

RÉGRESSIONS LINÉAIRES

Jérémie Cabessa

Laboratoire DAVID, UVSQ

MODÈLES LINÉAIRES

- ▶ Les **modèles linéaires** sont les plus simples, mais également les plus rapides et parmi les plus utiles.
- ▶ Une approche linéaire devrait toujours être envisagée avant de passer à des modèles plus complexes.

MODÈLES LINÉAIRES

- ▶ Les **modèles linéaires** sont les plus simples, mais également les plus rapides et parmi les plus utiles.
- ▶ Une approche linéaire devrait toujours être envisagée avant de passer à des modèles plus complexes.

RÉGRESSION LINÉAIRE

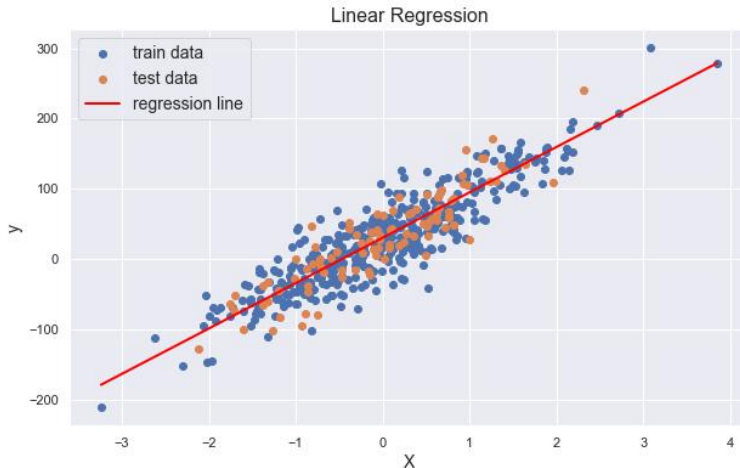


Figure 2 taken from [James et al., 2013]

RÉGRESSION LINÉAIRE

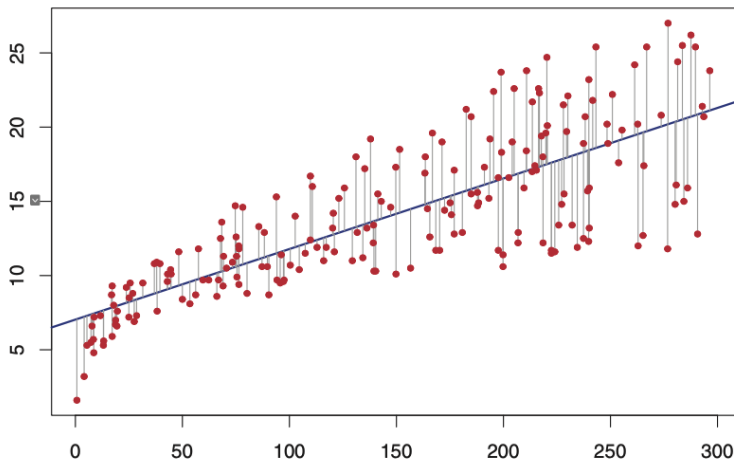


Figure 2 taken from [James et al., 2013]

RÉGRESSION LINÉAIRE SIMPLE

- Soient X une variables explicative et Y une variable réponse.
- Hypothèse forte: on suppose que la “vraie” relation entre X et Y est de la forme linéaire suivante:

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (1)$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

RÉGRESSION LINÉAIRE SIMPLE

- ▶ Soient X une variables explicative et Y une variable réponse.
- ▶ **Hypothèse forte:** on suppose que la “vraie” relation entre X et Y est de la forme linéaire suivante:

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (1)$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

RÉGRESSION LINÉAIRE SIMPLE

- ▶ Les “vrais” paramètres β_0, β_1 sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \hat{\beta}_1$ des ces paramètres.
- ▶ Une fois les estimateurs obtenus, la **prédiction** associée à une observation x est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (2)$$

- ▶ Pour obtenir les estimateurs $\hat{\beta}_0, \hat{\beta}_1$, on dispose d'observations (ou de data).

RÉGRESSION LINÉAIRE SIMPLE

- ▶ Les “vrais” paramètres β_0, β_1 sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \hat{\beta}_1$ des ces paramètres.
- ▶ Une fois les estimateurs obtenus, la **prédiction** associée à une observation x est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (2)$$

- ▶ Pour obtenir les estimateurs $\hat{\beta}_0, \hat{\beta}_1$, on dispose d'observations (ou de data).

RÉGRESSION LINÉAIRE SIMPLE

- ▶ Les “vrais” paramètres β_0, β_1 sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \hat{\beta}_1$ des ces paramètres.
- ▶ Une fois les estimateurs obtenus, la **prédiction** associée à une observation x est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (2)$$

- ▶ Pour obtenir les estimateurs $\hat{\beta}_0, \hat{\beta}_1$, on dispose d'observations (ou de data).

RÉGRESSION LINÉAIRE SIMPLE

- Soit un **training set** formé de N observations:

$$S_{\text{train}} = \{(x_1, y_1), \dots, (x_N, y_N)\}.$$

- On choisit les estimateurs $\hat{\beta}_0, \hat{\beta}_1$ qui minimisent la **somme des erreurs quadratiques (residual sum of squares (RSS))**, i.e., la somme des distances entre les prédictions et les réponses:

$$\begin{aligned} \text{RSS}(\beta_0, \beta_1) &= \sum_{i=1}^N (\hat{y}_i - y_i)^2 \\ &= \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 \end{aligned}$$

