

S8_2_Modèles linéaires mixtes

Table of contents

1. Régression Linéaire Multiple (RLM)	1
2. Régression Logistique	4
3. Régression de Poisson (comptages / GLM Poisson)	7
4. Modèles Linéaires Mixtes (effets fixes + aléatoires)	10

1. Régression Linéaire Multiple (RLM)

Régression Linéaire Multiple

Intérêt (utilité du modèle)	<p>La RLM est utilisée pour modéliser la relation entre une variable dépendante quantitative et plusieurs variables indépendantes (ou explicatives) X_k.</p> <p>Elle sert à l'analyse multivariable (ajustement) pour estimer l'effet d'un facteur (X_1) en maintenant les autres facteurs (X_2, \dots) constants.</p>
Nature de la variable dépendante Y	<p>La variable Y est attendue comme quantitative continue.</p> <p>Exemple typique : la pression artérielle, ou un score quantitatif de développement cognitif (BAS).</p>
Structure du modèle (formule mathématique)	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$ <p>Définition des termes : Y est la variable dépendante (continue). X_1, X_2, \dots, X_k sont les variables indépendantes (quantitatives ou catégorielles).</p> <p>β_0 est l'ordonnée à l'origine (intercept).</p> <p>β_i sont les coefficients de régression. ε est le terme d'erreur (résidu théorique).</p>
Fonction de lien et justification	<p>Lien utilisé : Lien identité. Le modèle suppose une relation linéaire directe et additive entre le prédicteur linéaire et la moyenne de Y.</p> <p>Formule inverse : $\mathbb{E}(Y X) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$.</p>

Régression Linéaire Multiple

Conditions de validité / hypothèses du modèle

Les hypothèses principales portent sur le terme d'erreur ε :

1. **Moyenne nulle** : $\mathbb{E}(\varepsilon) = 0$.

2. **Normalité** : ε suit une loi normale.

3. **Variance constante (Homoscédasticité)** : $\text{Var}(\varepsilon) = \sigma^2$ (variance non dépendante de X).

4. **Indépendance des résidus** : Les ε_i sont indépendants.

Vérification :

On utilise l'histogramme et le QQ-plot des résidus pour la normalité, et le graphique des résidus en fonction des valeurs prédites (\hat{Y}_i) pour l'homoscédasticité (vérifier l'absence d'hétéroscédasticité).

Gestion des variables catégorielles et interactions

Les variables catégorielles à K modalités ($K > 2$) doivent être recodées en $K - 1$ variables binaires indicatrices (variables *dummy*) en choisissant une catégorie de référence.

Le coefficient β_i d'une variable indicatrice représente alors la différence moyenne de Y entre cette modalité et la catégorie de référence.

Interactions : L'ajout d'un terme multiplicatif, comme $\beta_3(X_1X_2)$, modélise un **effet multiplicatif** (synergie) entre les variables explicatives.

Attention, l'ajout d'interactions rend l'interprétation des coefficients isolés β_1 et β_2 difficile, à moins d'utiliser un codage spécifique (-1/1).

Interprétation des coefficients β

Si X_i est continue : β_i est la variation moyenne de Y associée à une augmentation d'une unité de X_i , toutes les autres variables X_j étant maintenues constantes.

Si X_i est binaire (0/1) : β_i est la **différence moyenne** de Y entre les sujets avec $X_i = 1$ et les sujets avec $X_i = 0$, toutes les autres variables X_j étant maintenues constantes.

Mesure d'association clé (issue de e^β ou des β)

La mesure pertinente est la **Différence de Moyenne** (β), qui exprime de combien la valeur de Y augmente ou diminue.

Régression Linéaire Multiple

	Valeur neutre : 0 (zéro) signifie l'absence de différence de moyenne ou d'association.
Tests statistiques utilisés	Les tests sont typiquement basés sur la statistique t (ou Z pour de grands échantillons) pour évaluer $H_0: \beta_i = 0$. L'intervalle de confiance (IC) sur β est utilisé.
Spécificités / extensions importantes	Limitation majeure : l'hypothèse d'indépendance des observations est souvent violée avec les données longitudinales ou groupées. Les extensions incluent les Modèles Linéaires Mixtes (MLM) pour gérer la non-indépendance et décomposer les variances.
Applications possibles : Exemple	Étude de l'effet de la durée de l'allaitement maternel (variable catégorielle) sur le score cognitif BAS (variable quantitative) chez l'enfant, en ajustant pour d'autres facteurs (régression linéaire multiple).

