Chapitre 1

Corrections des exercices

1.1 Régression linéaire simple

Exercice 1.1 (Questions de cours)
B, A, B, A.

Exercice 1.2 (Biais des estimateurs)
Les $\hat{\beta}_j$ sont fonctions de Y (aléatoire), ce sont donc des variables aléatoires. Une autre façon d'écrire $\hat{\beta}_2$ en fonction de β_2 consiste à remplacer y_i dans (**??**) par sa valeur soit

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2} = \frac{\beta_1\sum(x_i-\bar{x})+\beta_2\sum x_i(x_i-\bar{x})+\sum(x_i-\bar{x})\varepsilon_i}{\sum(x_i-\bar{x})^2} \\ &= \beta_2 + \frac{\sum(x_i-\bar{x})\varepsilon_i}{\sum(x_i-\bar{x})^2}.\end{aligned}$$

Par hypothèse $\mathbb{E}(\varepsilon_i)=0$, les autres termes ne sont pas aléatoires, le résultat est démontré. Le résultat est identique pour $\hat{\beta}_1$ car $\mathbb{E}(\hat{\beta}_1)=\mathbb{E}(\bar{y})-\bar{x}\mathbb{E}(\hat{\beta}_2)=\beta_1+\bar{x}\beta_2-\bar{x}\beta_2=\beta_1$, le résultat est démontré.

Exercice 1.3 (Variance des estimateurs)
Nous avons

$$\mathrm{V}(\hat{\beta}_2) = \mathrm{V}\left(\beta_2 + \frac{\sum(x_i-\bar{x})\varepsilon_i}{\sum(x_i-\bar{x})^2}\right).$$

Or β_2 est inconnu mais pas aléatoire et les x_i ne sont pas aléatoires donc

$$\begin{aligned}\mathrm{V}(\hat{\beta}_2) &= \mathrm{V}\left(\frac{\sum(x_i-\bar{x})\varepsilon_i}{\sum(x_i-\bar{x})^2}\right) = \frac{\mathrm{V}\left(\sum(x_i-\bar{x})\varepsilon_i\right)}{\left[\sum(x_i-\bar{x})^2\right]^2} \\ &= \frac{\sum_{i,j}(x_i-\bar{x})(x_j-\bar{x})\,\mathrm{Cov}(\varepsilon_i,\varepsilon_j)}{\left[\sum(x_i-\bar{x})^2\right]^2}.\end{aligned}$$

Or $\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=\delta_{ij}\sigma^2$ donc

$$\mathrm{V}(\hat{\beta}_2) = \frac{\sum_i(x_i-\bar{x})^2\sigma^2}{\left[\sum_i(x_i-\bar{x})^2\right]^2} = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}.$$

Plus les mesures x_i sont dispersées autour de leur moyenne, plus $\mathrm{V}(\hat{\beta}_2)$ est faible et plus l'estimation est précise. Bien sûr, plus σ^2 est faible, c'est-à-dire plus les y_i sont proches de la droite inconnue, plus l'estimation est précise.
Puisque $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$, nous avons

$$\begin{aligned}
\mathrm{V}(\hat{\beta}_1) &= \mathrm{V}\left(\bar{y} - \hat{\beta}_2\bar{x}\right) = \mathrm{V}\left(\bar{y}\right) + V(\bar{x}\hat{\beta}_2) - 2\,\mathrm{Cov}(\bar{y}, \hat{\beta}_2\bar{x}) \\
&= \mathrm{V}\left(\frac{\sum y_i}{n}\right) + \bar{x}^2\frac{\sigma^2}{\sum(x_i - \bar{x})^2} - 2\bar{x}\,\mathrm{Cov}(\bar{y}, \hat{\beta}_2) \\
&= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum(x_i - \bar{x})^2} - 2\bar{x}\sum_i \mathrm{Cov}(\bar{y}, \hat{\beta}_2).
\end{aligned}$$

Calculons

$$\begin{aligned}
\mathrm{Cov}(\bar{y}, \hat{\beta}_2) &= \frac{1}{n}\,\mathrm{Cov}\left(\sum_i \left(\beta_1 + \beta_2 x_i + \varepsilon_i\right), \frac{\sum_j (x_j - \bar{x})\varepsilon_j}{\sum_j (x_j - \bar{x})^2}\right) \\
&= \frac{1}{n}\sum_i \mathrm{Cov}\left(\varepsilon_i, \frac{\sum_j (x_j - \bar{x})\varepsilon_j}{\sum_j (x_j - \bar{x})^2}\right) \\
&= \frac{1}{\sum_j (x_j - \bar{x})^2}\sum_i \frac{1}{n}\,\mathrm{Cov}\left(\varepsilon_i, \sum_j (x_j - \bar{x})\varepsilon_j\right) \\
&= \frac{\sigma^2\frac{1}{n}\sum_i (x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} = 0.
\end{aligned}$$

Nous avons donc

$$\mathrm{V}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2\sum x_i^2}{n\sum(x_i - \bar{x})^2}.$$

Là encore, plus σ^2 est faible, c'est-à-dire plus les y_i sont proches de la droite inconnue, plus l'estimation est précise. Plus les valeurs x_i sont dispersées autour de leur moyenne, plus la variance de l'estimateur sera faible. De même, une faible moyenne $\bar{x}$ en valeur absolue contribue à bien estimer β_1.

Exercice 1.4 (Covariance de $\hat{\beta}_1$ et $\hat{\beta}_2$)
Nous avons

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \mathrm{Cov}(\bar{y} - \hat{\beta}_2\bar{x}, \hat{\beta}_2) = \mathrm{Cov}(\bar{y}, \hat{\beta}_2) - \bar{x}\,\mathrm{V}(\hat{\beta}_2) = -\frac{\sigma^2\bar{x}}{\sum(x_i - \bar{x})^2}.$$

La covariance entre β_1 et β_2 est négative. L'équation $\bar{y} = \hat{\beta}_1 + \hat{\beta}_2\bar{x}$ indique que la droite des MC passe par le centre de gravité du nuage $(\bar{x}, \bar{y})$. Supposons $\bar{x}$ positif, nous voyons bien que, si nous augmentons la pente, l'ordonnée à l'origine va diminuer et vice versa. Nous retrouvons donc le signe négatif pour la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$.

Exercice 1.5 (Théorème de Gauss-Markov)
L'estimateur des MC s'écrit $\hat{\beta}_2 = \sum_{i=1}^n p_i y_i$, avec $p_i = (x_i - \bar{x})/\sum(x_i - \bar{x})^2$.
Considérons un autre estimateur $\tilde{\beta}_2$ linéaire en y_i et sans biais, c'est-à-dire

$$\tilde{\beta}_2 = \sum_{i=1}^n \lambda_i y_i.$$

Montrons que $\sum \lambda_i = 0$ et $\sum \lambda_i x_i = 1$. L'égalité $\mathbb{E}(\tilde{\beta_2}) = \beta_1 \sum \lambda_i + \beta_2 \sum \lambda_i x_i + \sum \lambda_i \mathbb{E}(\varepsilon_i)$ est vraie pour tout β_2 et $\tilde{\beta_2}$ est sans biais donc $\mathbb{E}(\tilde{\beta_2}) = \beta_2$ pour tout β_2, c'est-à-dire que $\sum \lambda_i = 0$ et $\sum \lambda_i x_i = 1$.
Montrons que $\mathrm{V}(\tilde{\beta_2}) \geq \mathrm{V}(\hat{\beta}_2)$.

$$\mathrm{V}(\tilde{\beta_2}) = \mathrm{V}(\tilde{\beta_2} - \hat{\beta}_2 + \hat{\beta}_2) = \mathrm{V}(\tilde{\beta_2} - \hat{\beta}_2) + \mathrm{V}(\hat{\beta}_2) + 2\,\mathrm{Cov}(\tilde{\beta_2} - \hat{\beta}_2, \hat{\beta}_2).$$

$$\mathrm{Cov}(\tilde{\beta_2} - \hat{\beta}_2, \hat{\beta}_2) = \mathrm{Cov}(\tilde{\beta_2}, \hat{\beta}_2) - \mathrm{V}(-\hat{\beta}_2) = \frac{\sigma^2 \sum \lambda_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} - \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = 0,$$

et donc

$$\mathrm{V}(\tilde{\beta_2}) = \mathrm{V}(\tilde{\beta_2} - \hat{\beta}_2) + \mathrm{V}(\hat{\beta}_2).$$

Une variance est toujours positive et donc

$$\mathrm{V}(\tilde{\beta_2}) \geq \mathrm{V}(\hat{\beta}_2).$$

Le résultat est démontré. On obtiendrait la même chose pour $\hat{\beta}_1$.

Exercice 1.6 (Somme des résidus)
Il suffit de remplacer les résidus par leur définition et de remplacer $\hat{\beta}_1$ par son expression

$$\sum_i \hat{\varepsilon}_i = \sum_i (y_i - \bar{y} + \hat{\beta}_2 \bar{x} - \hat{\beta}_2 x_i) = \sum_i (y_i - \bar{y}) - \hat{\beta}_2 \sum_i (x_i - \bar{x}) = 0.$$

Exercice 1.7 (Estimateur de la variance du bruit)
Récrivons les résidus en constatant que $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$ et $\beta_1 = \bar{y} - \beta_2 \bar{x} - \bar{\varepsilon}$,

$$\begin{aligned} \hat{\varepsilon}_i &= \beta_1 + \beta_2 x_i + \varepsilon_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \\ &= \bar{y} - \beta_2 \bar{x} - \bar{\varepsilon} + \beta_2 x_i + \varepsilon_i - \bar{y} + \hat{\beta}_2 \bar{x} - \hat{\beta}_2 x_i \\ &= (\beta_2 - \hat{\beta}_2)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}). \end{aligned}$$

En développant et en nous servant de l'écriture de $\hat{\beta}_2$ donnée dans la solution de l'exercice **??**, nous avons

$$\begin{aligned} \sum \hat{\varepsilon}_i^2 &= (\beta_2 - \hat{\beta}_2)^2 \sum (x_i - \bar{x})^2 + \sum (\varepsilon_i - \bar{\varepsilon})^2 + 2(\beta_2 - \hat{\beta}_2) \sum (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) \\ &= (\beta_2 - \hat{\beta}_2)^2 \sum (x_i - \bar{x})^2 + \sum (\varepsilon_i - \bar{\varepsilon})^2 - 2(\beta_2 - \hat{\beta}_2)^2 \sum (x_i - \bar{x})^2. \end{aligned}$$

Prenons en l'espérance

$$\mathbb{E}\left(\sum \hat{\varepsilon_i}^2\right) = \mathbb{E}\left(\sum (\varepsilon_i - \bar{\varepsilon})^2\right) - \sum (x_i - \bar{x})^2 \,\mathrm{V}(\hat{\beta}_2) = (n-2)\sigma^2.$$

Exercice 1.8 (Variance de $\hat{y}^p_{n+1}$)
Calculons la variance

$$\begin{aligned} \mathrm{V}\left(\hat{y}^p_{n+1}\right) &= \mathrm{V}\left(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}\right) = \mathrm{V}(\hat{\beta}_1) + x_{n+1}^2 \,\mathrm{V}(\hat{\beta}_2) + 2x_{n+1} \,\mathrm{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) \\ &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \left(\frac{\sum x_i^2}{n} + x_{n+1}^2 - 2x_{n+1}\bar{x}\right) \\ &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \left(\frac{\sum (x_i - \bar{x})^2}{n} + \bar{x}^2 + x_{n+1}^2 - 2x_{n+1}\bar{x}\right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right). \end{aligned}$$

Plus la valeur à prévoir s'éloigne du centre de gravité, plus la valeur prévue sera variable (i.e. de variance élevée).

Exercice 1.9 (Variance de l'erreur de prévision)
Nous obtenons la variance de l'erreur de prévision en nous servant du fait que y_{n+1} est fonction de ε_{n+1} seulement, alors que $\hat{y}^p_{n+1}$ est fonction des autres ε_i, $i = 1, \cdots, n$. Les deux quantités ne sont pas corrélées. Nous avons alors

$$\mathrm{V}(\hat{\varepsilon}^p_{n+1}) = \mathrm{V}\left(y_{n+1} - \hat{y}^p_{n+1}\right) = \mathrm{V}(y_{n+1}) + \mathrm{V}(\hat{y}^p_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right).$$

Exercice 1.10 (R^2 et coefficient de corrélation)
Le coefficient R^2 s'écrit

$$\begin{aligned}
\mathrm{R}^2 &= \frac{\sum_{i=1}^n \left(\hat{\beta}_1 + \hat{\beta}_2 x_i - \bar{y}\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n \left(\bar{y} - \hat{\beta}_2 \bar{x} + \hat{\beta}_2 x_i - \bar{y}\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\hat{\beta}_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right]^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2 \sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} = \rho^2(X, Y).
\end{aligned}$$

Exercice 1.11 (Les arbres)
Le calcul donne

$$\hat{\beta}_1 = \frac{6.26}{28.29} = 0.22 \qquad \hat{\beta}_0 = 18.34 - 0.22 \times 34.9 = 10.662.$$

Nous nous servons de la propriété $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ pour obtenir

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{20} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{20} (y_i - \bar{y})^2} = \frac{\sum_{i=1}^{20} (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})^2}{\sum_{i=1}^{20} (y_i - \bar{y})^2} = 0.22^2 \times \frac{28.29}{2.85} = 0.48.$$

Les statistiques de test valent 5.59 pour β_0 et 4.11 pour β_1. Elles sont à comparer à un fractile de la loi de Student admettant 18 ddl, soit 2.1. Nous rejetons dans les deux cas l'hypothèse de nullité du coefficient. Nous avons modélisé la hauteur par une fonction affine de la circonférence, il semblerait évident que la droite passe par l'origine (un arbre admettant un diamètre proche de zéro doit être petit), or nous rejetons l'hypothèse $\beta_0 = 0$. Les données mesurées indiquent des arbres dont la circonférence varie de 26 à 43 cm, les estimations des paramètres du modèle sont valides pour des données proches de $[26; 43]$.

Exercice 1.12 (Modèle quadratique)
Les modèles sont

$$\begin{aligned}
\texttt{O3} &= \beta_1 + \beta_2 \texttt{T12} + \varepsilon \quad \text{modèle classique,} \\
\texttt{O3} &= \gamma_1 + \gamma_2 \texttt{T12}^2 + \epsilon \quad \text{modèle demandé.}
\end{aligned}$$

L'estimation des paramètres donne

$$\begin{aligned}
\widehat{\texttt{O3}} &= 31.41 + 2.7\ \texttt{T12} \qquad \mathrm{R}^2 = 0.28 \quad \text{modèle classique,} \\
\widehat{\texttt{O3}} &= 53.74 + 0.075\ \texttt{T12}^2 \quad \mathrm{R}^2 = 0.35 \quad \text{modèle demandé.}
\end{aligned}$$

Les deux modèles ont le même nombre de paramètres, nous préférons le modèle quadratique car le R^2 est plus élevé.

1.2 Régression linéaire multiple

Exercice 2.1 (Questions de cours)
A, A, B, B, B, C.

Exercice 2.2 (Covariance de $\hat{\varepsilon}$ et de $\hat{Y}$)
Les matrices X, P_X et $P_{X^\perp}$ sont non aléatoires. Nous avons alors

$$\begin{aligned}\mathrm{Cov}(\hat{\varepsilon},\hat{Y}) &= \mathbb{E}(\hat{\varepsilon}\hat{Y}') - \mathbb{E}(\hat{\varepsilon})\mathbb{E}(\hat{Y}') \\ &= \mathbb{E}\left[P_{X^\perp}\varepsilon(P_X(X\beta+\varepsilon))'\right] \\ &= \mathbb{E}(P_{X^\perp}\varepsilon\beta' X') + \mathbb{E}(P_{X^\perp}\varepsilon\varepsilon' P_X) \\ &= 0 + P_{X^\perp}\sigma^2 P_X = 0.\end{aligned}$$

Exercice 2.3 (Théorème de Gauss-Markov)
Nous devons montrer que, parmi tous les estimateurs linéaires sans biais, l'estimateur de MC est celui qui a la plus petite variance. La linéarité de $\hat{\beta}$ est évidente. Calculons sa variance :

$$\mathrm{V}(\hat{\beta}) = \mathrm{V}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\,\mathrm{V}(Y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

Nous allons montrer que, pour tout autre estimateur $\tilde{\beta}$ de β linéaire et sans biais, $\mathrm{V}(\tilde{\beta}) \geq \mathrm{V}(\hat{\beta})$. Décomposons la variance de $\tilde{\beta}$

$$\mathrm{V}(\tilde{\beta}) = \mathrm{V}(\tilde{\beta}-\hat{\beta}+\hat{\beta}) = \mathrm{V}(\tilde{\beta}-\hat{\beta}) + \mathrm{V}(\hat{\beta}) - 2\,\mathrm{Cov}(\tilde{\beta}-\hat{\beta},\hat{\beta}).$$

Les variances étant définies positives, si nous montrons que $\mathrm{Cov}(\tilde{\beta}-\hat{\beta},\hat{\beta}) = 0$, nous aurons fini la démonstration.
Puisque $\tilde{\beta}$ est linéaire, $\tilde{\beta} = AY$. De plus, nous savons qu'il est sans biais, c'est-à-dire $\mathbb{E}(\tilde{\beta}) = \beta$ pour tout β, donc $AX = I$. La covariance devient :

$$\begin{aligned}\mathrm{Cov}(\tilde{\beta}-\hat{\beta},\hat{\beta}) &= \mathrm{Cov}(AY,(X'X)^{-1}X'Y) - \mathrm{V}(\hat{\beta}) \\ &= \sigma^2 AX(X'X)^{-1} - \sigma^2(X'X)^{-1} = 0.\end{aligned}$$

Exercice 2.4 (Représentation des variables)
Nous représentons les données dans $\mathbb{R}^2$ pour le premier jeu et dans $\mathbb{R}^3$ pour le second.

