Régression pénalisée

masedki.github.io

4 novembre 2021

Sélection de variables en régression linéaire

- Méthodes pas à pas
- Critères d'information
- Méthodes de régularisation, contraction de coefficients ou shrinkage

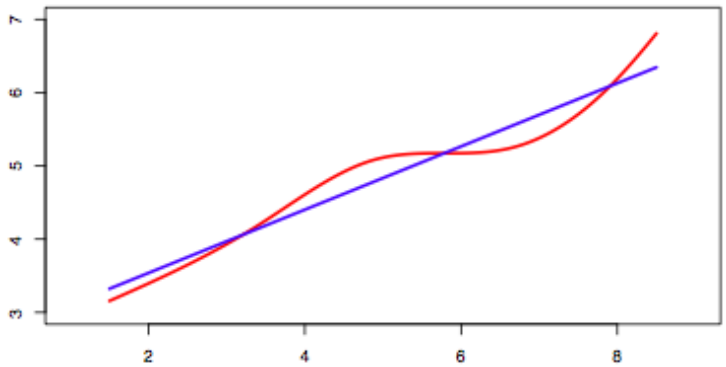
Classification

- Régression logistique
- Sélection de variables en régression logistique

Sélection de variables en régression linéaire

Régression linéaire : rappel

- Une approche simple pour faire de l'apprentissage supervisé. Elle suppose que Y dépend linéairement de $X_1, \ldots X_p$
- Les vraies fonctions de régression ne sont jamais linéaires



- Même si cela semble trop simple, la régression linéaire est extrêmement utile à la fois conceptuellement et en pratique.

Le cas de régression linéaire simple

- On pose un modèle de la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où β_0, β_1 inconnus sont ordonnée à l'origine (intercept) et pente (slope). Ce sont les *coefficients* du modèle

- Étant estimés ces coefficients par $\widehat{\beta_0}$ et $\widehat{\beta_1}$, on prédit y sachant x avec

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x,$$

Régression linéaire multiple

- Notre modèle

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

- On interprète β_j comme l'effet moyen sur Y d'un accroissement de X_j d'une unité *lorsque tous les autres prédicteurs sont fixés*.
- On ne peut faire aucune affirmation en terme de *causalité*.

Exemple. $Y =$ *serum triglycerides mg/dl*, $A =$ *age in years* et B *body-mass index*, kg/m^2.
$\widehat{Y} = -247.25 + 3.5A + 9.3B$.
Comment s'interprète 9.3 ?

Interpréter les coefficients de régression

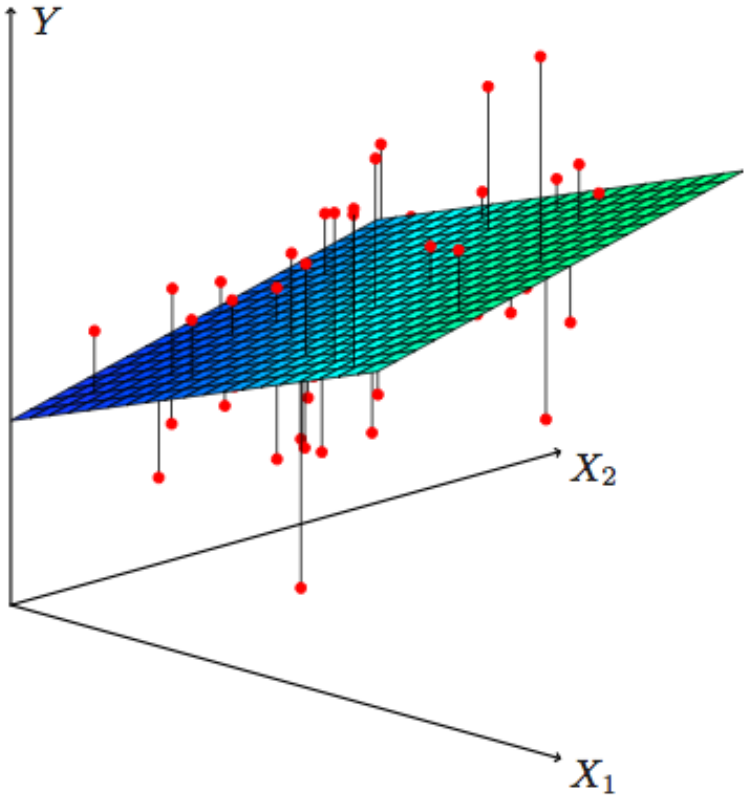
- Le scenario idéal lorsque les prédicteurs sont indépendants, et le design équilibré
  - chaque coeff peut être estimé et testé séparément
  - interprétation de gauche est OK
- La corrélation entre X_j pose des problèmes
  - la variance des estimateurs s'accroit
  - l'interprétation devient hasardeuse (lorsque X_j change, tout change !)

Estimation et prédiction pour la régression multiple

- À partir d'estimation des coef $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\ldots$, $\widehat{\beta_p}$, on peut prédire avec

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p$$

- Comme en dimension 1, on estime les β en minimisant la somme des carrées résiduelle. Formule théorique qui dépend d'une inversion de matrice produit. $\rightarrow$ utiliser un logiciel de statistique
- De même, $SE(\beta_j)$ pour chaque coefficient, t-test de nullité, test de Fisher,...



Résultats pour les données triglycérides

	Coeff	Std.Err	t-stat	p-value
Intercept	-247.25	21.24	-11.64	<2e-16
`BMI`	9.30	0.90	10.32	<2e-16
`age`	3.50	0.19	18.69	<2e-16

Corrélations

	`TG`	`BMI`	`age`
`TG`	1.00	0.56	0.73
`BMI`		1.00	0.34
`age`			1.00

Quelques rappels

On note

- La somme des carrés totale $\text{SST} = \text{SYY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- La somme des carrés résiduelle $\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, où

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \quad \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

- La somme des carrés expliquée par la régression $\text{SSreg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Des questions importantes

1. Y a t-il au moins un des X_j utile pour prédire Y ?
2. Sont-ils vraiment tous utiles ?
3. Comment le modèle s'ajuste aux données ?
4. Avec une nouvelle valeur de X, quelle réponse doit-on prédire ? Précision de la prédiction ?

Des questions importantes

1. Y a t-il au moins un des X_j utile pour prédire Y ?
2. Sont-ils vraiment tous utiles ?
3. Comment le modèle s'ajuste aux données ?
4. Avec une nouvelle valeur de X, quelle réponse doit-on prédire ? Précision de la prédiction ?

Pour la première question, on utilise la F-statistique

$$F = \frac{(\text{SST} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

Quantité	Valeur
Residual Std.Err	50.34
R^2	0.63
F-stat	343.7

Des questions importantes

1. Y a t-il au moins un des X_j utile pour prédire Y ?
2. Sont-ils vraiment tous utiles ?
3. Comment le modèle s'ajuste aux données ?
4. Avec une nouvelle valeur de X, quelle réponse doit-on prédire ? Précision de la prédiction ?

- Choix de co-variables : approche complète
  Comparer les modèles linéaires avec tous les sous-ensembles possibles de co-variables
- Souvent 2^p trop grand ($\log_{10}(2^{40}) \approx 12.0$
  On utilise une méthode que ne parcourt que certains sous-ensembles. Deux approches standard
  Sélections progressive, ou rétrograde
- Nécessite de répondre à la question suivante pour effectuer la comparaison.

