Sessió 9

Dades multivariants 3: relació entre 2 variables numèriques

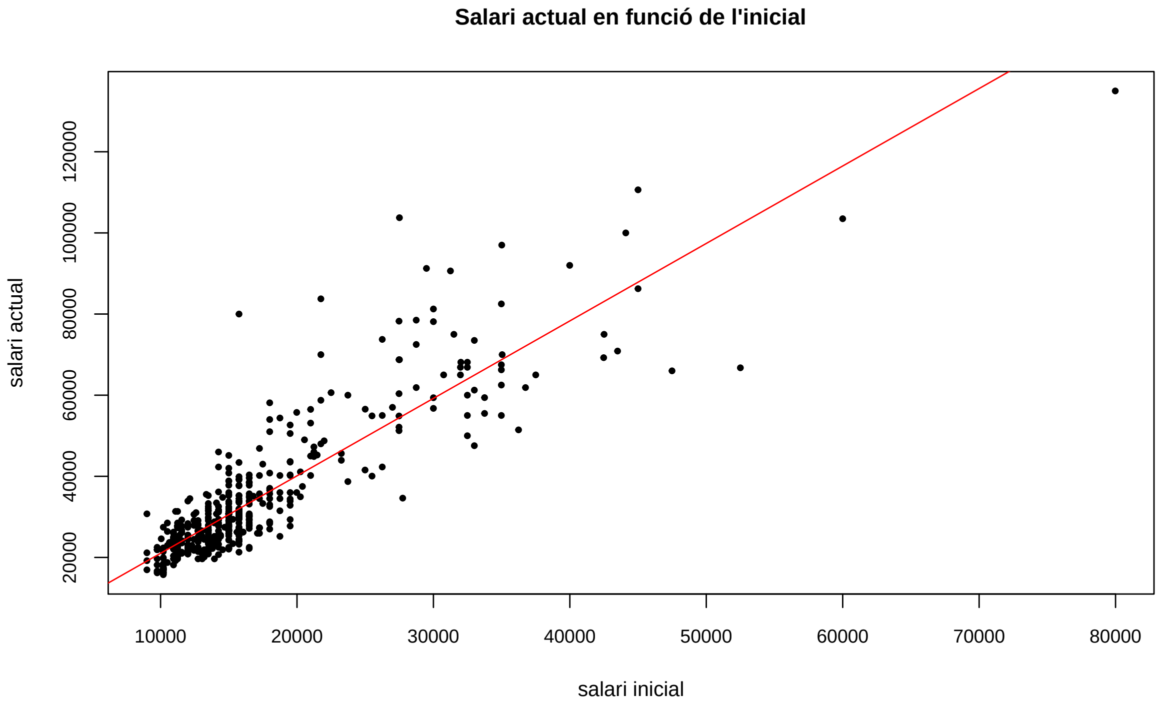
Tractarem l'anàlisi de les relacions lineals entre dues variables numèriques: regressió, correlació i diagrames de dispersió. Farem la representació mitjançant un núvol de punts (diagrama de dispersió) i el càlcul del coeficient de correlació i dels paràmetres de la recta de regressió. Treballarem amb el fitxer *Datosdeempleados.sav*. Per llegir-lo necessitem la funció `read.spss` de la llibreria `foreign`

```
library(foreign)
df<-read.spss("Datosdeempleados.sav",to.data.frame=T,max.value.labels=5)
# sense l'opció max.value.labels=5
# agafa les variables numèriques com a factors
summary(df)  #comprovem que tenim alguns factors
```

9.1 Gràfiques de dispersió: núvol de punts i recta de regressió

La funció bàsica per fer un núvol de punts de la distribució conjunta de dues variables numèriques és `plot`

```
attach(df)
plot(salario~salini) # sempre és y(dependent) ~ x(independent)

plot(salini,salario)  # equivalent, la x la primera, la y segona

plot(salario~salini, main="Salari actual en funció de l'inicial",
       xlab="salari inicial ", ylab="salari actual", pch=20)
abline(lm(salario~salini), col="red")
      # s'afegeix la recta de regressió al núvol de punts,
      # en el proper apartat veurem com calcular-la amb la funció lm
```



Pràctica: • Obre el fitxer de dades `enq.txt`. Fes un gràfic de dispersió de `Pes` i `Alt` del fitxer `enq.txt`. Què s'observa al gràfic?
• Amb el data.frame `df` fes un gràfic de dispersió de la variable `salario` (Y) respecte de la variable `tiempemp` (X). Afegeix-hi la recta de regressió. Diries pel gràfic que hi ha associació entre les variables?