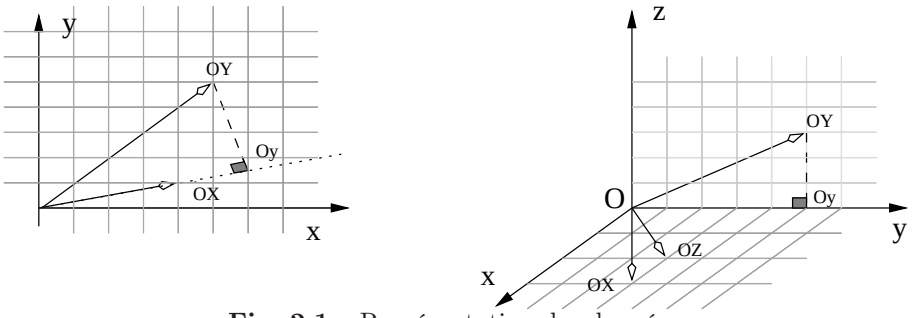

Fig. 2.1 – Représentation des données.

Dans le premier modèle, nous projetons Y sur l'espace engendré par X, soit la droite de vecteur directeur $\overrightarrow{OX}$. Nous trouvons par le calcul $\hat{\beta} = 1.47$, résultat que nous aurions pu trouver graphiquement car $\overrightarrow{OY} = \hat{\beta}.\overrightarrow{OX}$.
Considérons $\mathbb{R}^3$ muni de la base orthonormée $(\vec{i}, \vec{j}, \vec{k})$. Les vecteurs $\overrightarrow{OX}$ et $\overrightarrow{OZ}$ engendrent le même plan que celui engendré par $(\vec{i}, \vec{j})$. La projection de Y sur ce plan donne $\overrightarrow{OY}$. Il est quasiment impossible de trouver $\hat{\beta}$ et $\hat{\gamma}$ graphiquement mais nous trouvons par le calcul $\hat{\beta} = -3.33$ et $\hat{\gamma} = 5$.

Exercice 2.5 (Modèles emboîtés)
Nous obtenons

$$\hat{Y}_p = X\hat{\beta} \quad \text{et} \quad \hat{Y}_q = X_q\hat{\gamma}.$$

Par définition du R^2, il faut comparer la norme au carré des vecteurs $\hat{Y}_p$ et $\hat{Y}_q$. Notons les espaces engendrés par les colonnes de X_q et X, $\Im_{X_q}$ et $\Im_X$, nous avons $\Im_{X_q} \subset \Im_X$. Nous obtenons alors

$$\begin{aligned} \hat{Y}_p = P_{X_p}Y = (P_{X_q} + P_{X_q^\perp})P_{X_p}Y &= P_{X_q}P_{X_p}Y + P_{X_q^\perp}P_{X_p}Y \\ &= P_{X_q}Y + P_{X_q^\perp \cap X_p}Y \\ &= \hat{Y}_q + P_{X_q^\perp \cap X_p}Y. \end{aligned}$$

En utilisant le théorème de Pythagore, nous avons

$$\|\hat{Y}_p\|^2 = \|\hat{Y}_q\|^2 + \|P_{X_q^\perp \cap X_p}Y\|^2 \geq \|\hat{Y}_q\|^2,$$

d'où

$$\mathrm{R}^2(p) = \frac{\|\hat{Y}_p\|^2}{\|Y\|^2} \geq \frac{\|\hat{Y}_q\|^2}{\|Y\|^2} = \mathrm{R}^2(q).$$

En conclusion, lorsque les modèles sont emboîtés $\Im_{X_q} \subset \Im_X$, le R^2 du modèle le plus grand (ayant le plus de variables) sera toujours plus grand que le R^2 du modèle le plus petit.

Exercice 2.6
La matrice $X'X$ est symétrique, n vaut 30 et $\bar{x} = \bar{z} = 0$. Le coefficient de corrélation

$$\rho_{x,z} = \frac{\sum_{i=1}^{30}(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{30}(x_i - \bar{x})^2 \sum_{i=1}^{30}(z_i - \bar{z})^2}} = \frac{\sum_{i=1}^{30} x_i z_i}{\sqrt{\sum_{i=1}^{30} x_i^2 \sum_{i=1}^{30} z_i^2}} = \frac{7}{\sqrt{150}} = 0.57.$$

Nous avons

$$y_i = -2 + x_i + z_i + \hat{\varepsilon}_i$$

et la moyenne vaut alors

$$\bar{y} = -2 + \bar{x} + \bar{z} + \frac{1}{n}\sum_i \hat{\varepsilon}_i.$$

La constante étant dans le modèle, la somme des résidus est nulle car le vecteur $\hat{\varepsilon}$ est orthogonal au vecteur $\mathbb{1}$. Nous obtenons donc que la moyenne de Y vaut 2 car $\bar{x} = 0$ et $\bar{z} = 0$. Nous obtenons en développant

$$\begin{aligned} \|\hat{Y}\|^2 &= \sum_{i=1}^{30}(-2 + x_i + 2z_i)^2 \\ &= 4 + 10 + 60 + 14 = 88. \end{aligned}$$

Par le théorème de Pythagore, nous concluons que

$$\mathrm{SCT} = \mathrm{SCE} + \mathrm{SCR} = 88 + 12 = 100.$$

Exercice 2.7 (Régression orthogonale)
Les vecteurs étant orthogonaux, nous avons $\mathcal{S}_X = \mathcal{S}_U \overset{\perp}{\oplus} \mathcal{S}_V$. Nous pouvons alors écrire

$$\begin{aligned} \hat{Y}_X = P_X Y &= (P_U + P_{U^\perp})P_X Y \\ &= P_U P_X Y + P_{U^\perp} P_X Y = P_U Y + P_{U^\perp \cap X} Y \\ &= \hat{Y}_U + \hat{Y}_V. \end{aligned}$$

La suite de l'exercice est identique. En conclusion, effectuer une régression multiple sur des variables orthogonales revient à effectuer p régressions simples.

Exercice 2.8 (Centrage, centrage-réduction et coefficient constant)

1. Comme la dernière colonne de X, notée X_p vaut $\mathbb{1}_n$ sa moyenne empirique vaut 1 et la variable centrée issue de X_p est donc $X_p - 1 \times \mathbb{1}_n = \mathbf{0}_n$.
2. Nous avons le modèle sur variable centrée

$$\begin{aligned} \tilde{Y} &= \tilde{X}\tilde{\beta} + \varepsilon \\ Y - \bar{Y}\mathbb{1}_n &= \sum_{j=1}^{p-1}(X_j - \bar{X}_j\mathbb{1}_n)\tilde{\beta}_j + \varepsilon \\ Y &= \sum_{j=1}^{p-1}\tilde{\beta}_j X_j + \Big(\bar{Y} - \sum_{j=1}^{p-1}\bar{X}_j\tilde{\beta}_j\Big)\mathbb{1}_n + \varepsilon. \end{aligned}$$

   En identifiant cela donne

$$\begin{aligned} \beta_j &= \tilde{\beta}_j,\ \forall j \in \{1, \dots, p-1\}, \\ \beta_p &= \bar{Y}\mathbb{1}_n - \sum_{j=1}^{p-1}\bar{X}_j\tilde{\beta}_j. \end{aligned} \tag{2.1}$$

   Si l'on utilise des variables centrées dans le modèle de régression, on ne met pas de colonne $\mathbb{1}$ (pas de coefficient constant - *intercept*). Les coefficients du modèle sur les variables originales sont égaux à ceux sur les variables centrées et le coefficient constant est donné par la formule (2.1).
3. Maintenant les variables explicatives sont centrées et réduites :

$$\begin{aligned} \tilde{Y} &= \tilde{X}\tilde{\beta} + \varepsilon \\ Y - \bar{Y}\mathbb{1}_n &= \sum_{j=1}^{p-1}\frac{(X_j - \bar{X}_j\mathbb{1}_n)}{\hat{\sigma}_{X_j}}\tilde{\beta}_j + \varepsilon \\ Y &= \sum_{j=1}^{p-1}\frac{\tilde{\beta}_j}{\hat{\sigma}_{X_j}}X_j + \Big(\bar{Y} - \sum_{j=1}^{p-1}\bar{X}_j\frac{\tilde{\beta}_j}{\hat{\sigma}_{X_j}}\Big)\mathbb{1}_n + \varepsilon. \end{aligned}$$

En identifiant cela donne

$$\begin{aligned}\beta_j &= \frac{\tilde{\beta}_j}{\hat{\sigma}_{X_j}},\ \forall j \in \{1,\dots,p-1\},\\ \beta_p &= \bar{Y}\mathbb{1}_n - \sum_{j=1}^{p-1} \bar{X}_j \frac{\tilde{\beta}_j}{\hat{\sigma}_{X_j}}.\end{aligned}$$

Nous obtenons ici que les coefficients du modèle sur les variables originales sont égaux à ceux sur les variables centrées-réduites divisés par l'écart-type empirique des variables explicatives. Plus la variable explicative X_j est dispersée, plus son coefficient β_j sera réduit par rapport à $\tilde{\beta}_j$. Le coefficient constant est donné par la formule ci-dessus.

4. La variable à expliquer Y est elle aussi centrée-réduite :

$$\begin{aligned}\tilde{Y} &= \tilde{X}\tilde{\beta} + \tilde{\varepsilon}\\ \frac{Y - \bar{Y}\mathbb{1}_n}{\hat{\sigma}_Y} &= \sum_{j=1}^{p-1} \frac{(X_j - \bar{X}_j\mathbb{1}_n)}{\hat{\sigma}_{X_j}}\tilde{\beta}_j + \tilde{\varepsilon}\\ Y &= \hat{\sigma}_Y \sum_{j=1}^{p-1} \frac{\tilde{\beta}_j}{\hat{\sigma}_{X_j}} X_j + \Big(\bar{Y} - \hat{\sigma}_Y \sum_{j=1}^{p-1} \bar{X}_j \frac{\tilde{\beta}_j}{\hat{\sigma}_{X_j}}\Big)\mathbb{1}_n + \hat{\sigma}_Y\tilde{\varepsilon}.\end{aligned}$$

En identifiant cela donne

$$\begin{aligned}\beta_j &= \hat{\sigma}_Y \frac{\tilde{\beta}_j}{\hat{\sigma}_{X_j}},\ \forall j \in \{1,\dots,p-1\},\\ \beta_p &= \bar{Y}\mathbb{1}_n - \hat{\sigma}_Y \sum_{j=1}^{p-1} \bar{X}_j \frac{\tilde{\beta}_j}{\hat{\sigma}_{X_j}},\\ \varepsilon &= \hat{\sigma}_Y\tilde{\varepsilon}.\end{aligned}$$

L'écart-type empirique de Y entre en jeu et nous constatons que les résidus du modèle « centré-réduit » sont égaux à ceux initiaux divisés par l'écart-type empirique de Y.

Exercice 2.9 (Moindres carrés contraints)
L'estimateur des MC vaut

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

calculons maintenant l'estimateur contraint. Nous pouvons procéder de deux manières différentes. La première consiste à écrire le lagrangien

$$\mathcal{L} = S(\beta) - \lambda'(R\beta - r).$$

Les conditions de Lagrange permettent d'obtenir un minimum

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial\beta} = -2X'Y + 2X'X\hat{\beta}_c - R'\hat{\lambda} = 0,\\ \dfrac{\partial\mathcal{L}}{\partial\lambda} = R\hat{\beta}_c - r = 0,\end{cases}$$

Multiplions à gauche la première égalité par $R(X'X)^{-1}$, nous obtenons

$$\begin{aligned} -2R(X'X)^{-1}X'Y + 2R(X'X)^{-1}X'X\hat{\beta}_c - R(X'X)^{-1}R'\hat{\lambda} &= 0 \\ -2R(X'X)^{-1}X'Y + 2R\hat{\beta}_c - R(X'X)^{-1}R'\hat{\lambda} &= 0 \\ -2R(X'X)^{-1}X'Y + 2r - R(X'X)^{-1}R'\hat{\lambda} &= 0. \end{aligned}$$

Nous obtenons alors pour $\hat{\lambda}$

$$\hat{\lambda} = 2\left[R(X'X)^{-1}R'\right]^{-1}\left[r - R(X'X)^{-1}X'Y\right].$$

Remplaçons ensuite $\hat{\lambda}$

$$\begin{aligned} -2X'Y + 2X'X\hat{\beta}_c - R'\hat{\lambda} &= 0 \\ -2X'Y + 2X'X\hat{\beta}_c - 2R'\left[R(X'X)^{-1}R'\right]^{-1}\left[r - R(X'X)^{-1}X'Y\right] &= 0, \end{aligned}$$

d'où nous calculons $\hat{\beta}_c$

$$\begin{aligned} \hat{\beta}_c &= (X'X)^{-1}X'Y + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(r - R\hat{\beta}) \\ &= \hat{\beta} + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(r - R\hat{\beta}). \end{aligned}$$

La fonction $S(\beta)$ à minimiser est une fonction convexe sur un ensemble convexe (contraintes linéaires), le minimum est donc unique.
Une autre façon de procéder consiste à utiliser les projecteurs. Supposons pour commencer que $r = 0$, la contrainte vaut donc $R\beta = 0$. Calculons analytiquement le projecteur orthogonal sur $\Im_0$. Rappelons que $\dim(\Im_0) = p - q$, nous avons de plus

$$\begin{aligned} R\beta &= 0 \quad \Leftrightarrow \quad \beta \in Ker(R) \\ R(X'X)^{-1}X'X\beta &= 0 \\ U'X\beta &= 0 \quad \text{où} \quad U = X(X'X)^{-1}R'. \end{aligned}$$

Nous avons donc que $\forall\beta \in \ker(R)$, $U'X\beta = 0$, c'est-à-dire que $\Im_U$, l'espace engendré par les colonnes de U, est orthogonal à l'espace engendré par $X\beta$, $\forall\beta \in \ker(R)$. Nous avons donc que $\Im_U \perp \Im_0$. Comme $U = X[(X'X)^{-1}R']$, $\Im_U \subset \Im_X$. En résumé, nous avons

$$\Im_U \subset \Im_X \quad \text{et} \quad \Im_U \perp \Im_0 \quad \text{donc} \quad \Im_U \subset (\Im_X \cap \Im_0^{\perp}).$$

Afin de montrer que les colonnes de U engendrent $\Im_X \cap \Im_0^{\perp}$, il faut démontrer que la dimension des deux sous-espaces est égale. Or le rang de U vaut q (R' est de rang q, $(X'X)^{-1}$ est de rang p et X est de rang p) donc la dimension de $\Im_U$ vaut q. De plus, nous avons vu que

$$\Im_X = \Im_0 \overset{\perp}{\oplus} \left(\Im_0^{\perp} \cap \Im_X\right)$$

et donc, en passant aux dimensions des sous-espaces, nous en déduisons que $\dim(\Im_0^{\perp} \cap \Im_X) = q$. Nous venons de démontrer que

$$\Im_U = \Im_X \cap \Im_0^{\perp}.$$

Le projecteur orthogonal sur $\Im_U = \Im_X \cap \Im_0^{\perp}$ s'écrit

$$P_U = U(U'U)^{-1}U' = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'.$$

Nous avons alors

$$\begin{aligned} \hat{Y} - \hat{Y}_0 &= P_U Y \\ X\hat{\beta} - X\hat{\beta}_0 &= X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'Y \\ &= X(X'X)^{-1}R[R(X'X)^{-1}R']^{-1}R\hat{\beta}. \end{aligned}$$

Cela donne

$$\hat{\beta}_0 = \hat{\beta} - (X'X)^{-1}R[R(X'X)^{-1}R']^{-1}R\hat{\beta}.$$

Si maintenant $r \neq 0$, nous avons alors un sous-espace affine défini par $\{\beta \in \mathbb{R}^p : R\beta = r\}$ dans lequel nous cherchons une solution qui minimise les moindres carrés. Un sous-espace affine peut être défini de manière équivalente par un point particulier $\beta_p \in \mathbb{R}^p$ tel que $R\beta_p = r$ et le sous-espace vectoriel associé $\Im_0^v = \{\beta \in \mathbb{R}^p : R\beta = 0\}$. Les points du sous-espace affine sont alors $\{\beta_0 \in \mathbb{R}^p : \beta_0 = \beta_p + \beta_0^v, \beta_0^v \in \Im_0^v \quad et \quad \beta_p : R\beta_p = r\}$. La solution qui minimise les moindres carrés, notée $\hat{\beta}_0$, est élément de ce sous-espace affine et est définie par $\hat{\beta}_0 = \beta_p + \hat{\beta}_0^v$ où

$$\hat{\beta}_0^v = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\hat{\beta}.$$

Nous savons que $R\beta_p = r$ donc

$$R\beta_p = [R(X'X)^{-1}R'][R(X'X)^{-1}R']^{-1}r$$

donc une solution particulière est $\beta_p = (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}r$. La solution $\hat{\beta}_0$ qui minimise les moindres carrés sous la contrainte $R\beta = r$ est alors

$$\begin{aligned} \hat{\beta}_0 &= \beta_p + \hat{\beta}_0^v \\ &= (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}r + \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\hat{\beta} \\ &= \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}). \end{aligned}$$

1.3 Inférence dans le modèle gaussien

Exercice 3.1 (Questions de cours)
A, C, A, B, B.

