

# RÉGRESSIONS LINÉAIRES

Jérémy 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

- ▶ 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:

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

où  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ **Interprétation:** chaque  $\beta_i$  ( $i = 1, \dots, p$ ) 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

- ▶ 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:

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

où  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ **Interprétation:** chaque  $\beta_i$  ( $i = 1, \dots, p$ ) 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

- ▶ 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:

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

où  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ **Interprétation:** chaque  $\beta_i$  ( $i = 1, \dots, p$ ) 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

- ▶ Les “vrais” paramètres  $\beta_0, \dots, \beta_p$  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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p = \mathbf{x}^T \hat{\boldsymbol{\beta}} \quad (2)$$

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

- ▶ Les “vrais” paramètres  $\beta_0, \dots, \beta_p$  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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p = \mathbf{x}^T \hat{\boldsymbol{\beta}} \quad (2)$$

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

- ▶ Les “vrais” paramètres  $\beta_0, \dots, \beta_p$  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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p = \mathbf{x}^T \hat{\boldsymbol{\beta}} \quad (2)$$

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

- 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

- 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

- ▶ 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:

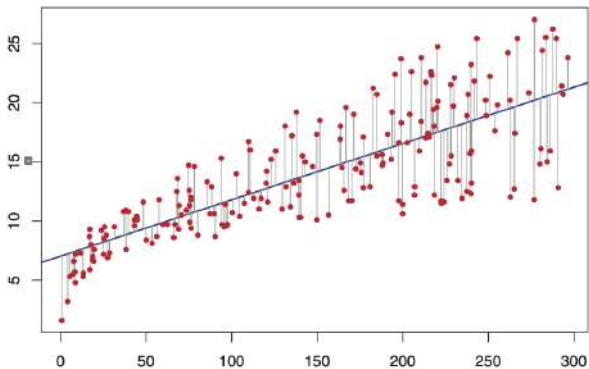
$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (x_i^T \beta - y_i)^2 \\ &= \|\mathbf{X}\beta - \mathbf{y}\|^2\end{aligned}$$

# RÉGRESSION LINÉAIRE

- ▶ 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:

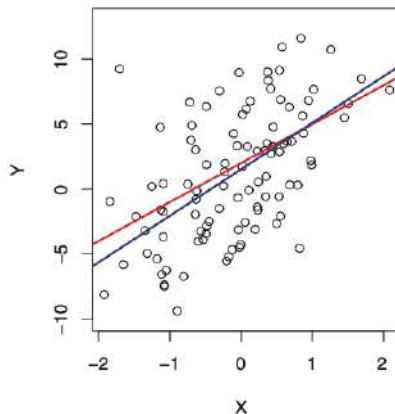
$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (\mathbf{x}_i^T \beta - y_i)^2 \\ &= \|\mathbf{X}\beta - \mathbf{y}\|^2\end{aligned}$$

# RÉGRESSION LINÉAIRE



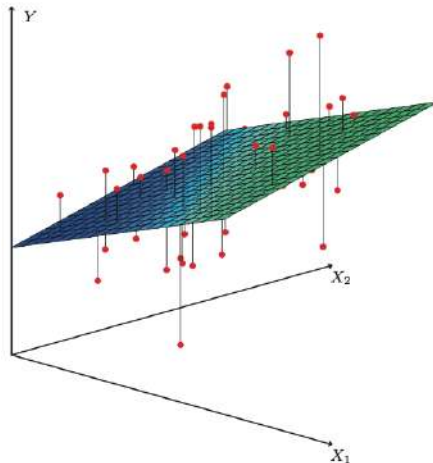
Figures taken from [James et al., 2013]

# RÉGRESSION LINÉAIRE



Figures taken from [James et al., 2013]

# RÉGRESSION LINÉAIRE



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  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$  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!):

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

## 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

- ▶ 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  $\beta$ :

# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# RÉGRESSION LINÉAIRE

- 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  $\beta$ :

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

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = 0 \quad \text{ssi} \quad \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\text{ssi} \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



# RÉGRESSION LINÉAIRE

- Soit un **test set** formé de  $N'$  observations:

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

- Une fois les estimateurs  $\hat{\beta}$  obtenus, les prédictions  $\hat{y}$  associés aux data  $X$  sont données par:

$$\hat{y}_i = \mathbf{x}_i^T \hat{\beta} \quad \text{pour } i = 1, \dots, N', \text{ i.e.,}$$

$$\hat{\mathbf{y}} = X \hat{\beta}$$

# RÉGRESSION LINÉAIRE

- Soit un **test set** formé de  $N'$  observations:

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

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

$$\begin{aligned}\hat{y}_i &= \mathbf{x}_i^T \hat{\beta} \text{ pour } i = 1, \dots, N', \text{ i.e.,} \\ \hat{\mathbf{y}} &= \mathbf{X} \hat{\beta}\end{aligned}$$

# RÉGRESSION LINÉAIRE

- Soit un **test set** formé de  $N'$  observations:

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

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

$$\begin{aligned}\hat{y}_i &= \mathbf{x}_i^T \hat{\beta} \text{ pour } i = 1, \dots, N', \text{ i.e.,} \\ \hat{\mathbf{y}} &= \mathbf{X} \hat{\beta}\end{aligned}$$

# RÉGRESSION LINÉAIRE

- Soit un **test set** formé de  $N'$  observations:

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

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

$$\begin{aligned}\hat{y}_i &= \mathbf{x}_i^T \hat{\beta} \text{ pour } i = 1, \dots, N', \text{ i.e.,} \\ \hat{\mathbf{y}} &= \mathbf{X} \hat{\beta}\end{aligned}$$

# 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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)
  - Lasso (penalized maximum likelihood)

- ▶ 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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 bénéfique de réduire la variance du modèle.
- ▶ En gros, les variables  $X_i$  les moins significatives voient leur estimateurs associés  $\hat{\beta}_i$  converger 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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 bénéfique de réduire la variance du modèle.
- ▶ En gros, les variables  $X_i$  les moins significatives voient leur estimateurs associés  $\hat{\beta}_i$  converger 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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 bénéfique de réduire la variance du modèle.
- ▶ En gros, les variables  $X_i$  les moins significatives voient leur estimateurs associés  $\hat{\beta}_i$  converger 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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 bénéfique de réduire la variance du modèle.
- ▶ En gros, les variables  $X_i$  les moins significatives voient leur estimateurs associés  $\hat{\beta}_i$  converger 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)**

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

• où  $\lambda$  est le *paramètre de régularisation*.

• Le terme  $\lambda \|\beta\|_2^2$  est une *penalité* (à la *penalty*).

# RÉGRESSION RIDGE

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

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

✱ où  $\lambda$  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)**

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

- ▶ où  $\lambda$  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)**

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

- ▶ où  $\lambda$  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  $\beta$ :

# 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  $\beta$ :

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

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

# 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  $\beta$ :

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

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

# RÉGRESSION RIDGE

- On a:

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

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

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

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

# 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  $\beta$ :

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

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

# 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  $\beta$ :

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

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

# 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  $\beta$ :

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

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



# RÉGRESSION RIDGE

- ▶  $\lambda$  est un *hyperparamètre* à optimiser: tester différentes valeurs de  $\lambda$  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  $\beta_i$  à converger vers 0.

# RÉGRESSION RIDGE

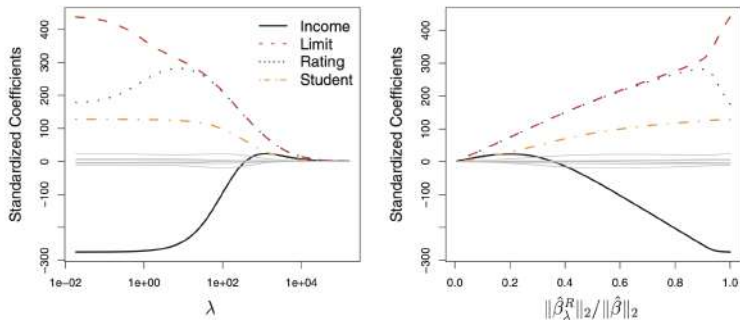
- ▶  $\lambda$  est un *hyperparamètre* à optimiser: tester différentes valeurs de  $\lambda$  jusqu'à obtenir les meilleurs résultats sur le test set.
- ▶  $\lambda = 0$  correspond au cas de la régression linéaire classique.
- ▶ Lorsque  $\lambda \rightarrow \infty$ , la régularisation force les coefficients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

# RÉGRESSION RIDGE

- ▶  $\lambda$  est un *hyperparamètre* à optimiser: tester différentes valeurs de  $\lambda$  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  $\beta_i$  à converger vers 0.

