

STT3030 - Cours #1

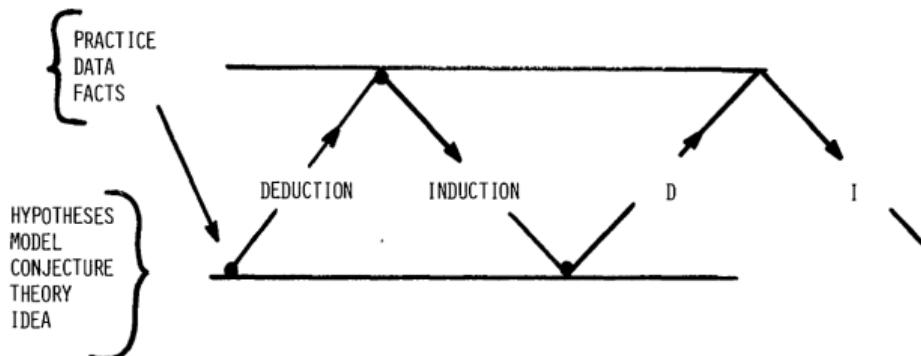
Arthur Charpentier

Automne 2024

Data Science

“The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning... Data-driven predictions can succeed—and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves.” Silver (2012)

cf Tukey (1962) sur l'analyse des données



cycle de raisonnement inductif-déductif de Box (1976)

THE FUTURE OF DATA ANALYSIS*

BY JOHN W. TUKEY

Princeton University and Bell Telephone Laboratories

- I. General Considerations
 1. Introduction
 2. Special growth areas
 3. How can new data analysis be initiated?
 4. Sciences, mathematics, and the arts
 5. Dangers of optimisation
 6. Why optimisation?
 7. The absence of judgment
 8. The reduction of judgment upon theory
 9. Teaching data analysis
 10. Practicing data analysis
 11. Facing uncertainty
 12. Spotty Data
 13. What is it?
 14. An appropriate step forward
 15. Trimming and Winsorizing samples
 16. How soon should such techniques be put into service?
 - II. Spotty Data
 16. Modified normal plotting
 17. Automated examination
 18. FUNOP
 19. FUNOP-FUNOM in a two-way table
 20. Example of use of FUNOP-FUNOM
 - IV. Multiple-Response Data
 21. Where are we, and why?
 22. The case of two samples
 23. Factor analysis: the two parts
 24. Factor analysis: regression
 25. Factor analysis: the middle lines
 26. Taxonomy; classification; incomplete data
 - V. Some Other Promising Areas
 27. Stochastic-process data
 28. Selection and screening problems
 29. External, internal, and confounded estimates of error
 30. The consequences of half-normal plotting
 31. Heterogeneous data
 32. Two samples with unequal variability
 - VI. Flexibility of Attack
 33. Choice of modes of expression
 34. Sizes, nomination, budgeting
 35. A caveat about indications
 36. FUNOP as an aid to group comparison
 37. Continuation
 - VII. A Specific Sort of Flexibility
 38. The vacuum cleaner
 39. Vacuum cleaning: the subprocedure
 40. The basic vacuum cleaner, and its attachments
 41. The vacuum cleaner: an example
 42. The example continued
 - VIII. How Shall We Proceed?
 43. What are the necessary tools?
 44. The role of empirical sampling
 45. What are the necessary attitudes?
 46. How might data analysis be taught?
 47. The impact of the computer
 48. What of the future?
- References

Data vs. Algorithmic Modeling

Statistical Science
2001, Vol. 16, No. 3, 199–231

Les “deux cultures”, de Breiman (2001)

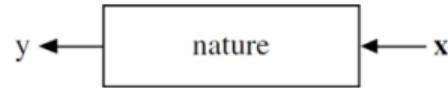
► Data Modeling

choisir un modèle simple (linéaire) basé sur l'intuition du mécanisme de génération des données. L'accent est mis sur l'interprétabilité du modèle et la validation, si elle a lieu, se fait par la qualité de l'ajustement.

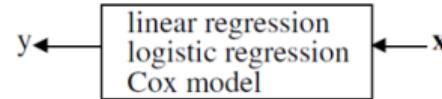
► Algorithmic Modeling

choisir le modèle dont la précision de validation prédictive est la plus élevée, sans tenir compte de l'explicabilité du modèle.

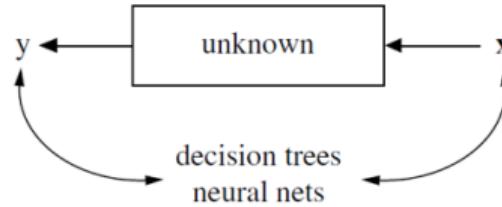
Statistical Modeling: The Two Cultures



The Data Modeling Culture



The Algorithmic Modeling Culture



Statistique et apprentissage machine

James et al. (2013), chapitre 2:

la statistique et l'apprentissage machine sont similaires:

- ▶ Analyse de données / apprentissage par données
- ▶ Estimer des paramètres / entraîner des modèles

La statistique et l'apprentissage machine sont différents:

- ▶ Fondé sur les probabilités / fondé sur l'algorithmique
- ▶ Modèles interprétables et analyse de la variance rigoureuse / modèles de prédiction très précis, mais difficilement interprétable.
- ▶ Inférence/ Prédiction
- ▶ Deux approches **complémentaires** pour l'analyse de données.

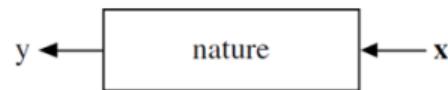
Les deux cultures de la modélisation statistique

Statistical Science
2001, Vol. 16, No. 3, 199–231

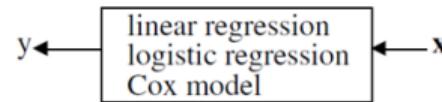
Statistical Modeling: The Two Cultures

Les “deux cultures”, de Breiman (2001)

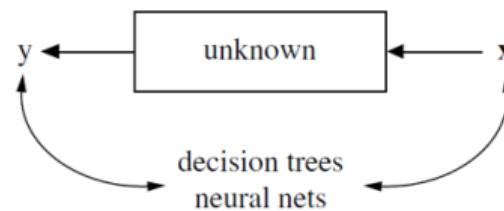
- ▶ La statistique: basée sur les distributions de probabilité.
- ▶ L'apprentissage machine: basé sur des approches d'optimisation algorithmique.
- ▶ L'apprentissage statistique: essaie d'utiliser le meilleur des deux mondes.



The Data Modeling Culture



The Algorithmic Modeling Culture



Apprentissage statistique

L'apprentissage statistique vise à garder ces deux cultures connectées; dans ce cours nous serons beaucoup appelés à faire la prédiction la plus précise possible.

Au cœur du cours est *l'application* de multiples modèles à des fins d'analyse de données. Conséquemment, nous verrons comment utiliser ceux-ci sur R et/ou Python.

Nous allons découvrir plusieurs modèles, apprendre comment les utiliser, les entraîner, leurs forces et leurs faiblesses et comment choisir le bon modèle étant donné un problème.

Apprentissage statistique

Nous allons travailler avec de vraies données et des données simulées.

Le concept du cours et d'en apprendre sur un modèle suffisamment pour l'employer, et immédiatement l'utilisé en laboratoire.

Il va donc de soi qu'une bonne partie du cours se tiendra dans les laboratoires informatiques, ainsi que les évaluations.

Machine Learning ?

RESEARCH CONTRIBUTIONS

Artificial
Intelligence and
Language Processing

David Waltz
Editor

A Theory of the Learnable

L. G. VALIANT

The main contribution of this paper is that it shows that it is possible to design *learning machines* that have all three of the following properties:

1. The machines can provably learn whole classes of concepts. Furthermore, these classes can be characterized.
2. The classes of concepts are appropriate and nontrivial for general-purpose knowledge.
3. The computational process by which the machines deduce the desired programs requires a feasible (i.e., polynomial) number of steps.

