Gestion des incertitudes et analyse de risque
PC1 - Régression linéaire et inférence bayésienne

Exercice 1. Régression linéaire.

On cherche à expliquer des observations scalaires $(y_i)_{i=1,\dots,n}$ à partir de variables observées $(x_i^1,\dots,x_i^p)_{i=1,\dots,n}$. Par exemple, on peut chercher à expliquer le salaire des individus[1] (les $(y_i)_{i=1,\dots,n}$, l'indice i désignant alors le i-ième individu) en fonction de diverses données $(x_i^1,\dots,x_i^p)_{i=1,\dots,n}$ sur les individus : l'âge, le nombre d'années d'éducation, le nombre d'années d'expérience, la taille, etc. On utilise un modèle linéaire pour essayer de reproduire les observations : pour tout $i \in \{1,\dots,n\}$,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i$$

où les $(\beta_j)_{j=0,\dots,p}$ sont des constantes que l'on cherche, et les $(\varepsilon_i)_{i=1,\dots,n}$ sont des variables aléatoires i.i.d. de loi normale centrée et de variance fixée σ^2. On suppose dans la suite que $n \geq p+1$ (il y a plus d'observations que de paramètres à estimer).

1. Vérifier qu'on peut réécrire le problème sous la forme matricielle

$$y = H\beta + \varepsilon \tag{1}$$

où $H = \begin{bmatrix} 1 & x_1^1 & \dots & x_1^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^1 & \dots & x_n^p \end{bmatrix} \in \mathbb{R}^{n\times(p+1)}$, $\beta = (\beta_0,\dots,\beta_p)^T \in \mathbb{R}^{p+1}$ et $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)^T \in \mathbb{R}^n$.

2. Montrer que dans le modèle (1), pour des variables explicatives H fixées et pour une valeur fixée de β, la densité du vecteur y est :

$$p_\beta(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|y-H\beta\|^2}{2\sigma^2}\right). \tag{2}$$

La fonction p_β est appelée la vraisemblance.

3. Montrer qu'à y fixé, le maximum de $\beta \mapsto p_\beta(y)$ est atteint en

$$\hat{\beta} \in \arg\min_{\beta\in\mathbb{R}^{p+1}} \|y-H\beta\|^2 \tag{3}$$

et que $\hat{\beta}$ satisfait

$$H^T H\hat{\beta} = H^T y. \tag{4}$$

On appelle $\hat{\beta}$ un estimateur de maximum de vraisemblance. Montrer que ce problème admet une solution en vérifiant que $\mathrm{Im}(H^TH) = \mathrm{Im}H^T$. *Indication : on pourra vérifier que* $\mathrm{Ker}(H^TH) = \mathrm{Ker}H$ *et utiliser le fait que pour toute matrice* M, $(\mathrm{Im}M)^\perp = \mathrm{Ker}(M^T)$. Décrire l'ensemble des solutions de (4). Montrer que $H\hat{\beta}$ est la projection orthogonale de y sur $\mathrm{Im}(H)$:

$$H\hat{\beta} = \pi_{\mathrm{Im}(H)}(y).$$

[1] En fait, en pratique, on remarque qu'il vaut mieux utiliser le logarithme des salaires pour que le modèle linéaire soit correct...

4. On note dans cette question H_p et $\hat{\beta}_p$ la matrice H et une solution $\hat{\beta}$ de (4) pour un nombre p de variables observées. Montrer que $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^n y_i$. Montrer que $\|y - H_0\hat{\beta}_0\|^2 = \|y - H_p\hat{\beta}_p\|^2 + \|H_p\hat{\beta}_p - H_0\hat{\beta}_0\|^2$. En déduire que le coefficient de détermination

$$R^2 = \frac{\|H_p\hat{\beta}_p - H_0\hat{\beta}_0\|^2}{\|y - H_0\hat{\beta}_0\|^2}$$

est un réel de $[0,1]$. Discuter les cas limites $R^2 \to 0$ et $R^2 \to 1$.

5. On suppose que

$$\mathrm{Rang}(H) = p + 1.$$

Comment se ramener à cette situation en pratique ? Montrer que la matrice $H^T H \in \mathbb{R}^{(p+1)\times(p+1)}$ est inversible. Comment calculer $\hat{\beta}$ en pratique ?

On suppose désormais que

$$\mathrm{Rang}(H) = r$$

avec $r \in \{1, \ldots, p+1\}$. On va chercher à décrire plus précisément les solutions de (4) dans le cas $r < p + 1$.

