Mise à niveau en statistique

Correction de la feuille de TD 4 : Modèle linéaire

M2 Mathématiques de l'Aléatoire – Université Paris-Saclay

Exercice 1 – Théorème de Cochran

1. Pour tout $x \in \mathbb{R}^n$, $Ax = u_F(x) \in F$, et la projection d'un élément de F sur F est lui-même. Ainsi $A^2x = Ax$ pour tout x, donc $A^2 = A$. La matrice $I - A$ représente $u_{F^\perp}$, puisque $x = u_F(x) + u_{F^\perp}(x)$.

   Pour tous x, y,
   $$y^\top (I-A)^\top Ax = ((I-A)y)^\top Ax = 0,$$
   car $(I-A)y \in F^\perp$ et $Ax \in F$. Donc $(I-A)^\top A = 0$. En développant et en utilisant $A^2 = A$,
   $$0 = (I-A)^\top A = A - A^\top A.$$

   En échangeant les rôles de x et y, ou directement par $\langle Ax, y\rangle = \langle Ax, Ay\rangle = \langle x, Ay\rangle$, on obtient $A^\top = A$.

2. Choisissons une base orthonormée $(e_1, \dots, e_p)$ de F puis une base orthonormée $(e_{p+1}, \dots, e_n)$ de $F^\perp$. Dans la base orthonormée obtenue,
   $$[u_F] = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}, \qquad [u_{F^\perp}] = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-p} \end{pmatrix}.$$

   Si $P = (e_1|\cdots|e_n)$ est la matrice de passage, ses colonnes sont orthonormées : $P^\top P = I_n$, donc $P^{-1} = P^\top$. En particulier,
   $$P^\top AP = \operatorname{diag}(I_p, 0).$$

3. (a) Comme $X \sim \mathcal{N}(0, I_n)$,
   $$\operatorname{Cov}((I-A)X, AX) = (I-A)\operatorname{Cov}(X)A^\top = (I-A)A = 0.$$

   Les deux vecteurs sont conjointement gaussiens, car ils sont des transformations linéaires de X. Pour des vecteurs conjointement gaussiens, une covariance croisée nulle implique l'indépendance. Ainsi $(I-A)X$ et AX sont indépendants.

   (b) Le vecteur $Z = P^\top X$ vérifie $Z \sim \mathcal{N}(0, P^\top P) = \mathcal{N}(0, I_n)$: ses coordonnées sont donc i.i.d. $\mathcal{N}(0,1)$. Comme
   $$P^\top AX = \operatorname{diag}(I_p, 0)Z, \qquad P^\top (I-A)X = \operatorname{diag}(0, I_{n-p})Z,$$

   le premier vecteur a pour p premières coordonnées $Z_1, \dots, Z_p$ et des zéros ensuite; le second a des zéros puis $Z_{p+1}, \dots, Z_n$.

   (c) Une matrice orthogonale conserve la norme :
   $$\|AX\|^2 = \|P^\top AX\|^2 = \sum_{j=1}^{p} Z_j^2 \sim \chi^2_p.$$

De même,

$$\|(I-A)X\|^2 = \sum_{j=p+1}^{n} Z_j^2 \sim \chi^2_{n-p}.$$

Ces deux variables sont indépendantes.

4. Posons $\mathbf{1} = (1,\dots,1)^\top$ et $F = \mathrm{Vect}(\mathbf{1})$. Alors

$$u_F(X) = \overline{X}_n \mathbf{1}, \qquad u_{F^\perp}(X) = X - \overline{X}_n \mathbf{1}.$$

Le théorème de Cochran donne l'indépendance de ces deux vecteurs, donc celle de $\overline{X}_n$ et de s_n^2, qui sont respectivement des fonctions mesurables de ces deux projections. Enfin,

$$\overline{X}_n \sim \mathcal{N}(0, 1/n), \qquad n s_n^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}.$$

Exercice 2 – Conditionnement gaussien

Écrivons

$$X = (X_1,\dots,X_n)^\top \sim \mathcal{N}(\mu, \Sigma), \qquad Y = AX.$$

Alors

$$\mathbb{E}[Y] = A\mu, \qquad \mathrm{Cov}(Y) = C := A\Sigma A^\top, \qquad \mathrm{Cov}(X,Y) = \Sigma A^\top.$$

Supposons d'abord C inversible et posons

$$m(Y) = \mu + \Sigma A^\top C^{-1}(Y - A\mu), \qquad R = X - m(Y).$$

Pourquoi $m(Y)$ est bien une espérance conditionnelle. Le couple (R, Y) est conjointement gaussien. De plus,

$$\begin{aligned}\mathrm{Cov}(R,Y) &= \mathrm{Cov}(X,Y) - \Sigma A^\top C^{-1}\,\mathrm{Cov}(Y,Y)\\ &= \Sigma A^\top - \Sigma A^\top C^{-1} C = 0.\end{aligned}$$

