Examen Final

Durée 2h. Les documents, la calculatrice, les téléphones portables, tablettes, ordinateurs ne sont pas autorisés. Les exercices sont indépendants. La qualité de la rédaction sera prise en compte.

Exercice 1. Régression Ridge

On considère le modèle de régression $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ où $\mathbf{Y}$ est un vecteur aléatoire de $\mathbb{R}^n$, $\mathbf{X}$ est une matrice de taille $n \times p$, $\boldsymbol{\beta}$ un vecteur de $\mathbb{R}^p$ et $\boldsymbol{\epsilon}$ un vecteur de $\mathbb{R}^n$ de variables aléatoires supposées indépendantes et identiquement distribuées, centrées et de variance σ^2.

1. **Régression des moindres carrés**

   (a) Rappelez la définition de l'estimateur des moindres carrés (MC) $\hat{\boldsymbol{\beta}}$ pour la régression multiple. Retrouvez son expression, en détaillant les étapes de calcul. Sous quelles hypothèses cette expression est-elle valide ?
   [On pourra utiliser la formule suivante : pour toute matrice $\mathbf{A}$ de taille $n \times p$ et vecteur $\mathbf{b}$ de taille n, le gradient de $f : \mathbf{x} \mapsto \|\mathbf{A}\mathbf{x} + \mathbf{b}\|^2$ est donné par : $\nabla_{\mathbf{x}} f(\mathbf{x}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} + \mathbf{b})$.]

   Voir le cours.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}' \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta'_j x_{ij} \right)^2 \right\} = \underset{\boldsymbol{\beta}' \in \mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}'\|^2$$

   En dérivant la fonction $f : \boldsymbol{\beta}' \mapsto \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}'\|^2$, on obtient:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

   Cette expression est valide si $\mathbf{X}^T\mathbf{X}$ est inversible, i.e. si $\mathbf{X}$ est de plein rang ($\operatorname{rg}(\mathbf{X}) = p$).

   (b) Retrouvez, en justifiant, l'espérance et la variance de cet estimateur (à σ^2 fixé).

   Voir le cours.

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{Y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$\mathbb{V}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{V}[\mathbf{Y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T[\sigma^2\mathbf{I}_n]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

   (c) On suppose que $\mathbf{X}^T\mathbf{X}$ admet la décomposition spectrale : $\mathbf{X}^T\mathbf{X} = \mathbf{V}^T\mathbf{D}\mathbf{V}$, où $\mathbf{V}$ est une matrice orthogonale, et $\mathbf{D}$ une matrice diagonale de coefficients diagonaux $(d_1, \cdots, d_p)$. Montrez que la variance de l'estimateur MC s'écrit $\mathbb{V}[\hat{\boldsymbol{\beta}}] = \sigma^2\mathbf{V}^T\mathbf{D}^{-1}\mathbf{V}$.

   Dans le cas où l'estimateur des moindres carrés est bien défini, $\mathbf{X}^T\mathbf{X}$ est symétrique inversible, donc symétrique définie positive, et toutes ses valeurs propres sont strictement positive. On peut donc écrire:

$$\mathbb{V}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{V}^T\mathbf{D}\mathbf{V})^{-1} = \sigma^2\mathbf{V}^T\mathbf{D}^{-1}\mathbf{V}$$

   car $\mathbf{V}$ est orthogonale : $\mathbf{V}^{-1} = \mathbf{V}^T$.

   (d) On définit le risque quadratique de l'estimateur MC par la formule : $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2]$. Interprétez cette définition. Qualitativement, que peut on dire d'un estimateur dont le risque quadratique est faible ?

   Le risque quadratique est l'espérance de la distance euclidienne au carré entre l'estimateur et la vrai valeur du paramètre. Si ce risque est faible, cela veut dire que l'estimateur est, en moyenne, "proche" pour la distance euclidienne de la vrai valeur du paramètre.

(e) Montrez l'égalité : $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] = \mathrm{tr}(\mathbb{V}[\hat{\boldsymbol{\beta}}])$, où on note tr la trace d'une matrice. En utilisant les notations de la question (1-c), en déduire que $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] = \sigma^2 \sum_{k=1}^{d} \frac{1}{d_k}$.

