EXAMEN DE MODELS LINEALS **20 de gener de 2014**

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.
Cada exercici es pot fer en una hora. El temps també és un factor en l'avaluació.

Problema 1

Una cadena d'empaquetatge està formada per cinc treballadors identificats amb els números 1 al 5, més un capatàs que treballa tot el temps. Cada dia registrem $X_j = 1$ si el treballador j és de servei i 0 en cas contrari, $Y =$ és el número de paquets despatxats.
La següent taula ens mostra les dades recollides. També les teniu en el fitxer `paquets.csv`.

X_1	X_2	X_3	X_4	X_5	Y
1	1	1	0	1	246
1	0	1	0	1	252
1	1	1	0	1	253
0	1	1	1	0	164
1	1	0	0	1	203
0	1	1	1	0	173
1	1	0	0	1	210
1	0	1	0	1	247
0	1	0	1	0	120
0	1	1	1	0	171
0	1	1	1	0	167
0	0	1	1	0	172
1	1	1	0	1	247
1	1	1	0	1	252
1	0	1	0	1	248
0	1	1	1	0	169
0	1	0	0	0	104
0	1	1	1	0	166
0	1	1	1	0	168
0	1	1	0	0	148

(a) Presenteu algun objectiu per recollir i analitzar aquestes dades. Plantegeu un model lineal que pugui assolir aquest objectiu.

(b) Feu la regressió de Y sobre $X_1, \ldots, X_5$. És possible que l'ordinador proporcioni algun resultat inesperat. Expliqueu-ne els motius. Trobeu una solució que eviti el problema anterior.

(c) Comproveu[1] que el rang de la matriu de disseny és 5 i identifiqueu les funcions paramètriques estimables.

(d) Podem saber la contribució estimada del capatàs? Quin és l'error estàndard d'aquesta estimació?

(e) Feu el contrast d'igualtat en la contribució dels treballadors 2,3 i 4.

(f) Creeu una nova variable $S = X_1 + \cdots + X_5$ i feu la regressió de Y sobre S. Quines són les estimacions dels paràmetres? Compareu els resultats amb els de l'apartat (d). Quina de les dues estimacions de la contribució del capatàs és correcte?

Problema 2

La concentració de lactat en sang serveix freqüentment per predir la resistència dels atletes. La identificació d'alguns predictors de la resistència pot ajudar als entrenadors i atletes a avaluar els canvis del seu rendiment.
Per identificar aquests predictors es van seleccionar vint-i-quatre dones ciclistes ben entrenades i se'ls va practicar una prova física en un cicloergonòmetre. La prova física va consistir en realitzar etapes de tres minuts fins l'esgotament. Als 30 segons d'haver finalitzat cada etapa de tres minuts, se'ls va

[1] Podeu fer servir les files 1,2,4,5,9.

treure sang capilar del dit índex per analitzar el lactat plasmàtic i obtenir els valors predictors següents: llindar de lactat (P-Tlac), DMax (DMax), LT-log-log (P-Tlac.II), P-4 mmol.L^{-1} (P-4mM) i taxa de treball corresponent a un increment de 1 mM del valor basal (Rise.1.PB). El rendiment de resistència (AV.Power) s'avaluà 7 dies més tard mitjançant una prova de bicicleta, d'una hora de durada, en la que les atletes havien d'aconseguir la potència de sortida més alta possible.
Els resultats es troben en el fitxer de dades `CyclingPower.xls`.

1. Calculeu l'hiperplà de regressió i el coeficient de correlació múltiple de AV.Power 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. Amb els gràfics o els estadístics adients, investigueu la diagnosi d'aquest 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.

4. Què podem dir del punt 21? En què millora el model si eliminem aquest punt de les dades? I què podeu dir del 14?

5. Contrasteu si els coeficients de regressió de les variables P-Tlac i P-Tlac.II són iguals.

6. Si una corredora té uns valors de P-Tlac=175, DMax=158, P-Tlac.II=131, P-4mM=206 i Rise.1.PB=182, quina és l'estimació del seu rendiment de resistència (AV.Power) i l'error estàndard d'aquesta estimació.

