hands-on ESTADÍSTICA AMB R

# Carregar LES DADES

El primer que farem serà anar a buscar la taula de dades que servirà per aquesta part pràctica. Teniu dues maneres de fer-ho (que segurament ja sabreu):

MANERA 1 – Per comandes de codi. Us situeu a la consola de l’R i

# Read file

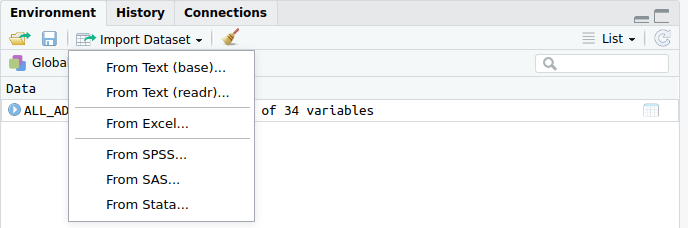
setwd(“My/working/directory”) Ens movem al directori on està la base de dades

ALL\_ADNI <- read.delim("ALL\_ADNI.xls", header = TRUE, sep = ”\t”) # llegim la base de dades

Si tot ha anat com havia d’anar, haurà aparegut una nova variable (un dataframe)



MANERA 2 – Per clics: Import Dataset 🡪 From Excel (NO RECOMANADA, amb la versió actual d’R de l’hospital no funciona. A partir de 2020, s’actualitzarà el sistema i ja funcionarà)

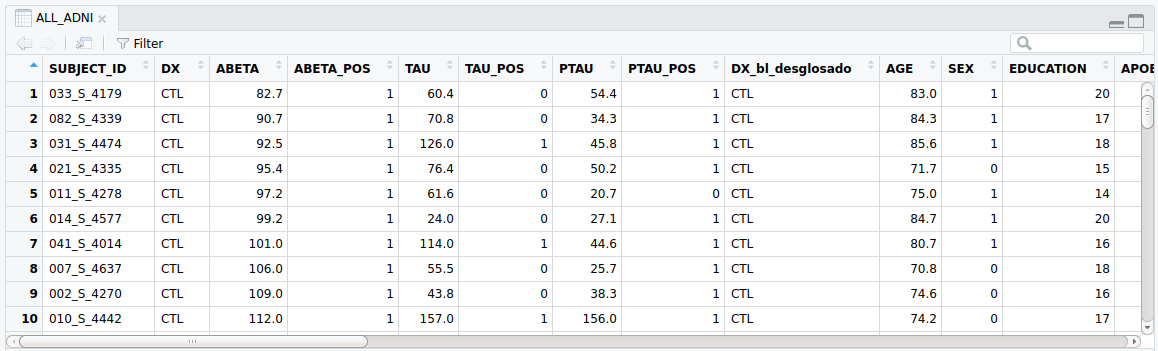


I podem navegar fins trobar la nostra taula.

Podem inspeccionar la taula visualment mitjançant:

View(ALL\_ADNI)

O bé clicant directament a sobre de la taula. En qualsevol dels casos s’obrirà la taula i hauríem de veure alguna cosa semblant a això:



La variable SUBJECT\_ID és la variable que identifica els subjectes. Les altres variables són dades dels subjectes. Concretament es tracta d’una base de dades americana: Alzheimer’s disease Neuroimaging Initiative (ADNI, <http://adni.loni.usc.edu/>). La taula està lleugerament processada perquè pogueu treballar-hi, cada fila és un subjecte únic. La taula inclou:

* SUBJECT\_ID: Identificador únic del subjecte
* DX : diagnòstic dels subjectes (CTL= healthy controls, MCI= mild cognitive impairment, i AD= Alzheimer’s disease)
* Biomarcadors en líquid cefalorraquídi (ABETA, TAU, PTAU)
* Sex (0=dona, 1=home)
* Education...
* I moltes altres variables que anireu veient.

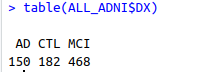
# Revisió de les dades

El primer pas en qualsevol anàlisi és revisar les dades, veure el nombre de casos, freqüències, explorar mitjanes, desviacions, en definitiva, analitzar les dades descriptives.

