Chapitre II

Le modèle linéaire gaussien

1 Le modèle linéaire général

Il existe beaucoup de phénomènes pour lesquels une certaine quantité d'intérêt y est supposée être une fonction d'un certain nombre d'autres quantités $x_1, \ldots, x_p$ et, dans de nombreuses situations, on peut penser que la liaison entre y et les x_i est linéaire, c'est à dire que $y = \sum_{j=1}^{p} \beta_j x_j$ pour des paramètres β_j fixes, c'est à dire indépendants des x_j. Mais ceci n'est qu'une approximation de la réalité et lorsque l'on fait une observation du phénomène, on obtient en fait un résultat aléatoire suite aux erreurs de mesure et aussi de modélisation, de sorte que l'on arrive à une relation stochastique de la forme $Y = \sum_{j=1}^{p} \beta_j x_j + \varepsilon$ où ε est une variable aléatoire que l'on suppose toujours centrée pour que les paramètres soient bien définis par $\mathbb{E}[Y] = \sum_{j=1}^{p} \beta_j x_j$. Si ε n'était pas centrée mais d'espérance $\beta_0 \neq 0$, il suffirait de rajouter la variable $x_0 \equiv 1$ au modèle afin d'écrire $Y = \sum_{j=0}^{p} \beta_j x_j + \varepsilon'$ avec $\varepsilon' = \varepsilon - \beta_0$ centrée.

1.1 Quelques exemples

Exemple 1 Supposons que l'on veuille étudier la résistance d'un alliage métallique en fonction de sa température de cuisson. On peut pour cela fabriquer un certain nombre d'échantillons de cet alliage à des températures respectives $t_1, \ldots, t_n$, pas nécessairement toutes distinctes, et mesurer pour chaque échantillon sa résistance $R_i, 1 \leq i \leq n$. On suppose qu'un modèle mathématique pour la dépendance $R(t)$ entre la résistance et la température peut être donné par un polynôme de degré trois (modèle approché validé, par exemple, par des essais antérieurs) d'où le modèle statistique correspondant :

$$R_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \varepsilon_i, \quad 1 \leq i \leq n, \tag{II.1.1}$$

où les β_j sont les coefficients (inconnus, à estimer) du polynôme et les ε_i des v.a. réelles représentant l'ensemble des erreurs liées au processus de fabrication. On suppose que ces erreurs sont i.i.d. (indépendantes de t en particulier), centrées (sinon il suffit de modifier le coefficient β_0 pour se ramener à ce cas) et de variance σ^2 inconnue. Le problème posé est celui de l'estimation des paramètres inconnus β_j en fonction des données (les valeurs des t_i et des R_i).

Exemple 2 On veut bâtir un modèle économétrique permettant d'évaluer (approximativement) la consommation des ménages en fonction d'un ensemble de variables économiques (salaire moyen, pression fiscale, charges sociales, encaisses des Caisses d'Épargne,

taux d'intérêts, ...) dont on pense qu'elles peuvent avoir une influence sur cette consommation. On note X^j ces diverses variables économiques et Y la consommation. On fait l'hypothèse d'une liaison linéaire entre Y et les X^j (ne pas oublier que les X^j peuvent comprendre des transformées des variables de base : puissances, logarithmes, etc.), ce qui conduit au modèle économétrique (approché, comme toujours) suivant :

$$Y = \beta_0 + \sum_{j=1}^{p-1} \beta_j X^j + \varepsilon,$$

avec des coefficients inconnus β_j et un terme d'erreur centré ε qui prend en compte les inévitables fluctuations aléatoires. Si l'on observe ces variables à des temps successifs $t_i, 1 \leq i \leq n$, on obtient pour Y des valeurs Y_i et pour les X^j des valeurs X_i^j liées approximativement par les relations

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_i^j + \varepsilon_i, \quad 1 \leq i \leq n, \tag{II.1.2}$$

avec des erreurs ε_i que l'on supposera ici, pour faire simple, i.i.d. On souhaite, bien évidemment, estimer les coefficients β_j, mais aussi détecter ceux qui sont très petits, c'est-à-dire correspondent à des variables économiques X^j qui influent très peu, voire pas du tout, sur la consommation des ménages.

Exemple 3 On souhaite maintenant évaluer l'influence de divers traitements sur la culture des tomates. Pour cela, on va sélectionner un certain nombre de plants de tomates, en principe identiques et les répartir en J groupes de tailles respectives K_j, $1 \leq j \leq J$. Chaque groupe subit un traitement différent et en fin de saison on mesure pour chaque plant de numéro (j,k), $1 \leq j \leq J$, $1 \leq k \leq K_j$, le poids $Z_{j,k}$ de tomates récolté. Le modèle statistique (simplifié) correspondant s'écrit

$$Z_{j,k} = \mu_j + \xi_{j,k}, \quad 1 \leq j \leq J, \ \ 1 \leq k \leq K_j,$$

où les μ_j représentent l'effet du traitement j et les $\xi_{j,k}$ les fluctuations aléatoires inévitables, toujours supposées i.i.d. et centrées. Ici on s'intéresse de nouveau à l'estimation des coefficients μ_j, mais surtout à détecter lesquels sont les plus grands, correspondant aux traitements les plus efficaces en terme de rendement.

1.2 Le modèle statistique

Comme indiqué ci-dessus, dans chacun des exemples précédents, on dispose d'observations Y_i qui correspondent à des fonctions linéaires de variables (pas au sens aléatoire mais au sens physique ou économique) X^j, X_i^j étant la valeur (numérique) que prend la variable X^j dans l'expérience qui produit Y_i. Ces expériences comportent des effets aléatoires qui se traduisent par des erreurs ε_i que l'on supposera ici i.i.d., centrées et de variance σ^2. L'observation Y_i est alors la somme de la fonction linéaire des X_i^j (déterministe et connue) et de l'erreur (aléatoire et inconnue) ε_i. Ceci correspond exactement au système de n équations données par (II.1.2), lequel peut être récrit sous forme matricielle $Y = X\beta + \varepsilon$, en posant $X_i^0 = 1$ et en notant X la matrice de coefficients X_i^j. On a donc, en désignant par A^t la transposée de la matrice A, comme nous le ferons systématiquement dans la suite,

$$\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^t \in \mathbb{R}^p, \quad Y = (Y_1, \ldots, Y_n)^t \quad \text{et} \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t \in \mathbb{R}^n.$$

Ce modèle est dit *modèle linéaire* car il l'est par rapport aux paramètres inconnus β_j. Par contre, la dépendance par rapport aux quantités connues n'a aucune raison d'être linéaire comme le montre la relation (II.1.1) avec une dépendance polynomiale par rapport aux t_i.

