

### Exercices faits en Td de statistiques fondamentales

**Ex 1.** Soient  $\varepsilon_1, \dots, \varepsilon_n$  des variables iid  $\mathcal{N}(0, \sigma^2)$ . Lesquels des modèles suivants sont linéaires ? (dans tous les cas, les  $\beta_i$  sont réels et  $i \in \{1, \dots, n\}$ ). Quand cela est possible, proposer une transformation du modèle qui le rende linéaire.

1.  $Y_i = \beta_1 + \beta_2 x_i^2 + \varepsilon_i$ .
2.  $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i^2$ .
3.  $Y_i = e^{\beta_1} e^{\beta_2 x_i} e^{\varepsilon_i}$ .
4.  $Y_i = \beta_1 + \beta_2 e^{\beta_2 x_i} + \varepsilon_i$ .
5.  $Y_i = (\sum_{j=1}^r \beta_j x_{ij} + \varepsilon_i)^{1/3}$ .
6.  $Y_i = \beta_1 + \beta_2 x_i + \sum_{j=1}^i \varepsilon_j$ .

#### Modèles de régression

**Ex 2.** On considère le modèle linéaire

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

où les  $\varepsilon_i$  sont i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ .

1. Déterminer les estimateurs  $\hat{\beta}_0, \hat{\beta}_1$  et  $\hat{\sigma}^2$  au sens des moindres carrés de  $\beta_0, \beta_1$ , et  $\sigma^2$ .
2. Construire un intervalle de confiance au niveau  $1 - \alpha$  pour  $\beta_0 + \beta_1 u, u \in \mathbb{R}$  fixé.
3. Construire un intervalle de prévision :  $[\underline{T}, \bar{T}]$  tel que  $\mathbb{P}(\underline{T} \leq Y' \leq \bar{T}) = 1 - \alpha$ , avec  $Y' = \beta_0 + \beta_1 u + \varepsilon'$ , où  $\varepsilon'$  est une gaussienne  $\mathcal{N}(0, \sigma^2)$  indépendante de  $\varepsilon_1, \dots, \varepsilon_n$ .
4. Tester l'hypothèse  $H_0 : \beta_1 = 0$  contre  $H_1 : \beta_1 \neq 0$  au niveau 5%.
5. On définit un estimateur de  $\beta_1$  par

$$\hat{\beta}_1(\lambda) = \frac{\text{cov}(x, Y)}{\lambda + \text{var}(x)}$$

avec  $\text{cov}(x, Y) = \frac{1}{n-1} (\sum_{i=1}^n x_i Y_i - n \bar{x}_n \bar{Y}_n)$  et  $\text{var}(x) = \frac{1}{n-1} (\sum_{i=1}^n x_i^2 - n \bar{x}_n^2)$ . Déterminer la valeur de  $\lambda$  qui minimise le risque, à  $\beta_0, \beta_1, \sigma^2$  fixés.

6. On note  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  et  $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ . On définit  $R^2$ , le coefficient de détermination de la régression de  $(y_1, \dots, y_n)$  sur  $(x_1, \dots, x_n)$ , par

$$R^2(x, Y) = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}$$

Montrer que

$$R^2(x, Y) = 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}.$$

7. Soit  $R'^2$  le coefficient de détermination associé à la régression de  $(x_1, \dots, x_n)$  sur  $(y_1, \dots, y_n)$ . Montrer que  $R^2(x, y) R'^2(y, x) = \rho^4$ , avec  $\rho$  le coefficient de corrélation linéaire empirique défini par

$$\rho = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \text{var}(y)}}$$

Application numérique : On observe :

$x_i$	1	4	5	9	11	13	23	23	28
$y_i$	64	71	54	81	76	93	77	95	109

où  $x$  est la quantité de phosphore répandu sur un champ et  $Y$  la quantité de phosphore retrouvé dans la plante.  $u = 25$ ,  $\alpha = 5\%$ .

