Raportul analizei statistice ce investigheaza relația dintre panta terenului și numărul de baraje observate de-a lungul unui șir de valori atribuite pantei terenului

1) Analiza de regresie

În ecologie se dorește de multe ori modelarea variabilelor răspuns de tipul variabilelor obtinute prin numărare (engl. count variables). Vatriabilele obtinute prin numărare sunt de obicei distribuite Poisson, ceea ce implică faptul că erorile modelelor nu sunt distribuite normal (Gaussian), făcând astfel imposibilă folosirea modelelor lineare (Kabacoff 2015, Quinn și Keough 2002, Wilkinson 2017).

Modelele lineare generalizate (engl. generalized linear models, GLMs) extind teoria de modelare lineară pentru a trata variabilele răspuns distribuite conform și altor distributii din familia exponentială în afară de cea Gaussiană (Kabacoff 2015, Wilkinson 2017). Aplicații des folosite ale modelelor lineare generalizate sunt regresia logisitică (cu variabilă răspuns binară), regresia Poisson (cu variabilă răspuns numărată) sau modelele log-lineare (cu variabilă răspuns categorică) (Quinn și Keough 2002). Pentru setul de date ce conține numărul de vizuini observate de-a lungul unui șir de valori atribuite pantei terenului analiza statistică a demarat cu o regresie Poisson implementată în R cu ajutorul funcției glm din pachetul stats (R Core Team 2017). Având în vedere faptul că variabila răspuns era supradispersată (engl. overdispersed), adică varianța mai mare decât media, s-a ajustat încă un model folosind familia quasipoisson ca și distribuție a variabilei răspuns. În modelul de regresie quasipoisson parametrul de dispersie este estimat din date, nefiind impus de distribuția Poisson (care presupune o varianță egală cu dispersia) (Quinn și Keough 2002). A doua soluție pentru problema supradispersiei observate în variabila răspuns este ajustarea unui model GLM ce folosește familia binomial negativă pentru descrierea erorii modelului (Crawley 2013, Quinn și Keough 2002). Pentru ajustarea modelului binomial negativ s-a folosit funcția glm.nb din pachetul MASS (Venables și Ripley 2002). Densitatea negativ binomială, comparativ cu densitatea Poisson, include un al doilea parametru (theta, parametrul de conglomerare) care este estimat din date, și folosește aceeași funcție de legătură (engl. Link functio) ca regresia Poisson (functia de legătură log).

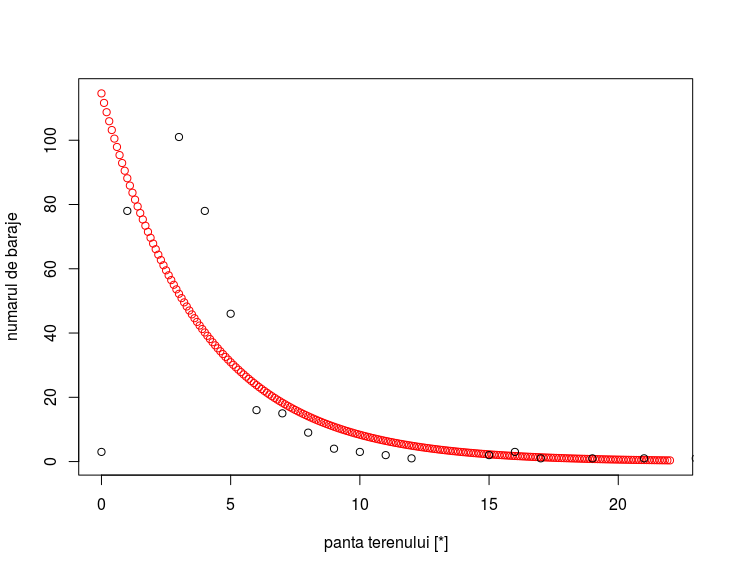
Coeficienții de regresie sunt la fel pentru modelul quasiposson și pentru cel poisson, așa cum arata tabelul 1, doar erorile standard ale coeficienților sunt ajustate pentru a lua în considerare supradispersarea (Kabacoff 2015). Modelul quasipoisson și cel negativ binomial se comportă în mare parte asemănător în ploturile de diagnoză. Totuși, observăm în ploturile ce prezintă valorile estimate și valorile observate o mai bună ajustare în cazul modelului negativ binomial dacă luăm în considerare informația a priori provenită de la echipa tehnică de prelevarea datelor (Pasca C) conform căreia vizuinele de castor apar cu precădere la o pantă a terenului între 0 și 1 grad și încetează să mai fie observate de la o panta de 5 grade în sus. Această informație va fi folosită în analiza Bayesiană prezentată la punctul 3.