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2\right] = \mathbb{E}\left[\mathrm{tr}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\right] = \mathbb{E}\left[\mathrm{tr}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right)\right]$$
$$= \mathrm{tr}(\mathbb{V}[\hat{\boldsymbol{\beta}}]).$$

D'après (1-c), on obtient:

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2\right] = \mathrm{tr}(\sigma^2 \mathbf{V}^T \mathbf{D}^{-1} \mathbf{V}) = \sigma^2 \mathrm{tr}(\mathbf{D}^{-1}) = \sigma^2 \sum_{k=1}^{d} \frac{1}{d_k}.$$

(f) Rappelez l'énoncé du théorème de Gauss-Markov pour une régression multiple (sans le démontrer). Que peut-on en déduire sur l'optimalité du risque quadratique de l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}}$?

Voir le cours. Comme l'estimateur des moindres carrés est le BLUE, c'est aussi l'estimateur dont le risque quadratique est le plus faible. En effet, si $\tilde{\boldsymbol{\beta}}$ est un estimateur linéaire sans biais, alors $\mathbb{V}[\hat{\boldsymbol{\beta}}] - \mathbb{V}[\tilde{\boldsymbol{\beta}}]$ est une matrice positive, donc

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2\right] - \mathbb{E}\left[\left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2\right] = \mathrm{tr}(\mathbb{V}[\hat{\boldsymbol{\beta}}]) - \mathrm{tr}(\mathbb{V}[\tilde{\boldsymbol{\beta}}]) = \mathrm{tr}\left(\mathbb{V}[\hat{\boldsymbol{\beta}}] - \mathbb{V}[\tilde{\boldsymbol{\beta}}]\right) \geq 0$$

car la trace d'une matrice symétrique positive est positive.

(g) On suppose que $p > n$. Que dire de l'estimateur des moindres carrés dans ce cas ?

Si $p > n$, alors la matrice $\mathbf{X}$ ne peut pas être de plein rang ($\mathrm{rg}(\mathbf{X}) < p$), et $\mathbf{X}^T\mathbf{X}$ n'est pas inversible. L'estimateur des moindres carrés n'est donc pas bien défini.

(h) On se place dans le cas où $n > p$, $p = 2$ et $\mathbf{X} = (x\ z)$, avec deux prédicteurs x et z centrés, réduits, et corrélés: $\sum_{i=1}^n x_i = \sum_{i=1}^n z_i = 0$, $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n z_i^2 = 1$ et $\rho(x,z) = \sum_{i=1}^n x_i z_i = \sqrt{1-\delta}$, avec δ un réel positif ($0 < \delta < 1$). Explicitez $\mathbf{X}^T\mathbf{X}$ dans ce cas, et vérifiez que $1 + \rho(x,z)$ et $1 - \rho(x,z)$ sont les valeurs propres de $\mathbf{X}^T\mathbf{X}$. En déduire que $\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] = 2\sigma^2/\delta$. Comment se comporte le risque quadratique lorsque les deux variables sont très corrélées ?

On a

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i z_i \\ \sum_{i=1}^n x_i z_i & \sum_{i=1}^n y_i^2 \end{pmatrix} = \begin{pmatrix} 1 & \rho(x,z) \\ \rho(x,z) & 1 \end{pmatrix}.$$

On vérifie facilement que $(1 + \rho(x,z))$ et $(1 - \rho(x,z))$ sont les valeurs propres associées aux vecteurs propres $(1,1)$ et $(1,-1)$. D'après la question (1-e),

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2\right] = \sigma^2 \mathrm{tr}((\mathbf{X}^T\mathbf{X})^{-1}) = \sigma^2\left(\frac{1}{1+\rho(x,z)} + \frac{1}{1-\rho(x,z)}\right) = \sigma^2 \frac{2}{1-\rho(x,z)^2} = \frac{2\sigma^2}{\delta}.$$

Lorsque les deux variables sont très corrélées, δ tend vers 0 et le risque quadratique n'est pas contrôlé. Dans le cas limite où $\delta = 0$, les deux prédicteurs sont identiques, et la matrice $\mathbf{X}^T\mathbf{X}$ n'est plus de plein rang : on sort du cadre de la régression des moindres carrés classique.

