Pierre Pudlo

Modèle linéaire

Cours de Master 1ère année

Table des matières

Chapitre 1

Motivations et régression linéaire simple

1.1 Motivations

Les premiers outils que vous avez vu en **statistique inférentielle** permettent d'obtenir de l'information sur la valeur de paramètres liés à des **populations** à partir d'un **échantillon**. Nous allons voir dans ce cours comment :

- prédire la valeur d'une variable Y au niveau d'individus, dite « **variable à expliquer** », « output » ou « sortie », (ou maladroitement « variable dépendante »)
- en s'appuyant sur les valeurs d'autres variables $X_1,\dots,X_p$ observées sur le même individu, appelées « **covariables** », « variables explicatives », « régresseurs » (ou maladroitement « variables indépendantes »)

Pour cela, nous allons devoir comprendre la régularité du phénomène qui lie la variable explicative aux covariable au niveau de la population à partir d'un échantillon d'individus, puis descendre au niveau du nouvel individu pour faire la prédiction. Par nature, c'est un problème compliqué puisqu'il y a deux niveaux, l'individu et la population, et plusieurs variables sur chacun des individus.

Le **modèle linéaire** est la façon la plus simple de répondre à cet objectif compliqué. Nous verrons qu'il repose sur des **hypothèses très fortes** sur le lien entre Y et les covariables au niveau de la population. Mais ces hypothèses permettent de construire une théorie complète d'inférence et de prédiction qui est la pierre de base du « **machine learning** » ou apprentissage automatique en français.

Il y a de très nombreuses choses à comprendre, et **le contenu de ce cours sera dense**.

Certains exercices de TP seront évalués par compte-rendu. Ils seront réalisés avec R ou SAS.

Dans les outils du statisticien, vous aurez besoin :

- d'algèbre linéaire : calcul matriciel, projection dans des espaces euclidiens
- d'analyse : calcul de gradient de fonctions de plusieurs variables
- du calcul des probabilités (variables aléatoires, vecteurs aléatoires, espérance, variance, corrélation, matrice de variance-covariance,...)
- de la notion d'indépendance, de conditionnement et déconditionnement
- de la loi normale ou gaussienne univariée, et multivariée
- des notions d'estimateurs et d'erreur d'inférence, d'intervalles de confiance et un peu de test

1.2 Prédire en l'absence de covariable

Comment peut-on prédire la valeur d'une variable Y en l'absence de toute autre information sur l'individu ?

Etape 1 On collecte un échantillon $y_1,\dots,y_n$ issu de la population

Etape 2 On tire de l'information sur la population à partir de cet échantillon

Etape 3 On utilise cette information pour prédire la valeur de Y sur un nouvel individu, en faisant éventuellement attention à l'erreur possible à l'étape 2

On veut prédire la valeur de Y sur un nouvel individu à l'aide d'**une seule valeur**, caractéristique de la population, notée θ. Par exemple, on peut prendre la moyenne de la population à cause du résultat ci-dessous.

Théorème 1.1. *Soit Y une variable aléatoire réelle L^2. Soit d la fonction définie pour tout $\theta \in \mathbb{R}$ par*

$$d(\theta) = \mathbb{E}\Big((Y-\theta)^2\Big).$$

Cette fonction admet un minimum global en $\theta = \mathbb{E}(Y)$. Et $d\big(\mathbb{E}(Y)\big) = \mathrm{Var}(Y)$.

Soit $y_1,\dots,y_n$ une série de n nombre. Soit d la fonction définie pour tout $\theta \in \mathbb{R}$ par

$$d(\theta) = \sum_{i=1}^{n} (y_i-\theta)^2.$$

Cette fonction admet une unique minimum global en $\theta = \bar{y}$.

Interprétation : le meilleur résumé numérique θ d'une v.a. Y au sens de **l'erreur des moindres carrés** est son espérance $\mathbb{E}(Y)$.

Ainsi, à l'étape 2, on peut chercher à estimer $\theta = \mathbb{E}(Y)$, la moyenne de la population par

$$\hat{\theta} = \frac{Y_1+\cdots+Y_n}{n},$$

où $Y_1,\dots,Y_n$ sont des v.a. qui modélise l'échantillon, que l'on suppose indépendant, tiré suivant la même loi (=dans la même population).

Pour réaliser le programme de l'étape 3, on doit alors répondre par la valeur observée de $\hat{\theta}$ quel que soit le nouvel individu. Pour comprendre l'erreur commise, on introduit une nouvelle variable aléatoire Y, indépendante de l'échantillon, tirée dans la même loi que les Y_i.

Proposition 1.2. *Soit Y une v.a. L^2 d'espérance θ et de variance σ^2. On a*

$$\mathbb{E}\Big((Y-\theta)^2\Big) = \mathrm{Var}(Y) = \sigma^2.$$

Et

$$\mathbb{E}\Big((Y-\hat{\theta})^2\Big) = \mathrm{Var}(Y) + \frac{\mathrm{Var}(Y)}{n} = \sigma^2\left(1+\frac{1}{n}\right)$$

Le premier terme vient de la **variabilité entre individus** de la population. Le second terme vient de **l'erreur d'inférence** sur θ. Ce terme correctif $+1/n$ est dû à la **propagation de l'incertitude** sur la valeur de θ dans la prédiction de Y.

Enfin, on peut fournir une réponse plus détaillée à l'étape 3, en renvoyant non pas une seule valeur, $\hat{\theta}$, mais un intervalle de valeurs, avec une certaine probabilité ou couverture. Pour cela, il faut faire une hypothèse plus forte sur la distribution de Y au sein de la population : on suppose que $Y \sim \mathcal{N}(\theta, \sigma^2)$.

Proposition 1.3. *Sous ces hypothèses de lois gaussiennes, on peut estimer σ^2 par l'estimateur usuel $\hat{\sigma}^2 = (n-1)^{-1}\sum_i (Y_i - \hat{\theta})^2$ et obtenir l'intervalle de confiance :*

$$\forall \alpha \in]0;1[, \quad \mathbb{P}\left(\theta \in \left[\hat{\theta} \pm \Phi^{-1}_{n-1}(1-\alpha/2)\frac{\hat{\sigma}}{\sqrt{n}}\right]\right) = 1-\alpha$$

où Φ^{-1}_{n-1} est la fonction quantile de la loi de Student t_{n-1}.

NB : si n est grand, $t_{n-1} \approx \mathcal{N}(0;1)$.

Pour un nouvel individu Y, indépendant des observations, on a un résultat équivalent.

Proposition 1.4. *Sous les mêmes hypothèses, et avec les mêmes notations, on a*

$$\forall \alpha \in]0;1[, \quad \mathbb{P}\left(Y \in \left[\hat{\theta} \pm \Phi^{-1}_{n-1}(1-\alpha/2)\hat{\sigma}\sqrt{1+\frac{1}{n}}\right]\right) = 1-\alpha.$$

Dans les deux cas, on notera :

- que la loi de Student apparaît car on remplace une variance au niveau de la population σ^2 par son estimateur sans biais
- que les résultats de la Prop 1.2 se trouve directement transcrits en facteur du quantile de cette loi.

Le second intervalle, qui tient compte à la fois de la variabilité individuelle et de l'erreur d'inférence est un intervalle de prédiction. Quand $n \to \infty$, la largeur de l'intervalle de confiance sur θ tend vers 0 (l'information est maximale). Mais l'erreur individuelle sur une prédiction pour un nouvel individu subsiste.

Au final, on peut imaginer d'autres modèles, qui supprime l'hypothèse gaussienne, mais cela dépasse les objectifs de ce cours.

Le dernier modèle peut s'écrire de différente façon pour un individu Y de la population.

1. $Y \sim \mathcal{N}(\theta, \sigma^2)$
2. $Y = \theta + \varepsilon$, où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

Cette dernière variable, ε, jamais observée car θ est inconnu, s'appelle **l'erreur**.

1.3 Régression linéaire simple

On s'intéresse à une population d'individus sur lesquels on observe maintenant deux variables, X et Y. On fait l'hypothèse que, pour un individu,

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.1}$$

où β_0, β_1 sont deux paramètres définis au niveau de la population qui caractérisent le lien entre X et Y, et ε est indépendant de X, d'espérance nulle. C'est une équation du même style que la forme 2 du modèle sans covariable.

Proposition 1.5. *Sous ces hypothèses, on a*

$$\beta_0 = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X) \quad \textit{et} \quad \beta_1 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}.$$

Cette proposition montre que β_0 et β_1 sont bien des paramètres définis au niveau de la population.

L'indépendance entre la covariable X et l'erreur ε permet de dire que l'on a décomposé la variable Y en somme de deux termes :

- un terme $\beta_0 + \beta_1 X$ qui ne dépend d'un individu à l'autre que de la valeur de X,
- un terme ε qui est indépendant de X, mais varie aussi d'un individu à l'autre.

Vue la décomposition de Y en deux termes indépendants, on a

Proposition 1.6.

$$\mathrm{Var}(Y) = \beta_1^2 \mathrm{Var}(X) + \sigma^2.$$

On peut donc comparer la variabilité de l'erreur avec la variabilité totale par la fraction de variance de Y expliquée par X avec :

$$R^2_{\text{pop}} = \frac{\beta_1^2 \mathrm{Var}(X)}{\mathrm{Var}(Y)} = \mathrm{Cor}^2(X,Y)$$

où l'on a utilisé la proposition **??** pour la dernière égalité. Ce nombre, entre 0 et 1, vaut 1 si et seulement si $\varepsilon = 0$. Et il faut 0 si et seulement si Y est indépendante de X.

Dans le même esprit que le théorème 1 en l'absence de covariable, on a

Théorème 1.7. *Sous les hypothèses qui précèdent, soit d la fonction définie pour tout b_0, b_1 par*

$$d(b_0,b_1) = \mathbb{E}\Big(\big(Y-(b_0+b_1X)\big)^2\Big).$$

Elle admet un unique minimum global en $(b_0,b_1)=(\beta_0,\beta_1)$.

Dans la présentation ci-dessus, il manque un terme important dans la modélisation qui définie la façon dont ε varie d'un individu à l'autre : $\sigma^2 = \mathrm{Var}(\varepsilon)$. Ainsi, on a formulé un modèle à trois paramètres, $\theta = (\beta_0, \beta_1, \sigma^2)$.

On peut ajouter une hypothèse sur la distribution de l'erreur : $\varepsilon \sim \mathcal{N}(0,\sigma^2)$. Il s'agit alors du modèle de régression linéaire simple gaussien. La forme 1 de ce modèle est plus compliquée à écrire que dans le cas où il n'y a pas de covariable. Ici, elle s'écrit sous la forme de la loi conditionnelle ci-dessous.

Proposition 1.8. *Sous les hypothèses du modèle linéaire simple gaussien, on a*

$$\Big[Y\Big|X=x\Big] \sim \mathcal{N}(\beta_0+\beta_1 x, \sigma^2).$$

Ce qui s'interprète comme : dans la sous-population où la variable X est fixée à la valeur x, la variable Y suit une loi normale centrée en $\beta_0+\beta_1 x$ et de variance σ^2.

1.4 Inférence

Pour estimer θ, on a recours à un échantillon indépendant de n individus tirés dans la population, sur lequel on observe les valeurs des deux variables simultanément. Ainsi, notre jeu de données est composé des n paires (x_i, y_i), $i = 1,\dots,n$.

Pour modéliser ces données, on introduit donc n paires indépendantes de v.a. (X_i, Y_i) qui vérifient :

$$Y_i = \beta_0+\beta_1 X_i + \varepsilon_i, \quad \mathrm{Var}(\varepsilon_i)=\sigma^2, \quad X_i \perp\!\!\!\perp \varepsilon_i.$$

Il faut noter ici qu'il existe **deux types de situations** :
- des cas où on peut **choisir la valeur de** X au moment où on collecte les données,
- et des cas où **les valeurs de** X **sont subies** au moment où on tire un individu, et où on a donc un échantillon où les valeurs de X sont représentative de leur distribution dans la population.

Tous les résultats théoriques dans la suite sont établis conditionnellement aux valeurs observées de X dans l'échantillon, donc ces deux situations sont équivalentes.

En s'inspirant du théorème 1, on peut introduire l'estimateur des moindres carrés pour inférer β_0 et β_1. Autrement dit, on cherche $(\hat\beta_0, \hat\beta_1)$ qui minimisent la fonction

$$d(b_0,b_1) = \sum_{i=1}^n \Big(Y-\big(b_0+b_1X\big)\Big)^2.$$

On écrit

$$(\hat\beta_0,\hat\beta_1) = \mathrm{argmin}_{b_0,b_1} d(b_0,b_1).$$

Proposition 1.9. *Sous les hypothèses qui précèdent, on a*

$$\hat\beta_1 = \frac{\sum_i (X_i-\bar X)(Y_i-\bar Y)}{\sum_i (X_i-\bar X)^2} \quad et \quad \hat\beta_0 = \bar Y - \hat\beta_1 \bar X.$$

Les formules d'inférences sont explicites. De plus, il est d'usage d'estimer σ^2 avec

$$\hat\sigma^2 = \frac{1}{n-2}\sum_{i=1}^n \Big(Y_i-\hat\beta_0-\hat\beta_1 X_i\Big)^2$$

qui a de bonnes propriétés.

Voici, avec ces hypothèses, les résultats que l'on peut démontrer.

Théorème 1.10. *Sous les hypothèses qui précèdent, les estimateurs $\hat\beta_0, \hat\beta_1, \hat\sigma^2$) sont sans biais conditionnellement à $(X_1,\dots,X_n)$ (noté $X_{1:n}$). C'est-à-dire*

$$\mathbb{E}\Big(\hat\theta\Big|X_{1:n}=x_{1:n}\Big)=\theta$$

Et les $\hat\beta_i$ sont et de variance minimale conditionnellement à $(X_{1:n})$. Et, toujours conditionnement à $X_{1:n}$, $\hat\sigma^2$ est indépendant de $(\hat\beta_0,\hat\beta_1)$.

La seconde partie est difficile à démontrer et doit être admise.

En revanche, on peut montrer le résultat suivant.

Proposition 1.11. *Sous les hypothèses précédente, on a*

$$\mathrm{Var}(\hat\beta_1|X_{1:n}=x_{1:n}) = \sigma^2\frac{1}{\sum_i (x_i-\bar x)^2}$$

et

$$\mathrm{Var}(\hat\beta_0|X_{1:n}=x_{1:n}) = \sigma^2\left(\frac{1}{n}+\frac{\bar x^2}{\sum_i (x_i-\bar x)^2}\right).$$

De plus, si $x\in\mathbb{R}$,

$$\mathrm{Var}\Big(\hat\beta_0+\hat\beta_1 x\Big|X_{1:n}=x_{1:n}\Big) = \sigma^2\left(\frac{1}{n}+\frac{(x-\bar x)^2}{\sum_i (x_i-\bar x)^2}\right)$$

Ce dernier terme, $\hat\beta_0+\hat\beta_1 x$ est l'estimateur de $\beta_0+\beta_1 x$, l'espérance de Y sur la sous-population où $X=x$. C'est le terme que l'on utilise si on veut **prédire la valeur de** Y pour un nouvel individu, indépendant de l'échantillon, pour lequel on a observé que $X=x$.