Choix de co-variables

Méthode progressive
(forward)

1. Commencer par le modèle nul (à zéro co-variables)
2. Ajuster les p régressions linéaires simples et ajouter au modèle nul la co-variable qui à le plus petit RSS
3. Ajouter à ce modèle à une co-variable la co-variable qui fait baisser le plus le RSS
4. Continuer jusqu'à un critère d'arrêt (par exemple sur la p-value du t-test)

Méthode rétrograde
(backward)

1. Commencer par le modèle avec tous les co-variables
2. Supprimer la variable avec la plus grande p-value —i.e., la co-variable la moins significative pour le modèle
3. Ré-ajuster le modèle, et enlever de nouveau la co-variable de plus grande p-value
4. Continuer jusqu'à un critère d'arrêt (par exemple portant sur la valeur de la p-value de la co-variable que l'on enlèverait)

Choix de co-variables

- Critère plus systématique pour choisir le modèle « optimal » dans ceux que l'on parcourt
- Avec C_p de Mallows, Akaike information criterion (AIC), Bayesian information criterion (BIC), R^2 ajusté et validation croisée (CV)

Évaluer un sous ensembles de covariables

On supposera dans la suite que nous avons m covariables au total et on entend par modèle un sous-ensemble de ces covariables de taille p.

Évaluer un sous ensembles de covariables

On supposera dans la suite que nous avons m covariables au total et on entend par modèle un sous-ensemble de ces covariables de taille p.

Rappelons que

$$R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$$

et

$$R^2_{\text{adj}} == 1 - \frac{\text{RSS}\,/(n-p-1)}{\text{SST}\,/(n-1)}$$

où p est le nombre de variables du modèle.

- On sélectionne le modèle avec le R^2_{adj} le plus élevé, cela revient à choisir le sous-ensemble de variables qui minimise

$$S^2 = \frac{\text{RSS}}{n-p-1}$$

où p est le nombre de variables du modèle.

Choix basé sur R^2_{adj}

- Souvent le choix basé sur R^2_{adj} montre un phénomène de ***over-fitting***.
- Supposons que la valeur maximale de $R^2_{\text{adj}} = 0.692$ pour un sous-ensemble $p = 10$ de covariables, $R^2_{\text{adj}} = 0.691$ pour $p = 9$ et $R^2_{\text{adj}} = 0.541$ pour un sous-ensemble de $p = 8$ covariables.
- Il est clairement préférable de choisir le modèle à $p = 10$ covariables.

Choix basé sur R^2_{adj}

- Souvent le choix basé sur R^2_{adj} montre un phénomène de ***over-fitting***.
- Supposons que la valeur maximale de $R^2_{\text{adj}} = 0.692$ pour un sous-ensemble $p = 10$ de covariables, $R^2_{\text{adj}} = 0.691$ pour $p = 9$ et $R^2_{\text{adj}} = 0.541$ pour un sous-ensemble de $p = 8$ covariables.
- Il est clairement préférable de choisir le modèle à $p = 10$ covariables.

On va faire appel à un critère(s) basé(s) sur la vraisemblance.

La vraisemblance

Rappelons que d'après les hypothèses du modèle linéaire

$$Y_i \mid x_{i1}, \ldots, x_{ip} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2\right).$$

Et

$$f\left(y_i \mid x_{i1}, \ldots, x_{ip}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(y_i - \left\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right\}\right)^2}{2\sigma^2}\right]$$

La vraisemblance

Rappelons que d'après les hypothèses du modèle linéaire

$$Y_i \mid x_{i1}, \ldots, x_{ip} \sim \mathcal{N}\big(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2\big).$$

Et

$$f\big(y_i \mid x_{i1}, \ldots, x_{ip}\big) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\Big(y_i - \big\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\big\}\Big)^2}{2\sigma^2}\right]$$

- Écrire la vraisemblance

$$L\big(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2; y_1, \ldots, y_n\big)$$

- Écrire la log-vraisemblance

$$\ell\big(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2; y_1, \ldots, y_n\big)$$

Mesurer l'ajustement ou l'adéquation du modèle

L'estimateur par maximum de vraisemblance de σ^2 est donné par

$$\hat{\sigma}^2_{\text{MLE}} = \frac{\text{RSS}}{n}.$$

$$\ell\big(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2_{\text{MLE}}; y_1, \ldots, y_n\big) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{\text{RSS}}{n}\right) - \frac{n}{2}$$

Critère d'information

Un critère d'information est une quantité qui réalise un compromis entre l'ajustement aux données (*la vraisemblance par exemple*) et la complexité du modèle.

Donc

- On peut chercher le modèle qui **maximise**

  Ajustement - Pénalité

- On peut chercher le modèle qui **minimise**

  - Ajustement + Pénalité

Critère d'information C_p de Mallows

Le critère d'information noté C_p de Mallows associé au modèle à p covariables est donné par

$$C_p = \frac{\text{RSS}_p}{S^2} + 2p - n$$

où RSS_p est la somme des carrés des résidus du modèle en question et S^2 est l'estimateur de σ^2 dans le modèle complet.

Critère d'information AIC

Le critère d'information noté AIC (*Akaike's Information Criterion*) associé à un modèle à p covariables est donné par

$$\text{AIC} = -2\ell\big(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2_{\text{MLE}}; y_1, \ldots, y_n\big) + 2K$$

où $K = p + 2$ est nombre de paramètre du modèle.
On peut montrer que

$$\text{AIC} = n \log \hat{\sigma}^2_{\text{MLE}} + 2p + \text{const.}$$

Ce critère est préférable pour la ***prédiction***.

Critère d'information BIC

Le critère d'information noté BIC (*Bayesian Information Criterion*) associé à un modèle à p covariables est donné par

$$\text{BIC} = -2\ell\big(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2_{\text{MLE}}; y_1, \ldots, y_n\big) + K \ln(n)$$

où $K = p + 2$ est nombre de paramètre du modèle.

- Ce critère possède de ***bonnes propriétés théoriques***.
- Ce critère est préférable pour ***l'explication***.

Sélection de modèle

L'idée de chercher le modèle (un sous-ensemble de covariables) qui minimise un des trois critères précédents.

Comment procéder ?

Sélection de modèle

L'idée de chercher le modèle (un sous-ensemble de covariables) qui minimise un des trois critères précédents.

Comment procéder ?

Recherche exhaustive : calculer la valeur du critère pour chaque modèle.

Sélection de modèle

L'idée de chercher le modèle (un sous-ensemble de covariables) qui minimise un des trois critères précédents.

Comment procéder ?

Recherche exhaustive : calculer la valeur du critère pour chaque modèle.

Problème combinatoire : $2^{\text{le nombre de covariables}}$ modèles en compétition ! !

