Annexe K

Produit scalaire et statistique

K.1 Cas avec deux variables

On appelle $\mathbb{R}^n$ l'espace des individus et $\mathbb{R}^2$ l'espace des variables.
On pose :

$$D_{\frac{1}{n}} = \frac{1}{n} I_n \tag{K.1}$$

avec I_n la matrice unité à n lignes et n colonnes.
$D_{\frac{1}{n}}$ est inclus dans l'espace des variables.
Le produit scalaire vaut :

$$\langle X|Y\rangle_{D_{\frac{1}{n}}} = \left\langle X|D_{\frac{1}{n}}|Y\right\rangle = \sum_{i=1}^{n} \frac{1}{n} x_i y_i = \frac{1}{n} \langle X|Y\rangle \tag{K.2}$$

avec $\langle X|Y\rangle$ le produit scalaire canonique de $\mathbb{R}^n$.

On note $\mathbf{1}_n = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$ le vecteur dont toutes les coordonnées sont égales à 1, appelé le **vecteur unité de** $\mathbb{R}^n$. Ce vecteur est normé. Sa longueur est $||\mathbf{1}_n||_{D_{\frac{1}{n}}} = \frac{1}{n} \sum_{i=1}^{n} 1 \times 1 = \frac{1}{n} n = 1$.

K.1.1 Moyenne d'une variable statistique

La moyenne $\bar{X}$ de la variable statistique X est donnée par :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{n} x_i \times 1 = \left\langle X|D_{\frac{1}{n}}|\mathbf{1}_n\right\rangle = \langle X|\mathbf{1}_n\rangle_{D_{\frac{1}{n}}} \tag{K.3}$$

La moyenne de X est le produit scalaire de X par le vecteur unité $\mathbf{1}_n$.

Soit $X_0 = X - \bar{X}$ la variable centrée correspondant à X. Pour chaque individu i de la population :

$$\begin{aligned} X_0 &= \begin{pmatrix} x_1 - \bar{X} \\ \dots \\ x_n - \bar{X} \end{pmatrix} = \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix} - \bar{X}\mathbf{1}_n \\ X_0 &= X - \bar{X}\mathbf{1}_n \\ \Leftrightarrow X &= X_0 - \bar{X}\mathbf{1}_n \end{aligned} \tag{K.4}$$

$$X = X_0 - \langle X|\mathbf{1}_n\rangle_{D_{\frac{1}{n}}}\mathbf{1}_n \tag{K.5}$$

K.1.2 Variance d'une variable statistique

$$s^2(X) = \bar{X_0^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2 = \left\langle X_0|D_{\frac{1}{n}}|X_0\right\rangle \tag{K.6}$$

$$s^2(X) = ||X_0||^2 \tag{K.7}$$

La variance de X est le carré de la norme de la variable centrée.

K.1.3 Covariance

La covariance de deux variables quantitatives réelles X et Y définies sur $\mathbb{R}^2$ est la moyenne du produit des variables centrées.

$$\operatorname{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right) = \left\langle X_0|D_{\frac{1}{n}}|Y_0\right\rangle = \langle X_0|Y_0\rangle_{D_{\frac{1}{n}}} \tag{K.8}$$

On pose Z la matrice des variables centrées $Z = [X_0, Y_0]$. La matrice de covariance C vaut alors :

$$C = {}^tZ.D_{\frac{1}{n}}.Z \tag{K.9}$$

et, dans ce cas :

$$C = \frac{1}{n}{}^tZ.Z \tag{K.10}$$

La covariance est le produit scalaire des variables centrées.

K.1.4 Coefficient de corrélation linéaire

$$r_{X,Y} = \frac{\mathrm{cov}\,(X,Y)}{s\,(X)\,s\,(Y)} = \frac{\left\langle X_0 | D_{\frac{1}{n}} | Y_0 \right\rangle}{||X_0||_\phi ||Y_0||_\phi} = \cos\,(X_0, Y_0) \qquad \text{(K.11)}$$

Le coefficient de corrélation linéaire est le cosinus de l'angle des variables centrées.

K.2 Prédicteur linéaire d'une régression linéaire

Soient Y la variable à expliquer, X la variable explicative, X_0 et Y_0 les variables centrées.

