Projet méthodes bayésiennes et modèles hiérarchiques
option mathématiques appliquées

Seeds : Random effects logistic regression

Razan ALTUJJAR
Clément REIS
Diego ACUÑA
Zihao GUO

Encadrant :
Mathieu Ribatet
mathieu.ribatet@ec-nantes.fr

25 mai 2023

Table des matières

1 Introduction

Le projet vise à créer un modèle pour évaluer la capacité de germination de deux types de graines différents, *O. aegyptiaco 75* et *O. aegyptiaco 73*, pour deux variétés distinctes, *haricot* et *concombre*. L'étude se base sur 21 assiettes dans lesquelles ont été plantées ces graines.

Les notations utilisées sont les suivantes : n_i représente le nombre total de graines plantées dans l'assiette i, r_i est le nombre de graines ayant germé dans cette même assiette et p_i est la probabilité de germination d'une graine dans l'assiette i.

TABLE 1 – Jeu de données

Seed O. aegyptiaco 75						Seed O. aegyptiaco 73					
Bean			Cucumber			Bean			Cucumber		
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
10	39	0.26	5	6	0.83	8	16	0.50	3	12	0.25
23	62	0.37	53	74	0.72	10	30	0.33	22	41	0.54
23	81	0.28	55	72	0.76	8	28	0.29	15	30	0.50
26	51	0.51	32	51	0.63	23	45	0.51	32	51	0.63
17	39	0.44	46	79	0.58	0	4	0.00	3	7	0.43
			10	13	0.77						

Le modèle utilisé est le suivant : r_i suit une loi binomiale de paramètres p_i et n_i, avec une transformation logistique de p_i en fonction de quatre coefficients α_0, α_1, α_2 et α_{12}, qui dépendent respectivement de l'intercept, du type de graine, de la variété et de leur interaction. Le modèle tient compte également de l'effet aléatoire b_i, qui suit une loi normale centrée en 0 avec une variance égale à τ. Cette dernière est définie comme l'inverse de la variance résiduelle σ^2. C'est-à-dire :

- $r_i \sim \text{Binomial}(p_i, n_i)$
- $\text{logit}(p_i) = \alpha_o + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i$
- $b_i \sim \text{Normal}(0, \tau)$, où $\tau = \frac{1}{\sigma^2}$

Dans ce contexte, x_{1i} représente le type de graine (0 pour O. aegyptiaco 75 et 1 pour O. aegyptiaco 73) et x_{2i} représente la variété (0 pour haricot et 1 pour concombre) pour l'assiette i et α_0, α_1, α_2, α_{12}, τ sont donnés avec des lois a priori "non informatives" indépendantes.

2 Modélisation mathématique

2.1 Choix de la loi a priori

Soit r_i le nombre de grains germées sur la $i-$ème plaque parmi le total de n_i, une grain plantée peut avoir que deux états : soit gérmée soit non gérmée, il est donc logique de faire le choix d'une distribution binomiale pour r_i qui décrit le nombre de fois où la graine vont germer pour les n_i expériences effectuées.

2.2 Choix du modèle

Commençons par voir notre approche, notre but principal étant d'étudier la germination de deux types de graines, la modélisation de p_i qui est la probabilité de germination sur la $i-$ème plaque sera basée sur une régression logistique à effet aléatoire afin d'étudier l'effet du type de graine (variable explicative x_1) et également l'extrait de racine (variable explicative x_2 $0 \leftrightarrow$ bean, $1 \leftrightarrow$ cucumber).