9.2 Càlcul de la recta de regressió i del coeficient de correlació

- Calculem i resumim el model lineal:

```
> model.salaris<-lm(salario~salini)
> summary(model.salaris)

Call:
lm(formula = salario ~ salini)

Residuals:
   Min      1Q Median      3Q     Max
-35424   -4031   -1154    2584   49293

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.928e+03  8.887e+02    2.17   0.0305 *
salini      1.909e+00  4.741e-02   40.28   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 8115 on 472 degrees of freedom
Multiple R-squared:  0.7746,   Adjusted R-squared:  0.7741
F-statistic:  1622 on 1 and 472 DF,  p-value: < 2.2e-16
```

A l'apartat `Coefficients` de l'anterior resum hi apareix la recta de regressió $y = bx + a$:

(Intercept)	$1.928 \cdot 10^3$	$= a$
salini	1.909	$= b$

A l'apartat `Multiple R-squared` hi apareix el coeficient de determinació r^2:

Multiple R-squared 0.7746 = r^2

- Un cop hem calculat el model i l'hem assignat a l'objecte `model.salaris` (que és tipus llista), podem recuperar directament els **coeficients de la recta de regressió**:

```
> model.salaris$coefficients
(Intercept)       salini
 1928.20576     1.90945
> model.salaris[[1]]   # equivalent a l'anterior
```

- El **coeficient de correlació** es calcula amb `cor`

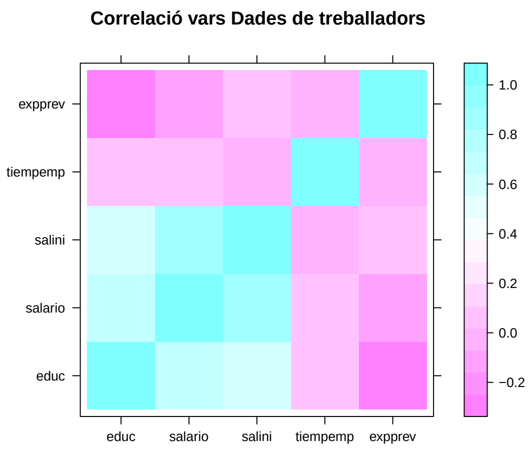
```
> cor(salario,salini)
[1] 0.8801175
> cor(salario,salini)^2
[1] 0.7746
# comprovem que si elevem al quadrat obtenim
    # el que apareixia al summary com  "Multiple R-squared:  0.7746"
```

- Per fer un seguit de **prediccions**, tantes com vulguem, farem el següent:

```
# Introduïm a la variable xllista un vector de valors de x (salini)
# per als quals volem predir y (salario)
xllista<- c(10500,12395,29900)
# Posem al vector coef els coeficients de la recta de regressió
coef<- model.salaris$coefficients
# Obtenim el vector de les tres prediccions:
pred<-coef[1]+coef[2]*xllista
[1] 21977.43 25595.84 59020.75

# també podem definir una funció que sigui la recta de regressió
f<-function(x) coef[1]+coef[2]* x
f(xllista)
```

- Quan tenim més d'una variable numèrica en un conjunt de dades, podem fer alhora la correlació entre les variables numèriques en l'anomenada **matriu de correlacions** i representar gràficament les correlacions en el que s'anomena un **gràfic de calor**. El gràfic de calor es pot fer amb `heatmap` de la llibreria `stats` bé amb la funció `levelplot` de la llibreria `lattice`.

```
vars.num<-c(4,6,7,8,9)  #les variables numèriques de df
df.num<-df[,vars.num]   # subset de df amb les variables numèriques
(cormat<-round(cor(df.num),2))  # matriu de correlació
          educ salario salini tiempemp expprev
educ      1.00    0.66   0.63     0.05   -0.25
salario   0.66    1.00   0.88     0.08   -0.10
salini    0.63    0.88   1.00    -0.02    0.05
tiempemp  0.05    0.08  -0.02     1.00    0.00
expprev  -0.25   -0.10   0.05     0.00    1.00
library(lattice)
levelplot(cormat,xlab="",ylab="",
           main="Correlació vars Dades de treballadors")
```



