Feuille de TP n°7
Vecteurs aléatoires et modèle linéaire gaussiens

1 Vecteurs aléatoires gaussiens

Définition 1.1. *Soit X un vecteur aléatoire défini sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, à valeurs dans $\mathbb{R}^d$ avec $d \geqslant 1$. X est dit gaussien si toute combinaison linéaire de ses composantes est une v.a. gaussienne.*

Théorème 1.2. *La loi d'un vecteur aléatoire gaussien est entièrement déterminée par son vecteur espérance $m = \mathbb{E}(X) \in \mathbb{R}^d$ et sa matrice de covariance $\Gamma = \mathbb{E}((X-m)(X-m)^\top) \in \mathbb{S}_d^+$ où $\mathbb{S}_d^+$ est le cône convexe des matrices carrées $d \times d$ symétriques semi-définies positives (pas forcément inversibles). Soit $X \sim \mathcal{N}(m, \Gamma)$. Alors, pour tout $u \in \mathbb{R}^d$, on a*

$$\mathbb{E}(\exp(i\langle u, X\rangle)) = \exp\left(i\langle u, m\rangle - \frac{1}{2}u^\top \Gamma u\right).$$

Théorème 1.3. *Soit $X \sim \mathcal{N}(m, \Gamma)$. X admet une densité f_X par rapport à la mesure de Lebesgue de $\mathbb{R}^d$ si et seulement si Γ est inversible et l'on a*

$$f_X(x) = ((2\pi)^d \det \Gamma)^{-1/2} \exp\left(-\frac{1}{2}(x-m)^\top \Gamma^{-1}(x-m)\right).$$

La loi particulière $\mathcal{N}(0, \mathrm{I}_d)$ où I_d est la matrice identité de $\mathbb{R}^d$ est appelée *gaussienne standard*. Dans tout ce texte, l'espérence des vecteurs et matrices est prise composante par composante. Voici quelques propriétés fondamentales des vecteurs gaussiens, cf. par exemple [1, IV.4 et VI.4]

P1. Soit $X = (X_1, \ldots, X_d)$ un vecteur gaussien de matrice de covariance Γ. Les trois propositions suivantes sont équivalentes : (a) $(X_1, \ldots, X_d)$ sont deux à deux indépendantes, (b) $(X_1, \ldots, X_d)$ sont indépendantes dans leur ensemble, (c) La matrice Γ est diagonale.

P2. Soit $m \in \mathbb{R}^d$, $A \in \mathcal{M}_d(\mathbb{R})$ et $X \sim \mathcal{N}(0, \mathrm{I}_d)$, alors[1] $m + AX \sim \mathcal{N}(m, AA^\top)$.

P3. Soit $Z = (X, Y)$ un vecteur gaussien de $\mathbb{R}^{d+1}$ avec $X = (X_1, \ldots, X_d)$ d'espérance m et de matrice de covariance inversible Γ. Alors, la loi conditionnelle de Y sachant X est gaussienne d'espérance affine en X $\mathbb{E}(Y|X) = a + b^\top X = \mathbb{E}(Y) + b^\top(X-m)$ et de variance $\mathbf{Var}(Y|X) = \mathbf{Var}(Y) - b^\top \Gamma b$ avec $a = \mathbb{E}(Y) - b^\top m$ et $b = \Gamma^{-1}\mathbf{Cov}(X, Y)$. De plus, $\varepsilon = Y - \mathbb{E}(Y|X)$ est indépendante de X.

Exercice 1.4 (Algorithme de Box-Muller). Soit (X, Y) un vecteur aléatoire de $\mathbb{R}^2$. Montrer que (X, Y) suit la loi normale $\mathcal{N}(0, \mathrm{I}_2)$ si et seulement si $X = r\cos\theta$ et $Y = r\sin\theta$ où r et θ sont deux variables aléatoires indépendantes avec r^2 de loi exponentielle $\mathcal{E}(1/2)$ et θ de loi uniforme $\mathcal{U}([0, 2\pi])$. En déduire un algorithme permettant de générer des réalisations de variables aléatoires indépendantes gaussiennes $\mathcal{N}(m, \Gamma)$. Cf. [7].

Exercice 1.5. Étudier et expliquer le programme suivant

```
N=input('Entrez la taille de l'échantillon N : ');
max=round(3*N/2);
m=input('Précisez la valeur de la moyenne m : ');
sigma=input('Précisez la valeur de écart type : ');
X=2*rand(max,1)-ones(max,1); Y=2*rand(max,1)-ones(max,1);
```

[1] Réciproquement, une matrice de covariance $\Gamma \in \mathbb{S}_d^+$ peut toujours s'écrire $AA^\top$ où $A \in \mathcal{M}_d(\mathbb{R})$. La matrice A n'est pas unique. La méthode de Choleski en fournit une (rapide), la racine carrée matricielle obtenue par diagonalisation en base orthonormée de Γ également (lent). Ce procédé ne requière pas l'inversibilité de Γ.

```
S=X.^2+Y.^2; X=X(find(S<1)); Y=Y(find(S<1));
r=sqrt(X.^2+Y.^2); R=2*sqrt(-log(r))./r; Z=R(1:N).*X(1:N);
T=m*ones(N,1)+sqrt(sigma^2)*Z;
```

