Projet Bayes : Seeds - Random effects logistic regression

Zakariae MAAYZOU, Souhail LYAMANI, Anass EL MOUBARAKI, Elyas BENYAMINA
Encadré par: Mathieu RIBATET

March 2023

Lien vers le code : https://github.com/Bayes-group/Seeds_Bayes_project/tree/main

1 Introduction

Le domaine de l'agroalimentaire est étroitement lié aux caractéristiques environnementales. En effet, des facteurs tels que les conditions météorologiques, l'irrigation et surtout la qualité du terrain agricole ont une influence considérable sur le succès de la production, en particulier la germination des graines utilisées. Dans ce contexte, notre étude vise à examiner l'effet de deux paramètres: le type de graine et l'extrait de racine sur le rendement agricole. Nous allons examiner ces facteurs de près afin de comprendre leur impact sur la production.

2 Présentation des données

Nous disposons de données sur la proportion de graines qui ont germé sur chacune des 21 plaques disposées pour chaque type de graine (seed type) et d'extrait de racine (root extract). Les données sont présentées ci-dessous, où r_i représente le nombre de graines germées et n_i représente le nombre total de graines sur la i-ème plaque, pour i allant de 1 à $N = 21$. Ces données nous permettent de comprendre la relation entre les types de graines et les extraits de racine utilisés dans la germination.

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						

3 Objectif et approche intuitive

Notre objectif est de comprendre le comportement statistique du taux de germination pour chaque terre cultivable présente dans notre jeu de données. Pour y parvenir, nous avons opté pour une approche MCMC appliquée à un modèle hiérarchique. Nous cherchons à estimer les paramètres de notre modèle, qui sont considérés comme des lois a priori, afin de déterminer le nombre de graines qui ont germé. Pour ce faire, nous allons utiliser un algorithme d'échantillonnage de Gibbs qui cible la loi postérieure issue de nos paramètres a priori. Cette méthode nous permettra de mieux comprendre les interactions entre les différents paramètres et leur impact sur le taux de germination.

4 Modèle mathématique

Nous modélisons la probabilité de germination sur chaque plaque comme étant une distribution binomiale, où r_i représente le nombre de graines germées sur la plaque i et n_i représente le nombre total de graines sur cette plaque:

$$r_i \sim \text{Binomial}(p_i, n_i)$$

Nous supposons que la probabilité de germination p_i dépend du type de graine (x_{1i}) et de l'extrait de racine (x_{2i}) de chaque plaque, ainsi que de leur interaction ($x_{1i}x_{2i}$). Nous utilisons une fonction logistique pour modéliser cette relation :

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i$$

Nous incluons également un terme d'erreur b_i qui suit une distribution normale avec une moyenne de 0 et une variance de τ :

$$b_i \sim N(0, \tau)$$

Les lois a priori des paramètres du modèle sont les suivantes :

α_0, α_1, α_2, et α_{12} suivent des distributions normales centrées en 0, avec des variances σ_0^2, σ_1^2, σ_2^2, et σ_{12}^2, respectivement.

τ suit une distribution gamma avec des paramètres a et b.

En utilisant cette modélisation, nous pouvons estimer les paramètres et la probabilité de germination pour chaque plaque en utilisant un algorithme d'échantillonnage de Gibbs et la loi a posteriori obtenue à partir des lois a priori et des données observées.

Les lois a priori des paramètres du modèle sont :

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

5 Interprétation du modèle

Nous pouvons observer que la distribution binomiale est le choix le plus adapté pour décrire le nombre de succès dans les n_i expériences enregistrées, étant donné que r_i représente le nombre de graines germées et que la probabilité de germination est représentée par p_i.

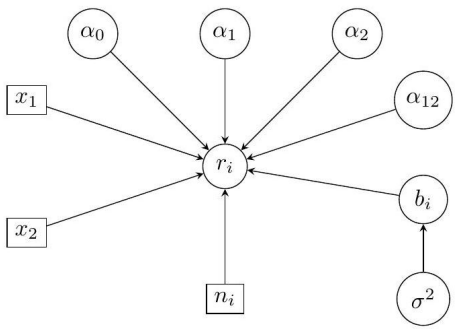
Quant à la modélisation de p_i, elle se base sur une régression logistique qui prend en compte les variables explicatives x_{1i} et x_{2i} et qui utilise un effet aléatoire. Dans ce contexte, α_0 représente l'intercept tandis que b_i représente les résidus du modèle. Les coefficients du modèle sont représentés par α_1 et α_2 qui indiquent l'impact respectif de x_{1i} et x_{2i}, ainsi que α_{12} qui décrit leur effet couplé. Il est important de noter que, contrairement à la régression logistique "classique", les paramètres du modèle sont considérés comme des variables aléatoires.