Sélection de modèle en forward

Le modèle de départ de la procédure de sélection *forward* est le modèle avec la constante seulement. La procédure consiste à

1. Ajouter séparément chaque variable au modèle actuel et calculer le critère d'intérêt (BIC, AIC, ou C_p).
2. Si aucun des nouveaux modèles n'améliore le critère, alors : **stop**.
3. Mettre à jour le modèle en incluant la covariable qui apporte la meilleure amélioration au sens du critère. Aller à **1.**

Sélection de modèle backward

Le point de départ de la procédure d'élimination *backward* est le modèle complet incluant toutes les covariables. La procédure consiste à

1. Si aucune élimination d'une covariable n'améliore le critère alors : **stop**.
2. Mettre à jour le modèle en éliminant la covariable qui réalise la meilleure amélioration du critère. Aller à **1.**

Données cancer de la prostate

lcavol	log(cancer volume)
lweight	log(prostate weight)
age	age
lbph	log(benign prostatic hyperplasia amount)
svi	seminal vesicle invasion
lcp	log(capsular penetration)
gleason	Gleason score
pgg45	percentage Gleason scores 4 or 5
lpsa	log(prostate specific antigen)

Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine,E.A. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate : II. radical prostatectomy treated patients, Journal of Urology 141(5), 1076–1083.

Comparaison de ces critères

- Je déconseille le R^2 ajusté.
- C_p et AIC sont des critères qui réalisent un compris biais-variance. Ils sont donc indiqués pour choisir un modèle que l'on souhaite utiliser pour prédire.
- BIC pénalise plus les modèles de grandes dimensions. C'est le seul critère à être consistent (i.e., à fournir un estimateur qui converge lorsque $n \to \infty$)
- BIC étant plus sélectif, on doit le préférer si l'on souhaite un modèle explicatif.
- Lorsque la taille de la base d'apprentissage est grande, préférer BIC (AIC fournit des modèles de trop grandes dimensions)

Sélection de variables : quelques remarques

Interprétabilité

- Si le vrai modèle ne contient que **quelques variables liées à la response** $\rightsquigarrow$ les algorithmes de sélection peuvent retrouver les prédicteurs pertinents.
- Si le vrai modèle contient **beaucoup de variables trés corrélées** $\rightsquigarrow$ les variables sélectionnées seront difficiles à interpréter.

Limites liées à la stabilité

En présence de prédicteurs très corrélés ou lorsque $n < p$, **de petites perturbations** des données peuvent provoquer **de grandes différences** entre les ensembles de variables sélectionnées.

Méthode de shrinkage

Régression ridge et *Lasso*

- Les méthodes précédentes de choix de sous-ensembles utilisent les moindres carrés pour ajuster chacun des modèles en compétition.
- Alternativement, on peut ajuster un modèle contenant toutes les p covariables en utilisant une technique que *contraint* ou *régularise* les estimations des coefficients, ou de façon équivalente, pousse les coefficients vers 0.
- Il n'est pas évidant de comprendre pourquoi de telles contraintes vont améliorer l'ajustement, mais il se trouve qu'elles réduisent la variance de l'estimation des coefficients.

Régression ridge

- Rappelons que la procédure d'ajustement par moindres carrés estime les coefficients β_0, β_1,..., β_p en minimisant

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

- En revanche, la régression ridge estime les coefficients en minimisant

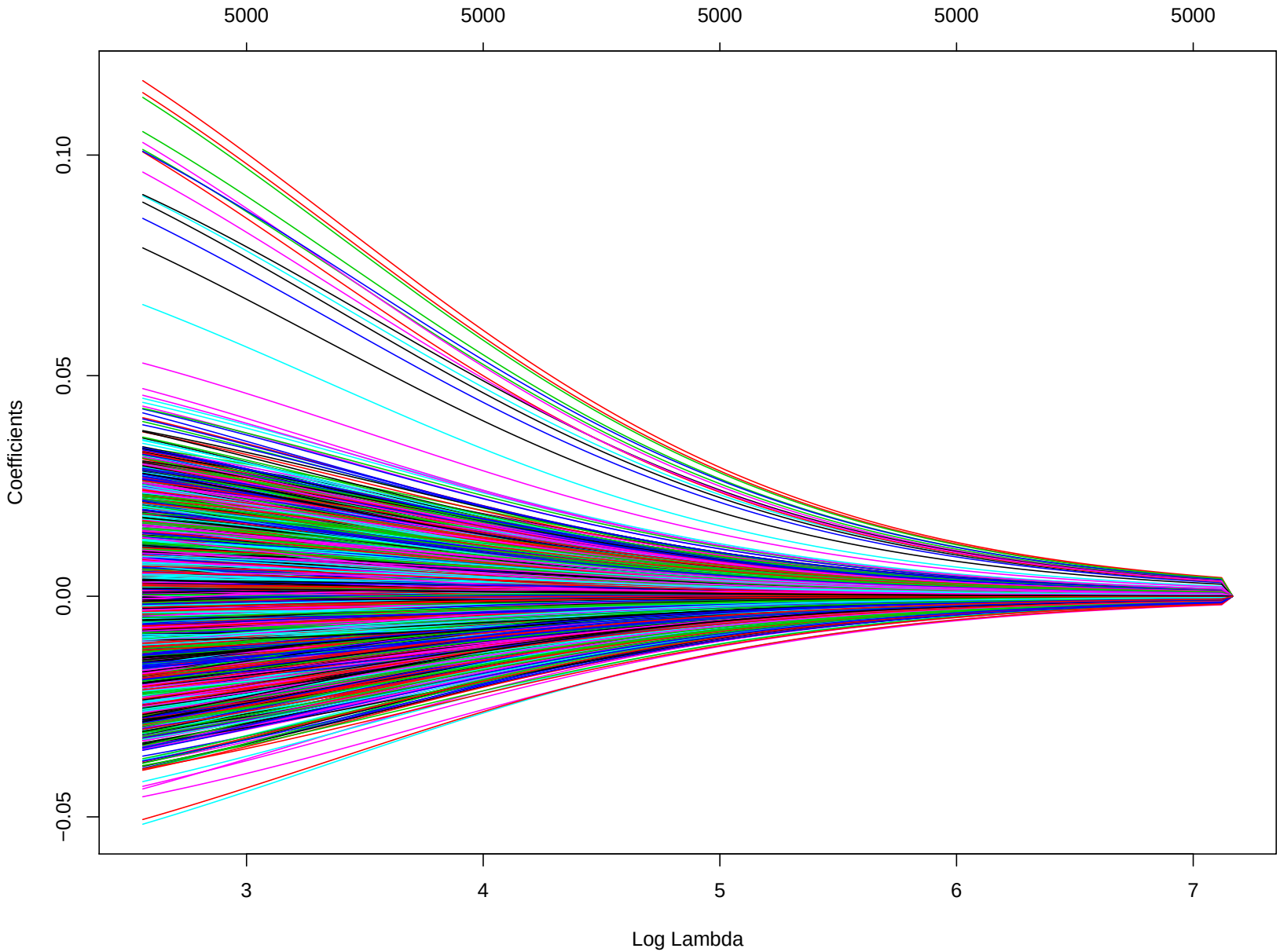
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2,$$

où λ est un *paramètre de réglage*, à déterminer par ailleurs.

Régression ridge (suite)

- Comme les moindres carrés, la régression ridge cherche des estimations des coefficients qui s'ajustent sur les données, donc à rendre $\mathrm{RSS}(\boldsymbol{\beta})$ petit.
- Cependant, le second terme $\lambda \sum_{j=1}^{p} \beta_j^2$, appelé *pénalité ridge* est petit lorsque les β_j sont proches de 0, et tire donc les estimations vers ce point.
- Le paramètre de réglage λ sert à contrôler l'impact de cette pénalité sur l'estimation.
- Choisir une bonne valeur de λ est critique pour construire un modèle acceptable. On utilise la validation croisée.