RÉGRESSION LINÉAIRE SIMPLE

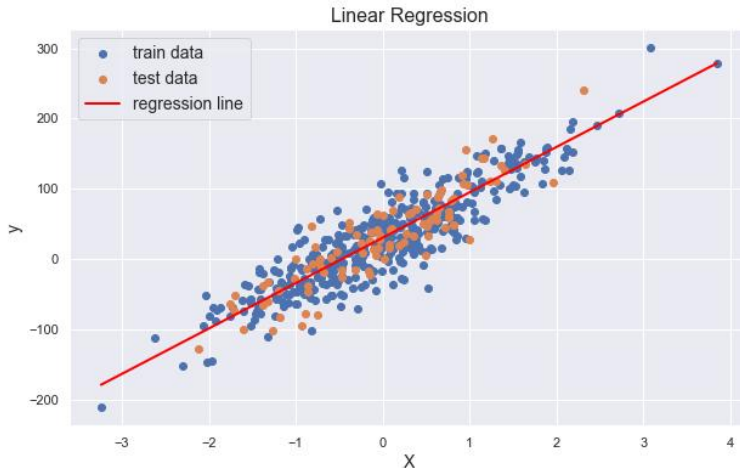
- Soit un **training set** formé de N observations:

$$S_{\text{train}} = \{(x_1, y_1), \dots, (x_N, y_N)\}.$$

- On choisit les estimateurs $\hat{\beta}_0, \hat{\beta}_1$ qui minimisent la **somme des erreurs quadratiques (residual sum of squares (RSS))**, i.e., la somme des distances entre les prédictions et les réponses:

$$\begin{aligned} \text{RSS}(\beta_0, \beta_1) &= \sum_{i=1}^N (\hat{y}_i - y_i)^2 \\ &= \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 \end{aligned}$$

RÉGRESSION LINÉAIRE SIMPLES



Figures taken from [James et al., 2013]

RÉGRESSION LINÉAIRE SIMPLE

► On a alors:

$$\begin{aligned}\hat{\beta}_0, \hat{\beta}_1 &= \arg \min_{\beta_0, \beta_1} \text{RSS}(\beta_0, \beta_1) \\ &= \arg \min_{\beta_0, \beta_1} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2\end{aligned}$$

RÉGRESSION LINÉAIRE SIMPLE

- Pour trouver le minimum de $\text{RSS}(\beta_0, \beta_1)$, on annule la dérivée de cette expression par rapport à β_0 et β_1 :

$$\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 = 0$$

$$\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 = 0$$

ce qui donne comme solutions

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^N x_i (y_i - \bar{y})}{\sum_{i=1}^N x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\ \text{où } \bar{x} &= \frac{1}{N} \sum_{i=1}^N x_i \text{ et } \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i\end{aligned}$$

RÉGRESSION LINÉAIRE SIMPLE

- Pour trouver le minimum de $\text{RSS}(\beta_0, \beta_1)$, on annule la dérivée de cette expression par rapport à β_0 et β_1 :

$$\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 = 0$$

$$\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 = 0$$

ce qui donne comme solutions

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^N x_i (y_i - \bar{y})}{\sum_{i=1}^N x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\ \text{où } \bar{x} &= \frac{1}{N} \sum_{i=1}^N x_i \text{ et } \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i\end{aligned}$$

RÉGRESSION LINÉAIRE SIMPLE

- Pour trouver le minimum de $\text{RSS}(\beta_0, \beta_1)$, on annule la dérivée de cette expression par rapport à β_0 et β_1 :

$$\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 = 0$$

$$\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 = 0$$

ce qui donne comme solutions

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^N x_i (y_i - \bar{y})}{\sum_{i=1}^N x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\ \text{où } \bar{x} &= \frac{1}{N} \sum_{i=1}^N x_i \text{ et } \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i\end{aligned}$$

RÉGRESSION LINÉAIRE SIMPLE

- Pour trouver le minimum de $\text{RSS}(\beta_0, \beta_1)$, on annule la dérivée de cette expression par rapport à β_0 et β_1 :

$$\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 = 0$$

$$\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{i=1}^N ((\beta_0 + \beta_1 x_i) - y_i)^2 = 0$$

ce qui donne comme solutions

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^N x_i (y_i - \bar{y})}{\sum_{i=1}^N x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\ \text{où } \bar{x} &= \frac{1}{N} \sum_{i=1}^N x_i \text{ et } \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i\end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ Soient X_1, \dots, X_p des variables explicatives et Y une variable réponse.
- ▶ **Hypothèse forte:** on suppose que la “vraie” relation entre X_1, \dots, X_p et Y est de la forme linéaire suivante:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon \\ &= \beta_0 + \sum_{i=1}^p \beta_i X_i + \epsilon \end{aligned} \tag{3}$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ **Interprétation:** chaque β_i ($i = 1, \dots, p$) représente l'effet moyen sur Y de l'accroissement d'une unité de X_i , si tous les autres X_j restent fixes.

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ Soient X_1, \dots, X_p des variables explicatives et Y une variable réponse.
- ▶ **Hypothèse forte:** on suppose que la “vraie” relation entre X_1, \dots, X_p et Y est de la forme linéaire suivante:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon \\ &= \beta_0 + \sum_{i=1}^p \beta_i X_i + \epsilon \end{aligned} \tag{3}$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ **Interprétation:** chaque β_i ($i = 1, \dots, p$) représente l'effet moyen sur Y de l'accroissement d'une unité de X_i , si tous les autres X_j restent fixes.

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ Soient X_1, \dots, X_p des variables explicatives et Y une variable réponse.
- ▶ **Hypothèse forte:** on suppose que la “vraie” relation entre X_1, \dots, X_p et Y est de la forme linéaire suivante:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon \\ &= \beta_0 + \sum_{i=1}^p \beta_i X_i + \epsilon \end{aligned} \tag{3}$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ **Interprétation:** chaque β_i ($i = 1, \dots, p$) représente l'effet moyen sur Y de l'accroissement d'une unité de X_i , si tous les autres X_j restent fixes.

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ Les “vrais” paramètres β_0, \dots, β_p sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \dots, \hat{\beta}_p$ des ces paramètres.
- ▶ Une fois les estimateurs obtenus, la **prédiction** associée à toute observation $\mathbf{x} = (x_1, \dots, x_p)$ est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p = \mathbf{x}^T \hat{\boldsymbol{\beta}} \quad (4)$$

où $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ et $\mathbf{x} = (1, x_1, \dots, x_p)$ (on a rajouté la composante 1).

- ▶ Pour obtenir les estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres, on dispose d'observations (ou de data).

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ Les “vrais” paramètres β_0, \dots, β_p sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \dots, \hat{\beta}_p$ des ces paramètres.
- ▶ Une fois les estimateurs obtenus, la **prédiction** associée à toute observation $\mathbf{x} = (x_1, \dots, x_p)$ est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p = \mathbf{x}^T \hat{\boldsymbol{\beta}} \quad (4)$$

où $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ et $\mathbf{x} = (1, x_1, \dots, x_p)$ (on a rajouté la composante 1).

