EXAMEN DE MODELS LINEALS **1 de febrer de 2013**

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

Considereu un model lineal amb $n = 4$ observacions i tres paràmetres β_1, β_2 i β_3. Suposem que

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 + \beta_3 + \epsilon_1 \\ Y_2 &= \beta_1 + \beta_3 + \epsilon_2 \\ Y_3 &= \beta_2 + \epsilon_3 \\ Y_4 &= 2\beta_1 - 3\beta_2 + 2\beta_3 + \epsilon_4 \end{aligned}$$

(a) Escriviu aquest model en la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ i resumiu les hipòtesis que cal fer per considerar-ho un model lineal normal.

(b) Caracteritzeu cadascuna de les següents funcions com estimable o no estimable: β_1, $\beta_1 - 2\beta_2 + \beta_3$, $\beta_1 - \beta_2 + \beta_3$. Justifiqueu la resposta.

(c) Escriviu les equacions normals d'aquest model i trobeu dues solucions $\boldsymbol{\beta}_1$ i $\boldsymbol{\beta}_2$ si $Y_1 = 3.03$, $Y_2 = 1.98$, $Y_3 = 1.02$ i $Y_4 = 0.97$.

(d) Calculeu les estimacions mínimo-quadràtiques de les funcions estimables de l'apartat (b). Verifiqueu que aquestes estimacions són invariants per a qualsevol solució com $\boldsymbol{\beta}_1$ o $\boldsymbol{\beta}_2$. Calculeu la covariància entre les estimacions.

(e) Feu el contrast de la hipòtesi $H_0 : \beta_1 + \beta_3 = 2$.

Problema 2

La base de dades `teengamb` del paquet `faraway` conté les dades d'un treball de Ide-Smith i Lea (1988) que pretén estudiar la despesa dels adolescents (nois i noies) en jocs.
Les variables considerades són:

`sex`	0 = noi, 1 = noia.
`status`	una puntuació basada en el nivell socio-econòmic dels pares.
`income`	ingressos en lliures per setmana.
`verbal`	puntuació sobre el número de paraules ben definides sobre 12.
`gamble`	(variable dependent) despesa en jocs en lliures a l'any.

1. Feu la regressió lineal múltiple sobre la variable `gamble` de les altres variables. Quina és l'estimació de la variància de l'error? I el coeficient de determinació ajustat?
2. És significativa la regressió? Què significa això? Quina és exactament la hipòtesi?
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? Quin són els FIVs?
4. Què podem dir del punt 24? Com millora el model si eliminem aquest punt de les dades?
5. Feu la predicció en forma d'interval de confiança de la despesa d'un noi conegut nostre que té els següents valors $(0, 60, 10, 11)$.

6. Si considerem el model amb la mateixa resposta i la variable `sex` com un factor i la variable `income` com variable concomitant, hi ha interacció? Llavors, les rectes de regressió de nois i noies són paral·leles? Què significa això?

Problema 3

Amb la mateixa base de dades del problema anterior.

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 `income` en els models, el complet i els seleccionats.
2. Ajusteu un model amb els següents mètodes:
   (a) Mínims quadrats *OLS* sense el punt 24.
   (b) Mètode de Huber.
   (c) *Least trimmed squares.*

Problema 4

En estadística, el test de Ramsey (1969) o *Ramsey Regression Equation Specification Error Test* (RESET) és un test per contrastar la linealitat d'un model de regressió. La idea és contrastar si combinacions de potències del vector de prediccions ens ajuden a explicar la variable resposta. La intuïció darrera el test és que si aquestes combinacions no lineals de les variables predictives tenen cap poder d'explicació en la resposta, llavors el model lineal inicial no està ben definit. Així doncs, el contrast RESET(k) per a un valor de $k > 1$ és

$$\begin{aligned} H_0: &\quad Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \\ H_1: &\quad Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \gamma_2 \hat{y}^2 + \cdots + \gamma_k \hat{y}^k + \epsilon \end{aligned}$$

on $\hat{y}$ és el vector de prediccions del model lineal múltiple inicial.
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 RESET[1] per a $k = 2, 3$ i 4 (són tres contrastos, un per a cada valor de k). A quina conclusió arribem?

[1] Encara que la funció `resettest` del paquet `lmtest` fa el test RESET, feu els vostres propis càlculs.