GRAU INTERUNIVERSITARI D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA (UB- UPC)

CURS 2012-2013 Q1 – EXAMEN PARCIAL : MODEL LINEAL GENERALITZAT

(Data: 29 d'Octubre del 2012 a les 15:00h Aula 004-FME)

Nom de l´alumne:	
DNI:	
Professors:	Lídia Montero – Josep Anton Sànchez
Localització:	Edifici C5 D217
Normativa de l'examen:	ÉS PERMÉS DUR APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	1h 00 min
Sortida de notes:	Abans del 13 de Novembre al Web Docent de MLGz
Revisió de l'examen:	13 de Novembre a 16:00 h a Sala Professors FME– Campus Sud

Problema 1 (3 punts): Forma Canònica

Il·lustreu si la següent llei de probabilitat pot rescriure's en la forma canònica de la família exponencial d'un paràmetre, tot detallant acuradament les ***expressions resultants*** per:

- El paràmetre canònic θ .
- La funció cumulant $b(\theta)$.
- $a(\phi)\ i\ \phi$.
- $c(y,\phi)$

La llei geomètrica $f_Y(y) = \pi(1-\pi)^y \quad 0 < \pi < 1 \quad y \geq 0 \quad enter$

Expressió	$f_Y(y) = \mathbf{exp}\left(\mathbf{ln}(y\theta) - \left(-\mathbf{ln}(1-e^{\theta})\right)\right)$
θ	$\mathbf{ln}(1-\pi)$
$b(\theta)$	$-\mathbf{ln}(1-e^{\theta})$
$a(\phi)\ i\ \phi$	$\phi = 1 \qquad a(\phi) = 1$
$c(y,\phi)$	0

Problema 2 (2 punts): ANCOVA

El fitxer PIBsp conté dades socio-econòmiques de les 52 províncies espanyoles. Volem determinar relacions amb el PIB per càpita de cada província amb les seves caràcterístiques. La variable "Immigrants" conté el percentatge de població immigrant i la variable categòrica "Costa" indica si la província està a l'interior (0) a la costa (1).

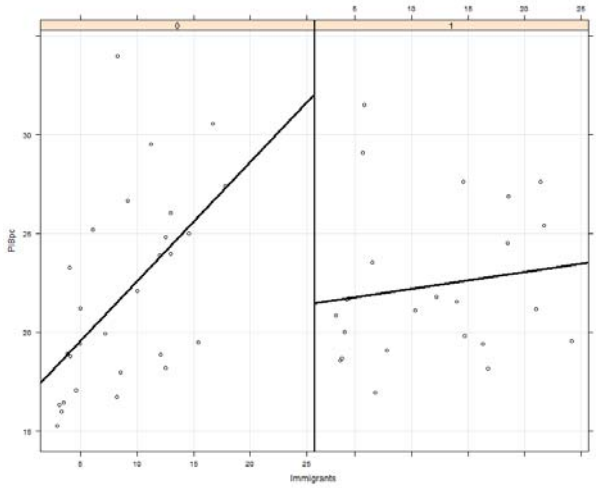

Fig. 1

A la figura 1. Hi ha el plot de la variable resposta amb la variable Immigrants, separat per si està o no a la costa. Si ajustem un model amb aquestes dues variables i la interacció, el resultat és el següent:

```
Call:
lm(formula = PIBpc ~ Immigrants * Costa, data = PIBsp)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3467 -2.8147 -0.4563  1.6495 12.4330

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      16.5626     1.6817   9.848 4.17e-13 ***
Immigrants        0.6026     0.1696   3.553 0.000866 ***
Costa             4.7977     2.3504   2.041 0.046750 *
Immigrants:Costa -0.5186     0.2086  -2.485 0.016479 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.033 on 48 degrees of freedom
Multiple R-squared: 0.2165,  Adjusted R-squared: 0.1676
F-statistic: 4.422 on 3 and 48 DF,  p-value: 0.007981
```

Expresseu els models lineals que relacionen el PIB per càpita amb el percentatge d'immigració, segons si la provincia està a la costa o no.