Une fois $X=x$ connu, cette espérance conditionnelle $\beta_0+\beta_1 x$ est un paramètre lié à la sous-population où $X=x$. La dernière variance est donc l'erreur d'inférence sur ce paramètre. Comme dans le cas où il n'y a pas de covariable, on a maintenant

Proposition 1.12. *Si on considère un nouvel individu* (X, Y),

$$\mathrm{Var}\left(Y - \hat\beta_0 - \hat\beta_1 X \middle| X = x\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\bar x)^2}{\sum_i (x_i - \bar x)^2}\right).$$

Dans cette erreur de prédiction de Y (inconnue) par $\hat\beta_0 + \hat\beta_1 x$ (connu), on constate à nouveau que l'on tient compte :
- de la variabilité de ε d'un individu à l'autre avec $\sigma^2(1+$
- de l'erreur d'inférence sur les vraies valeurs β_0 et β_1, erreur que l'on propage ici au niveau de la prédiction.

Le cas gaussien. On peut maintenant ajouter l'hypothèse que $\varepsilon \sim \mathcal{N}(0;\sigma^2)$, c'est-à-dire

$$\left[Y \middle| X = x\right] \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2).$$

Alors, la loi de l'échantillon, conditionnellement aux valeurs observées des covariables est une loi gaussienne multivariée :

$$\left[Y_{1:n} \middle| X_{1:n} = x_{1:n}\right] \sim \mathcal{N}(\beta_0 \mathbf{1} + \beta_1 x_{1:n}, \sigma^2 I_n),$$

où $\mathbf{1}$ est le vecteur dont les coordonnées sont toutes égales à 1 et I_n est la matrice identité d'ordre n.

Dans cette situation, la fonction de vraisemblance conditionnelle est :

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \varphi\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right),$$

où φ est la densité de la loi normale centrée réduite. Un rapide calcul montre que, la log-vraisemblance vaut

$$\ell(\beta_0, \beta_1, \sigma^2) = \log L(\beta_0, \beta_1, \sigma^2) = \frac{1}{2\sigma^2}\sum_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2 - \frac{n}{2}\log(\sigma^2).$$

Proposition 1.13. *L'estimateur du maximum de vraisemblance est donné par*

$$\hat\beta_1 = \frac{\sum_i (X_i - \bar X)(Y_i - \bar Y)}{\sum_i (X_i - \bar X)^2} \quad \text{et} \quad \hat\beta_0 = \bar Y - \hat\beta_1 \bar X.$$

Autrement, ce sont les estimateurs des moindres carrées pour β_0 *et* β_1. *De plus, l'estimateur du maximum de vraisemblance pour le dernier paramètre est*

$$\tilde\sigma^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \hat\beta_0 - \hat\beta_1 X_i\right)^2.$$

Ce dernier estimateur est différent de celui introduit précédemment, et est biaisé. L'estimateur $\hat\sigma^2$ *est non biaisé (conditionnellement aux valeurs de* X*) et on le préfère.*

Enfin, comme dans le cas où il n'y a pas de covariable, on peut quantifier les incertitudes par des intervalles :
- un intervalle de confiance sur $\beta_0 + \beta_1 x$, qui est un paramètre défini au niveau de la sous-population où $X = x$,
- un intervalle de prédiction sur Y restreint à la sous-population où $X = x$.

Proposition 1.14. *On note* φ^{-1}_{n-2} *la fonction quantile de la loi de Student* t_{n-2}. *Sous les hypothèses gaussiennes, on a, pour tout* $\alpha \in]0;1[$,

$$\mathbb{P}\left(\beta_0 + \beta_1 X \in \left[\hat\beta_0 + \hat\beta_1 X \pm \varphi^{-1}_{n-2}(1-\frac{\alpha}{2})\hat\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar x)^2}{\sum_i (x_i - \bar x)^2}}\right] \middle| X_{1:n} = x_{1:n}, X = x\right) = 1 - \alpha.$$

Et, pour tout $\alpha \in]0;1[$,

$$\mathbb{P}\left(Y \in \left[\hat\beta_0 + \hat\beta_1 X \pm \varphi^{-1}_{n-2}(1-\frac{\alpha}{2})\hat\sigma\sqrt{1+\frac{1}{n} + \frac{(x-\bar x)^2}{\sum_i (x_i - \bar x)^2}}\right] \middle| X_{1:n} = x_{1:n}, X = x\right) = 1 - \alpha.$$

À nouveau ici, on voit que l'on a propagé l'incertitude (liée à la substitution des β_i par leur estimateur) à la prédiction de Y sachant $X = x$.

1.5 Quelques remarques supplémentaires

Erreur n°1 Ne pas confondre le terme d'erreur $\varepsilon_i = Y_i - \beta_0 - \beta_1 X$ qui n'est jamais observé, avec le terme

$$e_i = Y_i - \hat\beta_0 - \hat\beta_1 X_i$$

que l'on peut calculer (puisque l'on a substitué les paramètres inconnus par leurs estimateurs). Ces derniers termes s'appellent les **résidus**.

Ces résidus sont utiles, puisqu'ils reconstruisent les valeurs non observées de ε_i. Mais les résidus ont tendance à être plus petits que les ε_i car les valeurs substituées $\hat\beta_0$ et $\hat\beta_1$ dépendent aussi des valeurs (X_i, Y_i) de l'échantillon.
[Pour s'en convaincre, il suffit de voir que $\hat\sigma^2 = \sum e_i^2/(n-2)$ alors que l'on estimerait cette variance avec $\sum \varepsilon_i^2/(n-1)$ si ces dernières variables étaient observables.]

Erreur n°2 Le R^2_{pop} est un nombre intéressant au niveau de la population, puisqu'il mesure à quel point X permet de prédire Y. Il donne la part de variance de Y expliquée par la meilleure fonction affine de X.
Son estimateur est :

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}, \text{où}$$

$$\mathrm{SST} = \sum_{i=1}^n \left(Y_i - \bar Y\right)^2$$

$$\mathrm{SSE} = \sum_{i=1}^n e_i^2, \quad \text{et}$$

$$\mathrm{SSR} = \sum_{i=1}^n \left(\hat\beta_0 + \hat\beta_1 X_i - \bar Y\right)^2$$

Ici, SST = *Total Sum of Square*, SSR = *Sum of Square due to Regression* and SSE = *Sum of Squares Error* (nom maladroit car il s'agit de la somme des carrés des résidus !)
On a toujours $SST = SSR + SSE$ et ces trois nombres sont positifs. Sur l'estimateur, on a à nouveau que le R^2 est une évaluation du lien entre Y et sa prédiction linéaire.
La qualité de l'ajustement du modèle aux données ne s'évalue pas avec R^2. C'est de la qualité des prédictions de Y avec X dont parlent ce R^2 !

Erreur n°3 Dans le cas gaussien, il existe des tests statistiques pour les jeux d'hypothèses ci-dessous.
Jeu n°1 :

$$\mathscr{H}_0 : \beta_0 = 0 \quad \text{vs} \quad \mathscr{H}_1 : \beta_0 \neq 0.$$

Jeu n°2 :

$$\mathscr{H}_0 : \beta_1 = 0 \quad \text{vs} \quad \mathscr{H}_1 : \beta_1 \neq 0.$$

Leur mise en œuvre revient à vérifier que 0 est dans l'intervalle de confiance. Mais, dans $99{,}9\%$ des cas, **ils sont sans intérêt** dans le sens où ils répondent, de façon péremptoire, à une question qui est mal posée.
La question de savoir si la covariable X a un intérêt pour prédire Y ne se traite pas avec le test du jeu d'hypothèses n°2, mais avec une question de choix de modèles, entre le modèle de la partie 1 de ce chapitre, et le modèle de régression linéaire simple.

Erreur n°4 La régression linéaire gaussienne ne repose pas sur le fait que X et Y suivent des lois gaussiennes. L'hypothèse de loi gaussienne est mise sur la variable non observée ε, ou, de manière équivalente, sur la loi de Y sachant $X = x$, c'est-à-dire sur la distribution de Y dans la sous-population où $X = x$.
On peut s'assurer que cette hypothèse est réaliste en représentant la distribution empirique des résidus, avec, par exemple, un **histogramme des e_i**. Si cet histogramme a une forme de cloche, c'est presque gagné.

Erreur n°5 En première approche, on peut être tenté de ne pas s'intéresser au paramètre σ^2, qui représente la variance de Y au sein de la sous-population où $X = x$. Cette variance ne dépend pas de x. Sans le dire explicitement, on a fait une hypothèse forte ici. Dans la régression linéaire, $\mathrm{Var}(Y|X = x)$ ne dépend pas de la valeur de x (de la sous-population que l'on regarde). On parle de modèles **homoscédastiques**. On peut s'affranchir de cette hypothèse et construire des modèles, mais qui sont alors beaucoup plus compliqués et dépassent de très loin le cadre de ce cours.

Erreur n°6 La dernière hypothèse, dont nous n'avons pas discuté, est de l'indépendance de ε et $\beta_0 + \beta_1 X$. Pour vérifier qu'elle soit vraie, et vérifier l'hypothèse d'homoscédasticité, il est courant de représenter le nuage des points de coordonnées $(\hat{\beta}_0 + \hat{\beta}_1 x_i, e_i)$. On met donc sur l'axe des abscisses les valeurs prédites de Y pour les individus de l'échantillon, et sur l'axe des ordonnées les résidus. Ce points doivent être répartis aléatoirement, sans structure, dans un rectangle parallèle à l'axe des abscisses.
Une erreur commune est de mettre sur l'axe des abscisses les valeurs y_i au lieu de mettre les prédictions. Le graphique où les y_i sont sur l'axe des abscisses n'a aucun intérêt (ou presque).

Chapitre 2

Régression linéaire multiple

2.1 Le modèle

On considère maintenant que l'on veut prédire une variable Y à l'aide de plusieurs covariables numériques $X_1, \ldots, X_p$. Un **individu** de la population est donc représenté par le vecteur $(X_1, \ldots, X_p, Y)$. On suppose que

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

où $\varepsilon \perp\!\!\!\perp (X_1, \ldots, X_p)$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$. Les paramètres de ce modèle sont $\beta_0, \ldots, \beta_p, \sigma^2$.

Comme dans le modèle de régression linéaire simple, on a

Proposition 2.1.

$$\mathbb{E}\Big(Y \Big| X_{1:p} = x_{1:p}\Big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

En revanche, l'**interprétation** des paramètres est plus délicate (mais toujours faisable) :

- σ^2 : variabilité entre individus lorsque les covariables sont fixées,
- β_0 : prédiction ponctuelle de Y lorsque $x_1 = \cdots = x_p = 0$,
- $\beta_j, j \geq 1$: quantité à ajouter à la prédiction lorsqu'on augmente x_j d'une unité, **sachant que toutes les autres covariables restent fixées**.

Il est remarquable (et c'est une hypothèse du modèle) que cette dernière quantité ne dépend pas des valeurs des autres covariables. **Attention**, lorsque $p > 1$, le signe de β_j n'a pas de lien direct avec le signe de la corrélation entre X_j et Y. En fait, on a :

Proposition 2.2. *Pour $j = 1, \ldots, p$,*

$$\beta_j = \frac{\mathrm{Cov}\Big(X_j, Y \Big| X_{(-j)} = x_{(-j)}\Big)}{\mathrm{Var}\Big(X_j \Big| X_{(-j)} = x_{(-j)}\Big)},$$

où $X_{(-j)} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$.

Imaginons un exemple où Y est la masse graisseuse, et X_1 et X_2 représente la taille et le poids d'un être humain. Alors, X_1 et X_2 ont une corrélation positive importante, et on peut avoir :

$$\mathrm{Cov}(X_1, Y) > 0,\ \mathrm{Cov}(X_2, Y) > 0 \quad \text{et} \quad \beta_1 < 0,\ \beta_2 > 0.$$

L'interprétation des effets β_j est donc toujours liée aux autres covariables présentes dans le modèle. On écrit parfois l'expression « **toute chose étant égale par ailleurs** », l'augmentation de X_j d'une unité induit une augmentation de Y de β_j en moyenne. Il faut faire attention au fait que l'expression entre guillemets cache les covariables que l'on a mis dans le modèle avec X_j. Si on change les covariables (que l'on enlève ou en ajoute une), la valeur de β_j, de même que son interprétation va changer.

On a décomposé Y en somme de deux variables indépendantes $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ et ε. On a donc

$$\mathrm{Var}(Y) = \mathrm{Var}\Big(\beta_1 X_1 + \cdots + \beta_p X_p\Big) + \sigma^2.$$

De même, on peut définir un R^2 au niveau de la population par

$$R^2_{\text{pop}} = \frac{\mathrm{Var}\Big(\beta_1 X_1 + \cdots + \beta_p X_p\Big)}{\mathrm{Var}(Y)}$$

qui représente la fraction de variabilité de Y expliquée par le modèle, c'est-à-dire $\beta_1 X_1 + \cdots + \beta_p X_p$.

Enfin, on peut aussi ajouter l'hypothèse que l'erreur ε soit gaussienne. Dans ce cas, le modèle s'écrit

$$\Big[Y \Big| X_{1:p} = x_{1:p}\Big] \sim \mathcal{N}\Big(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2\Big).$$

2.2 Estimation

Données : n individus $\Big(x_{i,1}, x_{i,2}, \ldots, x_{i,p}, y_i\Big)$ $i = 1, \ldots, n$. Conventionnellement, les données sont rangées dans un tableau où les individus sont en lignes, et les variables en colonne. On modélise les données par $\Big(X_{i,1}, X_{i,2}, \ldots, X_{i,p}, Y_i\Big)$. On suppose que ces n vecteurs aléatoires sont indépendants.

Ces individus étant tirés dans la population d'intérêt, on a

$$\forall i = 1, \ldots, n, \quad Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i.$$

On peut écrire matriciellement ces n équations : $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ où

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix}.$$

2.2.1 Moindres carrés

L'estimateur des moindres carrés est défini par

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 = \operatorname{argmin}_{\boldsymbol{\beta}} \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right)' \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right).$$

Proposition 2.3. *La fonction* $\boldsymbol{\beta} \mapsto \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2$ *atteint son minimum en tout* $\widehat{\boldsymbol{\beta}}$ *qui vérifie*

$$\left(\mathbf{X}'\mathbf{X} \right) \widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}.$$

Si $\operatorname{rg}(\mathbf{X}) = p+1$*, alors* $\mathbf{X}'\mathbf{X}$ *est inversible, ce minimum est unique et on a*

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Y}.$$

Cet estimateur a les propriétés suivantes.