More precisely we say that a class X of programs is *learnable* with respect to a given learning protocol if and only if there exists an algorithm A (the deduction procedure) invoking the protocol with the following properties:

1. The algorithm runs in time polynomial in an adjustable parameter h , in the various parameters that quantify the size of the program to be learned, and in the number of variables t .
2. For all programs $f \in X$ and all distributions D over vectors v on which f outputs 1, the algorithm will deduce with probability at least $(1 - h^{-1})$ a program $g \in X$ that never outputs one when it should not but outputs one almost always when it should. In particular, (i) for all vectors v , $g(v) = 1$ implies $f(v) = 1$, and (ii) the sum of $D(v)$ over all v such that $f(v) = 1$, but $g(v) \neq 1$ is at most h^{-1} .

via Valiant (1984)

Les fondations

Dans ce premier bloc thème cours il sera question de la base de l'apprentissage supervisé (**supervised learning**).

Il s'agit du problème de prédire/estimer la relation entre deux ensembles de variables.

On veut prédire la réponse Y avec les prédicteurs $\mathbf{X} = \{X_1, \dots, X_p\}$.

$\mathbf{X} = \{X_1, \dots, X_p\}$ forment une collection de prédicteurs, variables explicatives, covariables, entrées, etc.. Desfois simplement X (*predictors, features or input*).

On dit que $\mathbf{X} \in \mathcal{X}_1 \times \dots \times \mathcal{X}_p = \mathcal{X}$, où \mathcal{X}_j est le domaine de X_j .

Ce sont les variables que nous utilisons pour prédire.

$Y \in \mathcal{Y}$ est une réponse, variable expliquée, étiquette (si catégoriel) ou tout simplement la sortie.

C'est ce que nous cherchons à prédire (à l'aide des prédicteurs).

Supposons qu'il s'agit d'une variable continue pour l'instant.

Processus génératif

Comment les données sont-elles générées ? (dans la **nature**, dans le vrai monde)

- ▶ On suppose l'existence d'une fonction $f : \mathcal{X} \rightarrow \mathcal{Y}$
- ▶ Souvent de la forme la vrai fonction $Y = f(\mathbf{X}) + \varepsilon$

où ε est une supposée variabilité inexpliquée, résiduelle (erreur de mesure, variabilité naturelle, etc..).

En apprentissage supervisé, on veut estimer (ou apprendre) f , et \hat{f} sera la prévision.

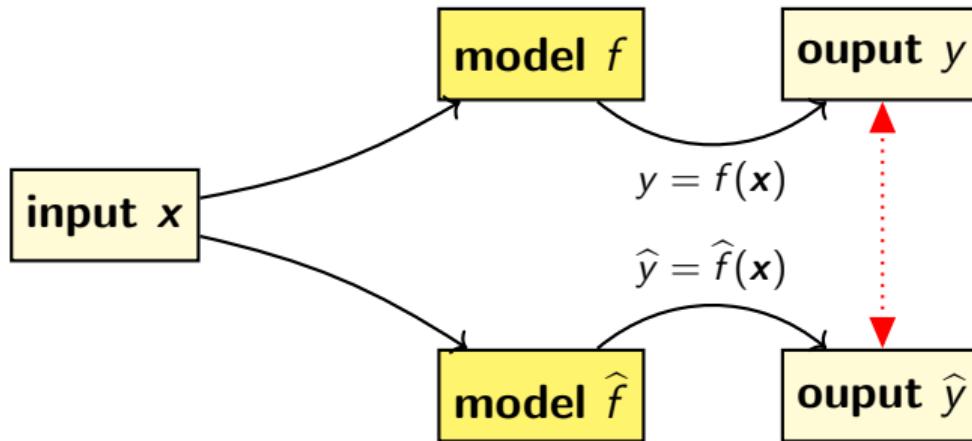
- ▶ Pour faire de la prédiction: $\hat{Y} = \hat{f}(\mathbf{X})$
- ▶ Étant donnée des prédicteurs \mathbf{X} quelle est notre meilleure estimation de la réponse.
- ▶ Pour faire de l'inférence.
- ▶ Quelle est la relation entre \mathbf{X} et Y ? Quel prédicteur est le plus utile ?

L'apprentissage au service de la prédiction.

- ▶ Dans ce cours, on va se concentrer sur la prédiction.
- ▶ La tâche pour laquelle il y a eu d'immenses progrès récemment en IA.
- ▶ On va parler un peu d'inférence, mais on en fait beaucoup en statistique déjà.

Modèle

On suppose qu'il existe une fonction $f : \mathcal{X} \rightarrow \mathcal{Y}$, telle que $y = f(\mathbf{x})$.



On dispose de données $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}$, réalisation de n variables i.i.d. (Y_i, \mathbf{X}_i) .

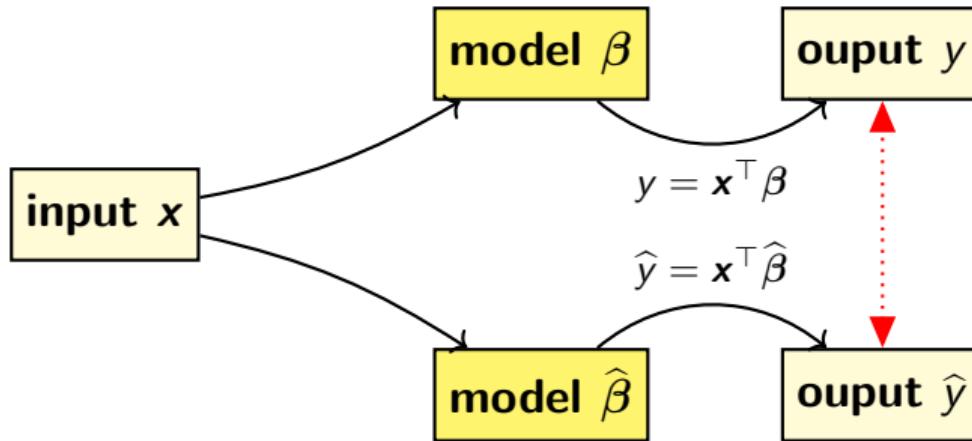
\mathcal{D}_n est une réalisation de $D_n = \{(Y_i, \mathbf{X}_i)\}$

On a associer un modèle $\hat{f} = f(\cdot | \mathcal{D}_n) \in \mathcal{M}$ à partir des données \mathcal{D}_n

- ▶ $\hat{y} = f(\mathbf{x} | \mathcal{D}_n)$ est la prévision associée à \mathbf{x}
- ▶ $\hat{Y} = f(\mathbf{X} | \mathcal{D}_n)$ est la prévision vue comme une variable aléatoire

Modèle

On suppose qu'il existe un paramètre β , tel que $y = \mathbf{x}^\top \beta + \epsilon$.



On dispose de données $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}$, réalisation de n variables i.i.d. (Y_i, \mathbf{X}_i) .

\mathcal{D}_n est une réalisation de $D_n = \{(Y_i, \mathbf{X}_i)\}$

On a associer un **modèle** (estimateur) $\hat{\beta}_n$ à partir des données \mathcal{D}_n

- ▶ $\hat{y} = \mathbf{x}^\top \hat{\beta}_n$ est la prévision associée à \mathbf{x}
- ▶ $\hat{Y} = \mathbf{X}^\top \hat{\beta}_n$ est la prévision vue comme une variable aléatoire

Données d'entraînement

Nous avons un ensemble de données

$$\mathcal{D}_n = \{(\mathbf{x}_i, y_i) | i \in (1, \dots, n)\} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

dit données **d'entraînement** (i.e. *training set*), ou **échantillon d'apprentissage**.

C'est grâce à ces données que nous allons entraîner notre modèle, apprendre à imiter f .

