Le modèle de régression linéaire

Master 2 Recherche SES-IES Analyse de données

Ana Karina Fermin

Université Paris Nanterre

http://fermin.perso.math.cnrs.fr/

Modèle de régression

On dispose de n observations $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ du couple $(\mathbf{X}, Y)$. On suppose que

$$y_i = f^\star(\mathbf{x}_i) + \varepsilon_i \text{ pour tout } i = 1, \ldots, n$$

- les $\mathbf{x}_i$ son des valeurs connues non aléatoires
- $f^\star$ est une fonction inconnue
- ε_i sont des réalisations inconnues d'une variable aléatoire.

Pour chaque individu i, la variable aléatoire ε_i représente l'erreur commise. Généralement pour étudier le modèle "le statisticien" formule des hypothèses sur la loi des erreurs ε_i.

Objectif

On souhaite "expliquer" une variable Y à partir de $\mathbf{X}$.
Nous allons chercher une fonction f telle que

$$Y \approx f(\mathbf{X}).$$

Pour définir $\approx$ il faut donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données:

$$(Y - f(X))^2$$

La vraie fonction $f^\star$ minimise en moyenne cette erreur... mais elle est inconnue!

En pratique

On va choisir f dans une classe de fonctions $\mathcal{S}$.
On va minimiser une erreur moyenne *sur les données*:

$$\widehat{f} = \arg\min_{f \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$$

La régression linéaire correspond à $\mathcal{S} = \{\mathbf{x} \mapsto \mathbf{x}^t \beta\}$.

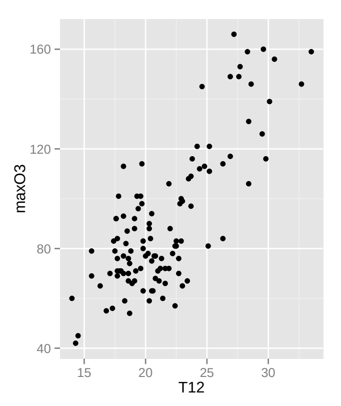
Attention:

- Il faut choisir $\mathcal{S}$ (le modèle)
- $\widehat{f} \neq f^\star$
- On est même pas sûr que $(Y - \widehat{f}(X))^2$ (ou $(f^\star(\mathbf{X}) - \widehat{f}(\mathbf{X})^2)$ soit petit en moyenne...

Exemple : Pollution l'ozone

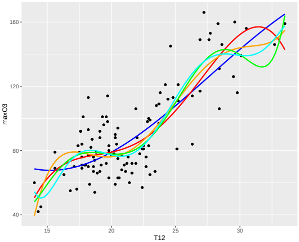
- X : température à midi
- Y : concentration maximale en ozone

mesurés en un lieu donné et une journée donnée pendant n jours.



Régression polynomiale

f est choisie dans une classe des fonctions $\mathcal{S}$ polynomiales
Modèles obtenus par des polynôme du degré 3, 4, 5, 6 et 7
Pb : Choisir "le bon" degré !

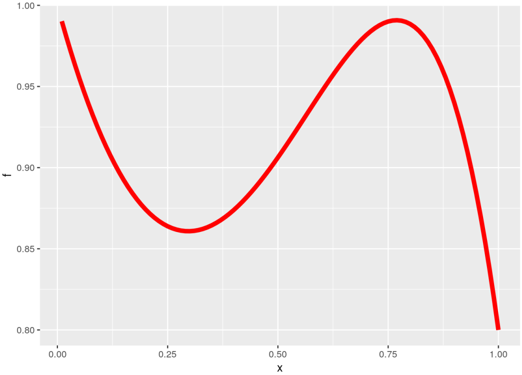


- Modèle polynomial: $f_\beta(\mathbf{X}_i) = \sum_{l=0}^{d} \beta_l \mathbf{X}_i^l$
- Linéaire en β!
- Ici $\mathbf{X}_i' = (1, \mathbf{X}_i, \ldots, \mathbf{X}_i^d)^t$
- Problème d'estimation de MC facile!

