**SUGGERIMENTI DALLA REGINA DELLA REGRESSIONE**

Stressori come parte parametrica, anno e spazio come parte non parametrica.

Parte non parametrica: prodotto tensoriale in mgcv di base sullo spazio (spline) e base nel tempo (spline) 🡪 altrimenti prova con altre basi.

Semplifica e alloca al centroide dello stato (stato come punto), ma nello specifico si dovrebbe considerare una nozione areale 🡪 dato riferito all’area dello spazio (non so se con mgcv riusciamo a fare dati areali).

Target: considerare come unbounded, usiamo il numero di colonie perse (valore assoluto e non percentuale).

Regressore: usa anche il numero di colonie (max) 🡪 indicatore di quanto grande è lo stato, facendo l’ipotesi che ci sia una proporzione diretta tra numero di colonie e dimensione dello stato.

Link lineare: va nella parte parametrica

Link non lineare: usa termini non parametrici 🡪 covariate bidimensionale, usa i GAM (f1(x) + f2(y)).

Fai scelta di covariate da usare nella parte parametrica basandosi su:

* Se ci pare di intravedere una relazione sufficientemente lineare con la risposta, considerando gli stressori un po' tutti insieme, non uno alla volta 🡪 perché altrimenti ci sarebbero confondenti
* Per il regressore parametrico vogliamo avere qualcosa di facile da interpretare

Più ci allontiamo dal link parametrico, meno facile è l’interpretazione 🡪 scelte modellistiche, se l’enfasi è nello stimare l’effetto degli stressore sulle colonie perse, gli stressori li considero nella parte parametrica.

Tutto quello che si riesce a mettere parametricamente, lo si mette, altrimente lo si mette nella parte non parametrica (esempio gli stati, a causa dell’ambiente climatico e legislazioni differenti, e anche il tempo per la differenza tra estate e inverno).

🡪 Esplora scatterplot e fit del modello nelle varie dimensioni, marginalmente, e capire se ha un senso oppure no.

Divisione per stato: considerare la distribuzione nello spazio come impatto non lineare 🡪 fit non parametrico nello spazio con t-spline (o solo spline????) (pacchetto mcgb).

Progetto stepwise:

fare vedere cosa sarebbe se buttassi dentro tutto in maniera lineare, poi faccio vedere il guadagno se inizio a considerare termini non parametrici.

Dipendenza nel tempo: link periodico, perché sulle stagioni nell’anno c’è stagionalità 🡪 base di Fourier.

Fregatene delle percentuali 🡪 usa i numeri assoluti, in termini di conteggio si potrebbe usare un modello di Poisson, non del tipo lineare come quello visto nella prima lezione della Sangalli ma del tipo generalizzato lineare (visto nella seconda lezione della Sangalli).

Ci sono modelli di regressione lineare (assumo che ‘errore abbia una distribuzione gaussiana, o meglio, non lo assumo ma il fit ai minimi quadrati è ideale se l’errore è gaussiano) o modelli di regressione generalizzati 🡪 con una funzione di link posso usare qualsiasi altro tipo di distribuzione, come Poisson, binomiale… quindi potrei anche considerare dei conteggi.

Non complicato con mgcb, anziché avere “family=gaussian”, abbiamo “family=poisson”.

Altra possibilità: utilizzare il numero di colonie per stato come peso in una regressione semiparametrica pesata 🡪 stati più grandi pesano di più.

Peso: numero di colonie, risposta: percentuale di colonie perse.

Però qua devo fare attenzione agli stati più numerosi (come colonie) che quelli meno numerosi, perchè un errore su uno stato più numeroso è più grave rispetto a quelli meno numerosi.