Statistiques avancées — Régression
Cours 2 : Régression linéaire multivariée

21 Septembre 2021

Vecteurs aléatoires

- On dit que $X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} \in \mathbb{R}^d$ est un vecteur aléatoire si $X_1, \ldots, X_d$ sont des variables aléatoires. On suppose $\mathbb{E}[X_i^2] < +\infty$.
- L'espérance de X est définie par $\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$.
- L'opérateur de covariance est défini par

$$\Sigma_X = \mathrm{Cov}(X) \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_d) \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \ldots & \mathrm{Var}(X_d) \end{pmatrix}$$

Transformation affine

Théorème

Soit $\alpha \in \mathbb{R}^d$ et $A \in \mathbb{R}^{k \times d}$. Le vecteur $Y = \alpha + AX$ est un vecteur aléatoire d'espérance

$$\mathbb{E}[Y] = \alpha + A\mathbb{E}[X],$$

et d'opérateur de covariance

$$\Sigma_Y = A\Sigma_X A^\top.$$

Fonction caractéristique

Soit $X \in \mathbb{R}^d$ un vecteur aléatoire. Sa fonction caractéristique est définie par $\Phi_X : \mathbb{R}^d \to \mathbb{C}$, avec pour $t \in \mathbb{R}^d$:

$$\Phi_X(t) = \mathbb{E}[e^{i\langle t, X\rangle}] = \mathbb{E}[e^{\sum_{j=1}^d t_i X_i}].$$

Vecteurs Gaussiens

Variable Gaussienne

Une variable aléatoire $z \in \mathbb{R}$ est Gaussienne si et seulement si sa fonction caractéristique est

$$\Phi(z) = e^{im - \frac{t^2\sigma^2}{2}}, \forall t \in \mathbb{R}.$$

Vecteur Gaussien

Un vecteur aléatoire $X \in \mathbb{R}^d$ est Gaussien si et seulement si $a^\top X$ est une variable Gaussienne pour tout $a \in \mathbb{R}^d$.

- Si $\mathbb{E}[X] = 0$ et $\mathrm{Cov}(X) = I_d$, on dit que X est un vecteur Gaussien standard.
- Si $X \in \mathbb{R}^d$ est un vecteur Gaussien, alors ses coefficients sont des variables Gaussiennes.

Indépendance et conditionnement

Indépendance

Soit $X \in \mathbb{R}^d$ un vecteur Gaussien. X_i et X_j sont indépendants si et seulement si $\mathrm{Cov}(X_i, X_j) = 0$.

Conditionnement

Soit $X \in \mathbb{R}^d$ un vecteur Gaussien, et soit $Y = (X_1, \ldots, X_m)^\top$ et $Z = (X_{m+1}, \ldots, X_d)^\top$. Alors $Y|Z \sim \mathcal{N}(\mu_c, \Sigma_c)$, avec :

$$m_c = m_Y + \Sigma_{YZ}(Z - m_Z), \quad \Sigma_c = \Sigma_Y - \Sigma_{YZ}\Sigma_Z^{-1}\Sigma_{YZ}.$$

Théorème de Cochran

Le théorème de Cochran est un outil essentiel pour les tests d'hypothèses en Régression linéaire

Théorème

Soit $Y \in \mathbb{R}^n$ un vecteur Gaussien avec $\mathbb{E}[Y] = \mu$ et $\Sigma_Y = \sigma^2 I_n$. Soient $M_1, \dots, M_k$ des sous-espaces linéaires mutuellement orthogonaux. Alors :

1. $P_{M_i}(Y)$, $i = 1, \dots, k$ *sont non corrélés. De plus :*

$$\mathbb{E}[P_{M_i}(Y)] = P_{M_i}(\mu), \quad \mathrm{Cov}(P_{M_i}(Y)) = \sigma^2 P_{M_i},$$

$$\mathbb{E}||P_{M_i}(Y)||^2 = \sigma^2 dim(M_i) + \mathbb{E}||P_{M_i}(\mu)||^2.$$