2. Régression Logistique

Régression logistique	
Intérêt (utilité du modèle)	<p>La régression logistique est un type de Modèle Linéaire Généralisé (GLM) utilisé pour modéliser une variable dépendante de nature binaire (dichotomique).</p> <p>Elle permet d'étudier l'association entre des variables explicatives et la probabilité p qu'un événement survienne.</p>
Nature de la variable dépendante Y	<p>La variable Y est binaire dichotomique (0/1). Elle suit une loi de Bernoulli.</p> <p>Exemple typique : Maladie oui/non (ex. HTA oui/non, rejet de greffe oui/non).</p>
Structure du modèle (formule mathématique)	<p>On modélise le logit de p comme une combinaison linéaire des variables explicatives :</p> $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$ <p>. </p> <p>Définition des termes : $p = P(Y = 1 X)$ est la probabilité que l'événement survienne. $\frac{p}{1-p}$ est la cote (Odds).</p> <p>$\log\left(\frac{p}{1-p}\right)$ est le logit de p.</p> <p>β_k sont les coefficients de régression.</p>
Fonction de lien et justification	<p>Lien utilisé : Logit (ou logistique).</p> <p>Formule du lien : $g(p) = \log\left(\frac{p}{1-p}\right)$.</p> <p>Justification : Le logit permet de transformer la probabilité p (qui est contrainte entre 0 et 1) en une variable définie entre $-\infty$ et $+\infty$, ce qui est adapté à un prédicteur linéaire.</p> <p>Fonction inverse (Expit) : $p = \frac{e^{\beta_0 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \dots + \beta_k X_k}}$.</p>
Conditions de validité / hypothèses du modèle	<p>La condition principale est que la taille de l'échantillon soit suffisante (au minimum 10 événements par variable explicative).</p> <p>Les coefficients sont estimés par la méthode du maximum de vraisemblance.</p>

Régression logistique

Gestion des variables catégorielles et interactions

Variables catégorielles : L'approche est la même qu'en RLM : on définit une classe de référence, et chaque coefficient β_k correspond au log-odds ratio entre la catégorie k et la référence.

Interactions : Elles sont gérées exactement de la même manière qu'en régression linéaire, en ajoutant un terme multiplicatif $\beta_3(X_1X_2)$. Le modèle multivarié suppose la **multiplicativité des effets** sur l'OR, sauf si une interaction est ajoutée.

Interprétation des coefficients β

Le coefficient β correspond au **logarithme de l'Odds Ratio** ($\ln(\text{OR})$). $\exp(\beta)$ est l'Odds Ratio (OR) ajusté sur les autres variables.

Si X_i est continue : $\exp(\beta_i)$ est l'OR associé à une augmentation d'une unité de X_i . **Si $\beta_i = 0$,** alors $\text{OR} = 1$, signifiant l'absence d'association.

Mesure d'association clé (issue de e^β ou des β)

La mesure clé est l'**Odds Ratio (OR)**.

Tests statistiques utilisés

Valeur neutre : 1 (OR = 1 signifie qu'il n'y a pas d'association).

1. **Test de Wald :** Permet de tester $H_0: \beta_k = 0$ (ou $\text{OR}_k = 1$) pour un coefficient individuel.

Il repose sur le fait que $\hat{\beta}/\text{SE}(\hat{\beta})$ suit approximativement une loi normale centrée réduite $\mathcal{N}(0,1)$ pour un grand échantillon.

2. **Test du Rapport de Vraisemblance (LRT) :** Compare la log-vraisemblance d'un modèle complet à celle d'un modèle réduit (sans la covariable testée).

La statistique $\Lambda = 2(\ell_{\text{complet}} - \ell_{\text{réduit}})$ suit une loi du χ^2 . Il est souvent plus robuste que le test de Wald, notamment pour les échantillons petits ou les événements rares.