Exercice 3.2 (Théorème ??)
L'IC (i) découle de la propriété (i) de la proposition **??**. La propriété (ii) donnant un IC pour σ^2 découle de la loi de $\hat{\sigma}^2$. Enfin, la propriété (iii) est une conséquence de la loi obtenue propriété (ii) de la proposition **??**.

Exercice 3.3 (Test et R^2)
En utilisant l'orthogonalité des sous-espaces (fig. **??** p. **??**) et le théorème de Pythagore, nous avons

$$\|\hat{Y}_0 - \hat{Y}\|^2 = \|\hat{\varepsilon}_0\|^2 - \|\hat{\varepsilon}\|^2.$$

Nous pouvons le démontrer de la manière suivante :

$$\begin{aligned}\|\hat{Y}_0 - \hat{Y}\|^2 &= \|\hat{Y}_0 - Y + Y - \hat{Y}\|^2 \\ &= \|\hat{\varepsilon}_0\|^2 + \|\hat{\varepsilon}\|^2 + 2\langle \hat{Y}_0 - Y, Y - \hat{Y}\rangle \\ &= \|\hat{\varepsilon}_0\|^2 + \|\hat{\varepsilon}\|^2 - 2\langle Y - \hat{Y}_0, Y - \hat{Y}\rangle \\ &= \|\hat{\varepsilon}_0\|^2 + \|\hat{\varepsilon}\|^2 - 2\langle P_{X_0^\perp} Y, P_{X^\perp} Y\rangle \\ &= \|\hat{\varepsilon}_0\|^2 + \|\hat{\varepsilon}\|^2 - 2\langle (P_{X^\perp} + P_X) P_{X_0^\perp} Y, P_{X^\perp} Y\rangle.\end{aligned}$$

Or $\Im(X_0) \subset \Im(X)$, nous avons donc $P_{X^\perp} P_{X_0^\perp} = P_{X^\perp}$. De plus, $\hat{\varepsilon} = P_{X^\perp} Y$, cela donne

$$\begin{aligned}\langle (P_{X^\perp} + P_X) P_{X_0^\perp} Y, P_{X^\perp} Y\rangle &= \langle P_{X^\perp} Y, P_{X^\perp} Y\rangle + \langle P_X P_{X_0^\perp} Y, P_{X^\perp} Y\rangle \\ &= \|\hat{\varepsilon}\|^2 + 0.\end{aligned}$$

Le résultat est démontré, revenons à la statistique de test. Introduisons les différentes écritures du R^2

$$\mathrm{R}^2 = \frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{Y}\|^2}.$$

La statistique de test vaut

$$\begin{aligned}F &= \frac{\|\hat{\varepsilon}_0\|^2 - \|\hat{\varepsilon}\|^2}{\|Y - \hat{Y}\|^2} \frac{n-p}{p-p_0} \\ &= \frac{\|\hat{\varepsilon}_0\|^2/\|Y - \bar{Y}\|^2 - \|\hat{\varepsilon}\|^2/\|Y - \bar{Y}\|^2}{\|Y - \hat{Y}\|^2/\|Y - \bar{Y}\|^2} \frac{n-p}{p-p_0},\end{aligned}$$

nous obtenons

$$F = \frac{\mathrm{R}^2 - \mathrm{R}_0^2}{1 - \mathrm{R}^2} \frac{n-p}{p-p_0},$$

soit le résultat annoncé. Cette dernière quantité est toujours positive car $\mathrm{R}_0^2 \leq \mathrm{R}^2$ et nous avons là un moyen de tester des modèles emboîtés *via* le coefficient de détermination.

Exercice 3.4 (Ozone)
Les résultats sont dans l'ordre

$$6.2, 0.8, 6.66, -1.5, -1, 50, 5, 124.$$

La statistique de test de nullité du paramètre se trouve dans la troisième colonne, nous conservons H_0 pour les paramètres associés à Ne9 et Ne12, et la rejetons pour les autres. La statistique de test de nullité simultanée des paramètres autres que la constante vaut 50. Nous rejetons H_0.
Nous sommes en présence de modèles emboîtés, nous pouvons appliquer la formule adaptée (voir l'exercice précédent) :

$$\begin{aligned}F &= \frac{\mathrm{R}^2 - \mathrm{R}_0^2}{1 - \mathrm{R}^2} \frac{n-p}{p-p_0} \\ &= \frac{0.66 - 0.5}{1 - 0.66} \frac{124}{2} = 29.\end{aligned}$$

Nous conservons H_0, c'est-à-dire le modèle le plus simple.

Exercice 3.5 (Equivalence du test T et du test F)
Récrivons la statistique de test F, en se rappelant que X_0 est la matrice X privée de sa j^{e} colonne, celle correspondant au coefficient que l'on teste :

$$F = \frac{\|X\hat{\beta} - P_{X_0}X\hat{\beta}\|^2}{\hat{\sigma}^2} = \frac{\|X_j\hat{\beta}_j - \hat{\beta}_j P_{X_0}X_j\|^2}{\hat{\sigma}^2} = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2}X_j'(I - P_{X_0})X_j.$$

Récrivons maintenant le carré de la statistique T en explicitant $\hat{\sigma}^2_{\hat{\beta}_j}$:

$$T^2 = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2[(X'X)^{-1}]_{jj}},$$

où $[(X'X)^{-1}]_{jj}$ est le j^{e} élément diagonal de la matrice $(X'X)^{-1}$. Afin de calculer ce terme, nous utilisons la formule permettant d'obtenir l'inverse d'une matrice bloc, formule donnée en annexe **??** p. **??**. Pour appliquer facilement cette formule, en changeant l'ordre des variables, la matrice X devient $(X_0|X_j)$ et $X'X$ s'écrit alors

$$X'X = \left(\begin{array}{c|c} X_0'X_0 & X_0'X_j \\ \hline X_j'X_0 & X_j'X_j \end{array}\right).$$

Son inverse, en utilisant la formule d'inverse de matrice bloc, est

$$[(X'X)^{-1}]_{jj} = \left(X_j'X_j - X_j'X_0(X_0'X_0)^{-1}X_0'X_j\right)^{-1} = \left(X_j'(I - P_{X_0})X_j\right)^{-1}.$$

Nous avons donc $T^2 = F$. Au niveau des lois, l'égalité est aussi valable et nous avons que le carré d'un Student à $(n-p)$ ddl est une loi de Fisher à $(1, n-p)$ ddl. Bien entendu, le quantile $(1-\alpha)$ d'une loi de Fisher correspond au quantile $1-\alpha/2$ d'une loi de Student. La loi $\mathcal{T}$ est symétrique autour de 0 et donc, lorsqu'elle est élevée au carré, les valeurs plus faibles que $t_{n-p}(\alpha/2)$, qui ont une probabilité sous H_0 de $\alpha/2$ d'apparaître, et celles plus fortes que $t_{n-p}(1-\alpha/2)$, qui ont une probabilité sous H_0 de $\alpha/2$ d'apparaître, deviennent toutes plus grandes que $t^2_{n-p}(1-\alpha/2)$. La probabilité que ces valeurs dépassent ce seuil sous H_0 est de α et correspond donc bien par définition à $f_{1,n-p}(1-\alpha)$.

Exercice 3.6 (Equivalence du test F et du test de VM)
Nous avons noté la vraisemblance en début du chapitre par

$$\begin{aligned}\mathcal{L}(Y,\beta,\sigma^2) &= \prod_{i=1}^n f_Y(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n\left(y_i - \sum_{j=1}^p \beta_j x_{ij}\right)^2\right] \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left[-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right].\end{aligned}$$

Cette vraisemblance est maximale lorsque $\hat{\beta}$ est l'estimateur des MC et que $\hat{\sigma}^2 = \|Y - X\hat{\beta}\|^2/n$. Nous avons alors

$$\begin{aligned}\max_{\beta,\sigma^2}\mathcal{L}(Y,\beta,\sigma^2) &= \left(\frac{n}{2\pi\|Y - X\hat{\beta}\|^2}\right)^{n/2}\exp\left(-\frac{n}{2}\right) \\ &= \left(\frac{n}{2\pi\,\text{SCR}}\right)^{n/2}\exp\left(-\frac{n}{2}\right) = \mathcal{L}(Y,\hat{\beta},\hat{\sigma}^2),\end{aligned}$$

où $\text{SCR} = \|Y - X\hat{\beta}\|^2$.

Sous l'hypothèse H_0 nous obtenons de façon évidente le résultat suivant :

$$\max_{\beta,\sigma^2} \mathcal{L}_0(Y,\beta_0,\sigma^2) = \left(\frac{n}{2\pi\,\mathrm{SCR}_0}\right)^{n/2} \exp\left(-\frac{n}{2}\right) = \mathcal{L}_0(Y,\hat{\beta}_0,\hat{\sigma}_0^2),$$

où SCR_0 correspond à la somme des carrés résiduels sous H_0, c'est-à-dire $\mathrm{SCR}_0 = \|Y - X_0\hat{\beta}_0\|^2$.
On définit le test du rapport de vraisemblance maximale (VM) par la région critique (Lehmann, 1959) suivante :

$$\mathcal{D}_\alpha = \left\{ Y \in \mathbb{R}^n : \lambda = \frac{\mathcal{L}_0(Y,\hat{\beta}_0,\hat{\sigma}^2)}{\mathcal{L}(Y,\hat{\beta},\hat{\sigma}^2)} < \lambda_0 \right\}.$$

La statistique du rapport de vraisemblance maximale vaut ici

$$\lambda = \left(\frac{\mathrm{SCR}}{\mathrm{SCR}_0}\right)^{n/2} = \left(\frac{\mathrm{SCR}_0}{\mathrm{SCR}}\right)^{-n/2}.$$

Le test du rapport de VM rejette H_0 lorsque la statistique λ est inférieure à une valeur λ_0 définie de façon à avoir le niveau du test égal à α. Le problème qui reste à étudier est de connaître la distribution (au moins sous H_0) de λ.
Définissons, pour λ positif, la fonction bijective g suivante :

$$g(\lambda) = \lambda^{-2/n} - 1.$$

La fonction g est décroissante (sa dérivée est toujours négative), donc $\lambda < \lambda_0$ si et seulement si $g(\lambda) > g(\lambda_0)$. Cette fonction g va nous permettre de nous ramener à des statistiques dont la loi est connue. Nous avons alors

$$\begin{aligned} g(\lambda) &> g(\lambda_0) \\ \frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}} &> g(\lambda_0) \\ \frac{n-p}{p-p_0}\frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}} &> f_0 \end{aligned}$$

où f_0 est déterminée par

$$P_{H_0}\left(\frac{n-p}{p-p_0}\frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}} > f_0\right) = \alpha,$$

avec la loi de cette statistique qui est une loi $\mathcal{F}_{p-p_0,n-p}$ (cf. section précédente). Le test du rapport de VM est donc équivalent au test qui rejette H_0 lorsque la statistique

$$F = \frac{n-p}{p-p_0}\frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}}$$

est supérieure à f_0, où f_0 est la valeur du fractile α de la loi de Fisher à $(p-p_0, n-p)$ degrés de liberté.

Exercice 3.7 (††Test de Fisher pour une hypothèse linéaire quelconque)
Nous pouvons toujours traduire l'hypothèse H_0 : $R\beta = r$ en terme de sous-espace de $\Im_X$. Lorsque $r = 0$, nous avons un sous-espace vectoriel de $\Im_X$ et lorsque $r \neq 0$ nous avons un sous-espace affine de $\Im_X$. Dans les deux cas, nous noterons ce sous-espace $\Im_0$ et

$\Im_0 \subset \Im_X$. Cependant nous ne pourrons plus le visualiser facilement comme nous l'avons fait précédemment avec $\Im_{X_0}$ où nous avions enlevé des colonnes à la matrice X. Nous allons décomposer l'espace $\Im_X$ en deux sous-espaces orthogonaux

$$\Im_X = \Im_0 \overset{\perp}{\oplus} (\Im_0^{\perp} \cap \Im_X).$$

Sous H_0, l'estimation des moindres carrés donne $\hat{Y}_0$ projection orthogonale de Y sur $\Im_0$ et nous appliquons la même démarche pour construire la statistique de test. La démonstration est donc la même que celle du théorème **??**. C'est-à-dire que nous regardons si $\hat{Y}_0$ est proche de $\hat{Y}$ et nous avons donc

$$\begin{aligned} F &= \frac{\|\hat{Y} - \hat{Y}_0\|^2 / \dim(\Im_0^{\perp} \cap \Im_X)}{\|Y - \hat{Y}\|^2 / \dim(\Im_{X^{\perp}})} \\ &= \frac{n-p}{q} \frac{\|Y - \hat{Y}_0\|^2 - \|Y - \hat{Y}\|^2}{\|Y - \hat{Y}\|^2} \\ &= \frac{n-p}{q} \frac{\mathrm{SCR}_0 - \mathrm{SCR}}{\mathrm{SCR}} \sim \mathcal{F}_{q,n-p}. \end{aligned}$$

Le problème du test réside dans le calcul de $\hat{Y}_0$. Dans la partie précédente, il était facile de calculer $\hat{Y}_0$ car nous avions la forme explicite du projecteur sur $\Im_0$.
Une première façon de procéder revient à trouver la forme du projecteur sur $\Im_0$. Une autre façon de faire est de récrire le problème de minimisation sous la contrainte $R\beta = r$. Ces deux manières d'opérer sont présentées en détail dans la correction de l'exercice **??**. Dans tous les cas l'estimateur des MC contraints par $R\beta = r$ est défini par

$$\hat{\beta}_0 = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}).$$

1.4 Validation du modèle

Exercice 4.1 (Questions de cours)
C si $\mathbb{1}$ fait partie des variables ou si $\mathbb{1} \in \Im(X)$, A, C, C, A.

Exercice 4.2 (Propriétés d'une matrice de projection)
La trace d'un projecteur vaut la dimension de l'espace sur lequel s'effectue la projection, donc $\mathrm{tr}(P_X) = p$. Le second point découle de la propriété $P^2 = P$.
Les matrices P_X et $P_X P_X$ sont égales, nous avons que $(P_X)_{ii}$ vaut $(P_X P_X)_{ii}$. Cela s'écrit

$$\begin{aligned} h_{ii} &= \sum_{k=1}^{n} h_{ik} h_{ki} \\ &= h_{ii}^2 + \sum_{k=1, k\neq i}^{n} h_{ik}^2 \\ h_{ii}(1 - h_{ii}) &= \sum_{k=1, k\neq i}^{n} h_{ik}^2. \end{aligned}$$

La dernière quantité de droite de l'égalité est positive et donc le troisième point est démontré. En nous servant de cet écriture les deux derniers points sont aussi démontrés.

Nous pouvons écrire

$$h_{ii}(1-h_{ii}) \quad = \quad h_{ij}^2 + \sum_{k=1,k\neq i,j}^{n} h_{ik}^2.$$

La quantité de gauche est maximum lorsque $h_{ii} = 0.5$ et vaut alors 0.25. Le quatrième point est démontré.

Exercice 4.3 (Lemme d'inversion matricielle)
Commençons par effectuer les calculs en notant que la quantité $u'M^{-1}v$ est un scalaire que nous noterons k. Nous avons

$$\begin{aligned} & \left(M+uv'\right)\left(M^{-1} - \frac{M^{-1}uv'M^{-1}}{1+u'M^{-1}v}\right) \\ = \quad & MM^{-1} - \frac{MM^{-1}uv'M^{-1}}{1+k} + uv'M^{-1} - \frac{uv'M^{-1}uv'M^{-1}}{1+k} \\ = \quad & I + \frac{-uv'M^{-1} + uv'M^{-1} + kuv'M^{-1} - ukv'M^{-1}}{1+k}. \end{aligned}$$

Le résultat est démontré.

Exercice 4.4 (Résidus studentisés)

1-2. Il suffit d'utiliser la définition du produit matriciel et de la somme matricielle et d'identifier les 2 membres des égalités.

