Llista 3: Problemes (amb R) -part 1

Models logístics

L'equació logística (Malthus)

Modelització del creixement de poblacions:

$$\frac{dN}{dt} = \frac{rN(N-K)}{K} \tag{1}$$

on: N (de fet, N_t $N(t)$) és la població a l'instant t, K és la població màxima sostenible i r és el ràtio de creixement màxim quan la població és petita (lluny de K). Definint la proporció:

$$p := \frac{N}{K} \in (0,1) \qquad \text{(equivalentment } p(t)\text{)},$$

es té l'equació diferencial

$$\frac{dp}{dt} = rp(1-p) \tag{2}$$

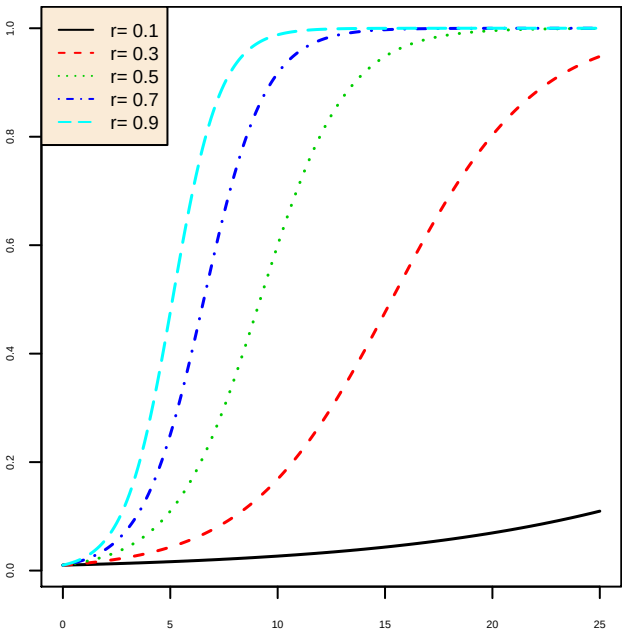
Les equacions (1) i (2) posen de manifest que en aquest model, el creixement de la població en l'instant t depén de la població existent en aquell moment, però també de com de lluny està del seu màxim sostenible.

La solució de l'equació (2) partint d'una proporció inicial p_0, ve donada per:

$$p = p(t) = \frac{1}{1+(\frac{1}{p_0}-1)e^{-rt}} = \frac{1}{1+e^{-\alpha-rt}} \tag{3}$$

on $\alpha = -\log(\frac{1}{p_0}-1)$. Aquesta solució, descriu una corba *sigmoide* anomenada *corba logística*. No es tracta d'un creixement exponencial sinó *sigmoide*, tal i com es pot veure fent una gràfica de la solució de l'equació.

Exercici 1. Amb R, fes la gràfica de la corba (3) per a valors diversos de $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, fixant $p_0 = 0.01$. Escull un rang de valors de t que permetin veure la forma *sigmoide*. *Indicació:* Per tal de posar totes les corbes al variar en una mateixa gràfica amb diversos colors i tipus de línia, es pot fer servir la funció `matplot()` i `legend()` per a la llegenda.



Model linealitzable logístic

La solució de l'equació de Malthus

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i + \epsilon_i)}} \tag{4}$$

pot ser apropiada per modelar proporcions (i anàlegs, com ara percentatges i recomptes amb petites transformacions), i la seva dependència d'una variable explicativa X. Aquest model ens assegura que les p_i estaran sempre entre 0 i 1, mentre que un model lineal del tipus $p_i = \beta_0 + \beta_1 x_i + \epsilon_i$ és no-acotat i donar valors fora de rang. Una manera equivalent d'expressar el model logístic és

$$\mathbf{p} = \frac{1}{1 + e^{-(\mathbf{X}\beta + \epsilon)}}$$

que tant sols és la forma matricial d'escriure totes les equacions del model variant i, amb $\mathbf{p}$ i ϵ vectors columna $n \times 1$, $\mathbf{X}$ matriu de disseny $n \times (k+1)$ (en regressió simple, $k = 1$) i β el vector columna de coeficients $(k+1) \times 1$.

