# 4, 2. Problèmes de régression.

4, 12. Soit $(\mathbf{x}_1, \ldots, \mathbf{x}_q)$ un échantillon simple et fortuit d'une population normale de moyenne $\mu_1$ et écart-type $\sigma$, $(\mathbf{x}_{q+1}, \ldots, \mathbf{x}_n)$ un échantillon simple et fortuit d'une population normale de moyenne $\mu_2$ et écart-type $\sigma$, les deux échantillons étant mutuellement indépendants. La théorie des modèles linéaires s'applique encore:

$$\mathbf{E} \left\| \begin{matrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_q \\ \mathbf{x}_{q+1} \\ \vdots \\ \mathbf{x}_n \end{matrix} \right\| = \left\| \begin{matrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{matrix} \right\| \quad \left\| \begin{matrix} \mu_1 \\ \mu_2 \end{matrix} \right\| ,$$

$$r = p = 2 , \qquad \mathfrak{b}_H^T = \| \mu_1, \mu_2 \| , \qquad \mathfrak{A}^T \mathfrak{A} = \left\| \begin{matrix} q & 0 \\ 0 & n - q \end{matrix} \right\| .$$

Si on pose

$$\sum_1^q \mathbf{x}_i^s = \mathbf{S}_{1,s} , \qquad \sum_{q+1}^n \mathbf{x}_i^s = \mathbf{S}_{2,s} ,$$

on a

$$\hat{\mu}_1 = \mathbf{S}_{1,1}/q , \qquad \hat{\mu}_2 = \mathbf{S}_{2,1}/(n - q) ,$$

$$\mathbf{SC}E = \frac{q\,\mathbf{S}_{1,2} - (\mathbf{S}_{1,1})^2}{q} + \frac{(n-q)\,\mathbf{S}_{2,2} - (\mathbf{S}_{2,1})^2}{n-q} ,$$

de sorte que, sous l'hypothèse $\mu_1 - \mu_2 = \Delta$, l'expression

$$\frac{(\hat{\mu}_1 - \hat{\mu}_2 - \Delta)\sqrt{(n-2)}}{\sqrt{\left[\left(\frac{1}{q} + \frac{1}{n-q}\right) \mathbf{SC}E\right]}}$$

est une aléatoire $\mathbf{t}_{n-2}$; $\mathbf{SC}E/\sigma^2$ est une aléatoire $\chi^2_{n-2}$.

## 4, 2. *Problèmes de régression.*

4, 211. Supposons que, $u_1, \ldots, u_s$ étant des constantes certaines deux à deux distinctes, on ait $n = \Sigma_1^s k_i$ variables aléatoires $\mathbf{x}_{i,j}$ $(i = 1, \ldots, s;\ j = 1, \ldots, k_i)$, normales, de variance commune $\sigma^2$, indépendantes, avec

$$\mathbf{E}\,\mathbf{x}_{i,j} = \alpha + \beta\, u_i . \tag{17}$$

Ce cas rentre dans le cadre général des modèles linéaires moyennant

$$\mathfrak{b}_H = \| \alpha \quad \beta \|^T, \quad \mathfrak{A} = \begin{Vmatrix} 1 & \dots & 1 & 1 & \dots & 1 & \dots & 1 & \dots & 1 \\ u_1 & \dots & u_1 & u_2 & \dots & u_2 & \dots & u_s & \dots & u_s \end{Vmatrix}^T$$

($\mathfrak{A}$ composée de $s$ groupes ayant respectivement $k_1, k_2, \dots, k_s$ lignes identiques entre elles); ici, $r = p = 2$. Si l'on procède comme au § 3, 232, en posant

$$\sum_{i=1}^{s} \sum_{j=1}^{k_i} u_i^q \, \mathbf{x}_{i,j}^t = \mathbf{S}_{q,t} \; ; \quad S_{0,0} \equiv n \; ; \quad S_{q,0} = \sum_1^i k_i \, u_i^q \; ;$$