- ▶ Pour obtenir les estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres, on dispose d'observations (ou de data).

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ Les “vrais” paramètres β_0, \dots, β_p sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \dots, \hat{\beta}_p$ des ces paramètres.
- ▶ Une fois les estimateurs obtenus, la **prédiction** associée à toute observation $\mathbf{x} = (x_1, \dots, x_p)$ est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p = \mathbf{x}^T \hat{\boldsymbol{\beta}} \quad (4)$$

où $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ et $\mathbf{x} = (1, x_1, \dots, x_p)$ (on a rajouté la composante 1).

- ▶ Pour obtenir les estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres, on dispose d'observations (ou de data).

RÉGRESSION LINÉAIRE MULTIPLE

- Soit un **training set** formé de N observations:

$$S_{\text{train}} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}.$$

- On définit les matrice et vecteur:

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \\ 1 & \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix} \quad \text{et} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

RÉGRESSION LINÉAIRE MULTIPLE

- Soit un **training set** formé de N observations:

$$S_{\text{train}} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}.$$

- On définit les matrice et vecteur:

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \\ 1 & \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix} \quad \text{et} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ On choisit les estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ qui minimisent une **fonction de coût (loss function)** $\mathcal{L}(\mathbf{X}, \mathbf{y}; \beta)$.
- ▶ On minimise la **somme des erreur quadratiques (residual sum of squares (RSS))**, i.e., distances entre prédictions et réponses:

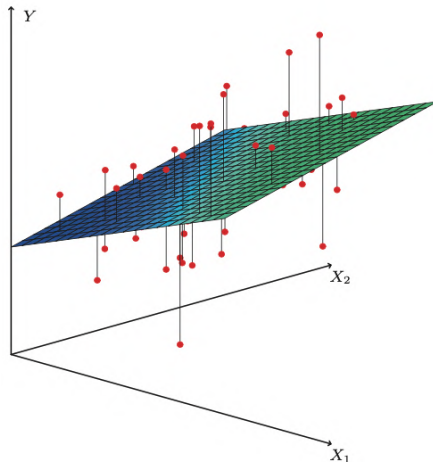
$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (\hat{y}_i - y_i)^2 \\ &= \sum_{i=1}^N (x_i^T \beta - y_i)^2 = \|\mathbf{X}\beta - \mathbf{y}\|^2\end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ On choisit les estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ qui minimisent une **fonction de coût (loss function)** $\mathcal{L}(\mathbf{X}, \mathbf{y}; \beta)$.
- ▶ On minimise la **somme des erreur quadratiques (residual sum of squares (RSS))**, i.e., distances entre prédictions et réponses:

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (\hat{y}_i - y_i)^2 \\ &= \sum_{i=1}^N (\mathbf{x}_i^T \beta - y_i)^2 = \|\mathbf{X}\beta - \mathbf{y}\|^2\end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE



Figures taken from [James et al., 2013]

APARTÉ: QUELQUES RAPPELS

- ▶ La norme au carrée d'un vecteur \mathbf{v} vaut $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$.
- ▶ Si r est un scalaire et $\beta = (\beta_0, \dots, \beta_p)$ un vecteur (r dépend potentiellement de β), on a le vecteur **gradient**:

$$\frac{\partial r}{\partial \beta} = \begin{pmatrix} \frac{\partial r}{\partial \beta_0} \\ \vdots \\ \frac{\partial r}{\partial \beta_p} \end{pmatrix}$$

- ▶ On a les résultats de dérivations suivants (à vérifier!):

$$\frac{\partial (\mathbf{y}^T \mathbf{y})}{\partial \beta} = \mathbf{0} \text{ (vecteur nul)}$$

APARTÉ: QUELQUES RAPPELS

- ▶ La norme au carrée d'un vecteur \mathbf{v} vaut $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$.
- ▶ Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ un vecteur (r dépend potentiellement de $\boldsymbol{\beta}$), on a le vecteur **gradient**:

$$\frac{\partial r}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial r}{\partial \beta_0} \\ \vdots \\ \frac{\partial r}{\partial \beta_p} \end{pmatrix}$$

- ▶ On a les résultats de dérivations suivants (à vérifier!):

$$\frac{\partial (\mathbf{y}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{0} \text{ (vecteur nul)}$$

APARTÉ: QUELQUES RAPPELS

- ▶ La norme au carrée d'un vecteur \mathbf{v} vaut $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$.
- ▶ Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ un vecteur (r dépend potentiellement de $\boldsymbol{\beta}$), on a le vecteur **gradient**:

$$\frac{\partial r}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial r}{\partial \beta_0} \\ \vdots \\ \frac{\partial r}{\partial \beta_p} \end{pmatrix}$$

- ▶ On a les résultats de dérivations suivants (à vérifier!):

$$\frac{\partial (\mathbf{y}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{0} \text{ (vecteur nul)}$$

$$\frac{\partial (\mathbf{y}^T \mathbf{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = (\mathbf{y}^T \mathbf{X})^T = \mathbf{X}^T \mathbf{y} \quad \text{et} \quad \frac{\partial (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$$

$$\frac{\partial (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$$

APARTÉ: QUELQUES RAPPELS

- ▶ La norme au carrée d'un vecteur \mathbf{v} vaut $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$.
- ▶ Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ un vecteur (r dépend potentiellement de $\boldsymbol{\beta}$), on a le vecteur **gradient**:

$$\frac{\partial r}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial r}{\partial \beta_0} \\ \vdots \\ \frac{\partial r}{\partial \beta_p} \end{pmatrix}$$

- ▶ On a les résultats de dérivations suivants (à vérifier!):

$$\frac{\partial (\mathbf{y}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{0} \text{ (vecteur nul)}$$

$$\frac{\partial (\mathbf{y}^T \mathbf{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = (\mathbf{y}^T \mathbf{X})^T = \mathbf{X}^T \mathbf{y} \quad \text{et} \quad \frac{\partial (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$$

$$\frac{\partial (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$$

APARTÉ: QUELQUES RAPPELS

- ▶ La norme au carrée d'un vecteur \mathbf{v} vaut $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$.
- ▶ Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ un vecteur (r dépend potentiellement de $\boldsymbol{\beta}$), on a le vecteur **gradient**:

$$\frac{\partial r}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial r}{\partial \beta_0} \\ \vdots \\ \frac{\partial r}{\partial \beta_p} \end{pmatrix}$$