Théorème 2.4. *Sous les hypothèses qui précèdent, si* $\operatorname{rg}(\mathbf{X}) = p+1$ *p.s., on a*

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \middle| \mathbf{X} = \mathbf{x} \right) = \boldsymbol{\beta}.$$

De plus, la matrice de variance-covariance est

$$\operatorname{Var}\left(\widehat{\boldsymbol{\beta}} \middle| \mathbf{X} = \mathbf{x} \right) = \sigma^2 (\mathbf{x}'\mathbf{x})^{-1}.$$

En outre, conditionnellement à $\mathbf{X} = \mathbf{x}$*,* $\widehat{\boldsymbol{\beta}}$ *est le meilleur estimateur linéaire sans biais de* $\boldsymbol{\beta}$

Dans toute la suite de cette section, on supposera que $\mathbf{x}'\mathbf{x}$ est inversible.

S'il y a de **fortes corrélations entre covariables**, la matrice symétrique $\mathbf{x}'\mathbf{x}$ définie positive a des valeurs propres proches de 0 relativement à la somme des valeurs propres, $\operatorname{tr}(\mathbf{x}'\mathbf{x}) = n + \sum_{i=1}^{n} \sum_{j=1}^{p} x_{i,j}^2$. L'inverse $\left(\mathbf{x}'\mathbf{x} \right)^{-1}$ a donc des valeurs propres très grandes. Cela veut dire que la variabilité de l'estimateur d'un échantillon à l'autre, mesurée par la matrice de variance-covariance $\operatorname{Var}\left(\widehat{\boldsymbol{\beta}} \middle| \mathbf{X} = \mathbf{x} \right)$ devient importante. L'**estimateur est donc instable** (intervalles de confiance très larges,...).

On peut **comprendre ce phénomène de façon plus pragmatique** ainsi en supposant qu'il existe une combinaison linéaire des covariable presque nulle, c'est-à-dire négligeable devant σ : il existe $\lambda_0, \ldots, \lambda_p$ tel que

$$\left(\lambda_0 + \lambda_1 X_1 + \cdots + \lambda_p X_p \right)^2 \ll \sigma^2$$

sur les données observées.
Alors, pour tout $a \in [-1;1]$,

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \approx (\beta_0 + a\lambda_0) + (\beta_1 + a\lambda_1) X_1 + \cdots + (\beta_p + a\lambda_p) X_p$$

dans le sens où la différence est négligeable devant σ^2. Les estimateurs $\widehat{\beta}_j$ ne savent pas s'ils doivent estimer β_j ou $\beta_j + a\lambda_j$ pour une valeur de $a \neq 0$. Ils vont **sur-apprendre** une valeur de a sur les données observées qui ne se généralise pas à l'ensemble de la population.
Nous proposerons différentes méthodes pour résoudre ce problème :

- la régression ridge : c'est la méthode qui permet de stabiliser les estimateurs lorsque les covariables sont fortement corrélées,
- la régression lasso : c'est une méthode qui permet d'annuler des coordonnées de $\widehat{\boldsymbol{\beta}}$ pour stabiliser l'estimateur,
- la sélection de variables : c'est une méthode qui permet de réduire l'ensemble des covariables.

Quant à σ^2, il est estimé avec

$$\widehat{\sigma}^2 = \frac{1}{n - (p+1)} \left\| \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right\|_2^2.$$

Cet estimateur est sans biais dans le sens ci-dessous.

Proposition 2.5.

$$\mathbb{E}\left(\widehat{\sigma}^2 \middle| \mathbf{X} = \mathbf{x} \right) = \sigma^2.$$

Comme dans le cas de la régression linéaire simple, $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ est le vecteur des **résidus**.

En cas de **sur-apprentissage**, le vecteur des résidus $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ est de norme plus faible qu'il ne faudrait. Et l'estimateur $\widehat{\sigma}^2$ ci-dessus sous-estime la variance de l'erreur σ^2. Nous sommes alors face à une double catastrophe :

- l'estimateur $\widehat{\boldsymbol{\beta}}$ risque de pointer vers des valeurs qui ne sont **pas généralisables** à la population, et
- la variance de **l'erreur est sous-estimée**.

De plus, s'il y a de corrélation dans les colonnes de $\mathbf{x}$ au point que $\mathbf{x}'\mathbf{x}$ soit mal conditionnée, alors l'algorithme d'inversion de cette matrice est instable et, troisième catastrophe, le calcul numérique $(\mathbf{x}'\mathbf{x})^{-1}$ n'est pas correct.

2.2.2 Modèle gaussien

Dans le modèle gaussien, on peut aller plus. L'équation matricielle $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ peut se ré-écrire avantageusement sous la forme

$$\left[\mathbf{Y} \middle| \mathbf{X} = \mathbf{x} \right] \sim \mathcal{N}_n\left(\mathbf{x}\boldsymbol{\beta}, \sigma^2 I_n \right),$$

où l'on n'a fait apparaître une loi gaussienne multivariée de dimension n.

La vraisemblance conditionnelle est donc donnée par :

$$L(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 \right)$$

et donc la log-vraisemblance est

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log\left(\sigma^2 \right) - \frac{1}{2\sigma^2} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \text{constante}.$$

L'estimateur du maximum de vraisemblance est l'estimateur des moindres carrées pour $\boldsymbol{\beta}$, mais il est biaisé sur σ^2. On préfère donc utiliser celui introduit ci-dessus.

Théorème 2.6. *Sous les hypothèses gaussiennes, si* $\mathrm{rg}(\mathbf{X}) = p+1$,

$$\left[\widehat{\boldsymbol{\beta}}\middle|\mathbf{X}=\mathbf{x}\right] \sim \mathscr{N}_p\left(\boldsymbol{\beta};\ \sigma^2(\mathbf{x}'\mathbf{x})^{-1}\right).$$

De plus,

$$\left[\frac{n-(p+1)}{\sigma^2}\widehat{\sigma}^2\middle|\mathbf{X}=\mathbf{x}\right] \sim \chi^2(n-p-1).$$

Et, conditionnellement à $\mathbf{X}=\mathbf{x}$, *on a* $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma}^2$.

Ce théorème est une application directe du théorème de Cochran, dont voici une version simple.

Théorème 2.7 (Cochran)**.** *Soit* $\mathbf{Z} \sim \mathscr{N}_d(\boldsymbol{\mu}, \sigma^2\mathbf{I}_d)$ *un vecteur gaussien de dimension* d, *dont les coordonnées sont indépendantes et de même variance* σ^2. *Soit* $\boldsymbol{\Pi}$ *une matrice de projection orthogonale sur un sous-espace* F *de dimension* d_1 : $\boldsymbol{\Pi}' = \boldsymbol{\Pi}$, $\boldsymbol{\Pi}^2 = \mathbf{I}_d$, $\mathrm{tr}(\boldsymbol{\Pi}) = d_1$.
Dans ce cas, on rappelle que $(\mathbf{I}_d - \boldsymbol{\Pi})$ *est la projection orthogonale sur l'orthogonale de* F, *noté* $F^\perp$, *de dimension* $d_2 = d - d_1$.
Alors,

— *les vecteurs aléatoires* $\boldsymbol{\Pi}\mathbf{Z}$ *et* $(\mathbf{I}_d - \boldsymbol{\Pi})\mathbf{Z}$ *sont deux vecteurs gaussiens indépendants, de lois respectives*

$$\boldsymbol{\Pi}\mathbf{Z} \sim \mathscr{N}_d\left(\boldsymbol{\Pi}\boldsymbol{\mu};\ \sigma^2\boldsymbol{\Pi}\right),$$
$$(\mathbf{I}_d - \boldsymbol{\Pi})\mathbf{Z} \sim \mathscr{N}_d\left((\mathbf{I}_d - \boldsymbol{\Pi})\boldsymbol{\mu};\ \sigma^2(\mathbf{I}_d - \boldsymbol{\Pi})\right)$$

— *les variables aléatoires*

$$S_1 = \sigma^{-1}\boldsymbol{\Pi}\left(\mathbf{Z}-\boldsymbol{\mu}\right), \quad S_2 = \sigma^{-1}(\mathbf{I}_d - \boldsymbol{\Pi})\left(\mathbf{Z}-\boldsymbol{\mu}\right)$$

sont indépendantes, et de lois respectives $\chi^2(d_1)$ *et* $\chi^2(d_2)$.
Donc, la variable aléatoire

$$F = \frac{S_1/d_1}{S_2/d_2}$$

suit une loi de Fisher $\mathscr{F}(d_1, d_2)$.

Grâce à ce théorème, on peut obtenir des intervalles de confiance sur les coordonnées de $\widehat{\boldsymbol{\beta}}$. Comme celles-ci sont corrélées, il est aussi intéressant de considérer l'ellipsoïde de confiance. On pose

$$\mathbf{v} = \left(\mathbf{x}'\mathbf{x}\right)^{-1}$$

de telle sorte que, la variance de $\sigma^{-1}\widehat{\beta}_j$ est donnée par le j-ième coefficient de la diagonale de cette matrice v_{jj}.

Proposition 2.8. *Conditionnellement à* $\mathbf{X}=\mathbf{x}$, *l'intervalle centré en* $\widehat{\beta}_j$, *de bornes*

$$\widehat{\beta}_j \pm u\widehat{\sigma}\sqrt{v_{jj}}$$

où u *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Student* t_{n-p-1} *est un intervalle de confiance de niveau* $(1-\alpha)$ *pour* β_j. *Autrement dit,*

$$\mathbb{P}\left(\beta_j \in \left[\widehat{\beta}_j \pm u\widehat{\sigma}\sqrt{v_{jj}}\right]\middle|\mathbf{X}=\mathbf{x}\right) = 1-\alpha.$$

Conditionnellement à $\mathbf{X}=\mathbf{x}$, *l'ellipsoïde « centré en »* $\widehat{\boldsymbol{\beta}}$, *défini par*

$$\mathscr{E} = \left\{\mathbf{b} \in \mathbb{R}^{p+1} \ :\ \left\|\mathbf{x}(\mathbf{b}-\widehat{\boldsymbol{\beta}})\right\|_2^2 \le u(p+1)\widehat{\sigma}^2\right\}$$

où u *est le quantile d'ordre* $1-\alpha$ *de la loi de Fisher* $\mathscr{F}(p+1, n-p-1)$ *est une région de confiance de niveau* $(1-\alpha)$ *pour* $\boldsymbol{\beta}$. *Autrement dit,*

$$\mathbb{P}\left(\boldsymbol{\beta} \in \mathscr{E}\middle|\mathbf{X}=\mathbf{x}\right) = 1-\alpha.$$

Il faut noter ici que :

— le « centre » de l'ellipsoïde $\widehat{\boldsymbol{\beta}}$ est aléatoire,
— la forme de l'ellipsoïde est donnée par la matrice $\mathbf{x}$, donc les corrélations entre les covariables,
— le « rayon au carré » $u(p+1)\widehat{\sigma}^2$ est lié à l'estimation de l'erreur, et est donc aléatoire.

De façon plus générale, on peut s'intéresser à la région de confiance du vecteur $\boldsymbol{\theta} = \mathbf{T}\boldsymbol{\beta}$, où $\mathbf{T}$ est une matrice $q \times (p+1)$, $q \le (p+1)$, de rang q. Ici, $\boldsymbol{\theta}$ est donc un vecteur de dimension q.

Proposition 2.9. *Sous les hypothèses qui précèdent, l'ellipsoïde « centré en »* $\mathbf{T}\widehat{\boldsymbol{\beta}}$, *défini par*

$$\mathscr{E}_{\mathbf{T}} = \left\{\mathbf{b} \in \mathbb{R}^{p+1} \ :\ (\mathbf{T}\mathbf{b} - \mathbf{T}\widehat{\boldsymbol{\beta}})'\left(\mathbf{T}(\mathbf{x}\mathbf{x}')^{-1}\mathbf{T}'\right)^{-1}(\mathbf{T}\mathbf{b} - \mathbf{T}\widehat{\boldsymbol{\beta}}) \le u\,q\widehat{\sigma}^2\right\}$$

où u *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Ficher* $\mathscr{F}(q, n-p-1)$ *est une région de confiance au niveau* $(1-\alpha)$ *pour* $\mathbf{T}\boldsymbol{\beta}$.

De ces intervalles et régions de confiance, on peut en déduire des tests statistiques (qui sont très souvent inutiles). Par exemple, pour $j \in \{0, 1, \dots, p\}$ fixé, la décision au niveau $(1-\alpha)$ du test :

$$H_0 : \beta_j = 0, \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

s'obtient en choisissant :

— de conserver H_0 si 0 est dans l'intervalle de confiance de niveau $1-\alpha$,
— de rejeter H_0 sinon.

Ce sont les t-tests de nullité des coefficients.

On peut également réalisation un test pour décider entre

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs H_1 : il existe j entre 1 et p tel que β

C'est le F-test d'intérêt du modèle. Pour cela, on doit utiliser les résultats de la proposition 2.9 avec $\boldsymbol{\theta} = (\beta_1, \dots, \beta_p)$, $q = p$. Dans ce cas, la matrice $\mathbf{T}$ est donnée par

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

ce qui permet de se débarrasser de β_0. Dans le test au niveau $(1-\alpha)$, on décide :

— de conserver H_0 si $0 \in \mathscr{E}_{\mathbf{T}}$, i.e. si 0 est dans l'ellipsoïde de confiance de niveau $(1-\alpha)$,
— de rejeter H_0 sinon.

2.3 Prédictions

Le modèle linéaire a deux objectifs.

1. **Comprendre et expliquer la variabilité de Y grâce aux covariables.** Dans ce cas, on ajuste un modèle linéaire et on interprète
   — les effets $(\beta_j,\ j \neq 0)$,
   — la variabilité quand on fixe toutes les covariables (σ^2) et
   — le pourcentage de variabilité de Y expliqué par le modèle (R^2)...
2. **Prédire la valeur de Y pour un nouvel individu.** Dans ce cas, on se donne de nouvelles valeurs des covariables pour un individu : $\tilde{x}_1,\dots,\tilde{x}_p$. Et on veut :
   — estimer la moyenne de Y sur la sous-population où les covariables sont fixées à $\tilde{\mathbf{x}}$, à savoir le paramètre $\theta = \beta_0 + \beta_1\tilde{x}_1 + \cdots + \beta_p\tilde{x}_p$,
   — prédire une valeur de la réponse, notée $\tilde{Y}$, pour un individu fixé, issu de cette sous-population.

Ces deux objectifs peuvent être mener simultanément. Souvent, l'un des deux objectifs est privilégié. Le premier objectif est lié à des problèmes d'estimation, d'erreur d'inférence, entièrement traité dans ce qui précède via les variances des estimateurs ou les intervalles/régions de confiance. On se concentre maintenant sur le second objectif. On notera que le paramètre θ, de dimension 1, est de la forme $\mathbf{T}\boldsymbol{\beta}$, où $\mathbf{T}$ est la matrice ligne

$$\mathbf{T} = \begin{pmatrix} 1 & \tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_p \end{pmatrix}.$$

On se place maintenant sous les hypothèses du **modèle linéaire gaussien** de la section 2. La proposition 2.9 permet de donner un intervalle de confiance pour θ, en utilisant la matrice ligne $\mathbf{T}$ introduite ci-dessus.

