Régressions linéaires

Jérémie Cabessa

Laboratoire DAVID, UVSQ

Modèles linéaires

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

Modèles linéaires

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

Régression linéaire

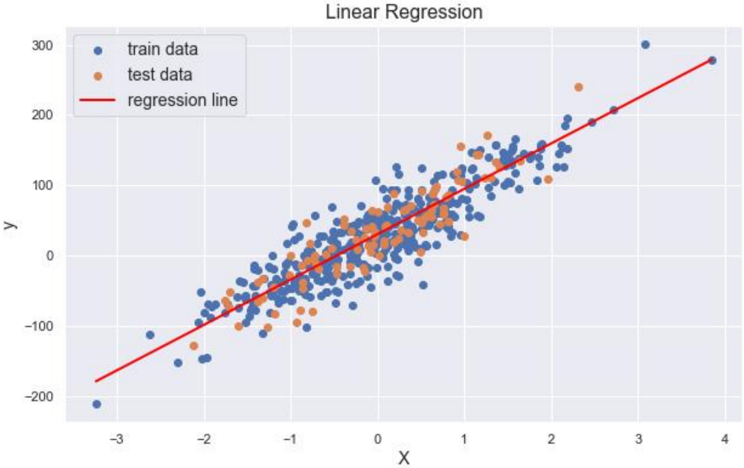


Figure 2 taken from [James et al., 2013]

Régression linéaire

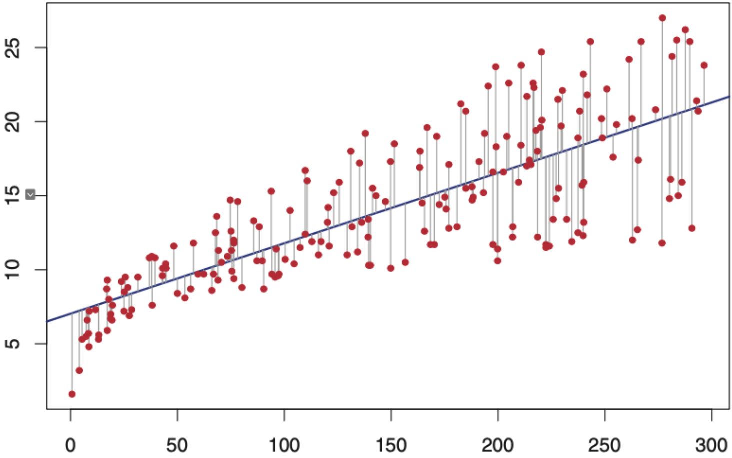


Figure 2 taken from [James et al., 2013]

Régression linéaire simple

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

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

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

Régression linéaire simple

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

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

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

Régression linéaire simple

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

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

Régression linéaire simple

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

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

Régression linéaire simple

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

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

Régression linéaire simple

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

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

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

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

Régression linéaire simple

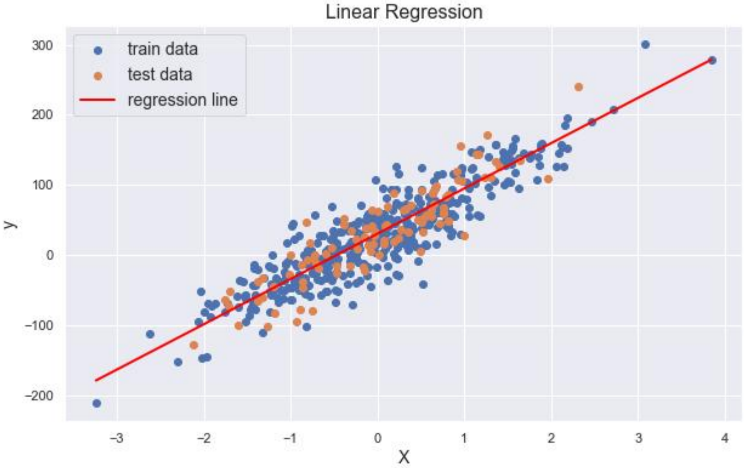
- Soit un **training set** formé de N observations:

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

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

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

Régression linéaire simples



Figures taken from [James et al., 2013]

Régression linéaire simple

- On a alors:

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

Régression linéaire simple

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

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

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

ce qui donne comme solutions (cf. notes):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i (y_i - \bar{y})}{\sum_{i=1}^{N} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

où $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ et $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$

Régression linéaire simple

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

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

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

ce qui donne comme solutions (cf. notes):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i(y_i - \bar{y})}{\sum_{i=1}^{N} x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

où $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ et $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$

Régression linéaire simple

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

$$\begin{aligned}\frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{N} \big((\beta_0 + \beta_1 x_i) - y_i\big)^2 = 0 \\ \frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\partial \beta_1} &= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{N} \big((\beta_0 + \beta_1 x_i) - y_i\big)^2 = 0\end{aligned}$$

ce qui donne comme solutions (cf. notes):

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

$$\text{où} \quad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad \text{et} \quad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Régression linéaire simple

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

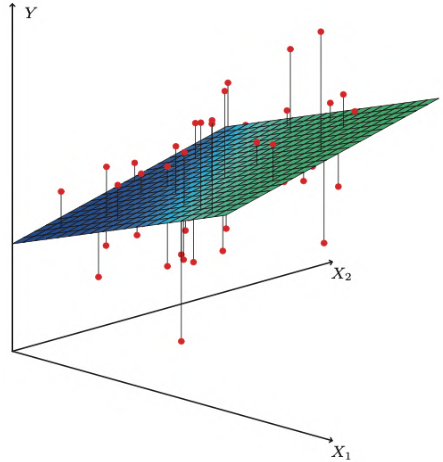
$$\begin{aligned}
\frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{N} \big((\beta_0 + \beta_1 x_i) - y_i\big)^2 = 0 \\
\frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\partial \beta_1} &= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{N} \big((\beta_0 + \beta_1 x_i) - y_i\big)^2 = 0
\end{aligned}$$

ce qui donne comme solutions (cf. notes):

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

$$\text{où} \quad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad \text{et} \quad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Régression linéaire multiple



Figures taken from [James et al., 2013]

Régression linéaire multiple

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

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

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

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

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

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

- **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 multiple

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

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

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

- 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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \boldsymbol{x}^T \hat{\boldsymbol{\beta}} \tag{4}$$

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 multiple

- 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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \boldsymbol{x}^T \hat{\boldsymbol{\beta}} \tag{4}$$

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

Régression linéaire multiple

- 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 $\boldsymbol{x} = (x_1, \dots, x_p)$ est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \boldsymbol{x}^T \hat{\boldsymbol{\beta}} \tag{4}$$

  où $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ et $\boldsymbol{x} = (1, x_1, \dots, x_p)$ (on a rajouté la composante 1).
- Pour obtenir les estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ des paramètres, on dispose d'observations (ou de data).

Régression linéaire multiple

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

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

- On définit les matrice et vecteur:

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

Régression linéaire multiple

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

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

- On définit les matrice et vecteur:

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

Régression linéaire multiple

- 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:

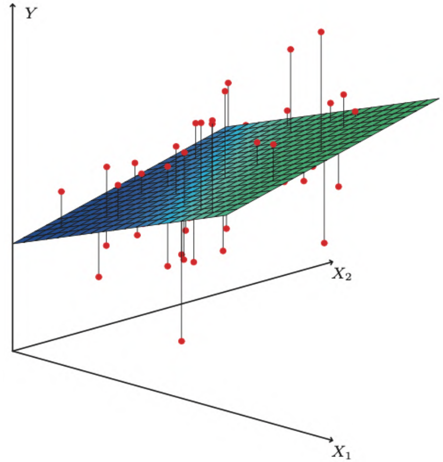
$$
\begin{aligned}
\mathrm{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \\
&= \sum_{i=1}^{N} \left(\boldsymbol{x_i}^T \boldsymbol{\beta} - y_i\right)^2 = \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2
\end{aligned}
$$

Régression linéaire multiple

- 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:

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

Régression linéaire multiple



Figures taken from [James et al., 2013]

Aparté: quelques rappels

- La norme au carrée d'un vecteur $\boldsymbol{v}$ vaut $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T\boldsymbol{v}$.
- Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ 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!):

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

Aparté: quelques rappels

- La norme au carrée d'un vecteur $\boldsymbol{v}$ vaut $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T\boldsymbol{v}$.
- Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ 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!):