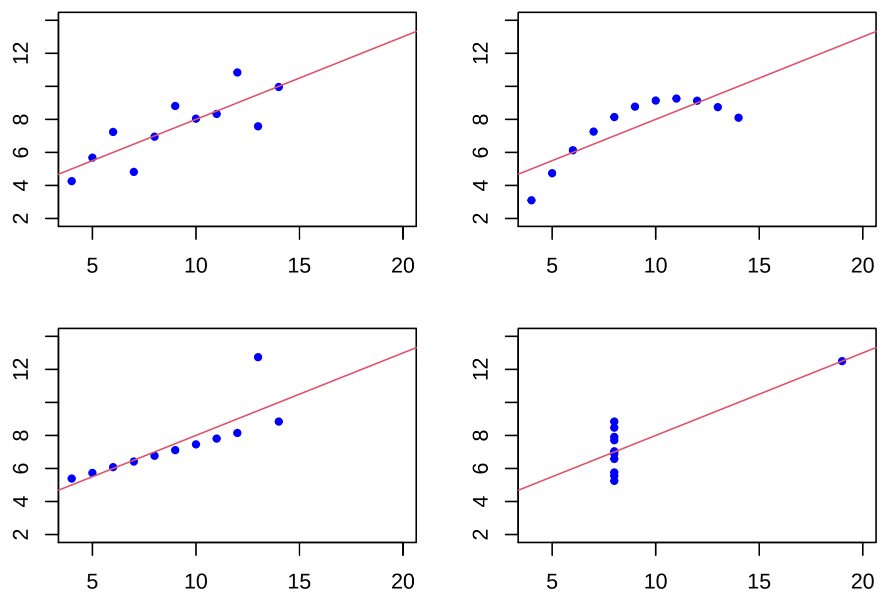
Pràctica: • Seguint amb el fitxer `enq.txt`. Calcula la recta de regressió de `Pes` respecte de l'altura `Alt` i utilitza-la per donar una estimació del pes que tindria una persona de 176 cm d'altura. Calcula el coeficient de correlació de les dues variables i interpreta el resultat.

• Troba la recta de regressió de la variable `salario` (Y) respecte de la variable `educ` (X) del data.frame `df` i utilitza-la per donar una estimació del salari d'una persona amb nivell educatiu de 13 anys. Calcula el coeficient de correlació de les dues variables i interpreta'l.

• El fitxer `AnscombeQuartet.csv` conté quatre parelles de variables $X_1, Y_1, X_2, Y_2, X_3, Y_3, X_4, Y_4$, que corresponen a l'exemple "quartet d'Anscombe". Comproveu que, arrodonint a dos decimals:

- les X_i tenen mateixes mitjanes i desviacions,
- les Y_i, entre elles, també coincideixen en les mitjanes i les desviacions,
- les correlacions de les parelles X_i, Y_i són iguals, i també les rectes de regressió

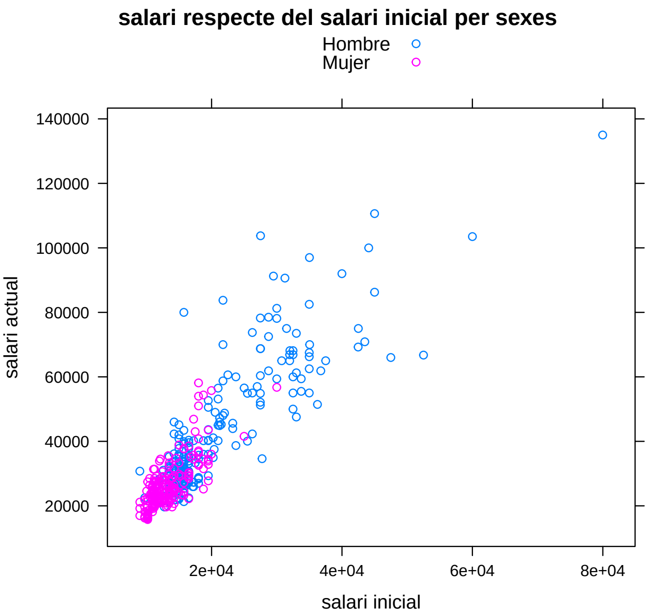
Feu un gràfic com el que hi ha a continuació, on es veu que les distribucions de les parelles són molt diferents.