2. **Régression ridge**

(a) On appelle estimateur *ridge* de paramètre λ ($\lambda > 0$) l'estimateur de $\boldsymbol{\beta}$ suivant:

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}' \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}'\|^2 + \lambda \|\boldsymbol{\beta}'\|^2 \right\}.$$

Dans toute la suite, on considère que $\lambda > 0$ est donné et fixé. Montrez l'égalité: $\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}$, avec $\mathbf{I}_p$ la matrice identité de taille p.

En utilisant la formule, on obtient:

$$\nabla_{\beta'} \left[\|\mathbf{Y} - \mathbf{X}\beta'\|^2 + \lambda \|\beta'\|^2 \right] = -2\mathbf{X}^T(-\mathbf{X}\beta' + \mathbf{Y}) + 2\lambda\beta'$$

On obtient l'estimateur en annulant ce gradient:

$$2\mathbf{X}^T\mathbf{X}\hat{\beta}_\lambda - 2\mathbf{X}^T\mathbf{Y} + 2\lambda\hat{\beta}_\lambda = 0 \iff \hat{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}$$

avec $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)$ toujours inversible lorsque $\lambda > 0$. C'est bien un estimateur linéaire en $\mathbf{Y}$.

(b) Calculez l'espérance et la variance de l'estimateur ridge (à σ^2 fixé).

$$\mathbb{E}[\hat{\beta}_\lambda] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{Y}] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}.$$
$$\mathbb{V}[\hat{\beta}_\lambda] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbb{V}[\mathbf{Y}]\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

(c) On suppose que $p > n$. Que dire de l'estimateur ridge dans ce cas ?

$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)$ est toujours inversible lorsque $\lambda > 0$, même lorsque $\mathbf{X}^T\mathbf{X}$ n'est pas de plein rang, et en particulier si $p > n$.

(d) On suppose la même décomposition spectrale $\mathbf{X}^T\mathbf{X} = \mathbf{V}^T\mathbf{D}\mathbf{V}$ que précédemment. Montrez que la variance de l'estimateur ridge s'écrit: $\mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda] = \sigma^2\mathbf{V}^T\mathbf{F}\mathbf{V}$, avec $\mathbf{F}$ une matrice diagonale telle que, pour tout $1 \leq k \leq p$: $F_{kk} = d_k/[(d_k + \lambda)^2]$.

On a $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1} = (\mathbf{V}^T(\mathbf{D} + \lambda\mathbf{I}_p)\mathbf{V})^{-1} = \mathbf{V}^T(\mathbf{D} + \lambda\mathbf{I}_p)^{-1}\mathbf{V}$ par orthogonalité de $\mathbf{V}$. Donc:

$$\begin{aligned}
\mathbb{V}[\hat{\beta}_\lambda] &= \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1} \\
&= \sigma^2(\mathbf{V}^T(\mathbf{D} + \lambda\mathbf{I}_p)^{-1}\mathbf{V})(\mathbf{V}^T\mathbf{D}\mathbf{V})(\mathbf{V}^T(\mathbf{D} + \lambda\mathbf{I}_p)^{-1}\mathbf{V}) \\
&= \sigma^2\mathbf{V}^T(\mathbf{D} + \lambda\mathbf{I}_p)^{-1}\mathbf{D}(\mathbf{D} + \lambda\mathbf{I}_p)^{-1}\mathbf{V} \\
&= \sigma^2\mathbf{V}^T\mathbf{F}\mathbf{V},
\end{aligned}$$

où $\mathbf{F} = (\mathbf{D} + \lambda\mathbf{I}_p)^{-1}\mathbf{D}(\mathbf{D} + \lambda\mathbf{I}_p)^{-1}$ est une matrice diagonale comme produit de matrices diagonales, dont les coefficients sont précisément donnés par la formule donnée dans l'énoncé.

(e) Déduire des questions précédentes que, lorsque les deux estimateurs sont bien définis, $\mathbb{V}[\hat{\boldsymbol{\beta}}] - \mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda]$ est une matrice définie positive.