Le modèle "classique" repose sur l'indépendance des observations. Cette hypothèse est applicable dans la plupart des expériences bien planifiées, mais elle ne s'applique que rarement aux données de la vie quotidienne, qui présentent de nombreuses corrélations. Le modèle à effets aléatoires prend en compte cette dépendance.

La distribution a priori de τ est choisie comme étant une Gamma, car τ est toujours positif et que la loi Gamma est pratique pour des raisons de calcul. De plus, elle est la distribution conjuguée de la vraisemblance associée à τ.

6 Graphe orienté acyclique

Pour $i = 1, \ldots, N$:



7 Lois conditionnelles

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

$$\pi\left(\alpha_0 \mid r, \alpha_1, \alpha_2, \alpha_{12}, b, \sigma^2\right) \propto \exp\left(-\frac{\alpha_0^2}{2.10^6}\right) \prod_{i=1}^{N} (p_i)^{r_i} (1-p_i)^{n_i - r_i}$$

De la même façon, on retrouve les lois conditionnelles de α_1, α_2 et α_{12} vu qu'elles ont les mêmes lois a priori. La loi conditionnelle de τ est une loi conjuguée, on trouve directement:

$$\tau \mid r, \alpha_0, \alpha_1, \alpha_2, \alpha_{12}, b \sim \text{Gamma}\left(10^{-3} + \frac{N}{2}, 10^{-3} + \frac{1}{2}\sum_{i=1}^{N} b_i^2\right)$$

Pour $i = 1, \ldots, N$:

$$\pi\left(b_i \mid \cdots\right) \propto \pi\left(b_i \mid \sigma^2\right) \pi\left(r_i \mid \ldots\right)$$

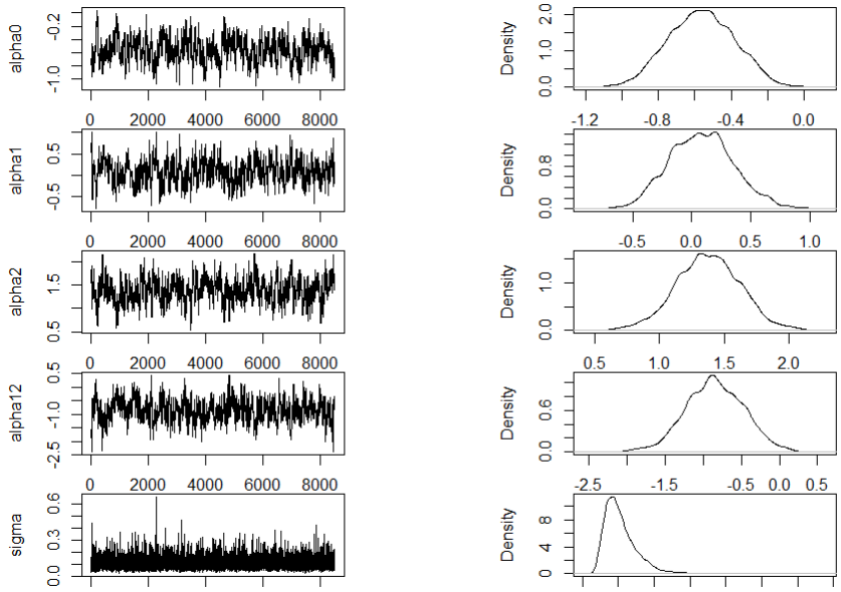
$$\pi\left(b_i \mid \cdots\right) \propto \exp\left(-\frac{b_i^2}{2\sigma^2}\right) (p_i)^{r_i} (1-p_i)^{n_i - r_i}$$

8 Résultats

Nous utilisons un échantillonneur Hastings-within-Gibbs pour générer une chaîne de Markov pour chaque variable aléatoire étudiée, avec un total de 104 réalisations et une période de "burn-in" de 1000 échantillons retirés. Nous commençons par comparer les moyennes et les écarts-types de nos résultats à ceux présentés dans l'énoncé, et nous obtenons des résultats assez similaires :