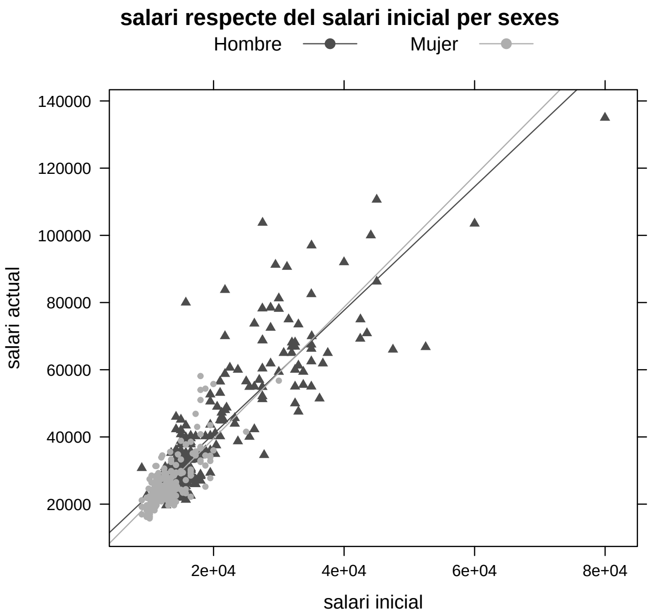
9.3 Estudi de la distribució conjunta de dues variables numèriques per grups

Anem ara a introduir un factor, el factor `sexo` i a comparar la relació entre les variables `salini` i `salario` en homes i dones. Hem de carregar la llibreria `lattice`

```
library(lattice)
# gràfic 1
xyplot(salario ~ salini, groups=sexo, data=df,
      main="salari respecte del salari inicial per sexes",
      xlab='salari inicial', ylab='salari actual',
      auto.key=T)
# gràfic 2, personalitzant colors, pch,
# i afegint les dues rectes de regressió amb type=c("p","r"),
xyplot(salario ~ salini, groups=sexo, data=df, col=gray.colors(3),
       main="salari respecte del salari inicial per sexes",
       xlab='salari inicial', ylab='salari actual',
       pch=c(17,20),type=c("p","r"),
       key=list(text=list(levels(sexo)),lines=list(pch=19, type="b",
              col=gray.colors(3)[1:2]), columns=2, divide=1))
```



gràfic 1



gràfic 2

Si volem calcular les **dues rectes de regressió**, de salario sobre salini en dones i homes, podem utilitzar la funció `by`

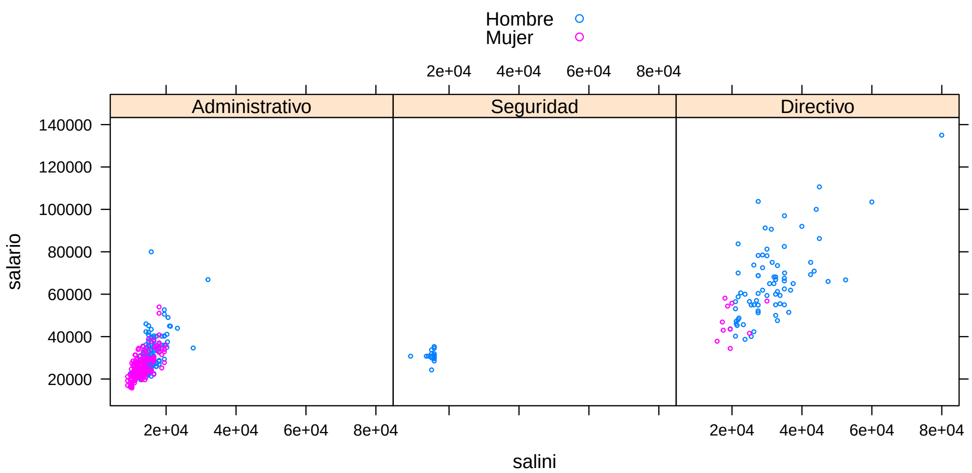
```
model.grups<-by(df, sexo, function(x) lm(salario ~salini, data = x))
sexo: Hombre

Call:
lm(formula = salario ~ salini, data = x)

Coefficients:
(Intercept)       salini
    4083.08         1.84
----------------------------------------------------------------------
sexo: Mujer
```

```
Call:
lm(formula = salario ~ salini, data = x)

Coefficients:
(Intercept)        salini
    438.510         1.955

#Si volem recuperar els coeficients per treballar-hi
model.grups$Mujer$coefficients
model.grups$Hombre$coefficients

library(lme4)
lmList(formula = salario ~ salini | sexo, data = df)
Coefficients:
       (Intercept)    salini
Hombre   4083.0758 1.840204
Mujer     438.5102 1.954894
```

