

Option Probabilités et Statistiques  
Vecteurs gaussiens, modèles linéaires

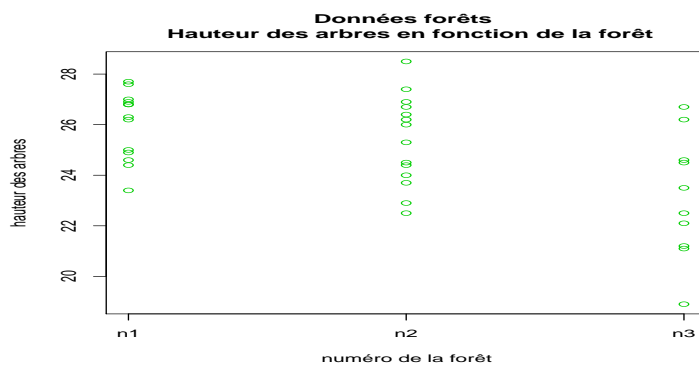
**Exercice 1 (Juste pour réviser)** On considère le modèle linéaire  $Y = M\beta + \epsilon$  où  $Y \in \mathbb{R}^n$ ,  $M$  est une matrice de  $\mathbb{R}^k$  dans  $\mathbb{R}^n$ ,  $\beta \in \mathbb{R}^k$  et les  $(\epsilon_i)_{1 \leq i \leq n}$  sont i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ .

1. Sous quelle(s) condition(s) sur  $M$ , l'estimateur de  $\beta$  des moindres carrés est unique ?
2. Lorsqu'il est unique, exprimer cet estimateur  $\hat{\beta}$  en fonction de  $X$  et  $Y$ .
3. Exprimer alors en fonction de  $M$  la valeur de  $\Gamma_{\hat{\beta}, \hat{\beta}}$ , matrice de variance-covariance de  $\hat{\beta}$ .
4. Donner une condition nécessaire et suffisante pour que les  $(\beta_i)_{1 \leq i \leq k}$  soient indépendants entre eux.
5. On suppose que  $\mathcal{M}M$  est une matrice diagonale de coefficients diagonaux  $(\lambda_i)_{1 \leq i \leq k}$ . Quelle est la loi de  $\sum_{i=1}^k \lambda_i (\hat{\beta}_i - \beta_i)^2 / \sigma^2$  ?

**Exercice 2 (Application du modèle ANOVA à un facteur (tiré du DDC1))**

Les données de moments qui vous sont données vous permettent de faire les calculs à la main en disposant de tables statistiques sur les loi de Students et Fisher. Ils peuvent aussi être traité à partir des données brutes (*ici*) sous matlab ou sous R

Le tableau suivant représente des mesures faites dans trois forêts de la hauteur de certains arbres.



	Forêt 1	Forêt 2	Forêt 3
$n_i$	13	14	10
$\sum Y_{ij}$	337.6	355.4	231.3
$\sum Y_{ij}^2$	8789.36	9062.96	5403.51

Dans la  $i^{\text{ième}}$  forêt,  $n_i$  arbres sont mesurés, de hauteur  $Y_{i1}, \dots, Y_{in_i}$ . On suppose que la hauteur des arbres se distribue selon le modèle

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

où  $\mu_i$  est déterministe, et les  $\epsilon_{ij}$  sont des variables aléatoires indépendantes normales centrées de variance  $\sigma^2$ .

1. Donner des estimateurs non biaisés des moyennes  $\mu_i$ .
2. Donner un estimateur sans biais de  $\sigma^2$ .
3. Donner un intervalle de confiance de niveau de confiance 0.99 pour chaque moyenne  $\mu_i$ .
4. Faire un test de niveau 0.05 pour l'égalité des trois moyennes.
5. Par un test de comparaison multiple au niveau global (FWER) 0.05, montrer que l'on peut rejeter  $\mu_1 = \mu_3$  et  $\mu_2 = \mu_3$ .