1.2.1 Rappels sur les vecteurs aléatoires

Un vecteur aléatoire $Y = (Y_1, \ldots, Y_n)^t$ dans $\mathbb{R}^n$ est un vecteur dont les composantes Y_i sont aléatoires. La loi du vecteur Y est la loi jointe des Y_i, son espérance $\mathbb{E}[Y]$ est le vecteur des espérances, sa matrice de covariance $\mathrm{Cov}(Y)$ est la matrice carrée $n \times n$ symétrique d'éléments $\mathrm{Cov}(Y_i, Y_j)$, donc diagonalisable dans une base orthonormée. On a aussi matriciellement

$$\mathrm{Cov}(Y) = \mathbb{E}\left[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^t\right].$$

Cette matrice est positive et définie positive si et seulement si les coordonnées de Y ne sont pas affinement liées p.s. Formellement, $\mathrm{Cov}(Y)$ est inversible si et seulement si

$$\sum_{i=1}^{n} a_i Y_i + b = 0 \quad \text{p.s.} \qquad \Longrightarrow \qquad a_i = 0 = b, \quad \forall i.$$

Si A est une matrice $p \times n$ et $B \in \mathbb{R}^p$, alors $Z = AY + B$ est un vecteur aléatoire dans $\mathbb{R}^p$ tel que

$$\mathbb{E}[Z] = A\mathbb{E}[Y] + B; \qquad \mathrm{Cov}(Z) = A\,\mathrm{Cov}(Y)A^t. \tag{II.1.3}$$

1.2.2 Formulation matricielle

Description de l'expérience statistique : on observe un vecteur aléatoire $Y = X\beta + \varepsilon \in \mathbb{R}^n$ et l'on connaît la matrice X de taille $n \times p$ avec $p < n$. Le vecteur aléatoire ε n'est pas observé mais on fait des hypothèses sur sa loi : ses composantes ε_i sont i.i.d. de moyenne 0 et variance σ^2 inconnue. Comme $Y = X\beta + \varepsilon$, la loi de Y est celle de ε translatée d'un vecteur déterministe mais inconnu $\mu = X\beta$ et $\mu = \mathbb{E}[Y]$. La loi de Y est entièrement déterminée par la loi de ε et le vecteur μ ou le paramètre β (car $\mu = X\beta$), ce qui définit l'espace des paramètres. Si l'on note X^j, $1 \leq j \leq p$, les colonnes de X, ce sont des vecteurs de $\mathbb{R}^n$ et par définition μ appartient à l'espace vectoriel V engendré par les colonnes de X et peut être n'importe quoi dans V puisque β est arbitraire dans $\mathbb{R}^p$. A partir de là deux cas sont possibles :

i) les colonnes de X sont linéairement indépendantes donc X est de rang p et V de dimension p. Alors l'application $\beta \mapsto \mu$ est injective et deux valeurs distinctes de β correspondent à deux lois distinctes pour Y (leurs espérances sont différentes).

ii) la matrice X est de rang $l < p$ et V de dimension l. Alors l'application $\beta \mapsto \mu$ n'est pas injective et deux valeurs distinctes de β peuvent correspondre à une même valeur de μ donc à une même loi pour Y. Ceci pose un sérieux problème si l'on s'intéresse à β puisque il n'est pas défini de manière unique, contrairement à μ. On pourra chercher à estimer μ mais pas β : cela n'a pas de sens. On dira que le modèle paramétré par β n'est *pas identifiable.*

Définition 7 *Un modèle statistique* $\{P_\theta, \theta \in \Theta\}$ *est dit* identifiable *si l'application* $\theta \mapsto P_\theta$ *est injective.*

Si un modèle n'est pas identifiable, on ne peut espérer estimer le paramètre correctement puisque même si l'on connaissait la vraie loi P_θ cela ne suffirait pas à déterminer θ lorsque $P_\theta = P_{\theta'}$. Il y a une vraie loi mais pas de vrai paramètre. On évitera donc d'utiliser des modèles non identifiables.

Il faut garder à l'esprit que ce qui a un sens, c'est c'est le modèle statistique, c'est à dire l'ensemble des lois possibles pour l'observation, pas la manière dont on les a paramétrées. Le paramétrage est souvent assez largement arbitraire et il doit être bijectif pour être utilisable. Le lemme suivant donne une condition nécessaire et suffisante d'identifiabilité du modèle linéaire :

Lemme 1 *Soit X une matrice de taille $n \times p$ avec $p \leq n$. Alors X^tX est inversible si et seulement si X est de rang p.*

1.3 Les estimateurs des moindres carrés

Supposons, comme dans notre exemple initial, que l'on observe, avec des fluctuations ou des erreurs aléatoires, un phénomène qui est une fonction polynomiale du temps, de degré $\leq k$ connu, à des instants successifs $t_1, \ldots, t_n$. On obtient ainsi une suite d'observations $Y_i = P(t_i) + \varepsilon_i$ où P désigne le polynôme (inconnu) de degré $\leq k$ qui donne la forme du phénomène. Pour reconstituter P, on peut chercher à ajuster une courbe polynomiale Q de degré k aux données (t_i, Y_i) de sorte que les $Q(t_i)$ soient proches des Y_i. Une méthode classique, pour ce faire, est la méthode dite des *moindres carrés* qui consiste à minimiser, parmi tous les polynômes de degré $\leq k$, la quantité $\sum_{i=1}^n [Q(t_i) - Y_i]^2$. On préfère minimiser ceci plutôt que la somme des valeurs absolues $\sum_{i=1}^n |Q(t_i) - Y_i|$, d'une part parce que cela pénalise davantage les grandes erreurs, mais surtout parce que c'est mathématiquement et numériquement un problème beaucoup plus facile à traiter.

En effet, géométriquement, $Y = (Y_1, \ldots, Y_n)^t$ est un vecteur dans $\mathbb{R}^n$ et si $Q(t) = \sum_{j=0}^k b_j t^j$, le vecteur $(Q(t_i))_{1 \leq i \leq n}$ peut s'écrire matriciellement Xb avec une matrice X d'éléments $X_i^j = t_i^j$. La résolution du problème des moindres carrés se ramène donc à la minimisation de $\|Y - Xb\|$ par rapport à $b \in \mathbb{R}^{k+1}$.

De manière plus générale, étant donné un modèle linéaire de la forme $Y = X\beta + \varepsilon$ avec $\beta \in \mathbb{R}^p$, on peut chercher à estimer l'espérance $X\beta$ de Y en minimisant $\|Y - Xb\|$ par rapport à b. Or les Xb engendrent un espace vectoriel V de dimension $\leq p$ (de dimension p si le modèle est identifiable) qui est l'espace engendré dans $\mathbb{R}^n$ par les colonnes de X et la solution du problème de minimisation de la distance de Y à V est bien connue, c'est la projection orthogonale de $Y = X\beta + \varepsilon$ sur V, laquelle s'écrit, puisque $X\beta \in V$, $X\beta + e$, où e est la projection orthogonale de ε sur V. On fera donc une erreur e par rapport au vrai vecteur $X\beta$, mais si p est petit devant n et si le vecteur ε des erreurs est bien réparti dans toutes les directions, on peut penser que la taille de l'erreur finale $\|e\|$ sera bien inférieure à celle de l'erreur initiale $\|\varepsilon\|$.

