Exemple d'analyse de données

Fabrice Rossi

Université Paris 1

Présentation des données

Contexte

- spectres proche infrarouge d'échantillons de vin
- but pratique : calculer le taux d'alcool dans le vin à partir du spectre

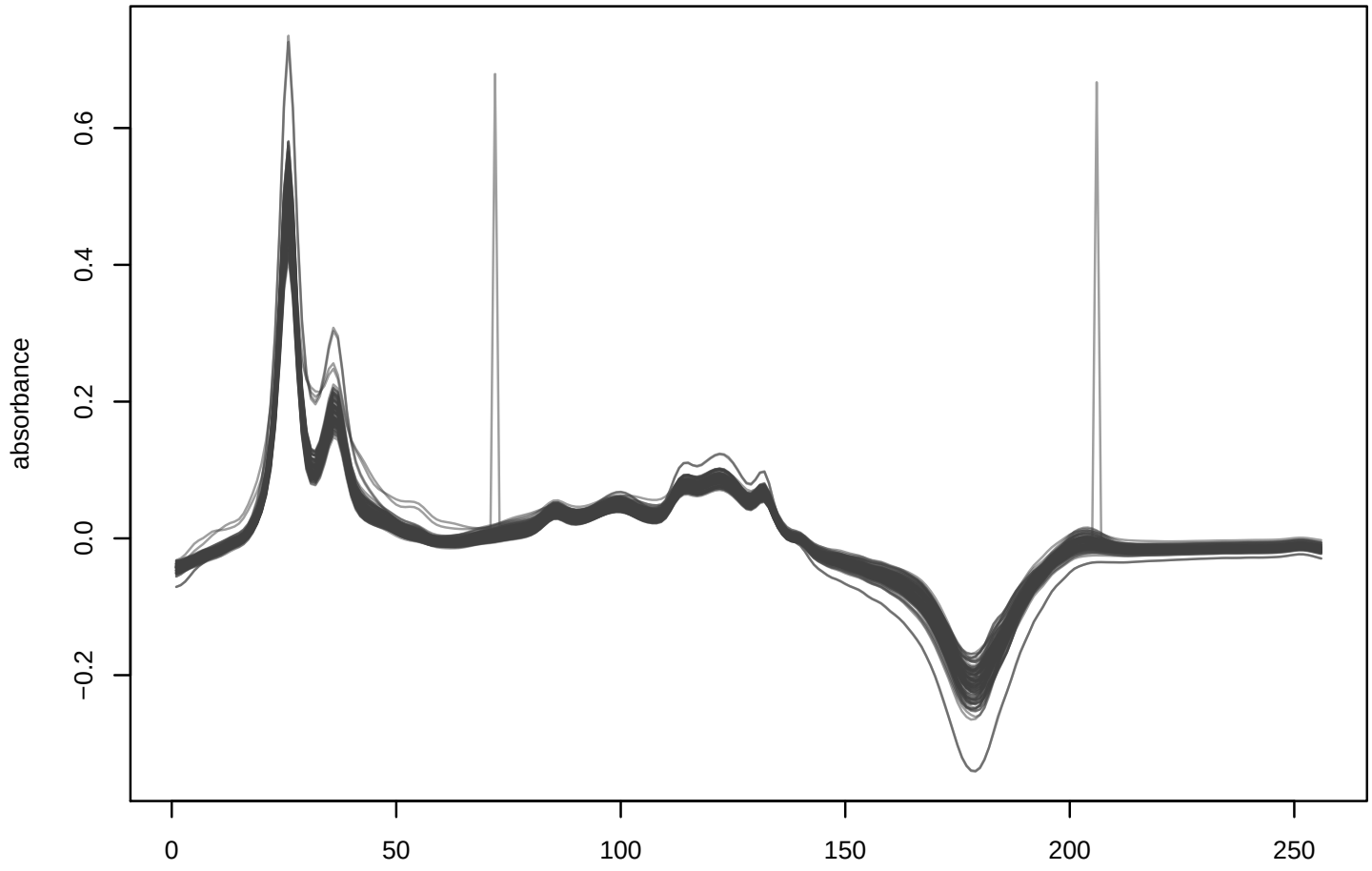
Caractéristiques

- 124 spectres, dont 30 réservés à l'évaluation finale
- 256 variables : nombre d'onde compris entre 400 et 4000 cm^{-1}

Analyse

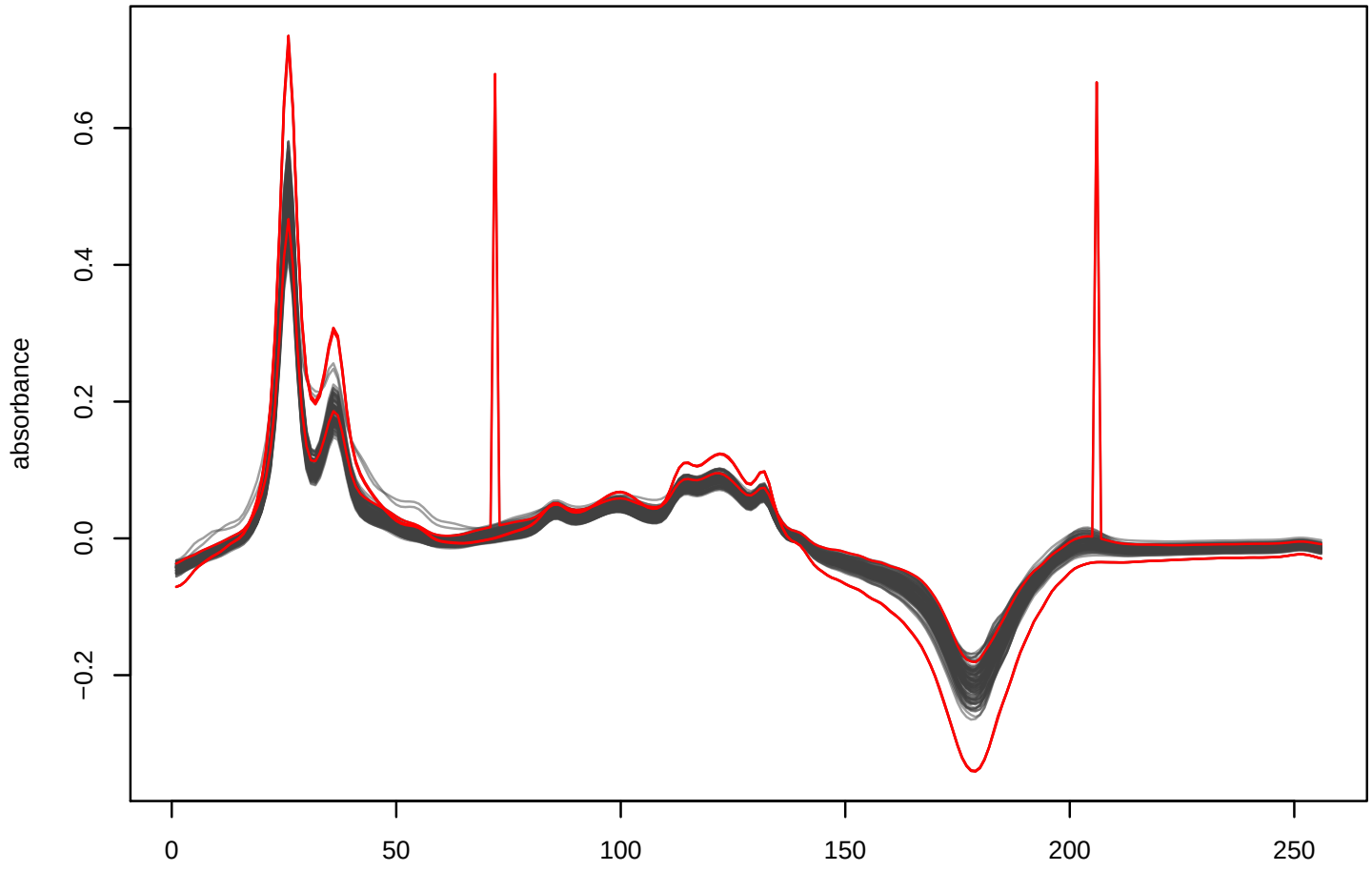
- trop de variables : 256 variables contre 94 objets (pour l'apprentissage)
- modèle linéaire simplifié