**Ex 3.** On considère le modèle linéaire

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

où les  $\varepsilon_i$  sont i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ . On observe :

$i$	1	2	3	4	5	6	7	8	9	10	11	12
$x_{1,i}$	4	4	4	8	8	8	12	12	12	16	16	16
$x_{2,i}$	2	4	6	2	4	6	2	4	6	2	4	6
$Y_i$	61	58	50	65	66	60	74	70	68	85	84	79

1. Calculer les estimateurs  $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$  des moindres carrés de  $(\beta_0, \beta_1, \beta_2, \sigma^2)$ .
2. Donner les variances de ces estimateurs et leurs intervalles de confiance.
3. Notons  $\beta' = (\hat{\beta}_0 - \beta_0, \hat{\beta}_1 - \beta_1, \hat{\beta}_2 - \beta_2)$ , et

$${}^t X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 4 & 8 & 8 & 8 & 12 & 12 & 12 & 16 & 16 & 16 & 16 \\ 2 & 4 & 6 & 2 & 4 & 6 & 2 & 4 & 6 & 2 & 4 & 6 & 6 \end{pmatrix}$$

Montrer que  $\beta' {}^t X X {}^t \beta'$  suit une loi du  $\chi_2$  à trois degrés de liberté (on pourra d'abord supposer que  ${}^t X X$  est diagonale). En déduire un ellipsoïde de confiance pour  $(\beta_0, \beta_1, \beta_2)$ .

4. On suppose qu'en construisant le modèle on a omis de de considérer l'influence du troisième facteur. Montrer qu'alors les estimateurs des moindres carrés de  $\beta_0$  et  $\beta_1$  sont biaisés.

### Modèles d'analyse de la variance

**Ex 4.** On observe des rendements en livre par acre pour trois concentrations différentes d'engrais :

Concentrations					
0	794	1800	576	411	
500	2012	2477	3498	2092	1808
1000	2118	1947	3361		

On note  $Y_{i,j}$  le rendement observé lors la  $j$ ème expérience avec la  $i$ ème concentration ;  $j = 1, \dots, 5$ , et  $i = 1, 2, 3$ . On considère le modèle d'analyse de la variance à un facteur

$$Y_{i,j} = \lambda_i + \varepsilon_{i,j}$$

avec  $\varepsilon_{i,j}$  i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ .

1. Déterminer les estimateurs des moindres carrés de  $(\lambda_1, \lambda_2, \lambda_3)$  et de  $\sigma^2$  à l'aide des formules matricielles.
2. On pose

$$\varphi(\lambda_1, \lambda_2, \lambda_3) = \sum_{i,j} (Y_{i,j} - \lambda_i)^2$$

Déterminer l'unique point critique de  $\varphi$ . En déduire les estimateurs des moindres carrés de  $(\lambda_1, \lambda_2, \lambda_3)$ .

3. Autre méthode : on écrit le modèle sous la forme

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$

avec  $\sum_{i=1}^3 n_i \alpha_i = 0$  et  $n_1 = 4$ ,  $n_2 = 5$  et  $n_3 = 3$ .

(a) Montrer que le modèle est identifiable.

Soit  $(\mu', \alpha'_1, \alpha'_2, \alpha'_3) \in \mathbb{R}^4$  tel que  $\sum_{i=1}^3 n_i \alpha'_i = 0$ . On pose

$$Y_{..} = \frac{1}{12} \sum_{i,j} Y_{i,j}$$

$$Y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$$

Montrer qu'on a l'égalité, dite équation normale :

$$\frac{1}{n_1 + n_2 + n_3} \sum_{i,j} (Y_{i,j} - \mu' - \alpha'_i)^2 = (Y_{..} - \mu')^2 + \sum_i \frac{1}{n_i} (Y_{i.} - Y_{..} - \alpha'_i)^2 + \frac{1}{n_1 + n_2 + n_3} \sum_{i,j} (Y_{i,j} - Y_{i.})^2$$

En déduire les estimateurs des moindres carrés de  $\mu, \alpha_1, \alpha_2, \alpha_3$  et  $\sigma^2$ , puis ceux de  $\lambda_1, \lambda_2$  et  $\lambda_3$ .

- On souhaite tester l'hypothèse ( $H_0$ ) d'égalité des rendements moyens pour les trois concentrations, au niveau  $\alpha = 0,01$ . Montrer que sous cette hypothèse  $Y$  vérifie un modèle linéaire dont on estimera les paramètres. En déduire une formule explicite pour le test de Fisher associé au problème de test posé.
- Même question, avec l'hypothèse d'égalité des traitements "500" et "1000".