Dans ce cas, on utilise l'estimateur $\widehat{\theta} = \mathbf{T}\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1\tilde{x}_1 + \cdots + \widehat{\beta}_p\tilde{x}_p$. Et on peut montrer que l'intervalle de confiance pour θ au niveau $(1-\alpha)$ est un intervalle centré en $\widehat{\theta}$, dont les bornes sont

$$\widehat{\theta} \pm u\widehat{\sigma}\sqrt{\mathbf{T}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{T}'},$$

où u est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p-1$ degrés de liberté.

Pour la prédiction de $\tilde{Y}$, on obtient un intervalle de prédiction de niveau $(1-\alpha)$ avec l'intervalle centré en $\widehat{\theta}$, dont les bornes sont

$$\widehat{\theta} \pm u\widehat{\sigma}\sqrt{1+\mathbf{T}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{T}'},$$

où u est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p-1$ degrés de liberté. Le terme $1+$ qui apparaît en rouge sous la racine carré vient du fait que l'on doit tenir compte de la variabilité individuelle d'un individu à l'autre dans la sous-population où les covariables sont fixées à $\tilde{\mathbf{x}}$. Cette variabilité individuelle est donnée par σ^2 (en variance), estimée par $\widehat{\sigma}^2$, et s'ajoute à la variabilité de l'estimateur, car le nouvel individu n'est pas dans l'échantillon des données.

2.4 Des covariables transformées

Variables catégorielles On ne peut pas ajouter directement dans une formule du type $\beta_0+\beta_1X_1+\cdots+\beta_pX_p$ une variable catégorielle. Lorsque l'on veut utiliser une variable catégorielle, il faut la transformer en une ou plusieurs **variables binaires** (=Dont les seules valeurs possibles sont 0 et 1).

Lorsqu'il s'agit d'une variable catégorielle à **deux modalités** (homme - femme ; diplômée - non diplômée ;...), on définit une variable X_j qui vaut 1 pour une modalité (ex : femme), et 0 pour l'autre modalité (ex : homme). Dans la formule linéaire, le terme $+\beta_jX_j$ peut donc prendre deux valeurs 0 ou β_j. On interprète alors β_j comme l'écart de moyenne entre les sous-populations pour la modalité codée par 1 (toutes les autres covariables étant fixées), et la modalité codée par 0 (toutes les autres covariables étant fixées aux mêmes valeurs). La modalité codée par 0 est appelée **modalité de référence**.

Lorsque la variable a K modalités, il faut introduire $K-1$ variables binaires :
— on fixe une modalité de référence a_0 parmi les K modalités,
— pour chacune des modalités a différentes de la modalité de référence, on introduit une variable binaire égale à 1 si cette modalité a est celle observée, 0 sinon.

Ces $(K-1)$ variables binaires portent la même information que la variable catégorielle à K modalités et,
— soit elles sont toutes nulles (c'est la modalité de référence qui est celle observée)
— soit une et une seule d'entre elle est égale à 1, les autres sont égales à 0.

On peut alors interpréter les coefficients en facteur de ces covariables binaires comme des différences de moyennes entre deux sous-populations, où, toutes les autres covariables différentes de ces $K-1$ variables binaires sont fixées, et l'une des deux sous-population correspond à la modalité de référence a_0, l'autre à la modalité liée à la covariable binaire a.

Interaction L'une des propriétés importantes des modèles linéaires est l'interprétation de β_j comme effet lorsqu'on augmente X_j d'une unité, sachant toutes les autres covariables fixées. Cet effet ne dépend pas des valeurs des valeurs des autres covariables. On peut chercher à rendre l'effet dépendant des autres covariables. Pour rester linéaire, si on veut faire dépendre cet effet de la valeur de X_k, on va

$$\text{remplacer} \quad \beta_j \quad \text{par} \quad \beta_j + \beta_{j,k}X_k.$$

Dans la formule linéaire, en reportant, on obtient

$$\begin{aligned}\beta_0+\beta_1X_1+\cdots+(\beta_j+\beta_{j,k}X_k)X_j+\cdots+\beta_pX_p \\ = \beta_0+\beta_1X_1+\cdots+\beta_pX_p+\beta_{j,k}X_kX_j.\end{aligned}$$

Cette opération revient donc à ajouter une nouvelle covariable, qui est le produit de deux covariables initiales, dans la formule linéaire. De telles nouvelles covariables s'appellent des interactions, et le nouveau coefficient en facteur $\beta_{j,k}$ s'interprète comme expliqué au dessus. (Notons que X_j et X_k jouent des rôles symétriques).

Attention, il ne faut jamais introduire une variable d'interaction (c'est-à-dire une variable produit) sans avoir utiliser les deux covariables dans le modèle. Sinon, on peut prendre pour une interaction ce qui serait en fait un effet direct de la covariable manquante...

Autres transformations non linéaires On peut introduire de nouvelles covariables à partir des anciennes en appliquant des transformations non linéaires aux covariables initiales. Par exemple, s'il n'y avait qu'une seule covariables X_1, on peut introduire les nouvelles covariables

$$X_2 = (X_1)^2,\ X_3 = (X_1)^3, \ldots, X_p = (X_1)^p.$$

Dans ce cas, la formule linéaire

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \beta_0 + \beta_1 X_1 + \beta_2 (X_1)^2 + \cdots + \beta_p (X_1)^p$$

est **un polynôme de degré p en X_1**. Mais les coefficients n'ont plus une interprétation simple...Le problème des polynômes de degré p est leur explosion vers $+\infty$ ou $-\infty$ à la vitesse $(X_1)^p$ quand $|X_1| \to +\infty$.

Plutôt qu'un polynôme, on peut utliiser une fonction **affine par morceau qui soit continue**, et où les jointures ont lieu en $\xi_1, \xi_2, \ldots, \xi_K$. Cela revient à introduire les nouvelles covariables :

$$X_2 = (X_1 - \xi_1)_+, \ldots, X_{K+1} = (X_1 - \xi_K)_+,$$

où, pour tout $u \in \mathbb{R}$, u_+ désigne la partie positive de u, c'est-à-dire $u_+ = \max(0, u)$. Notons que les ξ_k sont les points de discontinuité de la dérivée d'ordre 1.

Plutôt qu'une fonction affine par morceaux, on peut utiliser un polynôme par morceau de degré 3, C^2 sur tout $\mathbb{R}$, avec des discontinuité dans la dérivée d'ordre 3. De ce cas, cela revient à introduire dans le modèle linéaireles covariables :

$$X_2 = (X_1)^2,\ X_3 = (X_1)^3, X_4 = (X_1 - \xi_1)_+^3, \ldots, X_{K+3} = (X_1 - \xi_K)_+^3$$

On parle alors de **spline cubique**...Notons que les coefficients en facteur de ces covariables ne s'interprètent pas non plus.

Les splines cubiques dites naturelles sont des variantes de la situation ci-dessous où, sur les deux intervalles infinis $]-\infty; \xi_1[$ et $]\xi_K; +\infty[$, on impose que la fonction soit linéaire pour ne pas exploser trop vite vers l'infini. Cela ajoute $2 \times 2 = 4$ contraintes d'égalité sur les coefficients $\beta_1, \ldots, \beta_{K+3}$.

Chapitre 3

Compléments sur le modèle linéaire

Voici les hypothèses du modèle linéaire. On veut prédire une variable Y à l'aide de plusieurs covariables numériques $X_1, \ldots, X_p$. Un **individu** de la population est donc représenté par le vecteur $(X_1, \ldots, X_p, Y)$. On suppose que

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

où $\varepsilon \perp\!\!\!\perp (X_1, \ldots, X_p)$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$. Les paramètres de ce modèle sont $\beta_0, \ldots, \beta_p, \sigma^2$.

Sous l'hypothèse gaussienne, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

3.1 Autres méthodes d'inférence

3.1.1 Problèmes avec l'estimateur des moindres carrés

Dans toute cette partie, on supposera σ^2 connu. On rappelle que **l'estimateur des moindres carrés** est donné par

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Lorsque le nombre de covariables est important, et ou que celles-ci sont corrélées, l'estimateur devient instable. En effet, si les covariables sont centrées, toutes les colonnes de $\mathbf{X}$ sont centrées, sauf celle constante égale à 1 si on veut inclure β_0 dans le modèle. Le coefficient j, k de la matrice $\mathbf{X}'\mathbf{X}$ est alors

$$\sum_{i=1}^{n} (X_{i,j} - \bar{X}_j)(X_{i,k} - \bar{X}_k) = (n-1)\widehat{\sigma}(X_j, X_k).$$

Si la corrélation est importante, l'inversion de $\mathbf{X}'\mathbf{X}$ est instable. Et, par exemple

$$\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}^{-1} \approx \begin{pmatrix} 2.78 & -2.22 \\ -2..2 & 2.78 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}^{-1} \approx \begin{pmatrix} 5.26 & -4.74 \\ -4.74 & 5.26 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}^{-1} \approx \begin{pmatrix} 10.26 & -9.74 \\ -9.74 & 10.26 \end{pmatrix}$$

Toute erreur d'estimation dans la matrice de covariance due à l'échantillon des $X_{i,j}$ peut faire varier $(\mathbf{X}'\mathbf{X})^{-1}$ fortement.

L'estimateur des moindres carrées est donc non biaisé, mais de variance importante. Pour obtenir des estimateurs plus efficaces, il faut utiliser des estimateurs biaisés.

3.1.2 Estimation bayésienne

Le principe de l'estimation bayésienne est de coder l'information dont on dispose sur β **avant d'observer les données** sous forme d'une **loi a priori**. Puis de calculer une loi a posteriori, qui représente l'information après avoir observé les données.

Pour éviter que les valeurs de β n'explosent, on peut essayer de choisir une loi a priori qui favorise des valeurs de β autour de 0. Evidemment, l'unité dans laquelle la variable X_j influe sur la valeur de β_j :

$$\beta_j X_j = (\lambda\beta_j)\left(\frac{X_j}{\lambda}\right).$$

Autrement dit, quand on divise X_j par λ, la valeur de β_j est multipliée par λ.

Dans toute la suite de cette partie 1, on suppose que les variables X_j sont centrées et <u>réduites</u>.

L'a priori

$$\pi(\beta_0) \propto 1, \quad \beta_{1:p} \sim \mathcal{N}_p(0, \tau^2 I_d)$$

favorise les $\beta_{1:p}$ proches de 0. Comme toutes les covariables sont réduites, le fait que la variance a priori de β_i, $i = 1, \ldots, p$ soit identique, égale à τ^2 quelque soit i permet de tirer toutes les estimations de β_i vers 0 avec la même force.

Le modèle linéaire gaussien suppose que

$$\left[\mathbf{Y}\middle|\mathbf{X}, \boldsymbol{\beta}\right] \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$$

Dans ce cas, la loi a posteriori est donnée par

$$\pi\left(\boldsymbol{\beta}\middle|\mathbf{Y}, \mathbf{X}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 - \frac{1}{2\tau^2}\left\|\beta_{1:p}\right\|^2.\right)$$

La log-posterior est donc

$$\log\pi\left(\boldsymbol{\beta}\middle|\mathbf{Y}, \mathbf{X}\right) = -\frac{1}{2\sigma^2}\left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 - \frac{1}{2\tau^2}\left\|\beta_{1:p}\right\|^2 + \text{Cte.}$$

L'**estimateur du maximum a posteriori** (ou MAP) est celui qui maximie la log-posterior, donc qui minimise la fonction

$$\mathbf{b} \mapsto \left\|\mathbf{Y} - \mathbf{X}\mathbf{b}\right\|^2 + \frac{\sigma^2}{\tau^2}\left\|\mathbf{b}_{1:p}\right\|^2.$$

On voit ici que l'on se retrouve à minimiser le critère des moindres carrés, mais **pénalisé par la norme $\left\|\mathbf{b}_{1:p}\right\|^2$**. Cela revient à chercher un **compromis** entre une valeur de $\mathbf{b}$ qui minimise les moindres carrés, et une valeur qui minimise la norme. La valeur de τ règle se compromis :

— quand τ est grand, la loi a priori favorise peu les valeurs de $\beta_{1:p}$ autour de 0, et, de fait, la pénalisation du critère des moindres carrés est faible,
— quand τ est faible, l'a priori est fort, et la pénalisation est forte.

Il existe évidemment **d'autres lois a priori classiques** pour traiter le cas du modèle linéaire, comme les lois a priori de dites g**-prior de Zellner**, qui proposent

$$\left[\boldsymbol{\beta}\middle|\mathbf{X}\right] \sim \mathcal{N}(0, g\sigma^2\mathbf{X}'\mathbf{X}).$$

Cette loi a priori ne favorise pas les valeurs de $\boldsymbol{\beta}$ vers 0 en cas de forte corrélations entre celles-ci. Dans ce cas, la loi a posteriori est aussi explicite et est donnée par

$$\left[\boldsymbol{\beta}\middle|\mathbf{Y},\mathbf{X}\right] \sim \mathcal{N}\left(\frac{g}{g+1}\widehat{\boldsymbol{\beta}}_{\mathsf{ML}}, \frac{\sigma^2 g}{g+1}(\mathbf{X}'\mathbf{X})^{-1}\right)\dots$$

Quand $g = n$, cette loi a posteriori permet de retrouver des **résultats similaires à la statistique fréquentielle avec le maximum de vraisemblance**. Elle est plutôt utilisée pour faire du choix de modèle bayésien, voir plus tard dans le cours. . .

3.1.3 Estimateur ridge

Les **estimateurs ridges** sont exactement les estimateurs qui minimisent l'un des critères des moindres carrés pénalisés par la norme $\|\beta_{1:p}\|^2$. Ils permettent de **régler les problèmes de corrélations trop fortes dans les covariables.**

Remarquons que, comme les variables X_j sont centrées (et réduites), on a

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathbf{X}'_{-0}\mathbf{X}_{-0} & \\ 0 & & & \end{pmatrix}$$

où $\mathbf{X}'_{-0}\mathbf{X}_{-0}$ est la matrice de corrélation des covariables.

Par définition, les estimateurs ridge s'écrivent sous la forme

$$\widehat{\boldsymbol{\beta}}_\lambda^{\mathsf{ridge}} = \left(\mathbf{X}'\mathbf{X} + \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \lambda^2 I_p & \\ 0 & & & \end{pmatrix}\right)^{-1} \mathbf{X}'\mathbf{Y}, \quad (\lambda \geq 0).$$

Lorsque $\lambda = 0$, on retrouve l'estimateur des moindres carrés.