1. **Ex.1 - Inspeccionant les dades i veient descriptius: Intenteu treure el nombre de casos totals per grup diagnòstic:**

#library(MASS)

table(ALL\_ADNI$DX)



I si ara volem els percentatges com ho faríeu? Hauríeu de trobar un resultat així:



1. **Treballar les dades. Creació de noves variables.**

Voldríem trobar el % d’APOE e4 negatius/positius. Tenim una variable que es diu APOE4, que ens proporciona el nombre d’al·lels e4 de cada individuu. **Com creem una nova variable binaria que determini la positivitat? Amb la funció ifelse()**

ALL\_ADNI$APOE4\_positiu <- ifelse(ALL\_ADNI$APOE4 > 0 ,1,0)

ALL\_ADNI$APOE4\_positiu= crea la variable APOE4\_positiu dins la taula ALL\_ADNI

Ifelse( condició, [valor si es compleix la condició], [valor alternatiu]

**Extra job: Sabríeu trobar el percentatge d’APOE e4 negatius dins del grup d’Alzheimer? (la funció subset és la vostra solució)**

# Analitzar els descriptius: Freqüències, mitjanes, medianes, variances

En el punt anterior hem vist com podem calcular el nombre total de casos per grup. Això ens pot ser útil per donar la N mostral, subgrups, etc. El següent pas seria calcular els demogràfics: estudiar l’edat dels grups, proporció de sexe, educació, etc.

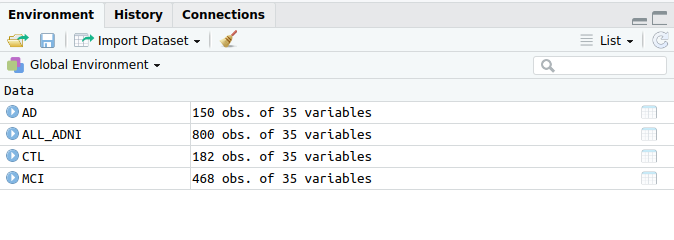
Començarem en crear un exemple senzill: calcularem mitjanes d’edat per grup (CTL, MCI i AD). Per fer-ho, primer crearem 3 taules auxiliars per a cada grup diagnòstic, de manera que els càlculs siguin més senzills. Comencem:

AD<-subset(ALL\_ADNI,ALL\_ADNI$DX=="AD")

MCI<-subset(ALL\_ADNI,ALL\_ADNI$DX=="MCI")

CTL<-subset(ALL\_ADNI,ALL\_ADNI$DX=="CTL")

Si tot ha anat bé, s’hauran creat 3 taules extres. Ho veieu?



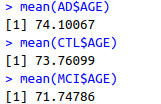
Ara calcularem les mitjanes per grup

mean(AD$AGE)

mean(CTL$AGE)

mean(MCI$AGE)

Outputs:



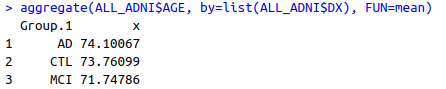
**Extra job: Podem fer el mateix per a les desviacions estàndard. Com? I per a les medianes?**

Fer això per a cada variable i per a cada grup és costós. Hi ha una manera més ràpida de treure mitjanes i desviacions estàndard de les dades i agrupades per grup, com per exemple la funció **aggregate()**. El seu funcionament és senzill:

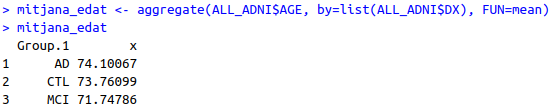
aggregate(vector\_amb\_la\_variable, by = list(vector\_amb\_la\_variable\_grup), FUN=operació a aplicar).

Exemple:

aggregate(ALL\_ADNI$AGE, by=list(ALL\_ADNI$DX), FUN=mean)

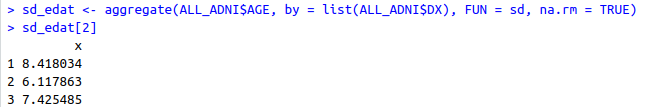


Podem posar el resultat dins d’una variable, per exemple.