Dans le cas où l'estimateur des moindres carrés est bien défini, $\mathbf{X}^T\mathbf{X}$ est symétrique inversible, donc symétrique définie positive, et toutes ses valeurs propres sont strictement positive. $\mathbb{V}[\hat{\boldsymbol{\beta}}] - \mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda]$ est alors une matrice symétrique, dont les valeurs propres dans la base $\mathbf{V}$ sont donnés par

$$g_k = \frac{1}{d_k} - \frac{d_k}{(d_k + \lambda)^2} = \frac{\lambda^2 + 2\lambda d_k}{d_k(d_k + \lambda)^2}$$

Comme $\lambda > 0$ par hypothèse, on obtient que $g_k > 0$ pour tout les $1 \leq k \leq p$. Les valeurs propres de $\mathbb{V}[\hat{\boldsymbol{\beta}}] - \mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda]$ sont donc strictement positives, et il s'agit bien d'une matrice définie positive.

(f) Est-ce que ce résultat est en contradiction avec le théorème de Gauss-Markov ?

Le théorème de Gauss-Markov nous assure que l'estimateur des moindres carré est l'estimateur ayant la plus petite variance parmi les estimateurs linéaires non biaisés. Or, l'estimateur ridge est bien linéaire, mais il est biaisé. Le fait que sa variance soit plus faible n'entre donc pas en contradiction avec le théorème de Gauss Markov.

(g) On se place dans le cadre de la question (1-h), où $p = 2$ et $\mathbf{X} = (x\ z)$, avec deux prédicteurs x et z centrés, réduits, et corrélés: $\sum_{i=1}^n x_i = \sum_{i=1}^n z_i = 0$, $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n z_i^2 = 1$ et $\rho(x,z) = \sum_{i=1}^n x_i z_i = \sqrt{1-\delta}$, avec δ un réel positif ($\delta < 1$). Montrez que dans ce cas: $\mathrm{tr}(\mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda]) \leq \frac{2\sigma^2}{(1+\lambda)^2} + \frac{\sigma^2}{\lambda^2}$. En quoi est-ce que ce comportement est différent de celui de l'estimateur des moindres carrés ?

D'après les questions précédentes:

$$\begin{aligned}\mathrm{tr}\left(\mathbb{V}\left[\hat{\beta}_\lambda\right]\right) &= \sigma^2\left(\frac{d_1}{(d_1+\lambda)^2} + \frac{d_2}{(d_2+\lambda)^2}\right)\\ &= \sigma^2\frac{1+\sqrt{1-\delta}}{(1+\sqrt{1-\delta}+\lambda)^2} + \sigma^2\frac{1-\sqrt{1-\delta}}{(1-\sqrt{1-\delta}+\lambda)^2}\\ &\leq \frac{2\sigma^2}{(1+\lambda)^2} + \frac{\sigma^2}{\lambda^2}.\end{aligned}$$

A λ fixé, la trace de la variance est donc bornée, même lorsque δ tend vers zéro, c'est à dire même lorsque la corrélation entre les deux variables explicatives s'approche de 1.

(h) Montrez l'égalité : $\mathbb{E}[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^T] = \mathbb{V}[\hat{\boldsymbol{\beta}}_\lambda] + (\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta})(\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta})^T$.
[On pourra remarquer que : $(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}}_\lambda - \mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda]) + (\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta})$*, et développer.]*

$$\begin{aligned}\mathbb{E}\left[(\hat{\beta}_\lambda - \beta)(\hat{\beta}_\lambda - \beta)^T\right] &= \mathbb{E}\left[(\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda] + \mathbb{E}[\hat{\beta}_\lambda] - \beta)(\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda] + \mathbb{E}[\hat{\beta}_\lambda] - \beta)^T\right]\\ &= \mathbb{E}\Big[(\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda])(\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda])^T + (\mathbb{E}[\hat{\beta}_\lambda] - \beta)(\mathbb{E}[\hat{\beta}_\lambda] - \beta)^T\\ &\qquad + (\mathbb{E}[\hat{\beta}_\lambda] - \beta)(\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda])^T + (\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda])(\mathbb{E}[\hat{\beta}_\lambda] - \beta)^T\Big]\\ &= \mathbb{V}\left[\hat{\beta}_\lambda\right] + (\mathbb{E}[\hat{\beta}_\lambda] - \beta)(\mathbb{E}[\hat{\beta}_\lambda] - \beta)^T.\end{aligned}$$