Exemple 1

Exemple du livre, James et al. (2013) :

- ▶ On veut prédire les ventes Y (variable continue).
- ▶ On utilise les dépenses en publicité \mathbf{X} pour prédire les ventes. Nous avons trois variables explicatives, les dépenses en publicité télévision X_1 , à la radio X_2 et dans les journaux X_3 .
- ▶ On a un échantillon \mathcal{D}_n de taille n entreprise pour lesquels on connaît leurs dépenses en publicité ($\{x_{i,j} : i \in (1, \dots, n), j \in (1, 2, 3)\}$) ainsi que leurs revenus ($\{y_i : i \in (1, \dots, n)\}$).
- ▶ Les données y_i et \mathbf{x}_i sont apparaillées

Exemple 1, James et al. (2013)

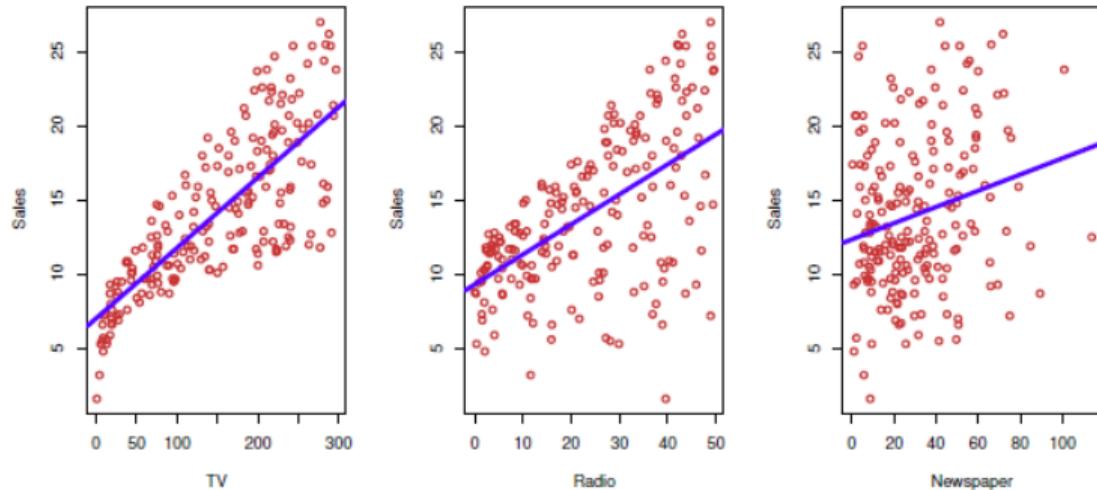


FIGURE 2.1. The Advertising data set. The plot displays sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict sales using TV, radio, and newspaper, respectively.

Exemple 2, Deng (2012)

- ▶ On a une collection d'images de chiffre écrit à la main ([MNIST](#)).
- ▶ Chaque image est composée de 784 pixels (28×28), allant du noir au blanc (niveau de gris).



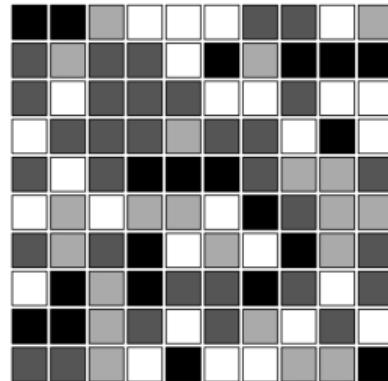
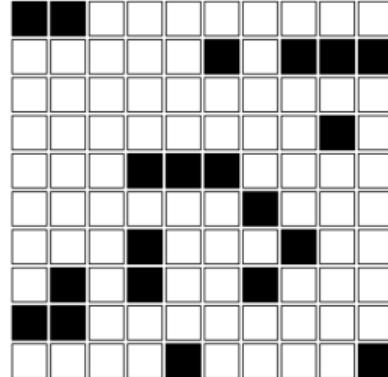
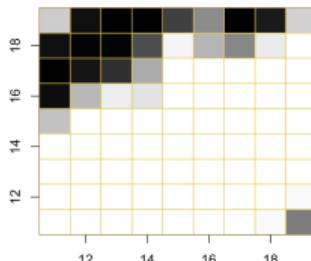
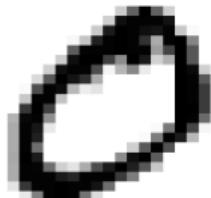
3	4	2	1	9	5	6	2	1	8
8	9	1	2	5	0	0	6	6	4
6	7	0	1	6	3	6	3	7	0
3	7	7	9	4	6	6	1	8	2
2	9	3	4	3	9	8	7	2	5
1	5	9	8	3	6	5	7	2	3
9	3	1	9	1	5	8	0	8	4
5	6	2	6	8	5	8	8	9	9
3	7	7	0	9	4	8	5	4	3
7	9	6	4	7	0	6	9	2	3

Images, matrices et tenseurs

Une image ($w \times h$), avec w en longueur (width)
et h en hauteur (height): **matrice**

Image noir et blanc, matrice M

avec $M_{i,j} \in [0, 1]$ (niveau de gris).



Tenseurs

Une image ($w \times h$), avec w en longueur (width)
et h en hauteur (height): **tenseur**

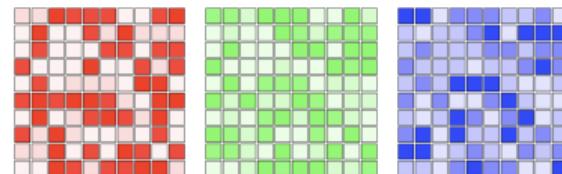
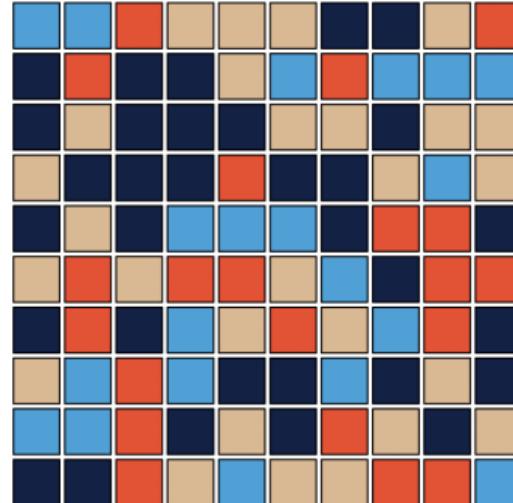
Image en couleur, tenseur T

$$T = (T[, , r], T[, , g], T[, , b])$$

(décomposition **RGB**)

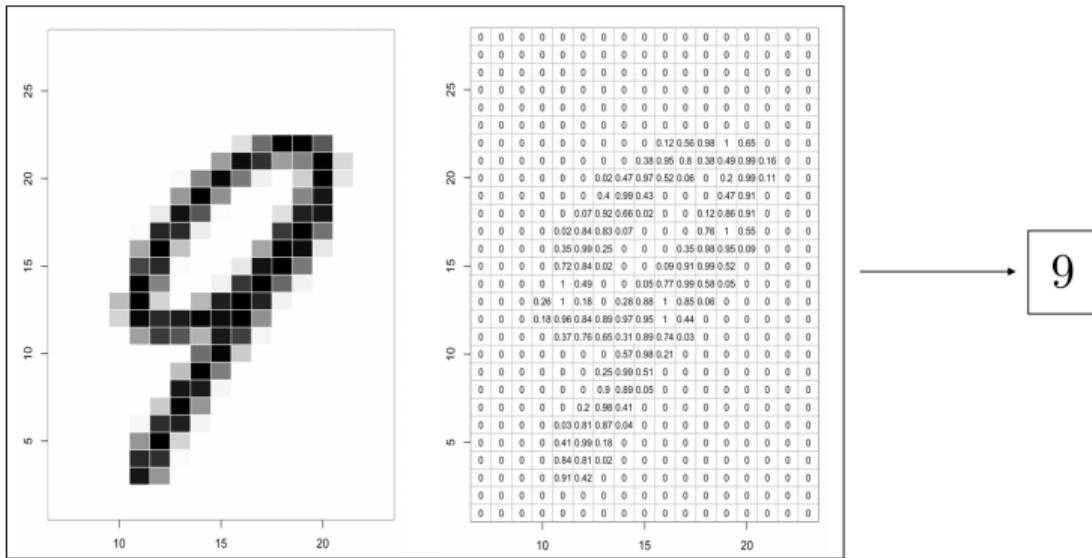
où $T_{i,j,c} \in [0, 1]$ est une intensité de couleur.