Par gaussianité, R et Y sont donc indépendants. Comme R est centré,

$$\mathbb{E}[R \mid Y] = \mathbb{E}[R] = 0.$$

Ici, la première égalité utilise l'indépendance; elle ne serait pas vraie pour un résidu simplement décorrélé dans un modèle non gaussien. Par ailleurs, $m(Y)$ est $\sigma(Y)$-mesurable, donc

$$\mathbb{E}[m(Y) \mid Y] = m(Y).$$

En conditionnant l'identité $X = m(Y) + R$, on obtient, terme à terme,

$$\boxed{\mathbb{E}[X \mid AX] = \mu + \Sigma A^\top (A\Sigma A^\top)^{-1}(AX - A\mu).}$$

Loi conditionnelle. La covariance du résidu vaut

$$\begin{aligned}\Gamma_R &= \mathrm{Cov}\big(X - \Sigma A^\top C^{-1} Y\big)\\ &= \Sigma - \Sigma A^\top C^{-1} A\Sigma.\end{aligned}$$

Puisque R est indépendant de Y, conditionner par $Y = y$ ne change pas sa loi. Dans $X = m(Y) + R$, seul le premier terme est alors figé :

$$\boxed{X \mid (AX = y) \sim \mathcal{N}\left(\mu + \Sigma A^\top (A\Sigma A^\top)^{-1}(y - A\mu),\, \Sigma - \Sigma A^\top (A\Sigma A^\top)^{-1} A\Sigma\right).}$$

Règles à retenir sur le conditionnement. Pour une matrice déterministe B, une fonction $h(Y)$ et une variable intégrable U,

$$\mathbb{E}[BU + h(Y) \mid Y] = B\mathbb{E}[U \mid Y] + h(Y), \qquad \mathbb{E}[\mathbb{E}[U \mid Y]] = \mathbb{E}[U].$$

Une quantité Y-mesurable sort donc de l'espérance conditionnelle; une quantité indépendante de Y peut être remplacée par son espérance ordinaire. Ces deux règles sont distinctes.

Si $C = A\Sigma A^\top$ est singulière, la même formule reste valable en remplaçant C^{-1} par la pseudo-inverse de Moore–Penrose C^+; la loi conditionnelle est alors éventuellement dégénérée sur un sous-espace affine.

Exercice 3 – Modèle linéaire gaussien

On note $X \in \mathbb{R}^{n\times p}$ la matrice dont la i-ème ligne est $x_i^\top$, et l'on suppose $\operatorname{rang}(X) = p$.

1. Le modèle s'écrit
$$Y = X\beta + \varepsilon, qquad\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

2. Ainsi,
$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n), \qquad Y_i \sim \mathcal{N}(x_i^\top \beta, \sigma^2).$$

3. Pour $h \neq 0$, $h^\top X^\top X h = \|Xh\|^2 > 0$, car X est injective. La matrice $X^\top X$ est donc définie positive et inversible. Le gradient de $h \mapsto \|Y - Xh\|^2$ vaut $-2X^\top(Y - Xh)$; l'unique minimiseur est
$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y.$$

4. La matrice
$$P = X(X^\top X)^{-1} X^\top$$
est symétrique et vérifie $P^2 = P$. De plus, $PXh = Xh$ pour tout h, et $Pz = 0$ lorsque $z \in \operatorname{Im}(X)^\perp$, car $X^\top z = 0$. C'est donc le projecteur orthogonal sur $D = \operatorname{Im}(X)$.

5. On en déduit
$$X\widehat{\beta} = PY :$$
le vecteur des valeurs ajustées est la projection orthogonale de Y sur D. Le résidu est $\widehat{\varepsilon} = (I - P)Y$.

6. Comme $PX\beta = X\beta$,
$$X\widehat{\beta} \sim \mathcal{N}(X\beta, \sigma^2 P), \qquad \widehat{\beta} \sim \mathcal{N}\big(\beta, \sigma^2 (X^\top X)^{-1}\big).$$
En particulier, $\widehat{\beta}$ est sans biais.

