Lliurament: **Divendres 28/02/2020**.

Introducció

En aquesta pràctica tractem algunes bases de dades que volem analitzar amb tècniques estadístiques. Previament verifiquem les bases de dades estudiant entre altres, la presencia de dades mancants, anomalies, zeros, i valorem la necessitat d'aplicar transformacions. A continuació treballem matrius bàsics de l'anàlisi multivariant, i realitzem estadística descriptiva multivariant.

Cada grup d'estudiants ha de realitzar el guió de la pràctica a continuació, realitzant els càlculs i gràfics necessaris en l'entorn R. Els resultats obtinguts i les respostes a les preguntes s'han de recollir en un document amb la vostre solució, fet per exemple en R markdown, Microsoft Word o Latex. Procureu posar els vostres noms i cognoms i número de grup a l'inici del document. Empleneu la mateixa numeració dels items de l'enunciat al document amb la vostre solució. S'ha de **lliurar la solució del qüestionari en format .pdf**, pujant-la a l'entorn Atenea (atenea.upc.edu), a l'apartat de la tasca corresponent, no mes tard que la data límit de lliurament.

Exercicis

1. (10p) Dades físiques d'estudiants i familiars.

   (a) En un estudi s'han recollit dades físiques (Pes, Edat, Alçada, Sexe entre d'altres) d'estudiants i els seus pares. Les dades es troben al fitxer `PesAlcada.txt`. Carregeu aquestes dades en l'entorn R amb l'instrucció `read.table`.

   (b) (1p) Quants estudiants hi ha a la base de dades? Quantes variables?

   (c) (1p) Investiguem la relació entre el pes de la mare (`Pesmare`) i l'alçada de la mare (`Alcmare`). Calculeu la mitjana i la desviació tipus d'aquestes dues variables.

   (d) (1p) Calculeu el coeficient de correlació entre les variables `Pesmare` i `Alcmare` amb la funció `cor`. Creeu que hi ha relació entrea les dues variables?

   (e) (1p) Feu boxplots de les variables `Pesmare` i `Alcmare` amb la funció `boxplot`. Què observeu?

   (f) (1p) Existeixen dades mancants en aquesta base de dades? Quantes existeixen a tota la base de dades? Quantes n'hi han per les variables `Pesmare` i `Alcmare`?

(g) (1p) Feu un diagram bivariant `Alcmare` versus `Pesmare`. Quants estudiants tenen dades mancants per alguna d'aquestes dues variables?

(h) (2p) Repetiu el càlcul del coeficient de correlació entre `Pesmare` i `Alcmare`, exclouent les dades mancants. Creeu que hi ha relació significativa entre les dues variables? (indicació: podeu realizar la regressió de `Pesmare` sobre `Alcmare` amb la funció `lm` i fer el contrast de hipòtesi corresponent).

(i) (1p) Substitiu les dades mancants de les variables `Pesmare` i `Alcmare` amb les mitjanes dels valors observats. Calculeu les mitjanes de les variables després de la substitució. Feu un diagrama bivariant de les noves variables.

(j) (1p) Calculeu de nou la correlació amb les dades mancants substituides. La substitució fa aumentar o disminuir la correlació? Argumenteu la resposta.

2. (10p) Dades demogràfiques de països.

Estudiem un segon joc de dades provenint d'un estudi de probresa utilitzant dades de les nacions units. Les dades es troben al fitxer `PovertyStudy.dat` (podeu descarregarles clicant en el nom del fitxer). El fitxer conté les variables:

- *Birth*: Live birth rate per 1,000 of population
- *Death*: Death rate per 1,000 of population
- *Infant*: Infant deaths per 1,000 of population under 1 year old
- *LifeEM*: Life expectancy at birth for males
- *LifeEF*: Life expectancy at birth for females
- *GNP*: Gross National Product per capita in U.S. dollars
- *Country*: Name of the country

(a) Llegiu les dades a l'entorn R amb `read.table`.

(b) (1p) Exploreu les relacions entre les variables amb la instrucció `pairs`

(c) (1p) La variable GNP té dades mancants. Com estan codificades? Quants països no tenen GNP?

(d) Substituiu el codi d'una data mancant pel codi que R utilitza per indicar dades mancants.

(e) (1p) Feu un boxplot i un normal probability plot de GNP. Comenteu la distribució de GNP (Feu servir les instruccions `boxplot` i `qqnorm`)

(f) (1p) Busquem una tranformació que normalitzi GNP, utilitzant la transformació de Box-Cox, mitjançant l'instrucció `boxcox(GNP~1,lambda = seq(-1, 1,by=0.1))`. Quina transformació cal utilitzar? Quin és el valor del parametre de transformació?