**Exercice 3 (Un détecteur élémentaire de changement de pente)** On dispose de  $2r + 1$  données  $r \geq 1$ , collectées à chaque pas de temps entre les instants  $-r$  et  $r$ . On cherche à détecter un changement de tendance au temps 0. Pour cela on considère le modèle linéaire suivant

$$Y = aJ + bK + \sigma\epsilon,$$

où  $\epsilon = (\epsilon_i)_{-r \leq i \leq r}$  est une suite de v.a.i.i.d. de loi  $\mathcal{N}(0, 1)$ ,

$$J = \begin{pmatrix} -r \\ -r+1 \\ \vdots \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \quad \text{et} \quad K = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ \vdots \\ r-1 \\ r \end{pmatrix}$$

Il s'agit alors de tester  $H_0 : a = b$  contre  $H_1 : a \neq b$ . On note  $E$  l'espace vectoriel engendré par  $J$  et  $K$  et  $E_0$  celui engendré par  $J + K$ .

1. Donner l'expression des estimateurs  $\hat{a}$  et  $\hat{b}$  aux moindres carrés. Sont-ils indépendants ?
2. Calculer la variance de  $\hat{a}$  et  $\hat{b}$  en fonction de  $r$  et  $\sigma^2$ .  
(On rappelle que  $\sum_{k=1}^n k^2 = n(n+1)(2n+1)/6$ .)
3. Donner l'expression de  $\hat{Y}_E$  projection orthogonale de  $Y$  sur  $E$ .
4. Donnez en fonction de  $\hat{a}$  et  $\hat{b}$  un estimateur noté  $S^2$  de  $\sigma^2$ . Quelle est la loi (en fonction de  $\sigma^2$ ) de  $S^2$  ?
5. Calculer la projection orthogonale  $\hat{Y}_{E_0}$  de  $Y$  sur  $E_0$  en fonction de  $\hat{a}$  et  $\hat{b}$  et donner une expression simplifiée de  $\hat{Y}_E - \hat{Y}_{E_0}$ .
6. Sous l'hypothèse  $H_0$ , quelle est loi de

$$\sqrt{\frac{r(r+1)(2r+1)}{12S^2}}(\hat{a} - \hat{b}) ?$$

(On justifiera clairement sa réponse).

7. En déduire un test de  $H_0$  contre  $H_1$  de niveau  $\alpha$ .

**Exercice 4 (Détection de valeurs aberrantes (et montagnes Ecossoises))** Dans un grand nombre de cas, pour améliorer l'adéquation d'un modèle linéaire avec des données particulières, il faut extraire certaines données aberrantes ou très singulières. Une façon d'aborder ce problème est d'examiner la valeur des résidus  $\hat{\epsilon}_i$  et de pointer les données pour lesquelles les résidus sont anormalement grands en valeur absolue. Cet exercice propose de quantifier cette notion de résidus "anormalement grand".

Soient  $k$  et  $n$  deux entiers tels que  $n > k+1$ . Soit  $M$  une matrice injective à coefficients réels à  $n$  lignes et  $k$  colonnes. On considère le modèle linéaire  $Y = M\beta + \epsilon$  où  $\epsilon \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$  et  $\beta \in \mathbb{R}^k$ . On note  $E = \{ M\beta' \mid \beta' \in \mathbb{R}^k \}$  et  $(e_1, \dots, e_n)$  la base canonique de  $\mathbb{R}^n$ . Pour tout  $1 \leq i \leq n$ , on suppose que  $e_i \notin E$  puis l'on note  $f_i = p_{E^\perp}(e_i)$  et  $F_i = E \oplus \mathbb{R}e_i$ . Enfin  $\hat{\epsilon}$  désigne le vecteur des résidus défini par  $\hat{\epsilon} = p_{E^\perp}(Y)$ .