**Ex 5.** On observe des pressions sanguines sur des hommes appartenant à trois groupes socio-économiques notés I, II et III, et suivant 3 tranches d'âges :

Groupes	30-45				46-59				60-75			
I	128	104	132	112	129	136	174	166	214	146	138	148
II	136	124	112	118	138	124	160	157	156	110	188	158
III	116	108	160	116	108	110	154	122	182	148	138	136

- Proposer un modèle linéaire additif d'analyse de la variance à deux facteurs. Etablir l'équation normale associée. En déduire les estimateurs des moindres carrés des paramètres.
- Tester au niveau 10% s'il y a des différences entre les trois groupes socio-économiques.
- Même question sur les trois tranches d'âges.
- Donner un intervalle de confiance au niveau 90% pour l'écart entre les influences des groupes I et III.
- Même question pour l'écart entre les influences des tranches d'âges 30 - 45 et 60 - 75.

**Ex 6.** Etude des doses croissantes d'azote sur un riz irrigué.

Il y a quatre traitements :

T : N=0

A : N=30 unités à l'hectare

B : N=45 unités à l'hectare

C : N=60 unités à l'hectare

La disposition des parcelles et les résultats (rendements en quintaux par ha) sont indiqués ci-dessous :

C ; 40,3	A ; 29,6	T ; 20,3	B ; 40,4
B ; 37,6	T ; 19,2	C ; 42,6	A ; 34,7
A ; 32,4	C ; 39,7	B ; 37,2	T ; 18,0
T ; 19,4	B ; 40,4	A ; 34,5	C ; 43,8

Ce dispositif expérimental, dit carré latin, élimine la plus grande partie de l'hétérogénéité due au sol. Les quatre traitements sont répartis en raison d'un par ligne et un par colonne, cette condition ayant été remplie strictement au hasard.

1. Proposer un modèle linéaire additif d'analyse de la variance à trois facteurs. Par la méthode du point critique ou de l'équation normale, déterminer les estimateurs des moindres carrés des paramètres.
2. Tester s'il y a des différences significatives entre les traitements. Le cas échéant, comparer les traitements deux à deux.

### Modèle d'analyse de la covariance

**Ex 7.** En septembre 2001, à la recherche d'un deux-pièces au centre de Paris, un quidam observe les prix suivants dans un journal d'annonces :

surface en $m^2$	43	45	40	30	42	54	24	25	46	47	35	40	49	36
prix en kF	1250	1395	1400	815	1620	2100	772	900	1590	1850	1290	1590	1650	1250
arrondissement	1	1	1	6	6	6	6	6	6	7	7	7	7	7

1. Un appartement est proposé à 840 kF dans le sixième arrondissement. Sa surface n'est pas précisée. Construire un intervalle de prévision à 95% pour cette surface à l'aide d'un modèle de régression à une variable.
2. Le quidam a le sentiment que les prix ne diffèrent pas sensiblement entre ces arrondissements voisins. Sans tenir compte des surfaces, réaliser un test de cette hypothèse à partir d'une analyse de la variance à un facteur.
3. Peu convaincu, le quidam souhaite que le test prenne en considération les superficies des appartements. Proposer alors un modèle linéaire, dit modèle d'analyse de la covariance, qui tienne compte de tous les facteurs, et résoudre formellement le problème de test posé.

### Exercices faits en Td de statistique des processus, légèrement récrits

**Ex 8.** On note

$$C_0 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad C_1 = \begin{pmatrix} 1875 \\ 1876 \\ \vdots \\ 1972 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 1875^2 \\ 1876^2 \\ \vdots \\ 1972^2 \end{pmatrix},$$

$$W = \text{Vect}(C_0, C_1, C_2), \quad V = \text{Vect}(C_0, C_1), \quad A = (C_0, C_1, C_2), \quad \Gamma = ({}^t A \cdot A)^{-1} = \begin{pmatrix} \Gamma_{1,1} & \Gamma_{1,2} & \Gamma_{1,3} \\ \Gamma_{2,1} & \Gamma_{2,2} & \Gamma_{2,3} \\ \Gamma_{3,1} & \Gamma_{3,2} & \Gamma_{3,3} \end{pmatrix}$$

On considère le modèle

$$\forall t \in \{1875, \dots, 1972\} \quad X_t = \alpha + \beta t + \gamma t^2 + \varepsilon_t,$$

avec  $\varepsilon_{1875}, \dots, \varepsilon_{1972}$  variables gaussiennes indépendantes de loi  $\mathcal{N}(0, \sigma^2)$ .