- ▶ On a les résultats de dérivations suivants (à vérifier!):

$$\frac{\partial (\mathbf{y}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{0} \text{ (vecteur nul)}$$

$$\frac{\partial (\mathbf{y}^T \mathbf{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = (\mathbf{y}^T \mathbf{X})^T = \mathbf{X}^T \mathbf{y} \quad \text{et} \quad \frac{\partial (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$$

$$\frac{\partial (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$$

APARTÉ: QUELQUES RAPPELS

- ▶ La norme au carrée d'un vecteur \mathbf{v} vaut $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$.
- ▶ Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ un vecteur (r dépend potentiellement de $\boldsymbol{\beta}$), on a le vecteur **gradient**:

$$\frac{\partial r}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial r}{\partial \beta_0} \\ \vdots \\ \frac{\partial r}{\partial \beta_p} \end{pmatrix}$$

- ▶ On a les résultats de dérivations suivants (à vérifier!):

$$\frac{\partial (\mathbf{y}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{0} \text{ (vecteur nul)}$$

$$\frac{\partial (\mathbf{y}^T \mathbf{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = (\mathbf{y}^T \mathbf{X})^T = \mathbf{X}^T \mathbf{y} \quad \text{et} \quad \frac{\partial (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$$

$$\frac{\partial (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$$

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- ▶ Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

RÉGRESSION LINÉAIRE MULTIPLE

- On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \\ &\quad \text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- ▶ On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- ▶ Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \\ &\quad \text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad &\text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \\ &\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \\ &\quad \text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \\ &\quad \text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta &= \mathbf{X}^T \mathbf{y} \\ \text{ssi} \quad \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta &= \mathbf{X}^T \mathbf{y} \\ \text{ssi} \quad \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta &= \mathbf{X}^T \mathbf{y} \\ \text{ssi} \quad \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- On a alors:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|\mathbf{X}\beta - \mathbf{y}\|^2)}{\partial \beta} = \frac{\partial ((\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y}))}{\partial \beta} \\ &= \frac{\partial (\beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \mathbf{y}^T \mathbf{y})}{\partial \beta} \\ &= 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \\ &\quad \text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- Soit un **test set** formé de M observations:

$$S_{\text{test}} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N'}, y_M)\}.$$

- Une fois les estimateurs $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ obtenus, les prédictions \hat{y} associés aux data \mathbf{X} sont données par:

$$\hat{y}_i = \mathbf{x}_i^T \hat{\beta} \quad \text{pour } i = 1, \dots, M, \text{ i.e.,}$$

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\beta} \quad (\text{vectors vectorisés})$$

RÉGRESSION LINÉAIRE MULTIPLE

- Soit un **test set** formé de M observations:

$$S_{\text{test}} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N'}, y_M)\}.$$

- Une fois les estimateurs $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ obtenus, les prédictions $\hat{\mathbf{y}}$ associés aux data \mathbf{X} sont données par:

$$\begin{aligned}\hat{y}_i &= \mathbf{x}_i^T \hat{\beta} \text{ pour } i = 1, \dots, M, \text{ i.e.,} \\ \hat{\mathbf{y}} &= \mathbf{X} \hat{\beta} \text{ (forme vectorielle)}\end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- Soit un **test set** formé de M observations:

$$S_{\text{test}} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N'}, y_M)\}.$$

- Une fois les estimateurs $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ obtenus, les prédictions $\hat{\mathbf{y}}$ associés aux data \mathbf{X} sont données par:

$$\begin{aligned}\hat{y}_i &= \mathbf{x}_i^T \hat{\beta} \text{ pour } i = 1, \dots, M, \text{ i.e.,} \\ \hat{\mathbf{y}} &= \mathbf{X} \hat{\beta} \text{ (forme vectorielle)}\end{aligned}$$

RÉGRESSION LINÉAIRE MULTIPLE

- Soit un **test set** formé de M observations:

$$S_{\text{test}} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N'}, y_M)\}.$$

- Une fois les estimateurs $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ obtenus, les prédictions $\hat{\mathbf{y}}$ associés aux data \mathbf{X} sont données par:

$$\begin{aligned}\hat{y}_i &= \mathbf{x}_i^T \hat{\beta} \text{ pour } i = 1, \dots, M, \text{ i.e.,} \\ \hat{\mathbf{y}} &= \mathbf{X} \hat{\beta} \text{ (forme vectorielle)}\end{aligned}$$

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ Polynomial regression
 - ▶ Logistic regression
 - ▶ Kernel functions
 - ▶ Support vector machines
 - ▶ Boosting algorithms
 - ▶ Ensemble methods
- ▶ On ne présentera pas ces méthodes en détail ici...

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ Polynomial regression
 - ▶ Step functions
 - ▶ Basis functions
 - ▶ Regression splines
 - ▶ Smoothing splines
 - ▶ Local regression
- ▶ On ne présentera pas ces méthodes en détail ici...

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ **Polynomial regression**
 - ▶ Step functions
 - ▶ Basis functions
 - ▶ Regression splines
 - ▶ Smoothing splines
 - ▶ Local regression
- ▶ On ne présentera pas ces méthodes en détail ici...

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ **Polynomial regression**
 - ▶ **Step functions**
 - ▶ Basis functions
 - ▶ Regression splines
 - ▶ Smoothing splines
 - ▶ Local regression
- ▶ On ne présentera pas ces méthodes en détail ici...

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ **Polynomial regression**
 - ▶ **Step functions**
 - ▶ **Basis functions**
 - ▶ Regression splines
 - ▶ Smoothing splines
 - ▶ Local regression
- ▶ On ne présentera pas ces méthodes en détail ici...

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ **Polynomial regression**
 - ▶ **Step functions**
 - ▶ **Basis functions**
 - ▶ **Regression splines**
 - ▶ Smoothing splines
 - ▶ Local regression
- ▶ On ne présentera pas ces méthodes en détail ici...

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ **Polynomial regression**
 - ▶ **Step functions**
 - ▶ **Basis functions**
 - ▶ **Regression splines**
 - ▶ **Smoothing splines**
 - ▶ **Local regression**
- ▶ On ne présentera pas ces méthodes en détail ici...