Pour résoudre le problème des moindres carrés, c'est-à-dire pour minimiser par rapport à $b \in \mathbb{R}^p$ la quantité $\|Y - Xb\|^2$, il suffit de trouver un $\widehat{\beta} \in \mathbb{R}^p$ tel que $Y - X\widehat{\beta}$ soit orthogonal à V, c'est-à-dire à tous les Xb, $b \in \mathbb{R}^p$. Ceci équivaut à dire que tous les produits scalaires $\langle Xb, Y - X\widehat{\beta} \rangle$ sont nuls, c'est-à-dire que

$$X^t(Y - X\widehat{\beta}) = X^tY - X^tX\widehat{\beta} = 0 \qquad \text{ou} \qquad X^tX\widehat{\beta} = X^tY.$$

Cette équation a toujours au moins une solution, puisque la projection existe, mais elle n'est pas forcément unique. Elle l'est si et seulement si la matrice X^tX est inversible, c'est-à-dire si X est de rang p ou si les colonnes de X forment une base de V. En effet,

dans ce cas la projection de Y sur V a une représentation $X\widehat{\beta}$ unique, ce qui n'est pas le cas si les colonnes de X ne sont pas linéairement indépendantes. Mais dans ce cas, comme on l'a vu, la représentation de $\mathbb{E}[Y] = X\beta$ n'est pas unique non plus et le modèle n'est pas identifiable. On excluera dorénavant cette situation en supposant toujours que la matrice X est de plein rang p. L'estimateur des moindres carrés de β s'écrit alors

$$\widehat{\beta} = (X^t X)^{-1} X^t Y$$

et c'est une fonction linéaire des observations $Y_1, \ldots, Y_n$. On remarquera que

$$X\widehat{\beta} = X(X^t X)^{-1} X^t Y$$

est la projection orthogonale de Y sur V, ce qui implique que $X(X^t X)^{-1} X^t$ est l'opérateur de projection sur V. Il s'ensuit que l'opérateur de projection sur $V^\perp$ est $I_n - X(X^t X)^{-1} X^t$.

1.4 Propriétés des estimateurs des moindres carrés

Les formules (II.1.3) impliquent que

$$\mathbb{E}_\beta[\widehat{\beta}] = (X^t X)^{-1} X^t \mathbb{E}[Y] = \beta; \quad \mathrm{Cov}_\beta(\widehat{\beta}) = (X^t X)^{-1} X^t \sigma^2 I_n X (X^t X)^{-1} = \sigma^2 (X^t X)^{-1}.$$

L'estimateur $\widehat{\beta}$ est donc sans biais. De plus le risque quadratique d'une coordonnée β_i de β s'écrit

$$\mathbb{E}_\beta\left[(\widehat{\beta}_i - \beta_i)^2\right] = \mathrm{Var}_\beta(\widehat{\beta}_i) = \sigma^2 \left((X^t X)^{-1}\right)_{i,i}.$$

Comme ce risque est proportionnel au paramètre inconnu σ^2, il est souhaitable d'estimer ce dernier. Pour ce faire on introduit la *somme des carrés résiduelle*

$$D^2(Y, \widehat{\beta}) = \|Y - X\widehat{\beta}\|^2 = \|\widehat{\varepsilon}\|^2,$$

qui est le carré de la norme de la partie de Y qui ne peut s'écrire comme fonction linéaire de X, $\widehat{\varepsilon}$ désignant la projection de ε (donc de Y) sur $V^\perp$. Comme

$$X\widehat{\beta} = X(X^t X)^{-1} X^t (X\beta + \varepsilon) = X\beta + X(X^t X)^{-1} X^t \varepsilon,$$

$Y - X\widehat{\beta} = A\varepsilon$ avec $A = I_n - X(X^t X)^{-1} X^t$ (opérateur de projection sur $V^\perp$), donc $A^t = A$ et $A^t A = A$, de sorte que,

$$\mathbb{E}\left[D^2(Y, \widehat{\beta})\right] = \mathbb{E}\left[\|A\varepsilon\|^2\right] = \mathbb{E}\left[\varepsilon^t A^t A \varepsilon\right] = \sum_{i,j} A_{i,j} \mathbb{E}[\varepsilon_i \varepsilon_j] = \sigma^2 \,\mathrm{Tr}(A),$$

puisque $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$. Les propriétés de la trace impliquent alors que

$$\begin{aligned} \mathrm{Tr}(A) &= \mathrm{Tr}\left(I_n - X(X^t X)^{-1} X^t\right) = \mathrm{Tr}(I_n) - \mathrm{Tr}\left(X(X^t X)^{-1} X^t\right) \\ &= \mathrm{Tr}(I_n) - \mathrm{Tr}\left((X^t X)^{-1} X^t X\right) = \mathrm{Tr}(I_n) - \mathrm{Tr}(I_p) = n - p. \end{aligned}$$

Il s'ensuit que $\widehat{s}_n^2 = (n-p)^{-1} D^2(Y, \widehat{\beta})$ est un estimateur sans biais de σ^2.

2 Le modèle linéaire gaussien

Si l'on veut aller plus loin, c'est-à-dire évaluer la variance ou la loi de l'estimateur $\widehat{s}_n^2$, il est nécessaire de faire sur le modèle des hypothèses supplémentaires. Le plus simple est de supposer que les erreurs ε_i sont indépendantes et gaussiennes, autrement dit que ε est un vecteur gaussien, ce qui colle assez bien à de nombreux problèmes pratiques (mais attention, pas tous !).

2.1 Rappels sur les vecteurs aléatoires gaussiens

La première remarque à faire est que, de même qu'une gaussienne de moyenne μ est la somme d'une gaussienne centrée et de μ, un vecteur gaussien général dans $\mathbb{R}^n$ d'espérance $\mu \in \mathbb{R}^n$ est la somme de μ et d'un vecteur gaussien centré, ce qui fait qu'en retranchant à un vecteur gaussien son espérance, on obtient un vecteur gaussien centré. Il suffit donc d'étudier les vecteurs gaussiens centrés, ce que nous allons faire tout d'abord.

Le vecteur gaussien standard X de $\mathbb{R}^n$ est tout simplement un vecteur aléatoire dont les n coordonnées sont i.i.d. $\mathcal{N}(0,1)$. Sa loi est notée $\mathcal{N}(0, I_n)$, I_n étant sa matrice de covariance.

Un vecteur gaussien général (centré) Y de $\mathbb{R}^n$ est une image linéaire d'un vecteur gaussien standard de $\mathbb{R}^p$: $Y = AX$, d'où $\Sigma = \mathrm{Cov}(Y) = AI_pA^t = AA^t$. Une condition équivalente est que toute combinaison linéaire des coordonnées de Y (de la forme a^tY pour $a \in \mathbb{R}^n$) est une variable gaussienne (de loi $\mathcal{N}(0, a^t\Sigma a)$). La loi de Y est alors entièrement déterminée par sa matrice de covariance Σ et notée $\mathcal{N}(0,\Sigma)$.