1. On souhaite tester au niveau 5%  $H_0 : \gamma = 0$  contre  $H_1 : \gamma \neq 0$ . Déterminer la statistique de Fisher  $F$  associée à ce problème de test d'hypothèses linéaires, et préciser la zone de rejet.

2. On sait que le vecteur aléatoire  $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\gamma} \end{pmatrix}$  suit la loi  $\mathcal{N}\left(\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}, \sigma^2 \Gamma\right)$ , et qu'il est indépendant de

$\frac{95\hat{\sigma}^2}{\sigma^2}$  qui suit une loi du  $\chi^2$  à 95 degrés de liberté. On en déduit que  $st = \frac{\hat{\gamma}}{\hat{\sigma} \Gamma_{3,3}}$  suit sous  $H_0$  une loi de Student à 95 degrés de liberté. En déduire un test de  $H_0$  contre  $H_1$  au niveau 5%. Montrer que  $st^2 = F$  et qu'il s'agit en fait du même test qu'à la question précédente. Indications :

- (a) Soient  $C_2^{\parallel} \in V$  et  $C_2^{\perp} \in V^{\perp}$  tels que  $C_2 = C_2^{\parallel} + C_2^{\perp}$ . Montrer que  $\text{proj}_W^{\perp}(X) - \text{proj}_V^{\perp}(X) = \hat{\gamma} C_2^{\perp}$  (remarquer que  $\text{proj}_V^{\perp} = \text{proj}_W^{\perp} \circ \text{proj}_V^{\perp}$ ).

(b) Montrer que  $\langle X, C_2^\perp \rangle = \langle \hat{X}, C_2^\perp \rangle$  (on rappelle qu'un projecteur orthogonal est auto-adjoint). En déduire que  $\text{var}(\hat{\gamma}) = \frac{\sigma^2}{\|C_2^\perp\|^2}$ .

3. L'application numérique donne  $st = 4,67$ . Quel est le degré de signification de cette valeur relativement à la famille de tests définie précédemment ?

**Ex 9.** Dans l'exercice précédent, on suppose que  $X_{1900}$  ne vérifie pas le modèle (i.e. il s'agit d'une valeur aberrante). Montrer que si on ne l'ôte pas des calculs, l'estimateur de  $\beta$  est biaisé, et donner une interprétation géométrique du biais.

**Ex 10.** Modèle linéaire avec bruit corrélé. Soit  $X = (X_t, t = 1, \dots, n)$  une famille de variables aléatoires. On suppose que  $X$  vérifie le modèle  $(*) : X = A\beta + \varepsilon$  avec  $A$  matrice de dimensions  $n \times p$  et de rang  $p$ ,  $\varepsilon$  vecteur aléatoire centré de matrice de covariance  $\sigma^2\Gamma$ , avec  $\Gamma$  connue, et  $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*$  paramètres.

1. Montrer que  $\Gamma$  est une matrice symétrique positive.
2. On suppose désormais  $\Gamma$  inversible. Déterminer une matrice orthogonale  $O$  telle que  $O\varepsilon$  soit un vecteur dont les composantes sont indépendantes.
3. En déduire une transformation du modèle  $(*)$  en modèle linéaire ordinaire, puis déterminer l'estimateur des moindres carrés de  $\beta$  correspondant et sa matrice de covariance.
4. On suppose  $\varepsilon$  gaussien. Déterminer la fonction de vraisemblance du modèle  $(*)$ , puis l'estimateur du maximum de vraisemblance de  $\beta$ .