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ **Polynomial regression**
 - ▶ **Step functions**
 - ▶ **Basis functions**
 - ▶ **Regression splines**
 - ▶ **Smoothing splines**
 - ▶ **Local regression**

▶ On ne présentera pas ces méthodes en détail ici...

AUTRES RÉGRESSIONS

- ▶ Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- ▶ Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y , on a:
 - ▶ **Polynomial regression**
 - ▶ **Step functions**
 - ▶ **Basis functions**
 - ▶ **Regression splines**
 - ▶ **Smoothing splines**
 - ▶ **Local regression**
- ▶ On ne présentera pas ces méthodes en détail ici...

RÉGULARISATION

- ▶ Rappel: on suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Il se peut que certaines des variables X_i soient peu ou pas du tout associées avec la réponse Y_i .
- ▶ Inclure ces variables accroît la complexité du modèle, affecte sa performance, et réduit son interprétabilité.
- ▶ Il existe alors des méthodes de réduction et/ou sélection des variables les plus significatives: shrinkage et feature selection.

RÉGULARISATION

- ▶ Rappel: on suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Il se peut que certaines des variables X_i soient peu ou pas du tout associées avec la réponse Y_i .
- ▶ Inclure ces variables accroît la complexité du modèle, affecte sa performance, et réduit son interprétabilité.
- ▶ Il existe alors des méthodes de réduction et/ou sélection des variables les plus significatives: **shrinkage** et **feature selection**.

RÉGULARISATION

- ▶ Rappel: on suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Il se peut que certaines des variables X_i soient peu ou pas du tout associées avec la réponse Y_i .
- ▶ Inclure ces variables accroît la complexité du modèle, affecte sa performance, et réduit son interprétabilité.
- ▶ Il existe alors des méthodes de réduction et/ou sélection des variables les plus significatives: **shrinkage** et **feature selection**.

RÉGULARISATION

- ▶ Rappel: on suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Il se peut que certaines des variables X_i soient peu ou pas du tout associées avec la réponse Y_i .
- ▶ Inclure ces variables accroît la complexité du modèle, affecte sa performance, et réduit son interprétabilité.
- ▶ Il existe alors des méthodes de réduction et/ou sélection des variables les plus significatives: **shrinkage** et **feature selection**.

RÉGULARISATION

Parmi ces méthodes, on a:

- ▶ Subset selection methods:

- ▶ best subset selection
- ▶ forward stepwise selection
- ▶ backward stepwise selection

- ▶ Regularization methods:

- ▶ Ridge regression (shrinkage)
- ▶ LASSO (subset selection)

- ▶ On s'intéresse ici aux Ridge regression et LASSO.

RÉGULARISATION

Parmi ces méthodes, on a:

- ▶ Subset selection methods:

- ▶ best subset selection
- ▶ forward stepwise selection
- ▶ backward stepwise selection

- ▶ Regularization methods:

- ▶ Ridge regression (penalized)
- ▶ Lasso (penalized)

- ▶ On s'intéresse ici aux Ridge regression et LASSO.

RÉGULARISATION

Parmi ces méthodes, on a:

▶ Subset selection methods:

- ▶ best subset selection
- ▶ forward stepwise selection
- ▶ backward stepwise selection

▶ Regularization methods:

Ridge regression (penalized
least squares)

▶ On s'intéresse ici aux Ridge regression et LASSO.

RÉGULARISATION

Parmi ces méthodes, on a:

- ▶ Subset selection methods:
 - ▶ best subset selection
 - ▶ forward stepwise selection
 - ▶ backward stepwise selection
- ▶ Regularization methods:
 - ▶ Ridge regression (shrinkage)
 - ▶ LASSO (shrinkage)
- ▶ On s'intéresse ici aux Ridge regression et LASSO.

RÉGULARISATION

Parmi ces méthodes, on a:

- ▶ Subset selection methods:
 - ▶ best subset selection
 - ▶ forward stepwise selection
 - ▶ backward stepwise selection
- ▶ Regularization methods:
 - ▶ Ridge regression (shrinkage)
 - ▶ LASSO (feature selection)
- ▶ On s'intéresse ici aux Ridge regression et LASSO.

RÉGULARISATION

Parmi ces méthodes, on a:

- ▶ Subset selection methods:
 - ▶ best subset selection
 - ▶ forward stepwise selection
 - ▶ backward stepwise selection
- ▶ Regularization methods:
 - ▶ Ridge regression (shrinkage)
 - ▶ LASSO (feature selection)
- ▶ On s'intéresse ici aux Ridge regression et LASSO.

RÉGULARISATION

Parmi ces méthodes, on a:

- ▶ Subset selection methods:
 - ▶ best subset selection
 - ▶ forward stepwise selection
 - ▶ backward stepwise selection
 - ▶ Regularization methods:
 - ▶ Ridge regression (shrinkage)
 - ▶ LASSO (feature selection)
- ▶ On s'intéresse ici aux Ridge regression et LASSO.

RÉGULARISATION

Parmi ces méthodes, on a:

- ▶ Subset selection methods:
 - ▶ best subset selection
 - ▶ forward stepwise selection
 - ▶ backward stepwise selection
- ▶ Regularization methods:
 - ▶ Ridge regression (shrinkage)
 - ▶ LASSO (feature selection)
- ▶ On s'intéresse ici aux **Ridge regression** et **LASSO**.

RÉGRESSION RIDGE

- ▶ On suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Le but est d'obtenir des estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres $\beta = (\beta_0, \dots, \beta_p)$.
- ▶ La **régression Ridge** permet de forcer les estimateurs à ne pas exploser, ce qui a comme effet positif de réduire la variance du modèle.
- ▶ En gros, les variables X_i les moins significatives ont leur estimateurs associés $\hat{\beta}_i$ qui convergent vers 0 (shrinkage method).

RÉGRESSION RIDGE

- ▶ On suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Le but est d'obtenir des estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres $\beta = (\beta_0, \dots, \beta_p)$.
- ▶ La régression Ridge permet de forcer les estimateurs à ne pas exploser, ce qui a comme effet positif de réduire la variance du modèle.
- ▶ En gros, les variables X_i les moins significatives ont leur estimateurs associés $\hat{\beta}_i$ qui convergent vers 0 (shrinkage method).

