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

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

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

  où ϵ est un bruit tel que $\mathrm{E}(\epsilon) = 0$.
- **Interprétation:** chaque β_i $(i = 1, \ldots, p)$ 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 $X_1, \ldots, X_p$ des variables explicatives et Y une variable réponse.
- **Hypothèse forte:** on suppose que la relation entre $X_1, \ldots, X_p$ et Y est de la forme linéaire suivante:

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

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

- **Interprétation:** chaque β_i $(i = 1, \ldots, p)$ 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 $X_1, \ldots, X_p$ des variables explicatives et Y une variable réponse.
- **Hypothèse forte:** on suppose que la relation entre $X_1, \ldots, X_p$ et Y est de la forme linéaire suivante:

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

  où ϵ est un bruit tel que $\mathrm{E}(\epsilon) = 0$.
- **Interprétation:** chaque β_i $(i = 1, \ldots, p)$ 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 $\beta_0, \ldots, \beta_p$ sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \ldots, \hat{\beta}_p$ des ces paramètres.
- Une fois les estimateurs obtenus, la **prédiction** associée à toute observation $\boldsymbol{x} = (x_1, \ldots, x_p)$ est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \boldsymbol{x}^T \hat{\boldsymbol{\beta}} \tag{2}$$

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

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

Régression linéaire

- Les "vrais" paramètres $\beta_0, \ldots, \beta_p$ sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \ldots, \hat{\beta}_p$ des ces paramètres.
- Une fois les estimateurs obtenus, la **prédiction** associée à toute observation $\boldsymbol{x} = (x_1, \ldots, x_p)$ est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \boldsymbol{x}^T \hat{\boldsymbol{\beta}} \tag{2}$$

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

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

Régression linéaire

- Les "vrais" paramètres $\beta_0, \ldots, \beta_p$ sont inconnus. On aimerait donc obtenir des estimateurs $\hat{\beta}_0, \ldots, \hat{\beta}_p$ des ces paramètres.
- Une fois les estimateurs obtenus, la **prédiction** associée à toute observation $\boldsymbol{x} = (x_1, \ldots, x_p)$ est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \boldsymbol{x}^T \hat{\boldsymbol{\beta}} \tag{2}$$

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

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

Régression linéaire

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

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

- On définit les matrice et vecteur:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_N} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} \quad \text{et} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

Régression linéaire

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

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

- On définit les matrice et vecteur:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_N} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} \quad \text{et} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