3. En utilisant maintenant l'égalité (**??**) sur les inverses, avec $u = -x_i$ et $v = x_i'$, nous avons

$$(X_{(i)}'X_{(i)})^{-1} = (X'X - x_ix_i')^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-x_i'(X'X)^{-1}x_i}.$$

La définition de $h_{ii} = x_i'(X'X)^{-1}x_i$ donne le résultat.

4. Calculons la prévision où $\hat{\beta}_{(i)}$ est l'estimateur de β obtenu sans la i^{e} observation

$$\begin{aligned} \hat{y}_i^p = x_i'\hat{\beta}_{(i)} \quad &= \quad x_i'(X_{(i)}'X_{(i)})^{-1}X_{(i)}'Y_{(i)} \\ &= \quad x_i'\left[(X'X)^{-1} + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_{ii}}\right]\left(X'Y - x_i'y_i\right) \\ &= \quad x_i'\hat{\beta} + \frac{h_{ii}}{1-h_{ii}}x_i'\hat{\beta} - h_{ii}y_i - \frac{h_{ii}^2}{1-h_{ii}}y_i \\ &= \quad \frac{1}{1-h_{ii}}\hat{y}_i - \frac{h_{ii}}{1-h_{ii}}y_i. \end{aligned}$$

5. Ce dernier résultat donne

$$\hat{\varepsilon}_i = (1-h_{ii})(y_i - \hat{y}_i^p).$$

Nous avons alors

$$\begin{aligned} t_i^* \quad &= \quad \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \\ &= \quad \frac{\sqrt{(1-h_{ii})}(y_i - \hat{y}_i^p)}{\hat{\sigma}_{(i)}}. \end{aligned}$$

Pour terminer, remarquons qu'en multipliant l'égalité de la question 3 à gauche par x_i' et à droite par x_i

$$\begin{aligned} x_i'(X_{(i)}'X_{(i)})^{-1}x_i &= h_{ii} + \frac{h_{ii}^2}{1-h_{ii}}. \\ 1 + x_i'(X_{(i)}'X_{(i)})^{-1}x_i &= 1 + \frac{h_{ii}}{1-h_{ii}} = \frac{h_{ii}}{1-h_{ii}}. \end{aligned}$$

6. Utilisons l'expression

$$t_i^* = \frac{y_i - \hat{y}_i^p}{\hat{\sigma}_{(i)}\sqrt{1 + x_i'(X_{(i)}'X_{(i)})^{-1}x_i}}.$$

Nous pouvons alors appliquer la preuve de la proposition **??** p. **??**, en constatant que la i$^{\text{e}}$ observation est une nouvelle observation. Nous avons donc $n-1$ observations pour estimer les paramètres, cela donne donc un Student à $n-1-p$ paramètres.

Exercice 4.5 (Distance de Cook)
Nous reprenons une partie des calculs de l'exercice précédent :

$$\begin{aligned} \hat{\beta}_{(i)} &= (X_{(i)}'X_{(i)})^{-1}X_{(i)}'Y_{(i)} \\ &= (X'X)^{-1}[X'Y - x_iy_i] + \frac{1}{1-h_{ii}}(X'X)^{-1}x_ix_i'(X'X)^{-1}[X'Y - x_iy_i] \\ &= \hat{\beta} - (X'X)^{-1}x_iy_i + \frac{1}{1-h_{ii}}(X'X)^{-1}x_ix_i'\hat{\beta} - \frac{h_{ii}}{1-h_{ii}}(X'X)^{-1}x_iy_i, \end{aligned}$$

d'où le résultat. Pour obtenir la seconde écriture de la distance de Cook, nous écrivons d'abord que

$$\hat{\beta}_{(i)} - \hat{\beta} = \frac{-\hat{\varepsilon}_i}{1-h_{ii}}(X'X)^{-1}x_i.$$

Puis nous développons

$$\begin{aligned} C_i &= \frac{1}{p\hat{\sigma}^2}(\hat{\beta}_{[i]} - \hat{\beta})'X'X(\hat{\beta}_{(i)} - \hat{\beta}) \\ &= \frac{1}{p\hat{\sigma}^2}\left(\frac{-\hat{\varepsilon}_i}{1-h_{ii}}\right)^2 x_i'(X'X)^{-1}(X'X)(X'X)^{-1}x_i. \end{aligned}$$

Le résultat est démontré.

Exercice 4.6 (Régression partielle)
Nous avons le modèle suivant :

$$P_{X_{\bar{j}}^\perp}Y = \beta_j P_{X_{\bar{j}}^\perp}X_j + \eta.$$

L'estimateur des moindres carrés $\tilde{\beta}_j$ issu de ce modèle vaut

$$\tilde{\beta}_j = (X_j'P_{X_{\bar{j}}^\perp}X_j)^{-1}X_j'P_{X_{\bar{j}}^\perp}Y.$$

La projection de Y sur $\Im(X_{\bar{j}})$ (i.e. la prévision par le modèle sans la variable X_j) peut s'écrire comme la projection Y sur $\Im(X)$ qui est ensuite projetée sur $\Im(X_{\bar{j}})$, puisque $\Im(X_{\bar{j}}) \subset \Im(X)$. Ceci s'écrit

$$P_{X_{\bar{j}}}Y \quad = \quad P_{X_{\bar{j}}}P_X Y = P_{X_{\bar{j}}}X\hat{\beta} = P_{X_{\bar{j}}}(X_{\bar{j}}\hat{\beta}_{\bar{j}} + \hat{\beta}_j X_j) = X_{\bar{j}}\hat{\beta}_{\bar{j}} + \hat{\beta}_j P_{X_{\bar{j}}}X_j,$$

et donc

$$X_{\bar{j}}\hat{\beta}_{\bar{j}} = P_{X_{\bar{j}}}Y - \hat{\beta}_j P_{X_{\bar{j}}}X_j.$$

Récrivons les résidus

$$\begin{aligned}
\hat{\varepsilon} &= P_{X^\perp}Y = Y - X\hat{\beta} = Y - X_{\bar{j}}\hat{\beta}_{\bar{j}} - \hat{\beta}_j X_j \\
&= Y - P_{X_{\bar{j}}}Y + \hat{\beta}_j P_{X_{\bar{j}}}X_j - \hat{\beta}_j X_j \\
&= (I - P_{X_{\bar{j}}})Y - \hat{\beta}_j (I - P_{X_{\bar{j}}})X_j \\
&= P_{X_{\bar{j}}^\perp}Y - \hat{\beta}_j P_{X_{\bar{j}}^\perp}X_j.
\end{aligned}$$

En réordonnant cette dernière égalité, nous pouvons écrire

$$P_{X_{\bar{j}}^\perp}Y \quad = \quad \hat{\beta}_j P_{X_{\bar{j}}^\perp}X_j + \hat{\varepsilon}.$$

Nous avons alors

$$\begin{aligned}
\tilde{\beta}_j &= (X_j' P_{X_{\bar{j}}^\perp}X_j)^{-1} X_j' P_{X_{\bar{j}}^\perp}Y \\
&= (X_j' P_{X_{\bar{j}}^\perp}X_j)^{-1} X_j' (\hat{\beta}_j P_{X_{\bar{j}}^\perp}X_j + \hat{\varepsilon}) \\
&= \hat{\beta}_j + (X_j' P_{X_{\bar{j}}^\perp}X_j)^{-1} X_j' \hat{\varepsilon}).
\end{aligned}$$

Le produit scalaire $X_j'\hat{\varepsilon} = \langle X_j, \hat{\varepsilon}\rangle$ est nul car les deux vecteurs appartiennent à des sous-espaces orthogonaux, d'où le résultat.

1.5 Régression sur variables qualitatives

Exercice 5.1 (Questions de cours)
A, A, C, B.

Exercice 5.2 (Analyse de la covariance)
Nous avons pour le modèle complet la matrice suivante :

$$X = \begin{bmatrix}
1 & \cdots & 0 & x_{11} & \cdots & 0 \\
\vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\
1 & \cdots & 0 & x_{1n_1} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 1 & 0 & \cdots & x_{I1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & 1 & 0 & \cdots & x_{In_I}
\end{bmatrix}$$

et pour les deux sous-modèles, nous avons les matrices suivantes :

$$X = \begin{bmatrix} 1 & \cdots & 0 & x_{11} \\ \vdots & \cdots & \vdots & \vdots \\ 1 & \cdots & 0 & x_{1n_1} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & x_{I1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & x_{In_I} \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & \cdots & x_{I1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & x_{In_I} \end{bmatrix}$$

Dans le modèle complet, nous obtenons par le calcul

$$X'X = \begin{pmatrix} n_1 & 0 & \cdots & \sum x_{i1} & 0 & \cdots \\ & \ddots & & & \ddots & \\ 0 & \cdots & n_I & 0 & \cdots & \sum x_{iI} \\ \sum x_{i1} & 0 & \cdots & \sum x_{i1}^2 & 0 & \cdots \\ & \ddots & & & \ddots & \\ 0 & \cdots & \sum x_{iI} & 0 & \cdots & \sum x_{iI}^2 \end{pmatrix} \quad X'Y = \begin{pmatrix} \sum y_{i1} \\ \vdots \\ \sum y_{iI} \\ \sum x_{i1}y_{i1} \\ \vdots \\ \sum x_{iI}y_{iI} \end{pmatrix}$$

Une inversion par bloc de $X'X$ et un calcul matriciel donnent le résultat indiqué.
Une autre façon de voir le problème est de partir du problème de minimisation

$$\begin{aligned} & \min \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \alpha_j - \beta_j x_{ij})^2 \\ = \quad & \min \sum_{j=1}^{n_i} (y_{j1} - \alpha_1 - \beta_1 x_{j1})^2 + \cdots + \sum_{j=1}^{n_I} (y_{jI} - \alpha_I - \beta_I x_{JI})^2 . \end{aligned}$$

Cela revient donc à calculer les estimateurs des MC pour chaque modalité de la variable qualitative.

Exercice 5.3 (Estimateurs des MC en ANOVA à 1 facteur)
La preuve de cette proposition est relativement longue mais peu difficile. Nous avons toujours Y un vecteur de $\mathbb{R}^n$ à expliquer. Nous projetons Y sur le sous-espace engendré par les colonnes de A_c, noté $\Im_{A_c}$, de dimension I, et obtenons un unique $\hat{Y}$. Cependant, en fonction des contraintes utilisées, le repère de $\Im_{A_c}$ va changer.
Le cas le plus facile se retrouve lorsque $\mu = 0$. Nous avons alors

$$(A_c'A_c) = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & n_I \end{bmatrix} \quad (A_c'Y) = \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \\ \vdots \\ \sum_{j=1}^{n_I} y_{Ij} \end{bmatrix}$$

d'où le résultat. La variance de $\hat{\alpha}$ vaut $\sigma^2(A_c'A_c)^{-1}$ et cette matrice est bien diagonale.
Pour les autres contraintes, nous utilisons le vecteur $\vec{e}_{ij}$ de $\mathbb{R}^n$ dont toutes les coordonnées sont nulles sauf celle repérée par le couple (i, j) qui vaut 1 pour repérer un individu. Nous notons $\vec{e}_i$ le vecteur de $\mathbb{R}^n$ dont toutes les coordonnées sont nulles sauf celles repérées par les indices i, j pour $j = 1, \cdots, n_i$ qui valent 1. En fait, ce vecteur repère donc les individus qui admettent la modalité i. La somme des $\vec{e}_i$ vaut le vecteur $\mathbb{1}$. Les vecteurs colonnes de la matrice A_c valent donc $\vec{e}_1, \cdots, \vec{e}_I$.

Considérons le modèle

$$Y = \mu\mathbb{1} + \alpha_1\vec{e_1} + \alpha_2\vec{e_2} + \cdots + \alpha_I\vec{e_I} + \varepsilon.$$

Voyons comment nous pouvons récrire ce modèle lorsque les contraintes sont satisfaites.

1. $\alpha_1 = 0$, le modèle devient alors

$$\begin{aligned} Y &= \mu\mathbb{1} + 0\vec{e_1} + \alpha_2\vec{e_2} + \cdots + \alpha_I\vec{e_I} + \varepsilon \\ &= \mu\mathbb{1} + \alpha_2\vec{e_2} + \cdots + \alpha_I\vec{e_I} + \varepsilon \\ &= [\mathbb{1}, \vec{e_2}, \cdots, \vec{e_I}]\beta + \varepsilon \\ &= X_{[\alpha_1=0]}\beta_{[\alpha_1=0]} + \varepsilon. \end{aligned}$$

2. $\sum n_i\alpha_i = 0$ cela veut dire que $\alpha_I = -\sum_{j=1}^{I-1} n_j\alpha_j/n_I$, le modèle devient

$$\begin{aligned} Y &= \mu\mathbb{1} + \alpha_1\vec{e_1} + \cdots + \alpha_{I-1}\vec{e_{I-1}} - \sum_{j=1}^{I-1}\frac{n_j\alpha_j}{n_I}\vec{e_I} + \varepsilon \\ &= \mu\mathbb{1} + \alpha_1(\vec{e}_1 - \frac{n_1}{n_I}\vec{e}_I) + \cdots + \alpha_{I-1}(\vec{e}_{I-1} - \frac{n_{I-1}}{n_I}\vec{e}_I) + \varepsilon \\ &= \mu\mathbb{1} + \alpha_1\vec{v}_1 + \cdots + \alpha_{I-1}\vec{v}_{I-1} + \varepsilon \quad \text{où} \quad \vec{v}_i = (\vec{e}_i - \frac{n_i}{n_I}\vec{e}_I) \\ &= X_{[\sum n_i\alpha_i=0]}\beta_{[\sum n_i\alpha_i=0]} + \varepsilon. \end{aligned}$$

3. $\sum \alpha_i = 0$ cela veut dire que $\alpha_I = -\sum_{j=1}^{I-1}\alpha_j$, le modèle devient

$$\begin{aligned} Y &= \mu\mathbb{1} + \alpha_1\vec{e_1} + \cdots + \alpha_{I-1}\vec{e_{I-1}} - \sum_{j=1}^{I-1}\alpha_j\vec{e_I} + \varepsilon \\ &= \mu\mathbb{1} + \alpha_1(\vec{e}_1 - \vec{e}_I) + \cdots + \alpha_{I-1}(\vec{e}_{I-1} - \vec{e}_I) + \varepsilon \\ &= \mu\mathbb{1} + \alpha_1\vec{u}_1 + \cdots + \alpha_{I-1}\vec{u}_{I-1} + \varepsilon \quad \text{où} \quad \vec{u}_i = (\vec{e}_i - \vec{e}_I) \\ &= X_{[\sum \alpha_i=0]}\beta_{[\sum \alpha_i=0]} + \varepsilon. \end{aligned}$$

Dans tous les cas, la matrice X est de taille $n \times I$, et de rang I. La matrice $X'X$ est donc inversible. Nous pouvons calculer l'estimateur $\hat{\beta}$ des MC de β par la formule $\hat{\beta} = (X'X)^{-1}X'Y$ et obtenir les valeurs des estimateurs. Cependant ce calcul n'est pas toujours simple et il est plus facile de démontrer les résultats *via* les projections.
Les différentes matrices X et la matrice A engendrent le même sous-espace, donc la projection de Y, notée $\hat{Y}$ dans ce sous-espace, est toujours la même. La proposition (**??**) indique que

$$\hat{Y} = \bar{y_1}\vec{e_1} + \cdots + \bar{y_I}\vec{e_I}.$$

Avec les différentes contraintes, nous avons les 3 cas suivants :

1. $\alpha_1 = 0$, la projection s'écrit

$$\hat{Y} = \hat{\mu}\mathbb{1} + \hat{\alpha}_2\vec{e_2} + \cdots + \hat{\alpha}_I\vec{e_I}.$$

2. $\sum n_i\alpha_i = 0$, la projection s'écrit

$$\hat{Y} = \hat{\mu}\mathbb{1} + \hat{\alpha}_1\vec{e_1} + \cdots + \hat{\alpha}_{I-1}\vec{e_{I-1}} - \sum_{j=1}^{I-1}\frac{n_j\hat{\alpha}_j}{n_I}\vec{e_I}.$$

3. $\sum \alpha_i = 0$, la projection s'écrit

$$\hat{Y} \quad = \quad \hat{\mu}\mathbb{1} + \hat{\alpha}_1\vec{e_1} + \cdots + \hat{\alpha}_{I-1}\vec{e_{I-1}} - \sum_{j=1}^{I-1}\hat{\alpha}_j\vec{e_I}.$$

Il suffit maintenant d'écrire que la projection est identique dans chaque cas et de remarquer que le vecteur $\mathbb{1}$ est la somme des vecteurs $\vec{e_i}$ pour i variant de 1 à I. Cela donne

1. $\alpha_1 = 0$

$$\begin{aligned} & \bar{y_1}\vec{e_1} + \cdots + \bar{y_I}\vec{e_I} \\ = \quad & \hat{\mu}\mathbb{1} + \hat{\alpha}_2\vec{e_2} + \cdots + \hat{\alpha}_I\vec{e_I} \\ = \quad & \hat{\mu}\vec{e_1} + (\hat{\mu}+\hat{\alpha}_2)\vec{e_2}\cdots(\hat{\mu}+\hat{\alpha}_I)\vec{e_I}. \end{aligned}$$

2. $\sum n_i\alpha_i = 0$

$$\begin{aligned} & \bar{y_1}\vec{e_1} + \cdots + \bar{y_I}\vec{e_I} \\ = \quad & \hat{\mu}\mathbb{1} + \hat{\alpha}_1\vec{e_1} + \cdots + \hat{\alpha}_{I-1}\vec{e_{I-1}} - \sum_{j=1}^{I-1}\frac{n_j\hat{\alpha}_j}{n_I}\vec{e_I} \\ = \quad & (\hat{\mu}+\hat{\alpha}_1)\vec{e_1} + \cdots + (\hat{\mu}+\hat{\alpha}_{I-1})\vec{e_{I-1}} + (\hat{\mu} - \sum_{i=1}^{I-1}\frac{n_i}{n_I}\hat{\alpha}_i)\vec{e_I}. \end{aligned}$$

3. $\sum \alpha_i = 0$

$$\begin{aligned} & \bar{y_1}\vec{e_1} + \cdots + \bar{y_I}\vec{e_I} \\ = \quad & \hat{\mu}\mathbb{1} + \hat{\alpha}_1\vec{e_1} + \cdots + \hat{\alpha}_{I-1}\vec{e_{I-1}} - \sum_{j=1}^{I-1}\hat{\alpha}_j\vec{e_I} \\ = \quad & (\hat{\mu}+\hat{\alpha}_1)\vec{e_1} + \cdots + (\hat{\mu}+\hat{\alpha}_{I-1})\vec{e_{I-1}} + (\hat{\mu} - \sum_{i=1}^{I-1}\hat{\alpha}_i)\vec{e_I}. \end{aligned}$$

En identifiant les différents termes, nous obtenons le résultat annoncé.