travailler avec des images signifie travailler avec
des matrices tridimensionnelles (tenseurs)



Exemple 2

- ▶ Les prédicteurs sont les 784 pixels ($p = 784$) et le domaine \mathcal{X} est continu entre 0 et 1: $\mathcal{X}_j = [0, 1] \forall j \in (1, \dots, 784)$.
 - ▶ La réponse est l'étiquette indiquant de quel chiffre il s'agit. $\mathcal{Y} = \{0, 1, \dots, 9\}$.
 - ▶ L'ensemble d'apprentissages comporte 60 000 images étiquetées.

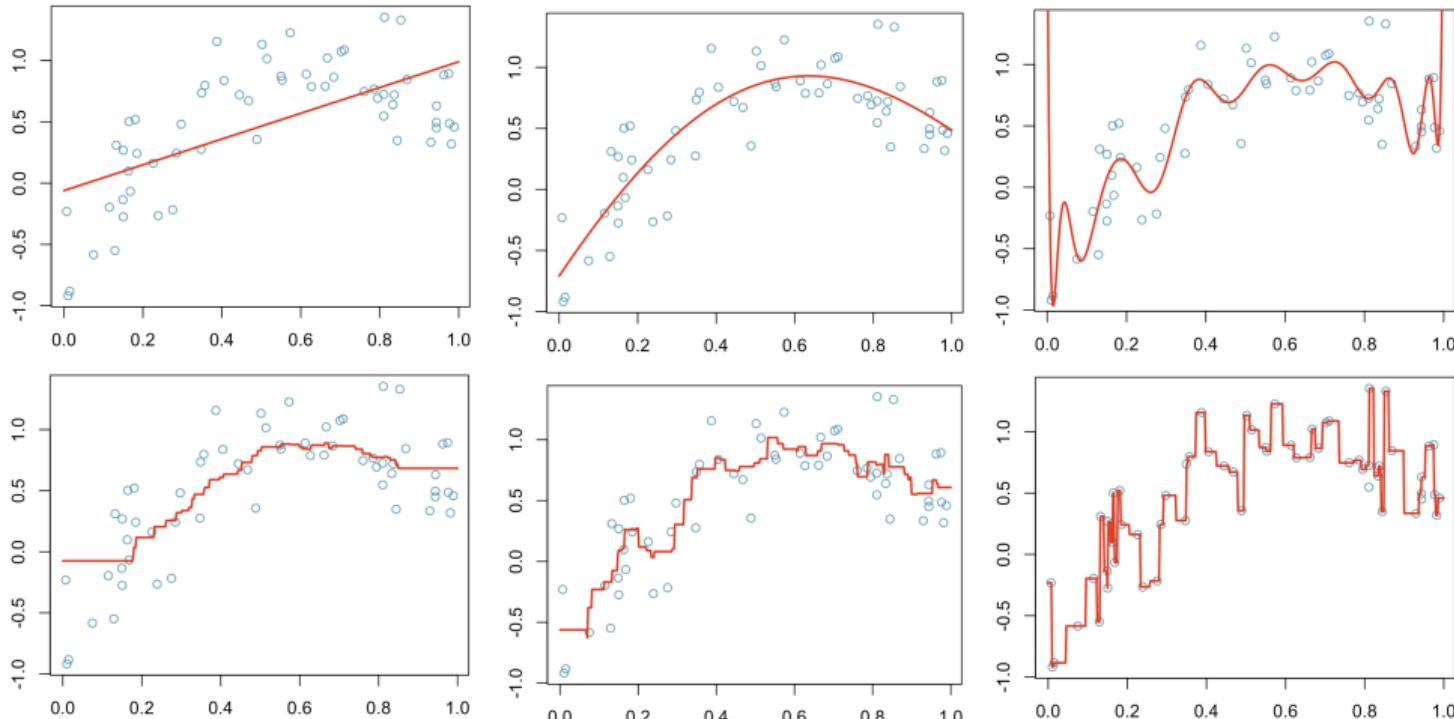


$$x_i \in \mathcal{M}_{28,28}$$

$$y_i \in \{0, 1, \dots, 9\}$$

Sous et sur-apprentissage

\mathcal{M}_k : régression polynomiale de degré k ou k -nearest neighbors



On apprend ce que l'on peut

Le concept qui est partagé pour tous les modèles est d'utiliser les données \mathcal{D}_n , l'échantillon d'entraînement, afin d'estimer f .

En supposant \hat{f} et X est fixe (non-aléatoire)

$$\mathbb{E}(Y - \hat{Y})^2 = \mathbb{E}(f(X) + \varepsilon - \hat{f}(X))^2 = (f(X) - \hat{f}(X))^2 + \text{Var}(\varepsilon)$$

Souvent on parle d'erreur réductible et irréductible.

En gros, on apprend ce qu'on peut, mais ce sera limité. Ce ne sera jamais parfait, il y aura toujours un peu d'aléatoire impossible à capturer.

Comment évalue-t-on \hat{f} ?

Considérant que nous nous concentrons sur la prédiction en apprentissage, nous évaluons \hat{f} quant à sa capacité à faire des prédictions précises.

On veut que $\hat{f}(X)$ soit proche de $f(X)$.

Si Y est une variable continue, on évalue l'erreur quadratique moyenne de \hat{f} :

$$\text{EQM}(\hat{f}) = \frac{1}{n} \sum_{i=1}^n (\hat{f}(X_i) - Y_i)^2 = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2.$$

Dans le cas où Y est une variable catégorielle, on évalue la précision de prédiction de \hat{f} :

$$\text{Precision}(\hat{f}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(\hat{f}(X_i) = Y_i) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(\hat{Y}_i = Y_i),$$

correspondant à la proportion d'observations correctement prédites/étiquetées.

Comment évalue-t-on \hat{f} ?

Un concept fondamental de l'apprentissage statistique est d'évaluer \hat{f} sur de nouvelles observations.

Le principe c'est que d'évaluer \hat{f} sur des observations utilisées lors de l'apprentissage c'est un peu comme tricher; on a déjà vu la réponse.

On n'utilise les données qu'une seule fois: soit pour apprendre ou soit pour évaluer.
Plus en détail, on évalue la performance de \hat{f} sur de nouvelles données afin de prévenir le surapprentissage.

Simplement le surapprentissage c'est d'apprendre une fonction \hat{f} qui reflète des caractéristiques exclusives aux données d'entraînement, à l'échantillon, mais qu'on ne retrouvera pas sur de nouvelles données.

C'est un problème, car on s'en fiche des données d'apprentissage, on connaît déjà Y dans ce cas, ce que l'on veut c'est une bonne performance sur de nouvelles données.

Le surapprentissage, James et al. (2013)

Donc en calculant l'EQM sur des données non observées lors de l'apprentissage, on évalue réellement la capacité de \hat{f} à prédire de nouvelles observations.