Régression linéaire

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

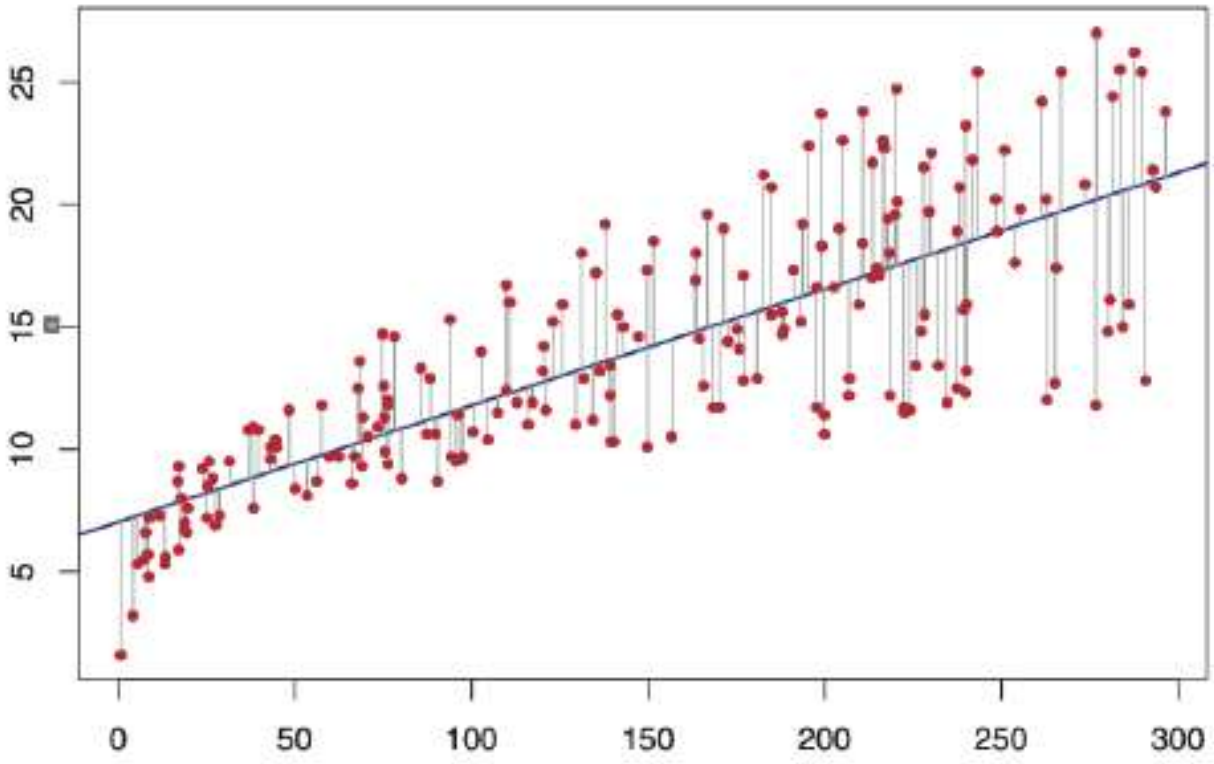
$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left(\boldsymbol{x_i}^T \boldsymbol{\beta} - y_i\right)^2 \\ &= \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 \end{aligned}$$

Régression linéaire

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

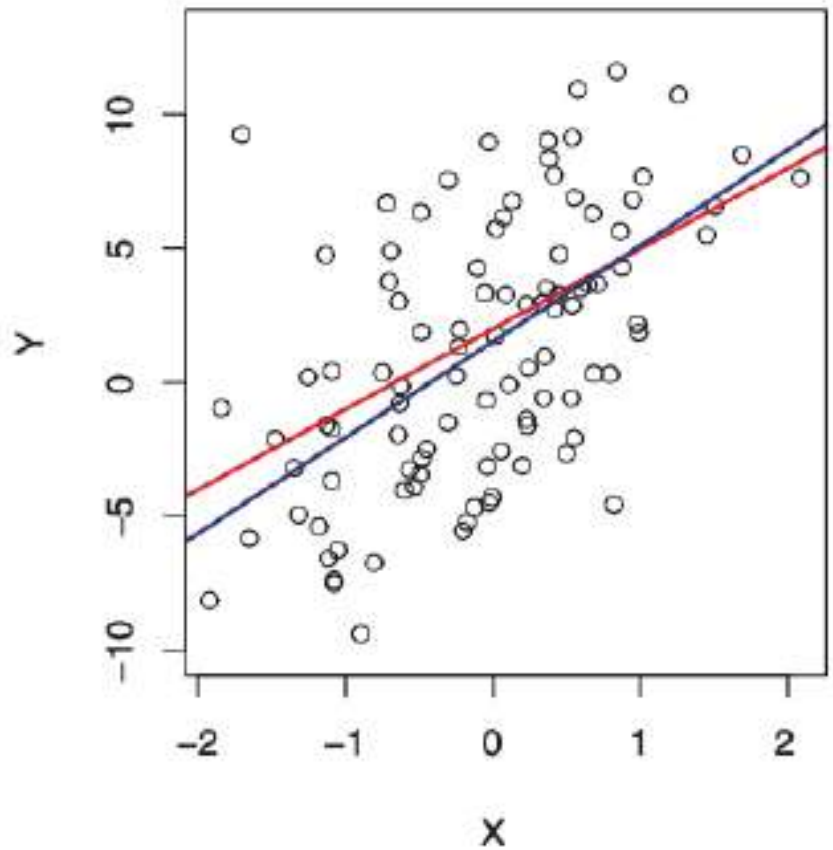
$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left(\boldsymbol{x_i}^T \boldsymbol{\beta} - y_i\right)^2 \\ &= \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 \end{aligned}$$