(g) (1p) Fem un boxplot de la variable transformada. Com ha canviat la seva distribució?

(h) (1p) Ajustem la regressió lineal múltiple de GNP (transformada) sobre les altres variables de l'estudi. Quin tan percent de GNP queda explicada pel model?

(i) (1p) Calculeu els valors predits del model pels països dels quals el GNP era mancant, utilitzant la funció `predict`.

(j) (1p) Calculeu, fent servir la funció `anova`, la variança dels residus del model ajustat.

(k) (2p) Doneu la instrucció `set.seed(123)`, per fixar la llavor del l'algorisme de generació de números aleatòris. Després genereu tantes observacions d'una distribució $N(0, s^2)$, amb s^2 és la variança dels residus calculada, com dades faltants teniu en GNP, fent servir `rnorm`. Calculeu noves prediccions per les dades mancants que inclouen aquest soroll.

3. (10p) Dades d'infarts de miocardi.

   Considerem dades de persones que van patir un infart de miocardi. Les variables recollides en l'estudi són:

   - *freqcard*: la frequencia cardiaca
   - *indcard*: l'index cardíac
   - *indsyst*: l'index systòlic
   - *prdiasto*: la pressió diastòlica (mínima)
   - *prartpul*: la pressió de l'arteria pulmonar
   - *prventri*: la pressió ventricular
   - *resipulm*: la resistència pulmonar
   - *pronosti*: 1=sobreviu, 2=decés

   Les dades es troben al fitxer `Infart.csv` (podeu descarregarles clicant en el nom del fitxer, o bé llegir el fitxer directament dins l'entorn R). Feu els càlculs i els gràfics que es demanen a continuació.

   (a) Podeu llegir el fitxer directcament dins R amb la funció `read.csv`. Utilizeu l'argurment `sep=;` per separar correctament els valors de les variables. Substitiu dades mancants codificats amb -1 per NA.

(b) (1p) Quantes persones conté la base de dades? Calculeu el vector de mitjanes de les 7 variables quantitatives. Calculeu aquest vector també per separat pels que sobreviuen (`pronosti==1`) i no.

(c) (1p) Calculeu la matriu de dades centrades. Mostreu les primeres 6 files d'aquesta matriu amb l'instrucció `head`

(d) (1p) Calculeu la matriu de variancies i covariancies. Quina variable té la major variancia? Són comparables les variancies de les variables?

(e) (1p) Calculeu la matriu de correlacions. Quines variables tenen una associació lineal forta?

(f) (1p) Feu el scatter plot matrix de la matriu de dades. Quines variables tenen una clara relació directa?

(g) (1p) Creeu que, a la vista dels diagrames bivariants, és recomanable transformar alguna/algunes de les variables? Argumenteu la resposta.

(h) (1p) Calculeu la matriu de dades estandaritzades, mostrant les primeres 6 files amb `head`.

(i) (1p) Calculeu la matriu de variancies i covariancies de les dades estandaritzades, amb l'instrucció `cov`. Que observeu?

(j) (1p) Calculeu la matriu de distàncies Euclideanes entre els individus, a partir de les dades originals, sense incloure la variable `pronosti`. Mostreu les distàncies de les primeres 5 observacions a la base de dades.

(k) (1p) Repetiu el càlcul de la matriu de distàncies Euclideanes entre els individus, a partir de les dades centrades, i també partir de les dades estandaritzades, sense la variable `pronosti`. Mostreu per cada cas les distàncies de les primeres 5 observacions. Que observeu?

4. (5p) Dades d'un qüestionari de salut amb variables categòriques.

   En un qüestionari de salut es va enregistrar la franja d'edat de l'entrevistat i la seva salut autopercebuda (VG=very good, G=good, R=regular, B=bad, VB=very bad). A partir de les dades, s'ha compilat la taula de contingencia a continuació.

	VG	G	R	B	VB
16-24	243	789	167	18	6
25-34	220	809	164	35	6
35-44	147	658	181	41	8
45-54	90	469	236	50	16
55-64	53	414	306	106	30
65-74	44	267	284	98	20
75+	20	136	157	66	17

   (a) Entreu les dades en R utilitzant la funció `matriu`.

   (b) (1p) Quina és la mida de la mostra d'aquest estudi?

   (c) (1p) Calculeu els percentatges de persones en les 5 categories de salut, i en les 6 franjes d'edat.

   (d) (1p) Feu servir la funció `barplot` per constuir un diagrama de barres dels comptatges a la taula de contingència.

   (e) (1p) Calculeu els percentatges de persones de certa franja d'edat dins cada categoria de salud. Feu de nou un diagrama de barres per comparar les diferentes categories.

   (f) (1p) Existeix relació entre les dues variables categòriques? Interpreteu els resultats obtinguts.