2 Bref rappel sur les échantillons gaussiens

Loi du χ^2. La loi $\chi^2(n)$ est la loi de $Z_1^2+\cdots+Z_n^2=\|(Z_1,\ldots,Z_n)\|_2^2$ où $Z_1,\ldots,Z_n$ sont des v.a.r. i.i.d. de loi $\mathcal{N}(0,1)$. Sa moyenne est n, sa variance $2n$, et sa densité est $x\mapsto(2^{n/2}\Gamma(n/2))^{-1}x^{n/2-1}\exp(-x/2)\,\mathrm{I}_{\mathbb{R}_+}(x)$. C'est donc une loi Gamma particulière de paramètres $\lambda=1/2$ et $a=n/2$. Il est clair que $\chi^2(n)*\chi^2(m)=\chi^2(n+m)$.
Loi de Student. Soit A et B des v.a.r. indépendantes de loi respectives $\mathcal{N}(0,1)$ et $\chi^2(n)$, alors la loi de Student $t(n)$ est la loi de $\sqrt{n}A/\sqrt{B}$. La loi $t(n)$ est centrée, de densité $x\mapsto\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}}(1+x^2/n)^{-(n+1)/2}$ sur $\mathbb{R}$, et de variance $n/(n-2)$. On a enfin que $t(n)$ converge étroitement vers $\mathcal{N}(0,1)$ quand n tend vers $+\infty$.
Loi de Fisher(-Snedecor). Soit A et B deux v.a.r. indépendantes de lois respectives $\chi^2(n)$ et $\chi^2(m)$, alors la loi de Fisher(-Snedecor) $F(n,m)$ est la loi de $(A/n)/(B/m)$. Sa moyenne et sa variance valent respectivement $m/(m-2)$ et $\frac{2m^2(m+n-2)}{n(m-4)(m-2)^2}$ et sa densité est $x\mapsto\frac{\Gamma((n+m)/2)n^{n/2}m^{m/2}}{\Gamma(n/2)\Gamma(m/2)}\frac{x^{2/n-1}}{(m+nx)^{(n+m)/2}}\,\mathrm{I}_{\mathbb{R}_+}(x)$. Il est immédiat que si $Z\sim t(n)$, alors $Z^2\sim F(1,n)$. De plus, si $Z\sim F(n,m)$ alors $1/Z\sim F(m,n)$.
Moyenne et variance empiriques. Soit $Z_1,\ldots,Z_n$ des v.a.r. i.i.d. de loi $\mathcal{N}(m,\sigma^2)$. Alors les estimateurs de la moyenne empirique $\widehat{m}$ et de la variance empirique $\widehat{\sigma^2}$ définis par

$$\widehat{m}:=\frac{Z_1+\cdots+Z_n}{n}\quad\text{et}\quad\widehat{\sigma^2}:=\frac{1}{n-1}\sum_{i=1}^n(Z_i-\widehat{m})^2$$

sont indépendants et sans biais. De plus, on a

$$\frac{1}{\sigma^2}\sum_{i=1}^n(Z_i-m)^2\sim\chi^2(n)\quad\text{et}\quad(n-1)\,\frac{\widehat{\sigma^2}}{\sigma^2}\sim\chi^2(n-1).$$

et

$$\sqrt{n}\,\frac{\widehat{m}-m}{\sigma}\sim\mathcal{N}(0,1)\quad\text{et}\quad\sqrt{n}\,\frac{\widehat{m}-m}{\sqrt{\widehat{\sigma^2}}}\sim t(n-1)\quad\text{et}\quad n\,\frac{(\widehat{m}-m)^2}{\widehat{\sigma^2}}\sim F(1,n-1).$$

Pour tout ce qui concerne ces lois et les échantillons gaussiens, on pourra consulter par exemple [2, Chap. 5].

Théorème 2.1 (de Cochran). *Soit $Y=(Y_1,\ldots,Y_n)$ des v.a.r. i.i.d. de loi $\mathcal{N}(0,\sigma^2)$, alors l'image de $(Y_1,\ldots,Y_n)$ par toute transformation orthogonale de $\mathbb{R}^n$ est encore un vecteur gaussien de loi $\mathcal{N}(0,\sigma^2\mathrm{I}_n)$. De plus, les projections orthogonales $P_{E_1}(Y),\ldots,P_{E_k}(Y)$ de Y sur des sous-espaces vectoriels $E_1,\ldots,E_k$ de $\mathbb{R}^n$ deux à deux orthogonaux sont indépendantes, et $\sigma^{-2}\|P_{E_i}(Y)\|_2^2$ suit la loi $\chi^2(\dim(E_i))$ pour tout $1\leqslant i\leqslant p$.*