Exemple jeu de données $n = 1000$ et $p = 5000$

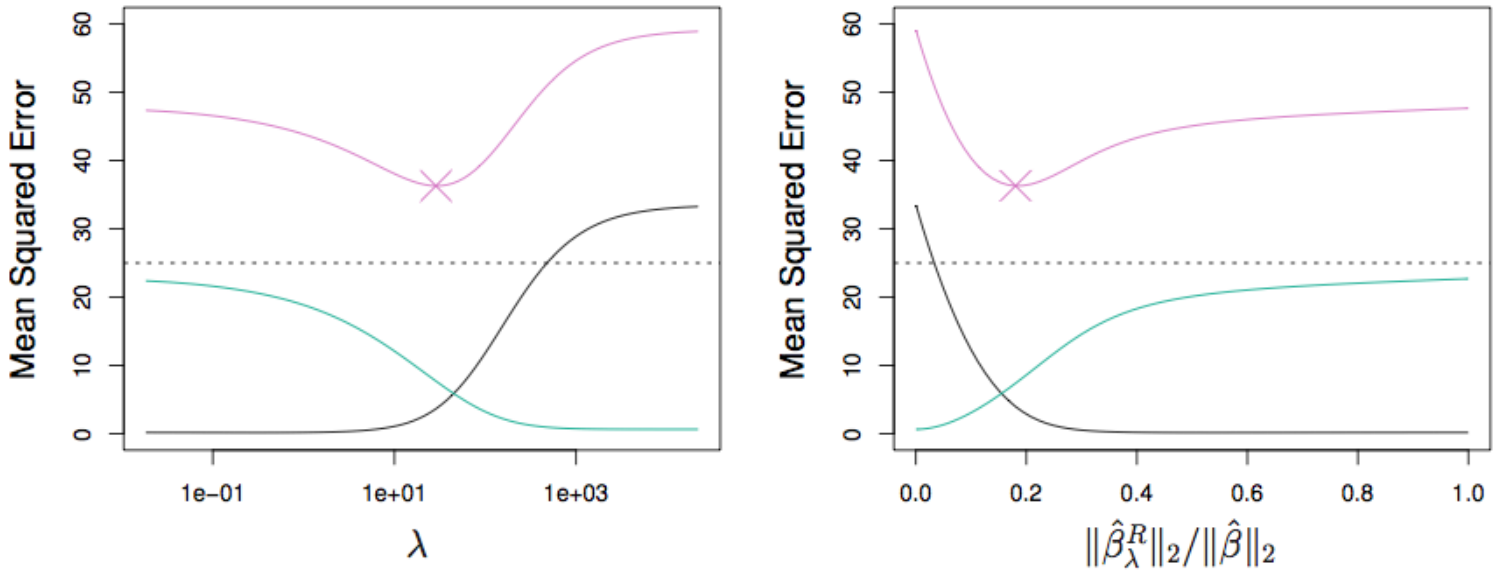


Régression ridge : normaliser les prédicteurs

- La méthode des moindres carrés standard est insensible à la normalisation des prédicteurs : si l'on multiplie X_j par c, le coefficient sera remplacé par $\hat{\beta}_j/c$.
- En revanche, la régression ridge peut changer *substantiellement* lorsque l'on multiplie un prédicteur par une constante, à cause de la norme quadratique dans le terme de pénalité.
- C'est pourquoi il est vivement recommander de toujours standardiser les prédicteurs (marginalement) avant d'utiliser la régression ridge.

Pour la régression ridge ?

Compromis biais-variance



Données simulées : $n = 50$, $p = 45$, tous de coefficients non nuls. Biais au carré (en noir), variance (en vert) et erreur de test quadratique (en violet) pour la régression ridge.
Droite horizontale : erreur minimale.

Le Lasso : *Least Absolute Shrinkage and Selection Operator*

- La régression ridge a un inconvénient évident : contrairement à la sélection de variable, la régression ridge inclut tous les prédicteurs dans le modèle final.
- Le Lasso est une alternative relativement récente qui répond à cette critique. On minimise en fait

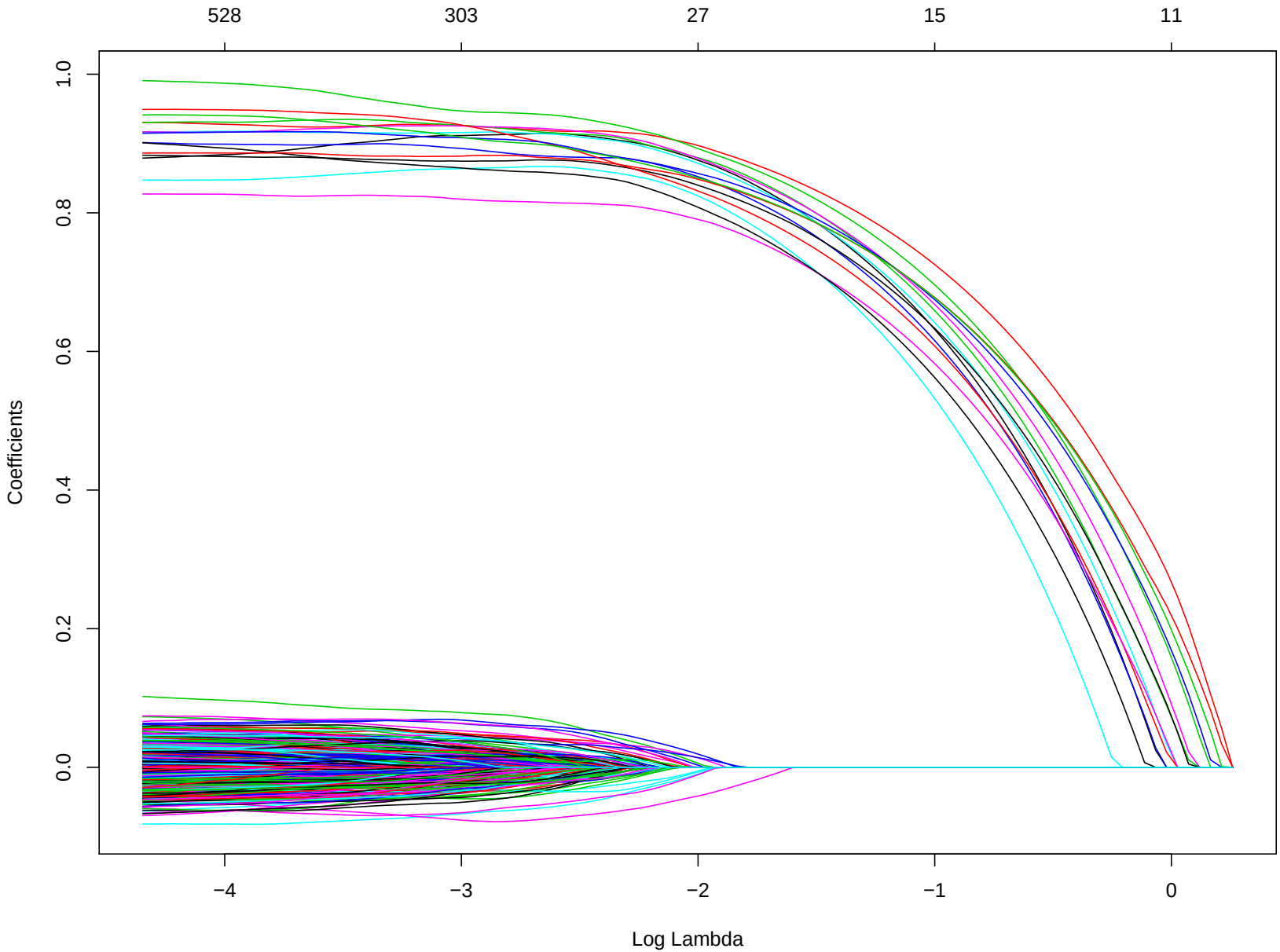
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|=\mathrm{RSS}(\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}\left|\beta_j\right|,$$

