39.5)^2}{39.5} + \frac{(47-39.5)^2}{39.5} + \frac{(29-21.5)^2}{21.5} + \frac{(14-21.5)^2}{21.5} = 8.08$$

c) Podem rebutjar la hipòtesi nul·la amb un grau de significació del 5%?

Com $\chi^2 = 8.09 > \chi^2_{1,0.95} = 3.84$, rebutgem la hipòtesi nul·la d'homogeneïtat.

La copisteria CUT&PASTE amb la intenció de verificar la eficiència de les seves fotocopiadores, analitza els temps d'impressió (en segons) en funció de la grandària del fitxer (en kb) i això ho fa per 100 fitxers diferents. La representació gràfica d'aquestes dades és a la figura 1

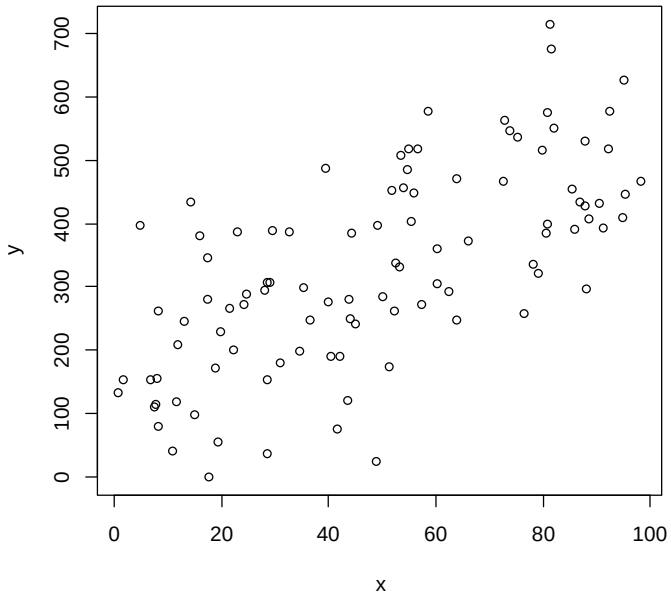


Figura 1

a) (3 punts)

Sabent que X=grandària del fitxer (en kb) i Y=temps d'impressió (en segons), en R s'han calculat els següents estadístics:

Mean(X)= 48.76 kb Var(X)= 796.76 kb^2

Mean(Y)= 330.44 seg Var(Y)= 24769.21 seg^2

i la correlació entre (X,Y), r=0.68

Estimeu els coeficients de la recta de regressió b_0 i b_1 i dibuixeu la recta de regressió damunt de la figura 1

Solució

$$b_1 = r\frac{S_y}{S_x} = 0.68\sqrt{\frac{24769.21}{796.76}} = 3.79$$

$$b_0 = \bar{y} - b_1\bar{x} = 330.44 - 3.79 * 48.76 = 145.64$$

I per tant, la recta de regressió estimada és:

$$\hat{Y} = 145.64 + 3.79X$$

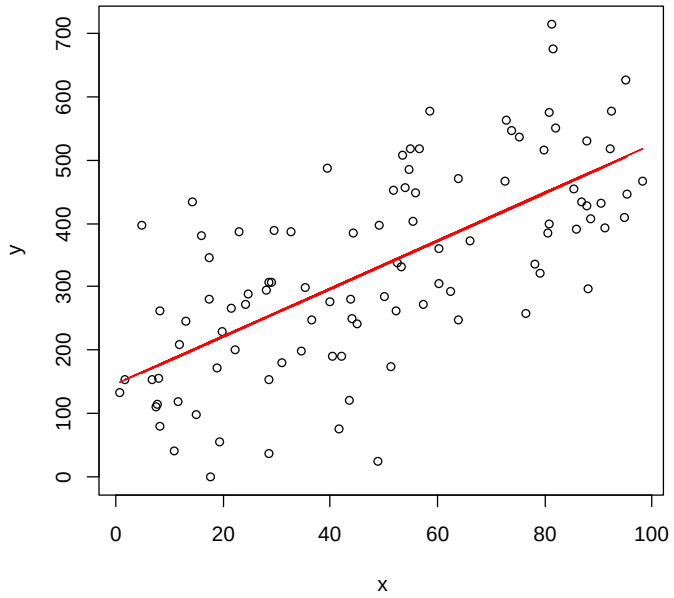


Figura 1, en vermell, la recta de regressió estimada

b) (2 punts)

Quant val la desviació tipus residual? Quina interpretació te en el model de regressió? Raoneu la vostra resposta.