RÉGRESSION RIDGE

- ▶ On suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Le but est d'obtenir des estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres $\beta = (\beta_0, \dots, \beta_p)$.
- ▶ La **régression Ridge** permet de forcer les estimateurs à ne pas exploser, ce qui a comme effet positif de réduire la variance du modèle.
- ▶ En gros, les variables X_i les moins significatives ont leur estimateurs associés $\hat{\beta}_i$ qui convergent vers 0 (shrinkage method).

RÉGRESSION RIDGE

- ▶ On suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Le but est d'obtenir des estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres $\beta = (\beta_0, \dots, \beta_p)$.
- ▶ La **régression Ridge** permet de forcer les estimateurs à ne pas exploser, ce qui a comme effet positif de réduire la variance du modèle.
- ▶ En gros, les variables X_i les moins significatives ont leur estimateurs associés $\hat{\beta}_i$ qui convergent vers 0 (shrinkage method).

RÉGRESSION RIDGE

- **Régression Ridge:** on choisit les estimateurs $\hat{\beta}$ qui minimisent une version *régularisée* la **residual sum of squares (RSS)**

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (x_i^T \beta - y_i)^2 + \lambda \sum_{j=0}^p \beta_j^2 \\ &= \|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2\end{aligned}$$

où λ est le *paramètre de régularisation*.

Le paramètre λ est une *hyperparamètre* à régler.

RÉGRESSION RIDGE

- **Régression Ridge:** on choisit les estimateurs $\hat{\beta}$ qui minimisent une version *régularisée* la **residual sum of squares (RSS)**

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (x_i^T \beta - y_i)^2 + \lambda \sum_{j=0}^p \beta_j^2 \\ &= \|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2\end{aligned}$$

► où λ est le *paramètre de régularisation*.

► Le terme $\lambda \|\beta\|_2^2$ est une *pénalité l_2* (l_2 penalty).

RÉGRESSION RIDGE

- **Régression Ridge:** on choisit les estimateurs $\hat{\beta}$ qui minimisent une version *régularisée* la **residual sum of squares (RSS)**

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (x_i^T \beta - y_i)^2 + \lambda \sum_{j=0}^p \beta_j^2 \\ &= \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_2^2\end{aligned}$$

- où λ est le *paramètre de régularisation*.
- Le terme $\lambda \|\beta\|_2^2$ est une *pénalité l_2* (l_2 penalty).

RÉGRESSION RIDGE

- **Régression Ridge:** on choisit les estimateurs $\hat{\beta}$ qui minimisent une version *régularisée* la **residual sum of squares (RSS)**

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (x_i^T \beta - y_i)^2 + \lambda \sum_{j=0}^p \beta_j^2 \\ &= \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_2^2\end{aligned}$$

- où λ est le *paramètre de régularisation*.
- Le terme $\lambda \|\beta\|_2^2$ est une *pénalité l_2* (l_2 penalty).

RÉGRESSION RIDGE

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

RÉGRESSION RIDGE

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2)}{\partial \beta} \\ &= 2X^T X\beta - 2X^T y + 2\lambda\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad (X^T X + \lambda I)\beta = X^T y \\ &\quad \text{ssi} \quad \hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \end{aligned}$$

RÉGRESSION RIDGE

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2)}{\partial \beta} \\ &= 2X^T X\beta - 2X^T y + 2\lambda\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad (X^T X + \lambda I)\beta = X^T y \\ &\quad \text{ssi} \quad \hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \end{aligned}$$

RÉGRESSION RIDGE

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2)}{\partial \beta} \\ &= 2X^T X\beta - 2X^T y + 2\lambda\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad (X^T X + \lambda I)\beta = X^T y \\ &\quad \text{ssi} \quad \hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \end{aligned}$$

RÉGRESSION RIDGE

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2)}{\partial \beta} \\ &= 2X^T X\beta - 2X^T y + 2\lambda\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad (X^T X + \lambda I)\beta = X^T y \\ &\quad \text{ssi} \quad \hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \end{aligned}$$

RÉGRESSION RIDGE

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2)}{\partial \beta} \\ &= 2X^T X\beta - 2X^T y + 2\lambda\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad (X^T X + \lambda I)\beta &= X^T y \\ \hat{\beta} &= (X^T X + \lambda I)^{-1} X^T y \end{aligned}$$

RÉGRESSION RIDGE

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β :

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} &= \frac{\partial (\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2)}{\partial \beta} \\ &= 2X^T X\beta - 2X^T y + 2\lambda\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 &\quad \text{ssi} \quad (X^T X + \lambda I)\beta = X^T y \\ &\quad \text{ssi} \quad \hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \end{aligned}$$

RÉGRESSION RIDGE

- ▶ λ est un *hyperparamètre* à optimiser: tester différentes valeurs de λ jusqu'à obtenir les meilleurs résultats sur le test set.
- ▶ $\lambda = 0$ correspond au cas de la régression linéaire classique.
- ▶ Lorsque $\lambda \rightarrow \infty$, la régularisation force les coefficients β_i à converger vers 0.

RÉGRESSION RIDGE

- ▶ λ est un *hyperparamètre* à optimiser: tester différentes valeurs de λ jusqu'à obtenir les meilleurs résultats sur le test set.
- ▶ $\lambda = 0$ correspond au cas de la régression linéaire classique.
- ▶ Lorsque $\lambda \rightarrow \infty$, la régularisation force les coefficients β_i à converger vers 0.

RÉGRESSION RIDGE

- ▶ λ est un *hyperparamètre* à optimiser: tester différentes valeurs de λ jusqu'à obtenir les meilleurs résultats sur le test set.
- ▶ $\lambda = 0$ correspond au cas de la régression linéaire classique.
- ▶ Lorsque $\lambda \rightarrow \infty$, la régularisation force les coefficients β_i à converger vers 0.

RÉGRESSION RIDGE

- Lorsque λ augmente, les coefficients diminuent.