1. Rappeler l'expression de l'estimateur  $\hat{\beta}$  de  $\beta$  aux moindres carrés ainsi que celle de l'estimateur usuel,  $S^2$ , de  $\sigma^2$ .
2. (a) Montrer que  $f_i \in F_i$  et que  $\hat{\epsilon}_i = \langle Y, f_i \rangle$ .  
 (b) Donner la loi de  $\hat{\epsilon}_i$  en fonction de  $\sigma^2$  et de  $f_i$ .  
 (c) Montrer précisément qu'il existe  $C_i \in \mathbb{R}_+$  tel que

$$\frac{C_i \hat{\epsilon}_i}{|p_{F_i^\perp}(Y)|}$$

ait une loi connue que l'on précisera.

- (d) Exprimer  $|p_{F_i^\perp}(Y)|^2$  en fonction de  $|p_{E^\perp}(Y)|^2$  et  $|p_{\mathbb{R}f_i}(Y)|^2$  puis  $|p_{\mathbb{R}f_i}(Y)|^2$  en fonction de  $|\hat{\epsilon}_i|^2$  et  $|f_i|^2$ . En déduire que le calcul de  $(C_i \hat{\epsilon}_i)/|p_{F_i^\perp}(Y)|$  ne nécessite que la connaissance de  $\hat{\epsilon}_i$ , de  $|f_i|$  et de  $S^2$ .
3. (a) Montrer que pour tout  $\alpha \in ]0, 1[$ , il existe  $s_\alpha$ , que l'on précisera, tel que, avec une probabilité au moins  $1 - \alpha$ , on a simultanément pour tout  $i \in \{1, \dots, n\}$

$$|\hat{\epsilon}_i| < s_\alpha \frac{|p_{F_i^\perp}(Y)|}{C_i}. \quad (0.1)$$

- (b) En montrant que  $|f_i| |p_{F_i^\perp}(Y)| < |p_{E^\perp}(Y)|$ , en déduire que pour tout  $\alpha \in ]0, 1[$ , on a simultanément pour tout  $1 \leq i \leq n$

$$|\hat{\epsilon}_i| < t_{n-k-1, \alpha/n} \sqrt{\frac{(n-k)S^2}{n-k-1}}. \quad (0.2)$$

où pour tout  $p \in \mathbb{N}^*$  et tout  $\eta \in ]0, 1/2[$ ,  $t_{p, \eta}$  est tel que  $P(|Z| > t_{p, \eta}) \leq \eta$  lorsque  $Z \sim t(p)$ .

4. Dans les données sur les courses ecossoises fournies par la table 1, on note  $Y$  le vecteur des temps,  $L$  et  $D$  respectivement le vecteur des longueurs et des dénivelés. On considère  $M$  la matrice  $(LD\mathbb{1})$  (3 colonnes qui sont  $L$ ,  $D$  et le vecteur dont toutes les coordonnées sur égale à 1). On analyse les données au travers du modèle linéaire  $Y = M\beta + \epsilon$  ou  $\beta \in \mathbb{R}^3$ . On obtient  $S^2 = 77532$  et les valeurs suivantes pour les résidus :

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$\hat{\epsilon}_i$	141	-455	-276	-52	-744	48	1875	-47	-71	255	261	68	-563	617	-517	-844	-236

18	19	20	21	22	24	24	25	26	27	28	29	30	31	32	33	34	35
-646	-757	43	178	-86	-298	272	143	-705	-163	-241	-15	-400	-972	-153	592	-16	3907

On donne  $t_{31,\alpha/n} = 3.501$  pour  $\alpha = 0.05$  et  $n = 35$ . Expliquer pourquoi les courses numéros 7 et 35 jouent un rôle particulier.