- On parle de pénalité ℓ^1 au lieu de pénalité ℓ^2 (ou quadratique)

Le Lasso (suite)

- Comme pour la régression ridge, le Lasso tire les estimations des coefficients vers 0.
- Cependant, dans le cas du Lasso, la pénalité ℓ^1 a pour effet de forcer certains coefficients à s'annuler lorsque λ est suffisamment grand.
- Donc, le Lasso permet de faire de la *sélection de variable*.
- On parle de modèle creux (sparse), c'est-à-dire de modèles qui n'impliquent qu'un sous ensemble des variables.
- Comme pour la régression ridge, choisir une bonne valeur de λ est critique. Procéder par validation ou validation croisée.

Exemple : $n = 1000$ et $p = 5000$



Qu'est qui fait marcher le Lasso ?

Avec les multiplicateurs de Lagrange, on peut voir

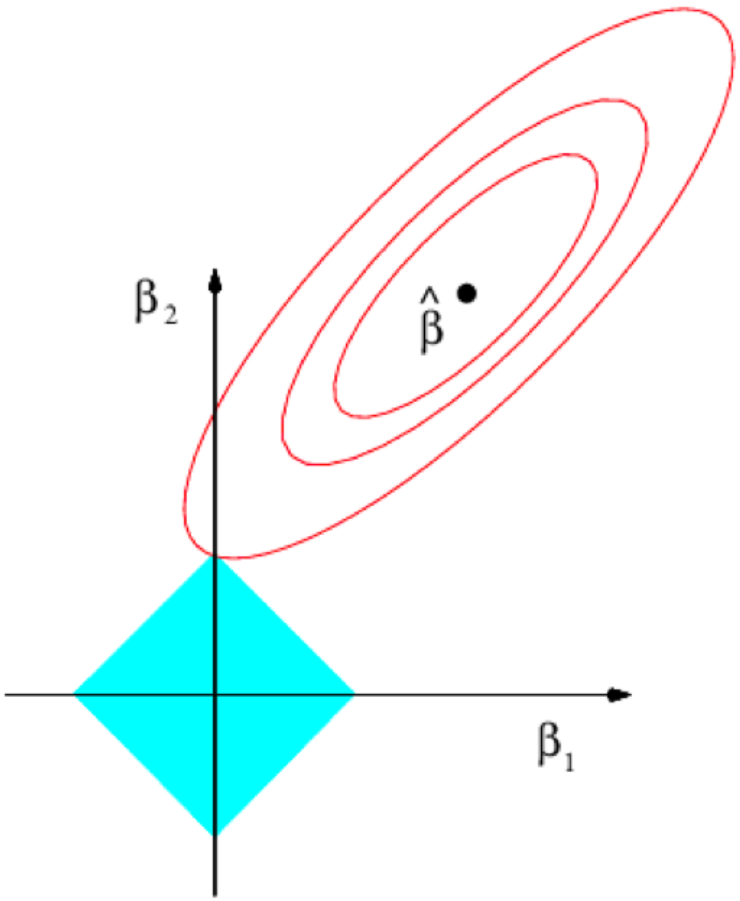
- La régression ridge comme

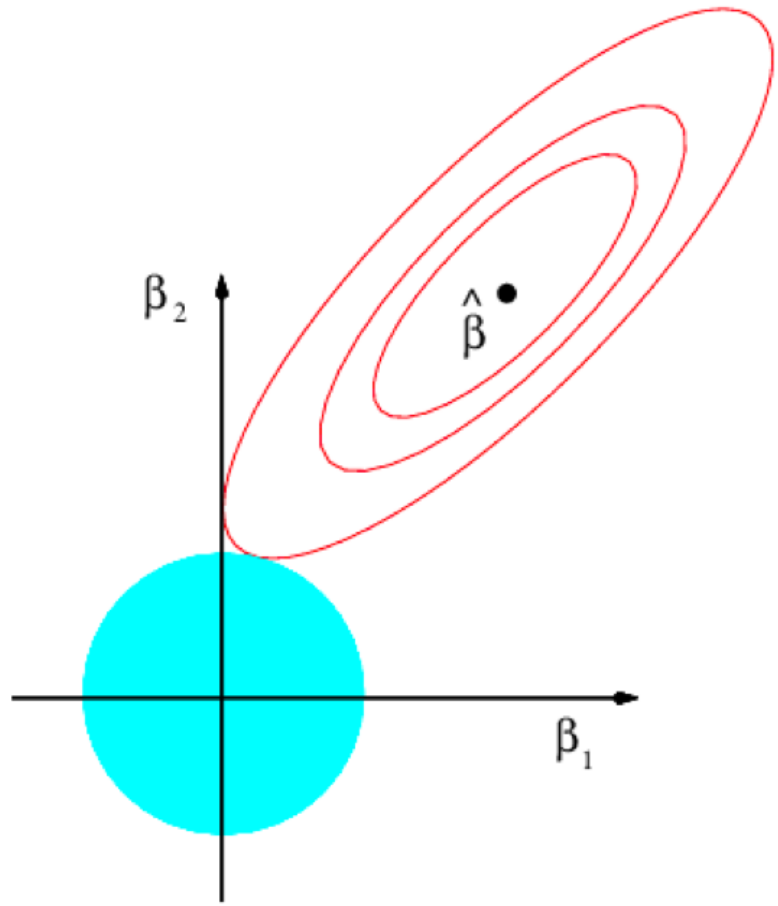
$$\text{minimise} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ sous la contrainte } \sum_{j=1}^{p} \beta_j^2 \leq s$$