Exemple Jouet

- Nous commencerons avec un exemple artificiel !
- Nous voulons estimer les valeurs de

$$f^\star(x) = 1 - x + 2x^2 - 0.8x^3 + 0.6x^4 - x^5$$

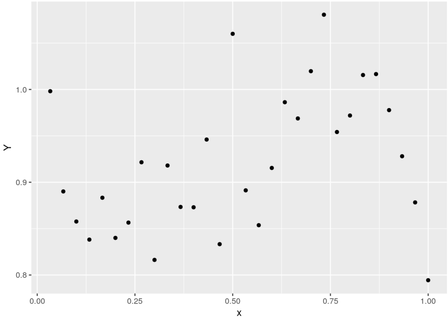


Modélisation

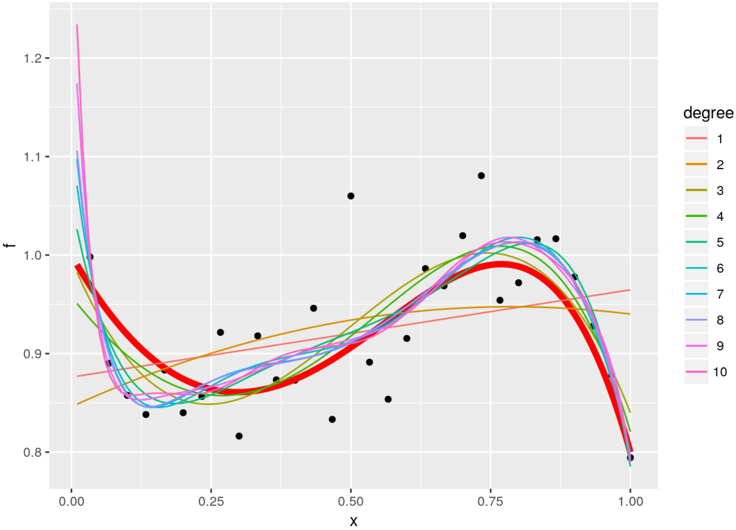
- Design fixé : $x_k = k/n$, with $1 \leqslant k \leqslant n$
- Nous observons les valeurs de $f^\star$ dans x_k contaminées par un bruit Gaussien

$$Y_k = f^\star(k/n) + \epsilon_k$$

- Ici, ϵ_k sont des réalisations i.i.d. centrées d'une v.a. Gaussienne of variance σ^2.

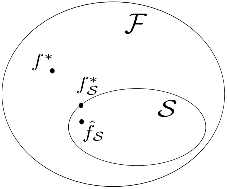


Quel degré?



Compromis Biais-Variance

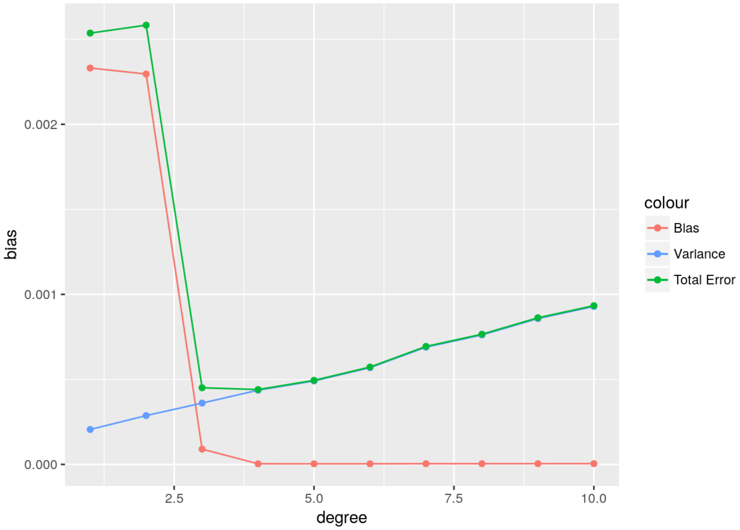
- Cadre général:
  - $\mathcal{F}$: Famille de toutes les fonctions
  - Meilleur solution dans $\mathcal{F}$: f^*
  - Sous-Famille $\mathcal{S} \subset \mathcal{F}$ de functions
  - Meilleure solution dans $\mathcal{S}$: $f^*_{\mathcal{S}}$
  - Estimée $\mathcal{S}$: $\widehat{f}_{\mathcal{S}}$ obtenue par moindre carré.



