Le modèle de régression linéaire

Master 2 IES

Ana Karina Fermin

Université Paris Nanterre

aferminrodriguez@parisnanterre.fr

D'un point de vue pratique l'objectif est double.

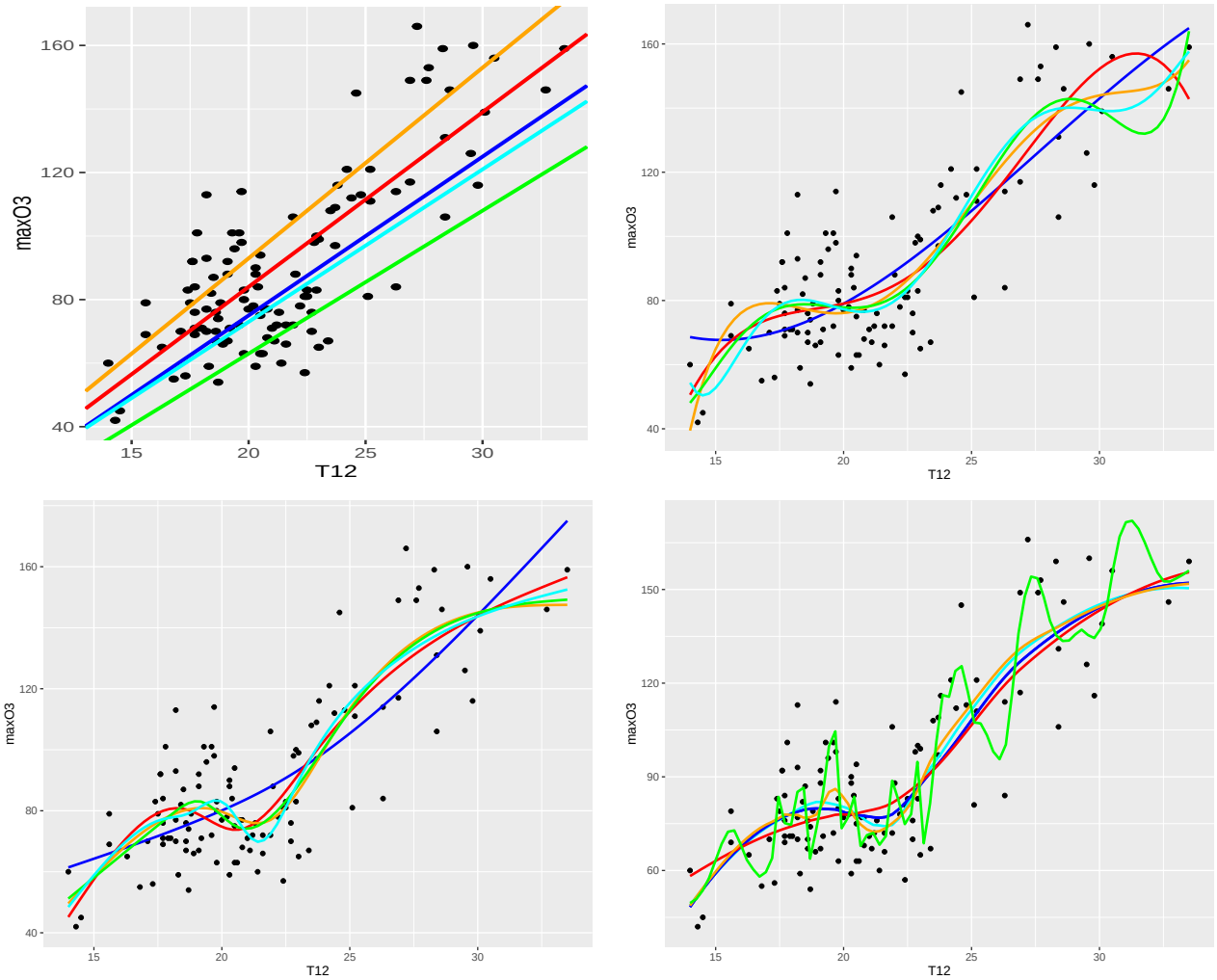
- Ajuster un modèle pour prédire Y en fonction de $\mathbf{X}$
- Prédire les valeurs de Y pour de nouvelles valeurs de $\mathbf{X}$.