TAB. 1 – On reporte ici les données fournies par la “Scottish Hill Runners Association” sur les courses à pieds dans les montagnes Ecossoises. On dénombre 35 courses, pour lesquelles on donne la longueur (en miles), le dénivélé (en pieds) et le meilleur temps (en secondes).

num.	long.	déniv.	temps
1	2.5	650.	965.
2	6.	2500.	2901.
3	6.	900.	2019.
4	7.5	800.	2736.
5	8.	3070.	3736.
6	8.	2866.	4393.
7	16.	7500.	12277.
8	6.	800.	2182.
9	5.	800.	1785.
10	6.	650.	2385.
11	28.	2100.	11560.
12	5.	2000.	2583.
13	9.5	2200.	3900.
14	6.	500.	2648.
15	4.5	1500.	1616.
16	10.	3000.	4335.
17	14.	2200.	5905.
../..	../..	../..	../..

num.	long.	déniv.	temps
18	20.	5000.	9590.
19	4.5	1000.	1045.
20	5.5	600.	1954.
21	3.	300.	957.
22	3.5	1500.	1674.
23	6.	2200.	2859.
24	2.	900.	1076.
25	3.	600.	1121.
26	4.	2000.	1573.
27	6.	800.	2066.
28	5.	950.	1714.
29	6.5	1750.	3030.
30	5.	500.	1257.
31	10.	4400.	5135.
32	6.	600.	1943.
33	18.	5200.	10215.
34	4.5	850.	1686.
35	3.	350.	4719.

### Solution (Ex 1)

1.  $X$  doit être injectif.
2.  $\hat{\beta} = ({}^tXX)^{-1}{}^tXY$ .
3.  $\Gamma_{\hat{\beta},\hat{\beta}} = \sigma^2 = ({}^tXX)^{-1}$ .
4.  $\Gamma_{\hat{\beta},\hat{\beta}}$  diagonale.
5.  $\sum_{i=1}^n \lambda_i (\hat{\beta}_i - \beta_i)^2 / \sigma^2 = \sum_{i=1}^n (\hat{\beta}_i - \beta_i)^2 / V(\hat{\beta}_i) \sim \chi^2(k)$ .

### Solution (Ex 2)

1.  $\hat{m}_i = \bar{Y}_i$ . AN :  $\hat{m} = (25.969, 25.386, 23.130)$ .
2.  $S^2 = |Y - \hat{Y}|^2 / (n - 3) = (|Y|^2 - |\hat{Y}|^2) / (n - 3) = (\sum_{ij} Y_{ij}^2 - \sum_i (\sum_j Y_{ij})^2 / n_i) / (n - 3)$ .  
AN :  $S^2 = 3.4284$ .
3. On note  $E = \{y \in \mathbb{R}^n \mid \exists \mu = (\mu_1, \mu_2, \mu_3), y_{ij} = \mu_i, \forall 1 \leq i \leq 3, 1 \leq j \leq n_i\}$  et  $\hat{Y} = p_E(Y)$ . On a  $V(\hat{\mu}_i) = \sigma^2 / n_i$ . L'intervalle de confiance est construit sur la variable  $T = \frac{\sqrt{n_i}(\hat{\mu}_i - \mu_i)}{S} \sim t_{n-3} : [\hat{\mu}_i - St_{n-3, 1-\alpha/2} / \sqrt{n_i}, \hat{\mu}_i + St_{n-3, 1-\alpha/2} / \sqrt{n_i}]$ . AN :  $\mu_1 : [24.568, 27.370]$   $\mu_2 : [24.036, 26.736]$   $\mu_3 : [21.532, 24.728]$
4. Il s'agit d'un test de sous-modèle. La projection de  $Y$  sur l'espace de  $E_0 \doteq \mathbb{R}\mathbb{1}$ , donne  $P_{E_0}(Y) = \bar{Y}\mathbb{1}$ . Le test se fait sur la variable  $F = |P_E(Y) - p_{E_0}(Y)|^2 / (2S^2)$  qui suit une loi de Fisher  $F_{2, n-3}$ . AN :  $\bar{Y} = 24.981$  et  $F = (|P_E(Y)|^2 - |p_{E_0}(Y)|^2) / (2S^2) = (n_1\hat{\mu}_1^2 + n_2\hat{\mu}_2^2 + n_3\hat{\mu}_3^2 - n\bar{Y}^2) / (2S^2) = 7.1824$ . On calcule le seuil  $F_{2, n-3, 1-\alpha} = 3.2759$  et donc on rejette l'hypothèse d'égalité des trois moyennes. La  $p$ -valeur ( $P(F_{2, n-3} > F)$ ) vaut ici 0.002501.
5. On utilise la méthode de Bonferroni pour faire tous les tests  $H_{ij} : \mu_i = \mu_j$  pour  $i < j$  au niveau  $\alpha/K$  où  $K$  est le nombre de tests simultanés. Ici  $K = 3$ . De façon équivalente si  $p_{ij}$  est la  $p$ -valeur de  $H_{ij}$  calculer sur la statistique  $|T_{ij}|$  où  $T_{ij} \doteq \frac{\sqrt{\frac{n_i n_j}{n_i + n_j}} (\hat{\mu}_i - \hat{\mu}_j)}{S} \sim t_{n-3}$ , on rejete  $H_{ij}$  lorsque  $K\hat{p}_{ij} \geq \alpha$ . Ci-dessous les valeurs de  $K\hat{p}_{ij} \wedge 1$  pour les données forêts.