Spectres



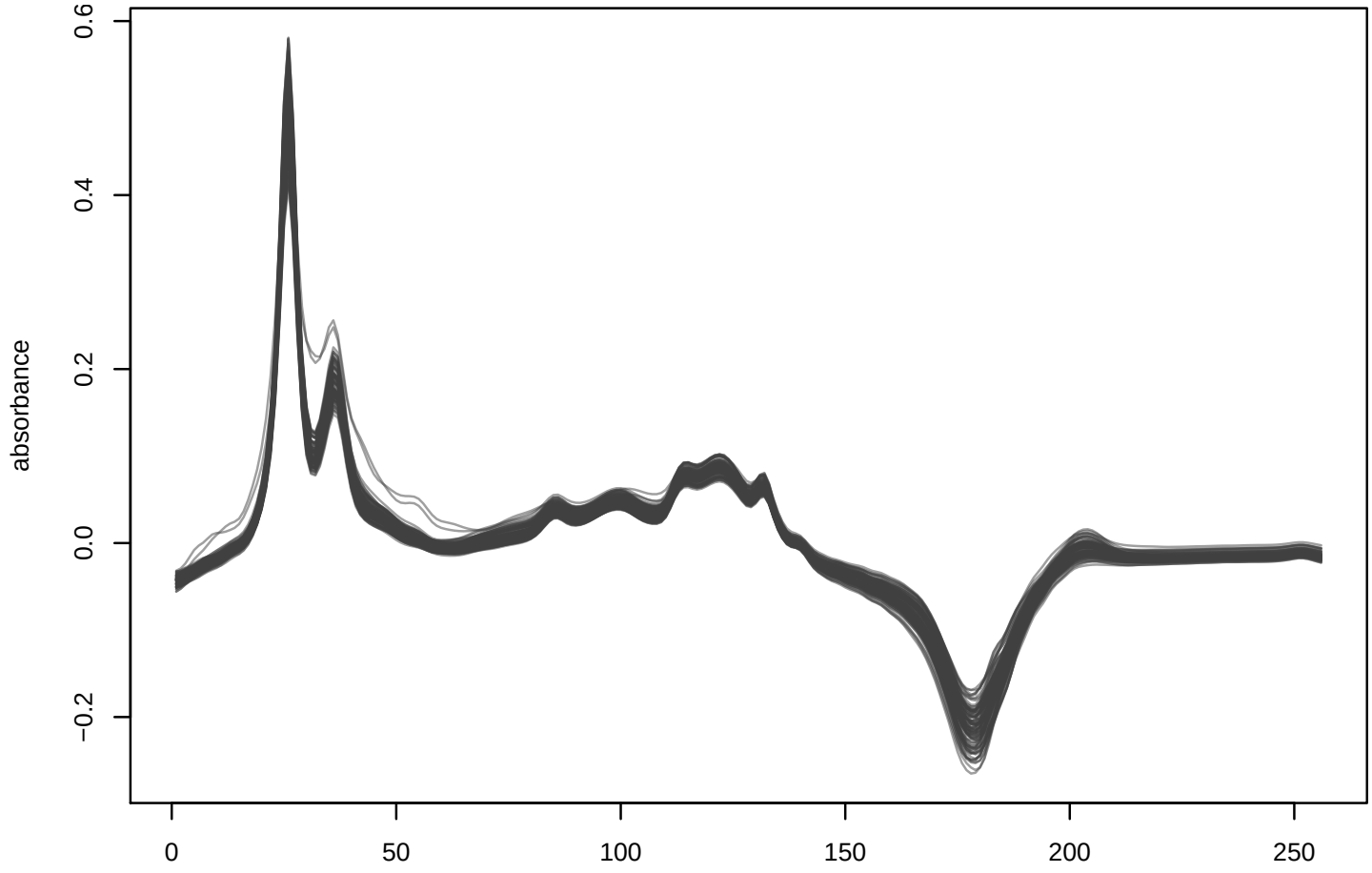
⇒ Quelques spectres atypiques

Spectres



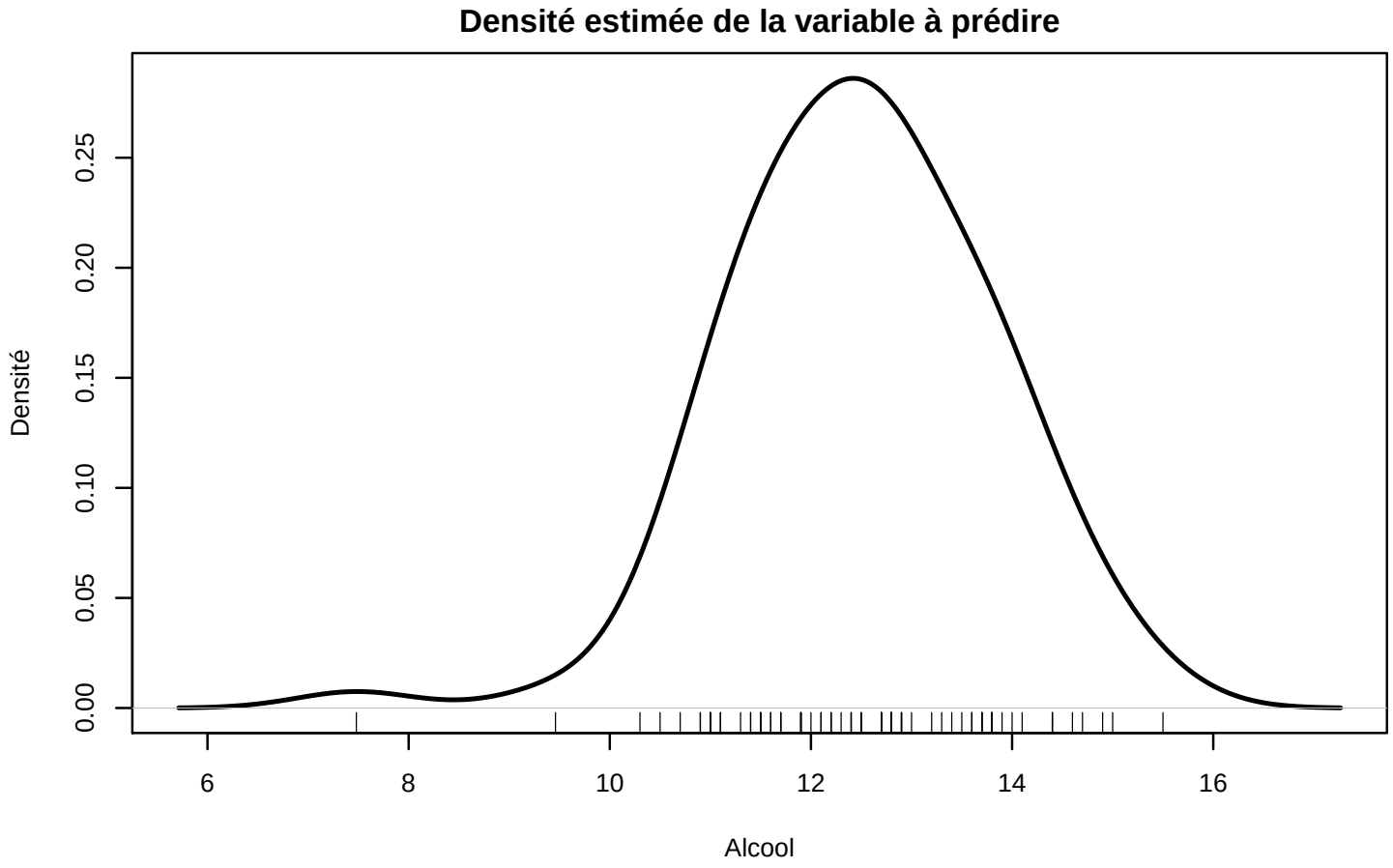
$\Rightarrow$ Trois spectres atypiques

Spectres « propres »



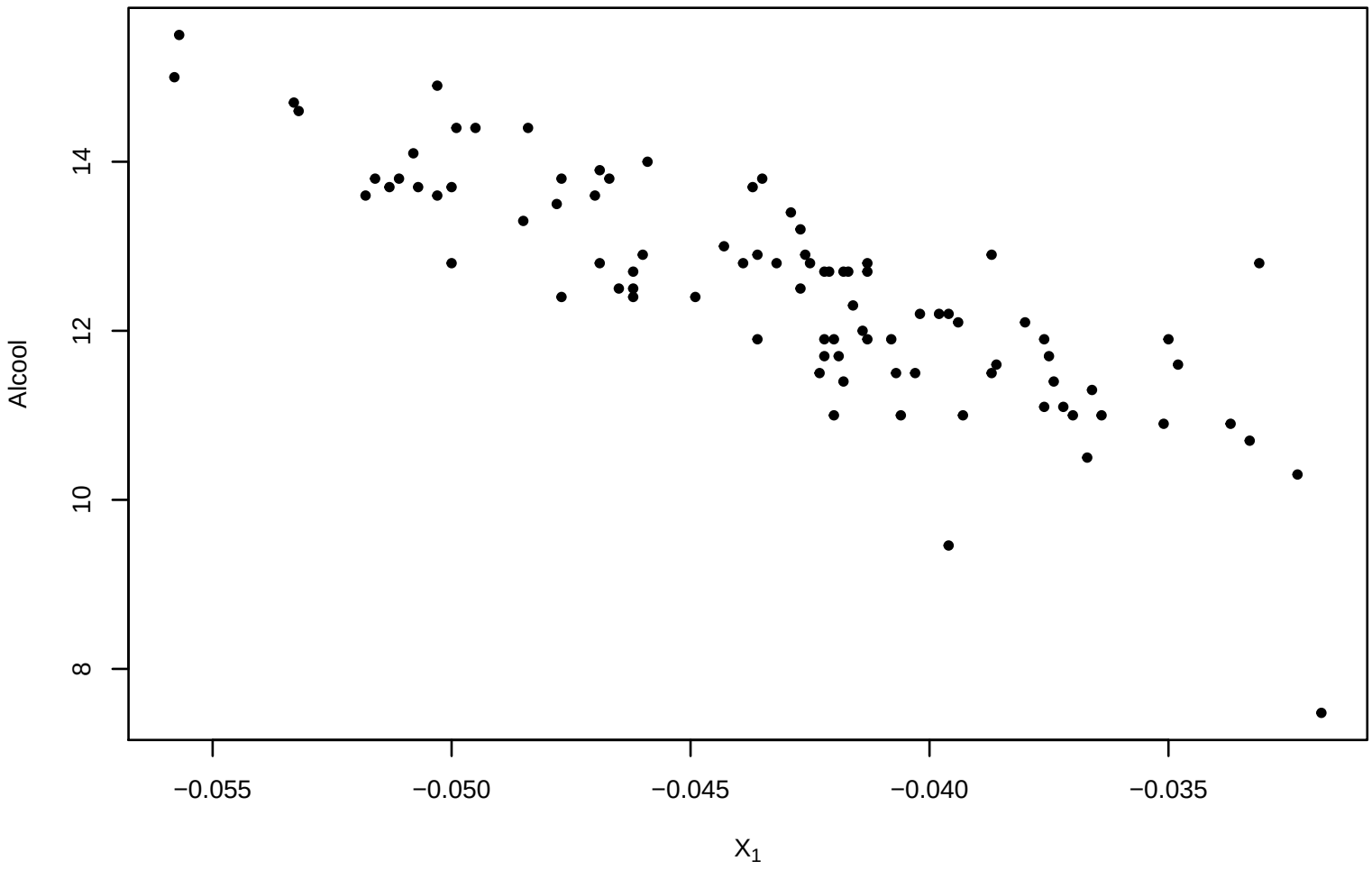
On conserve les spectres atypiques les moins extrêmes