Exercice 5.4 (Estimateurs des MC en ANOVA à 2 facteurs)
Nous notons $\vec{e}_{ijk}$ le vecteur de $\mathbb{R}^n$ dont toutes les coordonnées sont nulles sauf celle indicée par ijk qui vaut 1. Sous les contraintes de type analyse par cellule, le modèle devient

$$y_{ijk} \quad = \quad \gamma_{ij} + \varepsilon_{ijk},$$

et donc matriciellement

$$Y = X\beta + \varepsilon \qquad X = (\vec{e_{11}}, \vec{e_{12}}, \ldots, \vec{e_{IJ}}),$$

où le vecteur $\vec{e}_{ij} = \sum_k \vec{e}_{ijk}$. Les vecteurs colonnes de la matrice X sont orthogonaux entre eux. Le calcul matriciel $(X'X)^{-1}X'Y$ donne alors le résultat annoncé.

Exercice 5.5 (Estimateurs des MC en ANOVA à 2 facteurs, suite)
Nous notons $\vec{e}_{ijk}$ le vecteur de $\mathbb{R}^n$ dont toutes les coordonnées sont nulles sauf celle indicée par ijk qui vaut 1. Nous définissons ensuite les vecteurs suivants :

$$\vec{e}_{ij} = \sum_k \vec{e}_{ijk} \quad \vec{e}_{i.} = \sum_j \vec{e}_{ij} \quad \vec{e}_{.j} = \sum_i \vec{e}_{ij} \quad \vec{e} = \sum_{i,j,k} \vec{e}_{ijk}.$$

Afin d'effectuer cet exercice, nous définissons les sous-espaces suivants :

$$\begin{aligned}
E_1 &:= \{m\vec{e},\ m \text{ quelconque}\} \\
E_2 &:= \{\sum_i a_i \vec{e}_{i.},\ \sum_i a_i = 0\} \\
E_3 &:= \{\sum_j b_j \vec{e}_{.j},\ \sum_j b_j = 0\} \\
E_4 &:= \{\sum_{ij} c_{ij}\vec{e}_{ij},\ \forall j \sum_i c_{ij} = 0 \text{ et } \forall i \sum_j c_{ij} = 0\}.
\end{aligned}$$

Ces espaces E_1, E_2, E_3 et E_4 sont de dimension respective 1, $I-1$, $J-1$ et $(I-1)(J-1)$. Lorsque le plan est équilibré, tous ces sous-espaces sont orthogonaux. Nous avons la décomposition suivante :

$$E = E_1 \overset{\perp}{\oplus} E_2 \overset{\perp}{\oplus} E_3 \overset{\perp}{\oplus} E_4.$$

La projection sur E peut se décomposer en une partie sur $E_1, \cdots, E_4$ et l'estimateur des MC est obtenu par projection de Y sur E. Notons $P_{E^\perp}$, P_E, P_{E_1}, P_{E_2}, P_{E_3} et P_{E_4} les projections orthogonales sur les sous-espaces $E^\perp$, E, E_1, E_2, E_3 et E_4, nous avons alors

$$P_{E_1} Y \quad = \quad \bar{y}\mathbb{1},$$

puis, en remarquant que projeter sur le sous-espace engendré par les colonnes de $A = [\vec{e}_{1.}, \cdots, \vec{e}_{I.}]$ est identique à la projection sur $E_1 \overset{\perp}{\oplus} E_2$, nous avons alors avec $\mathbb{1} = \sum_i \vec{e}_{i.}$,

$$P_A Y = \sum_i \bar{y}_{i.}\vec{e}_{i.} \quad \text{donc} \quad P_{E_2} Y = \sum_i (\bar{y}_{i.} - \bar{y})\ \vec{e}_{i.}.$$

De la même façon, nous obtenons

$$\begin{aligned}
P_{E_3}(Y) &= \sum_j (\bar{y}_{.j} - \bar{y})\ \vec{e}_{.j}, \\
P_{E_4}(Y) &= \sum_{ij} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})\ \vec{e}_{i.}, \\
P_{E^\perp}(Y) &= \sum_{ijk} (y_{ijk} - \bar{y}_{ij})\ \vec{e}_{ijk},
\end{aligned}$$

où $\vec{e}_{ijk}$ est le vecteur dont toutes les coordonnées sont nulles sauf celle indicée par ijk qui vaut 1. En identifiant terme à terme, nous retrouvons le résultat énoncé.

Exercice 5.6 (Tableau d'ANOVA à 2 facteurs équilibrée)
Lorsque le plan est équilibré, nous avons démontré, que les sous-espaces E_1, E_2, E_3 et E_4 sont orthogonaux (cf. exercice précédent) deux à deux. Nous avons alors

$$Y \quad = \quad P_{E_1}(Y) + P_{E_2}(Y) + P_{E_3}(Y) + P_{E_4}(Y) + P_{E^\perp}(Y).$$

Nous obtenons ensuite par le théorème de Pythagore

$$\begin{array}{ccccccccc} \|Y-\bar{Y}\|^2 & = & \|P_{E_2}(Y)\|^2 & + & \|P_{E_3}(Y)\|^2 & + & \|P_{E_4}(Y)\|^2 & + & \|P_{E^\perp}(Y)\|^2 \\ \mathrm{SCT} & = & \mathrm{SC}_A & + & \mathrm{SC}_B & + & \mathrm{SC}_{AB} & + & \mathrm{SCR}, \end{array}$$

où

$$\begin{aligned} \mathrm{SCT} &= \sum_i \sum_j \sum k(y_{ijk}-\bar{y})^2 \\ \mathrm{SC}_A &= Jr\sum_i (y_{i..}-\bar{y})^2 \\ \mathrm{SC}_B &= Ir\sum_j (y_{.j.}-\bar{y})^2 \\ \mathrm{SC}_{AB} &= r\sum_i \sum_j (y_{ij.}-y_{i..}-y_{.j.}+\bar{y})^2 \\ \mathrm{SCR} &= \sum_i \sum_j \sum_k (y_{ijk}-\bar{y_{ij}})^2. \end{aligned}$$

1.6 Choix de variables

Exercice 6.1 (Questions de cours)
A, C, B en général. Un cas particulier de la dernière question est le suivant : si les variables sélectionnées ξ engendrent un sous-espace orthogonal au sous-espace engendré par les variables non sélectionnées $\bar{\xi}$, alors C est la bonne réponse.

Exercice 6.2 (Analyse du biais)
La preuve des deux premiers points s'effectue comme l'exemple de la section **??**. Nous ne détaillerons que le premier point. Supposons que $|\xi|$ soit plus petit que p, le « vrai » nombre de variables entrant dans le modèle. Nous avons pour estimateur de β

$$\hat{\beta}_\xi = (X_\xi' X_\xi)^{-1} X_\xi' Y = P_{X_\xi} Y.$$

Le vrai modèle étant obtenu avec p variables, $\mathbb{E}(Y) = X_p\beta$. Nous avons alors

$$\begin{aligned} \mathbb{E}(\hat{\beta}_\xi) &= P_{X_\xi} X_p \beta \\ &= P_{X_\xi} X_\xi \beta_\xi + P_{X_\xi} X_{\bar{\xi}} \beta_{\bar{\xi}}. \end{aligned}$$

Cette dernière quantité n'est pas nulle sauf si $\Im(X_\xi) \perp \Im(X_{\bar{\xi}})$. Comme $\hat{\beta}_\xi$ est en général biaisé, il en est de même pour la valeur prévue $\hat{y}_\xi$ dont l'espérance ne vaudra pas $X\beta$.

Exercice 6.3 (Variance des estimateurs)
L'estimateur obtenu avec les $|\xi|$ variables est noté $\hat{\beta}_\xi$ et l'estimateur obtenu dans le modèle complet $\hat{\beta}$. Ces vecteurs ne sont pas de même taille, le premier est de longueur $|\xi|$, le second de longueur p. Nous comparons les $|\xi|$ composantes communes, c'est-à-dire que nous comparons $\hat{\beta}_\xi$ et $[\hat{\beta}]_\xi$. Partitionnons la matrice X en X_ξ et $X_{\bar{\xi}}$. Nous avons alors

$$\mathrm{V}(\hat{\beta}) = \sigma^2 \begin{pmatrix} X_\xi' X_\xi & X_\xi' X_{\bar{\xi}} \\ X_{\bar{\xi}}' X_\xi & X_{\bar{\xi}}' X_{\bar{\xi}} \end{pmatrix}^{-1}.$$

En utilisant la formule d'inverse par bloc, donnée en annexe **??**, nous obtenons

$$\mathrm{V}([\hat{\beta}]_\xi) \quad = \quad \sigma^2 \left[X'_\xi X_\xi - X'_\xi X_{\bar{\xi}} (X'_{\bar{\xi}} X_{\bar{\xi}})^{-1} X'_{\bar{\xi}} X_\xi \right]^{-1},$$

alors que la variance de $\hat{\beta}_\xi$ vaut

$$\mathrm{V}(\hat{\beta}_\xi) \quad = \quad \sigma^2 \left[X'_\xi X_\xi \right]^{-1}.$$

Nous devons comparer $\mathrm{V}([\hat{\beta}]_\xi)$ et $\mathrm{V}(\hat{\beta}_\xi)$. Nous avons

$$X'_\xi X_\xi - X'_\xi X_{\bar{\xi}} (X'_{\bar{\xi}} X_{\bar{\xi}})^{-1} X'_{\bar{\xi}} X_\xi = X'_\xi (I - P_{X_{\bar{\xi}}}) X_\xi = X'_\xi P_{X_{\bar{\xi}}^\perp} X_\xi.$$

La matrice $P_{X_{\bar{\xi}}^\perp}$ est la matrice d'un projecteur, alors elle est semi-définie positive (SDP) (cf. annexe **??**), donc $X'_\xi P_{X_{\bar{\xi}}^\perp} X_\xi$ est également SDP. La matrice $X'_\xi P_{X_{\bar{\xi}}^\perp} X_\xi - X'_\xi P_{X_{\bar{\xi}}^\perp} X_\xi$ est définie positive (DP) puisque c'est $\mathrm{V}([\hat{\beta}]_\xi)/\sigma^2$. Utilisons le changement de notation suivant :

$$A = X'_\xi X_\xi - X'_\xi P_{X_{\bar{\xi}}} X_\xi \quad \text{et} \quad B = X'_\xi P_{X_{\bar{\xi}}} X_\xi.$$

La matrice A est DP et la matrice B SDP. La propriété donnée en annexe **??** indique que $A^{-1} - (A+B)^{-1}$ est SDP, or

$$\mathrm{V}([\hat{\beta}]_\xi) - \mathrm{V}(\hat{\beta}_\xi) = \sigma^2 (A^{-1} - (A+B)^{-1}).$$

Donc la quantité $\mathrm{V}([\hat{\beta}]_\xi) - \mathrm{V}(\hat{\beta}_\xi)$ est SDP. Le résultat est démontré. L'estimation, en terme de variance, de ξ composantes est plus précise que les mêmes ξ composantes extraites d'une estimation obtenue avec p composantes.
La variance des valeurs ajustées dépend de la variance de $\hat{\beta}$, le point 2 de la proposition se démontre de façon similaire.
Remarque : nous venons de comparer deux estimateurs de même taille *via* leur matrice de variance. Pour cela, nous montrons que la différence de ces deux matrices est une matrice SDP. Que pouvons-nous dire alors sur la variance de chacune des coordonnées ? Plus précisément, pour simplifier les notations, notons le premier estimateur (de taille p) $\tilde{\beta}$ de variance $\mathrm{V}(\tilde{\beta})$ et le second estimateur $\hat{\beta}$ de variance $\mathrm{V}(\hat{\beta})$. Si $\mathrm{V}(\tilde{\beta}) - \mathrm{V}(\hat{\beta})$ est SDP, pouvons-nous dire que $\mathrm{V}(\tilde{\beta}_i) - \mathrm{V}(\hat{\beta}_i)$ est un nombre positif pour i variant de 1 à p ? Considérons par exemple le vecteur $u'_1 = (1, 0, \ldots, 0)$ de $\mathbb{R}^p$. Nous avons alors

$$u'_1 \hat{\beta} = \hat{\beta}_1 \quad \text{et} \quad u'_1 \tilde{\beta} = \tilde{\beta}_1.$$

Comme $\mathrm{V}(\tilde{\beta}) - \mathrm{V}(\hat{\beta})$ est SDP, nous avons pour tout vecteur u de $\mathbb{R}^p$ que $u'(\mathrm{V}(\tilde{\beta}) - \mathrm{V}(\hat{\beta}))u \geq 0$, c'est donc vrai en particulier pour u_1. Nous avons donc

$$\begin{aligned} u'_1(\mathrm{V}(\tilde{\beta}) - \mathrm{V}(\hat{\beta}))u_1 &\geq 0 \\ u'_1 \mathrm{V}(\tilde{\beta}) u_1 - u'_1 \mathrm{V}(\hat{\beta}) u_1 &\geq 0 \\ \mathrm{V}(u'_1 \tilde{\beta}) - \mathrm{V}(u'_1 \hat{\beta}) &\geq 0 \\ \mathrm{V}(\tilde{\beta}_1) &\geq \mathrm{V}(\hat{\beta}_1). \end{aligned}$$

Nous pouvons retrouver ce résultat pour les autres coordonnées des vecteurs estimés ou encore pour des combinaisons linéaires quelconques de ces coordonnées.

Exercice 6.4 (Utilisation du R^2)
La correction de cet exercice est identique à la correction de l'exercice **??**, elle est donc omise.

Exercice 6.5 (Choix de variables)
Tous les modèles possibles ont été étudiés, la recherche est donc exhaustive. En prenant comme critère l'AIC ou le BIC, le modèle retenu est le modèle M134. Comme prévu, le R^2 indique le modèle conservant toutes les variables. Cependant le R^2 peut être utilisé pour tester des modèles emboîtés. Dans ce cas, le modèle retenu est également le M134.

1.7 Moindres carrés généralisés

Exercice 7.1 (Questions de cours)
A, A.

Exercice 7.2 (Régression pondérée)
Nous souhaitons minimiser

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 p_i,$$

où p_i est un réel positif.

Nous pouvons écrire ce critère sous la forme suivante :

$$\sum_{i=1}^{n}\left(\sqrt{p_i}y_i-\sum_{j=1}^{p}\beta_j\sqrt{p_i}x_{ij}\right)^2=\sum_{i=1}^{n}\left(y_i^\star-\sum_{j=1}^{p}\beta_j x_{ij}^\star\right)^2,$$

où $y_i^\star=\sqrt{p_i}y_i$ et $x_{ij}^\star=\sqrt{p_i}x_{ij}$. Notons $P^{1/2}$ la matrice des poids qui vaut $P^{1/2}=\mathrm{diag}(\sqrt{p_i})$. Ce dernier critère est un critère des MC avec comme observations $Y^\star$ et $X^\star$ où $Y^\star=P^{1/2}Y$ et $X^\star=P^{1/2}X$. L'estimateur vaut alors

$$\begin{aligned}\hat\beta_{pond} &= (X^{\star\prime}X^\star)^{-1}X^{\star\prime}Y^\star\\ &= (X'PX)^{-1}X'PY.\end{aligned}$$

Lorsque nous avons la constante comme seule variable explicative, $X=\mathbb{1}_n$, et nous avons alors

$$\hat\beta_{pond} = \frac{\sum p_i y_i}{\sum p_i}.$$

Lorsque les poids sont constants, nous retrouvons, non plus une moyenne pondérée, mais la moyenne usuelle.