Sabem que (4) és "linealitzable" mitjançant la transformació `logit` (o `log-odds`):

$$y_i := \mathbf{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i + \epsilon_i \tag{5}$$

amb la qual cosa, apartir de (5) es poden estimar els coeficients, i fer intervals i prediccions, etc. Hem de tenir en compte, però, que el model és lineal per les variables transformades (5), no per a les proporcions originals del model (4), els residus del qual és d'esperar que no tinguin un comportament centrat, homocedàstic i normal.

Exercici 2. Exercici de *simulació*: Com a valors de x, defineix una seqüència entre -5 i +5 amb increments de 0.01. Genera `p` un vector aleatori que satisfaci l'equació (4). Pots considerar els paràmetres $\beta_0 = 1, \beta_1 = 0.5$ i $\sigma^2 = 0.2^2$ (variància dels errors ϵ_i).[*Nota*: Com que hi ha generació de nombres aleatoris, si vols que els reusultats siguin exactament els mateixos cada cop que executes el codi, pots fixar una llavor.] Seguidament:

1. Crea un data.frame amb les columnes `x p`.
2. Fes la transformació linealitzadora obtenint el vector dels y_i i afegeix la columna `y` al data.frame anterior. *Nota:* per afegir una columna a un dataframe `dades` ja creat, pots usar `dades$nom<-columna`.
3. Estima els coeficients del model linealitzat (5) amb la funció `lm()`.
4. Calcula el vector de prediccions $\widehat{y}$ del model (5) - anomena 'l `haty` - i, fent servir la transformació inversa de *logit*, calcula el vector de prediccions $\widehat{p}$ del model (4) - anomena'l `hatp` -. Afegeix els dos vectors al data.frame. Mostra la capçalera del data.frame.
5. Calcula els residus del model (5) - anomena'ls `e` - i els redidus del moldel (4) - `ep` -. Afegeix-los al data.frame.
6. Fes una gràfica dels punts (x_i, y_i) amb la recta de regressió i al costat d'una altra amb els punts (x_i, p_i) amb la corba de regressió. Comenta-les.
7. Fes una gràfica de residus del model (5) respecte la resposta transformada $(\widehat{y}, e)$ i una per als residus del model (4) respecte de la resposta incial $(\widehat{p}, ep)$. Comenta-les.

Comentaris:

Apartat 6. Figura esquerra: La recta de regressió dibuixada sobre el núvol (x,y) mostra un cas clar de compliment de les hipòtesis del model que es poden visualitzar (com no pot ser d'altra manera atès que la simulació satisfà l'equació (5)): Hip1-centrament al voltant de la recta i Hip2-igual variància entorn de la recta. En aquesta gràfica la incorrelació i la Gaussianitat no són visibles.
Figura dreta: La recta de regressió dibuixada sobre el model logístic (x,p) mostra un comportament sigmoidal del núvol de punts (com no pot ser d'altra manera perquè així s'ha simulat). Com que aquest model respon a l'equació (4) la variància va disminuint quan x augmenta, perquè l'exponencial està al denominador.

Apartat 7. El mateix s'aprecia a la gràfica dels residus, on a la dreta es veu la no-igualtat de variàncies, que van disminuint amb x.

Llista 3: Problemes (amb R) -part 2

Probabilitats, *odds* i *odds-ratios*

Les *probabilitats* (i les proporcions) $p = p(A)$ prenen valors en (0,1) , exclosos els casos límit o no aleatoris. Una probabilitat superior a 1/2 indica que A és més probable que el seu complementari o contrari $\neg A$. Els *odds* són valors positius no acotats indiquen l'avantatge-desavantatge de A sobre $\neg A$, segons si el odds és més gran o menor que 1 i un log-odds positiu. Per exemple si $p = P(A) = 3/5$, aleshores $odds(p) = 3/2$, amb un avantatge de "3 a 2" per a l'esdeveniment A.

