Cours MPC - Partie 2

Régression : prédiction d'une variable à l'aide d'une ou plusieurs autres variables

Simon Malinowski

M1 Miage, Univ. Rennes 1

Sélection de variables : problème

Parmi les p variables prédictives ($X_1, \ldots X_p$), nous visons à sélectionner un sous-ensemble de variables qui conduisent au meilleur modèle.

Difficultés :

1. nombre de sous-ensembles d'un ensemble de p variables ?
   - je 20 variables : 1 million
   - je 30 variables : 1 milliard
2. Comment comparer des modèles avec un nombre différent de variables ?

Critères que nous avons déjà vus

1. coefficient de détermination R^2 → non adapté SCE
2. r → erreur de généralisation non adaptée → adapté
3.
4. probabilité critique du test de Student → adapté

Nous pouvons utiliser les critères suivants :

1. R-carré ajusté : $R^2_{adj} = R^2 \frac{(n-1)}{(n-p-1)}$
2. erreur de généralisation
3. probabilité critique du test de Student

Procédures de sélection des variables

Hypothèse : on sait comparer des modèles avec un nombre différent de variables (avec un critère adapté, cf. ci-dessus)

Plusieurs procédures :

1. Recherche exhaustive : nous essayons tous les sous-ensembles possibles. Jamais si $p > 15$
2. Recherche vers l'avant
3. Recherche en arrière
4. Recherche par étapes
5. Recherche par étapes

Nous avons p variables prédictives $x_1, \ldots, x_p$, et un critère de sélection C pour comparer des modèles (ex : erreur de généralisation)

- Rechercher le meilleur modèle avec 1 variable (selon C)
  - → Modèle $M_1 = \{x_1\}$, ses performances sont $C(M_1)$
  - → Le meilleur modèle trouvé jusqu'à présent est $M_b = M_1$
  - → La performance C_b du meilleur modèle est $C_b = C(M_1)$
- Nous recherchons ensuite la meilleure variable x_1 aller avec x_b
  - → Modèle $M_2 = \{x_b, x_2\}$, ses performances sont $C(M_2)$
  - → Si $C(M_2)$ est *mieux* que $C(M_1)$, alors $M_b = M_2$ et $C_b = C(M_2)$
- Répétez cette procédure jusqu'à ce qu'un critère d'arrêt soit atteint (expliqué ci-après)

Critères d'arrêt

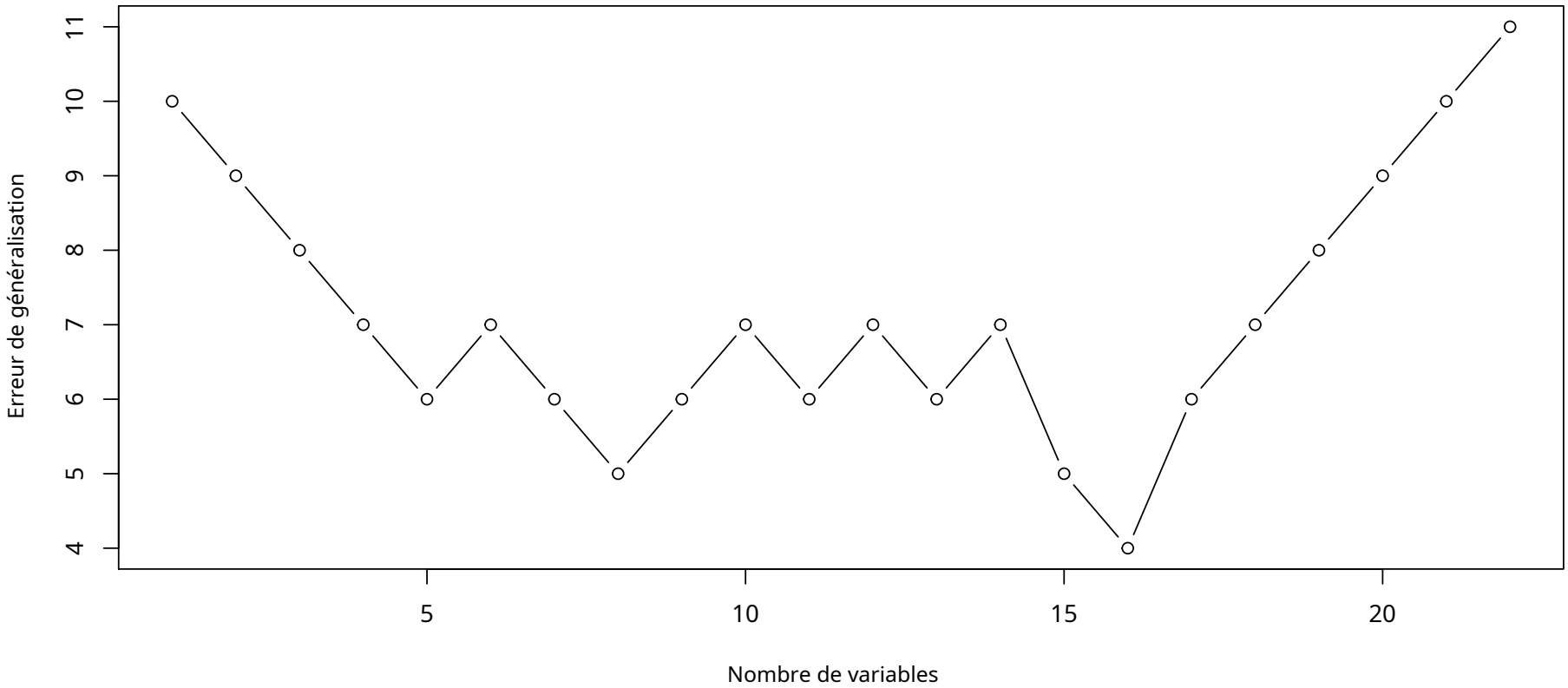
Le critère d'arrêt le plus utilisé est : arrêter dès qu'une itération n'améliore pas les performances (selon C).