- Le Lasso comme

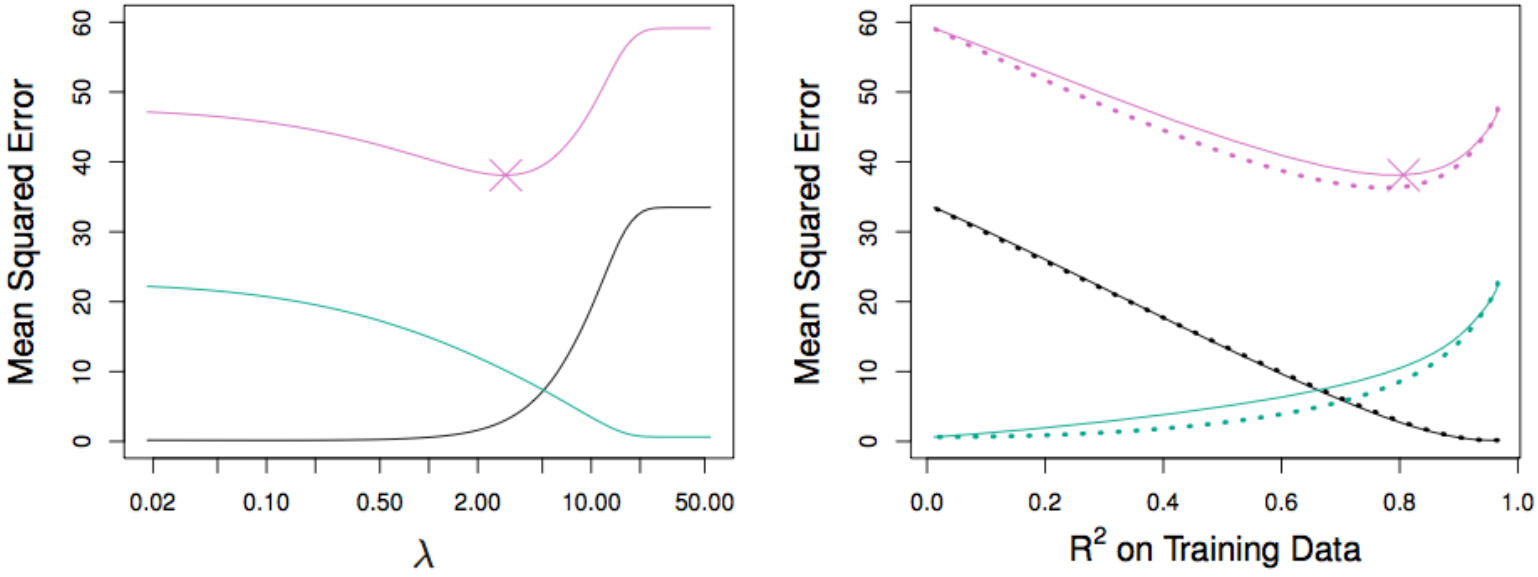
$$\text{minimise} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ sous la contrainte } \sum_{j=1}^{p} |\beta_j| \leq s$$

Le Lasso en image





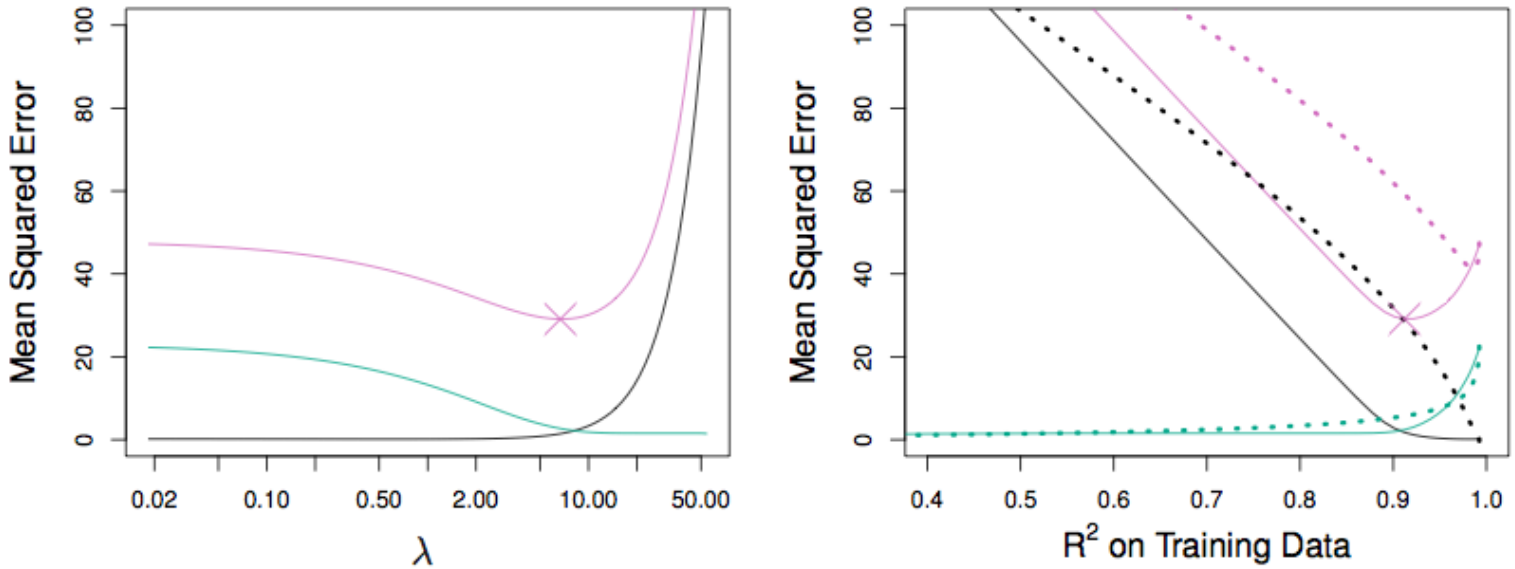
Comparaison du Lasso et de la régression ridge



À gauche, biais au carré (noir), variance (en vert) et erreur quadratique de test (violet) pour le Lasso sur données simulées.

À droite, comparaison du biais au carré, de la variance et de l'erreur de test quadratique pour le Lasso (traits plains) et la régression ridge (pointillés)

Comparaison du Lasso et de la régression ridge (suite)



À gauche, biais au carré (noir), variance (en vert) et erreur quadratique de test (violet) pour le Lasso sur données simulées (où seulement deux prédicteurs sont influents).

À droite, comparaison du biais au carré, de la variance et de l'erreur de test quadratique pour le Lasso (traits plains) et la régression ridge (pointillés)

Conclusions

- Ces deux exemples montrent qu'il n'y a pas de meilleur choix universel entre la régression ridge et le Lasso.
- En général, on s'attend à ce que le Lasso se comporte mieux lorsque la réponse est une fonction d'un nombre relativement faible de prédicteurs.
- Cependant, le nombre de prédicteurs reliés à la réponse n'est jamais connu *a priori* dans des cas concrets.
- Une technique comme la validation croisée permet de déterminer quelle est la meilleure approche.

Choisir le paramètre de réglage λ

- Comme pour les méthodes du début, la régression ridge et le Lasso doivent être calibré pour déterminer le meilleur modèle.
- C'est-à-dire qu'il faut une méthode qui choisisse une valeur du paramètre de réglage λ, ou de la contrainte s.
- La *validation croisée* fournit une façon simple d'attaquer ce problème. On fixe une grille de valeurs de λ possible et sur cette grille, on estime l'erreur de test par validation croisée.
- On choisit alors la valeurs de λ pour laquelle cette estimation de l'erreur de test est la plus faible.
- Enfin, le modèle est ré-ajusté pour utiliser toutes les observations de la base d'entrainement avec la valeur de λ précédemment obtenue.