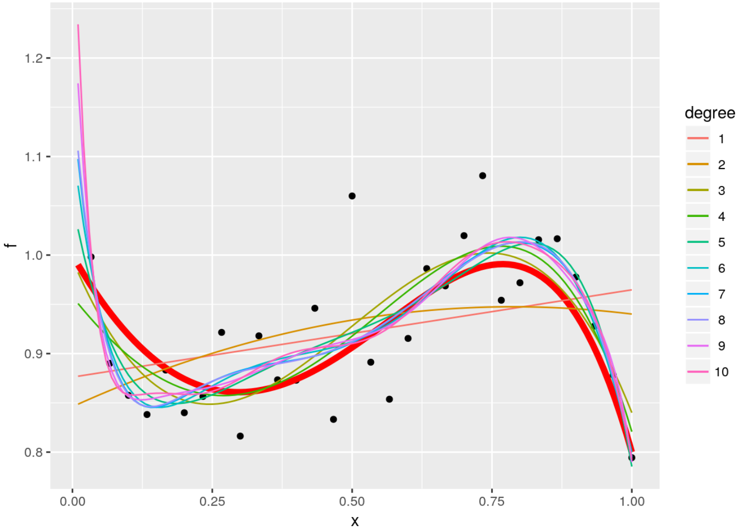
Erreur d'approximation et erreur d'estimation (Biais/Variance)

$$\|\widehat{f}_{\mathcal{S}} - f^\star\|^2 = \underbrace{\|f^*_{\mathcal{S}} - f^\star\|^2}_{\text{Erreur d'approximation}} + \underbrace{\|\widehat{f}_{\mathcal{S}} - f^*_{\mathcal{S}}\|^2}_{\text{Erreur d'estimation}}$$

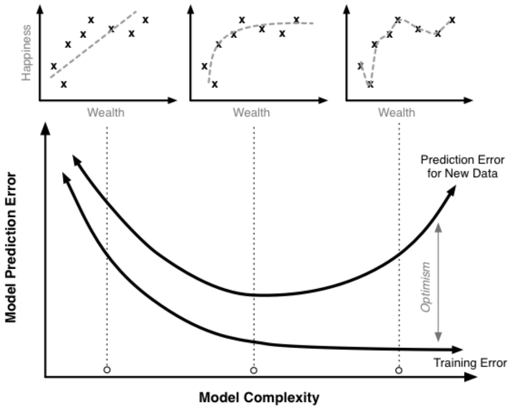
- L'erreur d'approximation peut être grande si le modèle $\mathcal{S}$ n'est pas adapté.
- L'erreur d'estimation est grande lorsque le modèle est complexe.


0.002
0.001
0.000
bias
2.5
5.0
7.5
10.0
degree
colour
Bias
Variance
Total Error

Quel degré?



Sur-Apprentissage



Validation croisée

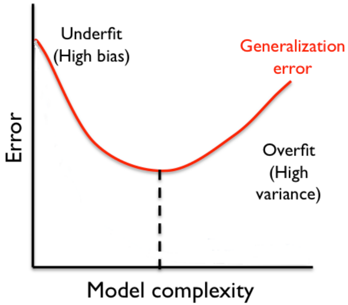
- **Idée très simple:** conserve une partie pour vérifier l'erreur.
- Suffisent pour éviter un sur-apprentissage!

Cross Validation

- Utiliser $\frac{V-1}{V}n$ observations pour apprendre et $\frac{1}{V}n$ pour vérifier!

- Variantes Classiques :
  - Leave One Out,
  - V-fold validation croisée.
- Souvent on choisi: $V = 5$ ou $V = 10$!

Sur-apprentissage / sous-apprentissage



- Différents comportements pour des complexités de modèles différentes

Compromis Bias-variance $\iff$ eviter sur-app. and sous-app.

$\widehat{f}_{\widehat{m}}$: régression avec un polynôme de degré 4