C'est un peu strict, parfois on peut trouver de meilleurs modèles en attendant un peu plus

Autres critères d'arrêt :

- s'arrêter lorsqu'aucun meilleur modèle n'a été trouvé depuis δ les itérations ne
- s'arrêtent pas (allez jusqu'au bout) et conservez le meilleur modèle trouvé

Arrêt critères



Recherche vers l'avant : pseudo-code

Nous choisissons l'erreur de généralisation comme critère de performance Critère d'arrêt : dès que la performance diminue

Entrée : Ensemble de données avec p variables prédictives $X_1, \ldots, X_p$ et une cible *Oui*

Sortir : $V_s = \{X_{\sigma_1}, \ldots, X_{\sigma_k}\}$, sous-ensemble de variables sélectionnées

Recherche vers l'avant : pseudo-code

Initialisation :

V_s= [] :variables sélectionnées
V_{nu}= [$X_1, \ldots, X_p$] : variables non (encore) utilisées C_f= ∞:performances du meilleur modèle trouvé *arrêt*=*F*: variable pour gérer le critère d'arrêt

Recherche vers l'avant : pseudo-code

ALORS QUE $arrêt = F$

$\forall x \in V_{nu}$, Calculer les performances des modèles avec des variables $[V_s, x]$ Laisser x_b sois le meilleur x (ci-dessus), et C_b ses performances

SI $C_b < C_f$,
$V_s = [V_s, x_b]$; $C_f = C_b$; $V_{nu} = V_{nu} \setminus \{x_b\}$

AUTRE $arrêt = T$

FIN PENDANT QUE

Recherche en arrière

Même principe mais dans l'autre sens :

- Nous commençons avec le modèle complet (avec p variables)
- Nous recherchons la meilleure variable à supprimer (celle qui conduit au meilleur modèle) $\rightarrow p - 1$ variable
- Parmi ceux-ci $p - 1$, nous recherchons le meilleur pour supprimer la répétition
- jusqu'à ce qu'un critère d'arrêt soit rempli

Sélection de variables en pratique

La recherche en avant et en arrière sont des approximations de la recherche exhaustive

Ils ne conduisent pas toujours au même modèle sélectionné (et donc pas toujours au meilleur)

Différents critères de performance (R^2_{adj}, erreur de généralisation, ...) peut conduire à différents modèles

Les différents modèles sélectionnés doivent ensuite être comparés (sur un nouvel ensemble)

Sélection de variables en pratique

1. Divisez l'ensemble de données en un ensemble d'entraînement et un ensemble de test (avec environ 20 % pour l'ensemble de test)
2. Appliquer l'algorithme de sélection de variables à l'aide de l'ensemble d'apprentissage
   → si l'erreur de généralisation est le critère de performance, vous devrez à nouveau diviser l'ensemble d'apprentissage (en apprentissage et validation)
3. Estimer l'erreur de généralisation du modèle sélectionné sur l'ensemble de test.

