Sessió de Laboratori 6: Resposta Binària

Eleccions Presidencials Nordamericanes del 92

L'arxiu Eleccions_92 conté els resultats d'un mostreig d'abast nacional de votants potencials. Concretament una mostra de 2198 electors a nivell nacional dels EEUU sobre l'actitut davant de les eleccions presidencials l'any 1992. Els camps disponibles són:

- pres: whether the respondent voted for President in the 1992 election
- age: repondent's age in years
- educ: respondent's years of education
- party: a measure of party preference with seven categories: 0 = strong democrat, 1 = weak democrat, 2 = independent democrat, 3 = independent independent, 4 = independent republican, 5 = weak republican, 6 = strong republican.
- inter: interest in the election: 1 = none, 2 = some, 3 = high.
- close: believes that the election will be close (1=yes, 0=no)
- sat: is satisfied with the candidates (1 = yes, 0 = no).

La variable de resposta és si l'individu va votar en les eleccions del 1992 o no (1=si, 0=no): pres.

1. Carregueu l'espai de treball Elecc_92.RData. Obriu els scripts subministrats a l'arxiu comprimit corresponent a la Sessió de Laboratori.
   a. Feu una ullada a les dades abans de desenvolupar l'exercici guiat per tal d'ajustar un bon model predictiu de la resposta, per millorar la llegibilitat podeu incorporar la definició de factors (pres, inter, sat, close etc) i veure per anàlisi exploratòria quines son les relacions esperables entre la resposta (pres) i la resta de variables.
   b. Crear una nova variable amb uns, indicant el pes de cada observació. Crear uns nous factors que continguin la discretització de l'edat (age) i els anys d'estudis (education) en els grups:

```
> table(elecc92$c.age)

(16,29] (29,39] (39,59] (59,91]
    414     573     664     547
> table(elecc92$edugroup)

(-1,11] (11,12] (12,15] (15,17]
    386     735     550     527
>
```

2. Estimeu el Model (M1): Predictor lineal amb covariable EDAT lineal.
   a. R: Feu l'estimació del model (M1) amb dades desagregades (li direm m1) i amb dades agregades (li direm m1a) . ***Cal crear un nou data frame que reculli la informació a nivell de dades agregades per les classes de la covariable definides pel model M1). Podeu usar la macro d'agregació dins dels scripts disponibles per la sessió d'avui: vigileu atentament la consistència de les explicatives (covariable).***
   b. Interpreteu els coeficients estimats en termes dels logodds, odds i probabilitats.
   c. Useu les eines standard de la llibreria car per l'anàlisi de residus glm() amb les dades desagregades: residualPlots(m1).

d. **Feu un diagrama bivariant múltiple solapant amb el format de dades agregades**: 1- logodds observats vs Edat més 2-logodds predits vs Edat. (Edat sempre en abscisses i Valors observats/predits en l'escala definida per la funció d'enllaç sempre en ordenades). **Pistes:**

Sigui el data.frame agregat. Calculeu:

i. Calculeu en una nova columna els logodds empírics/observats: log((ypos+0,5)/(m-ypos+0,5))= log((ypos+0,5)/(yneg+0,5))

ii. Calculeu en una nova columna els logodds predits/ajustats. R ja us dona directament el predictor lineal (predict(model)).

iii. Feu un diagrama bivariant múltiple solapat: 1- logodds observats vs Edat més 2- logodds predits vs Edat (Edat sempre en abscisses i Valors observats/predits en l'escala definida per la funció del link sempre en ordenades).

3. Ajusteu un model de regressió logística amb predictor lineal els termes d'ordre 1 i 2 de l'EDAT (Models M1-2, M1-3, variants de M1) (els ordres no lineals obteniu-los sempre respecte la tendència central de la variable o empreu la comanda poly(EDAT,k)).
    a. Contrasteu formalment la significació del terme d'ordre 3 de l'edat: segons l'estadístic de Wald i segons el contrast de la deviança.
    b. Contrasteu formalment la significació del terme d'ordre 2 de l'edat: segons l'estadístic de Wald i segons el contrast de la deviança.
    c. Useu sobre el conjunt de dades desagregat la comanda: marginalModelPlots() en m1 i model triat.
    d. Feu una diagnosi estàndard del model desagregat (m2): residualPlots(m2)

4. Considereu l'agrupació de l'EDAT en categories <30, 30-39, 40-59, 60+. Diem-li Model (M2). Tornarem a calcular el model m2 amb dades desagregades (i si en teniu ganes m2a amb dades agregades).
    a. Feu una diagnosi estàndard del model desagregat (m2c): residualPlots(m2c)
    b. Representeu el logits empírics/ajustats en funció de l'EDAT Categoritzada en les dades agregades (*diagnosi artesanal*). Pistes: ***Us ho podeu estalviar examinant atentament el sumari del model agregat m2a***.