Ce sont les **solutions du problème d'optimisation du critère des moindres carrés pénalisés**

$$\mathbf{b} \mapsto \left\|\mathbf{Y} - \mathbf{X}\mathbf{b}\right\|^2 + \lambda^2\left\|\mathbf{b}_{1:p}\right\|^2.$$

Notons qu'il s'agit exactement du critère qui définissait le MAP lorsque la loi a priori sur $\beta_{1:p}$ est gaussienne centrée en 0, en posant $\lambda^2 = \sigma^2/\tau^2$.

La valeur de λ fixe la force de la correction. Elle doit être choisi en fonction du jeu de données, par validation croisée (voir plus bas). Notons, par exemple, que

$$\begin{pmatrix} 1.2 & 0.8 \\ 0.8 & 1.2 \end{pmatrix}^{-1} \approx \begin{pmatrix} 1.5 & -1.0 \\ -1.0 & 1.5 \end{pmatrix}, \qquad \begin{pmatrix} 1.2 & 0.9 \\ 0.9 & 1.2 \end{pmatrix}^{-1} \approx \begin{pmatrix} 1.90 & -1.43 \\ -1.43 & 1.90 \end{pmatrix},$$

$$\begin{pmatrix} 1.2 & 0.95 \\ 0.95 & 1.2 \end{pmatrix}^{-1} \approx \begin{pmatrix} 2.23 & -1.77 \\ -1.77 & 2.23 \end{pmatrix}.$$

On voit ici, avec $\lambda^2 = 0.2$ que l'on se tromper dans l'estimation de la covariance (qui vaut 0.8 dans cet exemple), et utiliser des estimations de l'ordre de 0.9 ou 0.95 sans que l'inverse n'explose (à comparer avec les calculs numériques de 3.1.1).

Une autre façon de comprendre ces estimateurs est de regarder **les valeurs propres de $\mathbf{X}'_{-0}\mathbf{X}_{-0}$, la matrice de corrélation des covariables**. Rappelons que la matrice de corrélation $\mathbf{X}'_{-0}\mathbf{X}_{-0}$ est symétrique, semi-définie positive. Elle est donc diagonalisable, et toutes ses valeurs propres sont ≥ 0. Dire que les covariables sont fortement corrélées revient à dire que certaines de ces valeurs propres sont proches de 0. En ajoutant λ^2 à toutes les valeurs propres, on les éloigne de 0.

Une dernière façon de voir ces estimateurs est de comprendre le critère des moindres carrés pénalisés

$$\mathbf{b} \mapsto \left\|\mathbf{Y} - \mathbf{X}\mathbf{b}\right\|^2 + \lambda^2\left\|\mathbf{b}_{1:p}\right\|^2.$$

comme le lagrangien du problème d'optimisation du critère des moindres carrés sous la contrainte

$$\left\|\beta_{1:p}\right\|^2 \leq r^2.$$

La valeur de r^2 dépend bien sûr de la valeur de λ^2 : r^2 est grand quand λ^2 est petit, et réciproquement. Cette contrainte empêche l'estimateur des moindres carrés d'être trop grand. . .À nouveau, on voit qu'il faut rendre toutes les valeurs de $\boldsymbol{\beta}_i$ comparable donc centrer **réduire** les covariables avant d'appliquer ridge.

3.1.4 Pénalisation Lasso

La pénalisation Lasso est une autre pénalisation du critère des moindres carrés qui permet de forcer l'annulation de certains effets β_i. Initialement proposé par Tibshirani en 1996, Lasso veut dire « *least absolute shrinkage and selection operator* ». Les estimateurs Lasso sont définis par

$$\widehat{\boldsymbol{\beta}}_\lambda^{\mathsf{Lasso}} = \operatorname{argmin}_{\mathbf{b}}\left(\left\|\mathbf{Y} - \mathbf{X}\mathbf{b}\right\|^2 + \lambda\left\|\mathbf{b}_{1:p}\right\|_1\right)$$

où

$$\left\|\mathbf{b}_{1:p}\right\|_1 = \sum_{i=1}^{p} |b_i|$$

est la norme L^1 ou de Manhattan de $\mathbf{b}_{1:p}$. Comme pour la régression ridge, on peut voir ce critère pénalisé comme un lagrangien du problème d'optimisation du critère des moindres carrés sous la contrainte

$$\left\|\mathbf{b}_{1:p}\right\|_1 \leq r.$$

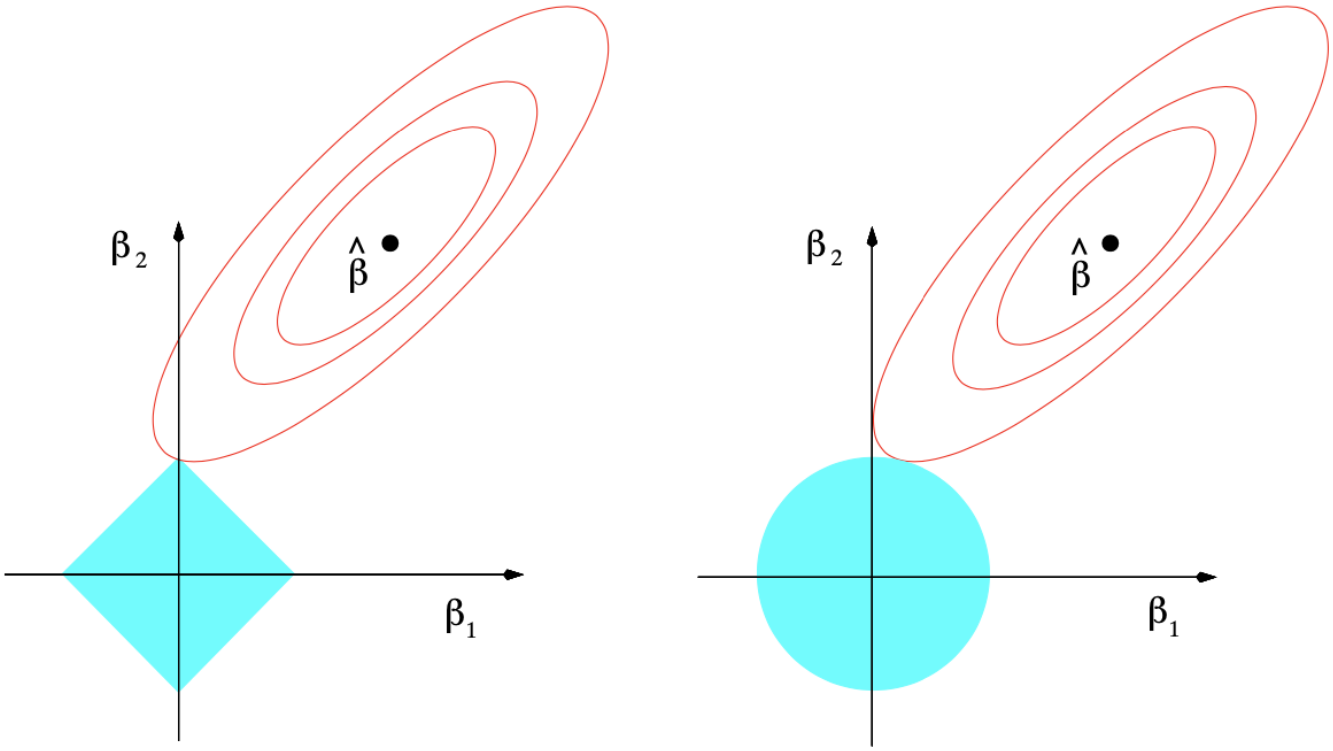

Figure 3.1 – **Estimation Lasso (à gauche) et ridge (à droite).** On s'intéresse à un problème où $\beta_0 = 0$. Les ellipses rouges sont les lignes de niveau du critère des moindres carrés, minimal en $\widehat{\beta}$, le maximum de vraisemblance.
On voit ici que l'estimateur Lasso va annuler $\widehat{\beta}_1^{\text{Lasso}}$, alors que l'estimateur ridge conserve une coordonnée $\widehat{\beta}_1^{\text{Ridge}}$ non nulle.

Comme la norme de Manhattan a des « boules » de rayon r qui sont des « losanges » avec des angles au niveau des axes, ce problème d'optimisation sous contraintes rend certaines coordonnées b_i nulles, voir Figure 3.1. Contrairement à l'estimateur ridge, il n'y a pas de formule explicite pour la solution de ce problème d'optimisation.

À nouveau, il existe autant d'estimateur Lasso que de valeurs de λ possibles. Il faudra trouver une façon de choisir cette valeur, par validation croisée par exemple.

3.2 Validation croisée

L'idée de la validation croisée est d'étudier l'erreur de généralisation d'une estimation sur de nouvelles données, en utilisant les données comme de nouvelles données. Autrement dit, on va

- enlever des observations,
- ajuster le modèle sur les observations conservées, puis finalement
- faire des prédications sur les observations enlevées au début.

On suppose ici que l'on veut comparer des estimateurs $\widehat{\boldsymbol{\beta}}_\lambda$ qui dépendent d'un paramètres λ à estimer. Comme dans les estimateurs ridge et lasso par exemple.

Fixons un jeu de données $\mathbf{X}, \mathbf{Y}$ sur lequel les $\widehat{\boldsymbol{\beta}}_\lambda$ sont calculés.

Et introduisons un nouvel individu $\widetilde{\mathbf{x}}$, $\widetilde{Y}$, indépendant du jeu de données. Si on veut prédire $\widetilde{Y}$ avec les covariables $\widetilde{\mathbf{x}}$ de l'individu, on utilise

$$\widehat{Y}_\lambda = \widetilde{\mathbf{x}}\widehat{\boldsymbol{\beta}}_\lambda.$$

L'erreur quadratique de prédiction est donnée par

$$SSE_{\text{gen}}(\lambda) = \mathbb{E}\left[(\widetilde{Y} - \widehat{Y}_\lambda)^2 \middle| \mathbf{X}, \mathbf{Y}\right] = \mathbb{E}\left[(\widetilde{Y} - \widetilde{\mathbf{x}}\widehat{\boldsymbol{\beta}}_\lambda)^2 \middle| \widehat{\boldsymbol{\beta}}_\lambda\right].$$

La somme des carrés des résidus est définie par

$$SSE(\lambda) = \sum_{i=1}^{n}\left(Y_i - \mathbf{X}_{i,\cdot}\widehat{\beta}_\lambda\right)^2 = \left\|\mathbf{Y} - \mathbf{X}\widehat{\beta}_\lambda\right\|^2,$$

où $\mathbf{x}_{i,\cdot}\widehat{\beta}_\lambda$ est la prédiction pour l'individu i. C'est une **estimation trop optimiste de l'erreur de prédiction** sur de nouvelles données.

Pour les estimateurs ridge ou lasso, si $0 \leq \lambda_1 < \lambda_2$, on a

$$SSE(\lambda_1) < SSE(\lambda_2).$$

3.2.1 Mettre une observation de côté

Il s'agit de la validation croisée dite « *leave-one-out* » ou LOO. Pour chaque observation de $i = 1$ à n, on va

- ajuster le modèle sans la i-ème observation Y_i, et donc calculer un $\widehat{\boldsymbol{\beta}}_\lambda$

— utiliser le modèle, donc la valeur de $\widehat{\boldsymbol{\beta}}_\lambda$ pour prédire Y_i,
— enregistrer la valeur prédite dans $\widehat{Y}_{i,\lambda}$.

Si on veut comparer différentes valeurs de λ pour l'estimateur ridge, ou l'estimateur lasso, on peut alors comparer les erreurs

$$SSE_{\text{LOO}}(\lambda) = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}_\lambda\right\|^2.$$

Sauf quand n est petit, on emploie plutôt d'autres méthodes. En effet, cette méthode de validation croisée est

— **lente** : on doit faire autant d'estimation qu'il y a d'observations
— **peu stable** : supprimer une seule observation ne permet pas de changer beaucoup le jeu de données.

3.2.2 K blocs

Cette fois-ci, on commence par diviser totalement au hasard le jeu de données en K blocs de tailles comparables. Et on prépare un vecteur $\widehat{\mathbf{Y}}_\lambda$ de dimension n, pour enregister les prédictions par validation croisée sur chaque observation. Puis, pour chaque bloc de $k = 1$ à K, on va

— ajuster le modèle sans le k-ème bloc d'observations, et donc calculer $\widehat{\boldsymbol{\beta}}_\lambda$
— utiliser le modèle, donc la valeur de $\widehat{\boldsymbol{\beta}}_\lambda$ pour prédire les Y_i du k-ème bloc,
— enregistrer les valeurs prédites de Y_i dans les bonnes coordonnées de $\widehat{\mathbf{Y}}_\lambda$.

À nouveau, l'erreur de généralisation de $\widehat{\boldsymbol{\beta}}_\lambda$ se calcule avec

$$SSE_K(\lambda) = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}_\lambda\right\|^2$$

En règle générale, on choisit $K = 5$ ou 10 blocs. Cette méthode de validation croisée est

— plus rapide : on ne doit faire l'estimation de $\boldsymbol{\beta}$ que K fois
— perturbe plus le jeu de données : à chaque fois, on enlève tout un bloc d'observations avant d'estimer (si $K = 5$, on enlève $\approx 20\%$ des observations)

3.2.3 Choisir λ

Une fois que l'on dispose d'un critère fiable, SSE_{LOO} ou SSE_K, on peut choisir λ avec

$$\widehat{\lambda} = \operatorname{argmin}_\lambda SSE_*(\lambda), \quad \text{où } * \in \{\text{LOO}, K\}.$$

L'estimateur de β retenu au final est donc $\widehat{\beta}_{\widehat{\lambda}}$ avec une valeur de λ calibrée sur le jeu de donnée par validation croisée.

Le problème d'optimisation de $SSE_*(\lambda)$ est résolu ainsi.

— On fixe une grille de valeurs de λ : $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_T$.
— On calcule $SSE_*(\lambda)$ en chacun des points de cette grille par validation croisée.
— On retient la valeur de λ qui donne la plus petite valeur.

Dernier problème : $SSE_*(\widehat{\lambda})$ est lui-même est une estimation trop optimiste de l'erreur de prédiction pour la valeur de λ choisie. Il faudrait donc avoir de nouvelles données pour évaluer son erreur de prédiction.

On ne peut plus utiliser la technique de validation croisée. Il faut donc garder **dès le début de l'étude** une petite partie des données comme jeu de données de **test** pour s'en servir comme « nouvelles données » à la fin.

Au final, le jeu de données est divisé est trois parties :

— **entraînement** : partie sur laquelle on ajuste les $\widehat{\beta}_\lambda$
— **validation** : partie sur laquelle on choisie λ
— **test** : partie sur laquelle on évalue l'erreur de prédication finale.

Les techniques de validation croisée permettent de regrouper entraînement et validation. Notons enfin que notre critère d'erreur ici était une différence au carré. On peut utiliser d'autres critères d'erreur (somme des valeurs absolues,...)