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

Aparté: quelques rappels

- La norme au carrée d'un vecteur $\boldsymbol{v}$ vaut $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T\boldsymbol{v}$.
- Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ 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!):

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

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

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

Aparté: quelques rappels

- La norme au carrée d'un vecteur $\boldsymbol{v}$ vaut $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T\boldsymbol{v}$.
- Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ 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!):

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

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

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

Aparté: quelques rappels

- La norme au carrée d'un vecteur $\boldsymbol{v}$ vaut $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T\boldsymbol{v}$.
- 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!):

$$\begin{aligned} \frac{\partial\left(\boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} &= \mathbf{0} \quad \text{(vecteur nul)} \\ \frac{\partial\left(\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} &= \left(\boldsymbol{y}^T\boldsymbol{X}\right)^T = \boldsymbol{X}^T\boldsymbol{y} \quad \text{et} \quad \frac{\partial\left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y} \\ \frac{\partial\left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} &= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

Aparté: quelques rappels

- La norme au carrée d'un vecteur $\boldsymbol{v}$ vaut $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T\boldsymbol{v}$.
- Si r est un scalaire et $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ 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!):

$$\begin{aligned}
\frac{\partial\left(\boldsymbol{y}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} &= \mathbf{0} \quad \text{(vecteur nul)} \\
\frac{\partial\left(\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} &= \left(\boldsymbol{y}^T\boldsymbol{X}\right)^T = \boldsymbol{X}^T\boldsymbol{y} \quad \text{et} \quad \frac{\partial\left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y} \\
\frac{\partial\left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} &= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}
\end{aligned}$$

Régression linéaire multiple

- 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 $\mathrm{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β:

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

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

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

Régression linéaire multiple

- 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 multiple

- 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 multiple

- 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 multiple

- 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 multiple

- 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 multiple

- 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 multiple

- 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 multiple

- 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 multiple

- 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 multiple

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

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

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

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

Régression linéaire multiple

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

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

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

$$\hat{y}_i = \boldsymbol{x_i}^T\hat{\boldsymbol{\beta}} \text{ pour } i = 1, \ldots, M, \text{ i.e.,}$$
$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} \text{ (forme vectorielle)}$$

Régression linéaire multiple

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

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

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

$$\hat{y}_i = \boldsymbol{x_i}^T\hat{\boldsymbol{\beta}} \text{ pour } i = 1, \ldots, M, \text{ i.e.,}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} \text{ (forme vectorielle)}$$

Régression linéaire multiple

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

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

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

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

Autres régressions

- Il existe bien d'autres méthodes non-linéaires qui généralisent la régression linéaire simple.
- Par exemple, pour modéliser la relation entre un seul prédicteur X et la réponse Y, on a:
  - **Polynomial regression**
  - **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 $\mathbb{E}(\epsilon) = 0$.
- Il se peut que certaines des variables X_i soient peu ou pas du tout associées avec la réponse Y_i.
- Inclure ces variables accroît la complexité du modèle, affecte sa performance, et réduit son interprétabilité.
- Il existe alors des méthodes de réduction et/ou sélection des variables les plus significatives: **shrinkage** et **feature selection**.

Régularisation

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

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

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

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

Régularisation

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

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

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

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

Régularisation

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

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

  où ϵ est un bruit tel que $\mathbb{E}(\epsilon) = 0$.
- Il se peut que certaines des variables X_i soient peu ou pas du tout associées avec la réponse Y_i.
- Inclure ces variables accroît la complexité du modèle, affecte sa performance, et réduit son interprétabilité.
- Il existe alors des méthodes de réduction et/ou sélection des variables les plus significatives: **shrinkage** et **feature selection**.

Régularisation

Parmi ces méthodes, on a:

- Subset selection methods:
  - best subset selection
  - forward stepwise selection
  - backward stepwise selection
- Regularization methods:
  - Ridge regression (shrinkage)
  - LASSO (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 $\mathbb{E}(\epsilon) = 0$.

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

Régression Ridge

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

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

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

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ 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 positif de réduire la variance du modèle.
- En gros, les variables X_i les moins significatives ont leur estimateurs associés $\hat{\beta}_i$ qui convergent vers 0 (shrinkage method).