Données ozone

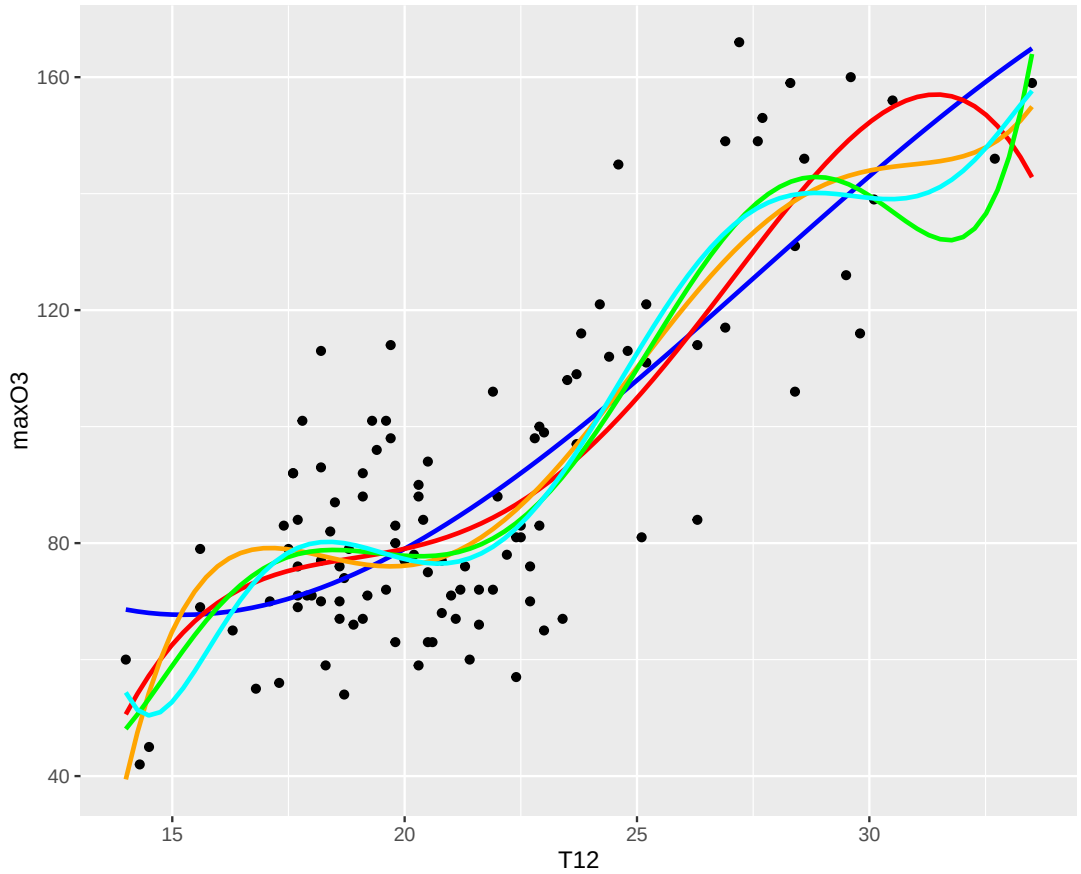
Nous commençons toujours par voir et représenter les données !

```
112 obs. of  13 variables:
maxO3 : int  87 82 92 114 94 80 79 79 101 106 ...
T9    : num  15.6 17 15.3 16.2 17.4 17.7 16.8 14.9 16.1 18.3 ...
T12   : num  18.5 18.4 17.6 19.7 20.5 19.8 15.6 17.5 19.6 21.9 ...
T15   : num  18.4 17.7 19.5 22.5 20.4 18.3 14.9 18.9 21.4 22.9 ...
Ne9   : int  4 5 2 1 8 6 7 5 2 5 ...
Ne12  : int  4 5 5 1 8 6 8 5 4 6 ...
Ne15  : int  8 7 4 0 7 7 8 4 4 8 ...
Vx9   : num  0.695 -4.33 2.954 0.985 -0.5 ...
Vx12  : num  -1.71 -4 1.879 0.347 -2.954 ...
Vx15  : num  -0.695 -3 0.521 -0.174 -4.33 ...
maxO3v: int  84 87 82 92 114 94 80 99 79 101 ...
vent  : Factor w/ 4 levels "Est","Nord","Ouest",..: 2 2 1 2 3 3 3 2 2 3 ...
pluie : Factor w/ 2 levels "Pluie","Sec": 2 2 2 2 2 1 2 2 2 2 ...
```

Quelle méthode choisir? Linéaire, Polynomiale, Spline, Noyau,....?