Régression linéaire



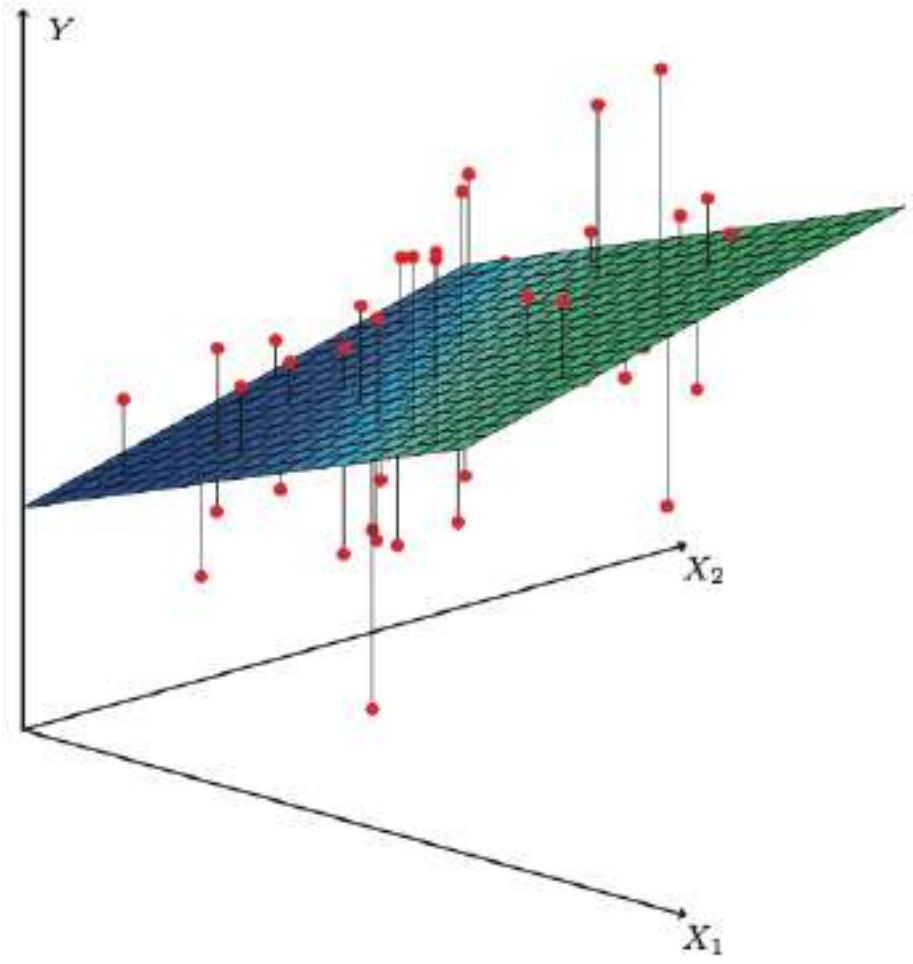
Figures taken from [James et al., 2013]

Régression linéaire



Figures taken from [James et al., 2013]

Régression linéaire



Figures taken from [James et al., 2013]

Régression linéaire

- On a alors:

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

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

$$\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial (\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2)}{\partial \boldsymbol{\beta}} = \frac{\partial (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})}{\partial \boldsymbol{\beta}}$$

$$= \frac{\partial (\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T \boldsymbol{y})}{\partial \boldsymbol{\beta}}$$

$$= 2\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T \boldsymbol{y}$$

$$\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \iff \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{y}$$

$$\iff \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

Régression linéaire

- On a alors:

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

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

$$\begin{aligned}
\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} \\
&= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

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

Régression linéaire

- On a alors:

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

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

$$\begin{aligned}
\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} \\
&= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

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

Régression linéaire

- On a alors:

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}}$$

$$= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}}$$

$$= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}$$

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

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

Régression linéaire

- On a alors:

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}}$$

$$= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}}$$

$$= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}$$

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

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

Régression linéaire

- On a alors:

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}}$$

$$= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}}$$

$$= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}$$

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

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

Régression linéaire

- On a alors:

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

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

$$\begin{aligned}\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}} \\ &= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} \\ &= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

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

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

Régression linéaire

- On a alors:

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

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

$$\begin{aligned}
\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} \\
&= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

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

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

Régression linéaire

- On a alors:

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

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

$$\begin{aligned}
\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} \\
&= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

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

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

Régression linéaire

- On a alors:

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

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

$$\begin{aligned}
\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \left((\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right)}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial \left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} \\
&= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

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

Régression linéaire

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

$$S_{\text{test}} = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_{N'}}, y_{N'})\}.$$

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

$$\hat{y}_i = \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}} \text{ pour } i = 1, \ldots, N', \text{ i.e.,}$$
$$\hat{y} = X\hat{\beta}$$

Régression linéaire

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

$$S_{\text{test}} = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_{N'}}, y_{N'})\}.$$

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

$$\hat{y}_i = \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}} \text{ pour } i = 1, \ldots, N', \text{ i.e.,}$$
$$\hat{y} = \boldsymbol{X}\hat{\beta}$$

Régression linéaire

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

$$S_{\text{test}} = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_{N'}}, y_{N'})\}.$$

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

$$\hat{y}_i = \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}} \text{ pour } i = 1, \ldots, N', \text{ i.e.,}$$
$$\hat{y} = \boldsymbol{X}\hat{\beta}$$

Régression linéaire

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

$$S_{\text{test}} = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_{N'}}, y_{N'})\}.$$

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

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

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
  - 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...

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 $\mathrm{E}(\epsilon) = 0$.

- 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 $\mathrm{E}(\epsilon) = 0$.

- 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 $\mathrm{E}(\epsilon) = 0$.

- 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 $\mathrm{E}(\epsilon) = 0$.

- 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 (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**.

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 $\mathrm{E}(\epsilon) = 0$.

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ des paramètres $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.
- 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 ϵ est un bruit tel que $\mathrm{E}(\epsilon) = 0$.

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ des paramètres $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.
- 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 ϵ est un bruit tel que $\mathrm{E}(\epsilon) = 0$.

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ des paramètres $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.
- 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 ϵ est un bruit tel que $\mathrm{E}(\epsilon) = 0$.

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ des paramètres $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.
- 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 $\hat{\boldsymbol{\beta}}$ qui minimisent une version *régularisée* la **residual sum of squares (RSS)**

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

- où λ est le *paramètre de régularisation*.
- le terme $\lambda \|\boldsymbol{\beta}\|_2^2$ est une *pénalité* (L2 penalty)

Régression Ridge

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

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

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

Régression Ridge

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

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

- 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 $\hat{\boldsymbol{\beta}}$ qui minimisent une version *régularisée* la **residual sum of squares (RSS)**

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

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

Régression Ridge

- On a:

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial (\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2)}{\partial \boldsymbol{\beta}}$$

$$= 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) + 2\lambda\boldsymbol{\beta}$$

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \Leftrightarrow (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\Leftrightarrow \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Régression Ridge

- On a:

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right)}{\partial \boldsymbol{\beta}}$$

$$= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y} + 2\lambda\boldsymbol{\beta}$$

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \quad \text{ssi} \quad (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\text{ssi} \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Régression Ridge

- On a:

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

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \quad \text{ssi} \quad (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\text{ssi} \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Régression Ridge

- On a:

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \left(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right)}{\partial \boldsymbol{\beta}}$$

$$= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y} + 2\lambda\boldsymbol{\beta}$$

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \quad \text{ssi} \quad (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\text{ssi} \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Régression Ridge

- On a:

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

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \quad \text{ssi} \quad (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\text{ssi} \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Régression Ridge

- On a:

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

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

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

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \quad \text{ssi} \quad (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\text{ssi} \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Régression Ridge

- On a:

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

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

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

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

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 $\lambda \to \infty$, 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 $\lambda \to \infty$, 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 $\lambda \to \infty$, la régularisation force les coefficients β_i à converger vers 0.