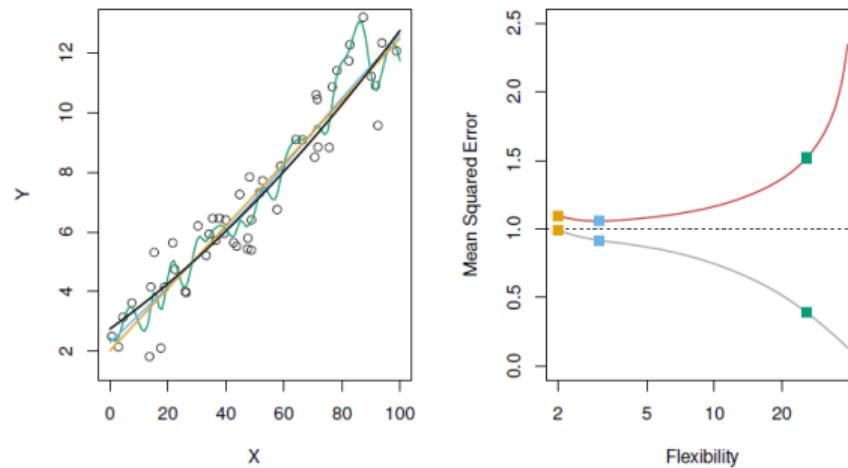


FIGURE 2.10. Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

Figure 2: Extrait du livre James et al. (2013)

Le surapprentissage, James et al. (2013)

Graphiques standard que nous produisons tous en pratique.

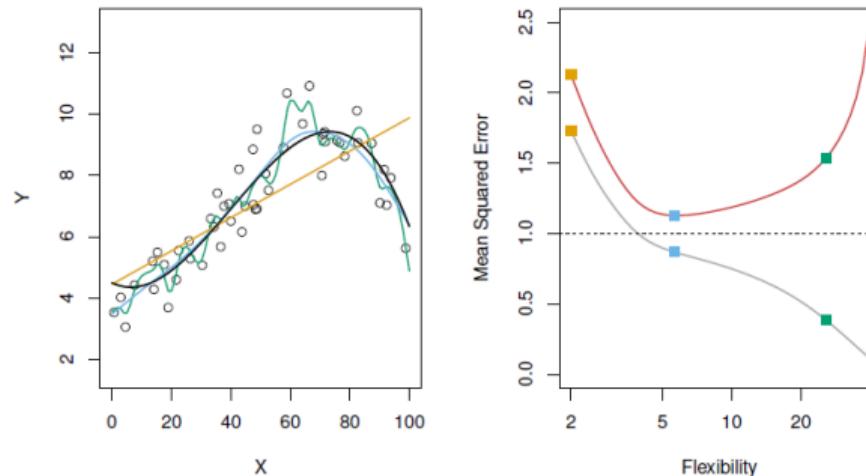


FIGURE 2.9. Left: Data simulated from f , shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

Figure 3: Extrait du livre James et al. (2013)

Comment évalue-t-on \hat{f} ?

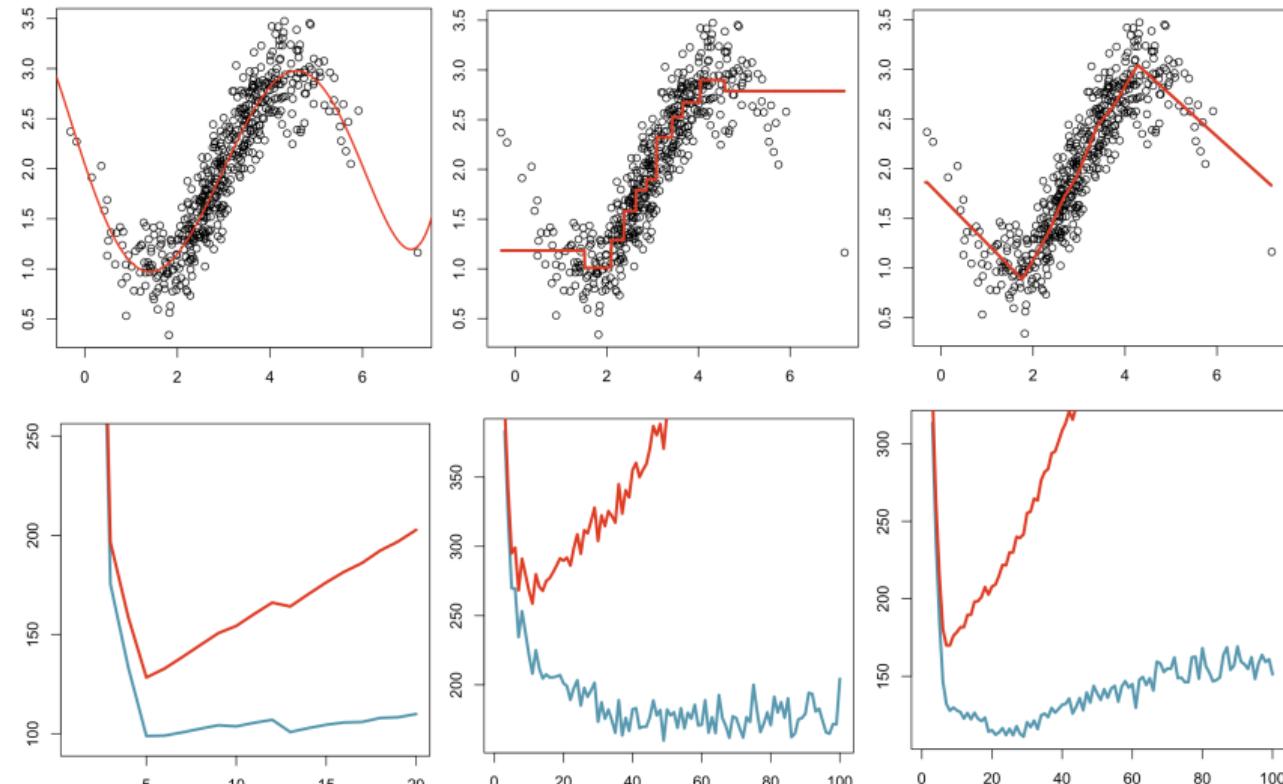
En pratique, on divise notre échantillon \mathcal{D}_n en deux sous-échantillons \mathcal{D}_{ent} et \mathcal{D}_{test} .

On entraîne plusieurs modèles sur \mathcal{D}_{ent} , la base d'entraînement.

On compare leur EQM sur \mathcal{D}_{test} la base de test (ou de validation)

Base d'entraînement et base de validation

Régression polynomiale, constante par morceaux, et linéaire par morceaux (splines)



Flexibilité de modèles

On n'en parlera qu'intuitivement dans le contexte de ce cours.

Plus \hat{f} est d'une forme flexible, plus on peut expliquer ou imiter une fonction f compliquée, mais plus on s'expose au surapprentissage.

Quand on augmente donc la flexibilité d'un modèle, c'est important de vérifier notre performance sur de nouvelles observations.

On calibre souvent la flexibilité en regardant les graphiques des slides précédentes.

On verra une procédure plus concrète au cours # 3.