$$L_{u,u} = \frac{n\, S_{2,0} - S_{1,0}^2}{n} \; ; \quad \mathbf{L}_{u,x} = \frac{n\, \mathbf{S}_{1,1} - S_{1,0}\, \mathbf{S}_{0,1}}{n} \; ;$$

on a, successivement,

$$\mathfrak{G} = \begin{Vmatrix} n & S_{1,0} \\ S_{1,0} & S_{2,0} \end{Vmatrix}$$

$$\mathfrak{S} = \begin{Vmatrix} \sqrt{n} & S_{1,0}/\sqrt{n} \\ 0 & \sqrt{L_{u,u}} \end{Vmatrix} \qquad \mathfrak{A}^T \mathfrak{x} = \begin{Vmatrix} \mathbf{S}_{0,1} \\ \mathbf{S}_{1,1} \end{Vmatrix} \qquad \hat{\mathfrak{b}}_K = \begin{Vmatrix} \mathbf{S}_{0,1}/\sqrt{n} \\ \mathbf{L}_{u,x} \end{Vmatrix}$$

$$\hat{\mathfrak{b}}_H^T = \begin{Vmatrix} \dfrac{\mathbf{S}_{0,1}}{n} - \dfrac{S_{1,0}}{n} \dfrac{\mathbf{L}_{u,x}}{L_{u,u}} & \dfrac{\mathbf{L}_{u,x}}{L_{u,u}} \end{Vmatrix}$$

d'où la table d'analyse de variance:

$$\begin{array}{lll} SCT = S_{0,2} & & n\; d.l. \\ & \left\{ \begin{array}{l} \text{red}\,[\alpha] = (S_{0,1})^2/n \\ \text{red}\,[\beta \mid \alpha] = (L_{u,x})^2 \end{array} \right. & \left\{ \begin{array}{l} 1\; d.l. \\ 1\; d.l. \end{array} \right. \\ SCN = \text{red}\,[\alpha] + \text{red}\,[\beta \mid \alpha] & & 2\; d.l. \\ SCE = SCT - SCN & & (n-2)\; d.l. \end{array}$$

4, 212. On peut traiter ce même problème d'une manière un peu différente, en posant

$$u_i' = u_i - S_{1,0}/n \; ,$$

$$\alpha + \beta\, u_i \equiv \alpha' + \beta'\, u_i' \qquad (\beta' = \beta \; , \quad \alpha' = \alpha + \beta\, S_{1,0}/n) \; ;$$

ceci revient à changer de base dans $\mathsf{B}$, et nous écrirons

$$\mathfrak{b}_{H'} = \| \alpha' \; \beta' \| \, .$$

On a alors, en marquant de l'apostrophe les expressions propres à la forme actuelle du modèle considéré,

$$\mathfrak{A}' = \left\| \begin{matrix} 1 & \dots & 1 & 1 & \dots & 1 & \dots & 1 & \dots & 1 \\ u_1' & \dots & u_1' & u_2' & \dots & u_2' & \dots & u_s' & \dots & u_s' \end{matrix} \right\|^T ,$$

$$S_{1,0}' = 0 \ , \quad S_{2,0}' = L_{u,u} \ , \qquad \mathbf{S}_{1,1}' = \mathbf{L}_{u,x} \ ,$$

$$\mathfrak{G}' = \mathrm{diag}\,(n, L_{u,u}) \ ; \quad \mathfrak{A}^T \mathfrak{x} = \| \ \mathbf{S}_{0,1} \quad \mathbf{L}_{u,x} \ \|^T$$

$$\hat{\alpha}' = \mathbf{S}_{0,1}/n \qquad \hat{\beta}' = \mathbf{L}_{u,x}/L_{u,u} \ (= \hat{\beta}) \ .$$

La table d'analyse de la variance ne change évidemment pas.