Variable à prédire

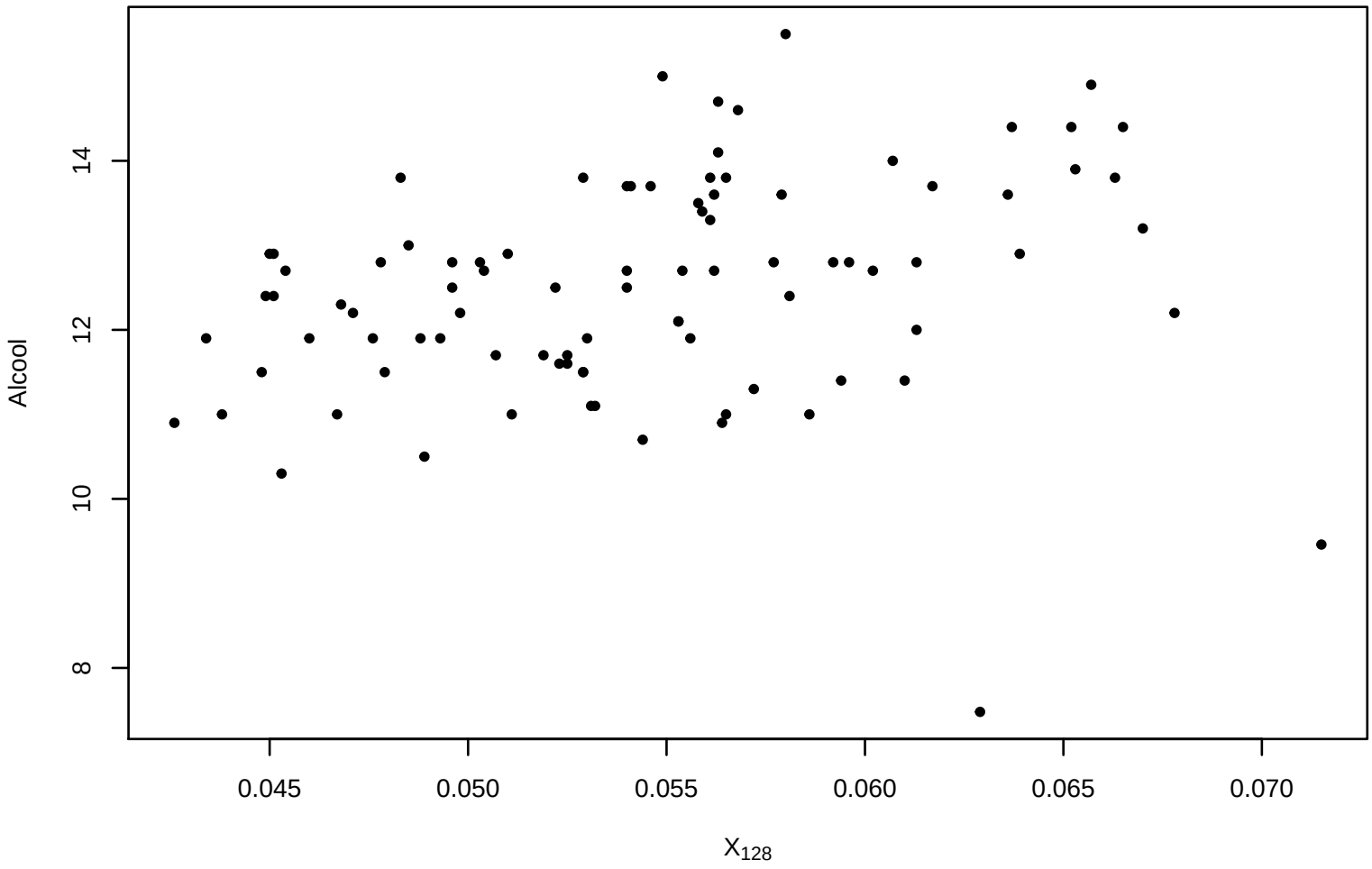


Quelques valeurs extrêmes : peuvent être utiles pour l'apprentissage

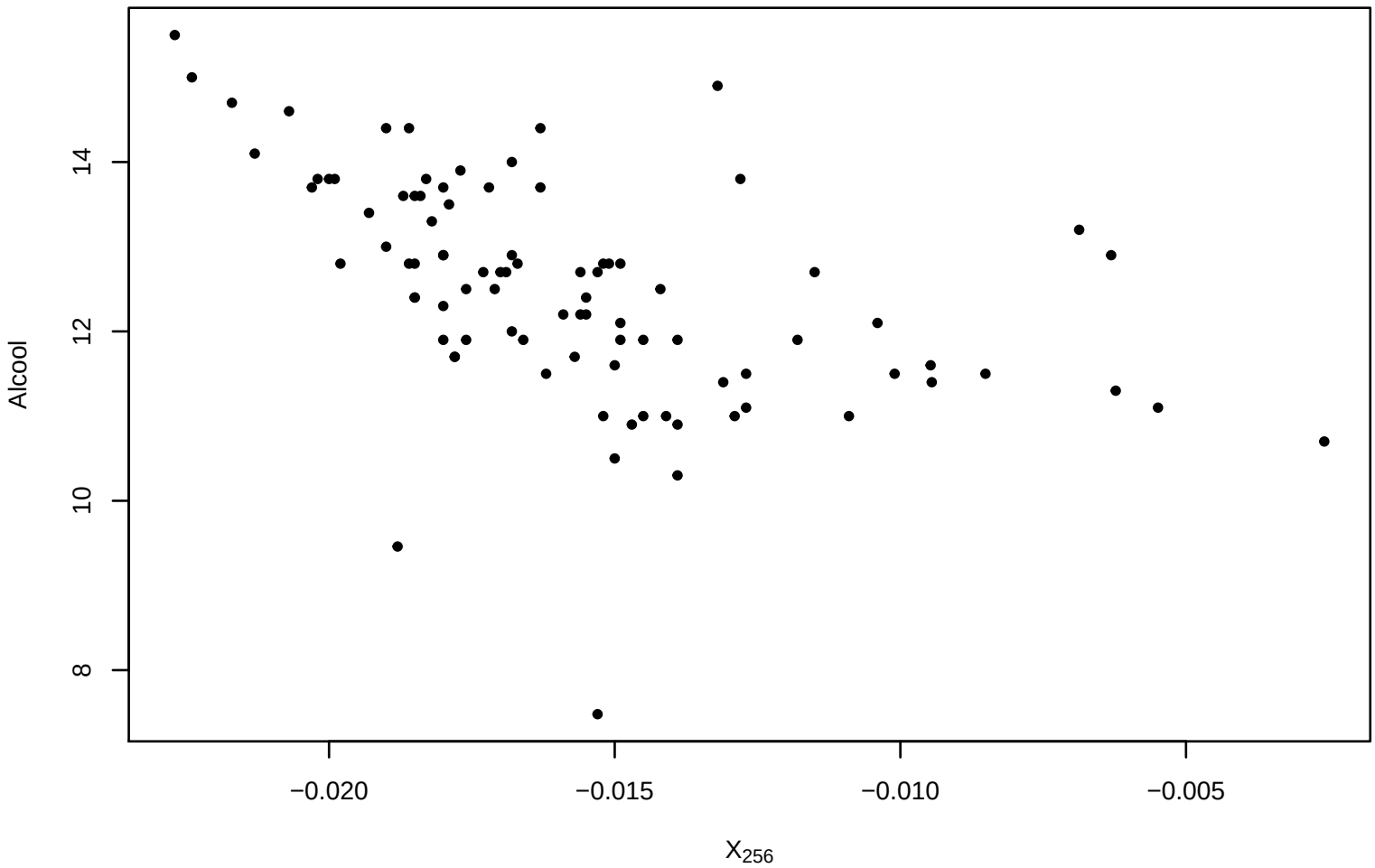
Lien variable à prédire / variables explicatives