**Extra job:** Podem fer el mateix per a les desviacions estàndard. Com? FUN=sd

sd\_edat <- aggregate(ALL\_ADNI$AGE, by = list(ALL\_ADNI$DX), FUN = sd, na.rm = TRUE)



Podem seleccionar només el vector de resultats seleccionant la 2a columna de l’element que hem creat.

sd\_edat[2]

**Super extra hard job**: Podem fer aquest procés de manera iterativa per N variables. Com? Fàcil 😊 Primer de tot definim un vector variables:

variables=c("AGE","ABETA","PTAU")

Després, creem una matriu buida que ens servirà per guardar els resultats. Concretament, ha de tenir 3 columnes (una per cada grup diagnòstic), i tantes files com variables tinguem:

sd\_matrix<-data.frame(matrix(0,ncol=length(variables),nrow=3))

colnames(sd\_matrix)<-variables

rownames(sd\_matrix)<-c("AD","CTL","MCI")

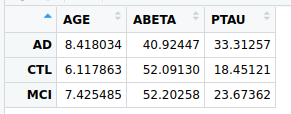
A continuació, fem un bucle que iteri en el nostre vector de variables

for (var in variables){

sd\_matrix[var] <- aggregate(ALL\_ADNI[var], by=list(ALL\_ADNI$DX), FUN=sd, na.rm=TRUE)[2]

}

I, equiliqüà! Tenim la nostra matriu amb variances



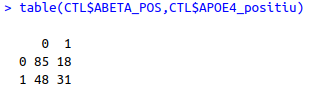
**Extra work – Feu el mateix per extreure les mitjanes**

# Anàlisi de proporcions: taules de contingència i chi-square

Crear una taula de contingència amb R és immediat. Per exemple, volem mirar si hi ha diferències de proporcions en positivitat d’Aß segons si els subjectes són portadors o no de l’al·lel e4 d’APOE.

table(CTL$ABETA\_POS,CTL$APOE4\_positiu)

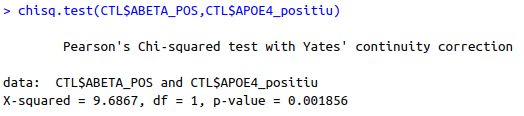
Això ens retorna una taula on podem veure les proporcions.



Variable 1

Variable 2

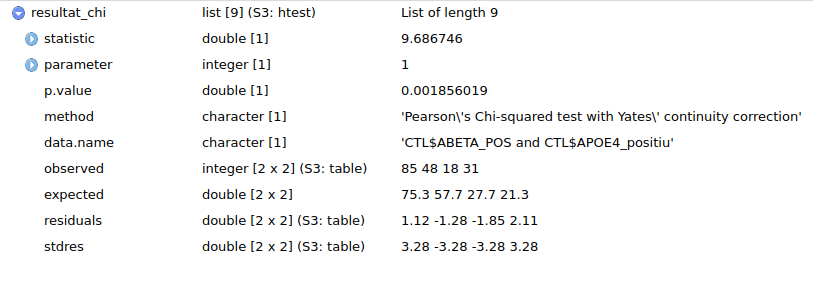
El pas obvi és seria calcular l’estadístic. Escollim la chi-square per avaluar la diferència de proporcions.

* chisq.test(variable1, variable2)
* 

En aquest cas, observem que existeix una diferència entre la proporció esperada d’Aß positius entre els portadors i no-portadors de l’al·lel e4 d’APOE.

Podem guardar el resultat en una variable. Per exemple

resultat\_chi <- chisq.test(CTL$ABETA\_POS,CTL$APOE4\_positiu)



Si explorem l’estructura que acabem de crear, podrem veure que té diferents paràmetres. Podem accedir a ells amb un “$”. Per exemple:



Guardar els resultats en altres variables és útil, sobretot i després volem re-comprovar resultats o per scriptar. Això ho anireu veient a mesura que us trobeu necessitats diverses.

Pro tip: En cas de tenir una mostra petita (i.e. N<5 a alguna de les caselles, la chi-square no ens serveix. En aquest cas s’haurà d’aplicar el test de Fisher.