$i \setminus j$	foret2	foret 3
foret1	1.0000	0.0026
foret2	-	0.0175

### Solution (Ex 3)

1. On remarque que  $\langle J, K \rangle = 0$  d'où  $\hat{a} = \langle Y, J / |J|^2 \rangle$  et  $\hat{b} = \langle Y, K / |K|^2 \rangle$  sont deux estimateurs sans biais de  $a$  et  $b$  indépendants (d'après Cochran puisque  $Y \sim \mathcal{N}(aJ + bK, \sigma^2 I)$ ).
2. On a  $V(\hat{a}) = \sigma^2 / |J|^2$  et  $V(\hat{b}) = \sigma^2 / |K|^2$  avec  $|J|^2 = |K|^2 = \sum_{k=1}^r k^2 = r(r+1)(2r+1)/6$ .
3. On a  $\hat{Y}_E = \hat{a}J + \hat{b}K$ .
4.  $S^2 = |Y - \hat{Y}_E|^2 / (2r - 1) \sim \sigma^2 \frac{\chi_{2r-1}^2}{2r-1}$ .

5.  $\hat{Y}_{E_0} = \langle Y, \frac{J+K}{\sqrt{|J|^2+|K|^2}} \rangle \frac{J+K}{\sqrt{|J|^2+|K|^2}} = \langle Y, \frac{J+K}{2|J|^2} \rangle (J+K) = \frac{1}{2}(\hat{a}+\hat{b})(J+K)$ . Par suite  $\hat{Y}_E - \hat{Y}_{E_0} = (\hat{a}J + \hat{b}K) - \frac{1}{2}(\hat{a} + \hat{b})(J + K) = \frac{\hat{b}-\hat{a}}{2}(J + K)$ .
6. Comme  $\hat{a}$  et  $\hat{b}$  sont indépendants,  $(\hat{a}-\hat{b}) \sim \mathcal{N}(\mathbf{a}-\mathbf{b}, 2\sigma^2/|J|^2)$  et donc  $\sqrt{\frac{r(r+1)(2r+1)}{12\sigma^2}}(\hat{a}-\hat{b}) \sim \mathcal{N}(\mathbf{a}-\mathbf{b}, 1)$ . Sous  $H_0$ , comme  $\mathbf{a} = \mathbf{b}$ , la gaussienne est centrée. En estimant  $\sigma^2$  par  $S^2$  indépendant de  $\hat{Y}_E$ , on déduit facilement par Cochran que  $T \doteq \sqrt{\frac{r(r+1)(2r+1)}{12S^2}}(\hat{a}-\hat{b}) \sim t(2r-1)$ .
7. On construit un test de niveau  $\alpha$  par  $d = \mathbb{1}_{|T| \geq t_{2r-1, \alpha/2}}$  où  $P(t > t_{2r-1, \alpha/2}) = \alpha/2$  lorsque  $t \sim t(2r-1)$ .