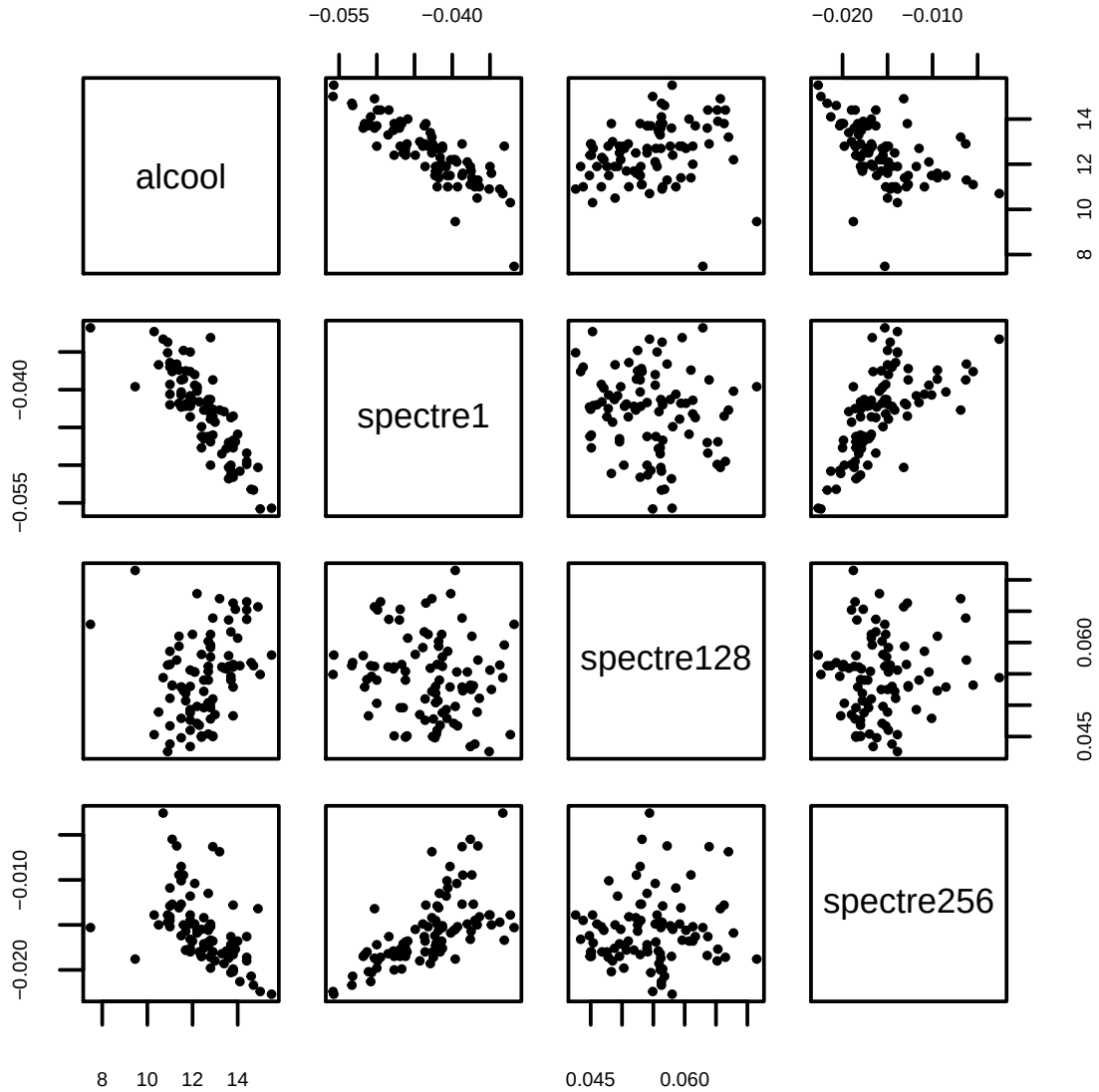
Lien variable à prédire / variables explicatives



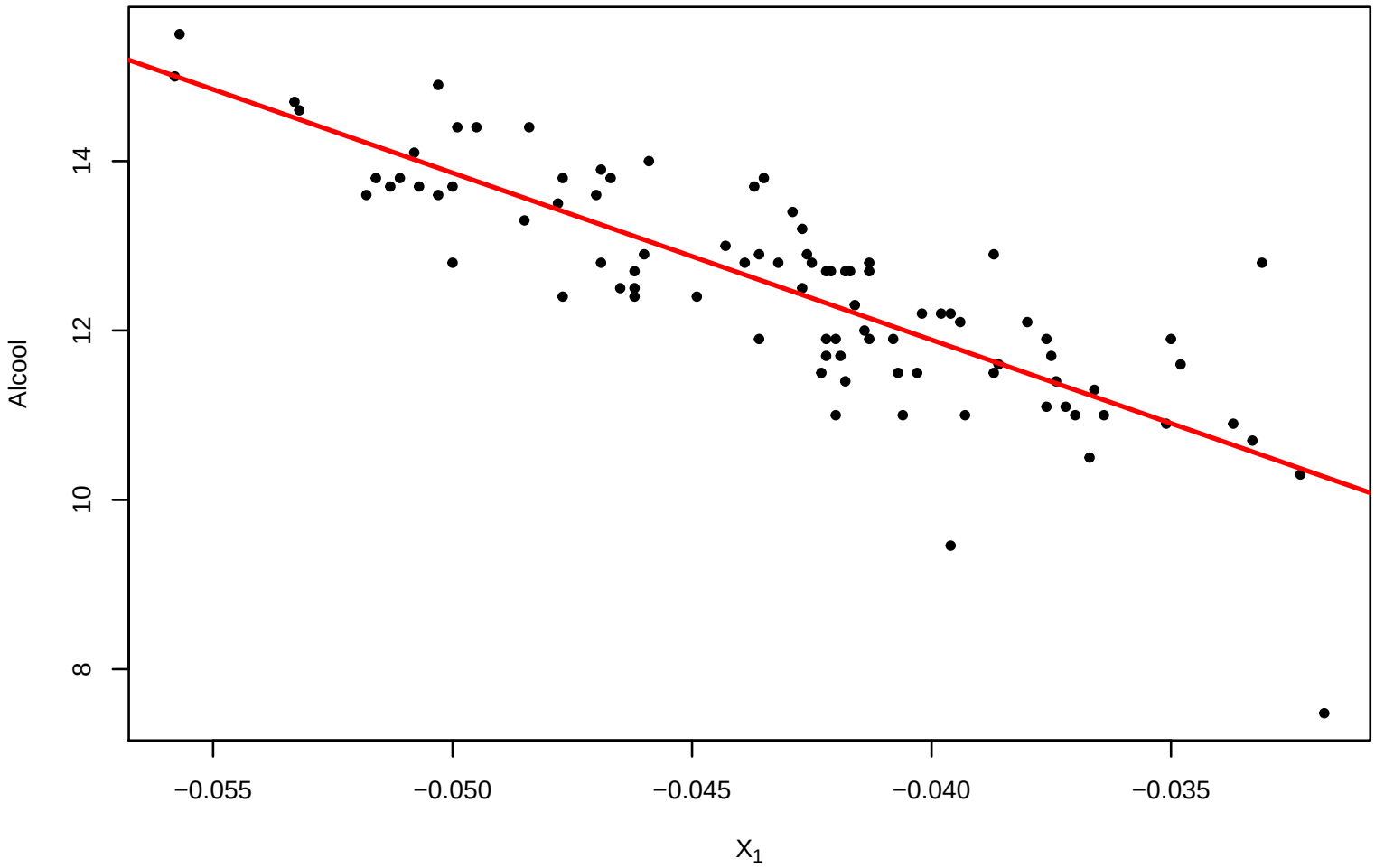
Lien variable à prédire / variables explicatives