La méthode du § 3, 232 constitue, en quelque sorte, une orthogonalisation a posteriori: on part de $\mathfrak{b}_H$, les vecteurs $\varepsilon_i^T \hat{\mathfrak{b}}_H$ ($\in \mathbf{V}^\star$) ne sont pas orthogonaux, mais les calculs introduisent d'eux-même une base $\mathfrak{K}$ telle que les vecteurs $\varepsilon_i^T \hat{\mathfrak{b}}_K$ soient orthogonaux. Ici, nous venons de procéder à une orthogonalisation a priori: nous avons d'emblée introduit une base $\mathfrak{H}'$ telle que la matrice $\mathfrak{G}'$ relative à cette base soit diagonale, ce qui garantit l'orthogonalité des vecteurs $\varepsilon_i^T \hat{\mathfrak{b}}_{H'}$. Cette seconde méthode est souvent préférable à la première. C'est sur elle que reposent, notamment, les procédés de « codage linéaire » utilisés, dans les manuels d'analyse statistique, pour l'étude des plans factoriels à facteurs quantitatifs (plans factoriels « de régression »).

4, 213. Il arrive que l'on désire contrôler, par les observations elles-mêmes, la validité de la relation (17). Le modèle basé sur (17) est alors considéré comme un cas particulier du modèle défini par

$$\mathbf{E}\,\mathbf{x}_{i,j} = M_i \ ;$$

dans ce modèle plus général, l'espace des erreurs, $\mathbf{V}_*$, est engendré par les fonctionnelles de $\mathfrak{x}$ qui sont de la forme $(x_{i,j} - x_{i,k})$; il admet donc la base suivante:

$$x_{i,1} - x_{i,2} \, , \, \dots, \, x_{i,1} - x_{i,k_i} \ , \qquad i = 1, 2, \dots, s \ ,$$

laquelle s'orthogonalise en

$$\begin{cases} \mathfrak{d}_{i,t}^\star \mathfrak{x} = x_{i,1} + \dots + x_{i,t-1} - (t-1)\, x_{i,t} \\ \qquad i = 1, \dots, s; \quad t = 2, \dots, k_i \ . \end{cases}$$

On a alors

$$SCint \equiv SC\,\mathsf{V}_* = \sum_{i=1}^{s} \sum_{t=2}^{k_i} SC\{\mathfrak{d}^\star_{i,t}\}$$

$$= \sum_{i=1}^{s} \sum_{t=2}^{k_i} (\mathfrak{d}^\star_{i,t}\,\mathfrak{x})^2/(t-1)\,t$$

$$= \sum_{i=1}^{s} \frac{k_i \sum_{j=1}^{k_i} x_{i,j}^2 - \left(\sum_{j=1}^{k_i} x_{i,j}\right)^2}{k_i} \tag{18}$$

avec $\Sigma_i\,(k_i - 1) = n - s$ degrés de liberté. Le contrôle envisagé n'est donc possible que si l'un au moins des entiers $k_i$ est $> 1$, et il ne présente, en pratique, quelque intérêt que si $n - s$ est, au moins, de l'ordre de $s$. On le fait alors en comparant $SCEM$ à $SCint$ au moyen des tables de **F**. Si l'on procède ainsi, il sied d'utiliser $SCint$, et non $SCE$, comme dénominateur des divers $F$ calculés.

**Remarque.** — Le calcul qui a conduit à l'expression (18) de $SCint$ est valide dans des conditions extrêmement générales.

4, 214. On peut évidemment éprouver des hypothèses très diverses relativement à $\alpha$ et $\beta$ (ou, ce qui revient au même, à $\alpha'$ et $\beta'$) [14]. Ainsi, l'on pourrait éprouver l'hypothèse $\beta = a$, $a$ étant un nombre donné; il suffit d'appliquer la formule du § 2, 23, en remplaçant, au besoin, $SCE$ et $(n - r)$ par $SCint$ et $(n - s)$. Le seul point un peu délicat est le calcul de $\mathfrak{l}^\star\,\mathfrak{l}$; or, on a

$$\hat{\beta} = \|\, l_{1,1}, \ldots, l_{1,k_1}, \ldots, l_{s,k_s} \,\|\, \mathfrak{x} \equiv \mathfrak{l}^\star\,\mathfrak{x}$$