Régression Ridge

- Lorsque λ augmente, les coefficients diminuent.

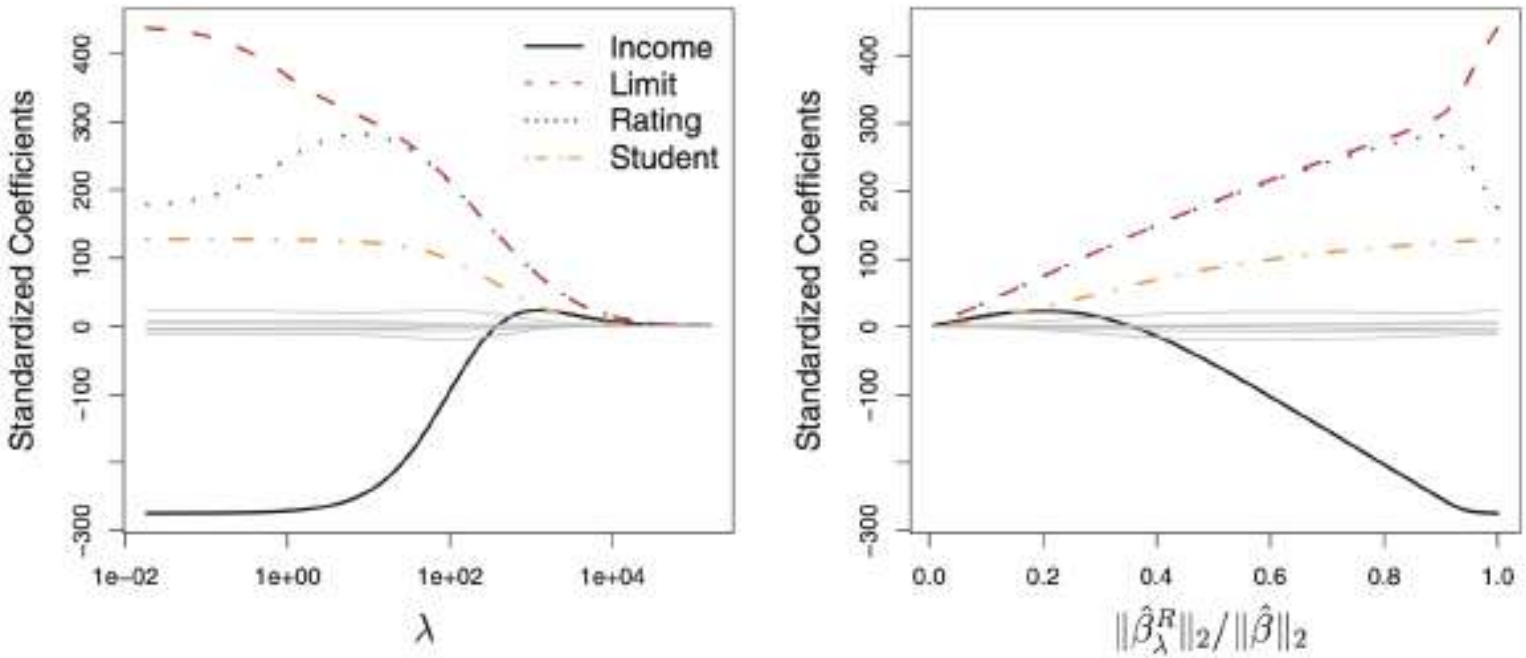


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

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 $\hat{f}(\boldsymbol{x})$, on a:

$$\begin{aligned} \text{Biais}\big[\hat{f}(\boldsymbol{x})\big] &= \mathrm{E}\big[\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\big] \\ \text{Var}\big[\hat{f}(\boldsymbol{x})\big] &= \mathrm{E}\Big[\big(\hat{f}(\boldsymbol{x}) - \mathrm{E}[\hat{f}(\boldsymbol{x})]\big)^2\Big] \end{aligned}$$