* fisher.test(variable1, variable2)

# Testejar la normalitat

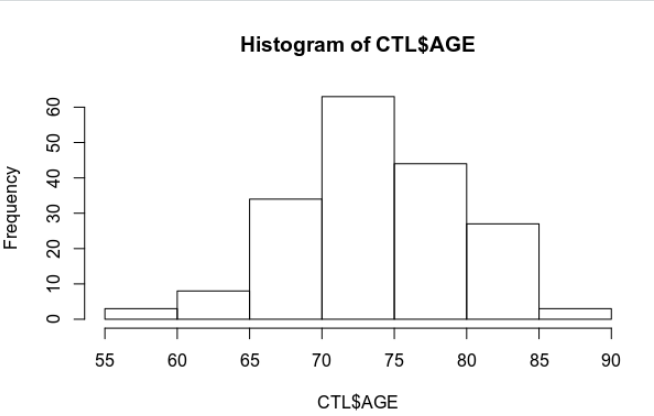
No cal dir que abans de passar a la feina i aplicar qualsevol mètode estadístic, cal assegurar-se que les dades compleixen certes assumpcions. Per exemple, si apliquem estadística normal, caldrà que les dades segueixin una distribució normal.

## TEST de SHAPIRO-WILK

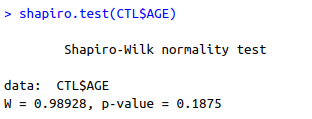
shapiro.test(dades)

En el nostre cas, imaginem que volem avaluar si la mitjana d’edat difereix entre grups. Primer, podem provar de fer un histograma per veure la distribució:

hist(CTL$AGE)

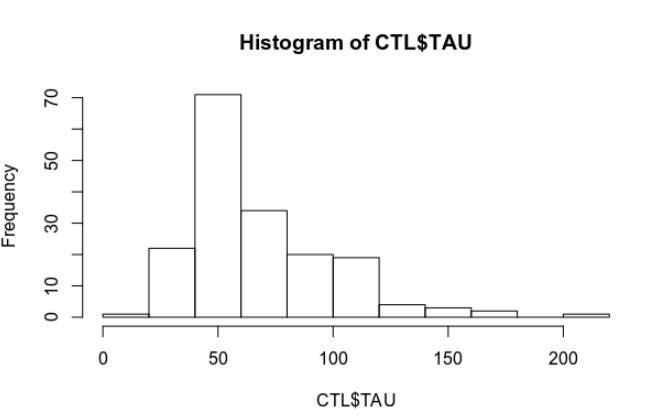


La qualitat visual del gràfic deixa molt a desitjar (estic segur que sabríeu millorar-lo :-) ), però l’important ara és veure com són les dades que tenim entre mans. Fa pinta de normal, però anem a aplicar estadística.



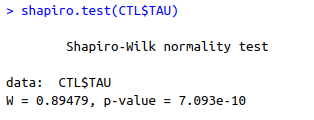
Efectivament, corroborem que la variable segueix una distribució normal (p no significativa). D’aquesta manera, podem passar a aplicar estadística normal.

Ara analitzarem una altra variable: TAU. Primer, l’histograma:



Anticipem que fa mala pinta...

A l’aplicar el test...



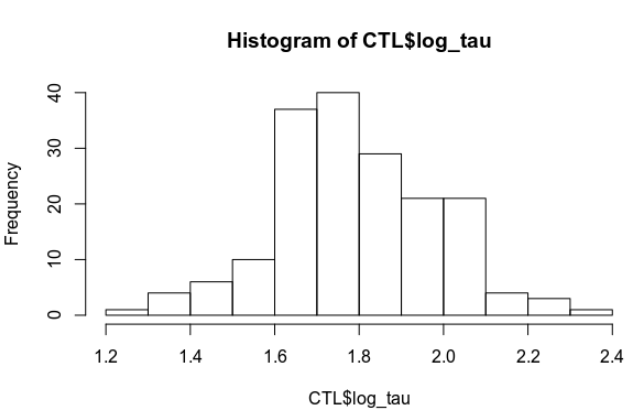
El test surt significatiu, demostrant que les dades difereixen significativament d’una distribució normal.

Vaja... Quines opcions plantegem?

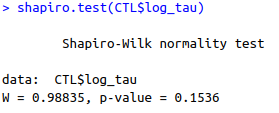
**Opció 1 – Transformar la variable i crear-ne una de nova**

CTL$log\_tau <- log10(CTL$TAU)

Si observem l’histograma de la variable transformada...