Le zoo des méthodes lasso I

- The eslaticnet *Zou et Hastie 2005* vise à activer les variables corrélées simultanément

$$\widehat{\boldsymbol{\beta}}^{\text{e-net}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\text{RSS}(\boldsymbol{\beta}) + \lambda\big(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2\big) \right\}$$

- Adaptive/Weighted-Lasso pondère chaque composante du vecteur de coefficients

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\text{RSS}(\boldsymbol{\beta}) + \lambda\|\mathbf{w}\circ\boldsymbol{\beta}\|_1 \right\}.$$

Le zoo des méthodes lasso II

- Group-Lasso *Yuan and Lin 2006* vise à activer les variables par groupes

$$\widehat{\boldsymbol{\beta}}^{\text{group}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\text{RSS}(\boldsymbol{\beta}) + \lambda \sum_{k=1}^{K} w_k \|\boldsymbol{\beta}_{\mathcal{G}_k}\|_1 \right\}$$

- Cooperative-Lasso *Chiquet et al. 2010* vise à activer les variables par groupes de même signe

$$\widehat{\boldsymbol{\beta}}^{\text{coop}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\text{RSS}(\boldsymbol{\beta}) + \lambda \sum_{k=1}^{K} w_k \left(\|\boldsymbol{\beta}^+_{\mathcal{G}_k}\|_1 + \|\boldsymbol{\beta}^-_{\mathcal{G}_k}\|_1 \right) \right\}$$

Bilan

- Les méthodes de sélection de modèles sont essentielles pour l'analyse de données, et l'apprentissage statistique, en particulier avec de gros jeu de données contenant de nombreux prédicteurs.
- Les questions de recherches qui donnent des solutions creuses (parcimonieuses, ou sparses), comme le Lasso, sont d'actualité.

Classification

Régression logistique

Notons $p(X) = \mathbb{P}(Y = 1|X)$ et considérons un seul prédicteur X. La régression logistique pose

$$p(X) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}$$

qui est toujours entre 0 et 1 ! On a alors

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

(transformation logit)

Estimation par maximum de vraisemblance La vraisemblance

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} [1-p(x_i)]$$

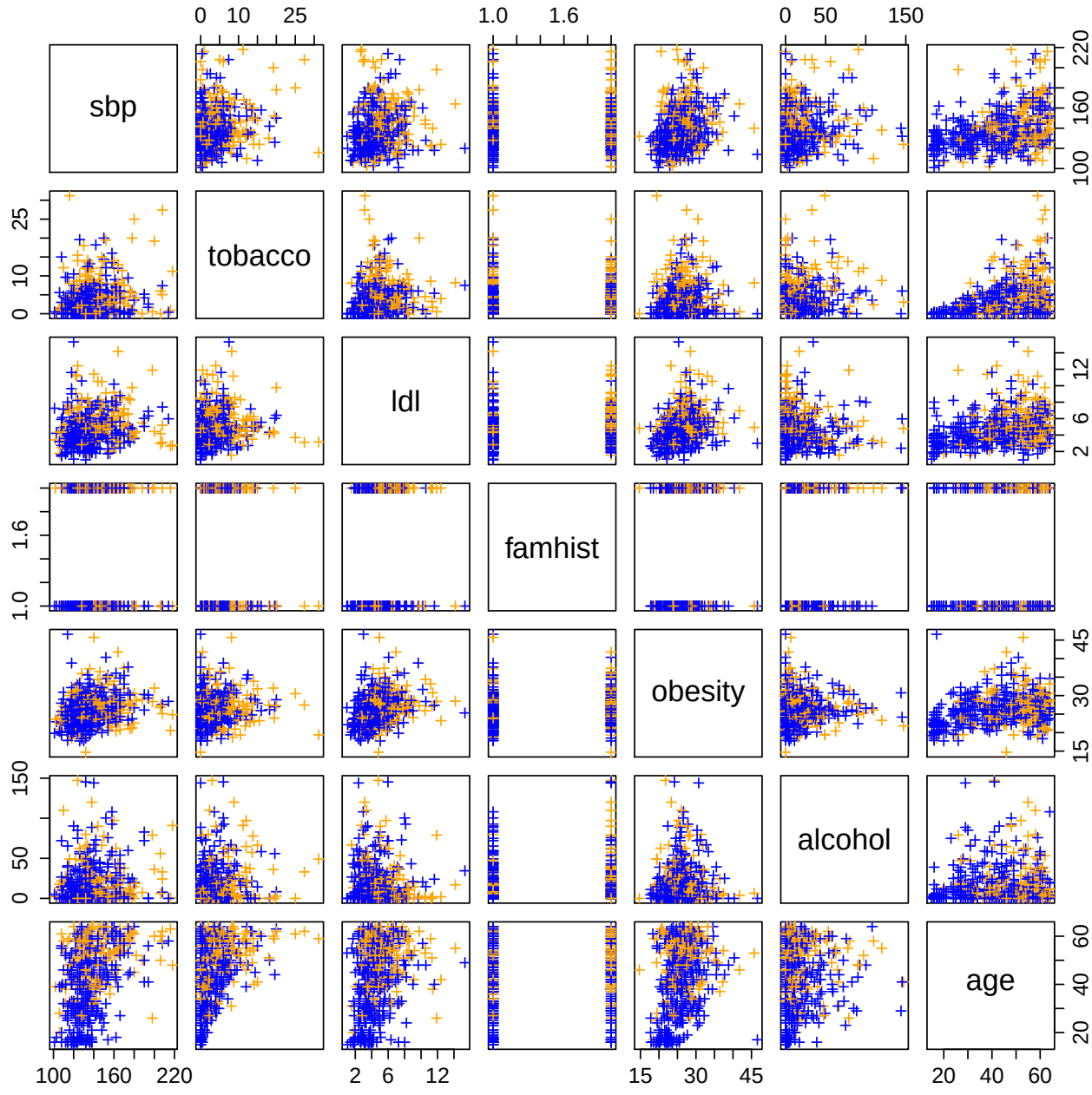
que l'on maximise pour obtenir β_0 et β_1 (Ordinateur)
La plupart des logiciels de statistique le font (glm de R par exemple)

Régression logisitque à plusieurs co-variables

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Exemple : maladie cardiaque en Afrique du Sud

- 160 cas d'infarctus du myocarde (MI) et 302 cas de contrôle (homme entre 15-64 ans), de la province de Cap-Occidental en Afrique du Sud, au début des années 80
- Prévalence très élevée dans cette région : 5.1 %
- Mesure de 7 prédicteurs (facteurs de risque), montrés dans la page suivante
- Le but est d'identifier l'influence et la force relative des facteurs de risque
- Cette étude fait partie d'un programme de santé publique dont le but était de sensibiliser la population sur une régime plus équilibré



orange : MI
bleu : contrôle
`famhist` : 1 si antécédents familiaux

Exemple (suite)

```
> heartfit <- glm(chd ~ .,data=heart ,family=binomial)
> summary(heartfit)

Call:
glm(formula = chd ~ ., family = binomial, data = heart)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp              0.0057607  0.0056326   1.023  0.30643
tobacco          0.0795256  0.0262150   3.034  0.00242 **
ldl              0.1847793  0.0574115   3.219  0.00129 **
famhistPresent   0.9391855  0.2248691   4.177 2.96e-05 ***
obesity         -0.0345434  0.0291053  -1.187  0.23529
alcohol          0.0006065  0.0044550   0.136  0.89171
age              0.0425412  0.0101749   4.181 2.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
```