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 $\hat{f}(\boldsymbol{x})$, on a:

$$\begin{aligned} \text{Biais}\big[\hat{f}(\boldsymbol{x})\big] &= \text{E}\big[\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\big] \\ \text{Var}\big[\hat{f}(\boldsymbol{x})\big] &= \text{E}\Big[\big(\hat{f}(\boldsymbol{x}) - \text{E}[\hat{f}(\boldsymbol{x})]\big)^2\Big] \end{aligned}$$

Régression Ridge

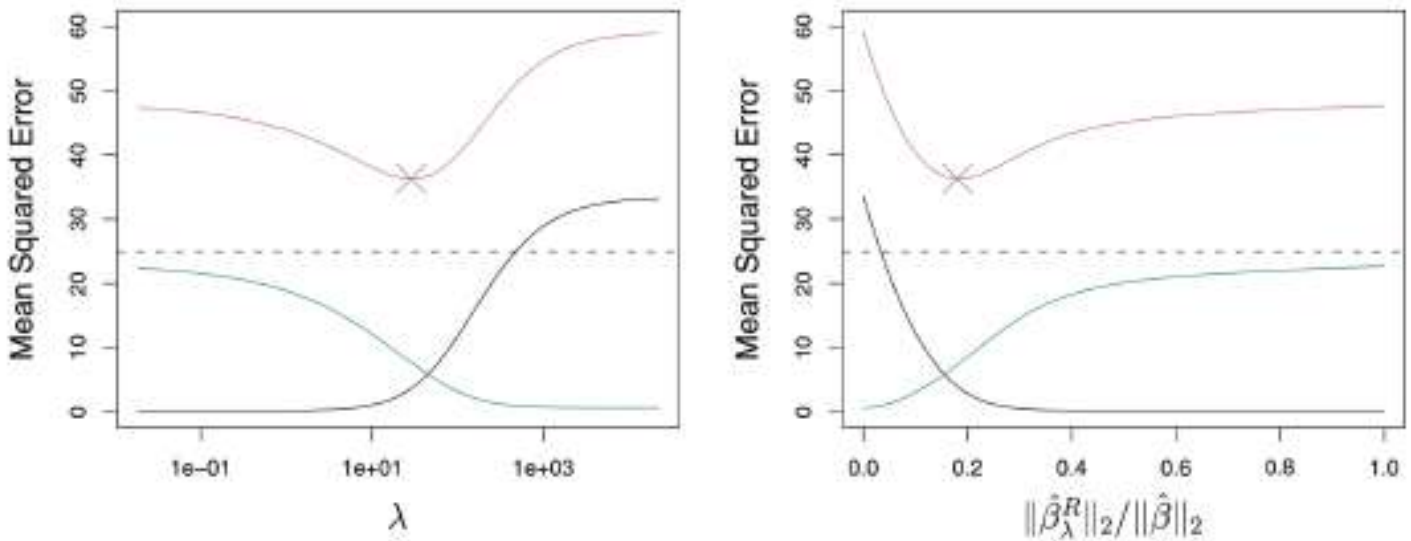


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple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

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).

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).

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).

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 $\mathrm{E}(\epsilon) = 0$.

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ des paramètres $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.
- 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 ϵ est un bruit tel que $\mathrm{E}(\epsilon) = 0$.

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ des paramètres $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.
- 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 ϵ est un bruit tel que $\mathrm{E}(\epsilon) = 0$.

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ des paramètres $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.
- 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 $\hat{\boldsymbol{\beta}}$ qui minimisent une autre version *régularisée* la **residual sum of squares (RSS)**

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

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

LASSO

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

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

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

LASSO

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

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

- où λ 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 $\hat{\boldsymbol{\beta}}$ qui minimisent une autre version *régularisée* la **residual sum of squares (RSS)**

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

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

LASSO

- Lorsque λ augmente, certains coefficients deviennent nuls.