I a l’aplicar el test...



Solucionat!

**Opció 2. Treballar amb estadística no-paramètrica**

# Comparant mitjanes

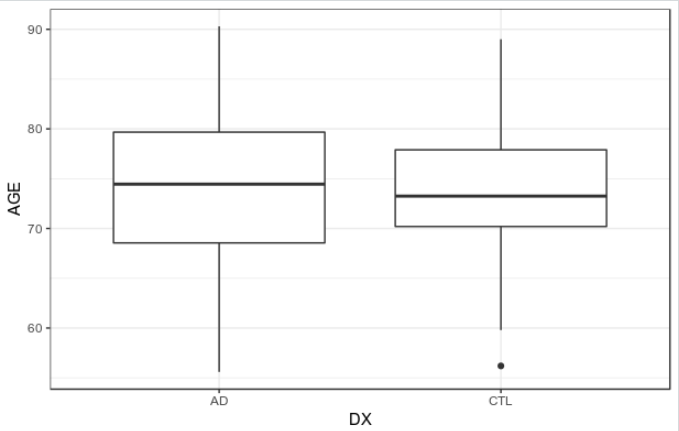
## Amb dos grups

Existeixen diferències entre les mitjanes de certa variable entre dos grups? Quantes vegades ens haurem formulat aquesta pregunta...

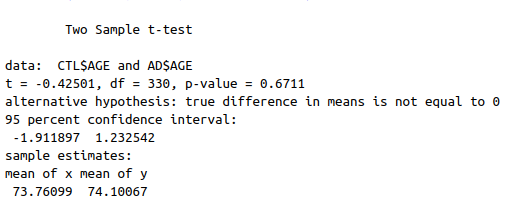
Per exemple, volem estudiar si existeixen diferències d’edat entre els pacients de control i els pacients amb demència tipo Alzheimer. Com ho faríeu? En primer lloc, intentem veure-ho gràficament.

**Ex.1 - Quin seria el codi per obtenir aquest box-plot?**

**Pista: podeu ajuntar dos *dataframes* amb la ordre new\_dataframe=rbind(dataframe1, dataframe2)...**



Gràficament sembla que no hi hagués d’haver diferències. Per calcular l’estadístic ens fa falta un test com el two-sample t-test. **Ex. 2 - Com ho calcularíeu?** El resultat hauria de sortir tal que això:



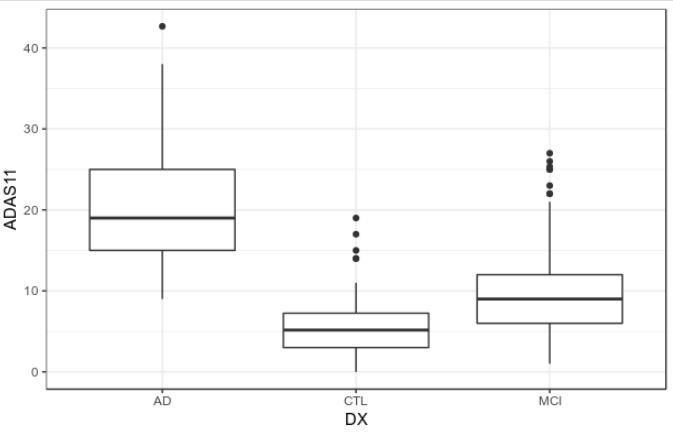
**Ex 3 - Ara, trobeu una variable on sí que hi hagi diferències significatives. Extraieu-ne el box-plot i l’estadístic.**

## Amb N grups

Podria ser que enlloc de tenir només dos grups, en tinguéssim N. En aquest cas, com ja sabeu, enlloc d’aplicar repetidament two-sample t-tests necessitem un altre test. Aquí ens centrarem amb la ANOVA.

Per exemple, volem estudiar les diferències d’ADAS11 (ADAS-cog, test neuropsicològic que mesura cognició general. Per a més info: <https://www.ncbi.nlm.nih.gov/pubmed/11218956>) a través dels diferents estadis de la malaltia d’Alzheimer (CTL, MCI i AD). Com ho faríeu?

