UNIVERSITAT DE BARCELONA

Departament d'Estadística
Facultat de Biologia

Diagonal, 643
08028 Barcelona
Tel. +34 934 021 560
Fax +34 934 111 733
fcarmona@ub.edu

Models Lineals **EXAMEN B** **9 de gener de 2018**

Grau d'Estadística

S'ha d'entregar un document PDF o HTML amb les respostes i un arxiu R ben ordenat per apartats i amb tots els comentaris (amb el símbol #) que considereu oportuns. En el nom dels fitxers no poseu accents, només símbols ASCII del 0 al 127 (sense ñ,ç,...).

Problema 1

S'ha realitzat un experiment amb quatre situacions experimentals diferents i algunes rèpliques per a cadascuna. Les dades són:

$$\begin{aligned} \alpha+\beta &\Rightarrow -1.25, -1.10 \\ \alpha+2\beta+\gamma &\Rightarrow 0.18, 0.20, 0.15 \\ 2\alpha+3\beta+\gamma &\Rightarrow -1.30, -1.28, -1.31 \\ \beta+\gamma &\Rightarrow 1.12, 1.15 \end{aligned}$$

contesteu les següents qüestions:

(a) Quina condició ha de verificar una funció paramètrica per a que sigui estimable en aquest model?

(b) Indiqueu si les funcions paramètriques següents són estimables i calculeu l'estimador MQ quan sigui possible:

$$\text{(i) } 3\alpha-\beta-4\gamma \qquad \text{(ii) } \alpha+2(\beta+\gamma)$$

(c) Calculeu l'estimació de la covariància entre els estimadors lineals òptims de $\alpha-\gamma$ i $\beta+\gamma$ i la variància de l'estimador lineal òptim de $\alpha+2\beta+\gamma$.

(d) Feu el contrast de la hipòtesi $H_0 : \beta = 2\alpha+3\beta$.

Problema 2

Les dades d'aquest problema formen part d'un estudi més gran del Dr. Rick Linthurst (1979) de la Universitat de Carolina del Nord que va ser la seva tesi doctoral. El propòsit d'aquesta investigació va ser identificar les característiques del sòl que influeixen en la producció de biomassa aèria de l'herba del pantà *Spartina alterniflora* a l'estuari del Cap Fear de Carolina del Nord.
El mostreig va incloure tres pantans en el estuari (Oak Island, Smith Island, and Snows Marsh) i tres llocs en cada pantà que representaven tres ecosistemes: una àrea on la Spartina va morir però recentment s'havia regenerat (dead), una àrea amb Spartina baixa (short) i una àrea amb Spartina alta (tall). En cadascú dels nou llocs, combinació de localització i vegetació, es van escollir aleatòriament cinc punts mostrals (45 en total) per mesurar la biomassa aèria i un conjunt de propietats fisicoquímiques del substrat en un programa mensual.
Les dades que estudiarem són només les de setembre i les mesures del substrat són:
1. free sulfide (H_2S), moles
2. salinity (SAL),
3. redox potentials at pH 7 ($Eh7$), mv
4. soil pH in water (pH), 1:1 soil/water
5. buffer acidity at pH 6.6 (BUF), meg/100 cm^3
6. phosphorus concentration (P), ppm

7. potassium concentration (K), ppm
8. calcium concentration (Ca), ppm
9. magnesium concentration (Mg), ppm
10. sodium concentration (Na), ppm
11. manganese concentration (Mn), ppm
12. zinc concentration (Zn), ppm
13. copper concentration (Cu), ppm
14. ammonium concentration (NH_4), ppm.

La variable dependent és la biomassa aèria gm^{-2}.

L'objectiu principal d'aquest estudi és identificar les variables del substrat amb major relació amb la biomassa. En una primera aproximació utilitzarem totes les dades sense tenir en compte les diferents localitzacions i vegetacions.
Les dades les obtenim així:

```
linthurst <- read.table("LINTHALL.DTA",
                        col.names=c("loc","type","biomass","H2S","salinity","Eh7",
                                    "pH","BUF","P","K","Ca","Mg","Na","Mn",
                                    "Zn","Cu","NH4"))
```

(a) Calculeu la regressió mínim-quadràtica ordinària de les 14 mesures del sòl sobre la biomassa.

Quines són les variables regressores amb coeficient de regressió significatiu?

Són aquestes les variables més importants per explicar els resultats de biomassa?

(b) Quin és el millor model amb només 3 variables regressores segons la funció `regsubsets`?

Quin és el millor model segons el valor AIC?

Quin és el millor model segons el coeficient de determinació ajustat?

Quin és el millor model segons el coeficient C_p?

(c) Quines són les variables regressores que queden al model si fem una selecció "backward"?

Quin és el criteri que has fet servir?

Quin és el resultat si fas servir una selecció "backward" amb el criteri del test F i el seu p-valor associat?

