PRÀCTICA: REGRESSIÓ LINEAL

TASCA PER A ENTREGAR: INSTRUCCIONS

Heu d'entregar al Campus Virtual un únic fitxer `.zip`, contenint:

- *L'informe*, un document amb format i extensió \texttt{.pdf} on hi ha d'haver les explicacions i els resultats "en net" dels càlculs. Aquest fitxer ha de ser precisament en format `.pdf`; NO valen els formats `.doc`, `.docx`, etc., de processadors de text, cal convertir-lo a `.pdf` abans d'entregar-lo.

- *El codi*, un fitxer de text ASCII planer, amb extensió .r, amb el codi R i, en línies consecutives comentades amb el caràcter #, els resultats tal i com surten del programa. El codi ha de ser complet, que es pugui executar a efectes de verificació i també ha de contenir els comentaris suficients (què són les variables, quins valors poden prendre) per fer-lo comprensible.

Alternativament, en cas que esteu familiaritzats amb els notebooks de python i la seva adaptació a un kernel de R i així ho desitgeu, podeu entregar un únic fitxer `.ipynb`, contenint els dos elements, informe i codi.

Per facilitar l'avaluació, convé que el nom de tots els fitxers que entregueu sigui el mateix (cadascun amb la corresponent extensió), format amb el vostre nom. Així:

`Cognom1.Cognom2.Nom.pdf` ,

sense lletres accentuades o altres caràcters >`ASCII128`.

Expliqueu amb prou detall què feu i els resultats que aneu obtenint. Si bé hi ha uns resultats correctes de cada qüestió, hi ha moltes maneres correctes d'arribar-hi, algunes molt planeres i d'altres més sofisticades o elegants. La qualificació de la tasca tindrà en compte la correcció dels resultats però també aquest aspecte.

TASCA PER A ENTREGAR:

1 – Amb el dataset `Boston` del package `MASS`. Dades per a la predicció del preu d'habitatges a suburbis de la ciutat que li dóna nom. La variable `medv` és la resposta a predir i les altres són els predictors. En el help del dataset hi ha la descripció de les variables. Observeu que algunes són característiques pròpies de l'habitatge, com nombre d'habitacions, altres són socioeconòmiques i altres de caire geogràfic o ambiental.

Feu una descripció de les variables, primer individualment, fent `summary` de cada variable, boxplot i histograma. Tenen aspecte de normalitat? O bé són variables molt asimètriques? Mireu les correlacions entre predictors individuals i resposta. Anirà bé fer:

`round(cor(Boston),2)` o, fins i tot, `round(cor(Boston),1)`

per veure d'un cop d'ull quines correlacions són més grans i més petites. I les correlacions entre parelles de predictors? Hi ha perill de multicolinealitat? Pot ser útil fer la mateixa operació amb els logaritmes de les variables, atès que de vegades la transformació logarítmica permet fer millors prediccions. En aquest dataset hi ha dues variables que poden ser 0. Quines són? Un cop separades podeu mirar les correlacions dels logaritmes de les variables.

Ajusteu un model de regressió lineal de la resposta `medv` sobre les altres variables. A partir del `summary` del model, es pot afirmar que el model ajusta bé? Quines variables apareixen més o menys bones predictores? Prepareu un model més simple amb els cinc predictors que tenen un p-valor més petit. Amb aquest, ha pitjorat molt la suma de quadrats residual? Observeu que aquest nou model ara té una variable predictora que surt no significativa. Descarteu-la. Ajusteu ara els dos models de regressió lineal, de la resposta `medv` i del logaritme `log(medv)` sobre els logaritmes dels quatre predictors. Quin és el millor model?

2 - De la pàgina del llibre An Introduction to Statistical Learning with Applications in R, descarregueu el fitxer `Advertising.csv`, que s'ha fet servir a la lliçó 3 de teoria. Llegiu-lo des de R fent:

```
Advertising<-read.csv("Advertising.csv")
```

Podeu suprimir la variable `X`, que és el número d'ordre de cada cas, fent:

```
Advertising<-Advertising[,-1],
```

de totes formes ja està contingut a `row.names(Advertising).`

Amb aquest conjunt de dades, resseguiu l'ajustament a diferents models, tal i com es fa a la lliçó 3 de teoria o al capítol 3 del llibre ISLR.

3 – Dataset `Auto` del package `ISLR`. Vegeu la descripció al help. Seguint línies semblants als exemples fets a classe i a l'estudi de les dades `Boston`, feu un estudi de les variables, individualment i per parelles, seguit d'una anàlisi de regressió lineal de la resposta `mpg` (consum del vehicle en milles per galó de combustible) sobre les altres variables com a predictores. Trobeu un subconjunt de predictors òptim, justificant els passos fins a arribar-hi. Proveu també la possibilitat de transformacions no lineals, com log o potències dels predictors i/o de la resposta.

Indicacions: La variable 9, `name`, és el nom del model de cotxe. Serà millor treure-la del dataset, atès que de la majoria de models només hi ha un exemplar. Una altra variable que requereix atenció és `origin` (1 = Americà, 2 = Europeu, 3 = Japonès). Comproveu, fent:

```
str(Auto$origin),
```

que ve codificada numèricament. Tot i que, plausiblement, aquesta variable no deu ser bona predictora del consum del vehicle, si voleu la podeu introduir a la regressió, però convertint-la en un factor:

```
Auto$origin<-as.factor(Auto$origin)
```

Segurament serà millor descartar `origin` en una primera fase, a fi de fer més còmodament la selecció de predictors. Eventualment se'l podria re-introduïr per confirmar que no és un bon predictor o bé decidir, per exemple, si els vehicles japonesos són més eficients que els americans i europeus.