**Ex1 – Boxplot ADAS11 per DX clínic**

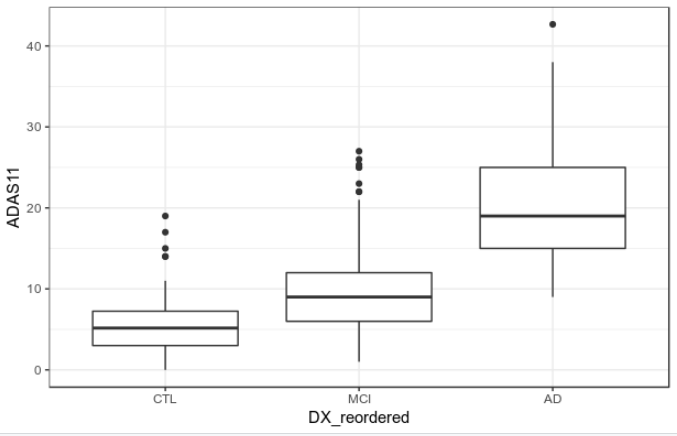


**PROBLEMA 1: l’ordre del gràfic no és el natural, no segueix la progressió de la malaltia (CTL – MCI – AD). Podem reordenar l’eix X? Si, amb la ordre factor:**

#Creem una nova variable que serà el DX reordenat

ALL\_ADNI$DX\_reordered <- factor(ALL\_ADNI$DX,levels=c("CTL","MCI","AD"))

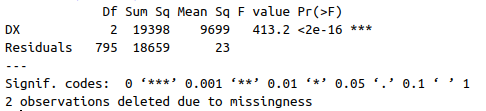
I ara sí, calculem el boxplot:



Visualment apreciem podria haver-hi diferències entre grups. Següent pas, calcular la ANOVA.

**Ex2 – ANOVA - ADAS11 per DX clínic**

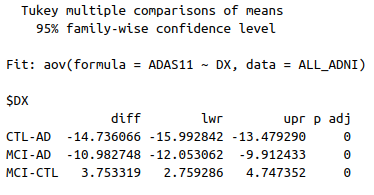
Llancem la ANOVA, que dona com a resultat el següent:



Com podeu observar surt significatiu, el que vol dir que hi ha diferències entre grups. Però quins són els grups que presenten diferències? Per respondre a això cal aplicar un test post-hoc. En el nostre cas farem un Tukey.

**Ex3 – Tukey post-hoc correction**

El resultat:



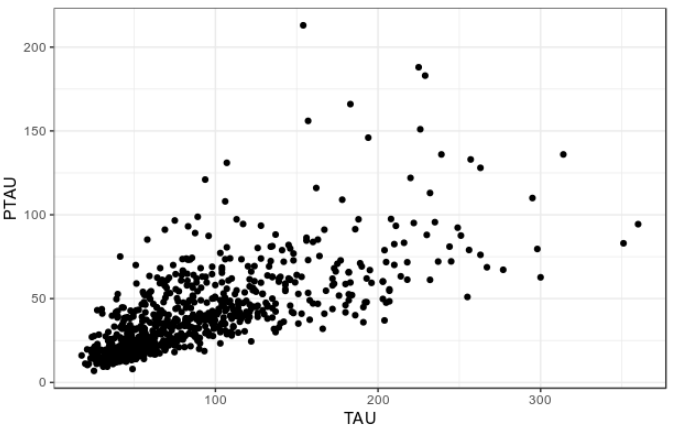
El two-sample t-test i la ANOVA són tests paramètric (i.e., per analitzar dades que segueixen una distribució normal). Si les dades no segueixen una distribució normal, què apliquem?

**Extra job: Realitza comparacions grupals per una variable no normal (per exemple, TAU). Quin seria el test a aplicar?**