Lien variable à prédire / variables explicatives

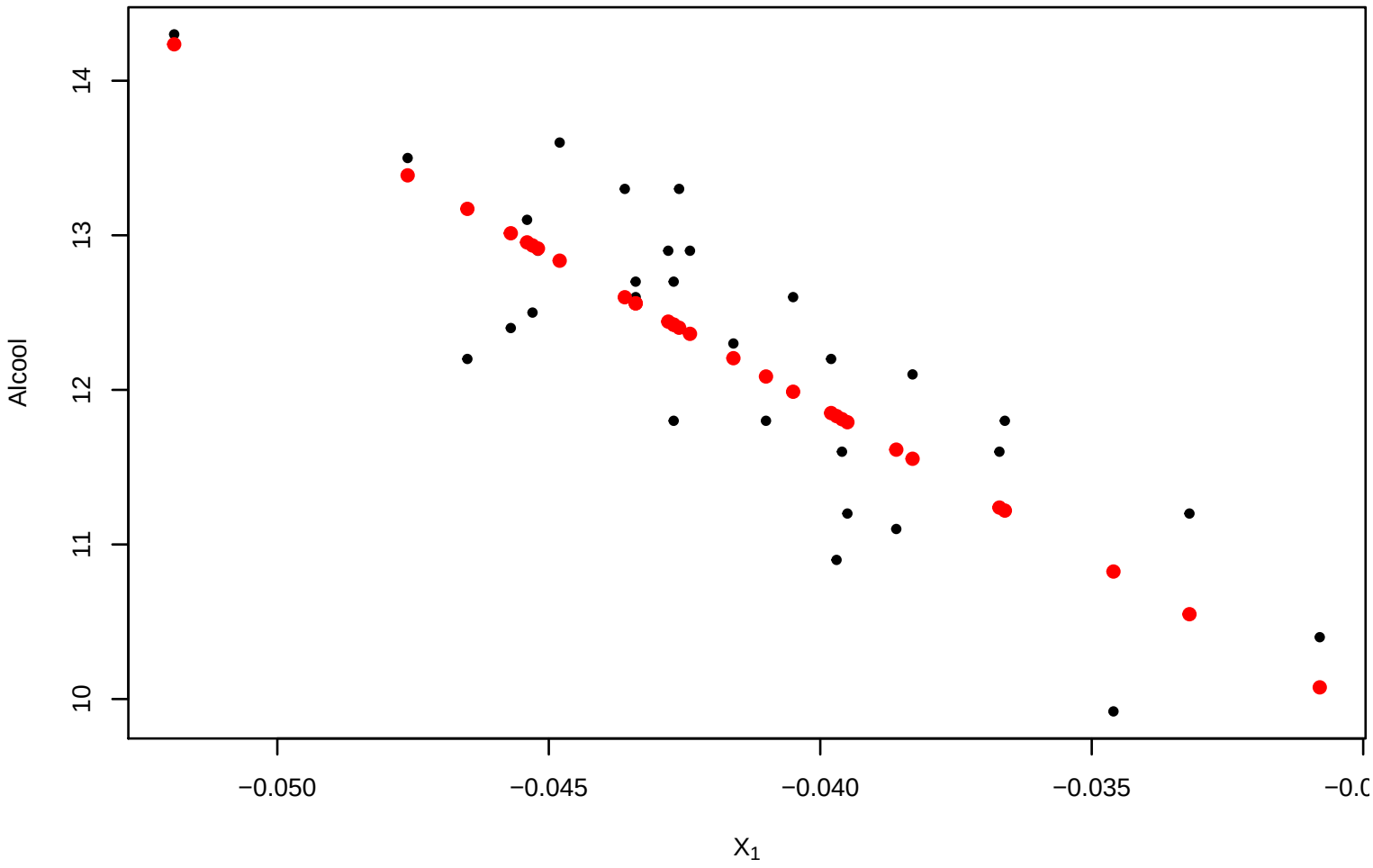


Modèle linéaire simple



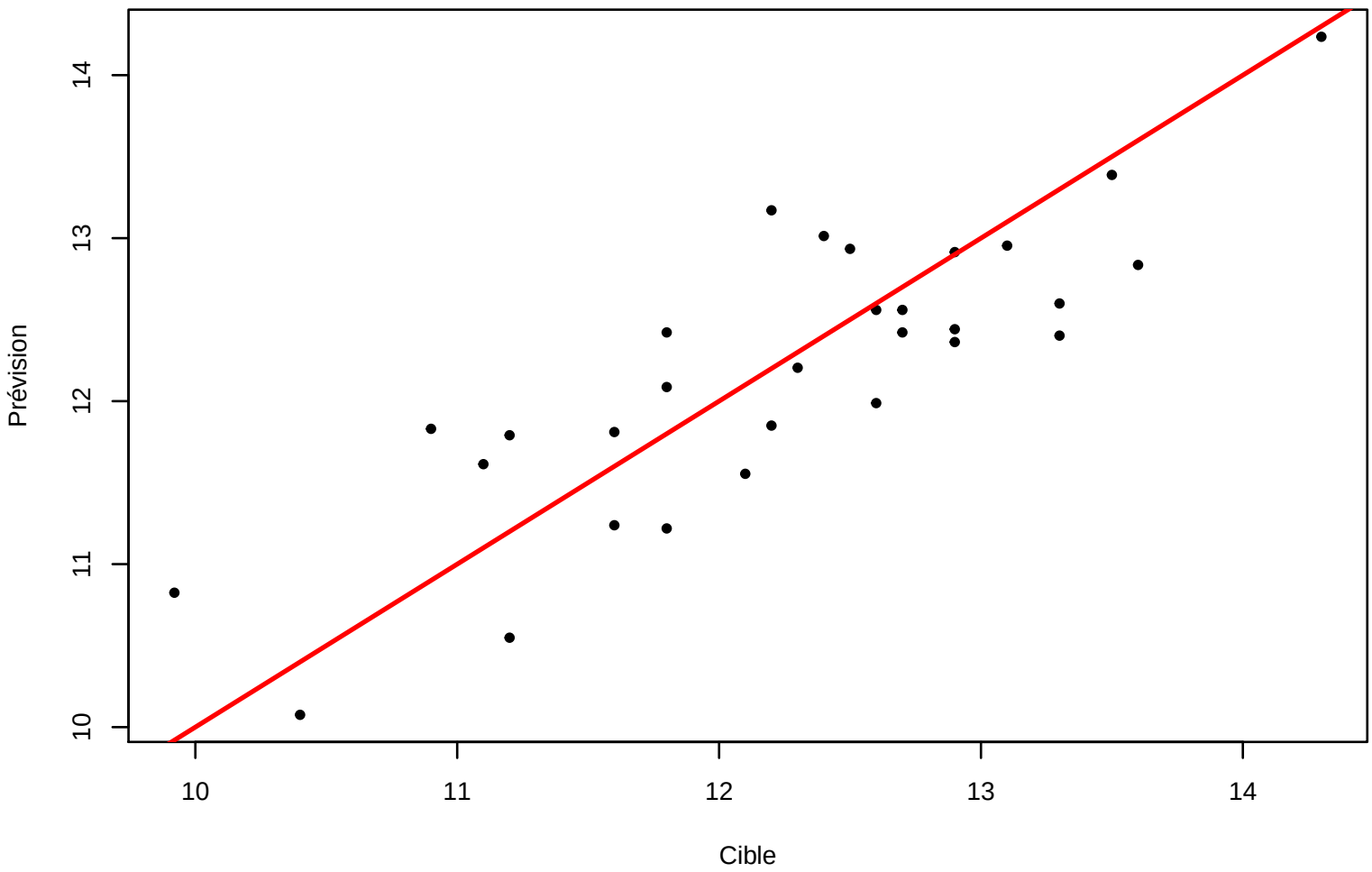
Modèle linéaire simple : $Alcool = \alpha X_1 + \beta$

