Régression pénalisée : exemple d'oncologie

15 septembre 2025

MSP *2025-2026*

On s'intéresse au jeu de données d'une étude sur les métastases du cancer du sein publiée ici. Les données (15 Mo) sont disponibles sur la pageweb du cours.

Dans cette étude, des échantillons biologiques ont été prélevés sur des tumeurs de femmes atteintes d'un cancer du sein. Ces échantillons ont été analysés à l'aide d'une puce à ADN, qui mesure simultanément l'expression de 10 000 gènes, c'est-à-dire la quantité de chaque produit génique produite par les cellules de l'échantillon. Les patientes ont été suivies afin de déterminer le temps nécessaire au cancer pour métastaser (se propager ailleurs, ce qui est une mauvaise nouvelle). Sur le plan clinique, l'objectif est d'identifier les patientes dont le pronostic est mauvais afin de leur administrer un traitement de suivi plus agressif. Sur le plan scientifique, la connaissance des gènes liés aux pires résultats peut aider à mieux comprendre la maladie et à développer des traitements à l'avenir.

Découverte du jeu de données

- Lire le jeu de données et vérifier sa dimension (nombre de lignes et nombre de colonnes).
- Vérifier que le jeu de données est sans données manquantes
- Extraire la première colonne du jeu de données comme variable réponse y et le reste des colonnes vont constituer la matrice des variables explicatives X. Penser à supprimer l'objet créer à la lecture du jeu de données.
- Créer une variable n qui correspond au nombre de patientes du jeu de données et p nombre de variables explicatives du jeu de données.
- Examiner (rapidement) la distribution de y à l'aide de la fonction `summary`.
- Examiner la distribution des moyennes des variables explicatives du jeu de données. Que remarque-t-on ?
- Examiner la distribution des écarts types des variables explicatives du jeu de données. Que remarque-t-on ?

Une étape de filtrage possible pourrait consister à supprimer les variables explicatives qui présentent une faible variance, car celles-ci sont moins susceptibles d'être informatives d'un point de vue statistique. Bien sûr, n'importe quel gène peut être important sur le plan biologique, même si son expression ne varie que très peu d'un individu à l'autre, mais nous disposons de moins de puissance statistique pour détecter ces effets. Par conséquent, si nous devions filtrer les variables explicatives, ces gènes peu informatifs pourraient être des candidats potentiels. Les méthodes sont suffisamment efficaces, donc pour l'instant, conservons toutes les variabels explicatives et normalisons les colonnes.

- Centrer uniquement la variable y.
- Centrer et réduire les colonnes de X à l'aide de la fonction `scale`.

Analyse univariée

Commençons par une version rapide de la méthode des moindres carrés ordinaires univariée pour chaque gène j à l'aide du modèle $y = x_j\beta_j + \varepsilon$. Nous allons éviter l'utilisation d'une boucle ou de la fonction lm() pour calculer les coefficients univariés $\beta_1^u, \ldots, \beta_p^u$.

- Sachant que les colonnes de X sont centrées et réduites, proposer une formule pour calculer les coefficients univariés en un seul produit matriciel.
- Calculer le vecteur composé des p estimateurs des écarts types

$$\widehat{\sigma}_j^2 = \frac{1}{n-2}\big(y - x_j\beta_j\big)^T\big(y - x_j\beta_j\big)$$

  où y est le vecteur des n observations de la variable à expliquer et x_j est la j ième colonne de la matrice X.
- Déduire le vecteur des estimateurs des écarts types des coefficients univariés $\beta_1^u, \ldots, \beta_p^u$.
- Vérifier les formules à l'aide de la fonction lm() appliquée à un ou deux gènes.
- Utiliser la fonction pchisq() pour calculer les pvalue des tests d'égalité à zéro des coefficients univariés.

Régressions LASSO et elasticnet de la librairie glmnet

- À l'aide de fonction cv.glmnet, utiliser une procédure de validation croisée à 5 folds pour choisir le meilleur modèle de régression LASSO. Afficher les noms des gènes sélectionnés ainsi que la pvalue correspondante (en univarié). Commenter
- À l'aide de fonction cv.glmnet, utiliser une procédure de validation croisée à 5 folds pour choisir le meilleur modèle de régression elasticnet ($\alpha = 0.8$. Afficher les noms des gènes sélectionnés ainsi que la pvalue correspondante (en univarié). Commenter
- Proposer une procédure pour choisir la valeur optimale de α.