Régression Ridge

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

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

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

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ 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 positif de réduire la variance du modèle.
- En gros, les variables X_i les moins significatives ont leur estimateurs associés $\hat{\beta}_i$ qui convergent vers 0 (shrinkage method).

Régression Ridge

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

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

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

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ 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 positif de réduire la variance du modèle.
- En gros, les variables X_i les moins significatives ont leur estimateurs associés $\hat{\beta}_i$ qui convergent vers 0 (shrinkage method).

Régression Ridge

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

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

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

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

- 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)**

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

- où λ est le *paramètre de régularisation*.
- Le terme $\lambda \left\| \boldsymbol{\beta} \right\|_2^2$ 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)**

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

- 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 $\mathrm{RSS}(\beta)$, on annule la dérivée de cette fonction par rapport à β:

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

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

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

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 $\boldsymbol{\beta}$:

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

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

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 $\boldsymbol{\beta}$:

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

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

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 $\boldsymbol{\beta}$:

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

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

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.

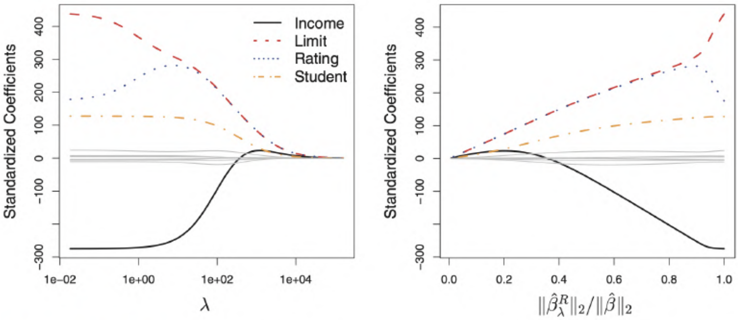


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* Credit *data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.*

Figures taken from [James et al., 2013]

Régression Ridge

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

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

Régression Ridge

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

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

Régression Ridge

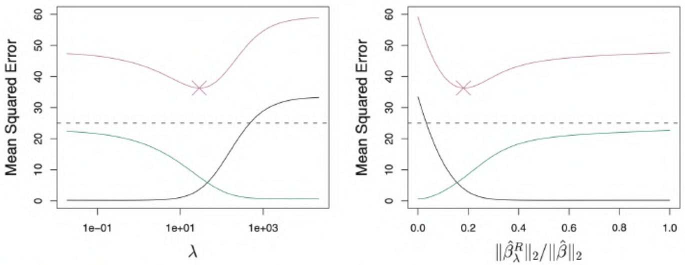


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple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

Figures taken from [James et al., 2013]

Régression LASSO

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

Régression LASSO

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

Régression LASSO

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

Régression LASSO

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

Régression LASSO

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

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

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

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

Régression LASSO

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

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

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

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ 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).

Régression LASSO

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

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

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

- Le but est d'obtenir des estimateurs $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ 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).

Régression LASSO

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

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

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

Régression LASSO

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

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

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

Régression LASSO

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

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

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

Régression LASSO

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

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

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

Régression LASSO

- On a:

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

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

Régression LASSO

- On a:

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

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

LASSO

- Lorsque λ augmente, certains coefficients deviennent nuls.

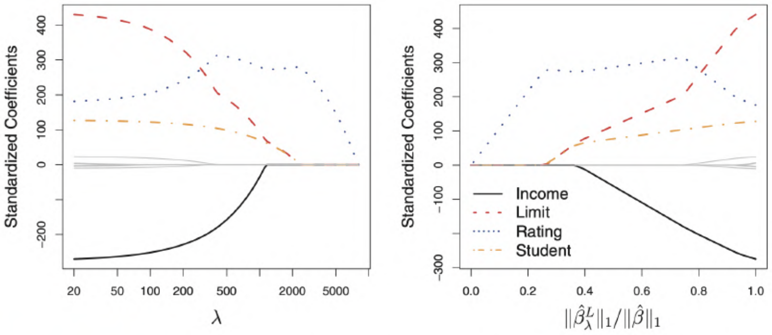


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

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:

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

- 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 $\hat{\beta}$ 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:

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

- 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:

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

- 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)**

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

- On a donc:

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

- 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.