car, par linéarité de l'espérance:

$$\mathbb{E}[(\mathbb{E}[\hat{\beta}_\lambda] - \beta)(\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda])^T] = (\mathbb{E}[\hat{\beta}_\lambda] - \beta)\mathbb{E}[(\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda])^T] = (\mathbb{E}[\hat{\beta}_\lambda] - \beta)(\mathbb{E}[\hat{\beta}_\lambda] - \mathbb{E}[\hat{\beta}_\lambda])^T = 0.$$

(i) En déduire que le risque quadratique de l'estimateur ridge s'écrit :

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|^2\right] = \mathrm{tr}(\mathbb{V}\left[\hat{\boldsymbol{\beta}}_\lambda\right]) + \left\|\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta}\right\|^2.$$

On utilise les mêmes formules que précédement:

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\beta}_\lambda - \beta\right\|^2\right] &= \mathrm{tr}\left(\mathbb{E}\left[(\hat{\beta}_\lambda - \beta)^T(\hat{\beta}_\lambda - \beta)\right]\right)\\ &= \mathrm{tr}\left(\mathbb{V}\left[\hat{\beta}_\lambda\right] + (\mathbb{E}[\hat{\beta}_\lambda] - \beta)(\mathbb{E}[\hat{\beta}_\lambda] - \beta)^T\right)\\ &= \mathrm{tr}\left(\mathbb{V}\left[\hat{\beta}_\lambda\right]\right) + \left\|\mathbb{E}[\hat{\beta}_\lambda] - \beta\right\|^2.\end{aligned}$$

(j) Que peut on en déduire (de manière qualitative) sur les risques quadratiques de estimateurs des moindre carrés et ridge ? Est-ce que l'un des deux estimateurs est préférable en terme de risque ? On pourra regarder la différence des risques quadratiques, et discuter d'un compromis biais-variance.

La différence des risques est donné par :

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2\right]-\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}\right\|^2\right] &= \operatorname{tr}\left(\mathbb{V}\left[\hat{\boldsymbol{\beta}}\right]\right)-\operatorname{tr}\left(\mathbb{V}\left[\hat{\boldsymbol{\beta}}_\lambda\right]\right)-\left\|\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda]-\boldsymbol{\beta}\right\|^2 \\ &= \operatorname{tr}\left(\mathbb{V}\left[\hat{\boldsymbol{\beta}}\right]-\mathbb{V}\left[\hat{\boldsymbol{\beta}}_\lambda\right]\right)-\left\|\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda]-\boldsymbol{\beta}\right\|^2.\end{aligned}$$

On sait que le premier terme est positif par les questions précédentes. Si le biais de l'estimateur ridge est petit, alors le risque de l'estimateur ridge est plus petit que celui de l'estimateur des moindres carrés. En revanche, si le biais de l'estimateur ridge est grand, alors le risque de l'estimateur ridge peut devenir plus grand que celui de l'estimateur des moindres carrés. La qualité en terme de risque de l'estimateur ridge dépend donc de son biais, et il y a un compromis biais-variance à trouver, qui dépend du choix de λ.

(k) D'après ce qui précède, dans quel(s) cas particulier(s) conseilleriez vous l'utilisation de l'estimateur ridge ?

En grande dimension, lorsque l'on a beaucoup de prédicteurs, avec potentiellement $p > n$, et potentiellement des prédicteurs corrélés, alors l'estimateur ridge peut donner des estimations bien définies, et avec une variance bornée, contrairement à l'estimateur des moindres carrés classique.