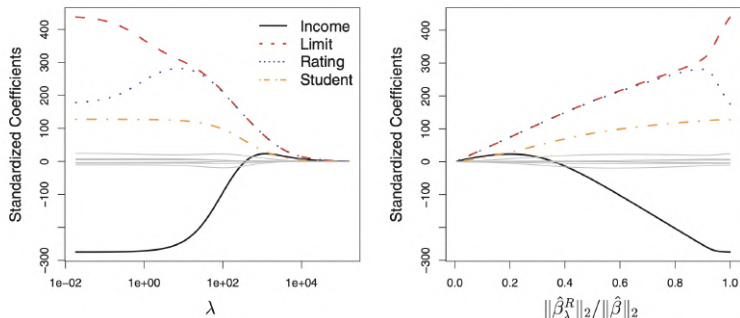


FIGURE 6.4. The standardized ridge regression coefficients are displayed for the **Credit** data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2 / \|\hat{\beta}\|_2$.

Figures taken from [James et al., 2013]

RÉGRESSION RIDGE

- ▶ La régression Ridge joue sur le bias-variance trade-off: lorsque λ augmente, la variance du modèle diminue, mais son bias augmente.
- ▶ Rappel: pour un modèle $\hat{f}(x)$, on a:

$$\begin{aligned}\text{Biais}[\hat{f}(x)] &= \mathbb{E}[\hat{f}(x) - f(x)] \\ \text{Var}[\hat{f}(x)] &= \mathbb{E}\left[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\right]\end{aligned}$$

RÉGRESSION RIDGE

- ▶ La régression Ridge joue sur le bias-variance trade-off: lorsque λ augmente, la variance du modèle diminue, mais son bias augmente.
- ▶ Rappel: pour un modèle $\hat{f}(\mathbf{x})$, on a:

$$\begin{aligned}\text{Biais}[\hat{f}(\mathbf{x})] &= \mathbb{E}[\hat{f}(\mathbf{x}) - f(\mathbf{x})] \\ \text{Var}[\hat{f}(\mathbf{x})] &= \mathbb{E}\left[(\hat{f}(\mathbf{x}) - \mathbb{E}[\hat{f}(\mathbf{x})])^2\right]\end{aligned}$$

RÉGRESSION RIDGE

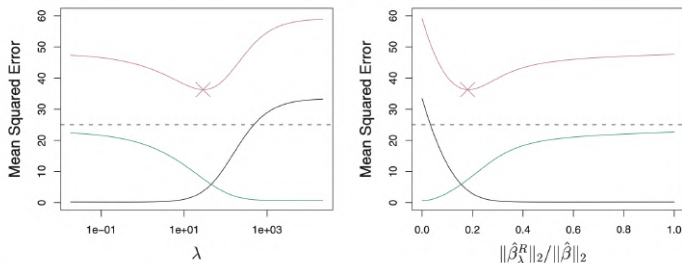


FIGURE 6.5. Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2^2 / \|\hat{\beta}\|_2^2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.

Figures taken from [James et al., 2013]

RÉGRESSION LASSO

- ▶ La régression Ridge réduit l'impact des prédicteurs X_i les moins significatifs, en leur assignant des paramètres β_i qui sont petits (shrinkage method).
- ▶ Mais elle n'élimine pas ces prédicteurs.
- ▶ La régression **LASSO** permet d'éliminer complètement les prédicteurs les moins significatifs.
- ▶ Ainsi, LASSO réalise une sélection des variables les plus pertinentes (feature selection).

RÉGRESSION LASSO

- ▶ La régression Ridge réduit l'impact des prédicteurs X_i les moins significatifs, en leur assignant des paramètres β_i qui sont petits (shrinkage method).
- ▶ Mais elle n'élimine pas ces prédicteurs.
- ▶ La régression **LASSO** permet d'éliminer complètement les prédicteurs les moins significatifs.
- ▶ Ainsi, LASSO réalise une **sélection des variables** les plus pertinentes (feature selection).

RÉGRESSION LASSO

- ▶ La régression Ridge réduit l'impact des prédicteurs X_i les moins significatifs, en leur assignant des paramètres β_i qui sont petits (shrinkage method).
- ▶ Mais elle n'élimine pas ces prédicteurs.
- ▶ La **régression LASSO** permet d'éliminer complètement les prédicteurs les moins significatifs.
- ▶ Ainsi, LASSO réalise une **sélection des variables** les plus pertinentes (feature selection).

RÉGRESSION LASSO

- ▶ La régression Ridge réduit l'impact des prédicteurs X_i les moins significatifs, en leur assignant des paramètres β_i qui sont petits (shrinkage method).
- ▶ Mais elle n'élimine pas ces prédicteurs.
- ▶ La **régression LASSO** permet d'éliminer complètement les prédicteurs les moins significatifs.
- ▶ Ainsi, LASSO réalise une **sélection des variables** les plus pertinentes (feature selection).

RÉGRESSION LASSO

- ▶ On suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Le but est d'obtenir des estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres $\beta = (\beta_0, \dots, \beta_p)$.
- ▶ La **régression LASSO** permet d'éliminer les estimateurs X_i qui sont le moins significativement associés avec la réponse Y (feature selection).

RÉGRESSION LASSO

- ▶ On suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Le but est d'obtenir des estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres $\beta = (\beta_0, \dots, \beta_p)$.
- ▶ La régression LASSO permet d'éliminer les estimateurs X_i qui sont le moins significativement associés avec la réponse Y (feature selection).

RÉGRESSION LASSO

- ▶ On suppose que la vraie relation entre les variables explicatives et la réponse est de la forme suivante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.

- ▶ Le but est d'obtenir des estimateurs $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres $\beta = (\beta_0, \dots, \beta_p)$.
- ▶ La **régression LASSO** permet d'éliminer les estimateurs X_i qui sont le moins significativement associés avec la réponse Y (feature selection).

RÉGRESSION LASSO

- **Régression LASSO**: on choisit les estimateurs $\hat{\beta}$ qui minimisent une autre version *régularisée* la **residual sum of squares (RSS)**

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (x_i^T \beta - y_i)^2 + \lambda \sum_{j=0}^p |\beta_j| \\ &= \|X\beta - y\|_2^2 + \lambda \|\beta\|_1\end{aligned}$$

► où λ est le *paramètre de régularisation*.

► λ est choisi par la *croûte* *cross-validation*.

RÉGRESSION LASSO

- **Régression LASSO**: on choisit les estimateurs $\hat{\beta}$ qui minimisent une autre version *régularisée* la **residual sum of squares (RSS)**

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (x_i^T \beta - y_i)^2 + \lambda \sum_{j=0}^p |\beta_j| \\ &= \|X\beta - y\|_2^2 + \lambda \|\beta\|_1\end{aligned}$$

► où λ est le *paramètre de régularisation*.

