

NOM: _____ COGNOM: _____

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

Problema 1 (Bloc C)

Volem estudiar unes dades sobre emissions de CO₂ (gCO₂/km) per models de vehicles dièsel (L), gasolina (G) i híbrids (H), tenint en compte tots els seus cicles de vida. Tenim unes mostres aleatòries simples de les emissions per a 24 models tant en versió dièsel com gasolina (L i G respectivament, i D per a la seva diferència) i per a uns altres 24 models híbrids (H), diferents tot i que comparables amb els anteriors.

$$\sum_{i=1}^{24} L_i = 3334.25 \quad \sum_{i=1}^{24} L_i^2 = 468018.8$$

$$\sum_{i=1}^{24} G_i = 3332.85 \quad \sum_{i=1}^{24} G_i^2 = 467669$$

$$\sum_{i=1}^{24} D_i = 1.4 \quad \sum_{i=1}^{24} D_i^2 = 1.315$$

$$\sum_{i=1}^{24} H_i = 2937$$

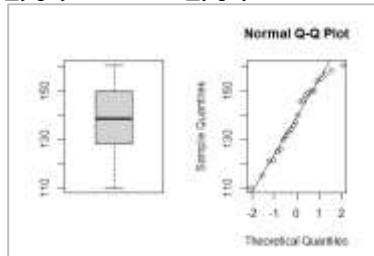


Figura 1

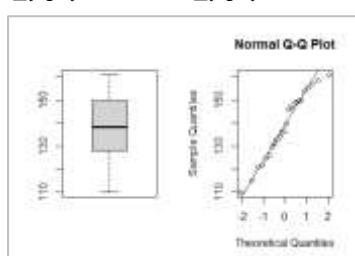


Figura 2

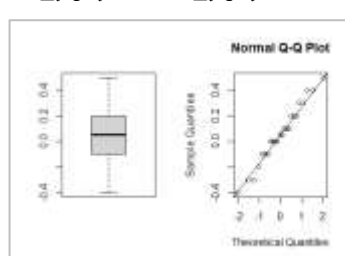


Figura 3

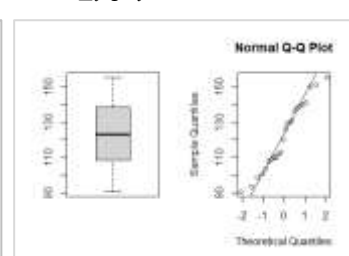


Figura 4

1.- (1 punt) Per una part volem comparar la mitjana d'emissions entre els dièsel i gasolina, i per altra entre els gasolina i híbrids. Indiqueu i justifiqueu si cadascuna d'aquestes dues proves de comparació de mitjanes serien de mostres aparellades o independents. Comenteu avantatges i inconvenients de les dues possibilitats.

Per comparar dièsel i gasolina usarem mostres aparellades ja que tenim 24 valors per uns mateixos models.

Per comparar gasolina i híbrid usarem mostres independents ja que tenim 24 vehicles de cada tipus però de models diferents. Mostres independents presenten més variabilitat, les aparellades són més controlades. Les mostres independents tenen més models diferents respecte les aparellades, ja que en aquestes cada model aporta dues observacions

2.- (1 punt) Calculeu una estimació puntual per a la diferència mitjana d'emissions entre dièsel i gasolina. Indiqueu la mitjana, la desviació, i el seu error tipus o estàndard i interpreteu-los

$$D_mean = \text{sum}(D)/24 \quad d = 1.4/24 = \mathbf{0.058}$$

$$sd = \sqrt{(\text{sum}(D*D) - ((\text{sum}(D)*\text{sum}(D)/24))) / (23))} = \mathbf{0.23}$$

$$se = sd/\sqrt{24} = \mathbf{0.047}$$

La diferència mitjana d'emissions entre dièsel i gasolina podem esperar que sigui de 0.058 punts més alta pels dièsel, amb una desviació de 0.23 punts i l'error estàndard baixa fins 0.047 per la mida de la mostra de 24 observacions