Exercice 2. Musique sur Spotify

On s'intéresse à la popularité des musiques sur Spotify. On dispose d'un jeu de données de 1466 pistes de musiques, toutes publiées avant l'an 2000, comportant les variables suivantes :

- `popularity` : popularité du morceau, entre 0 (non populaire) et 100 (très populaire).
- `danceability`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`: scores, compris entre 0 et 1, décrivant diverses caractéristiques du morceau.
- `tempo` : tempo en "beat par minutes".
- `duration` : durée du morceau en ms.
- `loudness` : volume sonore moyen du morceau en décibels.
- `genre` : style musical. On ne garde que les musiques "pop", "rap" et "r&b".
- `year` : année de publication.
- `artist` : interprète.

Table 1: Extrait de quelques lignes et colones du jeu de données.

	popularity	danceability	tempo	duration	loudness	genre	year	artist
Wannabe	79	0.768	110.008	173027	-6.145	pop	1996	Spice Girls
One Way Or Another	55	0.442	162.272	217364	-5.086	pop	1978	Blondie
Straight Outta Compton	70	0.834	102.848	258688	-9.484	rap	1988	N.W.A.
Still D.R.E.	75	0.816	93.431	270587	-3.323	rap	1999	Dr. Dre
All I Want for Christmas Is You	90	0.335	150.277	241107	-7.462	r&b	1994	Mariah Carey
Ain't No Sunshine	76	0.479	79.593	125093	-11.451	r&b	1971	Bill Withers

1. On cherche à prédire la popularité d'un morceau en fonction de ses caractéristiques. On exécute les commandes suivantes dans `R` :

```r
fit1 <- lm(popularity ~
             danceability + speechiness + acousticness + tempo +
             loudness + instrumentalness + liveness + duration + year,
           data = spotify_songs)
summary(fit1)
```

```
##
## Call:
## lm(formula = popularity ~ danceability + speechiness + acousticness +
##     tempo + loudness + instrumentalness + liveness + duration +
##     year, data = spotify_songs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.135 -11.783   1.112  12.322  45.047
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.517e+02  1.379e+02   5.452 5.85e-08 ***
## danceability      1.256e+00  3.785e+00   0.332 0.739948
## speechiness       2.626e-01  4.135e+00   0.064 0.949370
## acousticness      1.684e+00  2.347e+00   0.717 0.473252
## tempo            -6.799e-03  1.567e-02  -0.434 0.664376
## loudness          7.495e-01  1.435e-01   5.223 2.02e-07 ***
## instrumentalness -2.037e+00  3.195e+00  -0.638 0.523824
## liveness          1.790e+00  2.820e+00   0.635 0.525728
## duration         -2.713e-05  8.066e-06  -3.364 0.000789 ***
## year             -3.482e-01  6.939e-02  -5.019 5.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.52 on 1456 degrees of freedom
## Multiple R-squared:  0.04359, Adjusted R-squared:  0.03768
## F-statistic: 7.374 on 9 and 1456 DF,  p-value: 1.411e-10
```

(a) À quoi correspondent les colonnes `Std. Error` et `t value` ? Ecrivez de manière formelle les tests correspondants, avec les hypothèses, l'expression de la statistique, et sa loi sous H_0. Interprétez les différents coefficients.

Voir le cours. On se place dans le cadre des hypothèses gaussiennes classiques. Pour chaque coefficient k, le test de Student correspond à $\mathcal{H}_0$: le coefficient est nul vs $\mathcal{H}_1$: il est non nul. La statistique est donnée par

$$T_k = \frac{\hat{\beta}_k}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}}},$$

sous H_0, elle suit une loi de Student à $n - p = 1466 - 10 = 1456$ degrés de libertés. `Std. Error` correspond à $\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}}$, et `t value` à T_k^{obs}.

Trois coefficients sont significativement différents de zero: l'intercept, et ceux associés au volume sonore (`loudness`), à la durée en ms (`duration`) et à l'année de publications (`year`).

(b) À quoi correspond le `Multiple R-squared` ? Donnez son expression. Interprétez.

Voir le cours.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\left\|Y - \hat{Y}\right\|^2}{\left\|Y - \mathbf{1}\bar{Y}\right\|^2}.$$

Le R^2 est proche de 0 : les moindres carrés expliqués sont petits par rapport aux moindres carrés totaux. La regression n'est pas très bonne.