Si la província està a l'interior (Costa=0):

$$PIBpc = 16.5626 + 0.6026\, Immigrants + \varepsilon$$

Si la província està a la costa (Costa=1):

$$PIBpc = (16.5626 + 4.7977) + (0.6026 - 0.5186) Immigrants + \varepsilon$$
$$PIBpc = 21.3603 + 0.084\, Immigrants + \varepsilon$$

Interpreteu els p-valors que acompanyen als coeficients del model estimat. Podem afirmar que hi ha diferències segons la provincia estigui a la costa o a l'interior? Justifica la resposta.

El p-valor de l'intercept indica que el terme independent és significatiu. De forma ideal, si una província de interior tingues una taxa d'immigració nul·la, aquest valor equivaldria a l'estmació del PIB per càpita corresponent i seria diferent de zero.

El p-valor de la variable Immigrants indica que, en el cas de les províncies de l'interior, hi ha relació lineal significativa entre el percentatge d'immigrants i el PIB per càpita. A més, el signe positiu del coeficient indica una relació lineal directa (a més percentatge d'immigració, més PIB per càpita)

El p-valor de la variable Costa indica que hi ha una diferència de nivell entre les províncies situades a la Costa i a l'interior. En concret, pel fet de se una província costanera, s'incrementa el nivell del PIB per càpita en gairebé 5000 euros.

Finalment, el p-valor de la interacció fa referència a que la relació entre el PIBpc i el percentatge d'immigració és diferent de forma significativa segons s'estigui a l'interior o a la costa. La pendent que s'havia observat en les províncies de l'interior, gairebé s'anul·la quan considerem la interacció. Això implica que no podem considerar que les rectes de les relacions entre la resposta i la covariant per als dos grups de la variable Costa siguin paral·leles. D'aquesta taula no tenim un p-valor directe per establir si la pendent pel grup de províncies de la costa és significativa.

Problema 3 (2 punts): Inferència en el Model Lineal General

Decidim ajustar un model amb les variables Superfície, Natalitat (taxa per 1000 habitants) i Taxa d'atur. El model resultant és:

```
Call:
lm(formula = PIBpc ~ Sup + Natalitat + Atur, data = PIBsp)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6728 -1.5595 -0.0919  1.8373  6.8685

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.8793207  2.3834986  11.277 4.28e-15 ***
Sup         -0.0002732  0.0000864  -3.162  0.00271 **
Natalitat    0.7303308  0.2160507   3.380  0.00145 **
Atur        -0.5798599  0.0889793  -6.517 4.05e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.994 on 48 degrees of freedom
Multiple R-squared: 0.568,    Adjusted R-squared: 0.541
F-statistic: 21.04 on 3 and 48 DF,  p-value: 7.67e-09
```