Modèle linéaire simple : prévisions

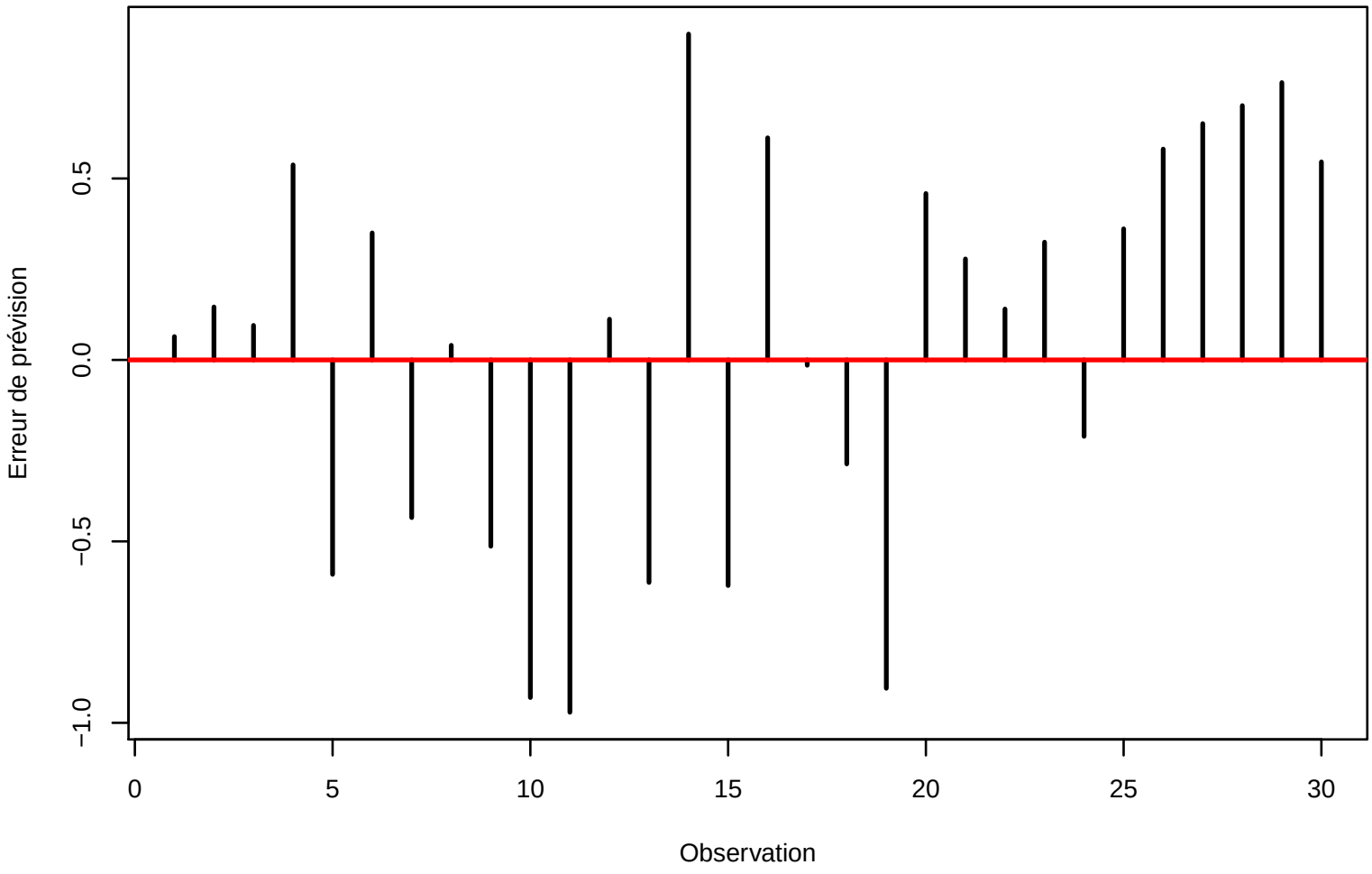


Prévisions du modèle

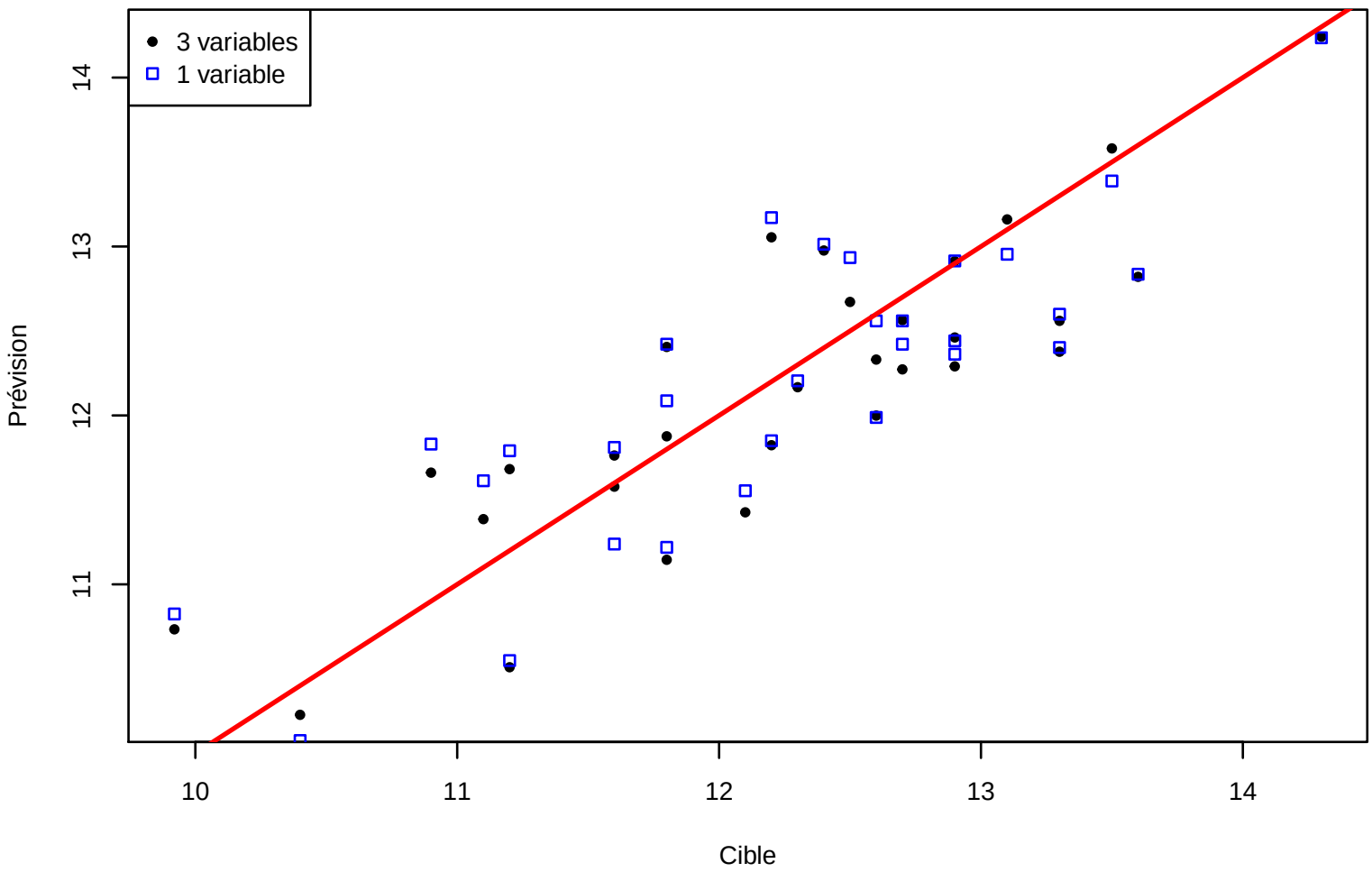
Représentation universelle



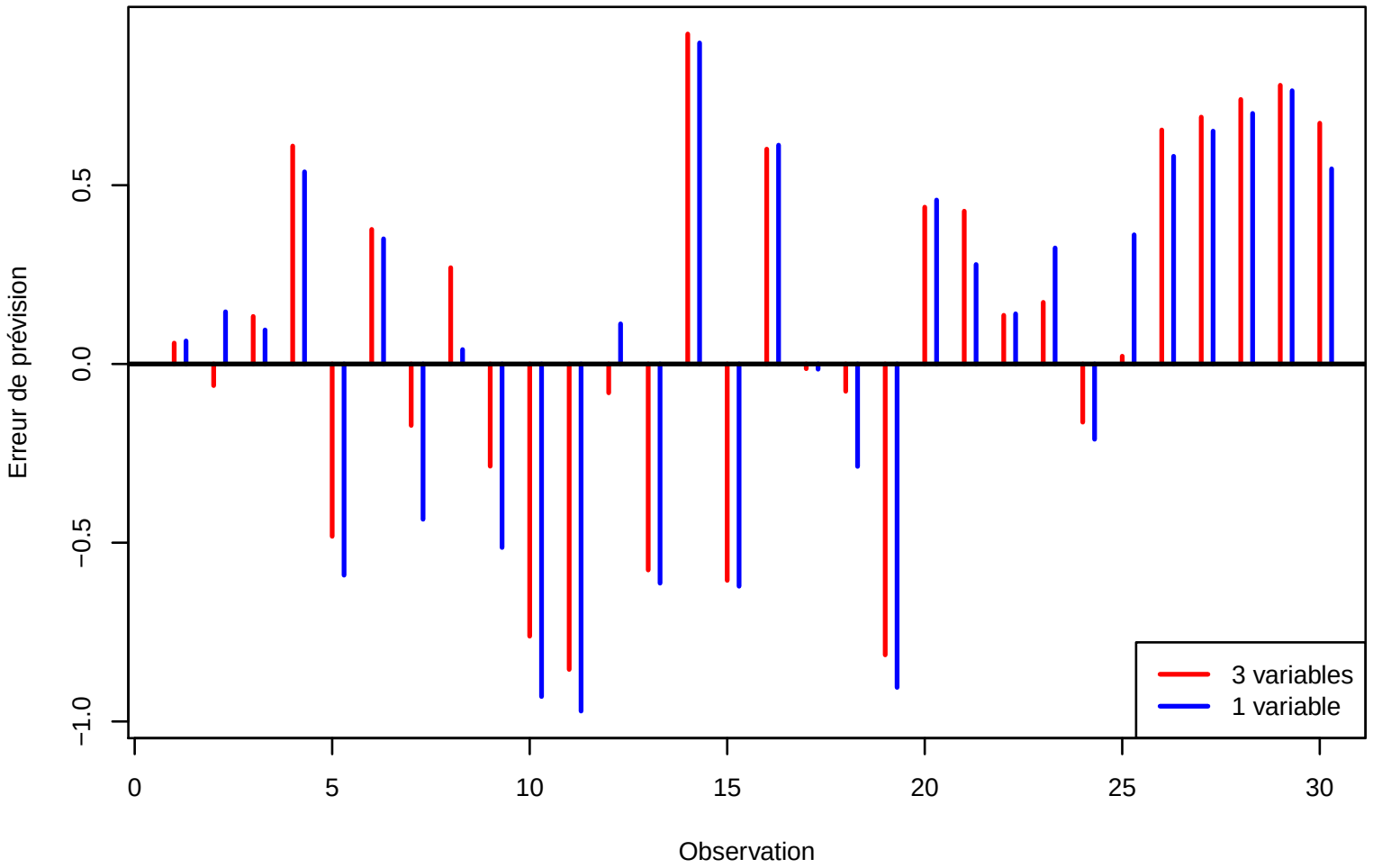
Représentation des erreurs



Modèle linéaire plus riche



Modèle linéaire plus riche



Choix du modèle

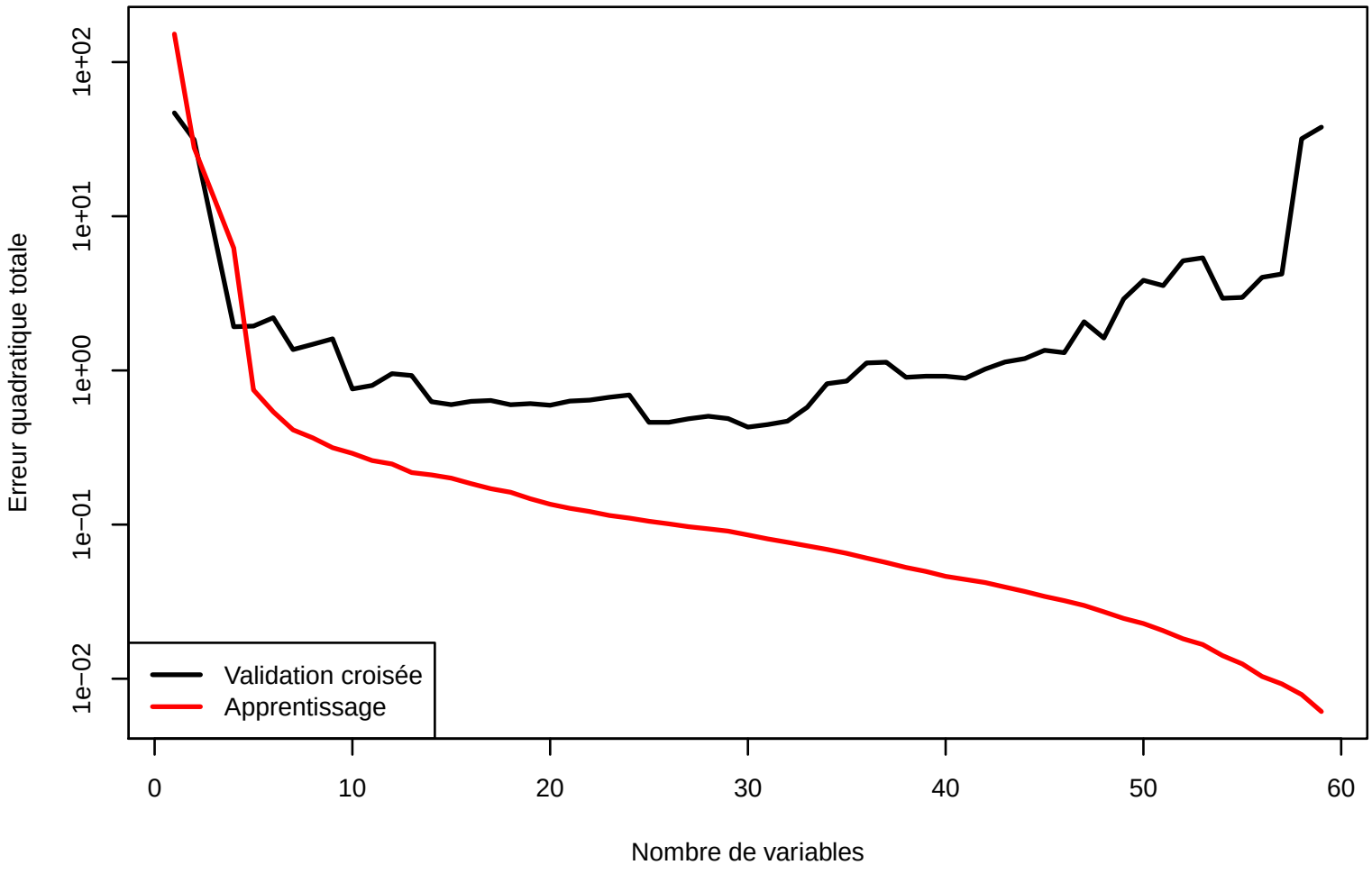
Candidats

- modèle linéaire réduit : choix des variables
- modèle linéaire régularisé *ridge*
- modèle linéaire régularisé *lasso*