Si l'esdveniment A és patir certa malaltia i B és pertànyer a un grup de risc, interessa comparar l'odds de A en el grup de risc en relació a l'odds de A en el grup de control, això s'anomena *odds-ratio* OR o raó d'avantatges:

$$OR = \frac{odds(A|B)}{odss(A|\neg B)} = \frac{P(A|B)/(1-P(A|B))}{P(A|\neg B)/(1-P(A|\neg B))}$$

Per exemple, si la probabilitat de patir la malaltia al grup de control és $P(A|\neg B) = 0.001$ mentre que en el grup de risc és $P(A|B) = 0.02$, tindrem

$$OR = \frac{0.02/0.98}{0.001/0.999} = \frac{0.02040816}{0.001001001} = 20.38776$$

és a dir, un quocient d'avantages de 20.4 aproximadament en el grup de risc. L'odds-ratio (OR) està relacionat amb un altre concepte, el de risc-relatiu (RR) que no explicarem, però OR i RR no coincidixen.

Regressió logística

Quan es parla de **regressió logística** no fa referència "exactament" a la conversió de la corba logística en una recta aplicant la funció logit anteriorment explicada, sinó que es refereix a un cas particular dels anomenenats **models lineals generalitzats (GLM)**.

Els GLM són una extensió del model lineal a variables no Gaussianes, però no a tot tipus de variables, sinó restringides a les distribucións de l'anomenada *família exponencial*. La família exponencial conté, pel cas unidimensional, les lleis exponencial, Bernoulli, binomial, Gaussiana i Poisson, entre altres.

Els GLM es basen en el plantejament següent:

Si Y és una v.a. amb llei de la família exponencial i el GLM suposa que alguna funció g invertible de l'esperança d'aquesta variable s'expressa com a funció lineal de les variables explicatives:

$$g(E(Y)) = \mathbf{X}\beta \qquad equivalentment \qquad E(Y) = g^{-1}(\mathbf{X}\beta) \tag{6}$$

La funció g que linealitza l'esperança s'anomena funció d'enllaç ("link-function").

Regressió logística com a cas particular del GLM:

Diem que estem davant d'una regressió logística si tenim dades de Bernoulli (o més general, binomials). En el cas Bernoulli:

- Cada cas de la variable Y, cada y_i. és un 0 o un 1, amb llei $B(p)$ (equiv. $B(1,p)$.
- Recordeu que $E(Y) = p$.
- La funció de link és la $g =$*logit* (log-odds), ja que suposem: $g(E(Y)) = g(p) = \log\left(\frac{p}{1-p}\right) = \mathbf{X}\beta$.

Observacions:

1. Aquí $g(p)$ no és aleatòria !!!!!!!!!!!!

2. L'estimació dels paràmetres β es resól pel mètode de màxima versemblança.

3. Els GLM també permeten estructures de covariància més general que la de variables incorrelacionades.

4. L'apartat *model linealitzable logístic* s'aplica quan cada cas és una proporció o un percentatge, mentre que l'apartat Regressió logística com a GLM s'aplica quan cada cas és una dada binària (o un recompte d'èxits).

Exercici 3. Exercici de simulació: (estudiants de matemàtiques)

1. Com a valors x_i, defineix una seqüència entre -5 i +5 amb increments de 0.01.

2. Calcula els valors p_i com a funció exacta dels x_i: $p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$, amb $\beta_0 = 1, \beta_1 = 0.5$.

3. Seguidament, genera el vector de y_i, cada una d'elles amb llei Bernoulli de probabilitat p_i.

4. Estima els paràmetres β amb la funció `mle()` de la llibreria `stats4`. **Indicació:** A l'ajuda de la funció hi ha fet un exemple de la llei de Poisson que pots imitar.

5. Obté el vector de coeficients. *Nota:* si `ajust<-mle(....)` és l'objecte ajustat, `ajust@coef` dóna les estimacions, noteu que aquí s'usa `@` en lloc de `$`. Obté també l'interval de confiança.

6. Aplica ara la funció `glm()` per obtenir de manera més còmoda les estimacions. Probablement no seran idèntiques a les anteriors.