Corollaire 2.2 (Du théorème de Cochran). *Soit $\varepsilon_1,\ldots,\varepsilon_n$ des v.a.r. i.i.d. de loi $\mathcal{N}(0,1)$ et V un sous-espace vectoriel de dimension p de $\mathbb{R}^n$. On observe $Y=m+\sigma\varepsilon$ où $\varepsilon:=(\varepsilon_1,\ldots,\varepsilon_n)$ et où $m\in V$ et $\sigma^2>0$ sont inconnus. Si $P_V\in\mathrm{End}(\mathbb{R}^n)$ désigne la projection orthogonale sur V, on a*

1. *$\widehat{m}:=P_V(Y)$ est un estimateur sans biais de m;*
2. *La variance résiduelle $\widehat{\sigma^2}:=(n-p)^{-1}\|Y-P_V(Y)\|_2^2$ est un estimateur sans biais de σ^2;*
3. *Les vecteurs aléatoires $Y-P_V(Y)$ et $P_V(Y)$ sont indépendants, et on a $\sigma^{-2}\|P_V(Y)-m\|_2^2\sim\chi^2(p)$ et $\sigma^{-2}\|Y-P_V(Y)\|_2^2\sim\chi^2(n-p)$.*

3 Modèle linéaire gaussien

Soit (X,Y) un vecteur aléatoire de $\mathbb{R}^{p-1}\times\mathbb{R}$. On suppose que la loi de X admet une densité de probabilité φ par rapport à une mesure positive μ sur $\mathbb{R}^{p-1}$. On pensera par exemple aux cas particuliers où μ est la mesure

de Lebesgue sur un borélien de $\mathbb{R}^{p-1}$, ou la mesure de comptage sur un sous-ensemble dénombrable de $\mathbb{R}^{p-1}$. On considère une modélisation gaussienne de la loi conditionnelle $\mathcal{L}(Y \mid X)$ sous la forme

$$\mathcal{L}(Y \mid X = x) = \mathcal{N}\big(\beta_1 f_1(x_1) + \cdots + \beta_{p-1} f_{p-1}(x_{p-1}) + \beta_p\,,\, \sigma^2\big)$$

où les fonctions réelles de la variable réelle $f_1, \ldots, f_{p-1}$ sont connues, et où les paramètres $\beta := (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ et $\sigma^2 \in \mathbb{R}_+^*$ sont inconnus. La moyenne est linéaire en β, d'où le qualificatif de *modèle linéaire gaussien*. On peut écrire $\mathcal{L}(Y \mid X = x) = \mathcal{N}(\mathbf{x}^\top \beta, \sigma^2)$ où $\mathbf{x} := (f_1(x_1), \ldots, f_{p-1}(x_{p-1}), 1)^\top$. Partant de l'observation d'un échantillon de taille $n \geqslant p$ noté $(x_1, y_1), \ldots, (x_n, y_n)$ du vecteur (X, Y), l'objectif est d'estimer β et σ^2 et d'effectuer des tests d'hypothèse sur β. On ne s'intéresse pas à la loi de X ici. En particulier, φ restera inconnue et inutile. L'espérence conditionnelle de Y sachant X est donnée pour tout $x \in \mathbb{R}^{p-1}$ par

$$\mathbf{E}(Y \mid X = x) = \beta_1 f_1(x_1) + \cdots + \beta_{p-1} f_{p-1}(x_{p-1}) + \beta_p.$$

L'ensemble paramétrique associé est $\Theta := \{(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*\}$. Le nombre entier p correspond à la dimension du paramètre β. On a coutume en statistique appliquée d'écrire ce modèle conditionnel sous la forme

$$Y_i = \beta_1 f_1(x_{i,1}) + \cdots + \beta_{p-1} f_{p-1}(x_{i,p-1}) + \beta_p + \varepsilon_i,$$

où les ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et où les $x_{i,j}$ sont des « variables déterministes » connues. Cette écriture n'est pas abusive si l'on considère que les Y_i sont des v.a.r. indépendantes, la loi de Y_i étant précisément $\mathcal{N}\big(\beta_1 f_1(x_1) + \cdots + \beta_{p-1} f_{p-1}(x_{p-1}) + \beta_p\,,\, \sigma^2\big)$. Le paramètre β_p est appelé *intercept*. On parle de *modèle linéaire simple* lorsque $p = 2$, et on écrit alors $Y_i = \beta_1 f(x_i) + \beta_2 + \varepsilon_i$. On parle de *modèle linéaire multiple* quand $p \geqslant 2$. De manière équivalente, on parle de *régression linéaire simple* et de *régression linéaire multiple*.

Pour le modèle linéaire simple ($p = 2$), l'estimateur des moindres carrés pour β consiste à minimiser $(\beta_1, \beta_2) \mapsto \sum_{i=1}^n (y_i - \beta_1 f(x_i) - \beta_2)^2$. Cela donne en notant $u_i = f(x_i)$, $\overline{u} := (u_1 + \cdots + u_n)/n$ et $\overline{y} := (y_1 + \cdots + y_n)/n$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^n (u_i - \overline{u})(y_i - \overline{y})}{\sum_{i=1}^n (u_i - \overline{u})^2} \quad \text{et} \quad \widehat{\beta_2} = \overline{y} - \widehat{\beta_1}\,\overline{u}.$$

Ces estimateurs sont sans biais. Les *valeurs ajustées* $\widehat{y_i} := \widehat{\beta_1} f(x_i) + \widehat{\beta_2}$ sont les estimées de $\mathbf{E}(Y_i | X = x_i)$. Les *résidus* correspondants sont les $e_i := y_i - \widehat{y_i}$. La courbe d'équation $y = \beta_1 f(x) + \beta_2$ est appelée *courbe de regression*. On observera que la nature gaussienne du problème n'intervient pas dans la déduction des estimateurs des moindres carrés. En revanche, c'est elle qui explique qu'il s'agit des estimateurs du maximum de vraisemblance.