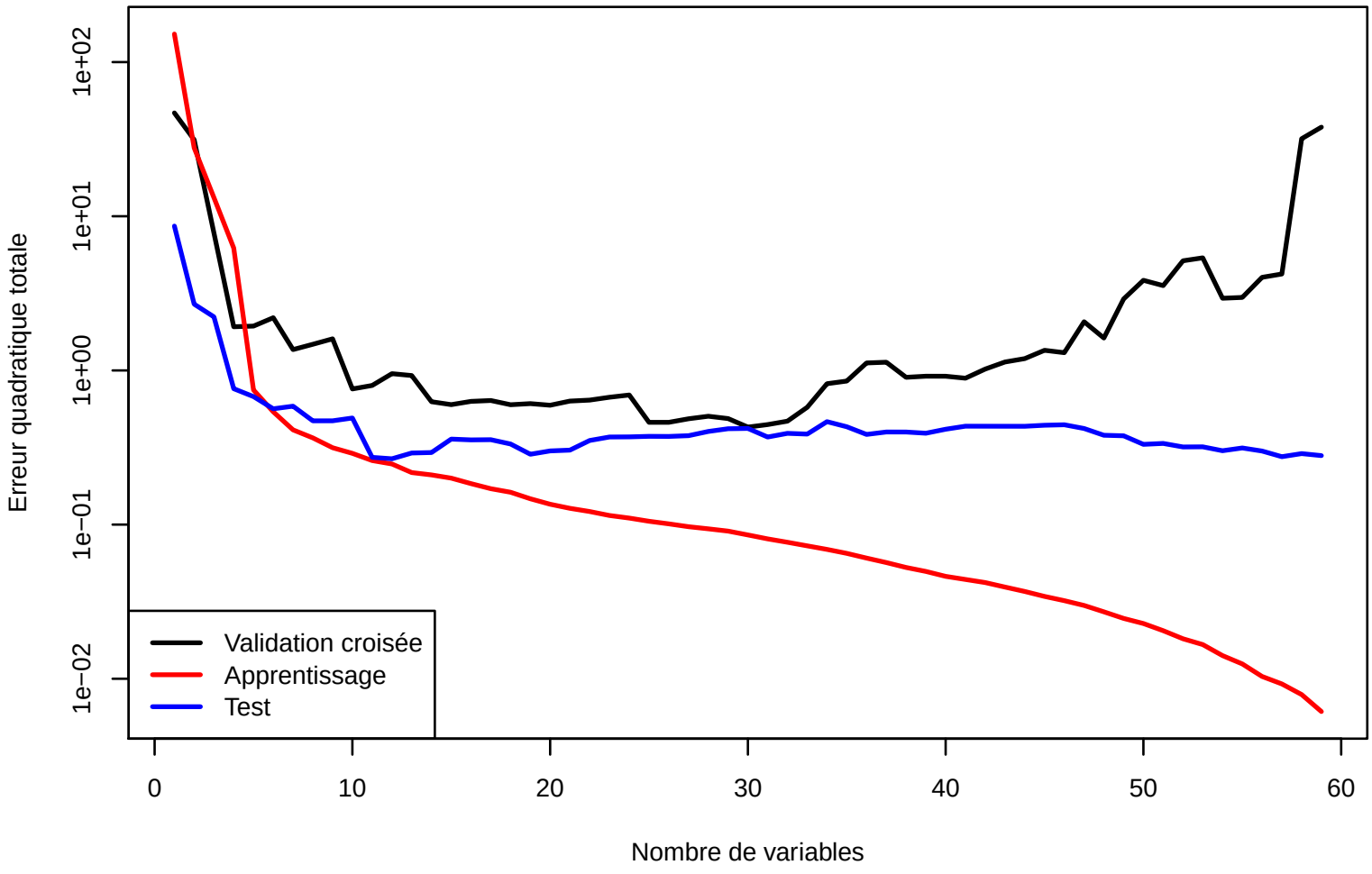
Méthode

- validation croisée (3 blocs, par exemple)
- choix du meilleur modèle
- et choix du paramètre du modèle (nombre de variables, paramètre de compromis)
- construction du modèle complet
- évaluation sur l'ensemble de test

Choix de variables



Choix de variables



Remarques

Validation croisée

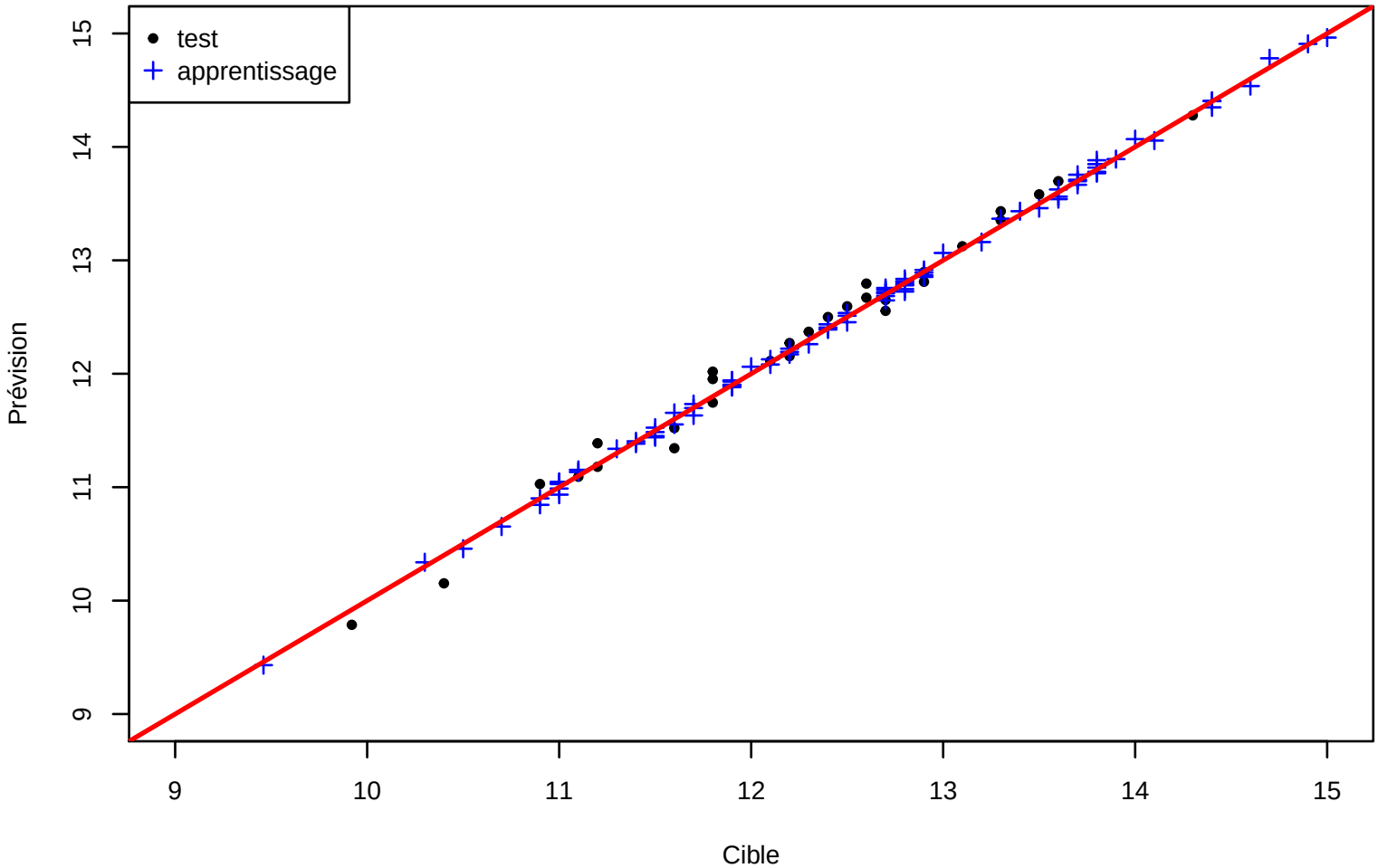
- chaque bloc correspond à un ordre spécifique pour les variables
- on agrège des prévisions pas des modèles
- on sélectionne une classe de modèles (des paramètres) pas un modèle
- on doit construire un modèle dans la classe après sélection
- le modèle final dépend de toutes les données d'apprentissage : on ne peut pas estimer ses performances sur l'ensemble d'apprentissage

Ensemble de test

- données indépendantes pour l'évaluation finale
- ne doit jamais être utilisé pour autre chose
- en particulier le choix du modèle final est fait sans utiliser l'ensemble de test