6. Montrer que toute matrice $A \in \mathbb{R}^{n\times(p+1)}$ peut s'écrire sous la forme

$$A = U\Sigma V^T \tag{5}$$

avec $U \in \mathbb{R}^{n\times n}$ et $V \in \mathbb{R}^{(p+1)\times(p+1)}$ des matrices orthogonales et $\Sigma \in \mathbb{R}^{n\times(p+1)}$ une matrice diagonale de diagonale $(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$ avec $\sigma_1 \geq \ldots \geq \sigma_r > 0$ et $r \leq p + 1 \leq n$. *Indication : On pourra considérer la matrice $A^T A$ pour construire V et Σ, puis ensuite construire U.* Quel est le rang de A ? Montrer qu'on peut écrire A sous la forme :

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

en identifiant les vecteurs $u_i \in \mathbb{R}^n$ et $v_i \in \mathbb{R}^{(p+1)}$. La décomposition (5) s'appelle la décomposition en valeurs singulières (SVD) de A. Il existe des algorithmes efficaces pour calculer cette décomposition.

7. On considère la SVD de la matrice $H \in \mathbb{R}^{n\times(p+1)}$:

$$H = U\Sigma V^T$$

avec $\Sigma \in \mathbb{R}^{n\times(p+1)}$ matrice diagonale de diagonale $(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$. Montrer que l'ensemble des solutions de (4) se décrit de la manière suivante :

$$\hat{\beta} \in \left\{ H^+ y + \sum_{i=r+1}^{p+1} \alpha_i v_i,\ (\alpha_i)_{i\in\{r+1,\ldots,p+1\}} \in \mathbb{R}^{p-r+1} \right\} = H^+ y + \mathrm{Ker}(H) \tag{6}$$

où H^+ est la pseudo-inverse de H défini par

$$H^+ = V\Sigma^+ U^T$$

où $\Sigma^+ \in \mathbb{R}^{(p+1)\times n}$ est la matrice diagonale de diagonale $(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0)$. Comparer au résultat obtenu à la Question 3. Vérifier que

$$HH^+ = \pi_{\mathrm{Im}H} \text{ et } H^+H = \pi_{(\mathrm{Ker}H)^\perp}$$

où (cf. Question 3), $\pi_{\mathrm{Im}H}$ (resp. $\pi_{(\mathrm{Ker}H)^\perp}$) est la projection orthogonale sur $\mathrm{Im}H$ (resp. $(\mathrm{Ker}H)^\perp$). Comment caractériser géométriquement H^+y parmi toutes les solutions ? Que se passe-t-il si $r = p + 1$?

8. On suppose qu'il existe un $\beta^* \in (\mathrm{Ker}H)^\perp$ tel que $y = H\beta^* + \varepsilon$. Noter que l'hypothèse $\beta^* \in (\mathrm{Ker}H)^\perp$ est naturelle : on ne peut pas apprendre les composantes de β^* sur $\mathrm{Ker}H$ en n'observant seulement $H\beta^* + \varepsilon$. Montrer que $\hat{\beta} = H^+ y$ est un vecteur gaussien centré en β^* (on dit que l'estimateur est *sans biais*) :

$$\mathbb{E}(\hat{\beta}) = \beta^*$$

et de matrice de covariance $\sigma^2 V \Sigma^+ (\Sigma^+)^T V^T$. En déduire que

$$\mathbb{E}(\|\hat{\beta} - \beta^*\|^2) = \sigma^2 \sum_{i=1}^{r} \frac{1}{\sigma_i^2}.$$

Construire un intervalle de confiance de niveau $(1-\alpha)$ pour β_k^*. Que se passe-t-il si H possède des valeurs propres singulières très petites ?

9. *(Pour aller plus loin...)* Si σ^2 n'est pas fixé, on peut chercher à maximiser la vraisemblance en le couple $(\beta, \sigma^2) \mapsto p_{\beta,\sigma^2}(y)$ (à y fixé). Montrer que le maximum est atteint en $(\hat{\beta}, \widehat{\sigma}^2)$ avec $\hat{\beta}$ satisfaisant (4) et

$$\widehat{\sigma}^2 = \frac{\|y - H\hat{\beta}\|^2}{n}.$$

Quelle est l'interprétation géométrique de $\|y - H\hat{\beta}\|^2$? On suppose qu'il existe un β^* tel que $y = H\beta^* + \varepsilon$ avec des variables aléatoires $(\varepsilon_i)_{i=1,\dots,n}$ de loi normale centrée et de variance $(\sigma^*)^2$. Montrer que $\frac{n\widehat{\sigma}^2}{(\sigma^*)^2}$ suit une loi du χ^2 de paramètre $n - r$. En déduire un estimateur sans biais de $(\sigma^*)^2$.