Le prédicteur linéaire $\Delta_{Y|X}$ est :

$$y^* = a + bx \qquad \text{(K.12)}$$

c'est-à-dire

$$y^* - \bar{Y} = b\left(x - \bar{X}\right) \qquad \text{(K.13)}$$

soit $y_0{}^* = bx_0$. Il est représenté par la droite de régression de Y en X dans l'espace des individus.

$$b = \frac{\mathrm{cov}\,(X,Y)}{s^2\,(X)} = \frac{\langle X_0 | Y_0 \rangle_{\frac{1}{n}}}{||X_0||_{D_{\frac{1}{n}}}{}^2} \qquad \text{(K.14)}$$

$bX_0 = \frac{\langle X_0|Y_0\rangle_{D\frac{1}{n}}}{||X_0||_{D_{\frac{1}{n}}}{}^2} X_0$ est le projeté orthogonal de Y_0 sur X_0, $Y_0 - bX_0$ est orthogonal à X_0, et b **est la valeur minimisant l'expression** :

$$\begin{array}{l} S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{0i} - bX_{0i}\right)^2 \\ S^2 = ||Y_0 - bX_0||_{D_{\frac{1}{n}}}{}^2 \\ S^2 = s^2\,(Y - bX) \\ S^2 = s^2\,(Y - a - bX) \\ S^2 = s^2\,(Y - Y^*) \\ S^2 = s^2\,(Y_0 - Y_0{}^*) \end{array} \qquad \text{(K.15)}$$

Le prédicteur linéaire de la variable centrée Y_0 est le projeté orthogonal de Y_0 sur X_0 dans $\mathbb{R}^n$, c'est-à-dire la variable $Y_0{}^*$ minimisant la variance $Y_0 - Y_0{}^*$.

$$\begin{aligned} s^2(Y) &= ||Y_0||_{D_{\frac{1}{n}}}{}^2 = ||Y_0 - bX_0 + bX_0||_{D_{\frac{1}{n}}}{}^2 \\ s^2(Y) &= ||Y_0 - bX_0||_{D_{\frac{1}{n}}}{}^2 + ||bX_0||_{D_{\frac{1}{n}}}{}^2 \end{aligned} \tag{K.16}$$

or $||Y_0 - bX_0||_{D_{\frac{1}{n}}}{}^2 = S_{\min}{}^2$

$$s^2(Y) = S_{\min}{}^2 + b^2 ||X_0||_{D_{\frac{1}{n}}}{}^2 \tag{K.17}$$

or $||X_0||_{D_{\frac{1}{n}}}{}^2 = s^2(X)$

$$\begin{aligned} s^2(Y) &= S_{\min}{}^2 + b^2 s^2(X) \\ s^2(Y) &= S_{\min}{}^2 + \left(\tfrac{\operatorname{cov}(X,Y)}{s^2(X)}\right)^2 s^2(X) \\ s^2(Y) &= S_{\min}{}^2 + \left(\tfrac{\operatorname{cov}(X,Y)s(X)}{s^2(X)}\right)^2 \\ s^2(Y) &= S_{\min}{}^2 + \left(\tfrac{\operatorname{cov}(X,Y)s(Y)}{s(X)s(Y)}\right)^2 \\ s^2(Y) &= S_{\min}{}^2 + \left(\tfrac{\operatorname{cov}(X,Y)}{s(X)s(Y)}\right)^2 s^2(Y) \end{aligned} \tag{K.18}$$

d'où

$$s^2(Y) = S_{\min}{}^2 + r_{XY}{}^2 s^2(Y) \tag{K.19}$$

$S_{\min}{}^2$ correspond à la **variance résiduelle**. $r_{XY}{}^2 s^2(Y)$ est la **variance expliquée par la régression**.

De manière symétrique, si Y est la variable explicative et X la variable à expliquer, on a :

$$s^2(X) = S'_{\min}{}^2 + r_{XY}{}^2 s^2(X) \tag{K.20}$$

K.3 Cas généralisé avec plusieurs variables

On appelle $\mathbb{R}^n$ l'espace des individus et $\mathbb{R}^m$ l'espace des variables.

Un n-uplet de variables est un vecteur dans l'espace des individus.

Un m-uplet de variables est un vecteur dans l'espace des variables.

Bibliographie