Expression des estimateurs. Le corollaire 2.2 donne des estimateurs de β et σ^2. En effet, il suffit de considérer le sous-espace vectoriel V de $\mathbb{R}^n$ engendré par les colonnes de la matrice $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ définie par

$$\mathbf{X} = \begin{pmatrix} f_1(x_{1,1}) & \cdots & f_{p-1}(x_{1,p-1}) & 1 \\ \vdots & & \vdots & \vdots \\ f_1(x_{n,1}) & \cdots & f_{p-1}(x_{n,p-1}) & 1 \end{pmatrix}.$$

On a alors $V = \{\mathbf{X}\beta,\, \beta \in \mathbb{R}^p\} = \mathrm{Im}(\mathbf{X})$. Dans la base canonique de $\mathbb{R}^n$, la matrice de la projection orthogonale P_V s'écrit $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. L'estimateur $\widehat{m}$ de $m = \mathbf{X}\beta$ s'écrit alors $\widehat{m} = P_V(Y) = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y$. Lorsque la matrice $\mathbf{X}$ est de rang plein p, et le paramètre β peut être estimé par $\widehat{\beta} := (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y$. Comme $m = \mathbf{X}\beta$ et comme $\mathbf{X}\widehat{\beta} = P_V(Y)$, on a par linéarité que $\widehat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2)$. On a également $(n-p)\widehat{\sigma^2}/\sigma^2 \sim \chi^2(n-p)$. Les estimateurs $\widehat{\beta}$ et $\widehat{\sigma^2}$ sont indépendants et sans biais. On note $\widehat{\sigma}$ la racine carrée de $\widehat{\sigma^2}$. Lorsque $p = 2$, on retrouve bien entendu les formules du modèle linéaire simple pour $\widehat{\beta_1}$ et $\widehat{\beta_2}$.

Test sur une hypothèse linéaire. Soit W un sous-espace vectoriel de V, de dimension $p - q < p$. On définit les hypothèses antagonistes $\mathcal{H}_0$:« $m \in W$ » et $\mathcal{H}_1$:« $m \in V \setminus W$ ». Soit U le sous-espace vectoriel de V tel que $V = W \perp U$. Le théorème de Cochran 2.1 appliqué au vecteur gaussien $Y - m \sim \mathcal{N}(0, \sigma^2\,\mathrm{I}_n)$ et à la décomposition

$\mathbb{R}^n = W \oplus U \oplus V^\perp$ entraîne alors en particulier que les vecteurs aléatoires $P_{V^\perp}(Y-m)$ et $P_U(Y-m)$ sont indépendants, et que $\sigma^{-2}\|P_{V^\perp}(Y-m)\|_2^2$ et $\sigma^{-2}\|P_U(Y-m)\|_2^2$ ont pour lois respectives $\chi^2(n-p)$ et $\chi^2(q)$. Or on a $P_V - P_W = P_U$ et $Id - P_V = P_{V^\perp}$. En particulier, $P_{V^\perp}(m) = 0$ donc $P_{V^\perp}(Y-m) = Y - P_V(Y)$, et sous $\mathcal{H}_0$, $P_U(m) = 0$ donc $P_U(Y-m) = P_V(Y) - P_W(Y)$. On en déduit que sous $\mathcal{H}_0$

$$Q = \frac{(n-p)}{q}\frac{\|P_V(Y) - P_W(Y)\|_2^2}{\|Y - P_V(Y)\|_2^2} \sim F(q, n-p). \tag{1}$$

Cela donne un test de niveau α consistant à rejetter $\mathcal{H}_0$ dès que $Q > F_\alpha(q, n-p)$ où $F_\alpha(q, n-p)$ est le quantile α de la loi de Fisher $F(q, n-p)$. En considérant une définition matricielle de W, il est possible d'exprimer très simplement Q en fonction de β, cf. [6, Chap. 6.4.3]. La statistique Q est à une transformation croissante près celle du rapport de maximum de vraisemblance, cf. [6, Chap. 6.4.2].