10. *(Pour aller plus loin...)* On souhaite dans cette question faire un lien entre le coefficient de détermination R^2 (cf. Question 4) et une statistique de test. On considère l'hypothèse nulle

$$\mathcal{H}_0 = \{\forall j \in \{1, \dots, p\},\ \beta_j^* = 0\}$$

et l'hypothèse alternative $\mathcal{H}_1 = \{\exists j \in \{1, \dots, p\},\ \beta_j^* \neq 0\}$. L'hypothèse $\mathcal{H}_0$ peut s'exprimer sous la forme : "aucune des variables observées $(x^1, \dots, x^p)$ n'est utile pour expliquer l'observation y". On introduit la statistique de test

$$F = \frac{\|H_p\hat{\beta}_p - H_0\hat{\beta}_0\|^2/(r-1)}{\|y - H_p\hat{\beta}_p\|^2/(n-r)}$$

en utilisant les notations de la Question 4.

(a) En utilisant le théorème de Cochran, montrer que $\frac{\|\varepsilon - \pi_{\mathrm{Im}H_p}\varepsilon\|^2}{\sigma^2}$ et $\frac{\|\pi_{\mathrm{Im}H_p}\varepsilon - \pi_{\mathrm{Im}H_0}\varepsilon\|^2}{\sigma^2}$ sont deux variables aléatoires indépendantes respectivement distribuées suivant les lois du χ^2 à $(n-r)$ et à $(r-1)$ degrés de liberté.

(b) Montrer que sous $\mathcal{H}_0$, $H_p\hat{\beta}_p = H_0\beta_0^* + \pi_{\mathrm{Im}(H_p)}(\varepsilon)$ et $H_0\hat{\beta}_0 = H_0\beta_0^* + \pi_{\mathrm{Im}(H_0)}(\varepsilon)$. En déduire que F suit la loi de Fischer $\mathcal{F}_{r-1,n-r}$.

(c) Montrer que sous $\mathcal{H}_1$, $H_p\hat{\beta}_p = H_p\beta^* + \pi_{\mathrm{Im}(H_p)}(\varepsilon)$ et $H_0\hat{\beta}_0 = \pi_{\mathrm{Im}(H_0)}(H_p\beta^*) + \pi_{\mathrm{Im}(H_0)}(\varepsilon)$. Expliquer pourquoi F prend des valeurs plus grandes sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$.

(d) On choisit donc comme région critique

$$W = \{F \geq \mathcal{F}_{r-1,n-r,1-\alpha}\},$$

où $\alpha \in (0,1)$ est le niveau du test, et $\mathcal{F}_{r-1,n-r,1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de la loi de Fisher $\mathcal{F}_{r-1,n-r}$. Si l'évènement W est réalisé, on rejette $\mathcal{H}_0$. Sinon, on accepte $\mathcal{H}_0$. Montrer que

$$F = \frac{R^2}{1-R^2}\frac{n-r}{r-1}$$

et faire un lien entre le test et les asymptotiques sur R^2 discutées en Question 4.

Exercice 2. Régularisation et inférence bayésienne.

On se place à nouveau dans le cadre de l'exercice précédent, avec un matrice H de rang r. Nous avons vu que si H n'est pas de rang plein, il existe plusieurs estimateurs de maximum de vraisemblance (cf. Question 3). De plus, si H possède des petites valeurs singulières, la variance de l'estimateur de maximum de vraisemblance de norme minimale $\hat{\beta} = H^+ y$ devient très grande (cf. Question 8). Pour résoudre cette difficulté, on considère en pratique un problème régularisé, lié à une approche bayésienne.

Plus précisément, on considère que le modèle est désormais le suivant. On suppose que pour tout $i \in \{1, \dots, n\}$,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i$$

où les $(\varepsilon_i)_{i=1,\dots,n}$ sont des variables aléatoires i.i.d. de loi normale centrée et de variance fixée σ^2 et où les $(\beta_j)_{j=0,\dots,p}$ sont également aléatoires, de loi un vecteur gaussien centré en $\bar{\beta}$ et de variance α^2 (c'est la loi *a priori*), où $\bar{\beta}$ et α^2 sont fixés.