Chapitre 4

Analyse de la variance

4.1 Variabilité expliquée par le modèle

Dans le modèle linéaire,

$$\begin{aligned}\widehat{\mathbf{Y}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \widehat{\beta}_0\mathbf{X}_{\cdot,0} + \widehat{\beta}_1\mathbf{X}_{\cdot,1} + \cdots + \widehat{\beta}_p\mathbf{X}_{\cdot,p} \in \mathrm{Im}\mathbf{X}.\end{aligned}$$

Sont équivalents :

(i) choisir $\widehat{\boldsymbol{\beta}}$ avec

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_\beta \left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|^2$$

(ii) choisir $\widehat{\boldsymbol{\beta}}$ tel que

$$\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \perp \mathrm{Im}\mathbf{X}$$

(iii) choisir $\widehat{\boldsymbol{\beta}}$ tel que $\mathbf{X}\widehat{\boldsymbol{\beta}}$ soit la projection orthogonale de $\mathbf{Y}$ sur $\mathrm{Im}\mathbf{X}$

Donc,

$$\begin{aligned}\mathbf{Y} &= \underbrace{\mathbf{Y} - \widehat{\mathbf{Y}}}_{\in \mathrm{Im}\mathbf{X}^\perp} + \underbrace{\widehat{\mathbf{Y}}}_{\in \mathrm{Im}\mathbf{X}} \quad \text{et} \\ \|\mathbf{Y}\|^2 &= \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 + \|\widehat{\mathbf{Y}}\|^2\end{aligned}$$

avec le théorème de Pythagore.

Pour obtenir la décomposition classique de la variance, il faut travailler dans le sous-espace orthogonal à $\mathbf{1} \in \mathbb{R}^n$. Notons que $\mathbf{1} \in \mathrm{Im}\mathbf{X}$. Si on utilise la première égalité, on obtient encore

$$\mathbf{Y} - \bar{\mathbf{Y}} = \underbrace{\mathbf{Y} - \widehat{\mathbf{Y}}}_{\in \mathrm{Im}\mathbf{X}^\perp} + \underbrace{\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}}_{\in \mathrm{Im}\mathbf{X}}$$

où $\bar{\mathbf{Y}} = (\bar{Y}, \ldots, \bar{Y})$. Donc

$$SST = SSR + SSE, \quad \text{où}$$

- $SST = \|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2$ est la variabilité totale,
- $SSR = \|\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2$ est la variabilité expliquée par le modèle et
- $SSE = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$ est la variabilité résiduelle

4.2 Analyse de la variance à un facteur

4.2.1 Le point de vue décomposition de la variance

On s'intéresse à un modèle où il n'y a qu'une seule covariable catégorielle G à $p+1$ modalités, que l'on doit remplacer par des variables binaires $X_1, \ldots, X_p$.

Il est usuel de représenter la décomposition de la variance comme dans la table 4.1

Les deux dernières colonnes n'ont de sens que sous les hypothèses gaussiennes, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Le F-test et sa p-value permet de tester si les effets en facteurs des binaires sont tous nuls ou si l'un au moins d'entre eux est non nul. On peut le voir comme un test de comparaison de $p+1$ groupes.

4.3 Tests de comparaison de groupes

4.3.1 Tests de Student

Supposons que la population est divisée en deux groupes, et que l'on s'intéresse à comparer les moyennes μ_0, μ_1 de Y au sein des deux groupes. On veut tester

$$H_0 \ : \ \mu_0 = \mu_1, \quad \text{contre} \quad H_1 \ : \ \mu_0 \neq \mu_1.$$

On dispose d'un échantillon de taille n que l'on peut renuméroter de telle sorte que $Y_{0,1}, \ldots, Y_{0,n_0}$ modélisent les observations au sein du premier groupe, et $Y_{1,1}, \ldots, Y_{1,n_1}$ au sein du second groupe, avec $n_0 + n_1 = n$.

Pour mimer les hypothèses gaussiennes et d'homoscédasticité, on suppose que les $Y_{i,j}$ sont indépendants et que

$$\begin{aligned}Y_{0,1}, \ldots, Y_{0,n_0} &\sim \mathcal{N}(\mu_0; \sigma^2), \\ Y_{1,1}, \ldots, Y_{1,n_1} &\sim \mathcal{N}(\mu_1; \sigma^2).\end{aligned}$$

Dans ce cas, on estime les deux moyennes et σ^2 par

$$\begin{aligned}\widehat{\mu}_0 &= n_0^{-1}(Y_{0,1} + \cdots + Y_{0,n_0}), \\ \widehat{\mu}_1 &= n_1^{-1}(Y_{1,1} + \cdots + Y_{1,n_1}), \\ \widehat{\sigma}^2 &= \frac{1}{n-2}\left(\sum_{j=1}^{n_0}(Y_{0,j} - \widehat{\mu}_0)^2 + \sum_{j=1}^{n_1}(Y_{1,j} - \widehat{\mu}_1)^2\right).\end{aligned}$$

Il est facile de voir que $\widehat{\mu}_0$ et $\widehat{\mu}_1$ sont indépendants, et que

$$\widehat{\mu}_i \sim \mathcal{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right), \quad i = 0, 1.$$

Table 4.1 – Table d'analyse de la variance à un facteur.

	DF	SS	MS	F-value	p-value
G	p	SSR	SSR/p	$\frac{SSR/p}{SSE/(n-p-1)}$	*
residuals	$n-p-1$	SSE	$SSE/(n-p-1)$		
total	$n-1$	SST			

De plus, avec le théorème de Cochran, $\widehat{\sigma}^2$ est indépendant de $(\widehat{\mu}_0, \widehat{\mu}_1)$ et

$$(n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Pour réaliser le test, on introduit la statistique

$$T = \frac{\widehat{\mu}_0 - \widehat{\mu}_1}{\widehat{\sigma}\sqrt{n_0^{-1} + n_1^{-1}}}.$$

Sous l'hypothèse nulle H_0, les résultats précédents montrent que T suit $t(n-2)$, la loi de Student à $n-2$ degrés de liberté. Sous l'hypothèse alternative, à mesure que μ_0 s'éloigne de μ_1, $|T|$ devient grand.

On décide donc en faveur de H_0 si la valeur observée de T est dans un intervalle de la forme $[-c;c]$ et en faveur de H_1 sinon. La valeur de c est choisie pour que le risque maximal de première espèce soit égal à α, ce qui conduit à

$$c = \Phi_{n-2}^{-1}(1-\alpha/2) = \text{le quantile d'ordre } 1-(\alpha/2) \text{ de } t(n-2).$$

Et la p-value du test est

$$p(t_{\text{obs}}) = \mathbb{P}_{T\sim t(n-2)}\Big(|T| > |t_{\text{obs}}|\Big) = 2\Big(1-\Phi_{n-2}(|t_{\text{obs}}|)\Big),$$

où Φ_{n-2} est la fonction de répartition de $t(n-2)$.

Remarque. Les équations

$$Y_{0,1},\dots,Y_{0,n_0} \sim \mathcal{N}(\mu_0;\sigma^2),$$
$$Y_{1,1},\dots,Y_{1,n_1} \sim \mathcal{N}(\mu_1;\sigma^2).$$

reviennent à dire que

$$Y_{i,j} = \mu_i + \varepsilon_{i,j}, \quad \text{où } \varepsilon_{ij} \sim \mathcal{N}(0,\sigma^2)$$

avec indépendance des $\varepsilon_{i,j}$. Quitte à poser $\beta_0 = \mu_0$ et $\beta_1 = \mu_1 - \mu_0$, on a

$$Y_{i,j} = \beta_0 + \beta_1 \mathbf{1}\{i=1\} + \varepsilon_{i,j}.$$

Il s'agit bien du modèle linéaire avec une seule variable explicative binaire, qui est l'indicatrice d'être dans le groupe n°1. Le t-test présenté plus haut est exactement le test de nullité de β. Dans ce cas où il n'y a qu'une seule covariable, ce t-test est équivalent au F-test (ou test de Fisher d'influence des covariables).

4.3.2 Le point de vue test de comparaison de plusieurs groupes

Lorsque la population est divisée en K groupes, cela veut dire qu'il existe une variable catégorielle avec K modalités (les noms des groupes). Pour l'introduire dans le modèle linéaire, il faut donc :

— choisir une modalité de référence (numérotée 0),
— introduire autant de variables binaires qu'il y a d'autres modalités : $X_1,\dots,X_{K-1}$.

Le modèle linéaire est donc, pour un individu $(X_1,\dots,X_{K-1},Y)$ pris au hasard dans la population :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \text{où } \varepsilon \sim \mathcal{N}(0,\sigma^2).$$

Si on note G le numéro du groupe auquel appartient cet individu, on a

$$\mathbb{E}\Big(Y\Big|G=i\Big) = \begin{cases} \beta_0 & \text{si } i=0, \\ \beta_0+\beta_i & \text{si } i=1,\dots,K-1. \end{cases}$$

Autrement dit, si on pose $\mu_i = \mathbb{E}\Big(Y\Big|G=i\Big)$ la moyenne de Y au sein du groupe numéro i,

Proposition 4.1. *On a $\beta_0 = \mu_0$ et pour $i=1,\dots,K-1$, on a $\beta_i = \mu_i - \mu_0$.*

Dans cette situation, si on a un jeu de données de n observations, il est courant de les re-numéroter de la façon suivante :

— $Y_{0,1},\dots,Y_{0,n_0}$ sont les n_0 observations du groupe 0 (groupe de référence),
— $Y_{i,1},\dots,Y_{i,n_i}$ sont les n_i observations du groupe i.

Le nombre total d'observations est $n = n_0 + \cdots + n_{K-1}$.

Le modèle statistique sur les observations est donc

$$Y_{0,j} \sim \mathcal{N}(\beta_0,\sigma^2), \quad Y_{i,j} \sim \mathcal{N}(\beta_0+\beta_i,\sigma^2), \text{ si } i=1,\dots,K-1.$$

Proposition 4.2. *Soit*

$$\widehat{\mu}_i = n_i^{-1}(Y_{i,1} + \cdots + Y_{i,n_i})$$

L'estimateur du maximum de vraisemblance de $(\beta_0,\dots,\beta_{K-1},\sigma^2)$ est donné par

$$\widehat{\beta}_0 = \widehat{\mu}_0, \quad \widehat{\beta}_i = \widehat{\mu}_i - \widehat{\mu}_0 \ (i=1,\dots,K-1)$$

et

$$\widehat{\sigma}^2 = \frac{1}{n-K}\sum_{i=0}^{K-1}\sum_{j=1}^{n_i}(Y_{i,j}-\widehat{\mu}_i)^2.$$

L'objectif de l'analyse de la variance est de tester

$$H_0 \ : \ \mu_0 = \mu_1 = \cdots = \mu_{K-1} \quad \text{contre} \quad H_1 \ : \ \exists i \neq i', \ \ \mu_i \neq \mu_{i'}.$$

Vue la proposition 4.1, il s'agit du F-test d'intérêt du modèle linéaire où H_0 est $\beta_1 = \cdots = \beta_{K-1}$.

Notons que si l'on décide en faveur de H_1 avec de F-test, on ne sait pas quels sont les deux groupes qui se distinguent par leurs moyennes. Pour répondre à cette question, il faut faire autant de t-tests que de paires de groupes, ce qui soulève un problème de tests multiples qui dépasse le contenu de ce cours.

4.4 Au delà de l'analyse de la variance à un facteur

4.4.1 Le problème de la décomposition de SST

Pour examiner un modèle linéaire, on le compare à des versions simplifiées de lui-même, de manière à comprendre l'intérêt des variables dans la prédiction de Y. Nous allons donc comparer deux modèles :

— le modèle M_q, ajusté sur la matrice de design $\mathbf{X}^{(q)}$ de rang q,
— le modèle M_r, ajusté sur la matrice de design $\mathbf{X}^{(r)}$ de rang $r > q$.

On suppose que les modèles sont emboîtés, c'est-à-dire que $\mathrm{Im}\mathbf{X}^{(q)} \subset \mathrm{Im}\mathbf{X}^{(r)} \subset \mathbb{R}^n$, où n est le nombre d'observations. Cette inclusion est vraie quand, par exemple, on enlève une colonne ou un ensemble de colonne à la matrice de design $\mathbf{X}$ originale.

Pour simplifier, on va supposer que $\mathbf{1} \in \mathrm{Im}\mathbf{X}^{(q)}$.

Si on ajuste le modèle M_r, on obtient une estimation $\widehat{\boldsymbol{\beta}}^{(r)}$ et une prédiction

$$\widehat{\mathbf{Y}}^{(r)} = \mathbf{X}^{(r)}\widehat{\boldsymbol{\beta}}^{(r)} \in \mathrm{Im}\mathbf{X}^{(r)}$$

De même pour le modèle M_q, on obtient une estimation $\widehat{\boldsymbol{\beta}}^{(q)}$ et une prédiction

$$\widehat{\mathbf{Y}}^{(q)} = \mathbf{X}^{(q)}\widehat{\boldsymbol{\beta}}^{(q)} \in \mathrm{Im}\mathbf{X}^{(q)}.$$

Malheureusement, dans la décomposition ci-dessous

$$\mathbf{Y} = \widehat{\mathbf{Y}}^{(q)} + \left(\widehat{\mathbf{Y}}^{(r)} - \widehat{\mathbf{Y}}^{(q)}\right) + \left(\mathbf{Y} - \widehat{\mathbf{Y}}^{(r)}\right)$$

n'est pas la somme de trois vecteurs orthogonaux en règle générale. Pour obtenir une décomposition de SST, il faut donc travailler différemment. La théorie complète est compliquée et dépasse largement le cadre de ce que l'on peut exposer dans ce cours.

Un cas simple où ces trois vecteurs sont orthogonaux est celui où l'on ajoute à $\mathbf{X}^{(q)}$ des colonnes qui sont orthogonale à $\mathrm{Im}\mathbf{X}^{(q)}$:

$$\mathbf{X}^{(r)} = \left(\mathbf{X}^{(q)} \middle| \mathbf{X}^{(r-q)}\right).$$

Cela revient à supposer que **les variables que l'on ajoute sont décorellées des variables de** M_q.

4.4.2 Analyse de la variance à deux variables indépendantes

Dans cette section, on suppose s'intéresse à **deux variables catégorielles indépendantes**. Autrement dit, il y a deux partitions de population :

— une première partition en K groupes,
— une seconde partition en L groupes.

On note G la variable qui décrit le numéro du groupe (de 0 à $K-1$) de la première partition. Et H celle pour la seconde partition, de 0 à $L-1$.

L'hypothèse d'indépendance revient à dire que, pour tout $i \geq 0$, tout $j \geq 0$,

$$\mathbb{P}(G = i, H = j) = \mathbb{P}(G = i)\mathbb{P}(H = j).$$

Introduire deux partitions revient en fait à introduire une seule partition à $K \times L$ groupes, chaque groupe de cette dernière partition est alors caractérisée par la valeur de la paire (G, H).