Si le rang de Σ est r ($\leq n$), on peut trouver A matrice $n \times r$ de rang r et $X \sim \mathcal{N}(0, I_r)$ tels que $Y = AX$ et $\Sigma = AA^t$. Il s'ensuit que, presque sûrement, Y appartient à un sous-espace de dimension r de $\mathbb{R}^n$ (l'image par A de $\mathbb{R}^r$). Si $r < n$, la loi de Y n'a pas de densité par rapport à la mesure de Lebesgue dans $\mathbb{R}^n$. Sinon ($r = n$), sa densité s'écrit

$$f(y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2}y^t\Sigma^{-1}y\right],$$

Si la matrice Σ est diagonale, d'éléments diagonaux σ_i^2, on peut factoriser f :

$$f(y) = \prod_{i=1}^{n} f_i(y_i) \quad \text{avec} \quad f_i(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{y_i^2}{2\sigma_i^2}\right],$$

ce qui implique que les coordonnées $Y_1, \ldots, Y_n$ de Y sont indépendantes de lois respectives $\mathcal{N}(0, \sigma_i^2)$. La réciproque est évidemment vraie.

Si X est un vecteur gaussien standard et Q une matrice orthogonale, QX est un vecteur gaussien standard : un vecteur gaussien standard conserve la même loi par changement de base orthonormée. Si $V_1, \ldots, V_k$ sont k sous-espaces vectoriels orthogonaux dans $\mathbb{R}^n$ et Z_j est la projection orthogonale de X sur V_j alors Z_j (considéré comme un vecteur aléatoire dans V_j) est un vecteur gaussien standard de même dimension que V_j et les Z_j sont indépendants.

Si Y est un vecteur gaussien de loi $\mathcal{N}(0,\Sigma)$ et $Z = Y + \mu$ alors la loi de Z est notée $\mathcal{N}(\mu,\Sigma)$.

Si X est un vecteur gaussien standard dans $\mathbb{R}^p$, sa norme euclidienne $\|X\|$ est donnée par $\|X\|^2 = \sum_{i=1}^p X_i^2$. Cette loi est dite loi du khi carré ou khi deux à p degrés de liberté, notée $\chi^2(p)$. Si Z a la loi $\chi^2(p)$ alors $\mathbb{E}[Z] = p$ et $\mathrm{Var}(Z) = 2p$. Cette loi est tabulée car sa fonction de répartition n'a pas d'expression explicite (comme celle de la loi normale). Les

tables sont des tables à double entrée avec en général d'un côté le nombre p de degrés de liberté et de l'autre des valeurs x comprises entre 0 et 1, telles que 0,001 ; 0,005 ; 0,01 ; 0,1 ; ... ; 0,9 ; ... A l'intersection de la ligne p et de la colonne x, on trouve la valeur t telle que $\mathbb{P}[\chi^2(p) \leq t] = x$ (ou, selon les tables, $\mathbb{P}[\chi^2(p) > t] = x$). Lorsque p est suffisamment grand (en pratique $p \geq 120$) on utilise l'approximation gaussienne du χ^2 pour faire les calculs en faisant comme si

$$\left(\sum_{i=1}^{p} X_i^2 - p\right) / \sqrt{2p} \sim \mathcal{N}(0;1).$$

Nous aurons besoin, dans la suite, des lois suivantes.

Définition 8 *Soit U, V, W trois variables aléatoires indépendantes de lois respectives $\mathcal{N}(0,1)$, $\chi^2(p)$ et $\chi^2(n)$. La loi de $U/\left(\sqrt{W/n}\right)$ est dite loi de Student à n degrés de liberté et notée $\mathcal{T}_n$. La loi de $(V/p)/(W/n)$ est dite loi de Fisher à p et n degrés de liberté et notée $\mathcal{F}_{p,n}$.*

Comme $W/n \to 1$ lorsque $n \to +\infty$, la loi de Student $\mathcal{T}_n$ est assimilée à une gaussienne standard lorsque n est grand (en pratique $n \geq 120$) et la loi de Fisher $\mathcal{F}_{p,n}$ à un $\chi^2(p)/p$ sous les mêmes conditions. On remarquera que si $Z \sim \mathcal{T}_n$, alors $Z^2 \sim \mathcal{F}_{1,n}$.

Théorème 3 (Cochran) *Si $V_1, \ldots, V_k$ sont k sous-espaces vectoriels orthogonaux dans $\mathbb{R}^n$ de dimensions respectives n_j et si X_j est la projection orthogonale d'un vecteur gaussien standard X sur V_j alors les X_j, $1 \leq j \leq k$ sont des vecteurs gaussiens indépendants et $\|X_j\|^2$ a une loi du chi-deux à n_j degrés de liberté.*

2.2 Propriétés du modèle linéaire gaussien

On part du modèle linéaire général $Y = X\beta + \varepsilon$ avec $\mathbb{E}[\varepsilon] = 0$ et $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$ et l'on ajoute l'hypothèse que ε est un vecteur gaussien, donc de loi $\mathcal{N}(0, \sigma^2 I_n)$. Autrement dit on peut écrire le modèle sous la forme

$$Y = X\beta + \sigma\xi = \mu + \sigma\xi \quad \text{avec } \beta \in \mathbb{R}^p, \quad \xi \sim \mathcal{N}(0, I_n).$$

On supposera que le modèle paramétré par (β, σ^2) est identifiable, c'est à dire que X est de rang p. On a donc un modèle paramétrique avec un paramètre de dimension $p+1$. La loi de Y est $\mathcal{N}(X\beta, \sigma^2 I_n)$.

2.2.1 Lois et propriétés des estimateurs

Sous l'hypothèse gaussienne, on peut calculer explicitement les lois des estimateurs des moindres carrés.

Proposition 3 *Les estimateurs des moindres carrés $\widehat{\beta}$ et $\widehat{s}_n^2 = (n-p)^{-1} D^2(Y, \widehat{\beta})$ de β et σ^2 sont indépendants et sans biais, $\widehat{\beta}$ suit une loi $\mathcal{N}(\beta, \sigma^2 (X^t X)^{-1})$ et $\sigma^{-2} D^2(Y, \widehat{\beta})$ une loi $\chi^2(n-p)$.*

Démonstration : Soit V le sous-espace de dimension p engendré par les colonnes de X et $V^\perp$ son orthogonal de dimension $n-p$. Le vecteur gaussien standard $\xi = \varepsilon/\sigma$ se décompose sur ces deux sous-espaces en $e/\sigma \in V$ et $e^\perp/\sigma \in V^\perp$, lesquels sont deux vecteurs gaussiens indépendants de dimensions respectives p et $n-p$ d'après le Théorème de Cochran. Qui

plus est $\sigma^{-2}D^2(Y,\widehat{\beta}) = \|e^{\perp}/\sigma\|^2$ est un $\chi^2(n-p)$. Par ailleurs, $X\widehat{\beta} = X\beta + e$, de sorte que $\widehat{\beta} = \beta + (X^tX)^{-1}X^te$ est un vecteur gaussien comme transformée affine de e/σ. On connaît déjà son espérance et sa covariance, d'où le résultat. □

Les premières conséquences de ce résultat sont les suivantes.