Per determiner la significació de les covariables fem server el mètode anova estándar i el mètode Anova implementat en la llibreria car de R.

```
> anova(mod2)
Analysis of Variance Table

Response: PIBpc
          Df Sum Sq Mean Sq F value    Pr(>F)
Sup        1 185.06  185.06 20.6387 3.748e-05 ***
Natalitat  1   0.01    0.01  0.0009    0.9759
Atur       1 380.81  380.81 42.4686 4.053e-08 ***
Residuals 48 430.41    8.97
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> Anova(mod2)
Anova Table (Type II tests)

Response: PIBpc
          Sum Sq Df F value    Pr(>F)
Sup        89.68  1  10.001  0.002712 **
Natalitat 102.46  1  11.427  0.001446 **
Atur      380.81  1  42.469 4.053e-08 ***
Residuals 430.41 48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indica que test s'associa a cada p-valor especificant quins models es comparen. En particular i amb els resultats obtinguts, consideres que la variable Natalitat s'ha d'incloure en el model? Raona la resposta.

En la primera taula, els test corresponen a models incrementals, comparant el model que té les covariables anteriors amb el que s'obté afegint a més la nova variable. Aquests p-valors no coincideixen amb els del test de Wald de cada coeficient en el summary (només el de la darrera variable). Aquests testos depenen de l'ordre d'introducció de les variables.

En la segona taula, la taula de test tipus II compara el model complert (amb totes les variables) amb el model amb totes menys la que s'indica. Aquests tests equivalen als test de Wald de significació dels coeficients.

En la primera taula, en comparar el model amb la variable Superfície amb el model que té la superfície i la natalitat, el p-valor ens indica que no hi ha diferència entre aquests 2 models. No obstant, en la segona taula, si comparem el model que té superfície i atur amb el que s'inclou també la Natalitat, aquesta passa a ser

significativa (com indica el p-valor del test de Wald). Això vol dir que la variable explica part de la variabilitat de la resposta si s'ajusta prèviament per superfície i atur.

Problema 4 (3 punts): Validació

Amb el model anterior, obtenim els plots per fer l'anàlisi de residus que apareixen representats a la figura 2.

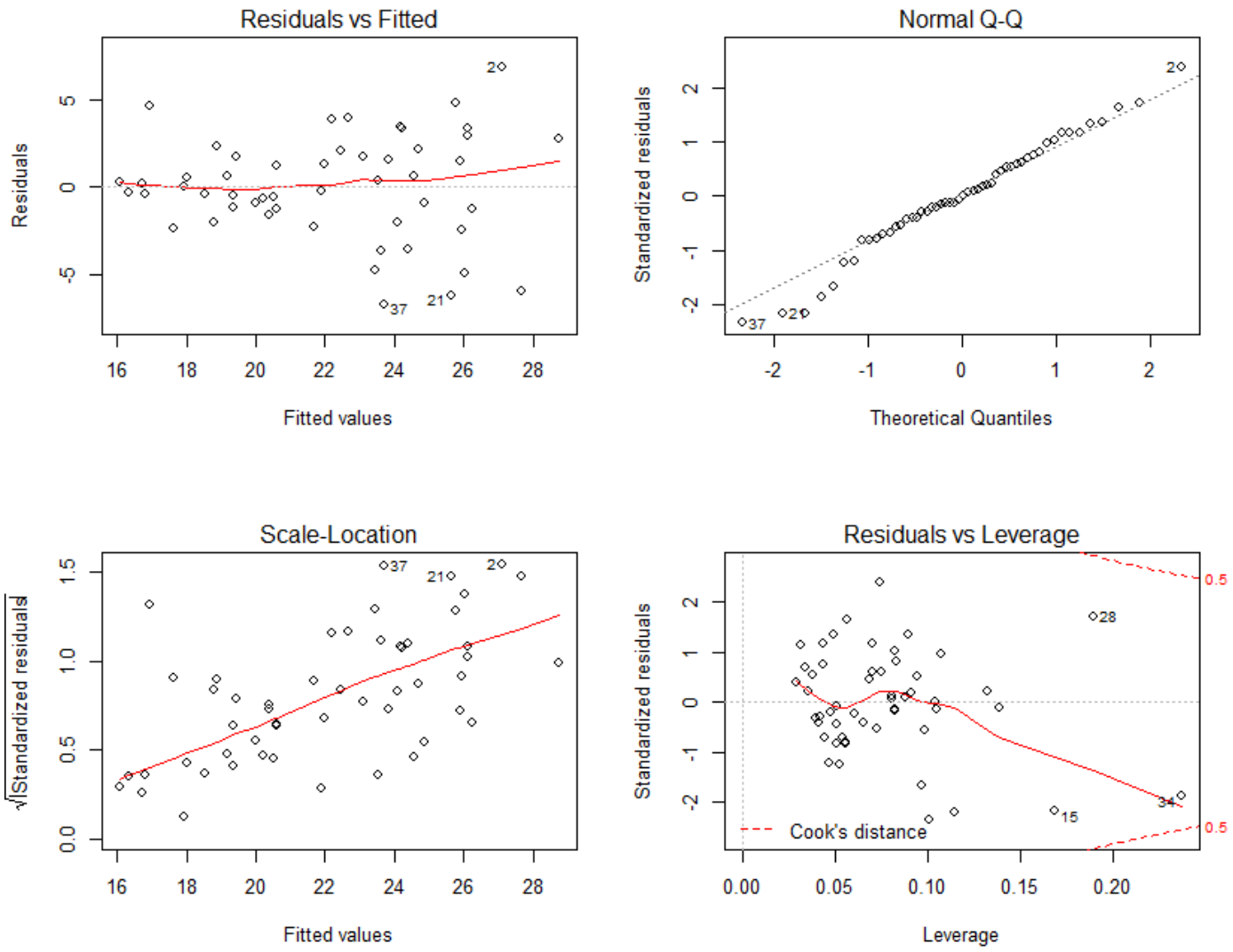

Fig 2. Plot del model

Feu la validació del model, indicant en cada gràfic, quina premissa s'analitza i a quina conclusió s'arriba en aquest cas. Considereu que el model és vàlid? En cas contrari, quin suggeriment faríeu per continuar la modelització?

Les premisses del model són: linealitat, homoscedasticitat, normalitat i independència.
El primer plot és el dels residus enfront les prediccions, permet veure si la disposició dels residus és aleatòria al voltant del zero, sense que s'observi cap patró que indiques desviacions de la relació lineal. L'ajust local (línea vermella) és pràcticament horitzontal, confirmant en aquest cas que no sembla haver patrons de no linealitat. En aquest plot també es pot verificar descriptivament si la variància es pot considerar constant, enfront de les prediccions. En aquest cas, sembla que s'observa un increment de la variabilitat dels residus a mesura que augmenta la predicció, indicant que pot haver heteroscedasticitat. També en aquest plot, apareixen etiquetades les observacions amb residus estandarditzats superior a 2 (aprox) en valor absolut (valors atípics)