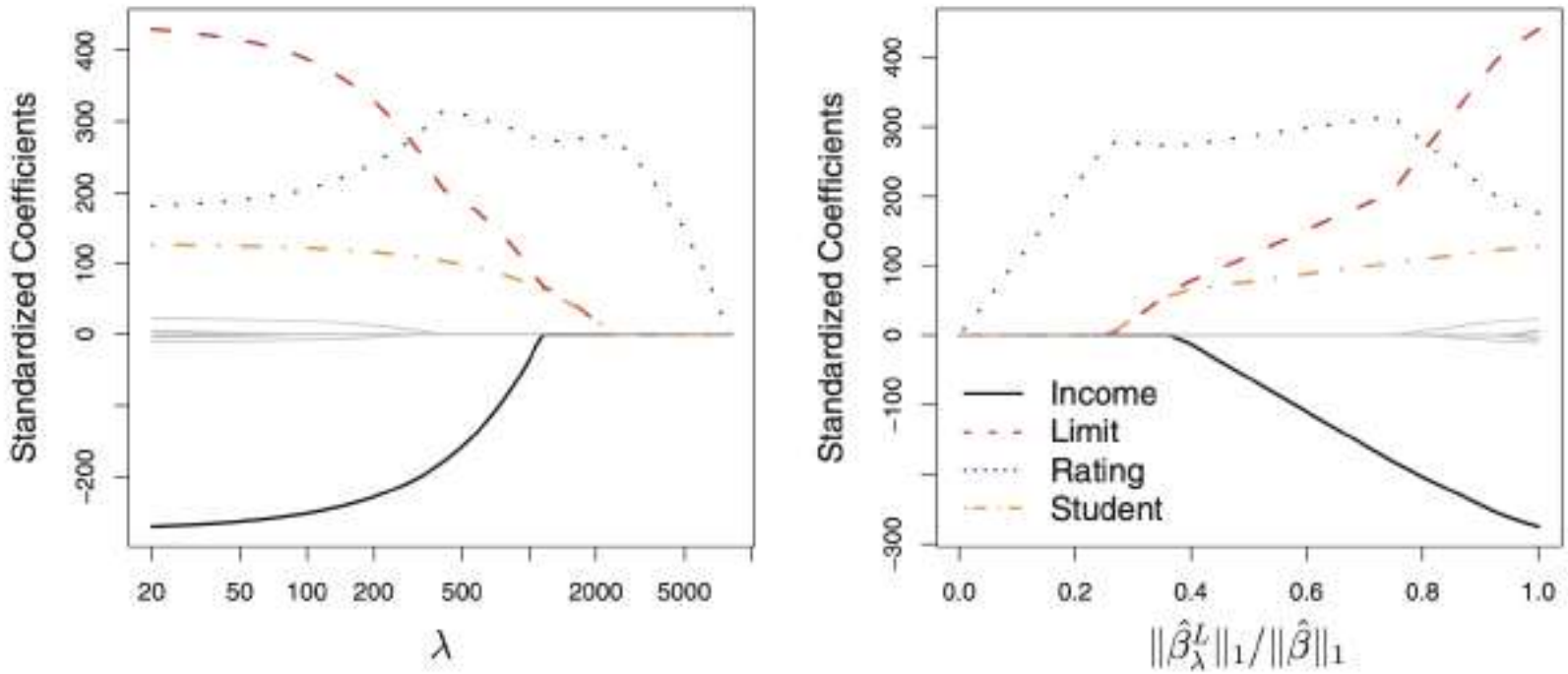


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$.

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)**

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

- On a donc:

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

- 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 $\hat{\boldsymbol{\beta}}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

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

- On a donc:

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

- 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 $\hat{\boldsymbol{\beta}}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

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

- On a donc:

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

- 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 $\hat{\boldsymbol{\beta}}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

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

- On a donc:

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

- 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 $\hat{\boldsymbol{\beta}}$ qui minimisent la version régularisée suivante de la **residual sum of squares (RSS)**

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

- On a donc:

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

- 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.