► Le terme $\lambda \|\beta\|_1$ est une *pénalité l_1* (l_1 penalty).

RÉGRESSION LASSO

- ▶ **Régression LASSO**: on choisit les estimateurs $\hat{\beta}$ qui minimisent une autre version *régularisée* la **residual sum of squares (RSS)**

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (\mathbf{x}_i^T \beta - y_i)^2 + \lambda \sum_{j=0}^p |\beta_j| \\ &= \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_1\end{aligned}$$

- ▶ où λ est le *paramètre de régularisation*.
- ▶ Le terme $\lambda \|\beta\|_1$ est une *pénalité l_1* (l_1 penalty).

RÉGRESSION LASSO

- ▶ **Régression LASSO**: on choisit les estimateurs $\hat{\beta}$ qui minimisent une autre version *régularisée* la **residual sum of squares (RSS)**

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (\mathbf{x}_i^T \beta - y_i)^2 + \lambda \sum_{j=0}^p |\beta_j| \\ &= \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_1\end{aligned}$$

- ▶ où λ est le *paramètre de régularisation*.
- ▶ Le terme $\lambda \|\beta\|_1$ est une *pénalité* l_1 (l_1 penalty).

RÉGRESSION LASSO

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_1 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on utilise des méthodes d'optimisation (pas de solution exacte, sauf dans certains cas spécifiques).

RÉGRESSION LASSO

- On a:

$$\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \left(\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_1 \right)$$

- Pour trouver le minimum de $\text{RSS}(\beta)$, on utilise des méthodes d'optimisation (pas de solution exacte, sauf dans certains cas spécifiques).

LASSO

- Lorsque λ augmente, certains coefficients deviennent nuls.

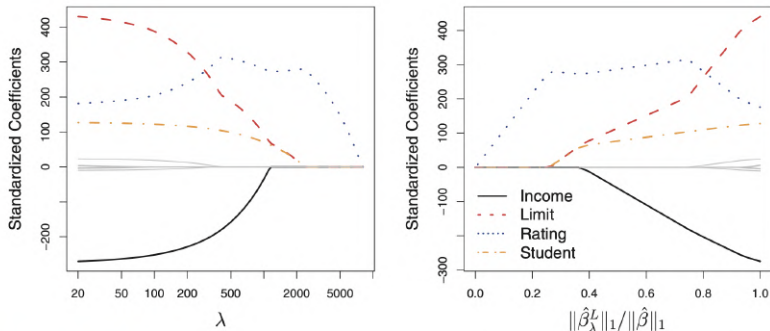


FIGURE 6.6. The standardized lasso coefficients on the **Credit** data set are shown as a function of λ and $\|\hat{\beta}_\lambda^L\|_1 / \|\hat{\beta}\|_1$.

Figures taken from [James et al., 2013]

LASSO

- ▶ Le fait que la Ridge régression diminue les coefficient alors que la LASSO les annulent est parfaitement explicable.
- ▶ Il existe une reformulation de ces méthodes en terme de problème d'optimisation sous contrainte et une interprétation parlante qui en découle...

LASSO

- ▶ Le fait que la Ridge régression diminue les coefficient alors que la LASSO les annulent est parfaitement explicable.
- ▶ Il existe une reformulation de ces méthodes en terme de problème d'optimisation sous contrainte et une interprétation parlante qui en découle...

ELASTIC-NET

- ▶ En combinant les méthodes Ridge et LASSO, on obtient une régression appelée **Elastic-Net**.
- ▶ Dans ce cas, on choisit les estimateurs $\hat{\beta}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

$$\text{RSS}(\beta) = \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

- ▶ On a donc:

$$\hat{\beta} = \arg \min_{\beta} (\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2)$$

- ▶ où λ_1 et λ_2 sont des *paramètres de régularisation*.
- ▶ On a donc introduit une *pénalité l_1* et une *pénalité l_2* .

ELASTIC-NET

- ▶ En combinant les méthodes Ridge et LASSO, on obtient une régression appelée **Elastic-Net**.
- ▶ Dans ce cas, on choisit les estimateurs $\hat{\beta}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

$$\text{RSS}(\beta) = \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

- ▶ On a donc:

$$\hat{\beta} = \arg \min_{\beta} (\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2)$$

- ▶ où λ_1 et λ_2 sont des *paramètres de régularisation*.
- ▶ On a donc introduit une *pénalité l_1* et une *pénalité l_2* .

ELASTIC-NET

- ▶ En combinant les méthodes Ridge et LASSO, on obtient une régression appelée **Elastic-Net**.
- ▶ Dans ce cas, on choisit les estimateurs $\hat{\beta}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

$$\text{RSS}(\beta) = \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

- ▶ On a donc:

$$\hat{\beta} = \arg \min_{\beta} (\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2)$$

- ▶ où λ_1 et λ_2 sont des *paramètres de régularisation*.
- ▶ On a donc introduit une *pénalité l_1* et une *pénalité l_2* .

ELASTIC-NET

- ▶ En combinant les méthodes Ridge et LASSO, on obtient une régression appelée **Elastic-Net**.
- ▶ Dans ce cas, on choisit les estimateurs $\hat{\beta}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

$$\text{RSS}(\beta) = \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

- ▶ On a donc:

$$\hat{\beta} = \arg \min_{\beta} (\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2)$$

- ▶ où λ_1 et λ_2 sont des *paramètres de régularisation*.
- ▶ On a donc introduit une *pénalité l_1* et une *pénalité l_2* .

ELASTIC-NET

- ▶ En combinant les méthodes Ridge et LASSO, on obtient une régression appelée **Elastic-Net**.
- ▶ Dans ce cas, on choisit les estimateurs $\hat{\beta}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

$$\text{RSS}(\beta) = \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

- ▶ On a donc:

$$\hat{\beta} = \arg \min_{\beta} (\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2)$$

- ▶ où λ_1 et λ_2 sont des *paramètres de régularisation*.
- ▶ On a donc introduit une *pénalité l_1* et une *pénalité l_2* .

BIBLIOGRAPHIE



James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
An Introduction to Statistical Learning: with Applications in R, volume 103 of
Springer Texts in Statistics.
Springer, New York.