Polynomiale : Quel modèle choisir ?



Objectif : Parmi tous les polynômes possibles, déterminer le bon degré du polynôme.

Modèle gaussien de la régression linéaire multiple

On observe des observations bruités

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{id} + \varepsilon_i, \qquad i = 1, \ldots, n$$

avec $\beta_0, \beta_1, \ldots, \beta_d$ inconnus.

On suppose que les ε_i sont les réalisations i.i.d. d'une variable aléatoire gaussienne centrée et de variance σ^2 inconnue.

1. Coefficients estimés (para le méthode de MC) : $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_d)$

$$\widehat{\beta} = (\mathbb{X}^t \mathbb{X})^{-1} \, \mathbb{X}^t Y$$

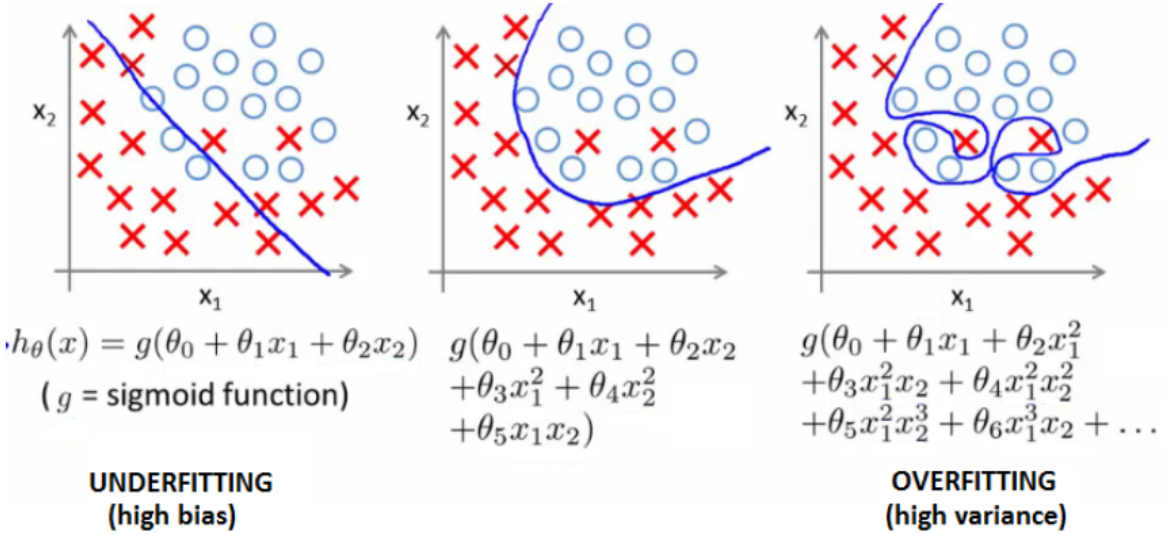
où $\mathbb{X}$ est la matrice augmentée.

Sélection de variables : motivations

1. Description : Quelles sont les variables les plus influentes?
2. On cherche un "bon" modèle : ni trop simple ni trop complexe
3. On recherche donc un modèle parcimonieux.