Spécificités / extensions importantes

Le test de Wald peut être peu fiable en cas de petits échantillons, d'événements très rares ou d'effets extrêmes (séparation complète).

L'OR est symétrique et peut être estimé dans les enquêtes cas-témoins (où le RR ne l'est pas).

Extension : Les **Modèles Linéaires Généralisés Mixtes (GLMM)**, incluant la régression logistique mixte, permettent d'introduire des

Régression logistique

effets aléatoires pour les données corrélées (longitudinales ou groupées).

**Applications
possibles :
Exemple**

Modélisation du risque de rejet de greffe (Y binaire) en fonction de la dose d'un médicament (X binaire) et d'autres facteurs (multivariable).

Étude des facteurs associés à la démence (Y binaire).

3. Régression de Poisson (comptages / GLM Poisson)

Régression de Poisson (comptages / GLM Poisson)

Intérêt (utilité du modèle)	<p>La régression de Poisson est un GLM adapté à la modélisation des données de comptage ($Y = 0, 1, 2, \dots$).</p> <p>Elle est fréquemment utilisée en épidémiologie pour modéliser les taux d'incidence et obtenir des Risques Relatifs (RR) ou Rapports de Taux d'Incidence (IRR).</p>
Nature de la variable dépendante Y	<p>La variable Y est une variable aléatoire discrète prenant des valeurs dans les entiers naturels positifs ($0, 1, 2, \dots$).</p> <p>Elle suit une loi de Poisson, $Y \sim \text{Poisson}(\lambda)$, où $\lambda > 0$ est le paramètre qui représente la moyenne et la variance théorique.</p> <p>Exemple typique : nombre de visites à l'hôpital, nombre de crises d'asthme.</p>
Structure du modèle (formule mathématique)	<p>On modélise le logarithme du paramètre de moyenne $\lambda = \mathbb{E}(Y X)$ par un prédicteur linéaire :</p> $\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ <p>Définition des termes : λ est la moyenne (et la variance théorique) du comptage.</p> <p>X_k sont les variables explicatives.</p> <p>β_k sont les coefficients de régression.</p> <p>Si l'on modélise un taux d'incidence, on inclut souvent $\log(\text{Temps d'exposition})$ comme un <i>offset</i> dans le modèle.</p>
Fonction de lien et justification	<p>Lien utilisé : Logarithmique (Log).</p> <p>Formule du lien : $g(\mu) = \log(\mu)$ avec $\mu = \lambda$.</p> <p>Justification : L'exponentielle du prédicteur linéaire assure que $\lambda = \mathbb{E}(Y X)$ est toujours positif ($\lambda > 0$), ce qui est nécessaire pour une moyenne de comptage.</p> <p>Fonction inverse : $\lambda = e^{\beta_0 + \dots + \beta_p X_p}$.</p>
Conditions de validité / hypothèses du modèle	<p>L'hypothèse fondamentale de la distribution de Poisson est l'équidispersion :</p> <p>la moyenne est égale à la variance ($\mathbb{E}(Y X) = \text{Var}(Y X) = \lambda$).</p>

Régression de Poisson (comptages / GLM Poisson)