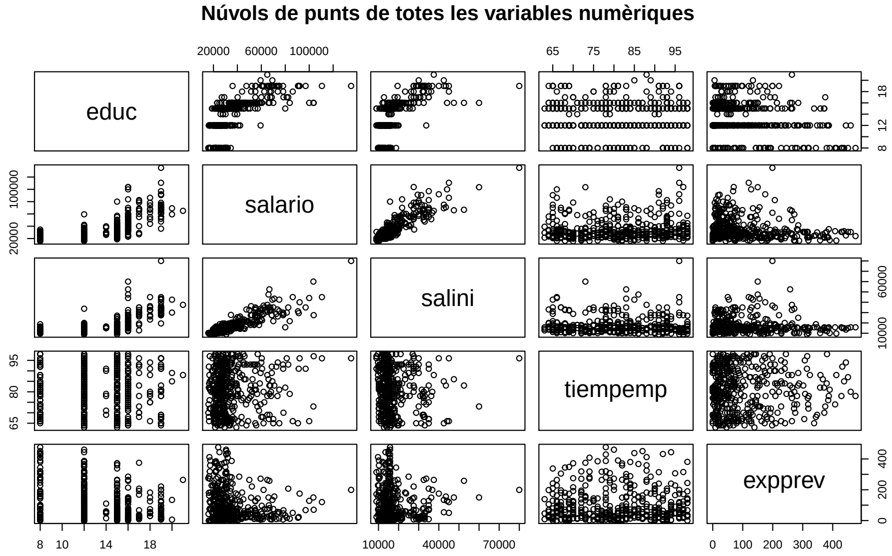
9.4 Miscel·lània

- Encara podem **representar més variables alhora** (dues numèriques i dues categòriques)

```
xyplot(salario~salini|catlab,data=df,groups=sexo,auto.key=T,cex=0.4)
```



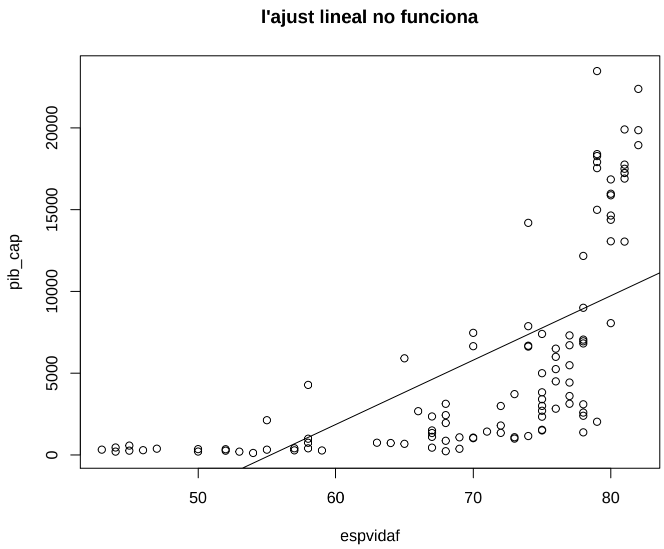
- Amb la funció `pairs` podem aconseguir núvols de punts de totes les parelles de variables. Va molt bé per fer-nos una idea de la relació aproximada entre elles.



- Exemple de com es faria una **regressió no lineal**, amb una transformació logarítmica.

