EXAMEN DE MODELS LINEALS **Prova A** **19 de gener de 2015**

Els alumnes amb el primer parcial aprovat han de fer els exercicis 2, 3 i 4.
La resta han de fer els exercicis 1, 2 i 3.

Problema 1 (3 punts)

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

$$\begin{array}{lcrrr} \alpha+\beta & \Rightarrow & -1.25 & -1.28 & -1.24 \\ \alpha+2\beta+\gamma & \Rightarrow & 0.18 & 0.19 & 2.00 \\ 2\alpha+3\beta+\gamma & \Rightarrow & -1.30 & -1.28 & -1.31 \\ \beta+\gamma & \Rightarrow & 1.12 & 1.13 & 1.11 \end{array}$$

contesteu les següents qüestions:

1. Quina condició ha de verificar una funció paramètrica per a que sigui estimable en aquest model?
2. Indiqueu si les funcions paramètriques següents són estimables i calculeu l'estimador MQ quan sigui possible:

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

3. Calculeu l'estimació de la covariància entre els estimadors lineals òptims de $\alpha-\gamma$ i $\alpha+\beta$ i la variància de l'estimador lineal òptim de $\alpha+2\beta+\gamma$.
4. Feu el contrast de la hipòtesi $H_0:\gamma=2\alpha+\beta$.

Problema 2 (4 punts)

En la base de dades UCI-Machine Learning Repository trobarem les dades d'un estudi sobre el càncer de mama a Wisconsin fet pel Dr. Wolberg sobre 1984 que inclou només aquells casos que presenten un càncer de mama invasiu sense evidència de metàstasi en el moment de la diagnosi. La pàgina web és

`http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29`

Les 35 variables anotades per a cada pacient són:

1. ID number
2. Outcome (R = recur, N = nonrecur)
3. Time (recurrence time if field 2 = R, disease-free time if field 2 = N)

4-33. Ten real-valued features are computed for each cell nucleus:

   (a) radius (mean of distances from center to points on the perimeter)
   (b) texture (standard deviation of gray-scale values)
   (c) perimeter
   (d) area
   (e) smoothness (local variation in radius lengths)
   (f) compactness (perimeter^2 / area - 1.0)
   (g) concavity (severity of concave portions of the contour)
   (h) concave points (number of concave portions of the contour)
   (i) symmetry
   (j) fractal dimension ("coastline approximation" - 1)

34. Tumor size
35. Lymph node

Les 10 característiques per a cada nucli cel·lular es van calcular a partir d'una imatge digitalitzada d'una aspiració amb agulla fina d'una massa de mama. Per a cada característica tenim $10 \times 3 = 30$ variables ja que mesurem la mitjana (mean), el error estàndard (se) i la "pitjor" o més gran (mitjana dels tres valors més grans) d'aquestes característiques per a cada imatge. D'aquí les 30 variables. Per exemple, el camp 4 és la mitjana del radi, el camp 14 és l'error estàndard del radi i el camp 24 és el pitjor radi.
Les següents instruccions de R permeten polir la base de dades original. Observeu que cal eliminar 4 files que contenen valors *missing*. A més, posem els noms a les variables i a les mostres i llavors eliminem la primera columna de identificadors.

```
link <-
"http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.data"
wpbc <- read.csv(link, header=FALSE, na.strings="?")
wpbc <- na.omit(wpbc)
names_wpbc <- read.table("wpbc_names.data", header=FALSE, colClasses = "character")
colnames(wpbc) <- names_wpbc[ ,1]
rownames(wpbc) <- wpbc$ID
wpbc <- wpbc[ ,-1]
```

L'objectiu és predir la variable `Time` en funció de les 10×3 característiques de la imatge digital més les variables `Tumor_size` i `Lymph_node`.

1. Calculeu l'hiperplà de regressió i el coeficient de correlació múltiple de `Time` sobre les altres variables. Quina és la variància estimada de l'error?
2. És un model amb un bon ajust? Vol dir això que és significativa la regressió? Concreta què significa cada pregunta.
3. Seleccioneu només els registres on `Outcome` és recurrent i repetiu la regressió. Valoreu l'ajust ara.
4. Amb els gràfics o els estadístics adients, investigueu la diagnosi d'aquest darrer model en el següents punts:
   (i) Variància constant dels errors.
   (ii) Hipòtesi de normalitat.
   (iii) Punts amb influència potencial (leverage).
   (iv) Outliers.
   (v) Punts influents.
   (vi) Creieu que pot haver un problema de multicolinealitat? En què us baseu?

   Concreteu els punts problemàtics.
5. Què podem dir del punt 859223? En què millora el model si eliminem aquest punt de les dades? I què podeu dir del 8611792?
6. Contrasteu si els coeficients de regressió de les variables `smoothness_mean` i `concavity_mean` són iguals.
7. Són significatives conjuntament les variables `Tumor_size` i `Lymph_node` o podem prescindir d'elles.

Problema 3 (3 punts)

Amb la mateixa base de dades del problema anterior (només amb els registres on `Outcome` és recurrent) i el mateix model de partida (amb totes les variables regressores), sembla que tenim un problema de multicolinealitat.

1. Trobeu el "millor" model per dos mètodes diferents de selecció de variables com, per exemple, AIC i C_p de Mallows.
   (a) Quines són les variables seleccionades?
   (b) Quins són els coeficients de determinació ajustats d'aquests models? Compareu-los amb el del model complet. Llavors, què hem guanyat?

(c) Calculeu l'interval de confiança al 95% per al coeficient de regressió de la variable `radius_mean` en els models, el complet i els seleccionats.

2. Un altre possibilitat és fer servir la Ridge Regression. Quins són els coeficients obtinguts?

   Expliqueu breument les avantatges i inconvenients d'aquest mètode front a la selecció de variables.

3. Amb el model reduït per AIC el punt 915143 és problemàtic. Ajusteu un model per un mètode robust adient per desactivar aquest problema.

Problema 4 (3 punts)

En estadística, el test Rainbow (Utts, 1982) és un test per contrastar la linealitat d'un model de regressió. La idea és comparar el model de regressió complet (full) amb un model idèntic però calculat només amb els punts de baixa influència (*low leverage points*), és a dir, els punts centrals de les dades.
Seleccionarem el subconjunt de les dades amb baixa influència o regió central de les dades amb el criteri[1]

$$leverage < \text{median}(leverage)$$

Llavors calcularem el següent estadístic:

$$F = \frac{(\text{SSE}_{\text{FULL}} - \text{SSE}_{\text{CENTRAL}})/(n-m)}{\text{SSE}_{\text{CENTRAL}}/(m-2)}$$

on n és el número total d'observacions en el model complet i m és el número de punts de baixa influència. Si la hipòtesi nul·la de linealitat és certa, llavors l'estadístic F segueix una distribució F de Fisher amb $n-m, m-2$ graus de llibertat.
Amb la base de dades `teengamb` del paquet `faraway` considerem el model lineal amb la variable `gamble` com a resposta i les variables `sex`, `income` i `verbal` com a regressores.
Contrasteu la linealitat d'aquest model amb el test Rainbow[2]. A quina conclusió arribem?

[1] Aquest no és l'únic criteri. També podríem escollir un altre percentatge de dades centrals.

[2] Encara que la funció `raintest` del paquet `lmtest` fa diferents tipus de tests del tipus Rainbow, feu els vostres propis càlculs.