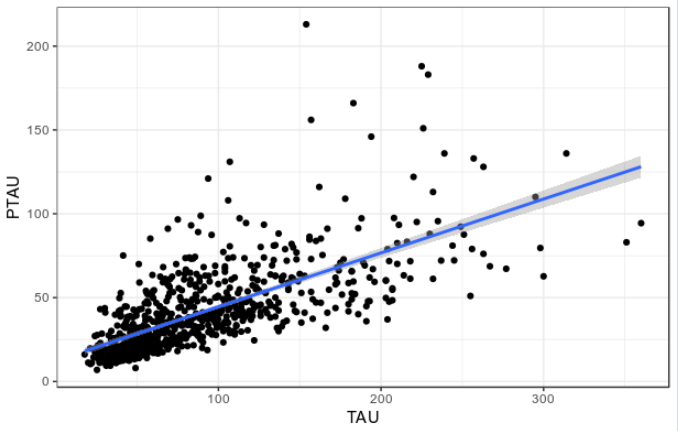
# Correlacions

La correlació ens proporciona la força i direcció d’una relació lineal i proporcional entre dos variables. Per exemple, plotegem en un gràfic les variables TAU (nivells de tau total en líquid cefalorraquídi) i PTAU (tau fosforilada) per al total de la mostra (CTL + MCI + AD), veiem que hi ha una forta relació:

**Ex1 – plotejar aquest gràfic**

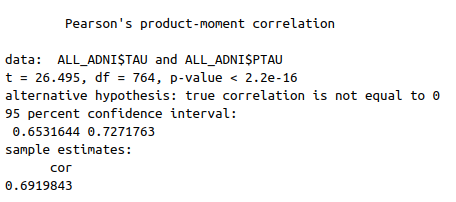


**Ex2 –** A més, podem afegir una línia de regressió amb una sola ordre mes. Com?



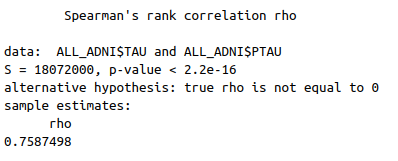
Efectivament, les dues variables estan correlacionades positivament. Com podem posar un estadístic?

**Ex2 – Calcula la r de Pearson i la p d’aquesta correlació**

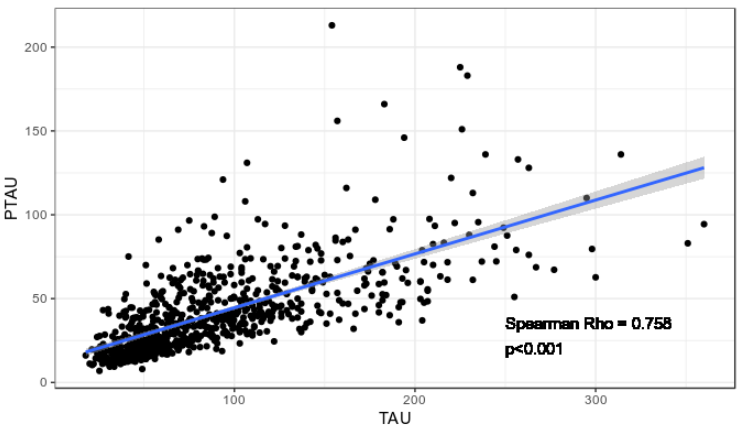


Tal com vèiem en el gràfic, existeix correlació entre les dues variables. Tot i això, alguns de vosaltres us n’haureu adonat de l’error. Hem aplicat un test paramètric a unes dades que no seguien una distribució normal. **ERROR!** El que havíem de fer és aplicar un Spearman.

**Ex3 – Calcula la rho de Spearman i la p:**



**Extra hard job: afegeix una etiqueta que mostri la R de correlació a dins del gràfic**



**Tip: Podem afegir una capa de text en un plot amb la següent ordre:**

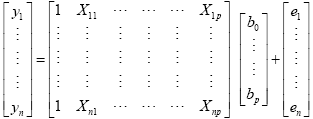
**geom\_text(aes(x=250, y= 35, hjust=0, label="text que vulguem escriure"))**

On:

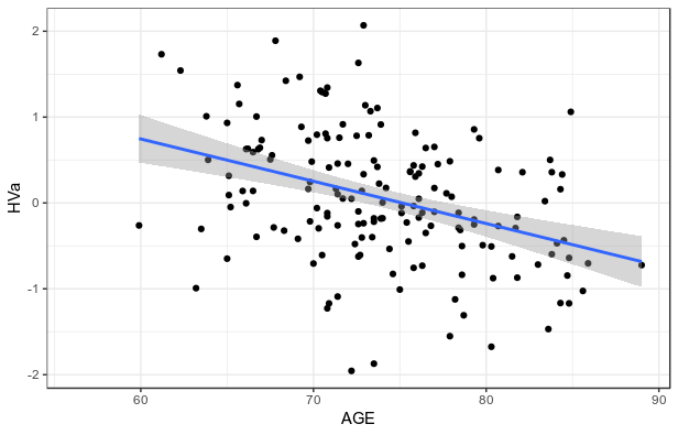
* x= coordenada x on col·loquem el text
* y= coordenada y on col·loquem el text
* hjust= justificació a l’esquerra
* label= text que volem escriure.