  Treballem amb les variables `espvidaf` i `pib_cap` del fitxer `mon.RData` que podeu trobar al Moodle. En representar-les veiem que no s'ajustaria bé una recta, i podem comprovar que el coeficient de determinació és baix:

```
load("mon.RData")   # o bé directament
mon<-read.csv2("mon.csv")
attach(mon)
plot(pib_cap~espvidaf)
abline(lm(pib_cap~espvidaf))
summary(lm(pib_cap~espvidaf))
...
Multiple R-squared:  0.4125,    Adjusted R-squared:  0.407
```

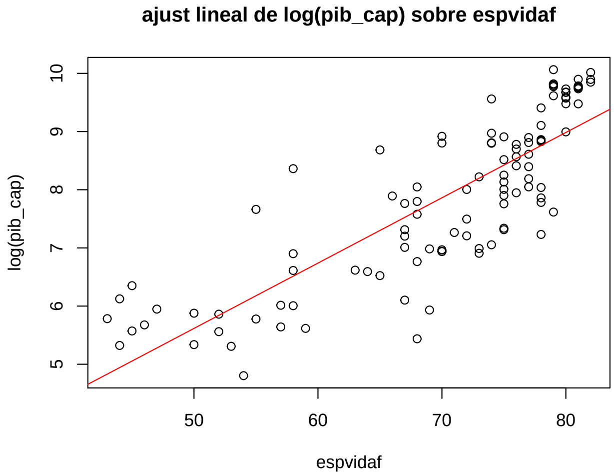


Com que el núvol de punts suggereix que la variable Y (`pib_capita`) és en mitjana una funció aproximadament exponencial de X (`espvidaf`), apliquem logaritme a la segona variable:

Si $Y \approx ce^{kX}$, aleshores $\log(Y) \approx \log(C) + kX$, és a dir $\log(Y)$ s'assemblarà a una funció lineal de X. Per aquest motiu estudiem $\log(Y)$ respecte de X:

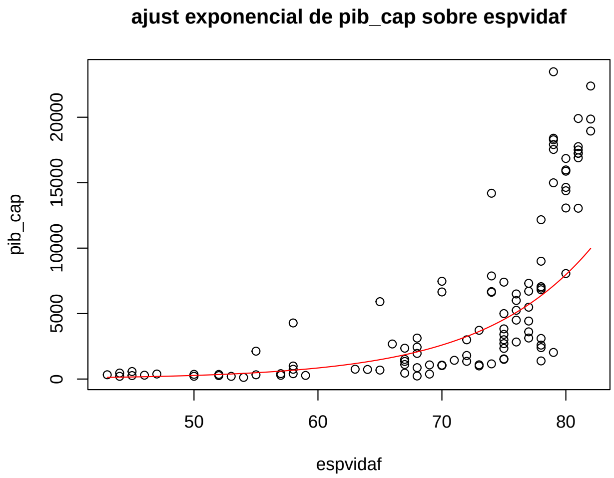
```
plot(log(pib_cap)~espvidaf)
abline(lm(log(pib_cap)~espvidaf))
summary(lm(log(pib_cap)~espvidaf))
...
Multiple R-squared:  0.6907,    Adjusted R-squared:  0.6878
```

D'entrada, el coeficient r^2 ha pujat de 0.41 a 0.69, per la qual cosa la recta de regressió s'ajustarà millor, com veiem al gràfic.



Si $\log(Y) \approx a + bX$, amb a i b els coeficients de la recta de regressió de $\log(Y)$ sobre X, aleshores $Y \approx e^{a+bX}$. Representem junt amb el núvol de punts de la distribució conjunta de X i Y, la funció d'ajust exponencial $y = e^{a+bX}$

```
# càlcul dels coefs recta de regressió de log(pib_cap) sobre espvidaf
logmod<-lm(log(pib_cap)~espvidaf)
coefs<-logmod$coefficients
# representem pib_cap sobre espvidaf i la corba d'ajust exponencial
plot(pib_cap~espvidaf,
           main="ajust exponencial de pib_cap sobre espvidaf")
x<-seq(min(espvidaf,na.rm=TRUE),max(espvidaf,na.rm=TRUE),by=0.05)
points(x,exp(coefs[1]+coefs[2]*x),type="l",col=2)
```



Pràctica: • Fes un gràfic de dispersió de `Pes` i `Alt` del fitxer `enq.txt` on s'hi vegi els efectes del factor Sexe. Interpreta-ho.
• Utilitzant la funció `pairs` amb un subconjunt de variables del fitxer `mon.RData`, tria dues variables per fer un estudi amb una transformació logarítmica.