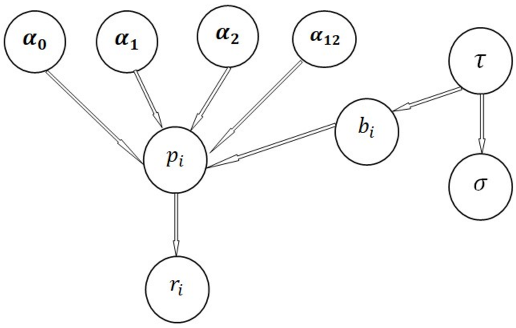
En effet, dans un modèle classique, on suppose que les erreurs sont indépendantes et identiquement distribuées, or quand les mesures sont répétées il est difficile de pouvoir faire chaque expérience avec exactement les mêmes conditions qu'une autre.

Un modèle à effet aléatoire va nous permettre de prendre cela en considération, pour généraliser l'approche par lequel on estime les paramètres d'un modèle de régression linéaire classique, pour qu'on puisse éstimer nos paramètres du modèle qui ne sont plus fixes, mais considérés comme des variables aléatoires.

Les coefficients de régression sont les suivants :

- α_1 et α_2 qui représentent l'impact respectif de x_1 et x_2 .
- α_{12} qui représente leur effect couplé.
- α_0 qui est l'intercept.
- b_i représente les résidus du modèle

2.3 DAG (Graphe orienté acyclique)



3 Lois conditionnelles

Rappelons les lois a priori des paramètres du modèle :

Paramètres	α_0	α_1	α_2	α_{12}	$\tau = \frac{1}{\sigma^2}$
Lois a priori	$\mathcal{N}(0, 10^6)$	$\mathcal{N}(0, 10^6)$	$\mathcal{N}(0, 10^6)$	$\mathcal{N}(0, 10^6)$	Gamma$(10^{-3}, 10^{-3})$

$$\pi\left(\alpha_0 \mid r, \alpha_1, \alpha_2, \alpha_{12}, b, \sigma^2\right) \propto \pi(\alpha_0) \prod_{i=1}^{N} \pi\left(r_i \mid \alpha_0, \alpha_1, \alpha_2, \alpha_{12}, b\right)$$

$$\pi\left(\alpha_0 \mid r, \alpha_1, \alpha_2, \alpha_{12}, b, \sigma^2\right) \propto \exp\left(-\frac{\alpha_0^2}{2 \times 10^6}\right) \prod_{i=1}^{N} (p_i)^{r_i} (1 - p_i)^{n_i - r_i}$$

α_1, α_2 et α_{12} elles ont les mêmes lois a priori, de la même façon, on retrouve leurs lois conditionnelles.

$$\pi\left(\alpha_1 \mid r, \alpha_0, \alpha_2, \alpha_{12}, b, \sigma^2\right) \propto \exp\left(-\frac{\alpha_1^2}{2 \times 10^6}\right) \prod_{i=1}^{N} (p_i)^{r_i} (1 - p_i)^{n_i - r_i}$$

$$\pi\left(\alpha_2 \mid r, \alpha_1, \alpha_0, \alpha_{12}, b, \sigma^2\right) \propto \exp\left(-\frac{\alpha_2^2}{2 \times 10^6}\right) \prod_{i=1}^{N} (p_i)^{r_i} (1 - p_i)^{n_i - r_i}$$

$$\pi\left(\alpha_{12} \mid r, \alpha_1, \alpha_2, \alpha_0, b, \sigma^2\right) \propto \exp\left(-\frac{\alpha_{12}^2}{2 \times 10^6}\right) \prod_{i=1}^{N} (p_i)^{r_i} (1 - p_i)^{n_i - r_i}$$

La loi conditionnelle de τ est une loi conjuguée, on trouve :

$$\pi\left(\tau_i \mid b_i\right) \propto \pi\left(\tau_i\right) \pi\left(b_i\right)$$

$$\pi\left(\tau_i \mid b_i\right) \propto \exp\left(-\tau.10^{-3}\right) \cdot \tau^{10^{-3}-1} \prod_{i=1}^{N} \exp\left(\frac{-b_i^2 \cdot \tau}{2}\right) \cdot \tau^{\frac{1}{2}}$$

$$\pi\left(\tau_i \mid b_i\right) \propto \exp\left\{-\tau\left(10^{-3} + \frac{\sum_{i=1}^{N} b_i^2}{2}\right)\right\} \cdot \tau^{10^{-3}-1+\frac{N}{2}}$$

$$\pi\left(\tau_i \mid b_i\right) \sim \text{Gamma}\left(10^{-3} + \frac{\sum_{i=1}^{N} b_i^2}{2}, 10^{-3} + \frac{N}{2}\right)$$