— La matrice X^tX peut se diagonaliser dans une base orthonormée sous la forme $Q\Delta^2Q^t$ où Δ est une matrice diagonale à coefficients diagonaux > 0 (racines des valeurs propres de X^tX) et $Q^t = Q^{-1}$. Il s'ensuit que $(X^tX)^{-1} = Q\Delta^{-2}Q^t$ et que $\sigma^{-1}\Delta Q^t(\widehat{\beta} - \beta)$ a une loi gaussienne standard dans $\mathbb{R}^p$. Donc $\sigma^{-2}\left\|\Delta Q^t(\widehat{\beta}-\beta)\right\|^2$ suit une loi $\chi^2(p)$ indépendante de celle de $(n-p)\widehat{s}_n^2/\sigma^2$ qui est un $\chi^2(n-p)$ et leur quotient correctement normalisé suit une loi de Fisher. Plus précisément :

$$\frac{\left\|\Delta Q^t(\widehat{\beta}-\beta)\right\|^2}{p\widehat{s}_n^2} \sim \mathcal{F}_{p,n-p}.$$

Or

$$\left\|\Delta Q^t(\widehat{\beta}-\beta)\right\|^2 = (\widehat{\beta}-\beta)^tQ\Delta^2Q^t(\widehat{\beta}-\beta) = (\widehat{\beta}-\beta)^tX^tX(\widehat{\beta}-\beta) = \left\|X(\widehat{\beta}-\beta)\right\|^2.$$

On peut alors obtenir ainsi des domaines de confiance de niveau $1-\alpha$ pour β de la manière suivante. On fixe t_α tel que $\mathbb{P}\left[\mathcal{F}_{p,n-p} \le t_\alpha\right] = 1-\alpha$. Alors avec une probabilité $1-\alpha$, $\left\|X(\widehat{\beta}-\beta)\right\|^2 / \left(p\widehat{s}_n^2\right) \le t_\alpha$ et

$$\left\{\beta \in \mathbb{R}^p \;\middle|\; \left\|X(\widehat{\beta}-\beta)\right\| \le \widehat{s}_n\sqrt{pt_\alpha}\right\}$$

est un domaine de confiance de niveau $1-\alpha$.

— Comme $(n-p)\widehat{s}_n^2/\sigma^2$ a une loi $\chi^2(n-p)$ d'espérance $n-p$ et variance $2(n-p)$,

$$\mathbb{E}\left[\widehat{s}_n^2\right] = \sigma^2 \qquad \text{et} \qquad \mathrm{Var}\left(\widehat{s}_n^2\right) = \frac{2\sigma^4}{(n-p)^2}.$$

Donc si l'on considère une suite de tels modèles linéaires avec p, β et σ^2 fixés, X variant avec n qui tend vers l'infini, les estimateurs des moindres carrés $\widehat{s}_n^2$ de σ^2 sont convergents car leur risque quadratique tend vers zéro.

2.2.2 Intervalles de confiance et tests associés

On peut construire des intervalles de confiance pour σ^2 à partir de la loi $\chi^2(n-p)$ de $\sigma^{-2}D^2(Y,\widehat{\beta}) = (n-p)(\widehat{s}_n/\sigma)^2$. Il suffit pour cela de choisir un intervalle $[a,b]$ tel que $\mathbb{P}[\chi^2(n-p) \in [a,b]] = 1-\alpha$ pour déduire que

$$\mathbb{P}\left[a \le (n-p)(\widehat{s}_n/\sigma)^2 \le b\right] = \mathbb{P}\left[\frac{(n-p)\widehat{s}_n^2}{b} \le \sigma^2 \le \frac{(n-p)\widehat{s}_n^2}{a}\right] = 1-\alpha. \qquad \text{(II.2.1)}$$

On peut choisir ainsi des bornes inférieures (respectivement supérieures) de confiance si $a = 0$ et $\mathbb{P}[\chi^2(n-p) \le b] = 1-\alpha$ (respectivement $b = +\infty$ et $\mathbb{P}[\chi^2(n-p) \ge a] = 1-\alpha$), avec un sens évident pour $1/a$ et $1/b$, ou fixer $\mathbb{P}[\chi^2(n-p) < a] = \mathbb{P}[\chi^2(n-p) > b] = \alpha/2$ pour obtenir un intervalle de confiance bilatère.

Les intervalles de confiance précédents permettent de construire des tests d'hypothèses sur σ. Supposons par exemple que l'on veuille tester que $\sigma^2 \le \sigma_0^2$, c'est à dire $\sigma^2 \in$

$]0;\sigma_0^2]$ au niveau α, on partira de l'intervalle de confiance unilatère de sens opposé $[b^{-1}(n-p)\widehat{s}_n^2, +\infty[$ fourni par (II.2.1) avec $\mathbb{P}[\chi^2(n-p) > b] = \alpha$ et l'on rejettera l'hypothèse si $b^{-1}(n-p)\widehat{s}_n^2 > \sigma_0^2$. On peut aussi, pour une valeur donnée de $\widehat{s}_n^2$ chercher le niveau de signification α_0 du test. On voit que la valeur critique de b pour le passage du rejet à l'acceptation est $b_0 = (n-p)\widehat{s}_n^2/\sigma_0^2$, laquelle valeur b_0 correspond à un niveau $\alpha_0 = \mathbb{P}[\chi^2(n-p) > b_0]$. Ceci fournit la p-value du test.

On peut aussi s'intéresser à trouver des intervalles de confiance pour un des paramètres, par exemple β_1, ou plus généralement pour une forme linéaire en β, $\langle a, \beta\rangle = a^t\beta = \sum_{j=1}^p a_j\beta_j$ avec $a \in \mathbb{R}^p$, ce qui inclut par exemple $\beta_2 - \beta_3$. Dans ce cas, $a^t\beta$ et $a^t\widehat{\beta}$ sont réels et, d'après la Proposition 3, $\sigma^{-1}(a^t\widehat{\beta} - a^t\beta)$ suit une loi $\mathcal{N}(0, \delta_a^2)$ avec $\delta_a^2 = a^t(X^tX)^{-1}a$, indépendante de celle de $\sigma^{-2}D^2(Y, \widehat{\beta})$. Il s'ensuit, d'après la Définition 8 (de la loi de Student), que

$$\frac{(\delta_a\sigma)^{-1}(a^t\widehat{\beta} - a^t\beta)}{\sqrt{\sigma^{-2}D^2(Y,\widehat{\beta})/(n-p)}} = \frac{a^t\widehat{\beta} - a^t\beta}{\widehat{s}_n\delta_a} \sim \mathcal{T}(n-p).$$