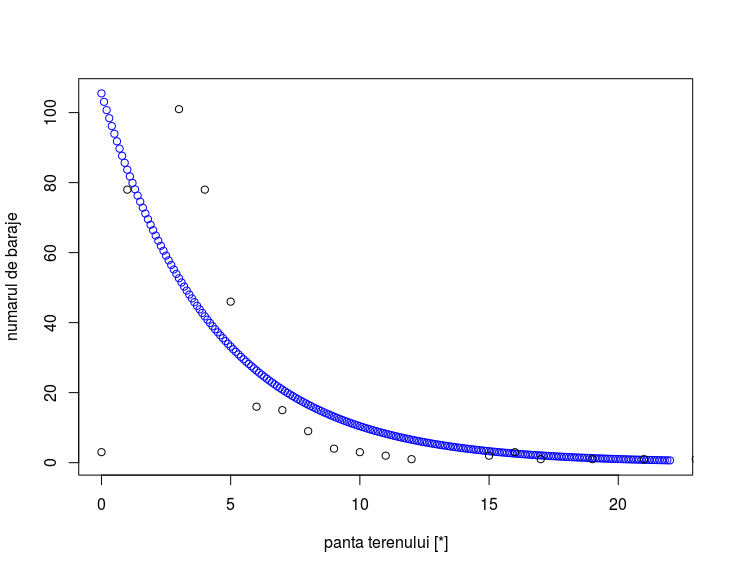
Interpretarea coeficienților de regresie pentru modelul negativ binomial ne arată că la o pantă a terenului de 0 grade se estimează în jur de 114 de baraje de castor (punctul de intersecție aducând în cazul acestui set de date o informație importantă). De asemenea, coeficientul pantei terenului ne indică faptul că pentru o creștere de o unitate în panta terenului avem o scădere în numărul în log al barajelor de castor de aprox. 0.262.

Interpretarea coeficienților de regresie pentru modelul quasipoisson ne arată că la o pantă a terenului de 0 grade se estimează în jur de 105.5 de baraje de castor (punctul de intersecție aducând și în cazul acesta o informație importantă). Coeficientul pantei terenului ne indică faptul că pentru o creștere de o unitate în panta terenului avem o scădere în numărul în log al barajelor de castor de aprox. 0.231.

Tabelul 1: Coeficienții modelelor de regresie, erorile standard ale coeficienților și valurile P pentru punctul de intersecție (cu axa y) și variabila explicativă (panta terenului)

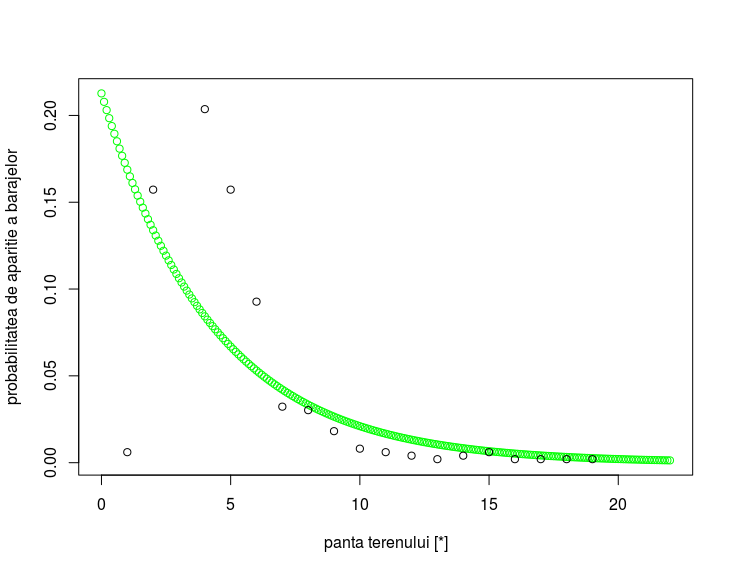
|  |  |  |  |
| --- | --- | --- | --- |
| Tipul GLM | Coef.  Intersectie  Panta terenului | Erori standard  Intersectie  Panta terenului | Valori P  Intersectie  Panta terenului |
| Poisson | 4.65870  -0.23143 | 0.06195  0.01234 | <0.001  <0.001 |
| Quasipoisson | 4.65870  -0.23143 | 0.24933  0.04965 | <0.001  <0.001 |
| Negativ binomială | 4.74121  -0.26201 | 0.35242  0.03771 | <0.001  <0.001 |

