PiE 2: Entregable 2

Instruccions :

- Les solucions a l'entregable s'han de pujar a ATENEA. La data límit és el dilluns 11 a les 23.50h.
- Feu la pràctica mantenint els mateixos grups (parelles) que vau fer servir a l'Entregable 1. En cas que no sigui possible, feu-nos-ho saber (com a molt tard dilluns dia 4) per correu electrònic, i si està ben justificat es pot fer una excepció.
- Hi ha cinc conjunts de dades diferents. Els arxius acaben amb un nombre que va de l'1 al 5. A l'arxiu parelles.pdf hi trobareu el conjunt de dades que heu d'analitzar, identificats amb el nombre corresponent.
- Heu d'entregar un informe en formar .pdf que inclogui la solució als exercicis i el codi que heu fet servir (integreu el codi a les vostres explicacions, NO bolqueu tot el codi en un apèndix). Feu servir Rmarkdown per crear el fitxer amb les solucions.
- Heu de JUSTIFICAR TOTES LES RESPOSTES, altrament no es donarà com a vàlida.
- Si teniu cap dubte podeu contactar-nos tant al Víctor com a mi.
- Els exercicis els podeu trobar a la segona pàgina d'aquest document. Que us vagi bé.

Exercici1: Es vol predir la longitud del sèpal de plantes de IRIS en funció de l'amplada del sèpal i de l'espècie de IRIS. Llegiu les dades i contesteu les preguntes següents (restringiu-vos a l'objectiu esmentat d'aquest exercici i a les variables esmentades):

1) Especifiqueu quina és la variable resposta i quines son les explicatives així com el tipus de variable del que es tracta.

2) Quines son les preguntes que té sentit contestar?

3) Té sentit parlar d'interacció en aquest exercici?. Si és que sí, com l'interpretarieu?

4) Dibuixeu de forma exploratòria el vostre conjunt de dades i en base al gràfic o gràfics realitzats, treieu unes primeres conclusions, especifiqueu-ne **tres** (caldrà comprovar-les més endavant).

5) Ajusteu un model de regressió lineal simple (model 1) i contesteu les següents preguntes:
   5.1) Quina és l'estimació de la variància que heu obtingut?
   5.2) Quina part de la variabilitat de la variable resposta és explicada pel model?
   5.3) Quin és el residu més gran que heu trobat?
   5.4) Hi ha valors que tinguin un leverage superior al permès (feu servir la fita més acurada)? Quants?
   5.5) Feu el plot dels residus i comenteu **tres** coses de les que considereu importants.

6) Ajusteu un model que tingui com a explicatives les dues variables esmentades a l'enunciat i que no son la variable resposta (model 2).
   6.1) Quants paràmetres té el vostre model? Tots els paràmetres son significatius?
   6.2) Com interpreteu els paràmetres obtnguts? (una frase per a cada paràmetre).
   6.3) Quina part de la variabilitat de les dades explica el vostre model?
   6.4) Quina és l'estimació de la variància residual que obteniu?
   6.5) Quina és la matriu del disseny (matriu X) associada al model que esteu ajustant?
   6.6) Feu els plots dels residus que creieu necessaris i treieu-ne tres conclusions que considereu importants.
   6.7) Creieu que és necessàri treure alguna observació del conjunt de dades?. Justifiqueu la vostra resposta. En cas que sigui necessàri feu-ho i torneu a ajustar el model.

7) Amb el conjunt de dades que us hagiu quedat al final de l'apartat 6), ajusteu un model que assumeixi que les dues variables explicatives del model 2 interaccionen (model3).
   7.1) Us surt significativa la interacció?
   7.2) Quina és l'estimació de la variància residual?
   7.3) Interpreteu els paràmetres que us surtin significatius (una frase per a cada paràmetre).
   7.4) Han millorat els gràfics de residus respecte als trobats amb el model 2? Justifiqueu la vostra resposta.

9) Feu una taula comparativa dels tres models ajustats i que us permeti decidir quin dels tres és el millor. Justifiqueu en paraules la vostra elecció.

10) Pel model escollit, prediu quina és la longitud esperada d'un sèpal que tingui una amplada de 38.00 per a cadascúna de les espècies que tenim. Expliciteu tant l'estimació puntual com l'intèrval de confiança al 95%.

Exercici2 : Treballarem el mateix conjunt de dades que a l'Exercici 1. Ara però volem predir la Longitud d'un pètal. Podeu fer servir la resta de variables que tenim com a explicatives.

1) Porteu a terme l'scatterplotmatrix de les dades. Què observeu? (esmenteu **tres** coses que creieu importants).

2) Ajusteu un model de regresió linel múltiple amb les explicatives que teniu (model 4). Per a aquest model contesteu les preguntes següents:

   2.1) Quin és el valor de R^2 i R^2_{adj}?

   2.2) Quins son els residus mínim i màxim?

   2.3) Hi ha alguna observació amb leverage?

   2.4) Hi ha alguna observació influent?

   2.5) Quina variable té el VIF més gran? És acceptable aquest valor?

   2.6) Interpreteu els paràmetres associats a totes les variables que creieu us surtin significatives (una frase per a cada paràmetre).

   2.7) Podem afirmar que el coeficient de la variable Amplada del pètal és estadísticament igual a 1.5?

3) Amplieu el model anterior afegint la variable espècie (model5), i justifiqueu quin dels dos models (m4 i m5) és millor (feu una taula comparativa).

4) Pel model que hagiu escollit,

   4.1) Quants residus estandaritzats cauen fora de l'inferval $(-2, 2)$? Creieu que és un percentatge escaient?

   4.2) Quina és la longitud del pètal esperada, per a una flor que tingui una amplada del pètal igual a 2.1, una longitud del sèpal igual a 5 i una amplada del sèpal igual a 3.5? Trobeu l'estimació puntual i per interval de confiança per a cadascúna de les espècies.