3.- (1.5 punt) Calculeu un IC al 95% de la diferència mitjana entre dièsel i gasolina. Interpreteu-lo i comenteu quina informació aporta de cara a concloure si les mitjanes d'emissions entre dièsel i gasolina són equivalents o no

$$D_mean - qt(0.975,23)*se = 0.058 - 2.069*0.047 = \mathbf{-0.04}$$

$$D_mean + qt(0.975,23)*se = 0.058 + 2.069*0.047 = \mathbf{0.16}$$

Podem esperar entre -0.04 i 0.16 punts de diferència en les emissions entre els dièsel i els gasolina amb 95% de confiança (entre 0.04 punts per sobre els gasolina i 0.16 per sobre els dièsel). Per tant a vegades és un o altre que podem esperar lleugerament per sobre, i 0 és un valor versemblant per a la diferència mitjana. Així aporta informació per concloure que les dades no evidencien que hi hagi diferència entre les mitjanes d'emissions

Per comparar la mitjana entre gasolina i híbrids hem obtingut els següent resultat en R:

```
t.test(G,H,var.equal=T,conf.level=0.99)
t = 3.3904, df = 46, p-value = 0.001442
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 3.421888 29.565612
sample estimates:
mean of x mean of y
138.8688 122.3750
```

4.- (1 punt) Indiqueu una estimació puntual per a la diferència mitjana d'emissions entre gasolina i híbrids. Indiqueu la mitjana i el seu error tipus o estàndard

$$G_mean = 138.8688 \quad (3332.85/24)$$

$$H_mean = 122.3750 \quad (2937/24)$$

$$Dif_mean = 138.8688 - 122.3750 = \mathbf{16.4938}$$

límit superior del IC és 29.565612 i és igual a $16.4938 + qt(0.995,46)*se \rightarrow se = (29.57 - 16.49)/2.687 \rightarrow se = \mathbf{4.87}$

5.- (1.5 punts) Indiqueu un IC al **99%** de la diferència mitjana entre gasolina i híbrids. Interpreteu-lo i comenteu quina informació aporta de cara a concloure si les mitjanes d'emissions entre gasolina i híbrids són equivalents o no

IC_min és 3.421888
IC_max és 29.565612

Podem esperar entre 3.42 i 29.57 punts de diferència en les emissions entre els gasolina i híbrids amb 99% de confiança (entre 3.42 i 29.57 punts per sobre els gasolina respecte els híbrids). Per tant, 0 no és un valor versemblant per a la diferència mitjana. Així aporta informació per concloure que les dades evidencien que es pot esperar una diferència entre 3.42 i 29.57 punts per sobre la mitjana d'emissions dels gasolina respecte els híbrids amb una confiança del 95%

6.- (1 punt) A partir dels resultats R i de les gràfiques, pel cas de comparació entre gasolina i híbrids, indiqueu si s'ha fet una comparació de mitjanes suposant homoscedasticitat o no i què vol dir

El t.test s'ha aplicat amb var.equal=T indicant que suposem homoscedasticitat
A les gràfiques els boxplot de G i H es veu que tenen una amplitud semblant indicant homoscedasticitat (més o menys entre 110 i 160 a la Fig 2 per G dels gasolina; i entre 90 i 150 a la Fig 4 per H dels híbrids)
Homoscedasticitat indica que estem suposant que la variabilitat de les emissions en gasolina i híbrids és semblant

7.- (1 punt) Indiqueu pels intervals de les preguntes 3 i 5 les premisses que han de complir, i si es compleixen o no

IC preg 3. És IC de mostres aparellades que són m.a.s. D és la diferència i podem assumir normalitat pel boxplot força simètric i per la normalitat en el qqnorm de la figura 3

IC preg 5. És IC de mostres independents que són m.a.s. Per G i H podem assumir normalitat pels boxplots força simètrics de i per la normalitat en els qqnorm de les figures 2 i 4 respectivament
També podem assumir homoscedasticitat entre G i H perquè els boxplot de G i H tenen una amplitud semblant (més o menys entre 110 i 160 a la Fig 2 per G dels gasolina; i entre 90 i 150 a la Fig 4 per H dels híbrids)

8.- (2 punts) Ara volem comparar el percentatge de vehicles que superen un cert valor crític. En els híbrids, dels 24 estudiats, 10 el superen; i en els dièsel i gasolina, dels 48 estudiats, 38 superen el valor crític. Calculeu un IC al **90%** per a la diferència de proporcions per concloure si són percentatges equivalents o no

$p1 = 10/24 = 0.42$
 $p2 = 38/48 = 0.79$
 $p1 - p2 = 0.42 - 0.79 = -0.37$
 $se = \sqrt{(p1*(1-p1)/24) + (p2*(1-p2)/26)} = 0.116$
 $(p1 - p2) - qnorm(0.95)*se = (0.42 - 0.79) - 1.645*0.116 = -0.57$
 $(p1 - p2) + qnorm(0.95)*se = (0.42 - 0.79) + 1.645*0.116 = -0.18$