El segon plot és el plot de normalitat, que permet determinar si podem considerar que la distribució Normal és adequada per als residus. Si els punts estan alineats podem assumir Normalitat dels residus. Aquest plot permetria veure patrons d'asimetria o cues pesades en els residus que anirien en contra de la hipòtesi de normalitat. També s'etiqueten els atípics.

El tercer plot fa l'arrel quadrada dels valors absoluts dels residus enfront de les prediccions. És un plot que permet determinar de forma més clara la presencia d'heteroscedasticitat. L'ajust local mitjançant la recta indica un clar increment dels valors que constitueixen una estimació de la variància dels residus. Aquest plot indica clarament que la variància augmenta, i per tant el model no será vàlid.

El quart model permet identificar i caracteritzar les dades influents. Representa els residus estandarditzats enfront del leverage. Amés inclou corbes de nivell per indicar la distancia de Cook de les observacions. Valors amb una distancia de Cook alta poden ser valors influents i s'ha d'analitzar el seu efecte en l'ajust del model. La distancia de Cook és una funció creixent dels residus al quadrat i del leverage. Les observacions que tenen un valor alt de la distancia de Cook apareixen etiquetades (poden ser per tenir molt leverage, o tenir un residus alt en valor absolut o una combinació d'ambdues situacions no tan extremes). Les observacions etiquetades com a influents sembla que tenen un leverage alt i a la vegada tenen un residu de magnitud elevada. Caldria analitzar quin efecte tenen en l'estimació del model.

En aquest cas, el model no seria vàlid perquè tot i que la linealitat i la normalitat sembla que es compleixen, hi ha evidències de variància no constant, així com a presencia de dades atípiques i/o influents. Una possibilitat seria considerar una transformació de la variable resposta (logaritme com a cas particular de transformació de Box-Cox que homogeneïtza la variància) o bé procedir a incloure nous predictors que puguin millorar la validació del model.

Indiqueu el motiu, com a resultat de l'ajust, pel qual apareixen etiquetats en els plots les següents observacions (no cal fer referència a les províncies concretes):

2 Àlava
15 Ceuta
21 Guadalajara
28 Las Palmas
34 Melilla
37 Ourense

2, 21 i 37 són valors atípics (residu estandarditzat gran). En concret, la observació 2 (Àlava) te un valor observat del PIBpc més gran del que prediu el model (residu positiu) i les altres dues (Guadalajara i Ourense) tenen un PIBpc per sota de l'esperat, d'acord a les variables incloses en el model

15, 21 i 28 són possibles dades influents en l'ajust del model. Són les 3 observacions amb un leverage més gran (les variables explicatives situen les observacions allunyades del centre de gravetat de la matriu de disseny). A més, tot i no aparèixer com a atípics, sembla que els residus són força grans. Són els valors situats més a prop de la corba de nivell que indica una distància de Cook de 0.5. L'efecte de la seva presencia sobre la estimació del model sembla que pot ser molt important.