(c) À quoi correspond la `F-statistic` ? Ecrivez de manière formelle le test correspondant, avec les hypothèses, l'expression de la statistique, et sa loi sous H_0. Interprétez.

Voir le cours. Il s'agit d'un test de Fisher emboîté. On teste H_0 : tous les coefficients sont nuls sauf l'intercept versus H_1 : au moins un des coefficients hors intercept est non nul. La statistisque de test est :

$$F = \frac{\|\hat{Y} - \bar{Y}\mathbf{1}\|^2/(p-1)}{\|Y - \hat{Y}\|^2/(n-p)}$$

où $p = 10$ le nombre de coefficients, et $n = 1466$ le nombre d'observations. En supposant le modèle classique gaussien, la statistique de test suit, sous H_0, une loi de Fisher à $p - 1 = 9$, $n - p = 1456$ degrés de libertés. On rejette ici l'hypothèse nulle: au moins l'un des coefficients hors intercept est non nul.

2. On complète l'analyse avec les commandes suivantes :

```
fit2 <- lm(popularity ~ loudness + duration + year, data = spotify_songs)
AIC(fit1, fit2)

##      df      AIC
## fit1 11 12395.34
## fit2  5 12385.17
```

(a) Donnez la définition de l'AIC. Quelle est son utilité ?

Voir le cours. Pour un modèle m donné avec k degrés de libertés, on a

$$AIC(m_k) = -2LL(\hat{\theta}_m) + 2k$$

où $LL(\hat{\theta}_m)$ est la log vraisemblance maximisée du modèle. L'AIC est un score de vraisemblance pénalisé : lorsque les modèles sont plus gros, leur vraisemblance est meilleure, mais la pénalité associée est auss plus élevée. On cherche à choisir un modèle qui a le plus petit AIC possible.

(b) Quel modèle de régression préférez-vous ?

On préfère le second modèle, qui a un plus petit AIC. C'est cohérent avec l'analyse précédente : on n'a gardé que ls coefficients significatifs.

3. On cherche à savoir si le style musical a une influence sur la popularité, avec l'analyse :

```
summary(lm(popularity ~ genre, data = spotify_songs))

##
## Call:
## lm(formula = popularity ~ genre, data = spotify_songs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.464 -12.626   1.374  13.374  46.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.464      1.124  44.904  < 2e-16 ***
## genrer&b      -7.313      1.288  -5.679 1.63e-08 ***
## genrerap      -5.837      1.332  -4.382 1.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.67 on 1463 degrees of freedom
## Multiple R-squared:  0.02163, Adjusted R-squared:  0.02029
## F-statistic: 16.17 on 2 and 1463 DF,  p-value: 1.133e-07
```

(a) Explicitez le modèle linéaire utilisé par `R`. À quoi correspondent les coefficients `genrer&b` et `genrerap` ?

Voir le cour. Soit P_{ik} la variable aléatoire représentant la popularité de ma musique i ($1 \leq i \leq 1466$) de genre musical k ($1 \leq k \leq 3$) avec $k = 1$ pour la pop, $k = 2$ pour le r&b et $k = 3$ pour le rap. Le modèle s'écrit:

$$P_{ik} = \mu + \beta_k + \epsilon_i$$

avec les ϵ_i iid gaussiens de variance inconnue σ^2. Pour l'identifiabilité, `R` impose la contrainte que $\beta_1 = 0$, si bien que β_k représente la différence de moyenne entre le groupe 1 de référence (pop) et le groupe $k > 1$.

`genrer&b` et `genrerap` correspondent aux β_2 et β_3 dans le modèle ci dessous : il s'agit de la différence de moyenne de popularité entre les musique pop et les musiques r&b et rap respectivement.

(b) D'après cette analyse, quelles sont les popularités moyennes respectives des musiques de pop, r&b et de rap ? Peut-on dire qu'elles sont significativement différentes ?