$$S^2 = \frac{(n-1)S_Y^2(1-r^2)}{n-2} = \frac{99 * 24769.21 * (1-0.68^2)}{98} = 13451.8$$

$$S = \sqrt{13451.8} = 115.98$$

La variancia residual és la variabilitat de la resposta no explicada pel model. Concretament, el numerador de l'equació S^2 és la suma dels residus al quadrat, obtinguts a partir de la recta de regressió.

c) (2 punts)

Poseu a prova mitjançant la prova d'hipòtesi adient si es pot admetre que aquesta recta passa pel punt (0,0), és a dir, si el paràmetre β_0 és significatiu. (Assumiu un risc α=0.05)

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$

$$S_{b0}^2 = S^2\left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}\right] = 13451.8\left[\frac{1}{100} + \frac{48.76^2}{99 * 796.76}\right] = 539.98$$

$$t = \frac{b_0 - \beta_0}{S_{b0}} = \frac{145.57 - 0}{\sqrt{539.98}} = 6.26$$

Com que el valor absolut de t: $|t|=6.26 > t_{98,0.975} = 1.984$, rebutgem la hipòtesis nul·la de que β_0 sigui nul i per tant, β_0 és diferent de zero.

d) (1 punt)

A la figura 2, teniu els gràfics necessaris per a poder validar la qualitat d'aquest model. Interpreteu quina premissa tracta de verificar cadascun d'aquests gràfics i si es verifica o no.

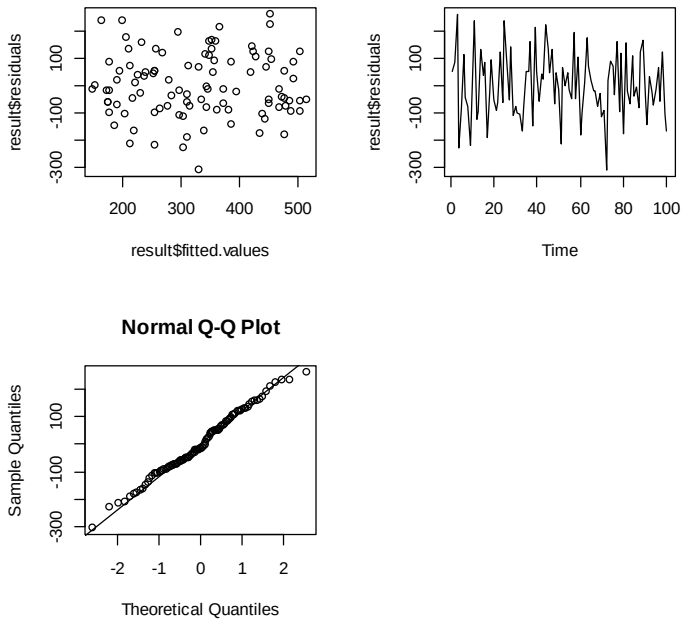


Figura 2. Anàlisi de les premisses

- La figura *result$residuals vs result$fitted.values* (dalt-esquerra) ens permet detectar la *homocedasticitat* i en aquest cas es verifica aquesta premissa, ja que la variabilitat dels residus versus els valors estimats no presenta cap tipus d'estructura. Aquesta figura també ens pot ajudar a verificar la linealitat entre les variable Y i X, que també es verifica ja que no existeix cap tipus de patró en aquest gràfic.

- La figura *result$residuals vs time* (dalt-dreta) ens permet detectar la independència dels residus al llarg del temps, ja que en aquest cas, aquest gràfic no mostra cap tipus d'estructura.

- Finalment, la figura *Normal Q-Q plot* (abaix) ens ajuda a detectar la normalitat, en aquest cas, dels residus de la regressió. Aquesta premissa es pot acceptar ja que els quantils empírics s'ajusten molt be als quantils teòrics provinents de la distribució normal.

e) (2 punts)

Doneu una previsió puntual amb el seu interval de confiança al 95% de quants segons trigarà en imprimir-se un fitxer de 60 kb.

- Previsió puntual:

$\hat{y}_h = b_0 + b_1 x_h = 145.64 + 3.79*60 = 373.05$ *seg*

$$V(\hat{y}_h) = S^2\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{(n-1)S_X^2}\right] = 13451.8\left[1 + \frac{1}{100} + \frac{(60 - 48.76)^2}{99 * 796.76}\right] =$$

$$= 13607.87\ seg^2$$

- Interval de confiança al 95%

IC95% = $\hat{y}_h \pm t_{n-2,0.975}\sqrt{V(y_h)} = 373.05 \pm 1.9844\sqrt{13607.87}$

És a dir:

IC95% =(141.56, 605.54) *seg*

- Conclusió:

La previsió puntual es de 373.05 *seg* amb un IC95% que oscil·la des de 141.56 a 604.54 *seg*.