Amb confiança 90% podem esperar entre un 18% i un 57% menys de vehicles amb emissions per sobre un valor crític en els híbrids respecte els de dièsel i gasolina

Valors que poden ser útils pels blocs C i D:

qt(0.975,11)= 2.201	qt(0.975,21)= 2.079	qt(0.975,46)= 2.013	qchisq(0.025,46)= 15.308	qnorm(0,85)= 1,036
qt(0.975,12)= 2.179	qt(0.975,22)= 2.074	qt(0.975,47)= 2.012	qchisq(0.025,47)= 16.047	qnorm(0,9)= 1,282
qt(0.975,13)= 2.160	qt(0.975,23)= 2.069	qt(0.975,48)= 2.011	qchisq(0.025,48)= 16.791	qnorm(0,95)= 1,645
qt(0.995,11)= 3.106	qt(0.995,21)= 2.831	qt(0.995,46)= 2.687	qchisq(0.975,46)= 44.461	qnorm(0,975)= 1,960
qt(0.995,12)= 3.054	qt(0.995,22)= 2.819	qt(0.995,47)= 2.685	qchisq(0.975,47)= 45.722	qnorm(0,99)= 2,326
qt(0.995,13)= 3.012	qt(0.995,23)= 2.807	qt(0.995,48)= 2.682	qchisq(0.975,48)= 46.979	qnorm(0,995)= 2,576

Contesteu cada pregunta en el seu lloc. Expliqueu i justifiqueu els càlculs.

Problema 2 (Bloc D)

Un estudi sobre vehicles considera dades pels anys de 2011 a 2023: la variable **A** com a “any–2010” (pren valors des de 1 fins a 13); la variable **t**, tipus de vehicle (BEV, híbrid convencional, o PHEV, híbrid endollable); i les vendes anuals a tot el mon. Aquí teniu els gràfics de les variables amb les vendes, abans i després d’aplicar *logaritme natural* (**sales** i **LOG(sales)**).

Hem ajustat amb R dos models (a sota). Pel model (1) la variable de resposta és el logaritme natural del total de vendes anual (l’etiqueta “both”, al gràfic). Pel model (2) la variable de resposta és el logaritme natural de les vendes anuals per als dos tipus.

(1) summary(lm(log(total_sales) ~ A))

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.67374	0.11956	97.64	<2e-16
A	0.44440	0.01506	29.50	8e-12

Residual standard error: 0.2032 on 11 deg. of fr.

Multiple R-squared: 0.9875, Adj. R-sq.: 0.9864

F-statistic: 870.4 on 1 and 11 DF, p-value: 7.998e-12

(2) summary(lm(log(sales) ~ A + t))

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.19873	0.15581	71.875	< 2e-16
A	0.44738	0.01776	25.196	< 2e-16
tPHEV	-0.59866	0.13287	-4.505	0.00016

Residual standard error: 0.3388 on 23 deg. of fr.

Multiple R-squared: 0.9661, Adj. R-sq.: 0.9631

F-statistic: 327.6 on 2 and 23 DF, p-value: < 2.2e-16

1. Expliqueu què és i com s’interpreta el valor 11.67374 del model (1).

És el terme independent del model lineal, i representa l’estimació puntual de la mitjana del logaritme del nombre de vendes totals per A=0 (any=2010); desfent la transformació, correspon a un total de vendes de 117447 vehicles.

2. Pel valor equivalent al model (2), 11.19873, subratlleu les similituds i les diferències.

Representa l’estimació de la mitjana del logaritme del nombre de vendes per A=0 específicament per a vehicles BEV (categoria de referència). Correspon a 73038 unitats de vehicles BEV. Aquest segon model permet diferenciar el nombre de vendes per tipus gràcies a una segona variable predictora.

3. Expliqueu què és i com s’interpreta el valor 0.44440 del model (1). *Pista: invertiu la transformació logarítmica i raoneu en conseqüència.*

És el pendent o terme lineal del model, i representa l’estimació puntual de l’increment del logaritme del nombre de vendes totals en un any. Desfent la transformació, obtenim un increment anual al voltant del 56%.

4. Què informació aporta el valor 0.01506?

És l’error tipus de l’estimació del terme lineal anterior (0.44440), la magnitud de la variació que podem esperar per l’error de mostreig. Aproximadament, una variació de ±0.03.