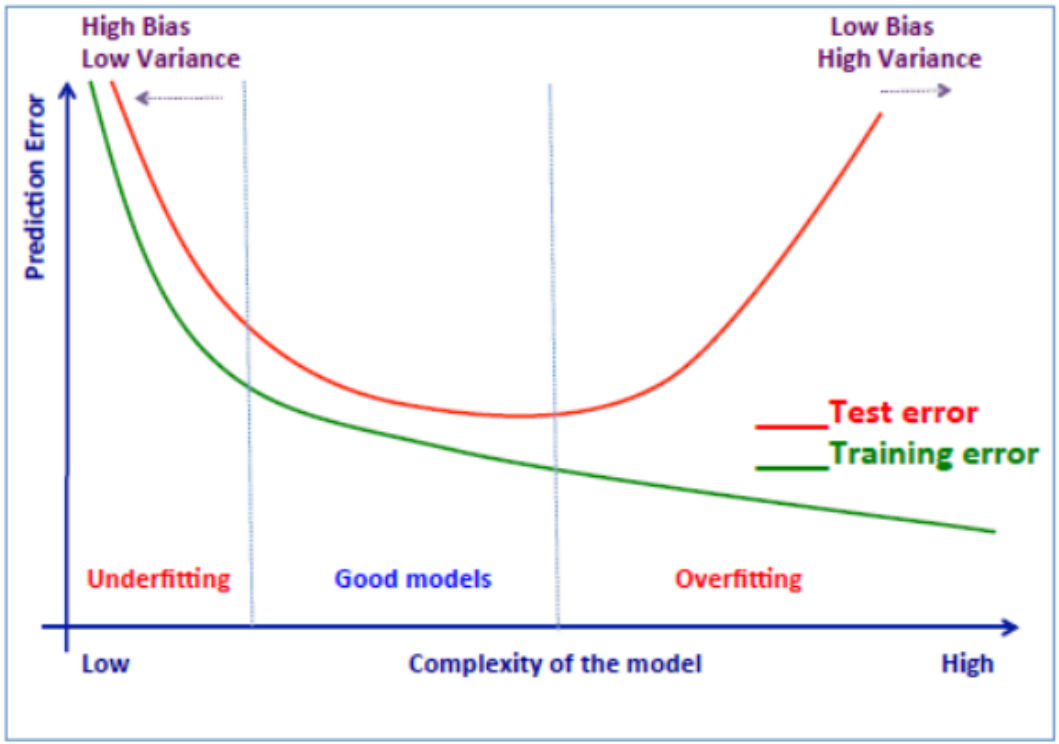
Problème : Sur ajustement/ sous-ajustements

Motivation : Sélection de variables, sélection de modèles



Problème : Sur ajustement/ sous-ajustements

Motivation : Sélection de variables, sélection de modèles



Recherche exhaustive

Démarche :

- On se donne un critère de qualité, qu'on calcule pour tous les sous-modèles comportant un intercept, et on retient le modèle qui optimise le critère.
- Critères :
  - R^2,
  - Par correction
    - R^2-ajusté
    - Cp Mallows
    - AIC, BIC, ...
  - Validation croisée

Estimation par correction

- Les méthodes par correction ajoutent à l'erreur empirique une correction qui dépend de la dimension du modèle et qui "corrige" l'erreur empirique pour être plus proche de l'erreur de prédiction

Estimation par correction

- Risque empirique :

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2$$

Cp de Mallow:

- Premier critère visant à une meilleure estimation de l'erreur de prévision que la seule considération de l'erreur d'ajustement (ou le R^2) dans le modèle linéaire.
- La méthode de Mallows consiste à chercher la fonction que minimise

$$R_n(f) + \frac{2\widehat{\sigma}^2(d+1)}{n}$$

Estimation par correction

Maximum de Vraisemblance (Likelihood)
$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} -\log \mathbb{P}(Y_i|X_i)$