1 Sélection de variables

2 Régression non linéaire

- Régression polynomiale

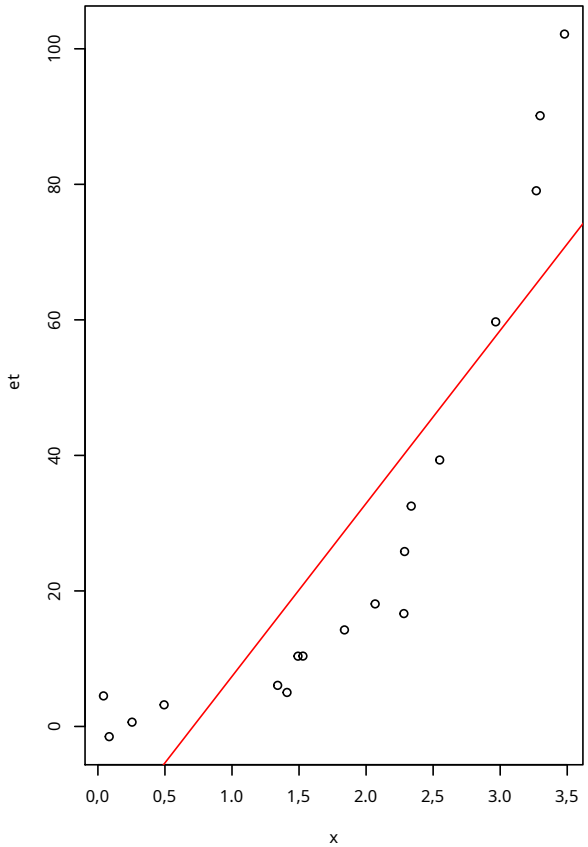
Régression polynomiale

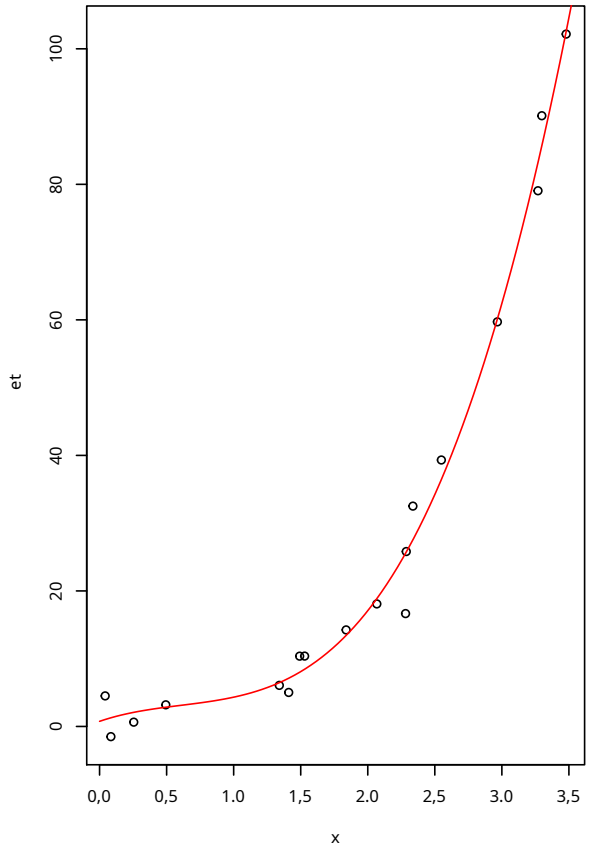
Régression linéaire classique : $et = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

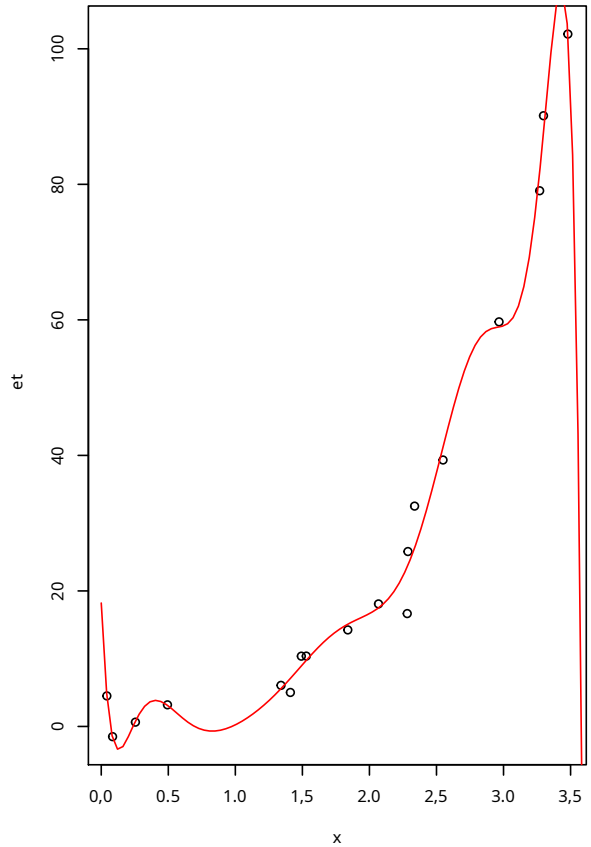
Régression polynomiale : même structure mais incluant les puissances des variables (en ajoutant de nouvelles colonnes dans le jeu de données)

Problème : Comment trouver quelle puissance doit être incluse pour quelle(s) variable(s) ?

Examen s'il vous plaît







En pratique

Régression avec des variables non numériques

Taille	District	Prix
80	Centre	120
92	Bourg-Lesvèque	180
75	Centre	145
110	Longchamps	220
85	Longchamps	140
150	Bourg-Lesvèque	225
55	Centre	100
105	Longchamps	??
95	Centre	??