7. Posons $v_0 = x_0^\top (X^\top X)^{-1} x_0$. Alors
$$\frac{x_0^\top \widehat{\beta} - x_0^\top \beta}{\sigma\sqrt{v_0}} \sim \mathcal{N}(0, 1).$$
Si $z_{1-\alpha/2}$ est le quantile normal usuel, un intervalle exact de niveau $1 - \alpha$ est
$$x_0^\top \widehat{\beta} \pm z_{1-\alpha/2}\sigma\sqrt{v_0}.$$

8. Supposons σ^2 inconnu et posons $\widehat{\sigma}^2 = \|(I - P)Y\|^2/(n - p)$.
   (a) La quantité $n^{-1}\|Y - X\beta\|^2$ serait naturelle si β était connu. En le remplaçant par $\widehat{\beta}$, on ajuste p paramètres et il reste $n - p$ degrés de liberté; cette correction rend l'estimateur sans biais.

(b) On a
$$(n-p)\widehat{\sigma}^2 = \|Y - X\widehat{\beta}\|^2 = \|(I-P)Y\|^2 = \|(I-P)\varepsilon\|^2.$$

(c) Les espaces D et $D^\perp$ sont orthogonaux. Le théorème de Cochran donne
$$\frac{\|P\varepsilon\|^2}{\sigma^2} \sim \chi^2_p, qquad\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p},$$
et l'indépendance de $P\varepsilon$ et $(I-P)\varepsilon$. Ainsi, $\widehat{\beta}$, fonction de $P\varepsilon$, est indépendante de $\widehat{\sigma}^2$, fonction de $(I-P)\varepsilon$.

(d) Cette indépendance donne
$$\frac{x_0^\top \widehat{\beta} - x_0^\top \beta}{\widehat{\sigma}\sqrt{v_0}} \sim t_{n-p}.$$
L'intervalle exact est donc
$$x_0^\top \widehat{\beta} \pm t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{v_0}.$$

(e) Pour $x_0 = e_j$, on rejette $H_0 : \beta_j = \beta_{j,0}$ lorsque
$$\left|\frac{\widehat{\beta}_j - \beta_{j,0}}{\widehat{\sigma}\sqrt{[(X^\top X)^{-1}]_{jj}}}\right| > t_{n-p,1-\alpha/2}.$$

(f) Si $q_{\nu,u}$ est le quantile d'ordre u de χ^2_ν, un intervalle exact pour σ est
$$\left[\sqrt{\frac{(n-p)\widehat{\sigma}^2}{q_{n-p,1-\alpha/2}}}, \sqrt{\frac{(n-p)\widehat{\sigma}^2}{q_{n-p,\alpha/2}}}\right].$$

(g) Sous $H_0 : \beta = \beta_0$,
$$F = \frac{\|X(\widehat{\beta} - \beta_0)\|^2/p}{\widehat{\sigma}^2} \sim F_{p,n-p}.$$
On rejette au niveau α lorsque $F > F_{p,n-p;1-\alpha}$.

9. Pour une nouvelle donnée $Y_{n+1} = x_{n+1}^\top \beta + \varepsilon_{n+1}$, où $\varepsilon_{n+1} \sim \mathcal{N}(0,\sigma^2)$ est indépendante de ε :

(a) Le prédicteur naturel est $\widehat{Y}_{n+1} = x_{n+1}^\top \widehat{\beta}$.

(b) En posant $h_{n+1} = x_{n+1}^\top (X^\top X)^{-1} x_{n+1}$,
$$Y_{n+1} - \widehat{Y}_{n+1} = \varepsilon_{n+1} - x_{n+1}^\top(\widehat{\beta} - \beta) \sim \mathcal{N}(0, \sigma^2(1+h_{n+1})).$$
Les deux termes sont indépendants : le premier est le nouveau bruit et le second dépend seulement de l'échantillon d'apprentissage.

(c) L'erreur de prédiction est indépendante de $\widehat{\sigma}^2$, car $\widehat{\beta}$ est indépendante des résidus et ε_{n+1} est indépendant de tout l'échantillon. Ainsi,
$$\frac{Y_{n+1} - \widehat{Y}_{n+1}}{\widehat{\sigma}\sqrt{1+h_{n+1}}} \sim t_{n-p},$$
et un intervalle de prédiction exact est
$$\widehat{Y}_{n+1} \pm t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{1+h_{n+1}}.$$

Exercice 4 – Moindres carrés, encore

On suppose ici, comme l'indique la première question, que $\text{rang}(X) = p$.

1. L'application $h \mapsto Xh$ est injective; son image est donc de dimension p. De plus, $h^\top X^\top Xh = \|Xh\|^2 > 0$ pour $h \neq 0$, donc $X^\top X$ est inversible.

2. L'annulation du gradient de $\|Y - Xh\|^2$ donne les équations normales $X^\top Xh = X^\top Y$, puis $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$.