moyennant

$$l_{i,j} = (n\,u_i - S_{1,0})/n\,L_{u,u}\ ;$$

on a donc

$$\mathfrak{l}^\star\,\mathfrak{l} = \textstyle\sum_i \sum_j l_{i,j}^2 = 1/L_{u,u}\ ;$$

par conséquent, sous l'hypothèse $\beta = a$, l'expression

$$\frac{(\hat{\beta} - a)\sqrt{(n-s)}}{\sqrt{(\mathfrak{l}^\star\,\mathfrak{l}.\ SC\,int)}} = (\hat{\beta} - a)\sqrt{\frac{(n-s)\,L_{u,u}}{SC\,int}}$$

est une valeur observée d'une aléatoire $\mathbf{t}_{n-s}$ .

On éprouverait de même, par exemple, l'hypothèse que, pour des valeurs données $u_0$ et $x_0$, on a $\alpha + \beta u_0 = x_0$ (on considérerait l'expression $\hat{\alpha} + \hat{\beta}\, u_0$, qui, sous cette hypothèse, a comme moyenne $x_0$).

4, 22. Supposons que, $u$ et $v$ étant deux variables certaines, on ait

$$\mathbf{E}\, \mathbf{x}_{u,v} = \beta_0 + \beta_1 u + \beta_2 v \ , \tag{19}$$

et que les observations aient été faites aux « points » (0, 0), (2, 0), (2, 1), (1, 2), (0, 2) et (1, 1); les observations sont, ici encore, censées être des valeurs observées d'aléatoires normales, indépendantes, de même variance $\sigma^2$. La théorie générale s'applique alors, avec

$$\mathfrak{b}_H = \| \beta_0, \beta_1, \beta_2 \|^T \ , \quad \mathfrak{A}_H = \left\| \begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 & 2 & 1 \end{matrix} \right\|^T , \quad n = 6 \ , \quad p = r = 3 \ .$$

Il est commode de traiter ce problème par orthogonalisation à priori; on rapporte donc $\mathsf{B}$ à une base $\mathfrak{K}$ telle que les colonnes de $\mathfrak{A}_K$ soient deux à deux orthogonales; si l'on pose $\mathfrak{b}_K = \| \gamma_0, \gamma_1, \gamma_2 \|^T$, cela revient à chercher deux polynômes du premier degré, $\varphi(u)$ et $\psi(u, v)$, tels que

$$\gamma_0 + \gamma_1 \varphi(u) + \gamma_2 \psi(u, v) \equiv \beta_0 + \beta_1 u + \beta_2 v \ ,$$

$$\sum_1^6 \varphi(u_i) = 0 \ , \quad \sum_1^6 \psi(u_i, v_i) = 0 \ , \quad \sum_1^6 \varphi(u_i)\, \psi(u_i, v_i) = 0 \ .$$

On peut prendre

$$\varphi(u) = u - 1 \ , \quad \psi(u, v) = -5 + u + 4v \ ,$$

ce qui correspond à

$$\mathfrak{b}_H = \left\| \begin{matrix} 1 & -1 & -5 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{matrix} \right\| \mathfrak{b}_K \ .$$

On a alors

$$\mathbf{E} \left\| \begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \end{matrix} \right\| = \left\| \begin{matrix} 1 & -1 & -5 \\ 1 & 1 & -3 \\ 1 & 1 & 1 \\ 1 & 0 & 4 \\ 1 & -1 & 3 \\ 1 & 0 & 0 \end{matrix} \right\| \left\| \begin{matrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{matrix} \right\|$$

$$\mathfrak{A}_K^T \mathfrak{x} = \left\| \begin{matrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \mathbf{A}_3 \end{matrix} \right\| = \left\| \begin{matrix} \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4 + \mathbf{x}_5 + \mathbf{x}_6 \\ -\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 \qquad - \mathbf{x}_5 \\ -5\mathbf{x}_1 - 3\mathbf{x}_2 + \mathbf{x}_3 + 4\mathbf{x}_4 + \mathbf{x}_5 \end{matrix} \right\|$$