	Moyenne		Écart-type	
Paramètres	Résultat	Énoncé	Résultat	Énoncé
α_0	-0.5567	-0.5525	0.1825	0.1852
α_1	0.09100	0.08382	0.3036	0.3031
α_2	1.3578	1.346	0.2597	0.2564
α_{12}	-0.8467	-0.8165	0.4381	0.4109
σ	0.3211	0.267	0.1226	0.1471

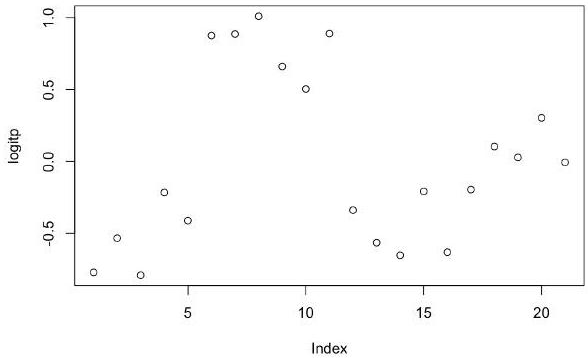
Les figures ci-dessous montrent les chaînes de Markov obtenues à gauche et leurs densités estimées à droite. Nous constatons des résultats satisfaisants, les valeurs de chaque chaîne étant bien distribuées autour des valeurs attendues. Nous avons également des taux d'acceptation adéquats, compris entre 30% et 40%, qui ne sont ni trop petits ni trop grands.



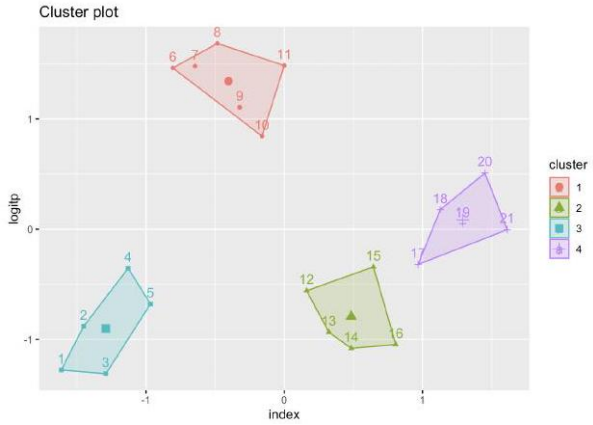
9 Interprétation

Après avoir tracé la fonction logit(p) en fonction de l'indice de l'échantillon, on observe que les échantillons semblent regroupés en plusieurs clusters. Pour mieux visualiser cette structure, nous allons utiliser l'algorithme de clustering k-means.

Dans notre cas, nous allons appliquer l'algorithme de k-means sur les échantillons de logit(p) en utilisant les moyennes empiriques de x_1 et x_2 obtenues précédemment comme variables explicatives. Le nombre de clusters sera déterminé de manière empirique en observant la structure des clusters obtenus pour différents choix de k.



(a) : logit(p) en fonction de l'indice



(b) : K-means clustering

Les résultats du clustering révèlent que les échantillons se regroupent en quatre clusters distincts en fonction des valeurs de x_1 et x_2.

- Le premier cluster correspond à l'association $x_1 = 0$ et $x_2 = 1$, c'est-à-dire pour les échantillons de type aegyptiao 75 et cucumber.
- Le deuxième cluster correspond à l'association $x_1 = 1$ et $x_2 = 0$, c'est-à-dire pour les échantillons de type aegyptiao 73 et bean.
- Le troisième cluster correspond à l'association $x_1 = 0$ et $x_2 = 0$, c'est-à-dire pour les échantillons de type aegyptiao 75 et bean.
- Enfin, le quatrième cluster correspond à l'association $x_1 = 1$ et $x_2 = 1$, c'est-à-dire pour les échantillons de type aegyptiao 73 et cucumber.

Ces résultats indiquent que la probabilité de germination varie en fonction du type de graine et des conditions environnementales. Les échantillons de type aegyptiao 75 ont une probabilité de germination plus élevée pour les graines de type cucumber, tandis que les échantillons de type aegyptiao 73 ont une probabilité de germination légèrement plus élevée pour les graines de type bean. Cependant, l'impact de ces facteurs sur la probabilité de germination n'est pas très drastique.