2. $P_{M_i}(Y) \sim \mathcal{N}(P_{M_i}(\mu), \sigma^2 P_{M_i})$ *are mutually independent.*
3. $||P_{M_i}(Y)||^2$ *are mutually independent and*

$$||P_{M_i}(Y)||^2 \sim \sigma^2 \chi^2_{\dim(M_i), ||P_{M_i}(\mu)||/\sigma}.$$

Le problème de régression

- Ce cours présentera un cadre statistique approprié pour étudier le problème de la régression multiple :
  - Un modèle probabiliste pour décrire les observations
  - La construction des estimateurs et leurs propriétés de base : (biais, variance, propriétés asymptotiques...)
  - Optimalité statistique (Gauss-Markov, Efficience, Minimax)
  - Inférence statistique (intervalles de confiance, tests)
- On observe un n-échantillon $((y_1, X_1), \ldots, (y_n, X_n))$, où $y_i \in \mathrm{R}$ est la réponse et $X_i \in \mathrm{R}^{\,p}$ est un vecteur de prédicteurs. Objectif : Expliquer et prédire la réponse y à partir des prédicteurs X.

Exemple : prédiction de la consommation d'essence

- Les données contiennent les caractéristiques suivantes pour $n = 31$ voitures différentes :
  - `Consommation` = Consommation en essence pour 100km.
  - `Prix` = Prix du véhicule en francs Suisses.
  - `Cylindree` = Cylindrée en cm3.
  - `Puissance` = Puissance en kW.
  - `Poids` = Poids en kg.
- Objectif : Expliquer et prédire la consommation d'essence de différents modèles de voitures en fonction d'autres caractéristiques :
  - La réponse Y est la variable `Consommation`
  - Les prédicteurs X_j correspondent aux quatre autres variables.

Modèle linéaire

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Précisions

- Les ε_i sont des erreurs stochastiques, en général $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Les coefficients de régression β_j mesurent l'effet de chaque prédicteur sur la réponse.

- Forme matricielle :

$$X = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{pmatrix}, \quad Y = X\beta + \varepsilon.$$

Quelques commentaires

1. Le modèle est identifiable (i.e. $X\beta = X\beta' \Rightarrow \beta = \beta'$) s.si X est de rang plein ($X^\top X$ est inversible).

2. On suppose la plupart du temps que ε est centré de variance finie, et l'homoscédasticité du bruit.

3. On peut faire du feature engineering pour créer de nouveaux X avec des fonctions non-linéaires (log, carré, etc.). Régression **linéaire** indique que le modèle dépend linéairement de β, mais pas forcément de X.

Hypothèses et postulats

Dans ce cours, on supposera toujours que :

1. Les erreurs sont centrées : $\forall i = 1, \dots, n$, $\mathbb{E}[\varepsilon_i] = 0$
2. Les erreurs sont de variance constante (homoscédasticité) : $\forall i = 1, \dots, n$, $\mathrm{Var}[\varepsilon_i] = \sigma^2$
3. Les erreurs sont décorrélées entre elles : $\forall i, j = 1, \dots, n$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$

On supposera en général par simplicité :

4. Les erreurs sont Gaussiennes : $\forall i = 1, \dots, n$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Définition

L'estimateur des moindres carrés (LSE, pour Least Squares Estimator) pour le modèle de régression

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

est défini comme suit.

LSE

$$\hat{\beta} = \operatorname{argmin}_{u \in \mathbb{R}^p} \|Y - Xu\|^2,$$

où $\|u\|^2 = \sum_{j=1}^{p} u_j^2$ est la norme Euclidienne.

Proposition

Si $X^\top X$ est inversible, $\hat{\beta}$ admet une forme close

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Propriétés statistiques

Biais et variance

Sous les hypothèses 1—3 discutées précédemment et si $X^\top X$ est inversible :

- For all $\beta \in \mathrm{R}^{p}$, $\mathbb{E}_\beta[\hat{\beta}] = \beta$.
- For all $\beta \in \mathrm{R}^{p}$, $\mathrm{Var}_\beta[\hat{\beta}] = \sigma^2 (X^\top X)^{-1}$.