Il est d'usage de re-numéroter les observations, et les variables aléatoires qui les modélisent, de telle sorte que

$$Y_{i,j,k}$$

est la k-ème observation appartenant aux groupes $G = i$ et $H = j$ simultanément. Le nombre de telles observations est noté $n_{i,j}$.

Si on note $\mu_{i,j} = \mathbb{E}\big(Y \big| G = i, H = j\big)$, le modèle où

$$Y_{i,j,k} \sim \mathcal{N}(\mu_{i,j}, \sigma^2), \quad k = 1, \ldots, n_{i,j}$$

correspond à un modèle linéaire utilisant les variables G, H **avec leurs interactions**. On introduit les variables binaires :

$$U_i = \mathbf{1}\{G = i\} \ (i = 1, \ldots, K-1) \quad \text{et} \quad V_j = \mathbf{1}\{H = j\} \ (j = 1, \ldots, L-1).$$

Alors,

$$Y = \mu + \sum_{i=1}^{K-1} \alpha_i U_i + \sum_{j=1}^{L-1} \beta_j V_j + \sum_{i=1}^{K-1}\sum_{j=1}^{L-1} \gamma_{i,j} U_i V_j + \varepsilon, \quad \text{où } \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Dans ce cas, on a

Proposition 4.3. *Pour $i = 1, \ldots, K-1$ et $j = 1, \ldots, L-1$,*

$$\mu_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$$

Pour $i = 1, \ldots, K-1$,

$$\mu_{i,0} = \mu + \alpha_i.$$

Pour $j = 1, \ldots, L-1$,

$$\mu_{0,j} = \mu + \beta_j.$$

Enfin

$$\mu_{0,0} = \mu.$$

Réciproquement, on a, si $i, j \geq 1$

$$\begin{cases} \mu = \mu_{0,0}, \quad \alpha_i = \mu_{i,0} - \mu_{0,0}, \quad \beta_j = \mu_{0,j} - \mu_{0,0} \\ \gamma_{i,j} = \mu_{i,j} - \mu_{i,0} - \mu_{0,j} + \mu_{0,0} \end{cases}$$

4.4.3 Le cas équilibré

On suppose que le nombre d'observations $n_{i,j}$ ne dépend pas de (i,j) et donc,

$$\forall i, \forall j, \quad n_{i,j} = n_i n_j$$

où $n_i > 1$ et $n_j > 1$ sont deux nombres entiers non nuls. C'est le cas le plus simple. Mais, le nombre total d'observations peut devenir rapidement grand.

On pose

$$\widehat{\mu}_{i,j} = n_{i,j}^{-1} \sum_{k=1}^{n_{i,j}} Y_{i,j,k}.$$

Proposition 4.4. *L'estimateur du maximum de vraisemblance du modèle avec interaction est donné par*

$$\begin{cases} \widehat{\mu} = \widehat{\mu}_{0,0}, \quad \widehat{\alpha}_i = \widehat{\mu}_{i,0} - \widehat{\mu}_{0,0}, \quad \widehat{\beta}_j = \widehat{\mu}_{0,j} - \widehat{\mu}_{0,0} \\ \widehat{\gamma}_{i,j} = \widehat{\mu}_{i,j} - \widehat{\mu}_{i,0} - \widehat{\mu}_{0,j} + \widehat{\mu}_{0,0} \end{cases}$$

et

$$\widehat{\sigma}^2 = \frac{1}{n - KL} \sum_{i=0}^{K-1} \sum_{j=0}^{L-1} \sum_{k=1}^{n_{i,j}} (Y_{i,j,k} - \widehat{\mu}_{i,j})^2.$$

Les tableaux d'analyse de la variance se présentent comme dans la table 4.2.

Le test, sur la ligne des interactions $G:H$, permet de savoir si celles-ci sont utiles. Les tests, sur les lignes G et H, permettent de savoir si ces variables ont un effet moyen, ou si leur effet n'apparait qu'au travers des interactions. Suivant les cas pratiques, ces test ont ou n'ont pas de sens...

4.4.4 Plans d'expérience déséquilibrés

Ce sont typiquement des cas où l'on n'a pas pu échantillonner tous les cas $G = i$ et $H = j$, pour $i = 0, \dots, K-1$ et $j = 0, \dots, L-1$. Bien souvent, on ne peut pas inclure les interactions dans les modèles sur de tels jeux de données par manque d'observations.

La section 2.5 du chapitre 2 de cours « *Le Modèle Linéaire et ses Extensions* » de L. Bel, J.J. Daudin *et al.* présente le cas d'un plan en blocs incomplets au travers d'un exemple...

4.5 Analyse de la covariance

Il existe tout un bestiaire d'analyse de la variance. Voir, par exemple
« *Analysis of Variance and Covariance* » de Doncaster et Davey.

Le plan du livre est disponible ici : https://www.southampton.ac.uk/

et les exemples là :
https://www.southampton.ac.uk/ cpd/anovas/datasets/index.htm

La section 2.6 du chapitre 2 de cours « *Le Modèle Linéaire et ses Extensions* » de L. Bel, J.J. Daudin *et al.* présente un cas où les covariables sont corrélées, et non contrôlables par plan d'expérience. ...

Table 4.2 – Table d'analyse de la variance à plusieurs facteurs.

	df	**Sum Sq**	**Mean Sq**	F-**value**	p-**value**
G	$K-1$	SSG	$SSG/(K-1)$		
H	$L-1$	SSH	$SSH/(L-1)$		
$G:H$	$(K-1)\times(L-1)$	$SSGH$	$SSGH/((K-1)(L-1))$		
residuals	$n-KL$	SSE	$\widehat{\sigma}^2$	—	—
total	$n-1$	SST			

Chapitre 5

Choix de modèle

On rappelle que, dans le modèle linéaire, un **individu** de la population est donc représenté par le vecteur $(X_1, \dots, X_p, Y)$. On suppose que

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

où $\varepsilon \perp\!\!\!\perp (X_1, \dots, X_p)$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$. Les paramètres de ce modèle sont $\beta_0, \dots, \beta_p, \sigma^2$.

Malgré son apparante simplicité, le modèle linéair a des avantages en termes d'**interprétabilité** et souvent il fournit de bonnes **performances prédictives**. L'ajustement de ces modèles se fait par moindres carrés ou par maximum de vraisemblance.

Nous avons déjà vu deux alternatives à l'ajustement par moindres carrés :

- **ridge**, pour traiter les cas où les covariables sont fortement corrélées ($\Longleftrightarrow$ matrice de corrélation a des valeurs propres proches de 0)
- **lasso**, pour supprimer les covariables les moins importantes.

Ici, nous allons chercher à supprimer des covariables de la liste, pour trouver un meilleur modèle. Nous allons distinguer deux cas ici :

- **interprétation** : on cherche le modèle qui ne contient que les covariables nécessaires. Dans ce cas, les variables exclues n'apportent aucune information complémentaires sur Y au delà des variables que l'on conserve.
- **prédiction** : on cherche le modèle qui a les meilleures qualités prédictives sur de nouvelles données.

5.1 Critères de sélection

On peut utiliser différents critères pour comparer des modèles. Pour l'instant, on peut imaginer que l'on compare deux modèles $M_{(1)}$ et $M_{(2)}$ construits à partir de deux sous-ensembles des covariables. Voici quelques critères.

- R^2 **ajusté** : on cherche à viser le meilleur modèle en termes de prédiction. Mais, même corrigé, le critère de R^2 reste trop optimiste quant aux qualités prédictives d'un modèle. Ce n'est pas un bon critère, sauf s'il y a peu de covariables, et beaucoup de données.
- **BIC** : *Bayesian Information Criterion* ou critère de Schwartz. Il s'agit d'un critère qui visible à trouver le « vrai » modèle qui a engendré les données, et doit donc être utiliser si on veut obtenir un modèle **interprétable**.
- **AIC** : *Akaike Information Criterion*. Il s'agit d'un critère qui construit un modèle avec de bonnes **qualités prédictives**, en réalisant un compromis entre
  - le **biais**, éventuellement introduit en otant des covariables qui auraient pu être utiles et
  - la **variance des coordonnées** de $\widehat{\boldsymbol{\beta}}$ lorsque le nombre de covariables est grand.
- **erreur de généralisation calculée par validation croisée** : ce genre de critères visent également des **qualités prédictives**, mais sont parfois un peu lent en temps de calcul.

Sur le modèle $M_{(i)}$, notons

- $\mathbf{X}_{(i)}$ la matrice de design du modèle,
- $\boldsymbol{\beta}_{(i)}, \sigma^2_{(i)}$ les paramètres du modèle,
- $\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\sigma}^2_{(i)}$ leurs estimations par maximum de vraisemblance.

Les critères BIC et AIC sont définis en partant de la valeur de la log-vraisemblance maximale. Rappelons que la **log-vraisemblance** est donnée par

$$\ell_{(i)}(\boldsymbol{\beta}_{(i)}, \sigma^2_{(i)}) = -\frac{1}{2\sigma^2_{(i)}} \left\| \mathbf{Y} - \mathbf{X}_{(i)} \boldsymbol{\beta}_{(i)} \right\|^2 - \frac{n}{2} \log \sigma^2_{(i)}.$$

et son maximum vaut

$$\widehat{\ell}_{(i)} = -\frac{1}{2\widehat{\sigma}^2_{(i)}} \left\| \mathbf{Y} - \mathbf{X}_{(i)} \widehat{\boldsymbol{\beta}}_{(i)} \right\|^2 - \frac{n}{2} \log \widehat{\sigma}^2_{(i)}.$$

L'estimateur du maximum de vraisemblance est également l'estimateur des moindres carrés. Ainsi, si le modèle $M_{(1)}$ est strictement inclus dans le modèle $M_{(2)}$, on a

$$\widehat{\ell}_{(2)} > \widehat{\ell}_{(1)}$$

On ne peut donc pas choisir i en maximisant la log-vraisemblance $\widehat{\ell}_{(i)}$.

On définit les **critères de vraisemblance pénalisés** par

$$\mathrm{AIC}_{(i)} = 2d_{(i)} - 2\widehat{\ell}_{(i)},$$
$$\mathrm{BIC}_{(i)} = d_{(i)} \log n - 2\widehat{\ell}_{(i)},$$

où $d_{(i)} = \mathrm{rg}(\mathbf{X}_{(i)})$. (Attention, le signe de la définition change d'un livre ou d'un logiciel à l'autre).

Le meilleur modèle au sens de ces critères est celui qui minimise leurs valeurs.

Remarque. Dans le cadre de la régression linéaire, le critère des « C_p de Mallow » est équivalent au critère AIC :

$$C_p = \frac{1}{n}\left(SSE_{(i)} + 2d_{(i)} \widehat{\sigma}^2_{(i)} \right).$$

Le critère du R^2 **ajusté** est défini par

$$R^2 \text{ ajusté} = 1 - \frac{SSE/(n - d_{(i)} - 1)}{SST/(n-1)}.$$

5.2 Algorithmes de sélections de variables

Dans toute cette partie, on suppose le critère fixé.

5.2.1 Parcours exhaustif

L'objectif de cet algorithme est de comparer tous les modèles possibles à partir des covariables $X_1, \ldots, X_p$.

1. Notons M_0 le modèle nul, qui ne contient que l'intercept. Ce modèle prédit Y avec $\bar{Y}$.
2. Pour chaque valeur de k entre 1 et p :
   (a) Ajuster chacun des $\binom{p}{k}$ modèles linéaires à k covariables,
   (b) Choisir parmi ces modèles le meilleur modèle au sens du plus petit SSE ou du plus grand R^2, et le nommer M_k
3. Parmi les modèles $M_0, M_1, \ldots, M_p$ choisir le meilleur modèle au sens du critère choisi.

Cette méthode ne convient que si le nombre total p de covariables est faible.
— Elle est coûteuse car elle nécessite d'ajuster de très nombreux modèles.
— Plus le nombre de modèles ajustés est grand, plus les chances de trouver un modèle qui semble correct uniquement sur les données, avec de mauvaises propriétés de généralisation est grand. Cela peut donc conduire à un problème de **sur-apprentissage**.
Pour ces deux raisons, on utilise plutôt des méthodes pas à pas qui ne regardent pas tous les sous-modèles du modèles complets.

5.2.2 Parcours pas-à-pas

Sélection progressive L'idée est de partir du modèle nul M_0 qui ne contient que l'intercept, et d'ajouter progressivement des covariables.

1. Notons M_0 le modèle nul, qui ne contient que l'intercept. Ce modèle prédit Y avec $\bar{Y}$.
2. Pour chaque valeur de k entre 0 et $p-1$:
   (a) Ajuster tous les $p-k$ modèles linéaires qui consistent à ajouter une nouvelle covariables à M_k.
   (b) Choisir parmi ces modèles le meilleur modèle au sens du plus petit SSE ou du plus grand R^2, et le nommer M_{k+1}
3. Parmi les modèles $M_0, M_1, \ldots, M_p$ choisir le meilleur modèle au sens du critère choisi.

Sélection rétrograde L'idée est de partir du modèle complet M_p qui contient toutes les covariables et d'enlever progressivement des covariables.

1. Notons M_p le modèle complet, qui contient toutes les covarialbes
2. Pour chaque valeur de k entre p et 1 :
   (a) Ajuster tous les k modèles linéaires qui consistent à enlever une covariables à M_k.
   (b) Choisir parmi ces modèles le meilleur modèle au sens du plus petit SSE ou du plus grand R^2, et le nommer M_{k-1}
3. Parmi les modèles $M_0, M_1, \ldots, M_p$ choisir le meilleur modèle au sens du critère choisi.

5.3 Réduction de dimension

On peut également utiliser des méthodes de réduction de dimension sur la matrice $\mathbf{X}$ pour réduire le nombre de covariables.

L'ACP permet de se concentrer sur quelques variables (les premiers axes). On parle alors de régression sur composantes principales. Mais ces axes sont obtenus par une méthode **non supervisée** puisque la réponse Y n'influe pas sur le calcul de ces axes. Il n'y a aucune garantie que les axes principaux soient les meilleures variables pour prédire la réponse Y.

La méthode **Partial Least Square** (PLS) est une méthode de réduction de la dimension qui construit de nouvelles variables Z_1, Z_2,. . .qui sont des combinaisons linéaires des variables originales comme dans l'ACP. On suppose ici que les covariables X_j sont centrées-réduites.

Voici une idée du fonctionnement de PLS. La première variable est

$$Z_1 = \sum_{i=1}^{p} \varphi_{1,j} X_j$$

où $\varphi_{1,j}$ est l'effet estimé lorsque l'on régresse Y sur la seule covariable X_j.

Les directions suivantes sont obtenues en prenant les résidus de la régression de Y sur Z_1 et en répétant la règle ci-dessus en utilisant ces résidus comme nouvelle réponse à prédire. . .

Dans la plupart des cas, PLS est la méthode la plus efficace pour trouver un modèle ayant de bonnes qualités prédictives.