Ceci permet de construire aisément des intervalles de confiance pour $a^t\beta$. Soit t_α tel que $\mathbb{P}[\mathcal{T}(n-p) > t_\alpha] = \alpha$. Par symétrie de la loi de Student, $\mathbb{P}[\mathcal{T}(n-p) < -t_\alpha] = \alpha$ et $\mathbb{P}[|\mathcal{T}(n-p)| \leq t_{\alpha/2}] = 1-\alpha$. On en déduit alors des intervalles de confiance de niveau $1-\alpha$ pour $a^t\beta$, unilatères ou bilatères, de la forme

$$[a^t\widehat{\beta} - t_\alpha\widehat{s}_n\delta_a; +\infty[\,, \quad]-\infty; a^t\widehat{\beta} + t_\alpha\widehat{s}_n\delta_a] \quad \text{et} \quad [a^t\widehat{\beta} - t_{\alpha/2}\widehat{s}_n\delta_a; a^t\widehat{\beta} + t_{\alpha/2}\widehat{s}_n\delta_a]\,.$$

En particulier, si δ_j^2 est le j-ième élément diagonal de $(X^tX)^{-1}$, $[\widehat{\beta}_j - t_{\alpha/2}\widehat{s}_n\delta_j; \widehat{\beta}_j + t_{\alpha/2}\widehat{s}_n\delta_j]$ est un intervalle de confiance bilatère de niveau $1-\alpha$ pour β_j.

Ces intervalles de confiance permettent de construire des tests sur les coefficients β_j ou, plus généralement, sur $a^t\beta$ de la manière déjà vue.

— Pour tester que $\beta_j = 0$, on part de l'intervalle de confiance bilatère $[\widehat{\beta}_j - t_{\alpha/2}\widehat{s}_n\delta_j; \widehat{\beta}_j + t_{\alpha/2}\widehat{s}_n\delta_j]$ et l'on rejette l'hypothèse si $|\widehat{\beta}_j| > t_{\alpha/2}\widehat{s}_n\delta_j$ avec $\mathbb{P}[|\mathcal{T}(n-p)| \leq t_\alpha] = 1-\alpha$. Par un raisonnement analogue on peut tester que $\beta_2 = \beta_3$ c'est-à-dire que $\beta_2 - \beta_3 = 0$ en rejetant cette hypothèse si $|\widehat{\beta}_2 - \widehat{\beta}_3| > t_{\alpha/2}\widehat{s}_n\delta_a$ avec $a = (0, 1, -1, 0, \ldots, 0)$. Ces tests sont dits tests de Student.

— Pour tester que $\beta_j > 0$, on part de l'intervalle de confiance unilatère $]-\infty, \widehat{\beta}_j + t_\alpha\widehat{s}_n\delta_j]$ et l'on rejette si $\widehat{\beta}_j \leq -t_\alpha\widehat{s}_n\delta_j$. Pour obtenir le niveau de signification du test, on remarque que la frontière entre acceptation et rejet est obtenue lorsque $\widehat{\beta}_j = -t_\alpha\widehat{s}_n\delta_j$, soit $t_\alpha = -\widehat{\beta}_j/(\widehat{s}_n\delta_j)$. La p-value du test s'écrit donc $\alpha_0 = \mathbb{P}[\mathcal{T}(n-p) > -\widehat{\beta}_j/(\widehat{s}_n\delta_j)]$.

2.2.3 Tests d'hypothèses linéaires

Problèmes liés aux tests multiples Pour tester simultanément deux hypothèses du type précédent, par exemple $\beta_1 = \beta_2 = 0$, il convient d'être prudent. D'abord, il ne suffit pas de combiner les tests précédents et rejeter si $|\widehat{\beta}_1| > t_\alpha\widehat{s}_n\delta_1$ ou $|\widehat{\beta}_2| > t_\alpha\widehat{s}_n\delta_2$. En effet, si l'hypothèse est vraie,

$$\mathbb{P}\left[|\widehat{\beta}_1| > t_\alpha\widehat{s}_n\delta_1 \text{ ou } |\widehat{\beta}_2| > t_\alpha\widehat{s}_n\delta_2\right] \leq \mathbb{P}\left[|\widehat{\beta}_1| > t_\alpha\widehat{s}_n\delta_1\right] + \mathbb{P}\left[|\widehat{\beta}_2| > t_\alpha\widehat{s}_n\delta_2\right] \leq 2\alpha.$$

Pour que le test ait le bon niveau, il faut remplacer t_α par $t_{\alpha/2}$ et, de manière plus générale, par $t_{\alpha/k}$ si on teste k égalités simultanées. Mais cette manière de combiner des tests est, le plus souvent, à éviter. Par exemple, pour tester $\beta = 0$ au niveau α, il faudrait combiner p tests individuels de niveau α/p, typiquement très petit, de sorte que chacun des tests aurait une puissance très faible. Ce n'est pas la bonne manière de procéder.

Tests de Fisher Lorsque l'on veut tester une hypothèse sur les coefficients qui correspond à plusieurs équations, comme $\beta_1 = \beta_2 = 0$, il n'est pas nécessaire de faire des tests multiples. En fait cette hypothèse revient à dire que $X\beta$ appartient à un sous-espace vectoriel W, de dimension $p-2$ de V, précisément celui qui est engendré par les colonnes de X d'indices > 2. De manière générale, comme on a une bijection linéaire entre $\mathbb{R}^p$ et V donnée par $\beta \longleftrightarrow X\beta$, toute hypothèse $\boldsymbol{H}_0$ de la forme $\beta \in S$, sous-espace de dimension $p-k$ de $\mathbb{R}^p$ avec $1 \leq k \leq p$, est équivalente par la bijection à $X\beta \in W$ sous-espace de même dimension $p-k$ de V, avec $W = \{X\beta \,|\, \beta \in S\}$. L'hypothèse $\boldsymbol{H}_0$ se traduit typiquement par un système de k équations de la forme $\Phi\beta = 0$ où Φ est une matrice $k \times p$ de rang k et de noyau S.

La solution de ce problème de test est essentiellement géométrique et fondée sur le Théorème de Cochran. On peut décomposer $\mathbb{R}^n$ en trois sous-espaces orthogonaux : $V^\perp$ de dimension $n-p$, W de dimension $p-k$ et W' qui est l'orthogonal de W dans V, donc de dimension k. Alors ε se décompose sur eux en trois vecteurs indépendants $e^\perp$, $e_1 \in W$ et $e_2 \in W'$ ($e = e_1 + e_2$).