Théorème de Gauss-Markov

Sous les hypothèses 1—3 discutées précédemment et si $X^\top X$ est inversible, l'estimateur LSE $\hat{\beta}$ est l'estimateur de variance minimale parmi tous les estimateurs linéaires.

Résidus

On définit les résidus empiriques $\hat{\varepsilon} = Y - X\hat{\beta}$. Sous les mêmes hypothèses :

Proposition

- $\mathbb{E}_\beta[\hat{\varepsilon}] = 0$
- $\mathrm{Var}_\beta[\hat{\varepsilon}] = \sigma^2 P_{X^\perp}$, où $P_{X^\perp}$ est la matrice de projection sur le sous-espace orthogonal à X.
- $\mathrm{Cov}(\hat{\varepsilon}, \hat{Y}) = 0$.
- $\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p}$ est un estimateur non biaisé de la variance σ^2.

Equivalence with the MSE

Dans le modèle Gaussien où $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, l'estimateur du Maximum de vraisemblance minimise la fonction de vraisemblance négative associée à l'échantillon :

$$\hat{\beta}_{ML} \in \operatorname{argmin}_{u \in \mathbb{R}^p} \frac{1}{2\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - X_i^\top u)^2.$$

L'estimateur MLE est donc équivalent au LSE.

- Notons que l'estimateur de la variance $\hat{\sigma}^2_{ML}$ est donné par $\hat{\sigma}^2_{ML} = \frac{\|\hat{\varepsilon}\|^2}{n}$. C'est donc un estimateur biaisé.

Propriétés du MLE

D'après le théorème de Cochran, sous les hypothèses 1—4 et si $X^\top X$ est inversible, alors

Proposition : propriétés du MLE

- $\hat{\beta}_{ML} \sim \mathcal{N}(\beta, \sigma^2 (X^\top X)^{-1})$,
- $\frac{(n-p)\hat{\sigma}^2_{ML}}{\sigma^2} \sim \chi^2(n-p)$,
- $\hat{\beta}_{ML}$ et $\hat{\sigma}^2_{ML}$ sont indépendants.

Problème d'inférence et tests d'hypothèse

- Jusqu'à présent nous avons discuté le problème d'estimation : approcher le paramètre inconnu β.

- On traite à présent le problème d'inférence, avec le calcul de régions de confiance et les tests d'hypothèse. Pour cela, on va devoir étudier la distribution des estimateurs LSE et MLE.

Distribution du LSE

On rappelle que $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ et $\sigma^2 = \frac{\|P_{X^\perp}(Y)\|^2}{n-p}$.

Théorème

Sous les hypothèses 1—4 et si $X^\top X$ est inversible, alors

1. $\forall c \in \mathbb{R}^p$,
$$\frac{c^\top \hat{\beta} - c^\top \beta}{\hat{\sigma}\sqrt{c^\top (X^\top X)^{-1} c}} \sim t_{(n-p)}.$$
2. $\forall C \in \mathbb{R}^{q \times p}$ de rang q $(q \leq p)$,
$$\frac{(C\hat{\beta} - C\beta)^\top (C(X^\top X)^{-1} C^\top)(C\hat{\beta} - C\beta)}{q\hat{\sigma}^2} \sim \mathcal{F}(q, n-p).$$

Régions de confiance

Le résultat précédent peut être utilisé pour construire des régions de confiance pour le paramètre inconnu β.

Théorème

Sous les hypothèses 1—4 et si $X^\top X$ est inversible, soit $\alpha \in]0,1[$. Alors

1. Pour tout $c \in \mathrm{R}^{\,p}$, l'intervalle ci-dessous est un intervalle de confiance exact au niveau $1-\alpha$:

$$I_{c,\alpha} = \left[c^\top \hat{\beta} \pm t_{n-p,1-\alpha/2} \hat{\sigma} \sqrt{c^\top (X^\top X)^{-1} c} \right]$$

Notons que, pour $c = e_i$ le i-ème vecteur de la base canonique de $\mathrm{R}^{\,p}$, on obtient directement un intervalle de confiance pour le coefficient de régression β_i.