Chapitre 6

Introduction aux modèles linéaires généralisés

Dans tout ce qui a précédé, on a cherché à prédire
— une réponse Y **numérique**,
— à l'aide de covariables $X_1, \dots, X_p$ **numériques**.
Et on a vu comment revenir à des covariables numériques si initialement catégorielles.

Dans ce chapitre, on va s'intéresser à deux autres cas :

(i) $Y \in \{0,1\}$ est **binaire** (Exemples : sain/malade ; remboursera/fera défaut ; républicain/démocrate, etc.)

(ii) $Y \in \mathbb{N}$ est un **comptage** (Exemples : nombres d'enfants ; nombres de décès causés par... ; nombres de buts dans un match)

Dans ces deux cas, l'hypothèse du **modèle linéaire gaussien**, à savoir

$$[Y|X_{1:p}] \sim \mathcal{N}\big(\mu(X_{1:p}), \sigma^2\big),$$

n'est pas réaliste. Il faut donc en changer. En outre, on rappelle que, dans le modèle linéaire gaussien, on a

$$\mathbb{E}(Y|X_{1:p}) = \mu(X_{1:p}) = \beta_0 + \sum_j \beta_j X_j.$$

On va donc regarder deux modèles.

(i) Si $Y \in \{0,1\}$ est **binaire**, il faut supposer que

$$[Y|X_{1:p}] \sim \mathscr{B}\big(\mu(X_{1:p})\big).$$

(ii) si $Y \in \mathbb{N}$ est un **comptage**, on va supposer que

$$[Y|X_{1:p}] \sim \mathscr{P}\big(\mu(X_{1:p})\big).$$

Dans les deux cas,

$$\mathbb{E}(Y|X_{1:p}) = \mu(X_{1:p})$$

mais on ne va pas supposer que cette espérance conditionnelle est de la forme $\beta_0 + \sum_j \beta_j X_j$ car il y a des **contraintes** :

(i) $0 \leq \mu(X_{1:p}) \leq 1$ dans le cas de la loi de Bernoulli ou

(ii) $\mu(X_{1:p}) \geq 0$ dans le cas de la loi de Poisson.

6.1 Régression logistique

On s'intéresse ici au premier cas, où $Y \in \{0,1\}$ est **binaire**, et on suppose que

$$[Y|X_{1:p}] \sim \mathscr{B}\big(\mu(X_{1:p})\big).$$

Dans ce cas, $\mu(X_{1:p}) = \mathbb{E}(Y|X_{1:p}) = \mathbb{P}(Y=1|X_{1:p}) \in [0;1]$.

La fonction logistique est définie, pour tout $x \in]0;1[$ par

$$\begin{aligned} \text{logit} : \,]0;1[&\to \mathbb{R} \\ x &\mapsto \log\Big(\frac{x}{1-x}\Big). \end{aligned}$$

C'est une bijection strictement croissante de $]0;1[$ sur $\mathbb{R}$, infiniment dérivable. Et la fonction inverse est donnée par

$$\begin{aligned} \text{logit}^{-1} : \, \mathbb{R} &\to]0;1[\\ x &\mapsto \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} \end{aligned}$$

qui est infiniment dérivable. Voici le graphe de cette fonction.

L'**hypothèse linéaire** devient maintenant

$$\text{logit}\big(\mu(X_{1:p})\big) = \beta_0 + \sum_j \beta_j X_j.$$

Autrement dit,

$$\begin{aligned} \mathbb{E}(Y|X_{1:p}) = \mathbb{P}(Y=1|X_{1:p}) &= \text{logit}^{-1}\Big(\beta_0 + \sum_j \beta_j X_j\Big) \\ &= \frac{1}{1+\exp\Big(-\beta_0 - \sum_j \beta_j X_j\Big)} \in]0;1[. \end{aligned}$$

Le **rapport de cote** (ou *odds ratio*) vaut alors

$$\frac{\mathbb{P}(Y=1|X_{1:p})}{\mathbb{P}(Y=0|X_{1:p})} = \exp(\beta_0) \prod_{j=1}^{p} \exp(\beta_j X_j).$$

Ce qui revient à supposer un **effet multiplicatif** des covariables sur le rapport de cote. En effet, toute autre covariable étant fixée, remplacer X_j par $X_j + 1$ revient à multipier le rapport de cote par $\exp(\beta_j)$.

Avec des données $\mathbf{Y} \in \{0,1\}^n$ et $\mathbf{X} \in \mathbb{R}^{n \times p}$, on ajuste le modèle par **maximum de vraisemblance**. Ce maximum est n'est pas explicite, il faut utiliser un algorithme d'optimisation numérique pour le trouver (Newton-Raphson). Ici, la vraisemblance est

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}, \quad \text{où } p_i = \frac{1}{1+\exp(-\beta_0 - \sum_j \beta_j X_{ij})}.$$

On note $\widehat{\boldsymbol{\beta}}$ l'estimateur du maximum de vraisemblance.

Pour un **nouvel individu** de covariables $\widetilde{X}_1, \dots, \widetilde{X}_p$ connues et de réponse $\widetilde{Y}$ inconnue, on peut **inférer/prédire** deux quantités différentes :
— la probabilité de la réponse 1, ou la réponse moyenne, à savoir $\mu(\widetilde{X}_{1:p}) = \mathbb{P}\Big(\widetilde{Y} = 1 \Big| \widetilde{X}_{1:p}\Big)$, avec

$$\widehat{\mu}(\widetilde{X}_{1:p}) = \frac{1}{1+\exp\Big(-\widehat{\beta}_0 - \sum_j \widehat{\beta}_j \widetilde{X}_j\Big)}$$

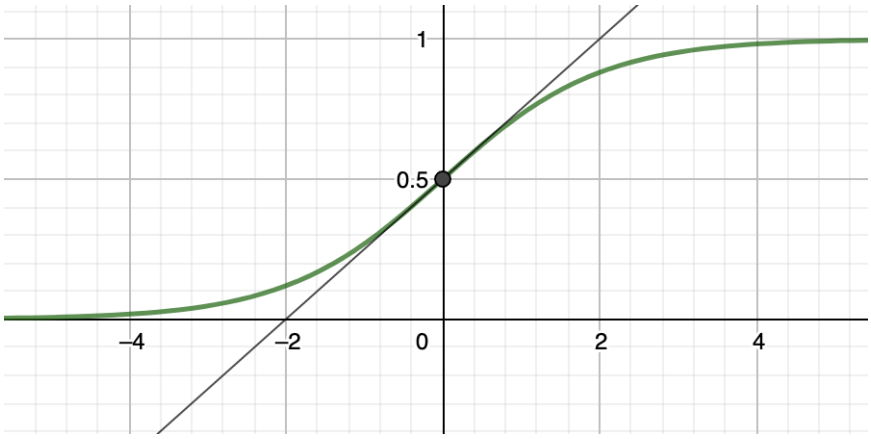

Figure 6.1 – *Graphe de la fonction logit^{-1}, et sa tangente en $x = 0$*

— la réponse $\widetilde{Y}$ elle-même, avec

$$\widehat{Y} = \begin{cases} 1 & \text{si } \mathbb{P}\left(\widetilde{Y} = 1 \middle| \widetilde{X}_{1:p}\right) > 0.5 \\ 0 & \text{sinon} \end{cases}$$

6.2 Régression de Poisson

Rappel. Si $Z \sim \mathscr{P}(a)$, alors,

$$\forall\ k \in \mathbb{N}, \quad \mathbb{P}(Z = k) = e^{-a}\frac{a^k}{k!} \quad \text{et} \quad \mathbb{E}(Z) = a.$$

6.2.1 Exposition constante

On s'intéresse ici au second cas, où $Y \in \mathbb{N}$ est un **comptage**, et on suppose que

$$[Y|X_{1:p}] \sim \mathscr{P}\big(\mu(X_{1:p})\big).$$

Dans ce cas, $\mu(X_{1:p}) = \mathbb{E}(Y|X_{1:p}) \in [0; +\infty[$.

Dans la régression de Poisson, la fonction **logarithme** va jouer le même rôle que la fonction logistique et on suppose que

$$\log \mu(X_{1:p}) = \log \mathbb{E}(Y|X_{1:p}) = \beta_0 + \sum_j \beta_j X_j.$$

Avec des données $\mathbf{Y} \in \mathbb{N}^n$ et $\mathbf{X} \in \mathbb{R}^{n\times p}$, on ajuste le modèle par **maximum de vraisemblance**. Ce maximum est n'est pas explicite, il faut utiliser un algorithme d'optimisation numérique pour le trouver (Newton-Raphson). Ici, la vraisemblance est

$$L(\beta) = \prod_{i=1}^{n} e^{-a_i}\frac{a_i^{y_i}}{y_i!}, \quad \text{où } a_i = \exp(\beta_0 + \sum_j \beta_j X_{ij}).$$

On note $\widehat{\boldsymbol{\beta}}$ l'estimateur du maximum de vraisemblance.

Pour un **nouvel individu** de covariables $\widetilde{X}_1, \ldots, \widetilde{X}_p$ connues et de réponse $\widetilde{Y}$ inconnue, on peut **inférer/prédire** deux quantités différentes :

— la réponse moyenne, à savoir $\mu(\widetilde{X}_{1:p}) = \mathbb{E}\left(\widetilde{Y}\middle|\widetilde{X}_{1:p}\right)$, avec

$$\widehat{\mu}(\widetilde{X}_{1:p}) = \exp\left(\widehat{\beta}_0 + \sum_j \widehat{\beta}_j \widetilde{X}_j\right)$$

— la réponse $\widetilde{Y}$ elle-même, avec

$$\widehat{Y} = G\left(\widehat{\mu}(\widetilde{X}_{1:p})\right)$$

où, pour tout $a > 0$,

$$G(a) = \operatorname*{argmax}_{k\in\mathbb{N}} e^{-a}\frac{a^k}{k!}.$$

Notons ici que $G(a)$ est la modalité la plus probable de la loi $\mathscr{P}(a)$.

6.2.2 Avec durée d'exposition variable

Une variable aléatoire de Poisson Z peut être un comptage **pendant une durée donnée** $D > 0$. (Exemple : nombre d'accident pendant une certaine durée). On suppose alors que

$\mathbb{E}(Z|D) = aD$, où a est une nombre moyen par unité de temps.

Donc,

$$\log \mathbb{E}(Z|D) = \log a + \log D.$$

Dans le modèle de Poisson avec durée d'exposition variable, on va donc supposer que

$$\log \mathbb{E}(Y|D, X_{1:p}) = \beta_0 + \sum_j \beta_j X_j + \log D.$$

Alors,

— $E = \log D$ intervient dans la formule comme une covariable, mais le coefficient en facteur est connu, égal à 1,

— $\exp\left(\beta_0 + \sum_j \beta_j X_j\right)$ est le nombre moyen par unité de temps (à covariables fixées).

L'ajustement se fait également par maximum de vraisemblance. La précision des estimateurs n'est pas liés au nombres d'observations, mais à la durée totale d'exposition

$$n_D = D_1 + D_2 + \ldots + D_n.$$

6.3 Exemples

6.3.1 Régression logistique simple

On s'intéresse à l'élection présidentielle américaine en 1992 qui opposait Georges W. Bush et Bill Clinton (vainqueur). S'ajoute un troisième candidat sans étiquette Ross Perot. Pour étudier le vote d'un électeur, on pose

$$Y = \begin{cases} 1 & \text{si vote en faveur de Bush} \\ 0 & \text{sinon} \end{cases}$$

On dispose d'une variable explicative $X \in \{1,2,\ldots,5\}$ qui indique le niveau de revenu de l'électeur. (Pauvre $\iff X = 1$). Données : 14031 votants. Voir tableau 6.1.

Si on ajuste un modèle de régression logistique sur ces données, on obtient

$$\text{logit}\Big(\mathbb{P}(Y=1|X)\Big) \approx -0.67 + 0.23X.$$

On peut représenter l'ajustement avec la figure 6.2.

Comment **s'interprète** 0.23 ? Quand on remplace income $= x$ par income $= x+1$, le **rapport de cote** est multiplié par

$$\exp(0.23) \approx 1.25.$$

Et les prédictions sont

6.3.2 Nombre d'accidents de la route

On souhaite prédire le nombre d'accidents de la route aux carrefours par année en fonction de deux covariables :

- X_1, vitesse moyenne en km/h des voitures sur les routes autour du carrefour,
- X_2, présence de feu tricolore au carrefour.

Avec des données, on a ajusté le modèle

$$\log\mathbb{E}\Big(Y\Big|X_1,X_2,D\Big) \approx 2.8 + 0.02X_1 - 0.20X_2 + \log D.$$

Interprétations :

- Si on augmente la vitesse moyenne de 10 km/h sans changer X_2, le nombre moyen d'accidents par années est multiplié par $\exp(0.02\times 10) \approx 1.22$. D'où une augmentation du nombre moyen d'environ 22%.
- Si on ajoute un feu tricolore sans changer X_1, le nombre moyen d'accidents par années est multiplié par $\exp(-0.2) \approx 0.82\%$. D'où une diminution de 18% du nombre moyen d'accidents.

Table 6.1 – *Fréquence des votes par classes de revenus ($Y = 1$ si vote Bush)*

income	1	2	3	4	5
$\%(Y = 1)$	41	44	49	54	70
$\%(Y = 0)$	59	56	51	46	30

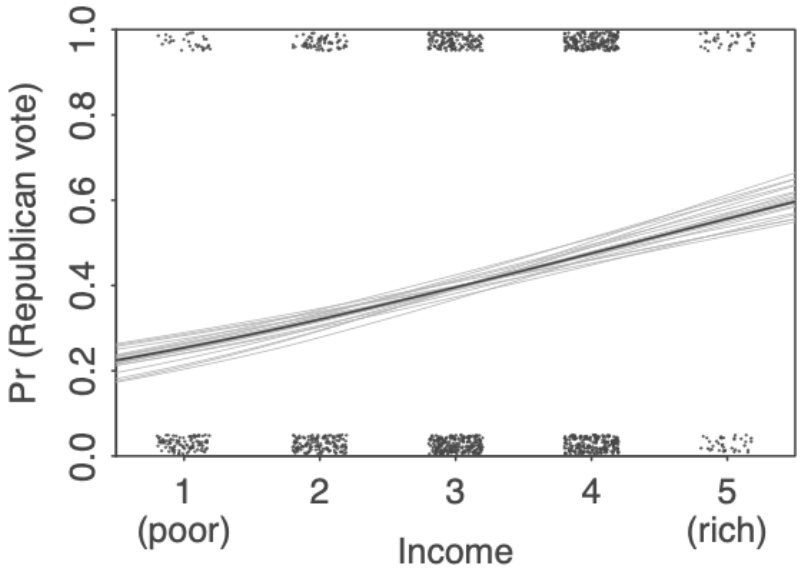

Figure 6.2 – Ajustement sur l'élection de 1992