3. Minimiser $\|Y - Xh\|$ revient à chercher le point de $\text{Im}(X)$ le plus proche de Y. Ainsi
$$P_X Y = X\widehat{\beta}.$$

4. La décomposition orthogonale $Y = P_X Y + (I - P_X)Y$ donne $X^\top(Y - X\widehat{\beta}) = 0$. On retrouve
$$P_X = X(X^\top X)^{-1} X^\top, \qquad \widehat{\varepsilon} = (I - P_X)Y.$$

5. Comme $\mathbb{E}[Y] = X\beta$,
$$\mathbb{E}[\widehat{\beta}] = (X^\top X)^{-1} X^\top X\beta = \beta.$$

6. Si $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, alors $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$.

7. Les vecteurs
$$\widehat{Y} = X\widehat{\beta} = X\beta + P_X\varepsilon, \qquad \widehat{\varepsilon} = (I - P_X)\varepsilon$$
sont gaussiens et leur covariance croisée vaut $\sigma^2 P_X(I - P_X) = 0$. Ils sont donc indépendants, avec
$$\widehat{Y} \sim \mathcal{N}(X\beta, \sigma^2 P_X), qquad\widehat{\varepsilon} \sim \mathcal{N}(0, \sigma^2(I - P_X)).$$

8. Le vecteur empilé est gaussien, de moyenne
$$\mathbb{E}\begin{pmatrix} X\widehat{\beta} \\ \widehat{\beta} \\ \widehat{\varepsilon} \end{pmatrix} = \begin{pmatrix} X\beta \\ \beta \\ 0 \end{pmatrix},$$
et de covariance
$$\sigma^2 \begin{pmatrix} P_X & X(X^\top X)^{-1} & 0 \\ (X^\top X)^{-1} X^\top & (X^\top X)^{-1} & 0 \\ 0 & 0 & I - P_X \end{pmatrix}.$$

Exercice 5 – Régression ridge

Le modèle est $Y = X\theta + \xi$, avec $\mathbb{E}[\xi] = 0$ et $\text{Cov}(\xi) = \sigma^2 I_n$.

1. Si $k > n$, alors $\text{rang}(X) \leq n < k$ et $\ker(X) \neq \{0\}$. Les moindres carrés ne sont pas identifiables : si $\widehat{\theta}$ est une solution, alors $\widehat{\theta} + v$ est encore une solution pour tout $v \in \ker(X)$. Le vecteur ajusté $X\widehat{\theta}$ reste néanmoins unique.

2. Le critère ridge
$$Q_\lambda(t) = \|Y - Xt\|^2 + \lambda\|t\|^2$$
a pour Hessienne $2(X^\top X + \lambda I_k)$, définie positive puisque
$$u^\top(X^\top X + \lambda I_k)u = \|Xu\|^2 + \lambda\|u\|^2 > 0$$
pour $u \neq 0$. Il admet donc l'unique minimiseur
$$\widehat{\theta}_\lambda = (X^\top X + \lambda I_k)^{-1} X^\top Y,$$
bien défini même si $k > n$.

3. En posant $G_\lambda = (X^\top X + \lambda I_k)^{-1}$,

$$\mathbb{E}[\widehat{\theta}_\lambda] = G_\lambda X^\top X\theta = \theta - \lambda G_\lambda \theta,$$

et

$$\mathrm{Cov}(\widehat{\theta}_\lambda) = \sigma^2 G_\lambda X^\top X G_\lambda.$$

Le biais vaut donc $-\lambda G_\lambda \theta$: la réduction de variance est obtenue au prix d'un biais.

4. Supposons $k = 1$ et posons $a = X^\top X > 0$. Alors

$$R(\widehat{\theta}_{\mathrm{MC}}, \theta) = \frac{\sigma^2}{a},$$

tandis que

$$R(\widehat{\theta}_\lambda, \theta) = \frac{\lambda^2\theta^2 + \sigma^2 a}{(a+\lambda)^2}.$$

Le gain de risque vaut

$$R(\widehat{\theta}_{\mathrm{MC}}, \theta) - R(\widehat{\theta}_\lambda, \theta) = \frac{\lambda\{2\sigma^2 a + \lambda(\sigma^2 - a\theta^2)\}}{a(a+\lambda)^2}.$$

Il est strictement positif pour tout $\lambda > 0$ assez petit. Plus précisément, si $a\theta^2 \leq \sigma^2$, tout $\lambda > 0$ convient; sinon il suffit que

$$0 < \lambda < \frac{2\sigma^2 a}{a\theta^2 - \sigma^2}.$$

La condition simple $0 < \lambda \leq 2\sigma^2/\theta^2$ est également suffisante. Il existe donc toujours une pénalisation ridge améliorant le risque quadratique des moindres carrés dans ce cadre unidimensionnel.