Tests d'hypothèse

- On veut tester les hypothèses

$$H_0 : \ c^\top \beta = a \text{ vs } H_1 : \ c^\top \beta \neq a,$$

pour $c \in \mathrm{R}^{\,p}$ et $a \in \mathrm{R}$.

- Ce cadre couvre le cas suivant :

$$H_0 : \ \beta_j = 0 \text{ vs } H_1 : \ c^\top \beta_j \neq 0.$$

Ce test est utilisé pour déterminer si une covariable X_j a un impact sur la variable réponse Y.

Tests d'hypothèse

Théorème

Soit $c \in \mathrm{R}^{\,p}$ et $\alpha \in]0,1[$. On considère les hypothèses

$$H_0 : \; c^\top \beta = a \text{ vs } H_1 : \; c^\top \beta \neq a.$$

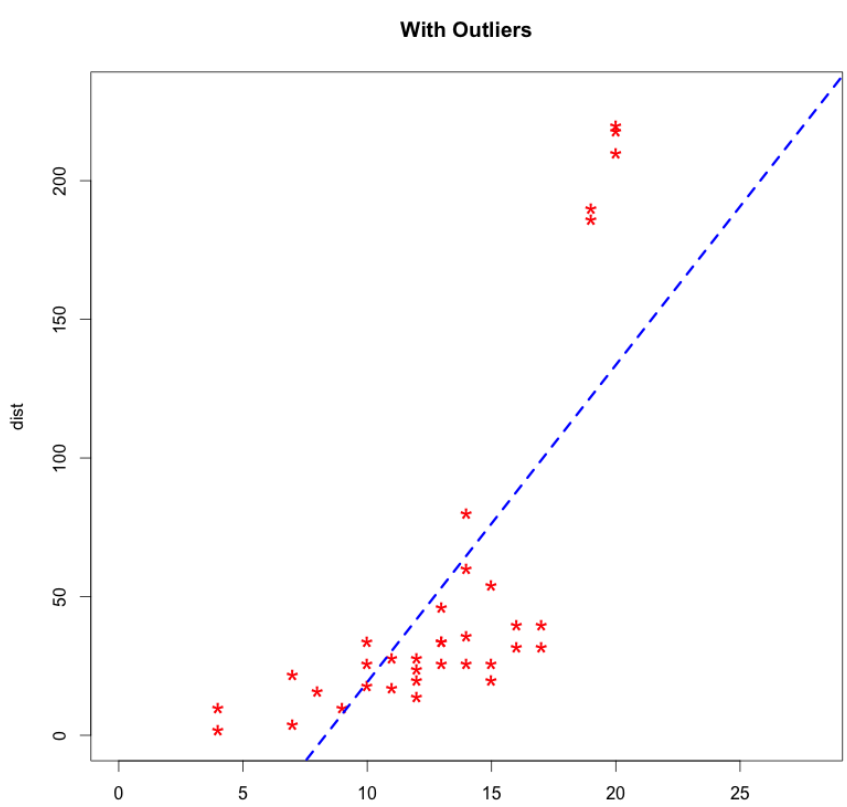
Le test suivant admet une erreur de type I de taille α :