Compromis Biais-Variance, James et al. (2013)

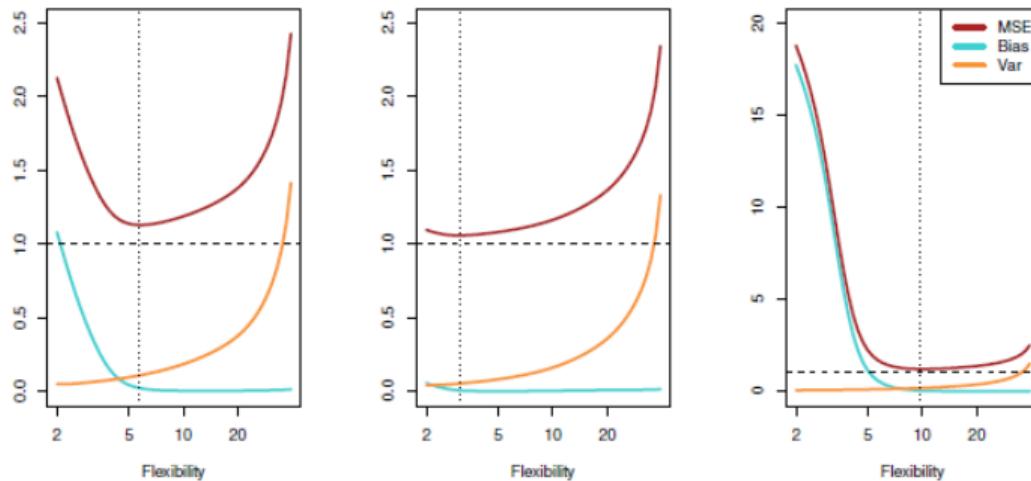


FIGURE 2.12. Squared bias (blue curve), variance ($\text{Var}(\epsilon)$ (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

Figure 4: Extrait du livre James et al. (2013)

Compromis Biais-Variance

Un simple calcul de MSE donne (W)

$$\text{MSE} = \mathbb{E}[(y - \hat{f})^2] = \mathbb{E}[y^2 - 2y\hat{f} + \hat{f}^2] = \mathbb{E}[y^2] - 2\mathbb{E}[y\hat{f}] + \mathbb{E}[\hat{f}^2]$$

où $y = f + \varepsilon$, de telle sorte que, tout d'abord

$$\begin{aligned}\mathbb{E}[y^2] &= \mathbb{E}[(f + \varepsilon)^2] \\&= \mathbb{E}[f^2] + 2\mathbb{E}[f\varepsilon] + \mathbb{E}[\varepsilon^2] \quad \text{par linéarité de } \mathbb{E} \\&= f^2 + 2f\mathbb{E}[\varepsilon] + \mathbb{E}[\varepsilon^2] \quad \text{comme } f \text{ ne dépend pas des données} \\&= f^2 + 2f \cdot 0 + \sigma^2 \quad \text{comme } \varepsilon \text{ est centré, et de variance } \sigma^2\end{aligned}$$

Compromis Biais-Variance

Ensuite,

$$\begin{aligned}\mathbb{E}[y\hat{f}] &= \mathbb{E}[(f + \varepsilon)\hat{f}] \\ &= \mathbb{E}[f\hat{f}] + \mathbb{E}[\varepsilon\hat{f}] \quad \text{par linéarité de } \mathbb{E} \\ &= \mathbb{E}[f\hat{f}] + \mathbb{E}[\varepsilon]\mathbb{E}[\hat{f}] \quad \text{comme } \hat{f} \text{ et } \varepsilon \text{ sont indépendants} \\ &= f\mathbb{E}[\hat{f}] \quad \text{comme } \mathbb{E}[\varepsilon] = 0\end{aligned}$$

Enfin,

$$\mathbb{E}[\hat{f}^2] = \text{Var}(\hat{f}) + \mathbb{E}[\hat{f}]^2$$

Comme $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ pour toute variable X En regroupant,

$$\begin{aligned}\text{MSE} &= f^2 + \sigma^2 - 2f\mathbb{E}[\hat{f}] + \text{Var}[\hat{f}] + \mathbb{E}[\hat{f}]^2 \\ &= (f - \mathbb{E}[\hat{f}])^2 + \sigma^2 + \text{Var}[\hat{f}] = \text{Bias}[\hat{f}]^2 + \sigma^2 + \text{Var}[\hat{f}]\end{aligned}$$

Compromis Biais-Variance

Aussi,

$$\mathbb{E}(y_i - \hat{f}(\mathbf{x}_i))^2 = \text{Var}(\hat{f}(\mathbf{x}_i)) + (\text{Biais}(\hat{f}(\mathbf{x}_i)))^2 + \text{Var}(\varepsilon),$$

où

- ▶ $\text{Var}(\hat{f}(\mathbf{x}_i)) = \mathbb{E}[(\hat{f}(\mathbf{x}_i) - \mathbb{E}(\hat{f}(\mathbf{x}_i)))^2]$
- ▶ $\text{Biais}(\hat{f}(\mathbf{x}_i)) = \mathbb{E}(\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)) = \mathbb{E}(\hat{f}(\mathbf{x}_i)) - f(\mathbf{x}_i)$
- ▶ $\text{Var}(\varepsilon) = \mathbb{E}[(Y - f(\mathbf{x}_i))^2]$

On voit donc que l'on peut avoir deux fonctions avec la même erreur quadratique moyenne ($\mathbb{E}(y_i - \hat{f}(\mathbf{x}_i))^2$) mais avec différentes variance et biais, c'est pour cela que l'on parle d'un compromis.

Compromis Biais-Variance

Par biais on veut dire, à quel point nous *biasons* le modèle dans une certaine direction.

Dans la régression linéaire on suppose (biais) que la relation est linéaire. C'est une forte hypothèse, c'est un grand biais.

En général, plus on fait d'hypothèse et plus les hypothèses sont fortes, plus le modèle est biaisé.

Par variance, on parle de la flexibilité et variabilité du modèle. Plus une fonction peut exprimer des choses différentes, plus elle est variable.

Donc en général, plus on biaise la fonction, plus on la dirige vers une forme précise, moins elle est variable. À l'inverse, si la modèle peut prendre n'importe quelle forme, il est peu biaisé mais très variable.

Compromis Biais-Variance

On se servira des notions de biais et variance tout au long de la session pour comparer des modèles et comprendre intuitivement certaines modifications à certains modèles. C'est un concept un peu abstrait au début, mais fondamental à comprendre; l'objectif du cours est que vous deveniez confortable à argumenter à l'aide du compromis biais-variance.

- ▶ Modèle #1 : $y = \beta_0 + \beta_1 x$
- ▶ Modèle #2 : $y = \beta_0 + \beta_1 x + \beta_2 x^2$

Quel modèle à le plus grand biais ? Plus grande variance ?

Méthode k plus proches voisins

La méthode des k plus proches voisins est une technique relativement simple **sans apprentissage**, dite non-paramétrique et relativement interprétable.

Par **sans apprentissage**, je veux dire que nous n'apprenons pas une forme fixe pour, \hat{f} mais bien apprenons $\hat{f}(x)$ à chaque fois que l'on prédit un nouveau x .

Étant donnée une nouvelle observation x_0 , on prédit la valeur y_0 par \hat{y}_0 étant donné les k voisins les plus proches dans l'ensemble d'entraînement S_{ent} .

C'est-à-dire, on calcule la distance (Euclidienne, ℓ_2 , si $\mathbf{X} \in \mathbb{R}^p$) entre x_0 et chaque point dans les données d'entraînement, puis on choisit les k points les plus proches.

Ces points forment le voisinage de x_0 , V_0 .

On utilise ceux-ci pour estimer y_0 par \hat{y}_0 .

Méthode k plus proches voisins

Si Y est une variable catégorielle, un problème de classification, chacun des k voisins émet un *vote* et on choisit la classe majoritaire.

$$\mathbb{P}(Y = j | \mathbf{X} = \mathbf{x}_0) = \frac{1}{k} \sum_{i \in V_0} \mathbb{1}(y_i = j)$$

Ensuite, on classifie \mathbf{x}_0 comme étant de la classe avec la plus large proportion/probabilité.