Test de validité du modèle. La première question que l'on se pose dans la pratique est de savoir si les x_i influent sur la moyenne de Y. Cela revient exactement à considérer l'hypothèse linéaire $\mathcal{H}_0$:« $\beta_1 = \cdots = \beta_{p-1} = 0$ » qui correspond à $q = p-1$ et donc à considérer la statistique de test (1). Comme dans ce cas $W = \mathbb{R}\mathbf{1}$ où $\mathbf{1} := (1, \ldots, 1)^\top$, on a $p_W(Y) = \frac{1}{n}(Y_1 + \cdots + Y_n)\mathbf{1} =: \overline{Y}\mathbf{1}$ et $p_V(Y) = \mathbf{X}\widehat{\beta} =: \widehat{Y}$. Cela conduit à la statistique de test

$$Q = \frac{(n-p)}{(p-1)}\frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2} = \frac{(n-p)}{(p-1)}\frac{\text{variance expliquée}}{\text{variance totale}}.$$

Le théorème de Pythagore correspond donc à la fameuse « formule de décomposition de la variance »

$$\underbrace{\frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2}_{\text{variance totale}} = \underbrace{\frac{1}{n}\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}_{\text{variance « expliquée »}} + \underbrace{\frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}_{\text{variance « résiduelle »}},$$

et en terme de lois : $\chi^2(n-1) = \chi^2(p-1) * \chi^2(n-p)$. Le praticien préfère parfois disposer d'un indice plutôt que d'un test. Un indice fréquement utilisé pour quantifier la qualité de la régression est l'indice $R^2 :=$ 1 − (variance résiduelle/variance totale) = variance expliquée/variance totale. On a toujours $R^2 \in [0,1]$, et R^2 est d'autant plus proche de 1 que la régression est « explicative » au sens de la variance.

Intervalle de confiance pour une composante β_i. Comme $\widehat{\beta}$ et $(n-p)\widehat{\sigma^2}/\sigma^2$ sont indépendants et de lois respectives $\mathcal{N}(\beta, (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2)$ et $\chi^2(n-p)$, on en déduit que

$$\frac{(\widehat{\beta}_j - \beta_j)}{\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{j,j}\,\widehat{\sigma^2}}} \sim t(n-p).$$

On peut ainsi effectuer des tests d'hypothèse sur une composante donnée de β. On a par exemple pour β_j l'intervalle de confiance exact de niveau $(1-\alpha)$ suivant

$$I_j(\alpha) := \widehat{\beta}_j + \left(t_{\alpha/2}(n-p)\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{j,j}\,\widehat{\sigma^2}}\right)[-1, +1],$$

où $t_{\alpha/2}(n-p)$ est le quantile $\alpha/2$ de la loi de Student $t(n-p)$. Ainsi, on peut construire un test de niveau α pour $\mathcal{H}_0 : \{\beta_j = 0\}$ en acceptant $\mathcal{H}_0$ dès que $0 \in I_j(\alpha)$. Sous $\mathcal{H}_0$, $\mathbb{P}(0 \notin I_j(\alpha)) = 1-\alpha$. La p-value de ce test, i.e. la plus grande valeur de α pour laquelle $\mathcal{H}_0$ est acceptée, se calcule très facilement en distinguant deux cas (selon le signe de $\widehat{\beta}_j$). On prendra garde au fait que ces intervalles de confiance $(I_j(\alpha))_{1\leqslant j\leqslant p}$ ne sont pas indépendants, et il en va de même pour les tests associés, d'où l'intérêt de la statistique (1).

Intervalle de confiance pour σ^2. Comme $\sigma^{-2}\|Y - P_V(Y)\|_2^2 \sim \chi^2(n-p)$, on en déduit qu'avec probabilité $1-\alpha$, le paramètre σ^2 se trouve dans l'intervalle

$$\|Y - P_V(Y)\|_2^2 \left[\frac{1}{\chi^2_{1-\alpha/2}(n-p)}\,,\ \frac{1}{\chi^2_{\alpha/2}(n-p)}\right]$$

où $\chi^2_\tau(n-p)$ est le quantile τ de la loi $\chi^2(n-p)$.

Région de confiance pour $\mathbf{H}\beta$. Rappelons que $m = \mathbf{X}\beta$. Comme $\sigma^{-2}\|P_V(Y)-m\|_2^2$ et $\sigma^{-2}\|P_V(Y)-Y\|_2^2$ sont indépendantes et de lois respectives $\chi^2(p)$ et $\chi^2(n-p)$, on en déduit que

$$\frac{(n-p)}{p}\frac{\|P_V(Y)-m\|_2^2}{\|Y-P_V(Y)\|_2^2} \sim F(p,n-p).$$

Ainsi, avec probabilité $1-\alpha$, le paramètre m est dans la boule $B_2(P_V(Y);r_\alpha)$ de rayon

$$r_\alpha = \|Y-P_V(Y)\|_2\sqrt{\frac{p}{n-p}F_\alpha(p,n-p)}$$

où $F_\alpha(p,n-p)$ est le quantile α de la loi $F(p,n-p)$. La traduction de cette région en terme de β et $\mathbf{X}$ est immédiate puisque $m=\mathbf{X}\beta$ et $P_V(Y)=\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y = \mathbf{X}\widehat{\beta}$. Si $\mathbf{H}$ est une matrice $q\times p$ de rang $q<p$, la région de confiance pour $\mathbf{H}\beta$ est donnée par

$$\Big\{\beta\in\mathbb{R}^p,\ (\widehat{\beta}-\beta)^\top\mathbf{H}^\top(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top)^{-1}\mathbf{H}(\widehat{\beta}-\beta)\leqslant q\,\widehat{\sigma^2}\,F_\alpha(q,n-p)\Big\}.$$