Les moyennes de popularité pour la pop, le r&b et le rap sont, respectivement, de 50.464, $50.464 - 7.313 = 43.151$ et $50.464 - 5.837 = 44.627$. Les coefficients `genrer&b` et `genrerap` sont significativement non nuls d'après le test de Student, la pop et le rap d'une part, et la pop et le r&b d'autre part ont bien des popularités significativement différents. En revanche, cette analyse ne nous permet pas de conclure quand à la significativité de la différence entre le r&b et le rap.

(c) D'après cette analyse, peut-on rejeter l'hypothèse suivant laquelle tous les genres de musique ont la même popularité ?

On regarde le test de Fisher du modèle, qui correspond au test de l'hypothèse $H_0 : \beta_2 = \beta_3 = 0$ vs au moins un des coefficients hors intercept est non nul. Sous H_0, tous les genres ont la même popularité moyenne. On constate que la p-valeur associée au `F-test` est très faible de `1.133e-07`. On rejette donc l'hypothèse nulle.

4. On effectue l'anova suivante:

```
anova(lm(popularity ~ loudness + year + duration * genre, data = spotify_songs))

## Analysis of Variance Table
##
## Response: popularity
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## loudness           1   5052  5051.6 18.8211 1.534e-05 ***
## year               1   9275  9274.8 34.5556 5.127e-09 ***
## duration           1   3286  3286.5 12.2446 0.0004807 ***
## genre              2   5852  2926.2 10.9021 1.997e-05 ***
## duration:genre     2    683   341.6  1.2729 0.2803353
## Residuals       1458 391331   268.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Explicitez le modèle utilisé par `R` dans cette analyse.

Soit P_{ik} la variable aléatoire représentant la popularité de ma musique i ($1 \leq i \leq 1466$) de genre musical k ($1 \leq k \leq 3$) avec $k = 1$ pour la pop, $k = 2$ pour le r&b et $k = 3$ pour le rap. On note le volume sonore du morceau i v_i, sa durée d_i, et son année a_i. On écrit :

$$P_{i,k} = \mu + \alpha_k + (\beta_d + \gamma_k) \times d_i + \beta_l l_i + \beta_a a_i + + \epsilon_i$$

Pour l'identifiabilité, `R` impose la contrainte que $\alpha_1 = 0$ et $\gamma_1 = 0$ (pop est la référence).

(b) En utilisant la commande précédente, pouvez-vous répondre aux questions suivantes ? Si vous n'avez pas assez d'élément pour répondre à une question, indiquez pourquoi. Sinon, indiquez précisément quelle partie de la sortie vous permet de conclure.

i. L'ajout du régresseur `loudness` à un modèle contenant uniquement l'intercept améliore-t-il significativement le modèle ?

Oui, il s'agit du premier test de Fisher, avec une p-valeur de `1.534e-05`.

ii. L'ajout du régresseur `year` à un modèle contenant uniquement l'intercept améliore-t-il significativement le modèle ?

On ne peut pas savoir : ici le test correspond à l'ajout de `year` a un modèle contenant déjà l'intercept et `loudness`.

iii. L'ajout du régresseur `duration` à un modèle contenant uniquement l'intercept améliore-t-il significativement le modèle ?

On ne peut pas savoir : ici le test correspond à l'ajout de `duration` a un modèle contenant déjà l'intercept, `loudness` et `year`.

iv. L'ajout du facteur `genre` à un modèle contenant déjà l'intercept et les trois regresseurs continus `loudness`, `year` et `duration` améliore-t-il significativement le modèle ?

Oui, il s'agit de l'avant-dernier dernier test de Fisher de la table, avec une p-valeur de `1.997e-05`.

v. L'ajout d'une pente spécifique à chaque genre dans la regression contre la `duration` améliore-t-il significativement le modèle ?

Non, il s'agit du dernier test de Fisher. La p-valeur des interactions est de $0.2803353 > 0.05$, on ne peut pas rejeter l'hypothèse nulle suivant laquelle tous les genres ont la même dépendance à la durée du morceau.

5. J'aimerais composer un morceau qui aie beaucoup de succès. Au vue des analyses précédentes, quels conseils pourriez vous me donner ?

Il faut produire un morceau de pop fort, court, et il y a longtemps. Mais ces variables n'expliquent qu'une très faible part du succès. (Il faut donc avoir du talent.)