1. Montrer que la densité du vecteur (y, β) s'écrit

$$q(y, \beta) = p_\beta(y) \frac{1}{(2\pi\alpha^2)^{(p+1)/2}} \exp\left(-\frac{\|\beta - \bar{\beta}\|^2}{2\alpha^2}\right)$$

où p_β est défini par (2).

2. En déduire que le maximum de la densité de la loi de β sachant les observations y (qu'on appelle la loi *a posteriori*) est atteint en un point $\hat{\beta}_\alpha$ satisfaisant

$$\hat{\beta}_\alpha \in \arg\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{\sigma^2}\|y - H\beta\|^2 + \frac{1}{\alpha^2}\|\beta - \bar{\beta}\|^2.$$

On appelle ce vecteur un mode de la loi a posteriori (MAP). Montrer qu'il existe un unique MAP défini par

$$\hat{\beta}_\alpha = \left(H^T H + \frac{\sigma^2}{\alpha^2}\mathrm{Id}\right)^{-1} \left(H^T y + \frac{\sigma^2}{\alpha^2}\bar{\beta}\right).$$

3. Montrer que pour tout $\hat{\beta}$ défini par (3), pour tout $\alpha > 0$, $\|\hat{\beta}_\alpha - \bar{\beta}\| \leq \|\hat{\beta} - \bar{\beta}\|$.

4. En déduire que

$$\lim_{\alpha \to \infty} \hat{\beta}_\alpha = \pi_{H^+ y + \mathrm{Ker}(H)}(\bar{\beta})$$

où $\pi_{H^+ y + \mathrm{Ker}(H)}$ désigne la projection orthogonale sur le sous espace affine $H^+ y + \mathrm{Ker}(H)$ (cf. (6)). Montrer en particulier que si $\bar{\beta} \in (\mathrm{Ker} H)^\perp$, alors $\lim_{\alpha \to \infty} \hat{\beta}_\alpha = H^+ y$.

5. On se place à nouveau dans le cadre de la Question 8 ci-dessus : on suppose que $y = H\beta^* + \varepsilon$ pour un $\beta^* \in (\mathrm{Ker} H)^\perp$ et des variables aléatoires $(\varepsilon_i)_{i=1,\dots,n}$ de loi normale centrée et de variance σ^2. Par souci de simplification, on suppose de plus que $\bar{\beta} = 0$. On introduit la SVD de H : $H = U\Sigma V^T$ (cf. Question 7). On note $(v_1, \dots, v_{p+1})$ les colonnes de la matrice V. Montrer qu'il existe $(b_1, \dots, b_r) \in \mathbb{R}^r$ tel que $\beta^* = \sum_{i=1}^r b_i v_i$. Montrer que

$$\mathbb{E}(\hat{\beta}_\alpha) = \sum_{i=1}^{r} \left(\frac{b_i \sigma_i^2}{\sigma_i^2 + \sigma^2/\alpha^2}\right) v_i.$$

L'estimateur $\hat{\beta}_\alpha$ de β^* est-il biaisé ? Discuter l'asymptotique $\alpha \to \infty$. On note $(l_1, \dots, l_n)$ les composantes de $\varepsilon \in \mathbb{R}^n$ dans la base $(u_1, \dots, u_n)$, où les $u_i \in \mathbb{R}^n$ sont les colonnes de la matrice U : $\varepsilon = \sum_{i=1}^n l_i u_i$. Montrer que $\left(H^T H + \frac{\sigma^2}{\alpha^2}\mathrm{Id}\right)^{-1} H^T \varepsilon = \sum_{i=1}^r \frac{\sigma_i l_i}{\sigma_i^2 + \sigma^2/\alpha^2} v_i$. En déduire que

$$\mathbb{E}(\|\hat{\beta}_\alpha - \mathbb{E}(\hat{\beta}_\alpha)\|^2) = \sigma^2 \sum_{i=1}^{r} \left(\frac{\sigma_i}{\sigma_i^2 + \sigma^2/\alpha^2}\right)^2.$$

Expliquer pourquoi le choix d'un bon paramètre α résulte d'un compromis entre biais et variance. Que se passe-t-il si $\bar{\beta} \neq 0$?