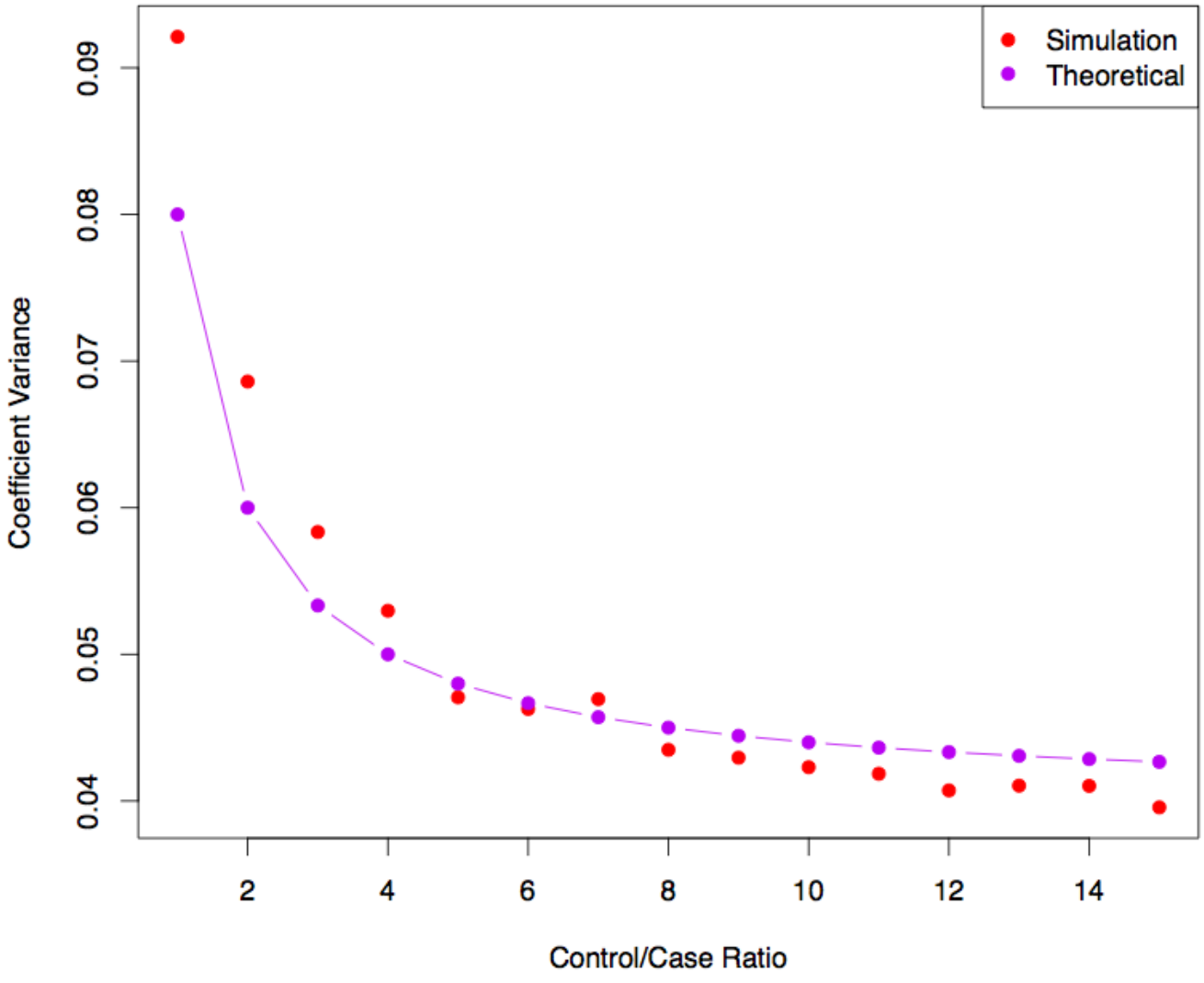
Échantillonnage du contrôle et régression logistique

- Dans les données d'Afrique du Sud, il y a 160 MI et 302 contrôle — $\tilde{\pi} = 0.35$ des cas. Cependant, la prévalence des MI dans la région est de $\pi = 0.05$.
- Ce biais d'échantillonnage permet d'estimer les β_j, $j \neq 0$, avec plus de précision (si modèle correct). Mais l'estimation de β_0 doit être corrigée.
- Une simple transformation permet de le faire :

$$\widehat{\beta_0}^* = \widehat{\beta}_0 + \log\left(\frac{\pi}{1-\pi}\right) - \log\left(\frac{\tilde{\pi}}{1-\tilde{\pi}}\right)$$

- Souvent, les cas pathologiques sont rares et on les prend tous. On peut sur-échantillonner jusqu'à 5 fois plus que les cas témoins. Au delà, peu de gain dans la variance d'erreur d'échantillonnage.

Gain de variance par biais d'échantillonnage de données binaires



Au delà d'un facteur 5 de sur-représentation des cas pathologiques, le gain n'est plus intéressant.

Régression logistique à plus de deux modalités

Jusqu'à maintenant, nous avons discuté de régression logistique pour expliquer un Y à deux modalités. Il est facile de généraliser à plus de deux classes. Une possibilité (utilisée dans la bibliothèque `glmnet` de R) est la forme symétrique

$$\mathbb{P}(Y = k|X) = \frac{\exp(\beta_{0k} + \beta_{1k}X_1 + \cdots + \beta_{pk}X_p}{\sum_{\ell=1}^{K} \exp(\beta_{0\ell} + \beta_{1\ell}X_1 + \cdots + \beta_{p\ell}X_p}$$

Il y a donc une fonction linéaire par classe ou modalité.
En fait, ce modèle est sur-paramétré, et comme dans le cas de 2 classes, on peut supprimer l'une des fonctions linéaires et seules $(K-1)$ sont utiles. *Le vérifier !*
La régression logistique multi-classe porte plusieurs noms. On parle parfois de régression multinomiale.

Sélection de modèles

Revenons à l'expression

$$\underset{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\text{minimiser}} \left\{ -\frac{1}{N}\ell\big(\beta_0,\boldsymbol{\beta}\big) + \lambda P_\alpha\big(\boldsymbol{\beta}\big) \right\}$$

- $\ell\big(\beta_0,\boldsymbol{\beta}\big) = \frac{1}{2\sigma^2}\|y - \beta_0\mathbf{1} - X\boldsymbol{\beta}\|_2^2 + c$ pour une régression linéaire multiple.
- Dans le cas d'une régression logistique

$$\ell\big(\beta_0,\boldsymbol{\beta}\big) = \sum_{i=1}^{N} \left\{ y_i\big(\beta_0 + \boldsymbol{\beta}' x_i\big) - \log\left(1 + e^{\beta_0+\boldsymbol{\beta}' x_i}\right) \right\}$$

- La terme de régularisation (de pénalité)

$$P_\alpha\big(\boldsymbol{\beta}\big) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2$$

Lasso si $(\alpha = 1)$ ridge si $\alpha = 1$.

Deux exemples en grande dimension

Allons faire des testes sous R

- Jeu de données leukemia du package spikeslab : les gênes exprimés différentiellement exprimés pour les deux types de leucémie *Acute Myeloblastic Leukemia* et *Acute Lymphoblastic Leukemia*. $n = 72$ et $p = 3571$.
- Jeu de données nki du package BreastCancerNKI de bioconductor : déterminer les gênes exprimés sous la condition *estrogen receptor status* actif. $n = 337$ et $p = 24481$