Si Y est une variable continue, un problème de régression, on retourne les moyennes du voisinage:

$$\hat{y}_0 = \frac{1}{k} \sum_{i \in V_0} y_i$$

Exemple du livre, James et al. (2013)

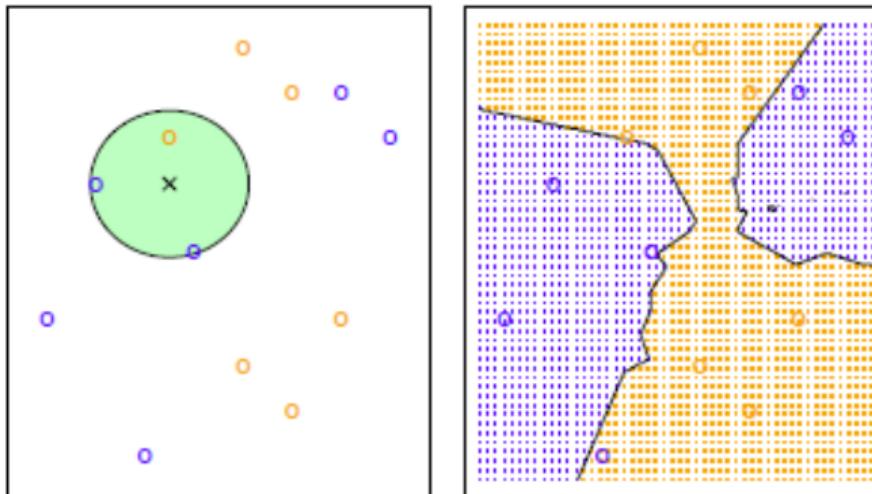


FIGURE 2.14. The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations. Left: a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue. Right: The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.

Exemple du livre, James et al. (2013)

KNN: K=10

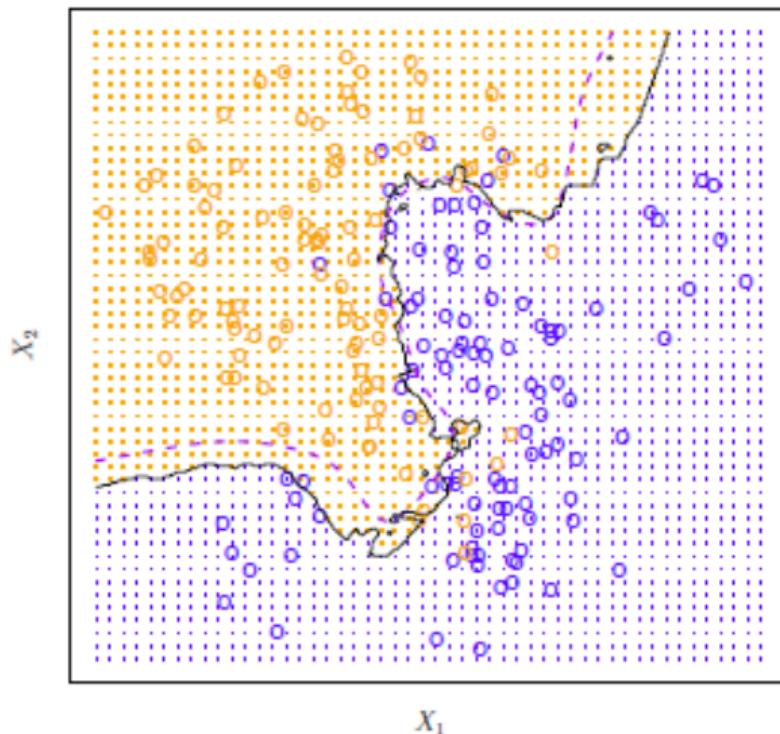


FIGURE 2.15. The black curve indicates the KNN decision boundary on the data from Figure 2.13 using $K=10$. The Rgg3030s decision boundary is shown as

Exemple du livre, James et al. (2013)

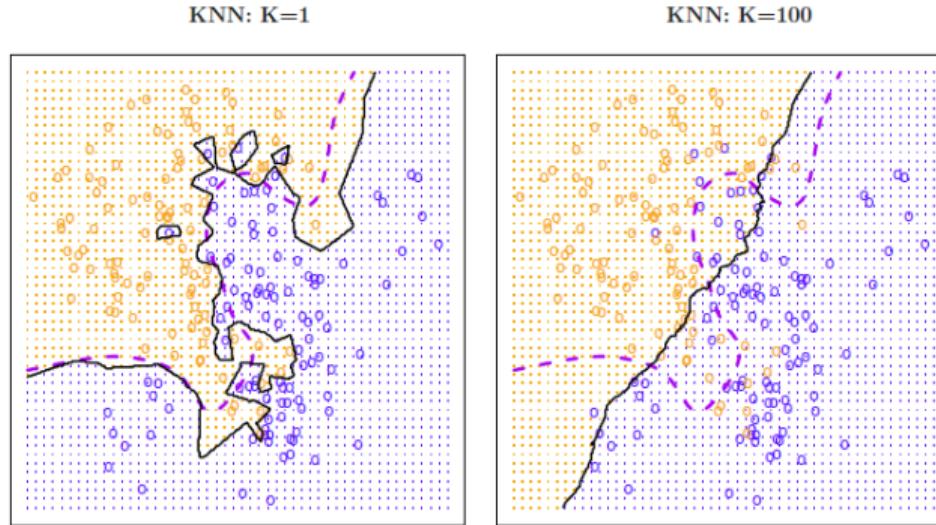


FIGURE 2.16. A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.

Figure 7: Extrait du livre James et al. (2013)

Comment choisir la bonne valeur de k ?

k est ce que l'on appelle un **hyper-paramètre**.

On n'apprend pas de paramètre pour \hat{f} lors de l'apprentissage.

k doit être choisi AVANT de faire de la prédiction, dans ce cas on appelle ça un hyper-paramètres. Presque tous les modèles en ont.

On choisit ceux-ci à l'aide de l'EQM sur S_{test} .

Exemple du livre, James et al. (2013)

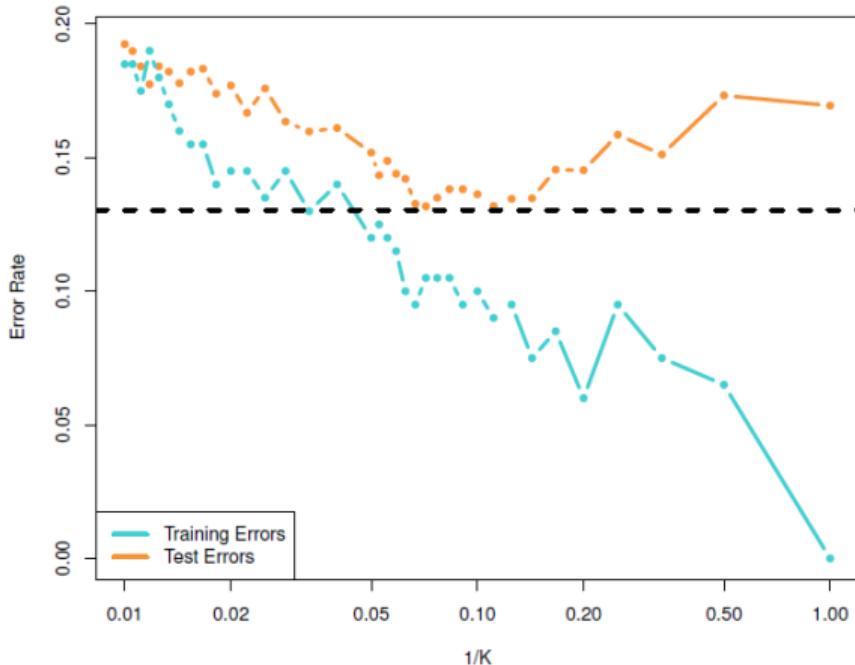


FIGURE 2.17. The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$ on the log scale) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training

Le modèle linéaire

Les modèles linéaires sont caractérisés par une fonction \hat{f} linéaire en X . C'est-à-dire que \hat{f} est une combinaison linéaire des variables explicatives:

$$\hat{Y}_i = \hat{f}(\mathbf{X}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \hat{\beta}_2 X_{i,2} + \cdots + \hat{\beta}_P X_{i,P} = \hat{\beta}_0 + \sum_{p=1}^P \hat{\beta}_p X_{i,p} = \sum_{p=0}^P \hat{\beta}_p X_{i,p} = \mathbf{X}_i^\top \boldsymbol{\beta}$$

(avec la convention usuelle $X_{i,0} = 1$)

On parle de **régression linéaire**, car la variable réponse prédictive \hat{f} est linéaire en $\boldsymbol{\beta}$.