Quin és el resultat si fas servir una selecció "forward"? Per això cal definir el model mínim i el paràmetre `scope` amb el model màxim.

(d) Si decidim combinar els resultats dels dos apartats anteriors i només volem retenir 4 variables, quines són les 4 variables més importants per explicar la biomassa? Com queda el coeficient de determinació ajustat respecte el model complert? Guanyem amb eficiència en l'estimació dels coeficients de regressió?

(e) La selecció de variables en regressió és molt sensible a valors atípics o influents. Feu una anàlisi dels residus i, en especial, un gràfic de bombolles amb les distàncies de Cook.

Problema 3

En les mateixes dades de l'estudi de Linthurst hi ha un greu problema de multicolinealitat per a les 14 variables regressores.

(a) Calculeu la matriu de correlacions entre les variables regressores. Quantes són en valor absolut superiors a 0.7 i quantes $|r| \geq 0.9$?

Nota: Observeu que la matriu de correlacions és simètrica i a la diagonal hi ha uns.

Quants factors d'inflació de la variància de les variables regressores són superiors a 4?

(b) Calculeu els valors propis i vectors propis de la matriu de correlacions de les 14 variables regressores. El resultat és equivalent a les components principals de les variables centrades i estandarditzades.

Quines variables regressores defineixen principalment a la primera component principal? I a la segona?

Quin és el percentatge de variabilitat explicat per les components amb valor propi superior a 1?

Quin és el percentatge de variabilitat explicat per les 6 darreres components principals?

Quin és el número de condició de la matriu de correlacions? (Es considera que un valor superior a 10 és símptoma de colinealitat.)

Calculeu la mesura de colinealitat de Thisted (1980)

$$\text{mci} = \sum_{j=1}^{14} \lambda_j^{-2} \lambda_{14}^2$$

(Valors de mci pròxims a 1.0 indiquen forta colinealitat, mentre que valors superiors a 2.0 indiquen poca o gens colinealitat.)

(c) Per tal de calcular la regressió per components principals, quin sembla el número adequat de components a fer servir?

Calculeu la PCR i doneu el seu R^2 i el màxim VIF. Valoreu el resultat.

(d) Calculeu la regressió PLS. Quin és el número de components adequat?

Donat l'objectiu principal d'aquest estudi explicat en el problema anterior, et sembla adequada la regressió PLS per assolir-ho?

Problema 4

En el paquet `faraway` hi ha una base de dades anomedada `globwarm` que conté temperatures anuals de l'escalfament global del planeta, conjuntament amb representacions a través dels anells dels arbres en diferents regions del planeta.
La variable resposta és `nhtemp`: Northern hemisphere average temperature (C) que proporciona la oficina meteorològica del Regne Unit (coneguda com HadCRUT2). Aquesta temperatura es calcula des de 1856 i en el *data.frame* tenim els valors fins l'any 2000. Així doncs, les observacions d'anys anteriors no es poden utilitzar en el model.
L'objectiu és calcular un model de regressió per a la predicció de la temperatura mitjana global en funció de les 8 "temperatures" basades en els anells dels arbres.
Sospitem que aquestes dades temporals poden tenir problemes d'autocorrelació que caldrà investigar i corregir. En la pàgina `https://onlinecourses.science.psu.edu/stat501/node/360` ens proposen algunes solucions. En particular el procediment de Cochrane-Orcutt.

(a) Calculeu el model de regressió `lmod` i els seus residus e_t que anomenarem `RESI1`.

Dibuixeu el gràfic de dispersió amb els residus per a cada any. Què observem?

Dibuixeu el gràfic de dispersió (e_{t-1}, e_t) amb l'ajuda de la funció `LAG` que definim així:

```
LAG <- function(x)  c(NA,x[-length(x)])
```

de forma que si `RESI1[t]`$= e_t$, llavors `LAG(RESI1)[t]`$= e_{t-1}$. Què observem?

Estimeu el pendent $r = \hat{\rho}$ de la regressió simple

$$e_t = \rho e_{t-1} + \text{error}$$

i dibuixeu la recta de regressió en el gràfic anterior.

(b) Contrasteu la presencia d'autocorrelació entre els residus e_t del model de regressió `lmod` anterior amb el test de Durbin-Watson.

(c) Calculeu les variables transformades $y_t^* = y_t - ry_{t-1}$ i $x_{t,j}^* = x_{t,j} - rx_{t-1,j}$ per a $j = 1, \ldots, 8$. Teniu un exemple de com fer-ho a la pàgina web esmentada. La funció `apply` de R us pot ajudar.

Nota: Millor si treballeu amb el *data.frame* `globwarm[857:1001, -10]`.

Calculeu llavors la regressió múltiple amb aquestes variables transformades i proveu amb els seus residus si hem corregit el problema d'autocorrelació.

(d) Calculeu la constant de regressió (intercept) segons el procediment de Cochrane-Orcutt i feu una predicció de la temperatura global del planeta l'any 1000.