Lien avec le maximum de vraisemblance et formulation conditionnelle. La vraisemblance s'écrit

$$L_n(\beta,\sigma^2)=(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\beta\|_2^2\right)\prod_{i=1}^n\varphi(x_{i,1},\dots,x_{i,p}),$$

où $\mathbf{Y}:=(y_1,\dots,y_n)^\top$ et où $\mathbf{X}:=(f_j(x_{i,j}))_{1\leqslant i\leqslant n,1\leqslant j\leqslant p}$. L'estimateur du maximum de vraisemblance $(\widehat{\beta},\tilde{\sigma^2})\in\Omega$ de (β,σ^2) est défini par

$$(\widehat{\beta},\tilde{\sigma^2}):=\arg\sup_{(\beta,\sigma^2)\in\Omega}L_n(\beta,\sigma^2)=\arg\sup_{(\beta,\sigma^2)\in\Omega}\log L_n(\beta,\sigma^2).$$

On montre sans difficulté qu'il est bien défini et qu'il ne dépend pas de φ, i.e. de la loi de X. Lorsque la matrice $\mathbf{X}$ est de rang plein, on a les formules explicites suivantes obtenues en annulant le gradient de L_n

$$\widehat{\beta}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}\quad\text{et}\quad\tilde{\sigma^2}=n^{-1}\left\|\mathbf{Y}-\mathbf{X}\widehat{\beta}\right\|_2^2.$$

La matrice $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}$ correspond à la projection orthogonale dans $\mathbb{R}^n$ sur l'espace vectoriel $\mathrm{Im}(\mathbf{X})$ des moyennes admissibles. Ainsi, $\mathbf{X}\widehat{\beta}$ est la projection orthogonale de $\mathbf{Y}$ sur $\mathrm{Im}(\mathbf{X})$. L'estimateur $\widehat{\beta}$ coïncide avec l'estimateur des moindres carrés obtenu en minimisant $z\in\mathbb{R}^p\to\|\mathbf{Y}-\mathbf{X}z\|_2^2$. On a de plus $\widehat{\beta}\sim\mathcal{N}(\beta,\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$, et $\widehat{\beta}$ est donc sans biais. En revanche, l'estimateur $\tilde{\sigma^2}$ fait intervenir $\widehat{\beta}$ au lieu de β, ce qui explique qu'il soit biaisé. Sa version débiaisée est donnée par $\widehat{\sigma^2}:=(n/(n-p))\tilde{\sigma^2}$. On a alors $(n-p)\widehat{\sigma^2}/\sigma^2\sim\chi^2(n-p)$. L'approche conditionnelle et le lien avec le maximum de vraisemblance à le mérite de placer le modèle linéaire gaussien dans le cadre de la statistique asymptotique paramétrique.

Intervalle de confiance sur l'espérance conditionnelle et intervalle de prédiction sur la loi conditionnelle. On sait que pour tout $x\in\mathbb{R}^{p-1}$, $\mathcal{L}(Y\,|\,X=x)=\mathcal{N}(\mathbf{x}^\top\beta,\sigma^2)$ et $\mathbf{E}(Y\,|\,X=x)=\mathbf{x}^\top\beta$, où $\mathbf{x}:=(f_1(x_1),\dots,f_{p-1}(x_{p-1}),1)^\top$. Pour un niveau α, on peut donner un intervalle de confiance qui contiendra la moyenne $\mathbf{x}^\top\beta$ avec probabilité $1-\alpha$. D'un autre côté, on peut également donner un intervalle de prédiction qui sera de probabilité $1-\alpha$ pour la loi conditionnelle $\mathcal{L}(Y\,|\,X=x)$. Ces deux intervalles sont différents[2].

[2]On peut par exemple retenir que les intervalles de confiance concernent des paramètres (moyenne,...) tandisque les intervalles de prédiction concernent des lois (ou des v.a.). Ces intervalles ne sont pas uniques pour un niveau donné. Idéalement, on cherchera toujours pour un niveau donné les intervalles les plus petits.

Pour construire un intervalle de confiance sur la moyenne, on écrit que $\mathbf{x}^\top\widehat{\beta} \sim \mathcal{N}(\mathbf{x}^\top\beta, \sigma^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{x})$. Cette v.a. est indépendante de $\widehat{\sigma^2}$. On obtient donc que $(x^\top\widehat{\beta} - x^\top\beta)/\widehat{\sigma} \sim t(n-p)$. Ainsi, avec probabilité $1-\alpha$, $\mathbf{E}(Y \mid X = x)$ se trouve dans l'intervalle

$$\mathbf{x}^\top\widehat{\beta} + t_{\alpha/2}(n-p)\sqrt{(\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})\widehat{\sigma^2}}\ [-1,+1],$$

où $t_{\alpha/2}(n-p)$ est le quantile $\alpha/2$ de la loi de Student $t(n-p)$.