- La méthode AIC cherche la fonction qui minimise

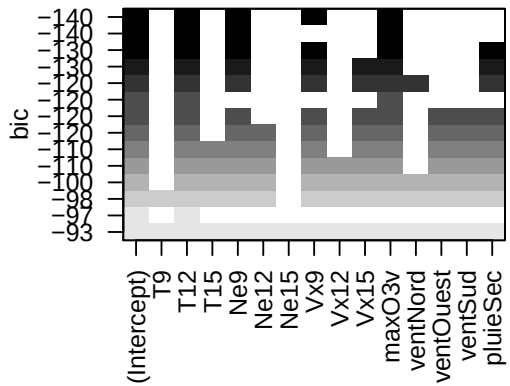
$$R_n(f) + \frac{D_{\mathcal{S}}}{n}$$

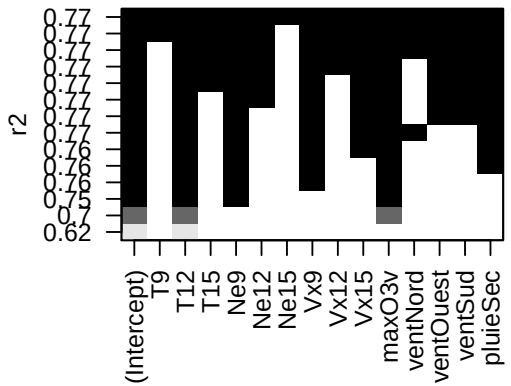
- La méthode BIC (Bayesian) cherche la fonction qui minimise:

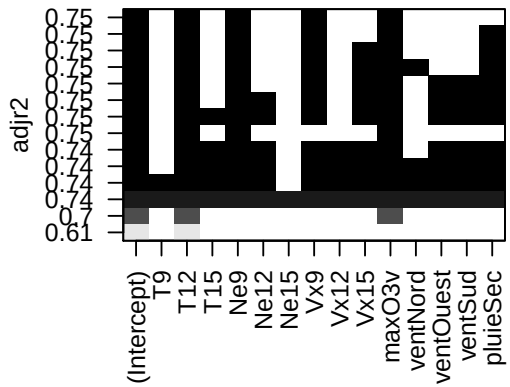
$$R_n(f) + \frac{\log n}{2}\frac{D_{\mathcal{S}}}{n}$$

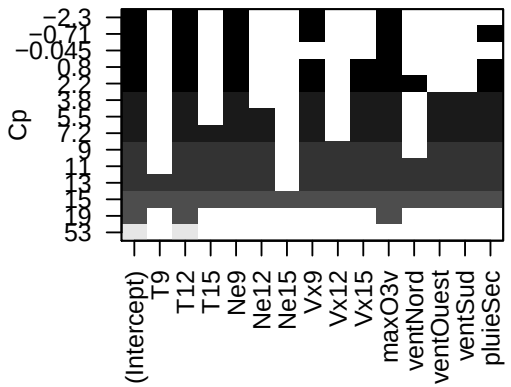
Le meilleur modèle est celui possédant l'AIC ou BIC le plus faible.

Recherche exhaustive









Recherches pas à pas

On les utilise lorsque le modèle complet est trop riche pour permettre facilement une recherche exhaustive. A chaque étape, on ajoute (ou on élimine) la variable la plus (ou la moins) significative, la variable considérée pouvant être fictive.

1. Descendante (Backward) part du modèle complet et élimine les variables une à une.
2. Ascendante (Forward) part du modèle constant et ajoute les variables une à une.
3. Mixte (Stepwise) combine les deux précédantes.

Recherche pas à pas (mixte)

Stepwise (mixte):
On peut aussi donner les directions "backward" ou "forward"

```
Start:  AIC=612.99
maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 +
maxO3v + vent + pluie

Step:  AIC=608.61
maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 +
maxO3v + pluie

.........

Step:  AIC=596.02
maxO3 ~ T12 + Ne9 + Vx9 + maxO3v
```

Sélection de Variable

coefficients du modèle

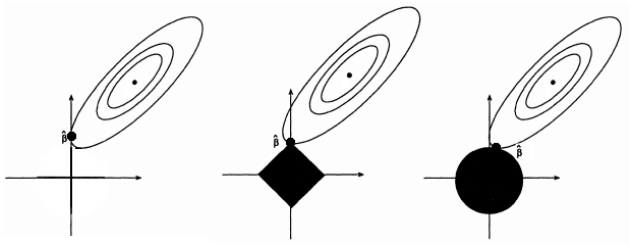
- Coefficients:
  - $\beta_i = 0$ signifie que la $i$ème variable n'est pas utilisé.
  - $\beta_i \sim 0$ signifie que la $i$ème variable est *peu* influente...

- Si certaines variables sont inutiles, il vaut mieux utiliser un modèle plus simple...