	<p>En cas de violation de cette hypothèse (surdispersion : $\text{Var}(Y) > \mathbb{E}(Y)$), les erreurs standards et les tests statistiques deviennent incorrects.</p>
Gestion des variables catégorielles et interactions	<p>Variables catégorielles : Chaque β_k représente le log du ratio des moyennes par rapport à la catégorie de référence.</p> <p>Interactions : Les effets sont additifs sur le $\log(\lambda)$, ce qui signifie qu'ils sont multiplicatifs sur la moyenne λ après exponentiation.</p> <p>Par exemple, si l'effet Fumeur et l'effet Diabète sont ajustés, leur effet combiné sur λ est la multiplication de leurs $\exp(\beta)$.</p>
Interprétation des coefficients β	<p>β_i représente le logarithme du Risque Relatif ($\ln(\text{RR})$) ou du Rapport de Moyennes/Taux. $\exp(\beta_i)$ est le Risque Relatif (RR) ajusté pour les autres variables.</p> <p>Si X_i est continu : $\exp(\beta_i)$ est le facteur multiplicatif (RR) sur la moyenne λ associé à une augmentation d'une unité de X_i.</p>
Mesure d'association clé (issue de e^β ou des β)	<p>La mesure clé est le Risque Relatif (RR) ou le Rapport de Taux d'Incidence (IRR).</p>
Tests statistiques utilisés	<p>Valeur neutre : 1 (RR = 1 signifie l'absence d'association).</p> <p>Test de Wald et Test du Rapport de Vraisemblance (LRT) sont utilisés pour l'inférence.</p> <p>L'estimation se fait par le maximum de vraisemblance.</p>
Spécificités / extensions importantes	<p>Attention : En cas de surdispersion, le test de Wald est biaisé (trop de faux positifs). Il faut alors utiliser la variance robuste (estimateur sandwich) ou des modèles alternatifs (Quasi-Poisson, Binomiale Négative).</p> <p>1. Surdispersion : $\text{Var}(Y) > \mathbb{E}(Y)$. Corrigée par l'estimateur de variance sandwich (corrige l'inférence sans changer $\hat{\beta}$) ou la régression Quasi-Poisson / Binomiale Négative.</p> <p>2. Excès de zéros : Lorsque le nombre de $Y = 0$ est supérieur à la prévision de Poisson. Traité par les modèles Hurdle ou Zero-Inflated (modèles à deux composantes).</p>
Applications possibles : Exemple	<p>Analyse du risque de saignement majeur (nombre d'événements) chez des patients exposés à des anticoagulants, en utilisant les</p>

Régression de Poisson (comptages / GLM Poisson)

trimestres-personnes comme exposition (pour calculer un taux d'incidence ajusté ou IRR).

4. Modèles Linéaires Mixtes (effets fixes + aléatoires)

Modèles Linéaires Mixtes (effets fixes + aléatoires)

Intérêt (utilité du modèle)	<p>Les Modèles Linéaires Mixtes (MLM) sont essentiels pour gérer la non-indépendance des observations.</p> <p>Ils sont utilisés dans les situations de données répétées (longitudinales) ou de données groupées (hiérarchiques), comme des patients suivis dans différents centres.</p> <p>Ils permettent d'estimer des effets fixes (covariables d'intérêt) tout en modélisant explicitement la corrélation au sein des groupes.</p>
Nature de la variable dépendante Y	<p>La variable Y est attendue comme quantitative continue.</p> <p>Exemple typique : Mesures biologiques répétées chez un patient, score cognitif mesuré à différents temps.</p>
Structure du modèle (formule mathématique)	<p>Un MLM combine des effets fixes et des effets aléatoires. Exemple avec intercept aléatoire de groupe (cluster i, observation j) :</p> $Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_i + \varepsilon_{ij}$ <p>Définition des termes :</p> <p>β_0, β_1 sont les effets fixes (effets moyens dans la population).</p> <p>X_{ij} est la covariable. u_i est l'effet aléatoire de groupe i (e.g., centre, patient).</p> <p>On suppose $u_i \sim \mathcal{N}(0, \tau^2)$. ε_{ij} est l'erreur résiduelle individuelle.</p> <p>On suppose $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. τ^2 est la variance inter-groupe, σ^2 est la variance intra-groupe (résiduelle).</p>
Fonction de lien et justification	<p>Lien utilisé : Lien identité (car Y est quantitative et suppose une distribution Gaussienne, comme en RLM).</p> <p>La relation est linéaire sur la moyenne conditionnelle aux effets aléatoires.</p>
Conditions de validité / hypothèses du modèle	<p>Les hypothèses principales sont la normalité des effets aléatoires u_i et la normalité des erreurs résiduelles ε_{ij}.</p> <p>On suppose aussi l'homoscédasticité des erreurs résiduelles.</p>

Modèles Linéaires Mixtes (effets fixes + aléatoires)