### Solution (Ex 4)

1.  $\hat{\beta} = (MM^T)^{-1}MY$ ,  $S^2 = |Y - M\hat{\beta}|^2/(n-k)$ .
2. (a)  $f_i = e_i - p_E(e_i) \in E \oplus \mathbb{R}e_i = F_i$ . De plus  $\hat{e}_i = \langle p_{E^\perp}(Y), e_i \rangle = \langle Y, p_{E^\perp}(e_i) \rangle$  car les projections orthogonales sont auto-adjointes.  
 (b)  $\hat{e}_i \sim \mathcal{N}(0, \sigma^2|f_i|^2)$ .  
 (c)  $p_{F_i^\perp}(Y)$  et  $\hat{e}_i$  sont orthogonaux d'après Cochran. De plus  $|p_{F_i^\perp}(Y)|^2 \sim \sigma^2\chi^2(n - (k+1))$ . Par suite  $\sqrt{\frac{(n-(k+1))}{|f_i|^2}} \frac{\hat{e}_i}{|p_{F_i^\perp}(Y)|} \sim t(n - (k+1))$ . On pose donc  $C_i \doteq \sqrt{\frac{(n-(k+1))}{|f_i|^2}}$ .  
 (d) On a  $p_{F_i^\perp} + p_{\mathbb{R}f_i} = p_{E^\perp}$  qui est une décomposition en projecteurs orthogonaux. Par suite  $|p_{F_i^\perp}(Y)|^2 = |p_{E^\perp}(Y)|^2 - |p_{\mathbb{R}f_i}(Y)|^2$ . De plus  $|p_{\mathbb{R}f_i}(Y)|^2 = \langle Y, f_i \rangle^2 / |f_i|^2 = \frac{\hat{e}_i^2}{|f_i|^2}$ . Comme  $|p_{E^\perp}(Y)|^2 = (n_k)S^2$  on déduit le résultat.
3. Il s'agit de tests multiples. On peut appliquer la méthode de Bonferroni en ajustant le niveau de chaque test au niveau  $\alpha/n$  pour avoir un FWER  $\alpha$ .
4. Il suffit de remarquer que  $|f_i| \leq 1$  et que par Pythagore,  $|p_{F_i^\perp}(Y)|^2 \leq |p_{E^\perp}(Y)|^2$ . On a alors

$$\sqrt{\frac{(n-(k+1))}{|f_i|^2}} \frac{|\hat{e}_i|}{|p_{F_i^\perp}(Y)|} \geq \sqrt{n-(k+1)} \frac{|\hat{e}_i|}{|p_{E^\perp}(Y)|} \geq \sqrt{\frac{n-(k+1)}{n-k}} \frac{|\hat{e}_i|}{S}.$$

Par suite  $P(\sqrt{\frac{n-(k+1)}{n-k}} \frac{|\hat{e}_i|}{S} \geq t_{n-k-1, \alpha/n}) \leq P(\sqrt{\frac{n-(k+1)}{n-k}} \frac{|\hat{e}_i|}{S} \geq t_{n-k-1, \alpha/n}) = \alpha/n$ .

5. En mettant en oeuvre l'approche précédente, on a  $n = 35$ ,  $k = 3$  et pour  $\alpha = 0.05$ , on a  $t_{n-k-1, \alpha/n} \sqrt{\frac{(n-k)S^2}{n-k-1}} = 990.48$ . On voit que seule les courses 7 et 35 passent le seuil. Un coup d'oeil sur les données montre assez facilement que ces deux courses sont atypiques.