Exercice 7.3 (Gauss-Markov)
L'estimateur $\hat\beta_{MCG}$ est bien linéaire. Calculons son biais et sa variance

$$\begin{aligned}\mathbb{E}(\hat\beta_{MCG}) &= \mathbb{E}((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y)=(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X\beta=\beta\\ \mathrm{V}(\hat\beta_{MCG}) &= (X'\Omega^{-1}X)^{-1}X'\,\mathrm{V}(Y)X(X'\Omega^{-1}X)^{-1}=(X'\Omega^{-1}X)^{-1}\sigma^2.\end{aligned}$$

Montrons maintenant que cet estimateur est de variance minimale parmi les estimateurs linéaires sans biais. Considérons un autre estimateur linéaire $\tilde{\beta} = CY$ sans biais de β. Posons

$$\tilde{\beta} = CPP^{-1}Y = CPY^*.$$

$\tilde{\beta}$ est linéaire en Y^* et sans biais dans le modèle (*). Or $\hat{\beta}_{MCG}$ est l'estimateur des MC dans le modèle $(*)$, il est donc de variance minimale. La variance de $\hat{\beta}_{MCG}$ est donc plus faible que la variance de $\tilde{\beta}$.

1.8 Ridge et Lasso

Exercice 8.1 (Questions de cours)
A, B, B, B.

Exercice 8.2 (Corrélation multiple et hypothèse $\mathcal{H}_1$)

1. Montrons que la moyenne empirique de $X\hat{\beta}$ vaut $\bar{Y}$. Le vecteur moyenne est obtenu en projetant sur $\mathbb{1}_n$. En effet, comme

$$P_{\mathbb{1}} = \mathbb{1}_n(\mathbb{1}_n'\mathbb{1}_n)^{-1}\mathbb{1}_n' = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n',$$

nous avons, pour une variable $Z = (Z_1, \ldots, Z_n)'$,

$$P_{\mathbb{1}}Z = = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n'Z = \frac{1}{n}\mathbb{1}_n\sum_{i=1}^{n} Z_i = \bar{Z}\mathbb{1}_n.$$

Comme $\mathbb{1}_n \in \Im(X)$, nous avons

$$\bar{Y} = P_{\mathbb{1}}Y = P_{\mathbb{1}}P_XY = P_{\mathbb{1}}X\hat{\beta},$$

c'est-à-dire que la moyenne empirique de $X\hat{\beta}$ vaut $\bar{Y}$.
Le coefficient de corrélation entre $\hat{Y}$ et Y élevé au carré s'écrit donc

$$\begin{aligned}\rho^2(\hat{Y}, Y) &= \frac{\langle \hat{Y} - \bar{Y}, Y - \bar{Y}\rangle^2}{\|\hat{Y} - \bar{Y}\|^2\|Y - \bar{Y}\|^2} \\ &= \frac{\langle \hat{Y} - \bar{Y}, Y - \hat{Y} + \hat{Y} - \bar{Y}\rangle^2}{\|\hat{Y} - \bar{Y}\|^2\|Y - \bar{Y}\|^2} \\ &= \left\{\frac{\langle \hat{Y} - \bar{Y}, Y - \hat{Y}\rangle}{\|\hat{Y} - \bar{Y}\|\|Y - \bar{Y}\|} + \frac{\langle \hat{Y} - \bar{Y}, \hat{Y} - \bar{Y}\rangle}{\|\hat{Y} - \bar{Y}\|\|Y - \bar{Y}\|}\right\}^2.\end{aligned}$$

Comme $(Y - \hat{Y}) \in \Im(X)^\perp$ et que $(\hat{Y} - \bar{Y}) \in \Im(X)$, nous avons $\langle \hat{Y} - \bar{Y}, Y - \hat{Y}\rangle = 0$ et donc

$$\rho^2(\hat{Y}, Y) = \frac{\|\hat{Y} - \bar{Y}\|^2\|\hat{Y} - \bar{Y}\|^2}{\|\hat{Y} - \bar{Y}\|^2\|Y - \bar{Y}\|^2} = \mathrm{R}^2.$$

2. (a) En effectuant le calcul nous trouvons que $Y - 2X_1 + 2X_2 = 3\eta$.

(b) En calculant les normes carrées, nous avons

$$\begin{aligned} \|X_1\|^2 &= 1^2 + 1^2 + 1^2 = 3, \\ \|X_2\|^2 &= 1/2 + 1/2 + 2 = 3, \\ \|X_3\|^2 &= 3/2 + 3/2 = 3. \end{aligned}$$

En calculant les produits scalaires, nous avons

$$\begin{aligned} \langle X_1, X_2 \rangle &= 1 \times 1/\sqrt{2} + 1 \times 1/\sqrt{2} + 1 \times (-\sqrt{2}) = \sqrt{2} - \sqrt{2} = 0, \\ \langle X_1, \eta \rangle &= \sqrt{3}/\sqrt{2} - \sqrt{3}/\sqrt{2} = 0, \\ \langle X_2, \eta \rangle &= 1/\sqrt{2} \times \sqrt{3}/\sqrt{2} - 1/\sqrt{2} \times \sqrt{3}/\sqrt{2} = 0. \end{aligned}$$

(c) La représentation graphique est :

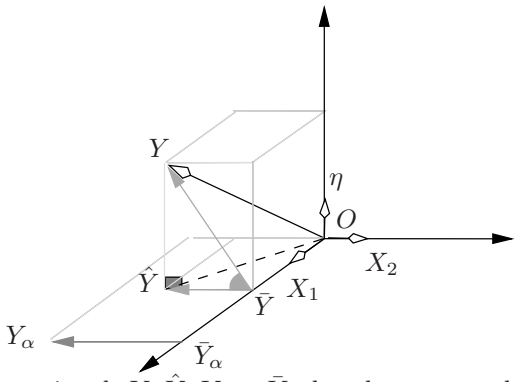

Fig. 8.2 – Représentation de Y, $\hat{Y}$, Y_α et $\bar{Y}_\alpha$ dans le repère orthogonal (X_1, X_2, η).

(d) Nous avons ici $X_1 \in \Im(X)$, $X_2 \in \Im(X)$ et $\eta \in \Im(X)^\perp$, ce qui permet de trouver $\hat{Y}$:

$$\begin{aligned} P_X Y &= P_X(2X_1 - 2X_2 + 3\eta) = 2P_X X_1 - 2P_X X_2 + 3P_X \eta = \\ &= 2X_1 - 2X_2 = (2 - \sqrt{2}, 2 - \sqrt{2}, 2 - 2\sqrt{2})'. \end{aligned}$$

(e) Puisque $\mathbb{1}$ fait partie des variables explicatives, nous avons

$$\rho(Y, \hat{Y}) = \frac{\langle Y - \bar{Y}, \hat{Y} - \bar{Y} \rangle}{\|\hat{Y} - \bar{Y}\| \|Y - \bar{Y}\|},$$

ce qui est la définition du cosinus de l'angle entre $\overrightarrow{\bar{Y}\hat{Y}}$ et $\overrightarrow{\bar{Y}Y}$.

(f) Notons par Y_α le vecteur $X\alpha$. Sa moyenne vaut $\bar{Y}_\alpha$. Nous avons maintenant le cosinus de l'angle entre $\overrightarrow{\bar{Y}Y}$ et $\overrightarrow{\bar{Y}_\alpha Y_\alpha}$. Graphiquement, la moyenne de Y_α est la projection sur $X_1 = \mathbb{1}_3$.

(g) La représentation graphique nous permet de voir que l'angle entre $\overrightarrow{\bar{Y}Y}$ et $\overrightarrow{\bar{Y}_\alpha Y_\alpha}$ est le même que celui entre $\overrightarrow{\bar{Y}Y}$ et $\overrightarrow{\bar{Y}\hat{Y}}$. L'angle est minimum (et le

cosinus maximum) quand $\alpha = \hat{\beta}$ ou pour tout α tel que $\overrightarrow{\bar{Y}_\alpha Y_\alpha} = k\overrightarrow{\bar{Y}\hat{Y}}$ avec $k > 0$.

Du fait de l'orthogonalité entre X_1 et X_2, $\overrightarrow{\bar{Y}_\alpha Y_\alpha}$ est toujours colinéaire à $\overrightarrow{\bar{Y}\hat{Y}}$, seul le signe change en fonction de l'orientation des vecteurs (même sens ou sens opposé).

3. Comme $\rho(X_j; X_k) = 1$ alors $R(X_j; (\mathbb{1}, X_j)) = 1$ et donc puisque la constante fait partie du modèle $R(X_j; X_{(j)}) = 1$. L'hypothèse $\mathcal{H}_1$ n'est donc pas vérifiée.

Exercice 8.3 (Nombre effectif de paramètres de la régression ridge)

1. Rappelons que pour une valeur κ donnée, le vecteur de coefficients de la régression ridge s'écrit

$$\hat{\beta}_{\text{ridge}}(\kappa) = (X'X + \kappa I)^{-1} X'Y.$$

et donc l'ajustement par la régression ridge est

$$\hat{Y}_{\text{ridge}}(\kappa) = X(X'X + \kappa I)^{-1} X'Y = H^*(\kappa)Y$$

2. Soit U_i le vecteur propre de A associé à la valeur propre d_i^2. Nous avons donc par définition que

$$\begin{aligned} AU_i &= d_i^2 U_i \\ AU_i + \lambda U_i &= d_i^2 U_i + \lambda U_i = (d_i^2 + \lambda) U_i \\ (A + \lambda I_p) U_i &= (d_i^2 + \lambda) U_i, \end{aligned}$$

c'est-à-dire que U_i est aussi vecteur propre de $A + \lambda I_p$ associé à la valeur propre $\lambda + d_i^2$.

3. Nous savons que $X = QDP'$ avec Q et P matrices orthogonales et $D = \text{diag}(d_1, \ldots, d_p)$. Puisque Q est orthogonale, nous avons, par définition, $Q'Q = I$. Nous avons donc que $X'X = (QDP')'QDP' = PDQ'QDP' = PD^2P'$. Puisque P est orthogonale $P'P = I_p$ et $P^{-1} = P$.

$$\begin{aligned} \text{tr}(X(X'X + \lambda I_p)^{-1} X') &= \text{tr}((X'X + \lambda I_p)^{-1} X'X) \\ &= \text{tr}((PD^2P' + \lambda PP')^{-1} PD^2P') \\ &= \text{tr}((P(D + \lambda I_p)P')^{-1} PD^2P'). \end{aligned}$$

Nous avons donc

$$\begin{aligned} \text{tr}(X(X'X + \lambda I_p)^{-1} X') &= \text{tr}((P')^{-1}(D + \lambda I_p)^{-1} P^{-1} PD^2P') \\ &= \text{tr}((P')^{-1}(D + \lambda I_p)^{-1} D^2P') \\ &= \text{tr}((D + \lambda I_p)^{-1} D^2). \end{aligned}$$

Selon la définition de $H^*(\kappa)$, nous savons que sa trace vaut donc

$$\text{tr}((D + \kappa I_p)^{-1} D^2).$$

Comme D et I_p sont des matrices diagonales, leur somme et produit sont simplement leur somme et produit terme à terme des éléments de la diagonale, et donc cette trace (somme des éléments de la diagonale) vaut

$$\sum_{i=1}^{p} \frac{d_j^2}{d_j^2 + \kappa}.$$

Exercice 8.4 (EQM de la régression ridge)

1. Les démonstrations figurent en p. **??** :

$$\begin{aligned} B(\hat{\beta}_{ridge}) &= -\kappa(X'X+\kappa I)^{-1}\beta, \\ V(\hat{\beta}_{\text{ridge}}) &= \sigma^2(X'X+\kappa I)^{-1}X'X(X'X+\kappa I)^{-1} \\ \text{EQM}(\hat{\beta}_{\text{ridge}}) &= (X'X+\kappa I)^{-1}\left[\kappa^2\beta\beta' + \sigma^2(X'X)\right](X'X+\kappa I)^{-1}. \end{aligned}$$

2. Puisque $X'X = P\,\text{diag}(\lambda_i)P'$, nous avons

$$(X'X+\kappa I) = P\,\text{diag}(\lambda_i)P' + \kappa PP' = P\,\text{diag}(\lambda_i+\kappa)P'.$$

En se rappelant que $P^{-1} = P'$, son inverse vaut

$$(X'X+\kappa I)^{-1} = P\,\text{diag}(1/(\lambda_i+\kappa))P'.$$

Nous avons donc

$$\begin{aligned} \text{EQM}(\hat{\beta}_{\text{ridge}}) &= P\,\text{diag}(\frac{1}{\lambda_i+\kappa})P'\left[\kappa^2\beta\beta' + \sigma^2(X'X)\right]P\,\text{diag}(\frac{1}{\lambda_i+\kappa})P' \\ &= P\,\text{diag}(\frac{1}{\lambda_i+\kappa})\left[\kappa^2(P'\beta\beta'P) + \sigma^2 I_p\right]\text{diag}(\frac{1}{\lambda_i+\kappa})P'. \end{aligned}$$

Nous en déduisons que sa trace vaut

$$\begin{aligned} \text{tr}\left\{EQM(\hat{\beta}_{\text{ridge}})\right\} = \text{tr}\Big\{&\text{diag}(\frac{1}{\lambda_i+\kappa})\left[\kappa^2(P'\beta\beta'P) + \sigma^2 I_p\right] \\ &\text{diag}(\frac{1}{\lambda_i+\kappa})P'P\Big\}, \end{aligned}$$

et, comme $P'P = I_p$, nous avons alors

$$\text{tr}\left\{EQM(\hat{\beta}_{\text{ridge}})\right\} = \text{tr}\left\{\left[\kappa^2(P'\beta\beta'P) + \sigma^2 I_p\right]\text{diag}(\frac{1}{(\lambda_i+\kappa)^2})\right\}.$$

Le i^{e} élément de la diagonale de la matrice $P'\beta\beta'P$ vaut $[P'\beta]_i^2$. Celui de $\left[\kappa^2(P'\beta\beta'P) + \sigma^2 I_p\right]$ vaut $\kappa^2[P'\beta]_i^2 + \sigma^2$ et celui de

$$\left[\kappa^2(P'\beta\beta'P) + \sigma^2 I_p\right]\text{diag}(\frac{1}{(\lambda_i+\kappa)^2})$$

vaut donc

$$\kappa^2[P'\beta]_i^2 + \sigma^2/(\lambda_i+\kappa)^2.$$

On en déduit le résultat annoncé car la trace est la somme des éléments diagonaux d'une matrice.

3. L'estimateur des MC est non biaisé et son EQM vaut sa variance :

$$\text{EQM}(\hat{\beta}_{\text{MC}}) = \sigma^2(X'X)^{-1}.$$

Nous avons alors

$$\begin{aligned} \text{EQM}(\hat{\beta}_{\text{MC}}) &= \sigma^2(X'X+\kappa I)^{-1}(X'X+\kappa I)(X'X)^{-1} \\ &= \sigma^2(X'X+\kappa I)^{-1}(X'X(X'X)^{-1} + \kappa I(X'X)^{-1}) \\ &= \sigma^2(X'X+\kappa I)^{-1}(I+\kappa(X'X)^{-1})(X'X+\kappa I)(X'X+\kappa I)^{-1} \\ &= \sigma^2(X'X+\kappa I)^{-1}(X'X+2\kappa I+\kappa^2(X'X)^{-1})(X'X+\kappa I)^{-1}. \end{aligned}$$

4. Le calcul de $\Delta = \text{EQM}(\hat{\beta}_{\text{ridge}}) - \text{EQM}(\hat{\beta}_{\text{MC}})$ est immédiat en utilisant l'expression précédente de $\text{EQM}(\hat{\beta}_{\text{MC}})$ et celle rappelée en question **??**.
5. En utilisant le théorème proposé avec $A = (X'X + \kappa I)^{-1}$ et $B = (\sigma^2(2I_p + \kappa^2(X'X)^{-1}) - \kappa\beta\beta')$ nous obtenons le résultat demandé. Cette condition dépend de β qui est inconnu, mais aussi de X, c'est-à-dire des mesures obtenues.
6. Intéressons-nous à la matrice $\gamma\gamma'$. Cette matrice est symétrique donc diagonalisable, de valeurs propres positives ou nulles. La somme de ses valeurs propres est égale à la trace de cette matrice

$$\text{tr}(\gamma\gamma') = \text{tr}(\gamma'\gamma) = \gamma'\gamma.$$

Montrons que cette matrice n'a qu'une seule valeur propre non nulle $\gamma'\gamma$. Pour cela, considérons le vecteur $\gamma \in \mathbb{R}^p$ et montrons qu'il est vecteur propre de $\gamma\gamma'$ associé à la valeur propre $\gamma'\gamma$:

$$(\gamma\gamma')\gamma \quad = \quad \gamma(\gamma'\gamma) = (\gamma'\gamma)\gamma.$$

Nous avons donc un vecteur propre de $\gamma\gamma'$ qui est γ associé à la valeur propre $\gamma'\gamma$. De plus, nous savons que la somme des valeurs propres positives ou nulles de $\gamma\gamma'$ vaut $\gamma'\gamma$. Nous en déduisons que les $p-1$ valeurs propres restantes sont toutes nulles.