Submodels

- *Simplifier* le modèle à partir d'une contrainte sur β!
- Exemples:
  - Imposer $\beta_i = 0$ pour $i \notin I$.
  - Imposer $\|\beta\|_0 = \sum_{i=1}^{d} \mathbf{1}_{\beta_i \neq 0} < C$
  - Imposer $\|\beta\|_p < C$ avec $1 \leqslant p$ (Fréquemment $p = 2$ ou $p = 1$)

Normes et "Sparsity"



Sparsity

- β est parcimonieux (sparse) si le nombre de coefficients non nuls (ℓ_0) est petit...
- Réduction de la dimension/complexité du modèle.

Contraintes et Pénalisation

Optimisation sous contrainte

- Choisir une constante C.
- Calculer β via

$$\underset{\beta\in\mathbb{R}^d,\|\beta\|_p\leqslant C}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - h(\beta^t \mathbf{X}_i))^2$$

Reformulation du Lagrangian

- Choisir λ et calculer β via

$$\underset{\beta\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - h(\beta^t \mathbf{X}_i))^2 + \lambda\|\beta\|_p^{p'}$$

avec $p' = p$ excepte si $p = 0$ avec $p' = 1$.

Estimation par pénalisation

- Les méthodes de régularisation (ou méthodes pénalisées) partent d'un problème d'optimisation dans lequel on cherche à minimiser la somme d'un terme d'erreur entre le réalité et le simulé (comme le moindres carrés ou risque empirique, vraisemblance) et d'une pénalisation

Pénalisation

Modèle linaire pénalisé

- Minimisation de

$$\underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i, h(\beta^t \mathbf{X}_i))^2 + \text{pen}(\beta)$$

- Sélection de variable si β est sparse.

Pénalités classiques

- AIC: $\text{pen}(\beta) = \lambda \|\beta\|_0$ (no convexe / sparsity)
- Ridge: $\text{pen}(\beta) = \lambda \|\beta\|_2^2$ (convex / no sparsity)
  Très utile dans les cas où $\mathbb{X}^t\mathbb{X}$ n'est pas inversible.
- Lasso: $\text{pen}(\beta) = \lambda \|\beta\|_1$ (convexe / sparsity)
- Elastic net: $\text{pen}(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$ (convexe / sparsity)

Pénalités classiques

- Ridge: $\text{pen}(\beta) = \lambda\|\beta\|_2^2$ (convex / no sparsity)
- Lasso: $\text{pen}(\beta) = \lambda\|\beta\|_1$ (convexe / sparsity)
- Elastic net: $\text{pen}(\beta) = \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$ (convexe / sparsity)
- L'optimisation est facile si pen est convexe...
- **On a besoin d'un "bon" λ!**

Dans `R`, il faut utiliser la fonction `glmnet` du package glmnet de la façon suivante

```
glmnet(data, family = "gaussian", alpha = 0)  (Pour Ridge)

glmnet(data, family = "gaussian",alpha = 1) (Pour Lasso)
```

Exemple Ozone : modèle retenu

$$\text{maxO3}_i = \beta_0 + \beta_1 \text{T12}_\text{i} + \beta_2 \text{Vx9}_\text{i} + \beta_3 \text{Ne9}_\text{i} + \beta_4 \text{maxO3v}_\text{i} + \varepsilon_\text{i}$$

```
---
Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
T12          2.76409    0.47450   5.825 6.07e-08 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
maxO3v       0.35483    0.05789   6.130 1.50e-08 ***

Residual standard error: 14 on 107 degrees of freedom
Multiple R-squared: 0.7622,     Adjusted R-squared: 0.7533
F-statistic: 85.75 on 4 and 107 DF,  p-value: < 2.2e-16
```

Choix de Lambda

- On peut utiliser la méthode de la validation croisée ce qui conduit à choisir $\lambda = \lambda_{CV}$.
- On peut utiliser la méthode de la "stability selection" qui consiste à
  1. considérer plusieurs sous-échantillons de Y,
  2. pour chacun de ces sous-échantillons, on applique Lasso avec $\lambda = \lambda_{CV}$
  3. On stocke les indices i tels que $\beta_i \neg 0$
  4. On ne retient que les indices qui au cours des différents sous-échantillonnages apparaissent avec une fréquence de 1.

.