5. Interpreteu els tres primers valors de la línia tPHEV, al model (2): quina informació proporciona aquesta estimació?

tPHEV

-0.59866

0.13287

-4.505

Estimate:

representa el canvi en el logaritme del nombre de vendes de vehicles PHEV per a un any donat respecte dels vehicles BEV (un descens de -0.6 equival a un 55%: es venen un 45% menys de vehicles PHEV que de BEV)

Std. Error:

representa una mesura de l’error d’estimació del paràmetre anterior.

t value: és el rati senyal/soroll del quocient dels valors anteriors. Indica que l’error de la mostra no és tan fort com per invalidar la informació que porta l’estimador; també, que hi ha evidència per dir que realment es venen menys híbrids endollables.

6. Calculeu un interval de confiança al 95% pel paràmetre associat a la variable **t**.

$IC(0, 95\%) = -0.59866 \pm t_{23,0.975} 0.13287 = (-0.8735, -0.3238)$. Desfent la transformació, el canvi real podria estar entre un 42% i un 72%, amb 95% de confiança.

7. Valoreu la capacitat/qualitat del primer model per fer prediccions: quin indicador és l'adequat, i què ens diu?

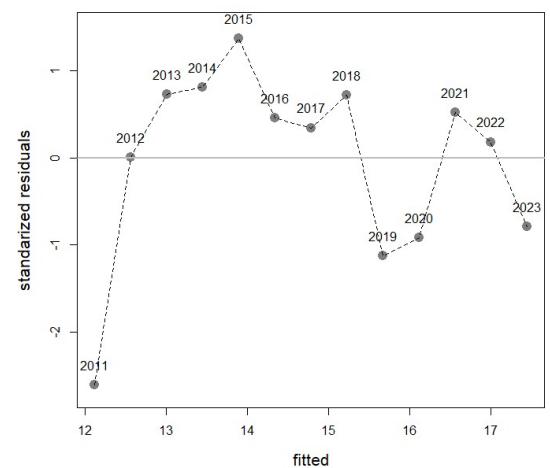
Utilitzarem el coeficient de determinació R^2 . Ens quantifica la fracció de variabilitat de les dades que ve explicada per la variable predictora en el model: 98.75%. L'any explica molt bé el creixement del logaritme de les vendes totals, només deixa un 1.25% de variabilitat a l'atzar.

8. Com ens retorna el model informació sobre les discrepàncies entre les observacions i les prediccions? Quantifiquem aquesta informació pel cas del segon model.

Les discrepàncies es poden analitzar amb l'estimació de la desviació residual s : 0.3388. Notem que els errors de predicció del logaritme de les vendes d'un tipus particular són majors que pel cas de vendes totals.

9. A la dreta teniu un gràfic obtingut pel model (1). Descriviu què està representant. Indiqueu les premisses del model i comenteu críticament i justificada quines es poden donar per assumibles.

A l'eix X, "fitted" vol dir les prediccions (per tant, logaritme de vendes totals); a l'eix Y tenim els residus estandaritzats, és a dir, els residus dividits per la desviació residual. Les prediccions (casualment) es situen en ordre temporal d'esquerra a dreta, i notem que d'un any al següent els residus són semblants en molts casos, indicant que no són gaire independents. La linealitat del model no sembla que es pugui posar en dubte, en general els punts es disposen al voltant de la línia horitzontal. Tenim una dada al 2011 bastant lluny de la resta, però no hi ha suficient informació per valorar normalitat o homocedasticitat. Tampoc podem dir que hi hagi evidències en contra.



10. Amb un dels models anteriors hem executat aquesta instrucció. Expliqueu quin resultat està proporcionant.

```
> predict(mod, data.frame(A=15,t='BEV'), int='prediction')
      fit      lwr      upr
1 17.90945 17.12508 18.69381
```

El model és el segon, perquè hem inclòs dades de dues covariants, l'any $A=15$, és a dir, el 2025, i el tipus BEV: volem obtenir una predicció del nombre de vehicles BEV que es vendran l'any 2025. El resultat puntual és 17.91 que en termes absoluts és 59975698 (uns 60 milions). Com que hem indicat 'prediction', vol dir que es vol fer una predicció de tipus individual amb interval de confiança al 95%, en aquest cas la xifra "realista" de vendes desfent el logaritme aniria de 27373337 a 131407009 (27.3 milions a 131.4 milions).