Nous pouvons donc dire que la matrice $\gamma\gamma'$ se décompose comme

$$\gamma\gamma' \quad = \quad UDU',$$

où U est la matrice orthogonale des vecteurs propres normés à l'unité de $\gamma\gamma'$ et $D = \text{diag}(\gamma'\gamma, 0, \ldots, 0)$. Nous avons donc

$$I_p - \gamma\gamma' \quad = \quad UU' - UDU' = U(\text{diag}(1 - \gamma'\gamma, 1, \ldots, 1)U'.$$

Les valeurs propres de $I_p - \gamma\gamma'$ sont donc $1 - \gamma'\gamma, 1, \ldots, 1$, qui sont toutes positives ou nulles dès que $\gamma'\gamma \leq 1$.

7. Une condition pour que $\sigma^2(2I_p - \kappa\beta\beta')$ soit semi-définie positive est que $(\kappa\beta\beta') \leq \sigma^2$ (cf. question précédente) et donc $(\sigma^2(2I_p + \kappa^2(X'X)^{-1}) - \kappa\beta\beta')$ est alors la somme de 2 matrices semi-définies positives donc semi-définie positive. Cela implique qu'il s'agit d'une condition suffisante pour que Δ soit semi-définie positive.
8. Nous venons de montrer 2 conditions, l'une nécessaire et suffisante, l'autre suffisante, afin que Δ soit semi-définie positive. Cette assertion signifie que, quelle que soit la combinaison linéaire du vecteur de paramètre (par exemple une coordonnée), l'estimateur ridge est meilleur que celui des MC au sens de l'EQM. Cela signifie aussi que, si une de ces conditions est vérifiée, globalement au sens de la trace de l'EQM, l'estimateur ridge est meilleur que celui des MC.

   Au niveau des conditions, cela permet de trouver la valeur optimale de κ. Malheureusement chacune des 2 conditions dépend de la valeur β inconnue et donc n'est pas réellement utilisable en pratique. La condition suffisante procure une amélioration, dans le sens où elle ne dépend pas de X donc de l'expérience. Le prix à payer est bien sûr qu'il s'agit seulement d'une condition suffisante et donc plus restrictive.

Exercice 8.5 (Estimateurs à rétrécissement -*Shrinkage*)

1. Soit le modèle de régression

$$Y = X\beta + \varepsilon.$$

En le pré-multipliant par P, nous avons

$$Z = PY = PX\beta + P\varepsilon = DQ\beta + \eta = D\gamma + \eta.$$

Puisque $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ et P fixé, nous avons que $\eta = P\varepsilon$ suit une loi normale de moyenne $\mathbb{E}(\eta) = P\mathbb{E}(\varepsilon) = 0$ et de variance $\mathrm{V}(\eta) = P\,\mathrm{V}(\varepsilon)P' = \sigma^2 PP' = \sigma^2 I_n$.
Par définition, Z vaut PY et nous savons que $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, donc $Z \sim \mathcal{N}(PX\beta, \sigma^2 PP')$, c'est-à-dire $Z \sim \mathcal{N}(DQ\beta, \sigma^2 I_n)$ ou encore $Z \sim \mathcal{N}(D\gamma, \sigma^2 I_n)$. En utilisant la valeur de D nous avons

$$D\gamma = \begin{pmatrix} \Delta\gamma \\ 0 \end{pmatrix}.$$

Donc $Z_{1:p} \sim \mathcal{N}(\Delta\gamma, \sigma^2 I_p)$.

2. Soit un estimateur de β linéaire en Y : $\hat{\beta} = AY$. Soit l'estimateur de $\gamma = Q\beta$ linéaire en Y : $\hat{\gamma} = QAY$. Pour calculer leur matrice de l'EQM, nous devons calculer leur biais et leur variance. Le biais de $\hat{\beta}$ est

$$B(\hat{\beta}) = \mathbb{E}(\hat{\beta}) - \beta = \mathbb{E}(AY) - \beta = A\mathbb{E}(Y) - \beta = AX\beta - \beta.$$

Le biais de $\hat{\gamma}$ s'écrit

$$B(\hat{\gamma}) = \mathbb{E}(\hat{\gamma}) - \gamma = \mathbb{E}(Q\hat{\beta}) - \gamma = Q\mathbb{E}(\hat{\beta}) - \gamma = QAX\beta - \gamma.$$

Comme $\gamma = Q\beta$ et $Q'Q = I_p$ nous avons

$$B(\hat{\gamma}) = QAXQ'\gamma - \gamma.$$

La variance de $\hat{\beta}$ s'écrit

$$\mathrm{V}(\hat{\beta}) = \mathrm{V}(AY) = A\,\mathrm{V}(Y)A' = \sigma^2 AA',$$

et celle de $\hat{\gamma}$ est

$$\mathrm{V}(\hat{\gamma}) = \mathrm{V}(Q\hat{\beta}) = Q\,\mathrm{V}(\hat{\beta})Q' = \sigma^2 QAA'Q'.$$

Nous en déduisons que les matrices des EQM sont respectivement

$$\begin{aligned} \mathrm{EQM}(\hat{\beta}) &= (AX\beta - \beta)(AX\beta - \beta)' + \sigma^2 AA', \\ \mathrm{EQM}(\hat{\gamma}) &= (QAXQ'\gamma - \gamma)(QAXQ'\gamma - \gamma)' + \sigma^2 QAA'Q', \end{aligned}$$

et enfin les traces de ces matrices s'écrivent

$$\begin{aligned} \mathrm{tr}(\mathrm{EQM}(\hat{\beta})) &= (AX\beta - \beta)'(AX\beta - \beta) + \sigma^2\,\mathrm{tr}(AA'), \\ \mathrm{tr}(\mathrm{EQM}(\hat{\gamma})) &= (QAXQ'\gamma - \gamma)'(QAXQ'\gamma - \gamma) + \sigma^2\,\mathrm{tr}(AA'). \end{aligned}$$

Rappelons que $\gamma = Q\beta$ et que $Q'Q = I_p$, nous avons donc

$$\begin{aligned}\operatorname{tr}(\mathrm{EQM}(\hat{\gamma})) &= \gamma'(QAXQ' - I_p)'(QAXQ' - I_p)\gamma + \sigma^2 \operatorname{tr}(AA') \\ &= \beta'(QAX - Q)'(QAX - Q)\beta + \sigma^2 \operatorname{tr}(AA') \\ &= \beta'(AX - I_p)Q'Q(AX - I_p)\beta + \sigma^2 \operatorname{tr}(AA') \\ &= \beta'(AX - I_p)(AX - I_p)\beta + \sigma^2 \operatorname{tr}(AA') = \operatorname{tr}(\mathrm{EQM}(\hat{\beta})).\end{aligned}$$

En conclusion, que l'on s'intéresse à un estimateur linéaire de β ou à un estimateur linéaire de γ, dès que l'on passe de l'un à l'autre en multipliant par Q ou Q', matrice orthogonale, la trace de l'EQM est identique, c'est-à-dire que les performances globales des 2 estimateurs sont identiques.

3. Nous avons le modèle de régression suivant :

$$Z_{1:p} = \Delta\gamma + \eta_{1:p},$$

et donc, par définition de l'estimateur des MC, nous avons

$$\hat{\gamma}_{\mathrm{MC}} = (\Delta'\Delta)^{-1}\Delta' Z_{1:p}.$$

Comme Δ est une matrice diagonale, nous avons

$$\hat{\gamma}_{\mathrm{MC}} = \Delta^{-2}\Delta' Z_{1:p} = \Delta^{-1} Z_{1:p}.$$

Cet estimateur est d'expression très simple et il est toujours défini de manière unique, ce qui n'est pas forcément le cas de $\hat{\beta}_{\mathrm{MC}}$.
Comme $Z_{1:p} \sim \mathcal{N}(\Delta\gamma, \sigma^2 I_p)$ nous avons que $\hat{\gamma}_{\mathrm{MC}} = \Delta^{-1} Z_{1:p}$ suit une loi normale d'espérance $\mathbb{E}(\Delta^{-1} Z_{1:p}) = \Delta^{-1}\mathbb{E}(Z_{1:p}) = \gamma$ et de variance $\mathrm{V}(\hat{\gamma}_{\mathrm{MC}}) = \sigma^2\Delta^{-2}$. Puisque $\hat{\gamma}_{\mathrm{MC}}$ est un estimateur des MC, il est sans biais, ce qui est habituel.

4. L'EQM de $\hat{\gamma}_{\mathrm{MC}}$, estimateur sans biais, est simplement sa variance. Pour la i^{e} coordonnée de $\hat{\gamma}_{\mathrm{MC}}$, l'EQM est égal à l'élément i, i de la matrice de variance $\mathrm{V}(\hat{\gamma}_{\mathrm{MC}})$, c'est-à-dire σ^2/δ_i^2. La trace de l'EQM est alors simplement la somme, sur toutes les coordonnées i, de cet EQM obtenu.

5. Par définition $\hat{\gamma}(c) = \operatorname{diag}(c_i) Z_{1:p}$ et nous savons que $Z_{1:p} \sim \mathcal{N}(\Delta\gamma, \sigma^2 I_p)$. Nous obtenons que $\hat{\gamma}(c)$ suit une loi normale d'espérance $\mathbb{E}(\operatorname{diag}(c_i) Z_{1:p}) = \operatorname{diag}(c_i)\Delta\gamma$ et de variance

$$\mathrm{V}(\hat{\gamma}(c)) = \operatorname{diag}(c_i)\,\mathrm{V}(Z_{1:p})\,\operatorname{diag}(c_i)' = \sigma^2 \operatorname{diag}(c_i^2).$$

La loi de $\hat{\gamma}(c)$ étant une loi normale de matrice de variance diagonale, nous en déduisons que les coordonnées de $\hat{\gamma}(c)$ sont indépendantes entre elles.

6. Calculons l'EQM de la i^{e} coordonnée de $\hat{\gamma}(c)$

$$\mathrm{EQM}(\hat{\gamma}(c)_i) = \mathbb{E}(\hat{\gamma}(c)_i - \gamma)^2 = \mathbb{E}(\hat{\gamma}(c)_i^2) + \mathbb{E}(\gamma_i^2) - 2\mathbb{E}(\hat{\gamma}(c)_i\gamma_i).$$

Comme γ_i et que $\mathbb{E}(\hat{\gamma}(c)_i^2) = \mathrm{V}(\hat{\gamma}(c)_i^2) + \{\mathbb{E}(\hat{\gamma}(c)_i^2)\}^2$, nous avons

$$\begin{aligned}\mathrm{EQM}(\hat{\gamma}(c)_i) &= \sigma^2 c_i^2 + (c_i\delta_i\gamma_i)^2 + \gamma_i^2 - 2\gamma_i\mathbb{E}(\hat{\gamma}(c)_i) \\ &= \sigma^2 c_i^2 + (c_i\delta_i\gamma_i)^2 + \gamma_i^2 - 2\sigma^2 c_i\delta_i\gamma_i = \sigma^2 c_i^2 + \gamma_i^2(c_i\delta_i - 1)^2.\end{aligned}$$

7. De manière évidente si γ_i^2 diminue, alors l'EQM de $\hat{\gamma}(c)_i$ diminue aussi. Calculons la valeur de l'EQM quand $\gamma_i^2 = \frac{\sigma^2}{\delta_i^2}\frac{(1/\delta_i)+c_i}{(1/\delta_i)-c_i}$. Nous avons, grâce à la question précédente,

$$\begin{aligned}
\mathrm{EQM}(\hat{\gamma}(c)_i) &= \sigma^2 c_i^2 + (c_i\delta_i - 1)^2 \frac{\sigma^2}{\delta_i^2}\frac{(1/\delta_i)+c_i}{(1/\delta_i)-c_i} \\
&= \sigma^2 c_i^2 + \frac{\sigma^2}{\delta_i^2}(1-c_i\delta_i)^2\frac{1+\delta_i c_i}{1-\delta_i c_i} \\
&= \sigma^2 c_i^2 + \frac{\sigma^2}{\delta_i^2}(1-c_i\delta_i)(1+\delta_i c_i) \\
&= \sigma^2 c_i^2 + \frac{\sigma^2}{\delta_i^2}(1-\delta_i^2 c_i^2) \\
&= \sigma^2 c_i^2 + \frac{\sigma^2}{\delta_i^2} - \sigma^2 c_i^2 = \frac{\sigma^2}{\delta_i^2} \\
&= \mathrm{EQM}(\hat{\gamma}_{\mathrm{MC}}),
\end{aligned}$$

d'où la conclusion demandée.

8. Par définition de $\hat{\gamma}(c)$, nous avons

$$\begin{aligned}
\hat{\gamma}(c) &= \mathrm{diag}(c_i)Z_{1:p} = \mathrm{diag}(\frac{\delta_i}{\delta_i^2+\kappa})Z_{1:p} \\
&= (\Delta'\Delta + \kappa I_p)^{-1}\Delta' Z_{1:p},
\end{aligned}$$

puisque Δ est diagonale. De plus nous avons $D = \binom{\Delta}{0}$, ce qui entraîne que $D'D = \Delta'\Delta$ et $D'Z = \Delta' Z_{1:p}$. Nous obtenons donc

$$\hat{\gamma}(c) = (D'D + \kappa I_p)^{-1} D'Z.$$

Rappelons que $D = PXQ'$ avec P et Q matrices orthogonales, nous avons alors

$$\begin{aligned}
\hat{\gamma}(c) &= (QX'P'PXQ' + \kappa I_p)^{-1}D'Z = (QX'XQ' + \kappa QQ')^{-1}D'Z \\
&= (Q(X'X + \kappa I_p)Q')^{-1}D'Z = (Q')^{-1}(X'X+\kappa I_p)^{-1}(Q)^{-1}D'Z \\
&= Q(X'X+\kappa I_p)^{-1}Q'D'Z.
\end{aligned}$$

Comme $Z = PY$ et $D = PXQ'$, nous avons

$$\hat{\gamma}(c) = Q(X'X+\kappa I_p)^{-1}Q'QX'P'PY = Q(X'X+\kappa I_p)^{-1}XY.$$

Enfin, nous savons que $Q\hat{\gamma} = \hat{\beta}$, nous en déduisons que $\hat{\gamma} = Q'\hat{\beta}$ et donc que dans le cas particulier où $c_i = \frac{\delta_i}{\delta_i^2+\kappa}$ nous obtenons

$$\hat{\beta} = Q\hat{\gamma}(c) = (X'X+\kappa I_p)^{-1}XY,$$

c'est-à-dire l'estimateur de la régression ridge.