   Amb una probabilitat del 0.95, entre quins valors es troba el rendiment de resistència (AV.Power) d'una corredora concreta que té els valors anteriors?

7. Calculeu la correlació parcial entre AV.Power i Rise.1.PB si eliminem la informació de P-Tlac, DMax, P-Tlac.II i P-4mM.

Problema 3

Amb la mateixa base de dades del problema anterior i el mateix model de partida, 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 Rise.1.PB 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 de tres variables regressores (P.Tlac.ll, DMax i Rise.1.PB) el punt 21 encara fa nosa. Ajusteu un model per un mètode robust adient.

Problema 4 (Test de Breusch-Pagan)

Sota les condicions de Gauss-Markov, que inclou l'homocedasticitat, els mínims quadrats ordinaris proporcionen el millor estimador lineal no esbiaixat (BLUE), és a dir, no esbiaixat i eficient. Malgrat això, l'eficiència es perd en presència d'heterocedasticitat. Per estudiar la presència d'heterocedasticitat podem fer servir el test de Breusch-Pagan.
Donat el model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ assumim que l'heterocedasticitat pren la forma:

$$E(\epsilon_i) = 0 \qquad \sigma_i^2 = E(\epsilon_i^2) = h(\mathbf{z}_i'\boldsymbol{\alpha}) \qquad \text{per a tota } i = 1, \ldots, n$$

on $\mathbf{z}_i' = (1, z_{1i}, z_{2i}, \ldots, z_{pi})$ i $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_p)$ és un vector de coeficients desconeguts i $h(\cdot)$ és una funció no especificada que només pot prendre valors positius.
La hipòtesi nul·la d'homocedasticitat és:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$$

i sota aquesta hipòtesi tenim que $\sigma_i^2 = h(\alpha_0)$ constant.
Considereu el model a estimar per mínims quadrats ordinaris i assumiu la normalitat dels errors.

El procediment a seguir és el següent:

1. Apliqueu el mètode dels mínims quadrats al model considerat i calculeu els seus residus e_i. Estimeu la variància dels errors $\widehat{\sigma}^2 = \sum e_i^2/n$.
2. Feu una regressió auxiliar de la variable $e_i^2/\widehat{\sigma}^2$ com a resposta i les variables $\mathbf{z}$ com regressores i calculeu la suma de quadrats explicada ESS o suma de quadrats de la regressió.
3. L'estadístic del test és BP = ESS/2 amb distribució asimptòtica χ_p^2 i l'homocedasticitat es rebutja si l'estadístic supera el valor crític de la taula.

El procediment explicat requereix el coneixement de las variables regressores $\mathbf{z}$ i no és necessari conèixer la funció h. Sovint les regressores $\mathbf{z}$ són algunes de les regressores del model original.

Amb la base de dades `Ornstein` del paquet `car` considerem el model lineal amb la variable `interlocks` com a resposta i les variables `assets`, `sector` i `nation` com a regressores.

(a) Contrasteu l'homocedasticitat d'aquest model amb el test de Breusch-Pagan segons el procediment explicat i amb les mateixes regressores per al model auxiliar que pel primer model.

Calculeu l'estadístic i el seu p-valor.

(b) Compareu el resultat amb la funció `ncvTest` del paquet `car`. Configureu correctament el paràmetre `var.formula`.

(c) Calculeu el test amb la funció `bptest` del paquet `lmtest`. Configureu correctament el paràmetre `studentize`.

Nota 1: Observeu que algunes variables regressores són factors, de forma que en el model de regressió es transformen en vàries dicotòmiques i això modifica els graus de llibertat.
Nota 2: Una versió simplificada del test[2] demana la regressió de e_i^2 sobre $\mathbf{z}$, llavors l'estadístic és nR^2 d'aquesta regressió auxiliar. Aquí no farem servir aquesta aproximació.

[2] `http://en.wikipedia.org/wiki/Breusch-Pagan_test`