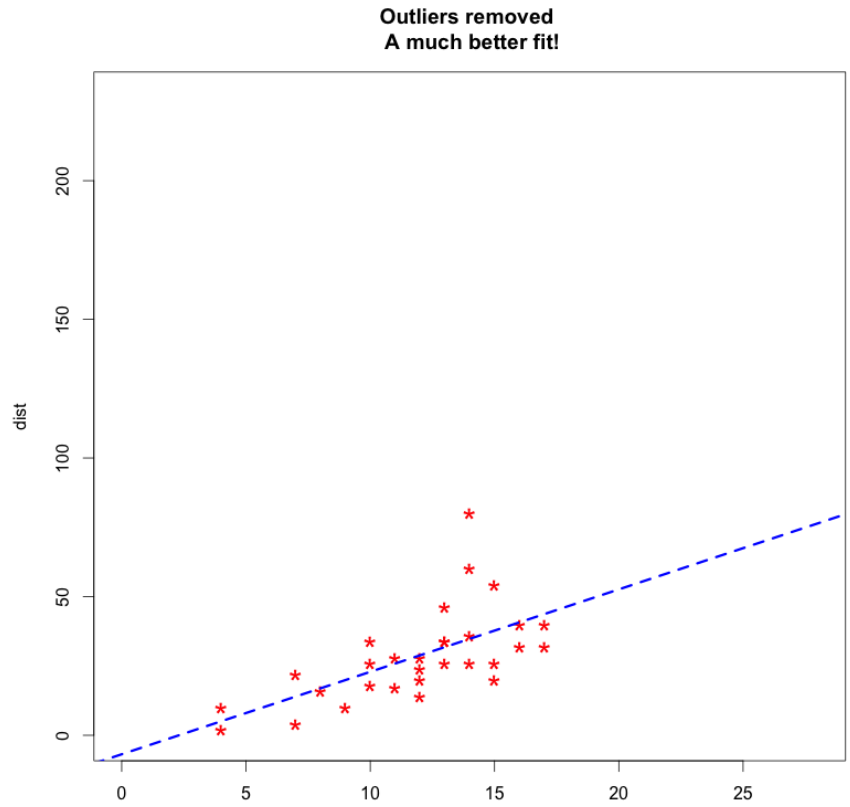
$$\phi(Y) = 1_{\{|T| > t_{n-p,1-\alpha/2}\}} \text{ with } T := \frac{c^\top \hat{\beta} - a}{\hat{\sigma}\sqrt{c^\top (X^\top X)^{-1} c}},$$

and $t_{n-p,1-\alpha/2}$ est le quantile $1-\alpha/2$ de $t(n-p)$.

Pénalisation et robustesse

- Les estimateurs des moindres carrés et du maximum de vraisemblance sont de "mauvais" estimateurs dans de nombreux cas, notamment lorsque le nombre de prédicteurs est grand, ou que les données contiennent des outliers. Dans ce cas on peut utiliser des extensions robustes.

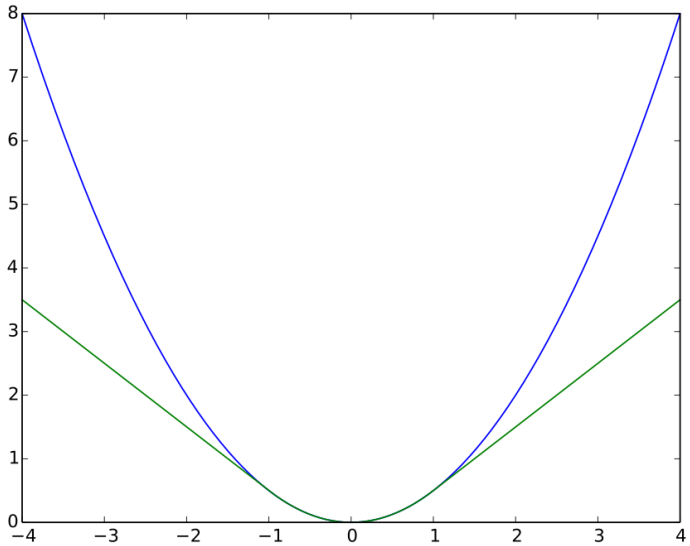




Huber loss

- La première possibilité est d'utiliser une autre perte que celle des moindres carrés. Pour le problème de régression, la fonction de Huber est moins sensible aux outliers. Elle est définie par

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{si } |a| \leq \delta, \\ \delta(|a| - \frac{1}{2}\delta) & \text{sinon.} \end{cases}$$