On peut alors considérer l'estimateur usuel des moindres carrés $\widehat{\beta}$ qui est donné par $X\widehat{\beta} = X\beta + e_1 + e_2$ avec $D^2(Y, \widehat{\beta}) = \|e^\perp\|^2$ et l'estimateur des moindres carrés sous l'hypothèse $\boldsymbol{H}_0$, $\widetilde{\beta}$, obtenu en projetant Y sur W, de sorte que, si $\boldsymbol{H}_0$ est vraie ($X\beta \in W$), $X\widetilde{\beta} = X\beta + e_1$ avec $D^2(Y, \widetilde{\beta}) = \|e^\perp + e_2\|^2 = \|e^\perp\|^2 + \|e_2\|^2$ d'après le Théorème de Pythagore. On en conclut que, sous $\boldsymbol{H}_0$, $D^2(Y, \widetilde{\beta}) - D^2(Y, \widehat{\beta}) = \|e_2\|^2$ et $D^2(Y, \widehat{\beta}) = \|e^\perp\|^2$ sont indépendants. De plus ces deux variables divisées par σ^2 ont des lois du chi-deux à k et $n-p$ degrés de liberté, respectivement. Il s'ensuit que

$$T(Y) = \frac{k^{-1}\left[D^2(Y, \widetilde{\beta}) - D^2(Y, \widehat{\beta})\right]}{(n-p)^{-1} D^2(Y, \widehat{\beta})} = \frac{D^2(Y, \widetilde{\beta}) - D^2(Y, \widehat{\beta})}{k\widehat{s}_n^2} \tag{II.2.2}$$

a une loi de Fisher $\mathcal{F}_{k,n-p}$. Donc le test qui rejette $\boldsymbol{H}_0$ lorsque $T > t_\alpha$ où t_α est défini par $\mathbb{P}[\mathcal{F}_{k,n-p} > t_\alpha] = \alpha$ a un niveau α. Si $\boldsymbol{H}_0$ est fausse, alors $X\beta \notin W$ et $X\beta$ a une composante z dans W', de sorte que $X\widetilde{\beta} = X\beta - z + e_1$ et qu'il faut remplacer, dans les calculs précédents, $\|e_2\|$ par $\|e_2 + z\|$ au numérateur de T. Si z est grand, T va alors dépasser t_α. C'est là toute l'idée de ce test.

Que faut-il retenir ? Si l'on veut tester que $\beta \in S$, sous espace de dimension $p-k$ de V, ce qui correspond au fait que β vérifie un système de k équations linéaires indépendantes : $\Phi\beta = 0$, on calcule les sommes des carrés résiduelles $D^2(Y, \widetilde{\beta})$ et $D^2(Y, \widehat{\beta})$, correspondant respectivement à l'estimation de β par la méthode des moindres carrés sous l'hypothèse $\boldsymbol{H}_0$ (projection de Y sur W) et sans l'hypothèse (comme d'habitude). On calcule alors la statistique $T(Y)$ donnée par (II.2.2) et l'on rejette l'hypothèse si elle est plus grande que t_α donné par $\mathbb{P}[\mathcal{F}_{k,n-p} > t_\alpha] = \alpha$. Le test a le niveau α et sa puissance est d'autant meilleure que la composante de $X\beta$ qui n'est pas dans W est plus grande, c'est à dire que l'hypothèse est plus fausse.

Si l'on revient à notre problème de départ : tester $\beta_1 = \beta_2 = 0$ qui correspond à $k = 2$ et $W = \{X_- b, b \in \mathbb{R}^{p-2}\}$ où X_- désigne la matrice $n \times (p-2)$ correspondant à X moins ses deux premières colonnes, on peut calculer d'une part $\widetilde{\beta}_- = (X_-^t X_-)^{-1} X_-^t Y \in \mathbb{R}^{p-2}$ et $\widetilde{\beta} = (0, 0, \widetilde{\beta}_-) \in \mathbb{R}^p$ est l'estimateur des moindres carrés sous l'hypothèse, puis $D^2(Y, \widetilde{\beta}) = \|Y - X\widetilde{\beta}\|^2$, enfin la statistique T ainsi que le t_α correspondant à $k = 2$ dans la table de Fisher.

2.2.4 Exemples

Le modèle gaussien ordinaire Relève en particulier du modèle linéaire gaussien le problème de l'estimation des paramètres à partir d'un n-échantillon $Y_1, \ldots, Y_n$ de loi $\mathcal{N}(\mu, \sigma^2)$. On peut écrire dans ce cas $Y_i = \mu + \varepsilon_i$ ce qui correspond à un modèle linéaire avec $p = 1$, $\beta_1 = \mu$ et $X = (1, \ldots, 1)^t$. Les estimateurs des moindres carrés se calculent immédiatement et l'on obtient

$$\widehat{\mu} = n^{-1} \sum_{i=1}^{n} Y_i = \overline{Y}_n \qquad \text{et} \qquad \widehat{s}_n^2 = (n-1)^{-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}_n\right)^2 .$$

Les deux estimateurs sont indépendants avec $\widehat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)$ et $(n-1)(\widehat{s}_n/\sigma)^2 \sim \chi^2(n-1)$ de sorte que $\sqrt{n}(\widehat{\mu} - \mu)/\widehat{s}_n$ a une loi $\mathcal{T}(n-1)$. On obtient alors des intervalles de confiance pour μ comme on l'a vu plus haut et l'on peut faire de même des tests de Student sur μ.

La régression linéaire simple (droite de régression) Dans ce cas on suppose le modèle

$$Y_i = a + bx_i + \varepsilon_i, \quad 1 \leq i \leq n,$$

ce qui donne matriciellement $Y = X\beta + \varepsilon$ avec

$$\beta = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 \\ \ldots & \ldots \\ 1 & x_n \end{pmatrix}.$$

Il sera utile ici, étant donné n quantités $z_1, \ldots, z_n$, de noter $\overline{z} = n^{-1} \sum_{i=1}^n z_i$. Ainsi $\overline{x^2} = n^{-1} \sum_{i=1}^n x_i^2$ et $\overline{xY} = n^{-1} \sum_{i=1}^n x_i Y_i$. On a alors, avec ces notations

$$X^t X = n \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix}; \qquad X^t Y = n \begin{pmatrix} \overline{Y} \\ \overline{xY} \end{pmatrix}$$

et

$$\widehat{\beta} = \begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = \frac{1}{\overline{x^2} - \overline{x}^2} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} \begin{pmatrix} \overline{Y} \\ \overline{xY} \end{pmatrix} = \frac{1}{\overline{x^2} - \overline{x}^2} \begin{pmatrix} \overline{x^2}\,\overline{Y} - \overline{x}\,\overline{xY} \\ \overline{xY} - \overline{x}\,\overline{Y} \end{pmatrix}.$$

On sait que $\widehat{\beta}$ est un vecteur gaussien d'espérance β et matrice de covariance $\sigma^2 (X^t X)^{-1}$ soit

$$\mathrm{Cov}(\widehat{\beta}) = \frac{\sigma^2}{n\left(\overline{x^2} - \overline{x}^2\right)} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}.$$