d'où

$$\mathfrak{G}_K = \left\| \begin{matrix} 6 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 60 \end{matrix} \right\| , \qquad \begin{aligned} \hat{\gamma}_0 &= \mathbf{A}_1/6 \\ \hat{\gamma}_1 &= \mathbf{A}_2/4 \\ \hat{\gamma}_2 &= \mathbf{A}_3/60 \end{aligned}$$

$$\mathbf{SC}\{\hat{\gamma}_0\} = A_1^2/6 , \quad \mathbf{SC}\{\hat{\gamma}_1\} = \mathbf{A}_2^2/4 , \quad \mathbf{SC}\{\hat{\gamma}_2\} = \mathbf{A}_3^2/60 .$$

$$\mathbf{SC}\{\hat{\gamma}_0\} = \mathbf{red}[\beta_0], \quad \mathbf{SC}\{\hat{\gamma}_1\} = \mathbf{red}[\beta_1 \mid \beta_0], \quad \mathbf{SC}\{\hat{\gamma}_2\} = \mathbf{red}[\beta_2 \mid \beta_0, \beta_1]$$

$$\mathbf{SC}N = \mathbf{SC}\{\hat{\gamma}_0\} + \mathbf{SC}\{\hat{\gamma}_1\} + \mathbf{SC}\{\hat{\gamma}_2\} ,$$

$$\mathbf{SC}E = \mathbf{SC}T - \mathbf{SC}N .$$

## 4, 3. *Problèmes de classification.*

4, 31. Supposons que l'on dispose des valeurs observées de douze aléatoires normales, indépendantes, de même variance $\sigma^2$, classées suivant deux critères: « lignes », de « valeurs » $L_1$ et $L_2$, et « colonnes », de « valeurs » $C_1$, $C_2$, $C_3$, suivant le schéma

| | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|
| $L_1$ | $x_1, x_2$ | $x_3, x_4$ | $x_5, x_6$ |
| $L_2$ | $x_7, x_8$ | $x_9, x_{10}$ | $x_{11}, x_{12}$ . |

On suppose a priori qu'il y a additivité, c'est-à-dire qu'il existe cinq nombres $\lambda_1$, $\lambda_2$, $\gamma_1$, $\gamma_2$, $\gamma_3$ tels que la valeur moyenne d'une observation de la ligne $L_i$ et de la colonne $C_k$ soit $\lambda_i + \gamma_k$ ($i = 1, 2$; $k = 1, 2, 3$). On a donc, par hypothèse,

$$\mathbf{E} \left\| \begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \\ \mathbf{x}_7 \\ \mathbf{x}_8 \\ \mathbf{x}_9 \\ \mathbf{x}_{10} \\ \mathbf{x}_{11} \\ \mathbf{x}_{12} \end{matrix} \right\| = \left\| \begin{matrix} 1 & & 1 & & \\ 1 & & 1 & & \\ 1 & & & 1 & \\ 1 & & & 1 & \\ 1 & & & & 1 \\ 1 & & & & 1 \\ & 1 & 1 & & \\ & 1 & 1 & & \\ & 1 & & 1 & \\ & 1 & & 1 & \\ & 1 & & & 1 \\ & 1 & & & 1 \end{matrix} \right\| \left\| \begin{matrix} \lambda_1 \\ \lambda_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{matrix} \right\| .$$

Si on appelle $S_{i,-}$, la somme des observations de la ligne $L_i$, et $S_{-,j}$ celle des observations de la colonne $C_k$, les équations normales s'écrivent:

$$6\hat{\lambda}_1 \qquad + 2\hat{\gamma}_1 + 2\hat{\gamma}_2 + 2\hat{\gamma}_3 = \mathbf{S}_{1,-} \qquad (a)$$

$$6\hat{\lambda}_2 + 2\hat{\gamma}_1 + 2\hat{\gamma}_2 + 2\hat{\gamma}_3 = \mathbf{S}_{2,-} \qquad (b)$$