Pour $i = 1, ..., N$:

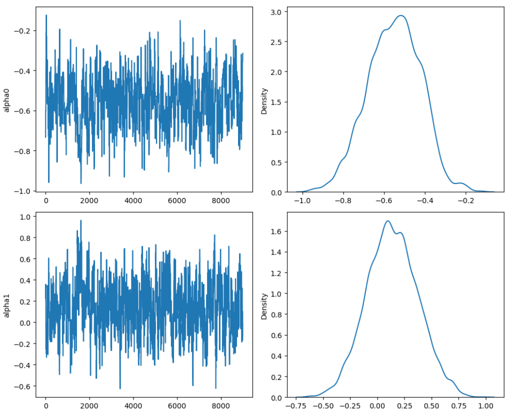
$$\pi\left(b_i \mid r, \alpha_0, \alpha_1, \alpha_2, \alpha_{12}, \sigma^2\right) \propto \pi\left(b_i \mid \sigma^2\right) \pi\left(r_i \mid b_i \mid \alpha_0, \alpha_1, \alpha_2, \alpha_{12}, \sigma^2\right)$$

$$\propto \exp\left(-\frac{b_i^2}{2\sigma^2}\right) (p_i)^{r_i} (1 - p_i)^{n_i - r_i}$$

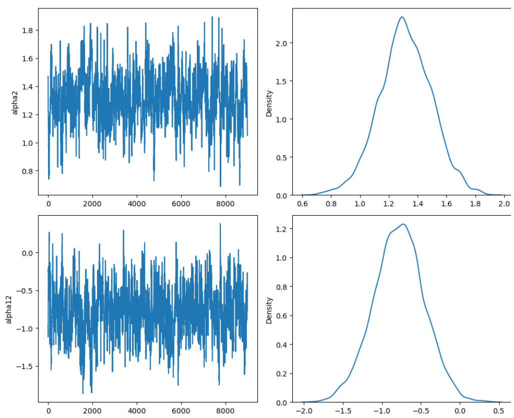
4 En Résumé

4.1 Les résultats obtenus

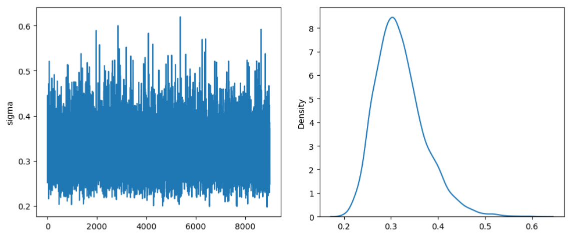
Nous générons une chaîne de taille 10000 et en retirons les 1000 premières valeurs. L'image ci-dessous montre la chaîne de Markov obtenue à gauche et la densité estimée à droite. Nous pouvons voir que les valeurs de chaque chaîne sont bien distribuées autour des valeurs attendues, nous pensons donc que les résultats sont satisfaisants.