Exercice 8.6 (Généralisation de la régression ridge)
Soit la fonction objectif à minimiser

$$\mathcal{L}(\beta) = \|Y - X\beta\|^2 - \sum_{j=1}^{p}\tau_j(\beta_j^2).$$

Dérivons cette fonction par rapport à β_j et nous obtenons

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = 2[X'(Y - X\beta)]_j - 2\tau_j \beta_j.$$

Cette équation se récrit comme

$$\frac{\partial \mathcal{L}}{\partial \beta} = 2X'(Y - X\beta) - 2\Delta\beta.$$

A l'optimum cette dérivée est nulle et nous avons

$$2X'(Y - X\hat{\beta}_{\text{RG}} = 2\Delta\hat{\beta}_{\text{RG}},$$

c'est-à-dire

$$\hat{\beta}_{\text{RG}} = (X'X + \Delta)^{-1}X'Y.$$

Comme le nombre effectif de paramètres est égal à la trace de la matrice H permettant d'obtenir $\hat{Y}$ à partir de Y, nous avons

$$\hat{Y}_{\text{RG}} = X\hat{\beta}_{\text{RG}} = X(X'X + \Delta)^{-1}X'Y = HY$$

Donc le nombre effectif de paramètres est ici

$$\text{tr}(H) = \text{tr}(X(X'X + \Delta)^{-1}X').$$

Exercice 8.7 (IC pour la régression ridge)

1. Nous savons que $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$.
2. La définition de l'estimateur ridge est $\hat{\beta}_{\text{ridge}}(\tilde{\kappa}) = (X'X + \tilde{\kappa})^{-1}X'Y$.
3. Grâce à $\mathcal{H}_3$ nous savons que $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Le vecteur $\hat{\beta}_{\text{ridge}}$ est une fonction fixée de Y, il suit donc une loi normale d'espérance et de variance à déterminer. L'espérance vaut

$$\begin{aligned}\mathbb{E}(\hat{\beta}_{\text{ridge}}) &= \mathbb{E}((X'X + \tilde{\kappa})^{-1}X'Y) = (X'X + \tilde{\kappa})^{-1}X'\mathbb{E}(Y) \\ &= (X'X + \tilde{\kappa})^{-1}X'X\beta,\end{aligned}$$

   et la variance

$$\begin{aligned}\text{V}(\hat{\beta}_{\text{ridge}}) &= \text{V}((X'X + \tilde{\kappa})^{-1}X'Y) = (X'X + \tilde{\kappa})^{-1}X'\,\text{V}(Y)X(X'X + \tilde{\kappa})^{-1} \\ &= \sigma^2(X'X + \tilde{\kappa})^{-1}X'X(X'X + \tilde{\kappa})^{-1}.\end{aligned}$$

4. Rappelons que la projection orthogonale de Y sur $\Im(X)$, notée $\hat{Y}$ ou encore $P_X Y$ est unique. Par définition, nous avons $P_X Y \perp (Y - \hat{Y})$.
   Par construction $\hat{Y}_{\text{ridge}} = X\hat{\beta}_{\text{ridge}}$ appartient à $\Im(X)$. Selon l'énoncé $\hat{Y}_{\text{ridge}} \neq P_X Y$ donc $(Y - \hat{Y})$ est différent de $Y - \hat{Y}_{\text{ridge}}$ et ils ne sont pas colinéaires. En conclusion, $\hat{Y}$ n'est pas orthogonal à $(Y - \hat{Y}_{\text{ridge}})$.
5. Il faut pouvoir démontrer l'indépendance de $\hat{\sigma}_{\text{ridge}}$ et $\hat{\beta}_{\text{ridge}}$. Pour le théorème **??**, on montre l'indépendance entre $\hat{\beta}$ et $\hat{\sigma}$ en considérant les 2 vecteurs $\hat{\beta}$ et $\hat{\varepsilon} = (Y - \hat{Y})$. Comme nous pouvons écrire $\hat{\beta} = (X'X)^{-1}X'P_X Y$, $\hat{\beta}$ est donc une fonction fixe (dépendante uniquement des X) de $P_X Y$. De plus, $\hat{\varepsilon} = P_{X^\perp} Y$ est orthogonal à

$P_X Y$. Ces 2 vecteurs suivent des lois normales et sont donc indépendants. Il en résulte que $\hat{\beta}$ et $Y - \hat{Y}$ sont indépendants et de même pour $\hat{\beta}$ et $\hat{\sigma}$.
Ici, $\hat{\sigma}_{\text{ridge}}$ est une fonction de $Y - \hat{Y}_{\text{ridge}}$. Le vecteur $\hat{\beta}_{\text{ridge}} = (X'X + \tilde{\kappa} I_p)^{-1} X'Y = (X'X + \tilde{\kappa} I_p)^{-1} X' P_X Y$ est une fonction fixe ($\tilde{\kappa}$ est considéré comme fixé) de $P_X Y$. Par contre, $P_X Y$ n'est pas orthogonal à $(Y - \hat{Y}_{\text{ridge}})$, comme nous l'avons montré, nous ne pouvons donc montrer l'indépendance de $\hat{\beta}_{\text{ridge}}$ et $\hat{\sigma}_{\text{ridge}}$.
Une autre idée serait d'utiliser $\hat{\sigma}$ mais en général si l'on utilise la régression ridge c'est que l'on se doute que $\hat{Y}$ n'est pas un bon estimateur de $X\beta$ et donc *a priori* $\hat{\sigma}$ qui est une fonction de $Y - \hat{Y}$ risque de ne pas être un bon estimateur de σ. L'estimateur $\hat{\sigma}$ peut même être nul, ce qui pratiquement peut arriver quand $p > n$.

6. L'idée repose sur le bootstrap.
   Require: $\tilde{\kappa}$ fixé, α fixé, B choisi.
   Ensure: IC, au niveau α, coordonnée par coordonnée de β .
   Estimer $\beta_{\text{ridge}}(\tilde{\kappa})$.
   En déduire $\hat{\varepsilon}_{\text{ridge}} = Y - X\hat{\beta}_{\text{ridge}}$.
   for $k = 1$ à B **do**
   tirer avec remise n résidus estimés parmi les n coordonnées de $\hat{\varepsilon}_{\text{ridge}}$;
   notons ces résidus (réunis dans 1 vecteur) $\hat{\varepsilon}^{(k)}_{\text{ridge}}$;
   construire 1 échantillon $Y^{(k)} = X\beta_{\text{ridge}}(\tilde{\kappa}) + \hat{\varepsilon}^{(k)}_{\text{ridge}}$;
   $\tilde{\kappa}^{(k)} \leftarrow \tilde{\kappa}$;
   estimer le vecteur de paramètre $\beta^{(k)}_{\text{ridge}}(\tilde{\kappa}^{(k)}) = (X'X + \tilde{\kappa}^{(k)} I_p)^{-1} X' Y^{(k)}$;
   end for
   for $j = 1$ à p **do**
   calculer les quantiles empiriques de niveau $\alpha/2$ et $1 - \alpha/2$ pour la coordonnée j, sur tous les vecteurs $\{\beta^{(k)}_{\text{ridge}}(\tilde{\kappa})\}$;
   end for
7. L'algorithme est presque le même. Cependant comme $\tilde{\kappa}$ n'est pas fixé, pour estimer $\beta_{\text{ridge}}(\tilde{\kappa})$ il faut déterminer $\tilde{\kappa}$ par une méthode choisie. Ensuite, à chaque estimation de $\beta^{(k)}_{\text{ridge}}(\tilde{\kappa}^{(k)})$, il est nécessaire au préalable de déterminer $\tilde{\kappa}^{(k)}$ par la même méthode que celle utilisée pour déterminer $\tilde{\kappa}$.

Exercice 8.8 (Algorithme LARS)

1.9 Régression sur composantes : PCR et PLS

Exercice 9.1 (Questions de cours)
A, B, C, A, B, C.

Exercice 9.2 (Théorème ??)
Elle s'effectue par récurrence. Nous allons ajouter à cette propriété un résultat intermédiaire qui constituera la première partie de la propriété :

$$X^{(j)} = X \prod_{i=1}^{j-1} (I - w^{(i)} (t^{(i)\prime} t^{(i)})^{-1} t^{(i)\prime} X).$$

La seconde partie sera bien sûr de vérifier que $\tilde{w}^{(j)}$ s'écrit bien sous la forme annoncée.

La propriété pour $j = 1$: la première partie n'a pas de sens et, concernant $\tilde{w}^{(j)}$, par construction $X = X^{(1)}$ et donc $\tilde{w}^{(1)} = w^{(1)}$.
La propriété pour $j = 2$ est-elle vraie ?
Nous savons que par définition $X^{(2)} = P_{t^{(1)}\perp} X^{(1)}$ et $X^{(1)} = X$ donc

$$\begin{aligned} X^{(2)} &= P_{t^{(1)}\perp} X^{(1)} = X - P_{t^{(1)}} X = X - t^{(1)}({t^{(1)}}'t^{(1)})^{-1}{t^{(1)}}'X \\ &= X(I - w^{(1)}({t^{(1)}}'t^{(1)})^{-1}{t^{(1)}}'X), \end{aligned}$$

car $t^{(1)} = Xw^{(1)}$. Ceci démontre la première partie de la propriété.
Ensuite, puisque $t^{(2)} = X^{(2)}w^{(2)} = X\tilde{w}^{(2)}$, en remplaçant $X^{(2)}$ par $X(I - w^{(1)}({t^{(1)}}'t^{(1)})^{-1}{t^{(1)}}'X)$ nous avons démontré la propriété pour le rang $j = 2$.
Supposons la propriété vraie au rang $(j-1)$. Nous avons par définition : $X^{(j)} = P_{t^{(j-1)}\perp} X^{(j-1)}$ donc $X^{(j)} = X^{(j-1)} - P_{t^{(j-1)}} X^{(j-1)}$. Or par construction les $\{t^{(k)}\}_{k=1}^{j}$ sont toutes orthogonales donc $P_{t^{(j-1)}} X^{(j-1)} = P_{t^{(j-1)}} X$. Nous avons, grâce à la propriété vraie pour le rang $(j-1)$, que

$$\begin{aligned} X^{(j)} &= X^{(j-1)} - t^{(j-1)}({t^{(j-1)}}'t^{(j-1)})^{-1}{t^{(j-1)}}'X \\ &= X^{(j-1)} - X^{(j-1)}w^{(j-1)}({t^{(j-1)}}'t^{(j-1)})^{-1}{t^{(j-1)}}'X \\ &= X\prod_{i=1}^{j-2}(I - w^{(i)}({t^{(i)}}'t^{(i)})^{-1}{t^{(i)}}'X)(I - w^{(j-1)}({t^{(j-1)}}'t^{(j-1)})^{-1}{t^{(j-1)}}'X) \end{aligned}$$

démontrant la première partie de la proposition. Ensuite, puisque $t^{(j)} = X^{(j)}w^{(j)} = X\tilde{w}^{(j)}$, en remplaçant $X^{(j)}$ par $X\prod_{i=1}^{j-1}(I - w^{(i)}({t^{(i)}}'t^{(i)})^{-1}{t^{(i)}}'X)$, nous avons démontré la propriété pour le rang j.

Exercice 9.3 (Géométrie des estimateurs)

1-4. Les quatre premières réponses sont évidentes, les coordonnées de $\hat{Y}$ valent $1.5, 0.5$ et 0. Ici p vaut 2 et B_1 est un cercle de centre O de rayon 1, alors que B_2 est un losange.

5. Intuitivement, l'image d'un cercle par une application linéaire est une ellipse et l'image d'un losange est un parallélogramme.

6. Le dessin suivant représente les ensembles C_1 et C_2 et $\hat{Y}$ grâce aux ordres GNU-R suivants :

```
X <- matrix(c(1,0,0,1/sqrt(3),2/sqrt(3),0),3,2)
sss <- 1
iter <- 1
coord <- matrix(0,500,2)
for (tt in  seq(-pi,pi,length=500)) {
   coord[iter,] <- (X%*%as.matrix(sqrt(sss)
                        *c(cos(tt),sin(tt))))[1:2,]
   iter <- iter+1
 }
iter <- 1
coord2 <- matrix(0,500,2)
for (tt in  seq(-1,1,length=250)) {
   coord2[iter,] <- (X%*%as.matrix(c(tt,1-abs(tt))))[1:2,]
```

```
   coord2[iter+250,] <- (X%*%as.matrix(c(tt,
                                abs(tt)-1)))[1:2,]
   iter <- iter+1
 }
plot(coord,type="l",xlab="",ylab="")
lines(coord2)
```

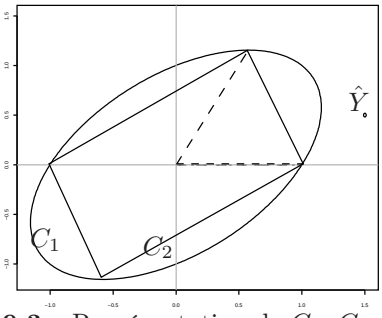

Fig. 9.3 – Représentation de C_1, C_2 et $\hat{Y}$.

7. Par définition, $X\hat{\beta}_{\text{ridge}}$ est l'élément de C_1 le plus proche de $\hat{Y}$. De même, $X\hat{\beta}_{\text{lasso}}$ est l'élément de C_2 le plus proche de $\hat{Y}$. Cela donne graphiquement

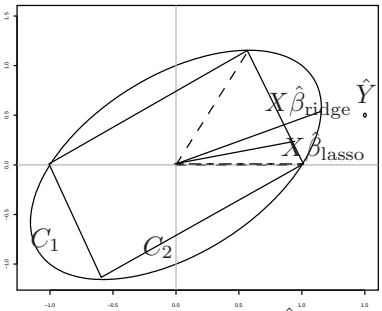

Fig. 9.4 – Représentation de $X\hat{\beta}_{\text{ridge}}$ et $X\hat{\beta}_{\text{lasso}}$.

8. L'ensemble C_1, composé de vecteurs de la forme $u = X_1\alpha_1 + X_2\alpha_2$ avec la norme du vecteur α valant 1, peut être vu comme l'ensemble des composantes dans lequel on va choisir la composante PLS. La première composante PLS est le vecteur de C_1 dont le produit scalaire avec Y (et donc $\hat{Y}$) est le plus grand. Graphiquement, c'est le vecteur de C_1 dont l'extrémité sur l'ellipse est le pied de la tangente à l'ellipse perpendiculaire à $O\hat{Y}$. La prévision de Y par la régression PLS est la projection de Y et donc de $\hat{Y}$ sur la composante PLS.
9. La calcul donne simplement

$$X'X = \begin{pmatrix} 1 & \sqrt{3}/3 \\ \sqrt{3}/3 & 5/3 \end{pmatrix}.$$

Les valeurs propres sont 2 et 2/3. Le premier axe principal correspond au vecteur propre associé à la valeur 2. Pour trouver la première composante principale, il faut

pré-multiplier ce vecteur par X. Cela donne le vecteur de coordonnées $(1, 1, 0)'$. Les commandes GNU-R sont

```
X <- matrix(c(1,0,0,1/sqrt(3),2/sqrt(3),0),3,2)
svd <- eigen(t(X)%*%X)
X%*%svd$vect[,1]
```

10. La prévision de Y par la régression PCR est la projection de Y (et donc de $\hat{Y}$) sur la composante PCR. Dans cet exemple, la projection de $\hat{Y}$ sur la composante PCR est un point de l'ellipse, mais cela est uniquement dû aux données de cet exercice. Le graphique suivant représente les 4 projections :

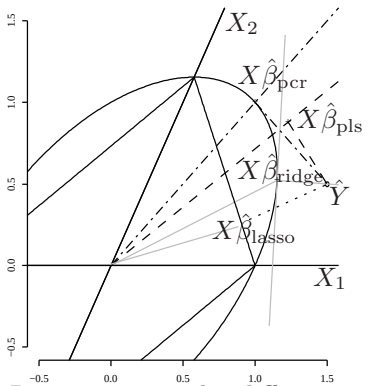

Fig. 9.5 – Représentation des différentes projections.

1.10 Régression spline et régression à noyau

Exercice 10.1 (Questions de cours)
A, A, A.

Exercice 10.2
Les poids changent en fonction du point x, mais à x fixé, nous effectuons simplement une régression pondérée comme indiqué à l'exercice précédent.

Exercice 10.3

Exercice 10.4

Exercice 10.5

Exercice 10.6 (Estimateur de Nadaraya-Watson)
La matrice P vaut

$$P = \text{diag}\left(K(\frac{x - X_i}{h})\right),$$

et la matrice X vaut

$$X' = \left[\begin{array}{ccc} 1 & \cdots & 1 \end{array} \right].$$

L'estimateur vaut

$$\hat{\beta}(x) = (X'PX)^{-1}X'PY.$$

Le calcul donne le résultat proposé.

Exercice 10.7 (Polynômes locaux)
La matrice P vaut

$$P = \text{diag}\left(K(\frac{x - X_i}{h})\right),$$

et la matrice X vaut

$$X' = \left[\begin{array}{ccc} 1 & \cdots & 1 \\ (X_1 - x) & \cdots & (X_n - x) \end{array} \right].$$

L'estimateur vaut

$$\hat{\beta}(x) = (X'PX)^{-1}X'PY.$$

Le calcul donne

$$\begin{pmatrix} \sum K(\frac{x-X_i}{h}) & \sum (X_i - x)K(\frac{x-X_i}{h}) \\ \sum (X_i - x)K(\frac{x-X_i}{h}) & \sum (X_i - x)^2 K(\frac{x-X_i}{h}) \end{pmatrix}^{-1} \begin{pmatrix} \sum K(\frac{x-X_i}{h})Y_i \\ \sum (X_i - x)K(\frac{x-X_i}{h})Y_i \end{pmatrix}.$$

Posons

$$\begin{aligned} S_0 &= \sum K(\frac{x - X_i}{h}) \\ S_1 &= \sum (X_i - x)K(\frac{x - X_i}{h}) \\ S_2 &= \sum (X_i - x)^2 K(\frac{x - X_i}{h}). \end{aligned}$$

Cela nous donne, après calcul de l'inverse,

$$\begin{aligned} \hat{\beta}(x,h) &= \frac{1}{S_0 S_2 - S_1^2} \begin{pmatrix} S_2 & -S_1 \\ -S_1 & S_0 \end{pmatrix} \begin{pmatrix} \sum K(\frac{x-X_i}{h})Y_i \\ \sum (X_i - x)K(\frac{x-X_i}{h})Y_i \end{pmatrix} \\ &= \frac{1}{S_0 S_2 - S_1^2} \begin{pmatrix} S_2 \sum K(\frac{x-X_i}{h})Y_i - S_1 \sum (X_i - x)K(\frac{x-X_i}{h})Y_i \\ S_0 \sum (X_i - x)K(\frac{x-X_i}{h})Y_i - S_1 \sum K(\frac{x-X_i}{h})Y_i \end{pmatrix} \end{aligned}$$

et finalement, en ne prenant que la première composante de $\hat{\beta}$, nous obtenons le résultat énoncé.

Bibliographie

Lehmann E.L. (1959). *Testing statistical hypotheses.* John Wiley.