# Model general lineal

Un model general lineal és un model multivariant on avaluem la relació d’una variable resposta com a funció de múltiples variables. Concretament:



Per exemple, sabem que el l’edat provoca pèrdua de volum hipocampal en la població general:



Però, tenim la sospita que altres variables poden afectar el volum hipocampal, a part de la pròpia edat. Creem un model multivariant que inclogui les següents variables:

* TAU
* ICV (El volum total intracranial)
* Positivitat d’APOE e4 (portador vs no-portador)
* Edat
* Educació

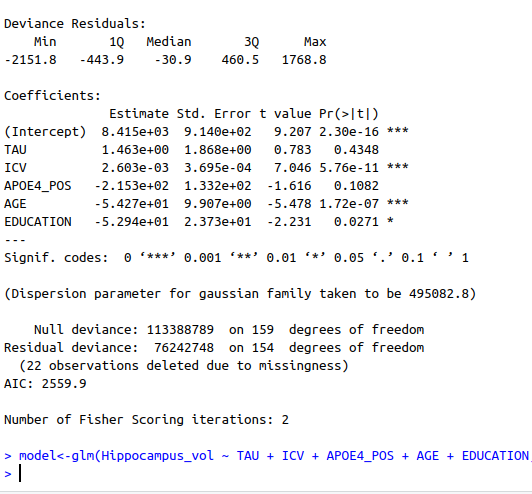
El codi seria el següent:

# Calcular un glm

model<-glm(Hippocampus\_vol ~ TAU + ICV + APOE4\_POS + AGE + EDUCATION, family = gaussian(), data=CTL)

Després, avaluem el resultat:

summary (model)



Veiem que hi ha 3 variables que són significatives:

* ICV: Volum total intracranial. A més volum intracranial, més volum hipocampal
* Age: la variable que ja sabíem
* Education: l’educació podria afectar al volum hipocampal

# Corbes ROC

Una corba ROC és una representació gràfica de la sensibilitat vs especificitat per un sistema de classificació binari en funció de com varia el llindar de classificació.

**Ex1 – Calcular la corba ROC de la captació de AV45 per discriminar pacients sans de controls amb pacients amb malaltia Alzheimer i demència. Utilitzar l’índex Youden per trobar el punt de tall òptim.**

#Calculem la corba ROC

rocobj<-roc(CTLAD$DX ,CTLAD$AV45\_quant\_Landau)

#Calculem l’àrea sota la corba

auc(rocobj)



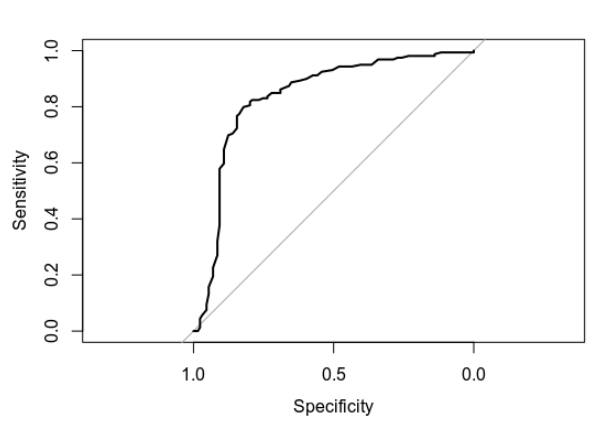
#Trobem el punt de tall

coords(rocobj,"best",best.method = "youden")



Podem plotejar la corba

plot(rocobj)



Si heu arribat fins aquí vol dir que heu resolt tota la pràctica. Enhorabona! A la 5ª sessió tornem amb més tasques!