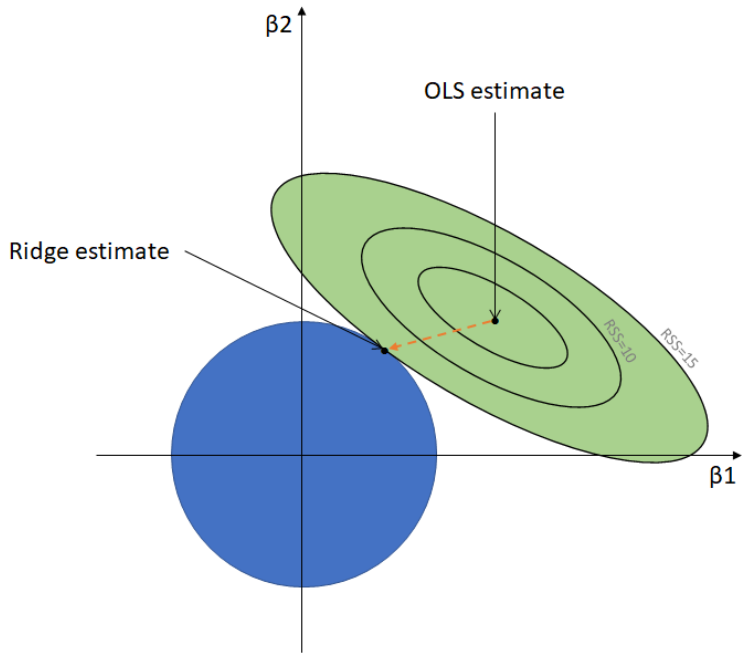


Pénalisation et régression ridge

- La deuxième possibilité est de "pénaliser" les estimateurs qui prennent de trop grande valeurs à l'aide d'une fonction de régularisation/pénalisation. En régression, la méthode pénalisée la plus simple est la régression ridge, définie par :

$$\hat{\beta}_{ridge} \in \operatorname{argmin}_{u \in \mathbb{R}^p} \underbrace{\|Y - Xu\|^2}_{\text{fonction d'attache aux données}} + \underbrace{\lambda \|u\|^2}_{\text{pénalité ridge}} .$$

La pénalité ridge ramène l'estimateur $\hat{\beta}$ vers des régions de petite norme euclidienne, comme sur l'image :



Régression ridge

- Pour les prédicteurs corrélés, le LSE de $\hat{\beta}$ peut avoir une grande variance (et donc prendre de grandes valeurs).
- Pour contrôler la variance, nous pouvons imposer des contraintes sur la taille des estimateurs. Cela peut produire un estimateur avec une meilleure précision de prédiction.

$$\hat{\beta}_{ridge} \in \operatorname{argmin}_{u \in \mathbb{R}^p} \|Y - Xu\|^2 + \lambda \|u\|^2,$$

- Le paramètre $\lambda > 0$ est appelé paramètre de régularisation.

Régression ridge

$$\hat{\beta}_{ridge} = \text{argmin}_{u \in \mathbb{R}^p} \|Y - Xu\|^2 + \lambda \|u\|^2,$$

- La fonctionnelle à minimiser est fortement convexe, la solution est donc unique.
- Il esxiste une forme close pour $\hat{\beta}_{ridge}$, c'est donc une solution rapide.
- La pénalité ridge améliore la stabilité numérique de l'estimateur,
- Elle mène à de meilleures erreurs de prédiction en général.

Exemple jouet

On considère le modèle

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

avec X_1 et X_2 fortement corrélés.

- Si $X \simeq X_2$, alors pour tout $\gamma > 0$,

$$Y \simeq X_1(\beta_1 + \gamma) + X_2(\beta_2 - \gamma) + \varepsilon.$$

  On en déduit qu'il existe de nombreuses solutions au pbm des moindres carrés.

- La pénalité ridge contraint la norme euclidienne de β à n'être pas trop grande. Cela permet de "sélectionner" parmi les LSE possible celui qui minimise la norme euclidienne. Plus λ augmente, plus la norme de $\hat{\beta}_{ridge}$ diminue.

Chemin de régularisation