5. **Quin tractament us sembla més adient per la variable EDAT: com a covariable (fins a quin terme d'ordre) o com a factor. Justifiqueu estadísticament la resposta.**

6. **Ara estudiarem la introducció de la variable EDUCACIÓ en el model que ja conté l'EDAT (en el seu millor tractament). Per començar es treballarà amb l'EDUCACIÓ com a covariant.**
    a. Afegiu un terme lineal d'EDUCACIÓ en el millor model anterior amb l'EDAT.
    b. Interpreteu el coeficient estimat en termes dels logodds, odds i probabilitats.
    c. Contrasteu la hipòtesi que l'efecte de EDUCACIÓ és lineal mitjançant la introducció d'un terme d'ordre 2 en el model. I el terme d'ordre 3, és significatiu? Pareu a l'ordre 3.
    d. Trieu el millor model que empra l'EDAT i la covariant EDUCACIÓ: li direm (M3). No cal treballar amb la versió agregada tret que vulgueu fer diagnosi artesanal.
    e. Useu les eines standard d'anàlisi de residus: marginalModelPlots() i residualPlots().
7. Considereu una agrupació dels anys d'educació (EDUCACIÓ Categoritzada) en 4 grups: <12, 12, 13-15, 16+. Estimeu el model de regressió logística amb els termes i tractament adient de l'edat i el factor EDUCACIÓ. Quina és la millor manera de tractar els anys d'educació? Li direm (M4).
        i. Assageu rectes amb pendents idèntics per cada categoría d'EDUCACIO

ii. Assageu rectes amb pendents diferents per cada categoría d'EDUCACIO
iii. Assageu paràboles amb idèntica corbatura per cada categoría d'EDUCACIO
iv. Assageu paràboles amb diferent corbatura per cada categoría d'EDUCACIO
v. Estrictament per inferència, quin us sembla el tractament més apropiat.

8. **Quina és la millor manera de tractar els anys d'educació, un cop l'EDAT ja ha estat incorporada en el model? Resultat Model (M5)**
9. Afegir al model les **preferències partidistes**. Analitzar els coeficients estimats i suggerir com es podria recodificar aquesta variable per simplificar la interpretació del model i estalviar un quants graus de llibertat (heu de veure que amb màxim 3 categories és suficient). Reajustar el model i reinterpretar els coeficients de la variable codificada.
10. Introduïu la variable 'grau d'interès' en les eleccions en el model. Contrasteu la significació del seu efecte principal emprant: test de Wald (si es paquet estadístic us ho permet) i el test de la deviança. Afirmaríeu que la principal raó per la que la gent jove vota menys és que no estan interessats en les eleccions?
11. Introduïu en el model un terme d'interacció entre el 'grau d'interès' i les 'preferències partidistes'. Contrasteu la significació de la interacció mitjançant el test de deviança. Expliqueu detalladament, si hi ho trobeu evidència, com funciona la interacció.
12. Hi ha alguna evidència que, després d'introduir el factor de '*proximitat entre els candidats'* en el model TREBALLAT FINS EL MOMENT (EDAT, EDUCACIO, PREFERÈNCIES PARTIDISTES, GRAU D'INTERÈS), les persones que creuen que les eleccions són ajustades tinguin una major incidència de vot?
13. Considereu la satisfacció amb les candidatures. La gent que no està satisfeta amb les candidatures té una menor incidència de vot? Canviaria la conclusió si el 'grau d'interès' en les eleccions no estigués inclòs en el model?
14. Feu la diagnosi del vostre model final.
15. Empreu el vostre model final per predir el comportament d'un votant. Classifiqueu com a votant probable aquells individus amb probabilitat superior o igual a 0.5. Feu una taula de contingència amb el vot probable i el vot real i analitzeu-la. Quina és l'explicabilitat del model final?
16. Calculeu el pseudo coeficient de determinació del model final i el coeficient de Naglekerke. Feu un goodness of fit test sobre el model final. Calculeu el goodness of fit emprant l'estadístic proposat per Hosmer-Lemershow.
17. Determineu la capacitat predictiva mitjançant l'anàlisi de la corba ROC.
18. Considereu una metodologia de modelització que s'iniciés amb un model complert i gran, on el coneixement del tractament de les variables EDAT i EDUCACIÓ es suposa conegut i apliqueu algun procediment tipus step() com a heurístic per la tria del millor model. Analitzeu els resultats.
19. Escriure un paràgraf de resum de les conclusions i el que heu après treballant amb el present arxiu.