	<p>L'indépendance est modélisée : les Y_{ij} du même groupe sont corrélés via u_i, mais ε_{ij} sont indépendants.</p>
Gestion des variables catégorielles et interactions	<p>Effets fixes : Les variables catégorielles sont traitées comme en RLM (variable indicatrice par rapport à une catégorie de référence). Les interactions sont ajoutées comme des termes multiplicatifs.</p> <p>Effets aléatoires : Ils modélisent la structure de corrélation. On utilise la notation (1 groupe) pour un intercept aléatoire (niveau de Y qui varie entre les groupes) et (temps patient) pour une pente aléatoire (l'effet du temps varie entre les patients).</p> <p>Les effets peuvent être emboîtés (patient dans centre) ou croisés (patient vu par plusieurs médecins).</p>
Interprétation des coefficients β	<p>Les β sont les effets fixes : ils représentent l'effet moyen d'une covariable sur Y dans l'ensemble de la population.</p> <p>L'interprétation est celle d'une différence de moyenne (si X est binaire) ou d'une variation de la moyenne (si X est continu), ajustée pour la corrélation interne des données.</p>
Mesure d'association clé (issue de e^β ou des β)	<p>La mesure principale est la Différence de Moyenne (β) pour les effets fixes.</p> <p>Une autre mesure clé est le Coefficient de Corrélation Intraclasse (ICC), calculé à partir des composantes de variance.</p> <p>Formule de l'ICC : $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$.</p> <p>Valeur neutre : 0 pour β (pas d'effet fixe).</p>
Tests statistiques utilisés	<p>1. Test de Wald : Pour les coefficients des effets fixes (β_k). Utilise des approximations (e.g., méthode de Satterthwaite) pour calculer les degrés de liberté.</p> <p>2. Test du Rapport de Vraisemblance (LRT) : Pour comparer la pertinence de l'ajout d'un effet fixe ou d'un effet aléatoire. Nécessite l'estimation par Maximum de Vraisemblance (ML) pour comparer des modèles avec des effets fixes différents. L'estimation par défaut est souvent le REML (Maximum de Vraisemblance Restreint), qui donne des variances non biaisées.</p>

Modèles Linéaires Mixtes (effets fixes + aléatoires)

Spécificités / extensions importantes

1. **Décomposition de la variance** : Les MLM permettent de quantifier la variabilité inter-groupe (τ^2) et intra-groupe (σ^2), ce qui est impossible avec des modèles séparés.

2. Extensions :

- **Modèles Linéaires Généralisés Mixtes (GLMM)** pour des Y non-Gaussiens (e.g., logistique mixte pour Y binaire, Poisson mixte pour les comptages).
- **Modèles Non Linéaires Mixtes (NLMM)** pour les trajectoires de croissance complexes.

Applications possibles : Exemple

Analyse de la distance dentaire (mesure quantitative) mesurée à 8, 10, 12, 14 ans chez un même enfant.

Le modèle inclut l'âge (effet fixe) et un intercept aléatoire par enfant (pour tenir compte de la non-indépendance des mesures répétées).

Analogie de cloture

Imaginez que vous essayez d'estimer l'effet d'une nouvelle méthode d'enseignement sur la performance des élèves.

Le **Modèle Linéaire Multiple classique** ne verrait qu'une masse unique d'élèves, comme si tous étudiaient dans la même pièce et étaient totalement indépendants.

Si cette méthode marche mieux pour certains professeurs ou dans certaines écoles, le modèle classique serait trompé, car il mélangerait l'effet de la méthode avec la qualité de l'école (l'erreur de type I serait gonflée).

Le **Modèle Linéaire Mixte** agit comme un filtre sophistiqué : il considère les élèves comme étant **emboîtés** dans des classes, elles-mêmes emboîtées dans des écoles.

Les **effets fixes** mesurent l'effet moyen de la méthode d'enseignement pour *tous* les élèves (le résultat qui vous intéresse), tandis que les **effets aléatoires** mesurent à quel point chaque école ou chaque classe s'écarte de cette moyenne globale.

C'est comme déterminer la hauteur moyenne des vagues sur l'océan (effet fixe), tout en reconnaissant que la marée varie d'une côte à l'autre (effet aléatoire inter-côtes) et que les vagues locales sont différentes selon la plage (effet aléatoire intra-côte).