Pour construire un intervalle de prédiction sur la loi conditionnelle, on considère une variable aléatoire Y_x de loi $\mathcal{L}(Y \mid X = x)$ indépendante de l'échantillon. On a alors

$$\mathbf{x}^\top\widehat{\beta} - Y_x \sim \mathcal{N}(0, \sigma^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}) * \mathcal{N}(0,\sigma^2) = \mathcal{N}(0, \sigma^2(1+\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})).$$

Par suite, par indépendance de $\widehat{\beta}$, $\widehat{\sigma^2}$, et Y_x, il vient

$$\frac{\mathbf{x}^\top\widehat{\beta} - Y_x}{\sqrt{(1+\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})\widehat{\sigma^2}}} \sim t(n-p).$$

Ainsi, avec une probabiltié $1-\alpha$, la variable aléatoire Y_x prend sa valeur dans l'intervalle

$$\mathbf{x}^\top\widehat{\beta} + t_{\alpha/2}(n-p)\sqrt{(1+\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})\widehat{\sigma^2}}\ [-1,+1].$$

Résidus et contrôle. Si le modèle est exact, le vecteur des résidus $\mathbf{Y} - \mathbf{X}\widehat{\beta}$ est gaussien centré de matrice de covariance $\sigma^2(\mathrm{I}_n - \mathbf{X}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X})$. Nous avons supposé que les composantes de $\mathbf{Y} - \mathbf{X}\beta$ étaient de moyenne nulle, indépendantes, de variance constante, et de loi gaussienne. Ces hypothèses sont rarement toutes satisfaites dans la pratique malheureusement. On peut par exemple tracer un `qqplot` des quantiles de la la gaussienne versus les quantiles empiriques des résidus (droite de Henri), cf. par exemple [4, Chap. 15.4.1.C]. On peut alors s'intéresser aux résidus studentisés, cf. par exemple [4, Chap. 17.3].

Références bibliographiques. Par exemple [4, Chap. 15 et 16], [2, Chap. 5], [6, Chap. 6], [3], [5].

Les données et outils de Stixbox. Stixbox possède une collection de jeux de données empiriques. La commande `getdata` permet d'y accéder, et en donne également un bref descriptif. Stixbox fournit également trois fonctions utiles pour le modèle linéaire, qui sont

- `identify`. Permet d'itentifier des points sur un nuage de point issu d'un échantillon. Un click sur le bouton du mileu fait sortir.
- `linreg`. Effectue une *régression linéaire simple*[3]. Un graphique est automatiquement produit. Ce graphique comporte la courbe de régression de y en fonction de x, les zônes de confiance et de prédiction du niveau désiré, ainsi que le nuage de points du jeu de données.
- `lsfit`. Effectue une *régression linéaire multiple*. La syntaxe est identique à `linreg`, pour les trois paramètres obligatoires. Aucun graphique n'est produit.
- `cmpmod`. Quantifie la différence entre un modèle et un sous-modèle pour une *régression linéaire multiple*, au moyen d'un test de Fisher.

Il est très instructif de lire le code de ces fonctions en utilisant la commande `type`.

Exercice 3.1 (Régression linéaire simple). Créer un échantillon de taille 100 du couple (X,Y) de telle sorte que $\mathcal{L}(X) = \mathcal{U}([0,1])$ et $\mathcal{L}(Y \mid X = x) = \mathcal{N}(1+x,\, 1)$. Étudier la régression linéaire simple $Y_i = 1 + x_i + \varepsilon_i$ au moyen de `linreg`. Les résultats sont-ils compatibles avec les vraies valeurs ? Voici un programme à titre d'exemple et de correction.

[3] `linreg` permet également de faire de la régression linéaire polynomiale.

```
% On fait appel aux agents de surface... :-)
clear; figure(1); clf; figure(2); clf;
% Création des données !
n = 100; sigma = 1; beta(1) = 1; beta(2) = 1;
X = rand(n,1);
Y = beta(1)*X + beta(2) + sigma*randn(n,1);
% Pour voir le nuage et identifier des points... (facultatif)
% On sort avec un click sur le bouton du mileu.
v = identify(X,Y);

% Tracé de la courbe de régression et du nuage de points.
% Attention, le modèle ici s'écrit Y = beta(1)*X + beta(2) + erreur.
% Appeler linreg avec [] au lieu de 1-alpha pour obtenir un graphique
% dépourvu des régions de confiance et de prédiction.
figure(1);
alpha = 0.05;
[beta, ICbeta, residus, sigma, ICsigma] = linreg(Y,X,1-alpha)
ylabel('Y'); xlabel('X');
% beta    = estimation de beta
% ICbeta  = intervalles de confiance pour les composantes de beta
% sigma   = estimation de sigma
% ICsigma = intervalle de confiance pour sigma

% Le vrai beta(1) se trouve avec probabilité 1-alpha dans l'intervalle de
% confiance ICbeta(1,:) donné par linreg. Considérons l'hypothèse
% H0: « beta(1)=0 ». On a alors un test de niveau alpha pour H0 qui consiste
% à rejeter H0 si 0 n'est pas dans ICbeta(1,:). Le alpha critique pour lequel
% cela se produit est appelé p-value du test. La p-value n'est pas calculée
% par linreg... Exercice !

% Tracé du graphe des résidus.
figure(2); plot(X,residus,'*'); xlabel('X'); ylabel('Résidus');
```