On remarque que, de manière générale, $Y - X\widehat{\beta} = Y - X(X^tX)^{-1}X^tY = AY$, où A a été utilisé précédemment, de sorte que

$$D^2(Y, \widehat{\beta}) = Y^t A Y = \|Y\|^2 - Y^t X (X^t X)^{-1} X^t Y.$$

Dans notre cas, cela donne

$$\begin{aligned} \frac{D^2(Y,\widehat{\beta})}{n} &= \overline{Y^2} - \frac{\overline{x^2}\,\overline{Y}^2 + \overline{xY}^2 - 2\overline{x}\,\overline{Y}\,\overline{xY}}{\overline{x^2} - \overline{x}^2} = \overline{Y^2} - \frac{\left(\overline{xY} - \overline{x}\,\overline{Y}\right)^2 - \overline{x}^2\overline{Y}^2 + \overline{x^2}\,\overline{Y}^2}{\overline{x^2} - \overline{x}^2} \\ &= \left(\overline{Y^2} - \overline{Y}^2\right) - \left(\overline{xY} - \overline{x}\,\overline{Y}\right)^2 \left(\overline{x^2} - \overline{x}^2\right)^{-1}, \end{aligned}$$

de sorte que

$$\widehat{s}_n^2 = n(n-2)^{-1}\left[\left(\overline{Y^2}-\overline{Y}^2\right)-\left(\overline{xY}-\overline{x}\,\overline{Y}\right)^2\left(\overline{x^2}-\overline{x}^2\right)^{-1}\right].$$

Si l'on veut tester que x n'a aucune influence sur Y, c'est à dire que $b=0$, on peut faire un test de Student. Sous $\boldsymbol{H}_0$, $\widehat{b}$ est normal, de loi $\mathcal{N}\left(0, n^{-1}\sigma^2 v^{-2}\right)$ avec $v^2=\overline{x^2}-\overline{x}^2$ et il est indépendant de $\widehat{s}_n$, de sorte que $\sqrt{n}v\widehat{b}\widehat{s}_n^{-1}$ suit une loi de Student à $n-2$ degrés de liberté. Si t_α est déterminé par le fait que $\mathbb{P}[\mathcal{T}_{n-2}\geq t_\alpha]=\alpha/2$, on acceptera l'hypothèse si $\sqrt{n}v|\widehat{b}|\leq t_\alpha\widehat{s}_n$ et le test aura un niveau α.

L'analyse de la variance à un facteur Le problème pratique correspond à l'étude simultanée de plusieurs groupes sur lesquels on effectue des mesures. On a rencontré un tel exemple en début de chapitre : on cultive des tomates avec divers traitements, pour les plants d'une certaine parcelle, on adopte un certain traitement et l'on a un certain nombre de plants, pas forcément le même, par parcelle. A la fin on mesure le poids de tomates récolté sur chaque plant. Ce poids est évidemment aléatoire, donc fluctuant sur une même parcelle mais en plus il dépend a priori du traitement et l'on peut donc supposer que la loi de ce poids varie d'une parcelle à l'autre.

Plus généralement, on effectue une mesure pour chaque "individu" (au sens statistique) d'une population et l'on suppose que chaque mesure est la réalisation d'une variable aléatoire gaussienne de variance inconnue, mais fixe (la même pour tous les individus) et dont la moyenne dépend du groupe auquel appartient l'individu. Formellement on a des observations $Z_{j,k}$ où j est un indice de groupe avec $1\leq j\leq J$ et k le numéro de l'individu dans le groupe, $1\leq k\leq K_j$. On a donc J groupes de tailles éventuellement variables, celle du groupe j étant K_j. La variable $Z_{j,k}$ représente la quantité observée ou mesurée sur l'individu k du groupe j et elle a une loi $\mathcal{N}\left(\mu_j,\sigma^2\right)$. Ici μ_j traduit l'effet de l'appartenance au groupe j. On peut donc écrire que $Z_{j,k}=\mu_j+\xi_{j,k}$ où les variables $\xi_{j,k}$ sont i.i.d. $\mathcal{N}\left(0,\sigma^2\right)$.

Pour mettre ceci sous la forme d'un modèle linéaire, il suffit de renuméroter les variables. Soit $n_0=0$ et $n_j=\sum_{l=1}^j K_j$. Alors $n_J=n$ est le nombre total d'observations et l'on peut poser $Y_i=Z_{j,k}$ si $n_{j-1}<i\leq n_j$ et $i=n_{j-1}+k$. On notera de même $\varepsilon_i=\xi_{j,k}$ de sorte que l'on peut écrire

$$Y_i=\sum_{j=1}^J \mu_j \mathbb{1}_{]n_{j-1};n_j]}(i)+\varepsilon_i \quad \text{ou matriciellement} \quad Y=X\mu+\varepsilon,$$

avec $\mu\in\mathbb{R}^J$, $Y,\varepsilon\in\mathbb{R}^n$, X est une matrice $n\times J$ et

$$X_i^j=\mathbb{1}_{]n_{j-1};n_j]}(i);\quad \mu=(\mu_1,\ldots,\mu_J)^t;\quad Y=(Y_1,\ldots,Y_n)^t;\quad \varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)^t.$$

On peut alors facilement vérifier que

$$X^tX=\mathrm{Diag}(K_j)\qquad\text{et}\qquad (X^tY)_j=\sum_{i=n_{j-1}+1}^{n_j}Y_i=\sum_{k=1}^{K_j}Z_{j,k}.$$

Finalement $\widehat{\mu}_j=K_j^{-1}\sum_{k=1}^{K_j}Z_{j,k}$ est la moyenne (au sens usuel) des mesures correspondant au groupe j et comme la matrice de covariance de $\widehat{\mu}$ est diagonale, les $\widehat{\mu}_j$ sont des

variables indépendantes de lois respectives $\mathcal{N}\left(\mu_j, K_j^{-1}\sigma^2\right)$. Quant à l'estimateur de la variance σ^2, il s'écrit

$$\widehat{s}_n^2 = \frac{1}{n-J}\sum_{j=1}^{J}\sum_{k=1}^{K_j}(Z_{j,k}-\widehat{\mu}_j)^2\,.$$

A partir de là, on peut se poser diverses questions. L'une d'elles est la suivante : y a-t-il ou non un effet lié à l'appartenance aux différents groupes ? Si la réponse est non, tous les μ_j doivent être égaux. C'est une hypothèse que l'on peut tester et qui correspond au modèle très simple de variables Y_i i.i.d. de loi $\mathcal{N}\left(\mu,\sigma^2\right)$. Sous cette hypothèse, notée $\boldsymbol{H}_0$, l'estimateur de μ est $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ et la somme des carrés résiduelle s'écrit $D_0^2 = \sum_{i=1}^n \left(Y_i - \overline{Y}_n\right)^2$(cf. le modèle gaussien ordinaire). Sans l'hypothèse, la somme des carrés résiduelle est $D_1^2 = \sum_{j=1}^J \sum_{k=1}^{K_j}(Z_{j,k}-\widehat{\mu}_j)^2$. D'après ce que l'on a vu sur les tests de Fisher, la quantité

$$T(Y) = \frac{(n-J)\left[D_0^2 - D_1^2\right]}{(J-1)D_1^2} = \frac{D_0^2 - D_1^2}{(J-1)\widehat{s}_n^2}$$

suit une loi de Fisher $\mathcal{F}_{J-1,n-J}$. On rejettera donc $\boldsymbol{H}_0$ lorsque $T > t_\alpha$ où t_α est défini par $\mathbb{P}[\mathcal{F}_{J-1,n-J} > t_\alpha] = \alpha$ et le test correspondant aura un niveau α.