#### Exercices qui n'étaient pas donnés, et qui doivent dater du poly de Christine Graffigne

**Ex 11.** Soit  $Y = (Y_1 \dots Y_n)'$  un vecteur aléatoire gaussien de matrice de covariance égale à  $\sigma^2 I_n$ ,  $\sigma^2 > 0$ ; on suppose que, pour tout  $i = 1 \dots n$ ,  $E(Y_i) = a + bx_i + cx_i^2$ , où  $a, b$  et  $c$  sont des paramètres réels et  $x_1 \dots x_n$  des réels tous distincts tels que  $\sum_{i=1}^n x_i = 0$  et  $\sum_{i=1}^n x_i^3 = 0$ . On pose  $\tau = \frac{1}{n} \sum_{i=1}^n x_i^2$  et  $\nu = \frac{1}{n} \sum_{i=1}^n (x_i^2 - \tau)^2$ .

1. Montrer que, si on pose  $\lambda = a + c\tau$  et  $\beta = (\lambda, b, c)'$ , on peut écrire  $E(Y) = X\beta$  où  $X$  est une matrice que l'on déterminera. Calculer  $X'X$  et en déduire les estimateurs des moindres carrés de  $a, b$  et  $c$ .
2. On suppose  $\sigma^2$  connu, donner un intervalle de confiance à  $(1 - \alpha)$  de  $c$ .
3. On suppose  $c = 0$ , trouver les estimateurs des moindres carrés de  $a$  et  $b$ .
4.  $\sigma^2$  étant inconnu, montrer que, pour le test d'hypothèse  $H_0 : c = 0$  contre  $H_1 : c \neq 0$ , la statistique du test de Fisher est égale à

$$\frac{(\sum_{i=1}^n (x_i^2 - \tau)Y_i)^2}{n\nu\hat{\sigma}^2}$$

où  $\hat{\sigma}^2$  est l'estimateur du maximum de vraisemblance sans biais de  $\sigma^2$ . Sous l'hypothèse  $H_0$ , donner la loi de cette statistique.

**Ex 12.** On considère un modèle d'analyse de la variance à deux facteurs, avec interactions.

1. Donner l'estimateur UVMB de  $\gamma_{ij}$  et un intervalle de confiance de  $\gamma_{ij}$  basé sur cet estimateur.
2. Montrer que la statistique de Fisher pour tester  $\gamma_{ij} = 0, \forall i, j$ , est

$$\frac{(n - pb)c \sum_{i=1}^p \sum_{j=1}^b (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2}{(b - 1)(p - 1) \sum_{i=1}^p \sum_{j=1}^b (Y_{ijk} - Y_{ij.})^2}$$

3.  $\beta_i - \beta_j$  est souvent utilisé pour comparer les facteurs  $i$  et  $j$ . Montrer que  $Y_{i..} - Y_{j..}$  est l'estimateur UVMB pour  $\beta_i - \beta_j$  et qu'un intervalle de confiance au niveau  $(1 - \alpha)$  pour  $\beta_i - \beta_j$  est donné par :

$$Y_{i..} - Y_{j..} \pm St\sqrt{2/cb} \text{ où } S^2 = \frac{1}{n - bp} \sum_i \sum_j \sum_k (Y_{ijk} - Y_{ij.})^2$$

et  $t$  est lu dans la table de la loi de Student.

4. Pour  $p = b = 2$ ,  $c = 4$ ,  $\beta_{11} = 0$ ,  $\beta_{21} = 1$ ,  $\beta_{12} = 3$ ,  $\beta_{22} = 1$  et  $\sigma^2 = 1$ , on observe :

-1.33	1.28	0.62	0.70	4.39	4.23	3.76	2.14
1.10	-0.38	0.43	1.10	0.22	1.04	3.61	0.58

Estimer  $\alpha_i$ ,  $\lambda_j$  et  $\gamma_{ij}$ ,  $i = 1, 2$  et  $j = 1, 2$ . On suppose  $\gamma_{ij} = 0$ , pour  $i = 1, 2$  et  $j = 1, 2$ , tester  $\alpha_1 = \alpha_2 = 0$ . Dans le cas général, tester  $\gamma_{ij} = 0$ ,  $i = 1, 2$  et  $j = 1, 2$ .