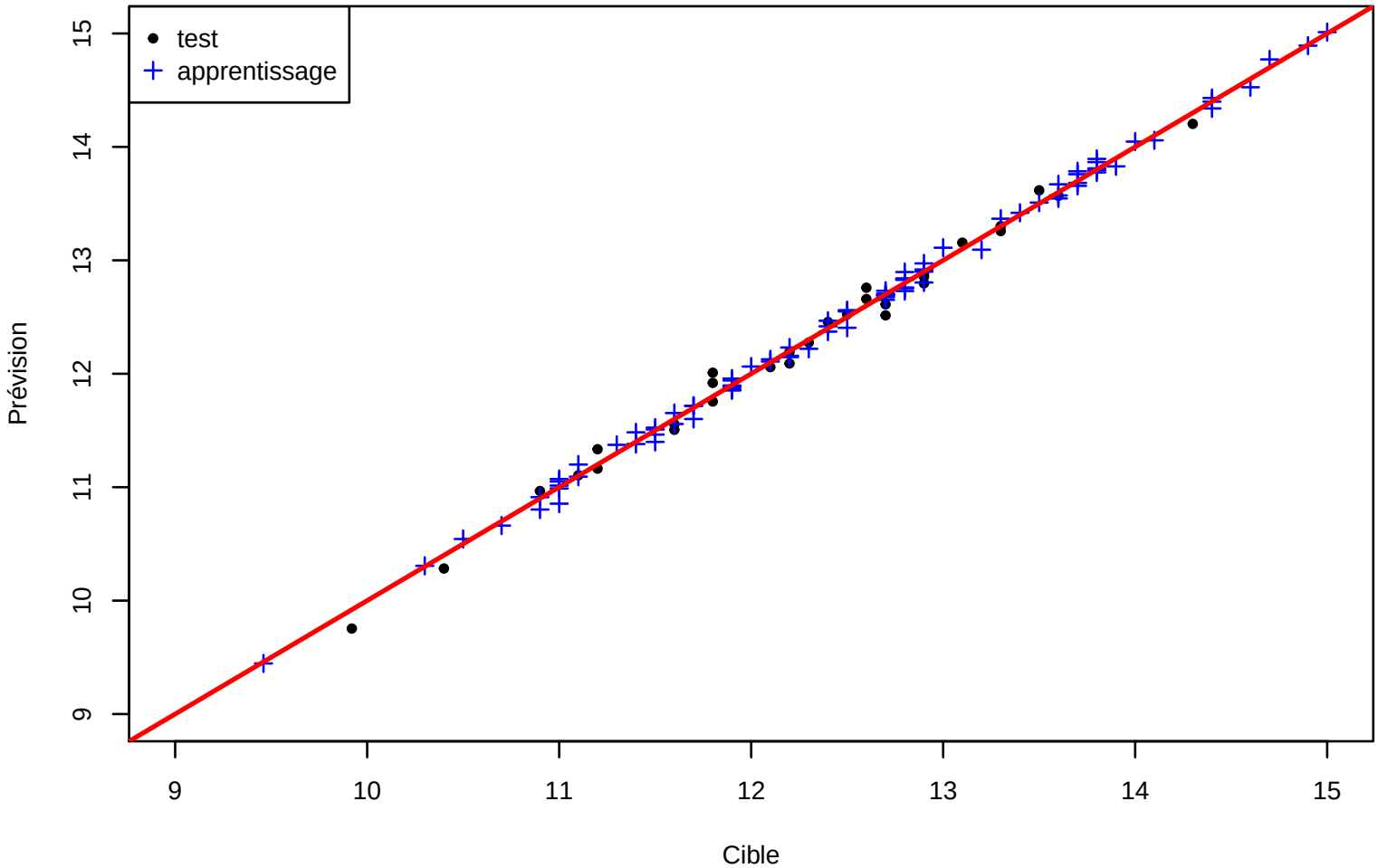
Meilleur modèle

30 variables



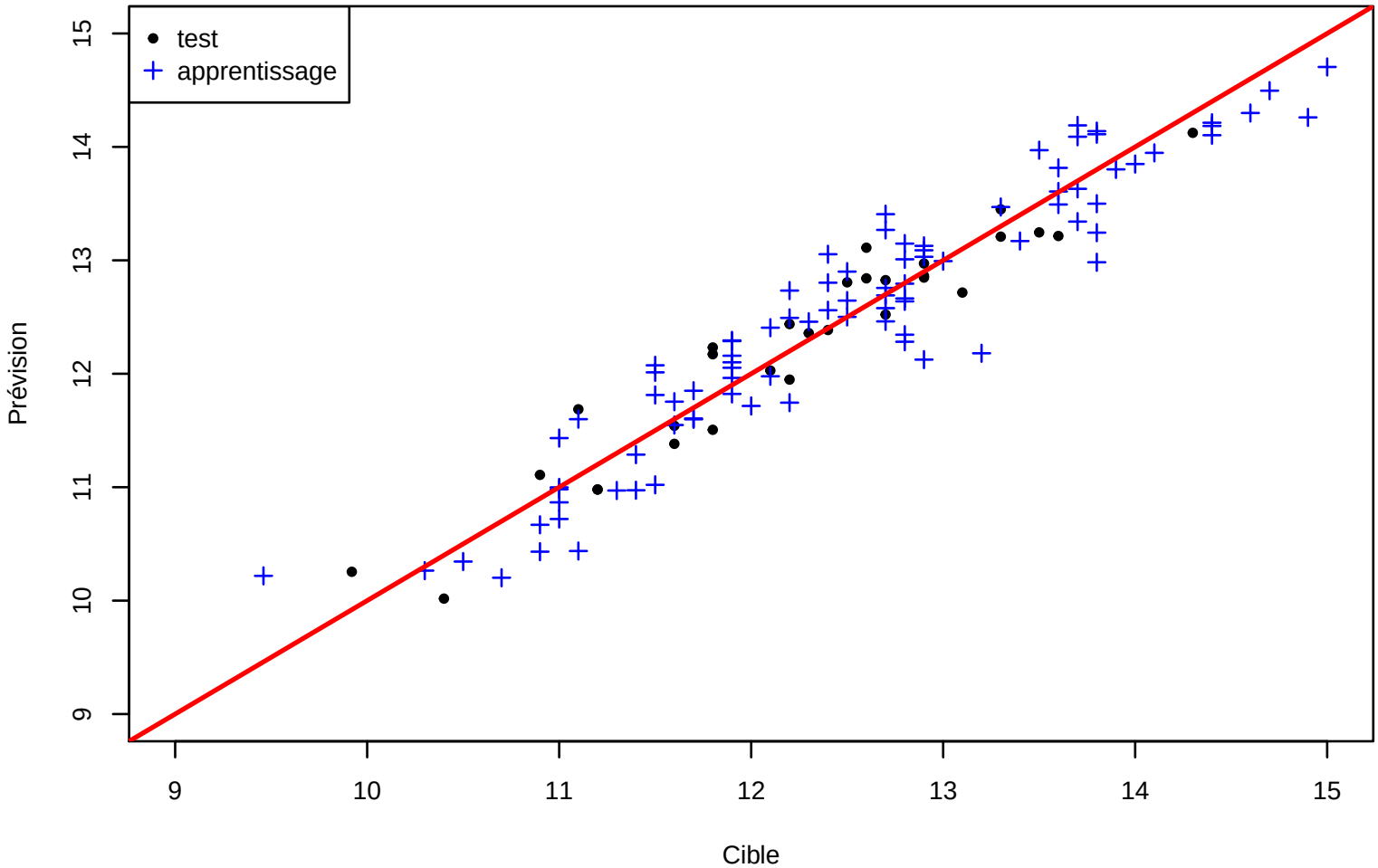
Meilleur modèle (tricherie)

12 variables



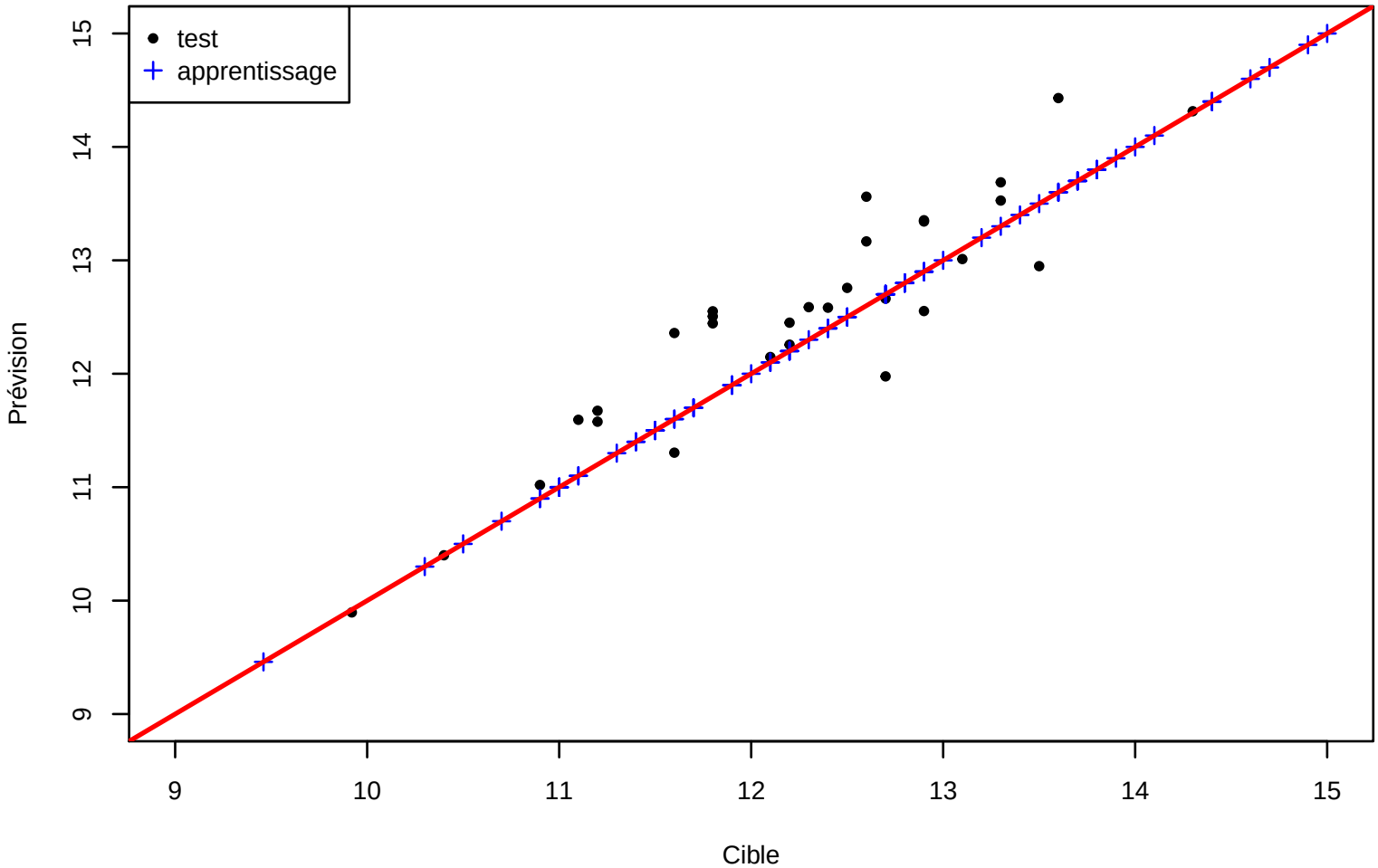
Modèle trop faible

3 variables



Modèle trop puissant

90 variables



Meilleur modèle

30 variables