  
Fig. 1: Valorile estimate de modelul negativ binomial (rosu) si cele observate

  
Fig. 2: Valorile estimate de modelul poisson (albastru) si cele observate

2) Distribuția de probabilități ajustată și observată a apariției barajelor de castor de-a lungul unui șir de valori atribuite pantei terenului

În mod similar modelării numărului de baraje de castor de-a lungul unui șir de valori atribuite pantei terenului, un al doilea model negativ binomial a fost ajustat, de data aceasta folosind ca variabilă răspuns probabilitatea observată de apariție a barajelor de castor de-a lungul unui șir de valori atribuite pantei terenului. Coeficientii de regresie au fost: -1.5479 cu eroarea standard de 1.3797 și valoare P de 0.262 (pentru punctul de intersectie) și -0.2314 cu eroarea standard de 0.2748 și valoare P de 0.400 (pentru panta terenului). Distribuția de probabilități ajustată și distribuția de probabilități observată sunt prezentate în plotul următor.

  
Fig. 3: Distribuția de probabilități ajustată (verde) și observată a apariției barajelor de castor de-a lungul șirului de valori atribuite pantei terenului

3) Folosirea modelului Poisson-Gamma pentru estimarea de valori viitoare

a) Modelul Gamma-Poisson

Pentru a obține o funcție predictivă posterioră (engl. posterior predictive mass function) pentru estimarea de apariții viitoare a barajelor de castor de-a lungul unui șir de valori atribuite pantei terenului, a fost ajustat un model Poisson-Gamma. Modelul Poisson-Gamma, cunoscut și sub numele de analiză conjugată a variabilelor distribuite Poisson, are trei componente: o densitate conjugată a priori Gamma (reprezentând părerile a priori despre parametrul de intensitate lambda) (engl. prior density), o verosimilitudine Poisson (o variabilă observată distribuită Poisson) (engl. Poisson likelihood) și o densitate Gamma a posteriori (reajustarea părerilor despre lamda via Regula lui Bayes) (Jackman 2009). Conform datelor observate lambda MLE (Maximum Likelihood Estimation a parametrului lambda a distributiei Poisson) este de 26.105, verosimilitudinea (engl. likelihood) este: baraje\_castor ~ Poisson(26.105).

Din expertiza echipei tehnice (Pasca C), se presupune că în medie numărul de baraje de castor de-a lungul unui șir de valori atribuite pantei terenului este a/b=0.5, din care rezultă o densitate Gamma(a,b) a priori: p(lambda) ~ Gamma(9.5, 19). Cele două densităti (likelihood și densitatea a priori), determină următoarea densitate a posteriori: p(lambda | baraje\_castor) ~ Gamma(505.5, 38). Calculul exact al parametrilor densitătilor de probabilităti se găseste în scriptul R anexat raportului.

Densitatea a posteriori obtinută a fost folosită pentru a determina funcția predictivă posterioră (engl. posterior predictive mass function) pentru estimarea de apariții viitoare a barajelor de castor de-a lungul unui șir de valori atribuite pantei terenului. Această funcție are forma unei densităti binomial negative (Jackman 2009).

b) Funcția predictivă a posteriori pentru modelul Poisson-Gamma

Funcția predictivă a posteriori pentru o variabilă distribuită Poisson cu păreri a priori/ a posteriori despre lamda reprezentate de o densitate Gamma este o distributie binomial negativă caracterizată de theta (probabilitatea succesului) și un parametru a (numărul succeselor) (Jakman 2009).

Folosind regula lui Bayes, a fost obtinută următoarea funcție predictivă a posteriori:

p(baraje\_viitoare, baraje) ~ NB(505.5, 0.974359).

Calculul exact al acestei funcții poate fi urmărit în scriptul R atașat raportului.

Din cauza supradispersării (engl. overdispersion) din datele observate, este o discrepanță foarte mare intre funcția predictivă a posteriori și densitatea a priori. Pentru a produce rezultate aplicabile, trebuiesc adresate sursele de supradispersie. Supradispersia este un fenomen cu multe surse, însă, în cazul de față, erorile din date ar putea fi cauzate de modul în care au fost măsurate variabilele (o resolutie scazuta a modelului digital de elevatie) sau de variatia neexplicata cauzata de variabile neincluse în model (cu este vegetation cover). Cercetările viitoare ar trebui să genereze o funcția predictivă a posteriori pentru aparitia barajelor de castor de-a lungul unui șir de valori atribuite pantei terenului. Deci, este important ca în viitor să se reducă aceste doua surse de torsionarea datelor pentru a exploata modelul Poisson-Gamma în scopuri predictive.

Autor: Roxana Tesileanu, INCDS Romania, [roxana.te@web.de](mailto:roxana.te@web.de), 12.06.2017