# RÉGRESSION RIDGE

- Lorsque  $\lambda$  augmente, les coefficients diminuent.



**FIGURE 6.4.** The standardized ridge regression coefficients are displayed for the **Credit** data set, as a function of  $\lambda$  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  $\lambda$  augmente, la variance du modèle diminue, mais son bias augmente.
- ▶ Rappel: pour un modèle  $\hat{f}(x)$ , on a:

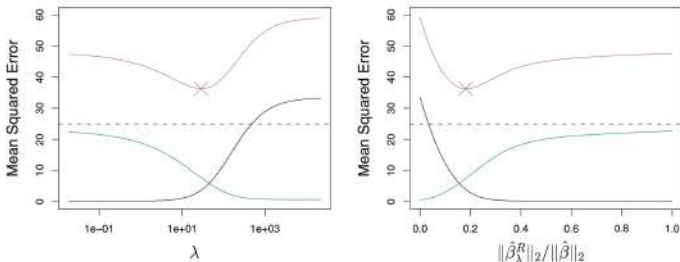
$$\begin{aligned}\text{Biais}[\hat{f}(x)] &= \mathbb{E}[\hat{f}(x) - f(x)] \\ \text{Var}[\hat{f}(x)] &= \mathbb{E}\left[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\right]\end{aligned}$$

# RÉGRESSION RIDGE

- ▶ La régression Ridge joue sur le bias-variance trade-off: lorsque  $\lambda$  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  $\lambda$  and  $\|\hat{\beta}_\lambda^R\|_2 / \|\hat{\beta}\|_2$ . 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]

# LASSO

- ▶ La régression Ridge réduit l'impact des prédicteurs  $X_i$  les moins significatifs, en leur assignant des paramètres  $\beta_i$  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).



# LASSO

- ▶ La régression Ridge réduit l'impact des prédicteurs  $X_i$  les moins significatifs, en leur assignant des paramètres  $\beta_i$  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).

# LASSO

- ▶ La régression Ridge réduit l'impact des prédicteurs  $X_i$  les moins significatifs, en leur assignant des paramètres  $\beta_i$  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).

# LASSO

- ▶ La régression Ridge réduit l'impact des prédicteurs  $X_i$  les moins significatifs, en leur assignant des paramètres  $\beta_i$  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).

# 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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).

# 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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).

# 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  $\epsilon$  est un bruit tel que  $E(\epsilon) = 0$ .

- ▶ 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).

# LASSO

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

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

► où  $\lambda$  est le *paramètre de régularisation*.

- $\lambda \geq 0$  est une variable à optimiser

# LASSO

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

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

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



# LASSO

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

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

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

# LASSO

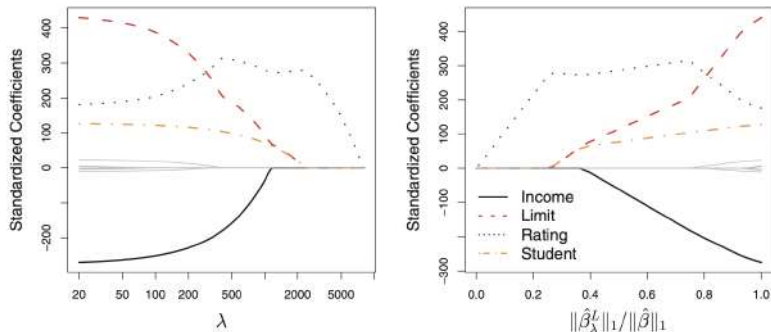
- **Régression LASSO:** on choisit les estimateurs  $\hat{\beta}$  qui minimisent une autre version *régularisée* la **residual sum of squares (RSS)**

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

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

# LASSO

- ▶ Lorsque  $\lambda$  augmente, certains coefficients deviennent nuls.



**FIGURE 6.6.** The standardized lasso coefficients on the **Credit** data set are shown as a function of  $\lambda$  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  $\lambda_1$  et  $\lambda_2$  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  $\lambda_1$  et  $\lambda_2$  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  $\lambda_1$  et  $\lambda_2$  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  $\lambda_1$  et  $\lambda_2$  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  $\lambda_1$  et  $\lambda_2$  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.