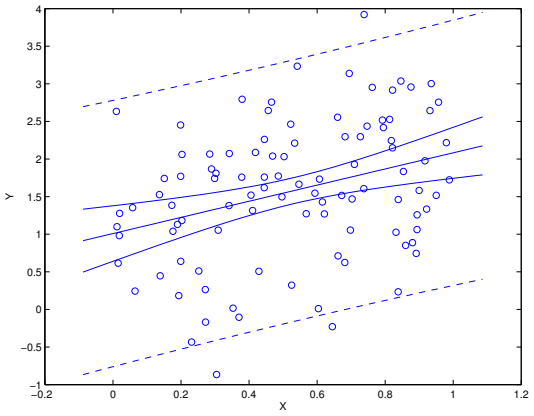

Fig. 1 – Graphique de la régression linéaire de l'exercice 3.1, comprenant la droite de regression, le nuage de points, et les régions de confiance et de prédiction (cette dernière étant la plus large).

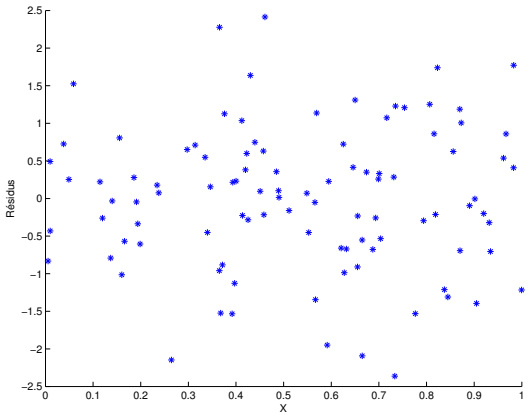

Fig. 2 – Résidus de la régression linéaire de l'exercice 3.1.

Exercice 3.2 (Régression linéaire simple). On s'intéresse au jeu de données numéro 13 de Stixbox. Il correspond à un échantillon de taille 14 sur deux variables aléatoires : le nombre de composants électriques en panne dans un ordinateur et la durée de la réparation en minutes. Visualisez les données avec `identify`. Effectuez une régression linéaire simple $\text{durée}_i = \beta_2 + \beta_1 \text{nombre}_i + \varepsilon_i$ au moyen de `linreg`. Tester l'hypothèse de nullité de β_1 à 5%. Idem pour β_2. Conclusion ? Que vous inspire le graphe des résidus ? Sont-ils corrélés ? Nous avons supposé que la durée de réparation sachant le nombre de composant en panne était gaussienne, cela vous semble-t-il plausible ? Peut-on effectuer la régression inverse ?

Exercice 3.3 (Régression linéaire multiple). On souhaite étudier la variation du taux d'hémoglobine dans le sang au cours d'une opération chirurgicale en fonction de la durée de l'opération et du volume de sang perdu pendant l'opération. On dispose des résultats suivants où y_i représente la valeur observée en pourcentage de la variation du taux d'hémoglobine, $x_{i,1}$ est la durée de l'opération en heures décimales et $x_{i,2}$ est le volume en litres de sang perdu.

y_i	-1.70	-4.61	-5.82	-1.17	-4.23	-3.31	+0.42	-2.98
$x_{i,1}$	1.75	1.33	1.43	1.86	1.81	1.66	1.60	2.00
$x_{i,2}$	0.52	0.59	0.61	0.50	0.54	0.49	0.27	0.47

On suppose que y_i est une réalisation d'une variable aléatoire Y_i de loi $\mathcal{N}(\beta_3+\beta_2 x_{i,2}+\beta_1 x_{i,1}, \sigma^2)$. Étudier cette régression linéaire multiple grâce à `lsfit`. Tester l'hypothèse suivant laquelle la variation du taux d'hémoglobine ne dépend ni de la durée de l'opération ni du volume de sang perdu ou encore l'hypothèse suivant laquelle la variation du taux d'hémoglobine ne dépend pas de la durée de l'opération.

Exercice 3.4 (Analyse de la variance à un facteur). L'analyse de la variance est la comparaison de k échantillons gaussiens indépendants de même variance. On se ramène à la formule de décomposition de la variance de la regression linéaire, cf. [4, Chap. 15.6] et [2, Chap. 5.1.7] par exemple.

Références

[1] P. Barbe et M. Ledoux – *Probabilités*, De la licence à l'agrégation, Belin, 1998.
[2] D. Dacunha-Castelle et M. Duflo – *Probabilités et statistiques. Tome 1*, Masson, Paris, 1982, Problèmes à temps ↓xe.
[3] X. Guyon – *Du modèle linéaire aux modèles non-linéaires*, Ellipses, 2001.
[4] G. Saporta – *Probabilités, analyse des données et statistique*, Technip, 1989.
[5] H. Scheffé – *The analysis of variance*, John-Wiley, 1999.
[6] P. Toulouse – *Thèmes de probabilités et statistique*, Dunod, 1999, INTERDIT À L'ORAL.
[7] B. Ycart – *Modèles et algorithmes markoviens*, SMAI - Mathématiques et Applications 39, Springer, 2002.