(a) Alpha0 et Alpha1



(b) Alpha12 et Alpha2



(c) Sigma

FIGURE 1 – Représentation graphique des chaînes de Markov et des densités de variables aléatoires

TABLE 2 – Comparaison des résultats

	Moyenne		Écart-type		Médiane	
	Résultat	Référence	Résultat	Référence	Résultat	Référence
alpha0	-0.5531	-0.5525	0.129	0.1852	-0.5481	-0.5505
***alpha1**	0.1424	0.08382	0.2339	0.3031	0.1367	0.09076
alpha12	-0.7817	-0.8165	0.3243	0.4109	-0.7813	-0.8073
alpha2	1.3184	1.346	0.1825	0.2564	1.3145	1.34
sigma	0.3194	0.267	0.0523	0.1471	0.3124	0.2552

En comparaison, nous avons constaté que les moyennes pour Alpha1 variaient considérablement. Nous avons approfondi cette question en comparant les intervalles de confiance [Table. 3], dont nous avons constaté qu'ils variaient considérablement dans notre modèle ; nous en avons donc déduit que l'éventail des intervalles de confiance était responsable de la variation des moyennes, une conclusion qui peut également être vérifiée sur la base de l'approximation de la médiane.

Table 3 – Comparaison des intervalles de confiance

	Résultat	Référence
val2.5pc	-0.31	-0.52
val97.5pc	0.61	0.68

4.2 Analyse des résultats

Sur la base des résultats obtenus, on peut conclure ce qui suit. Les résultats sont affichés comme suit : <Seed 1> correspond aux haricots de la graine O. aegyptiaco 75 et <Seed 2> correspond aux haricots de la graine O. aegyptiaco 73, les semences 3 et 4 se rapportent au groupe des concombres.

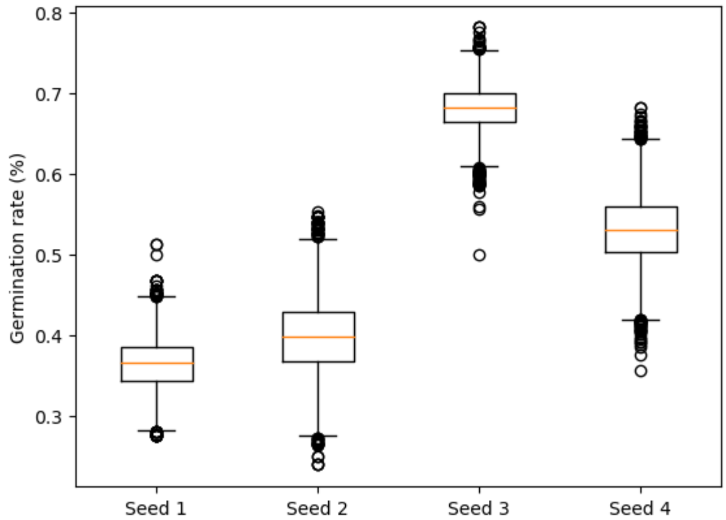


Figure 2 – Représentation graphique des taux de germination

— Le taux de germination des haricots est plus élevé pour les semences d'O. aegyptiaco 73 que pour les semences d'O. aegyptiaco 75, tandis que le taux de germination des concombres est plus élevé pour les semences d'O. aegyptiaco 75 que pour les semences d'O. aegyptiaco 73.

— Pour les graines de haricots, le taux de germination de la semence O. aegyptiaco 73 était de 39,96%, tandis que le taux de germination de la semence O. aegyptiaco 75 était de 36,57%. Cela indique que les graines de haricots de la semence O. aegyptiaco 73 sont de meilleure qualité et peuvent mieux croître et se développer.

— Pour les semences de concombre, le taux de germination des semences O. aegyptiaco 75 était de 68,19%, tandis que celui des semences O. aegyptiaco 73 était de 53,12%. Cela indique que les graines de concombre de la graine O. aegyptiaco 75 sont de meilleure qualité et peuvent mieux croître et se développer.

Par conséquent, si vous envisagez de cultiver des haricots, la semence O. aegyptiaco 73 est un meilleur choix. Si vous souhaitez cultiver des concombres, les semences d'O. aegyptiaco 75 sont un meilleur choix.

Notre code est ici : Cliquez ici!
En cas d'invalidation : https ://github.com/zihao-guo/Projet_seeds