Le modèle linéaire

Il s'agit d'un modèle **paramétrique**, où chacun des β est un paramètre que nous devons apprendre. Les modèles linéaires ont donc une phase d'apprentissage où nous estimons la valeur des coefficients.

Le modèle est aussi **interprétable**, nous pouvons aisément interpréter l'effet des variables explicatives. Si $Y = 2 + 3X$, alors chaque augmentation de 1 unité de la variable X implique donc une augmentation de 3 de la variable réponse.

C'est un modèle que vous devriez tous connaître (étant donné les préalables), faisons un petit rappel néanmoins en employant une perspective de prédiction et d'apprentissage.

Ce modèle nous offre une belle manière de discuter de différentes approches/perspective d'apprentissage.

Maximisation de vraisemblance

Soit $Y = f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_P X_p + \varepsilon = \mathbf{X}^\top \boldsymbol{\beta}$.

Une technique possible d'apprentissage suppose que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, cela implique donc que $Y \sim \mathcal{N}(\mathbf{X}^\top \boldsymbol{\beta}, \sigma^2)$.

Il est possible d'estimer le vecteur de coefficient $\boldsymbol{\beta}$ en maximisant la vraisemblance.

C'est une optimisation relativement simple avec une **solution exacte**

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Ce processus implique naturellement une distribution (normale) sur les estimateurs des β , soit $\hat{\boldsymbol{\beta}}$ et cela nous permet de faire de l'inférence:

- ▶ On peut se demander si les β sont significativement différents de $\mathbf{0}$.
- ▶ On peut se demander si k composantes de $\boldsymbol{\beta}$ sont significativement différentes de $\mathbf{0}$

Par contre, cela vient avec un certain prix, nous avons supposé $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Si cette hypothèse est fausse, toute notre inférence est fausse. Il faut donc vérifier cette hypothèse, et il peut être difficile de le faire.

Plan des moindres carrés

Ici on choisit les paramètres β qui minimise l'EQM sur les données d'entraînement.

On veut minimiser :

$$S(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{X}_i^\top \hat{\beta})^2$$

$$\hat{\beta}^* = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{X}_i^\top \beta)^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

En effet, la somme des carrés des résidus est $S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta)$

$$\mathbf{0} = \frac{\partial S}{\partial \beta}(\hat{\beta}) = \frac{\partial}{\partial \beta} \left(\mathbf{Y}^\top \mathbf{Y} - \beta^\top \mathbf{X}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta \right) \Big|_{\beta=\hat{\beta}} = -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\hat{\beta}$$

ce qui donne l'expression écrite, si $\mathbf{X}^\top \mathbf{X}$ est inversible.

C'est le même estimateur qu'en maximisant la vraisemblance, c'est un *hasard* causé par l'hypothèse $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Exemple du livre, James et al. (2013)

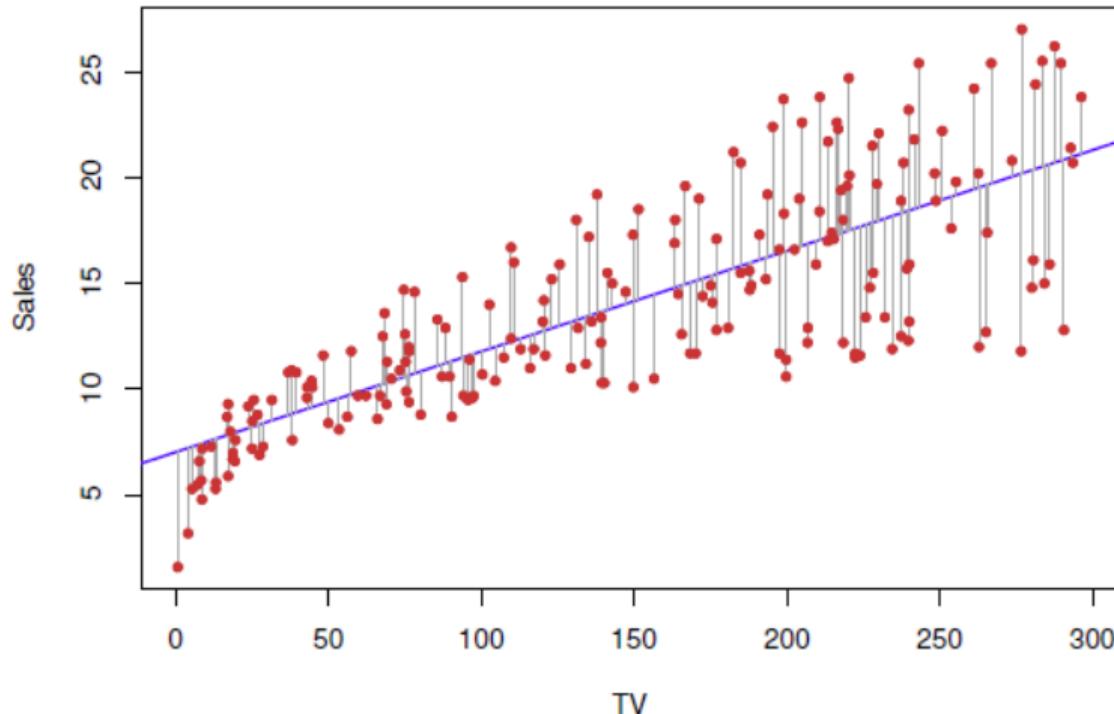


FIGURE 3.1. For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear fit

Exemple du livre, James et al. (2013)

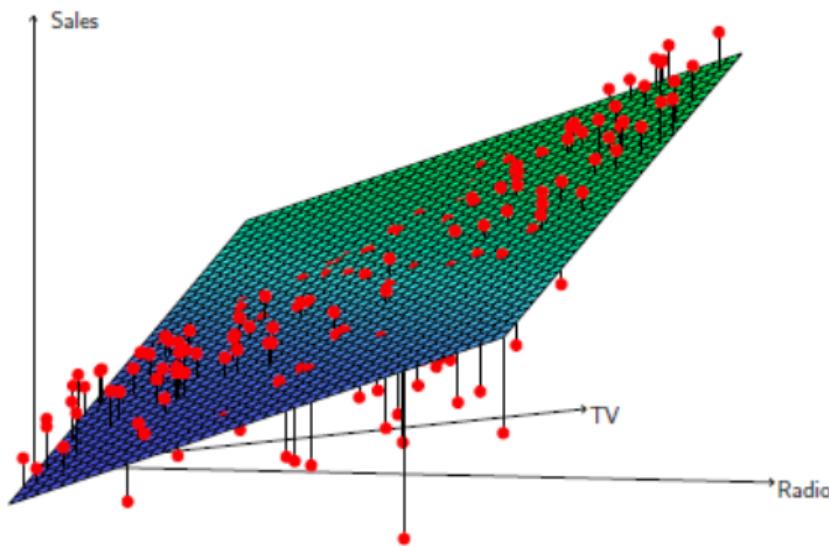


FIGURE 3.5. For the Advertising data, a linear regression fit to sales using TV and radio as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data. The positive residuals (those visible above the surface), tend to lie along the 45-degree line, where TV and Radio budgets are split evenly. The negative residuals (most not visible), tend to lie away from this line, where budgets are more lopsided.

Régression Linéaire : commentaires

- ▶ C'est un bon modèle qui fonctionne presque toujours.
- ▶ On peut faire de la prédiction ET de l'inférence, et le modèle est robuste (l'inférence fonctionne correcte même si les hypothèses sont fausses)
- ▶ C'est jamais un mauvais choix et vous devriez TOUJOURS considérer la régression linéaire par défaut.
- ▶ Pour calculer $(\mathbf{X}^\top \mathbf{X})^{-1}$ il faut pouvoir inverser $(\mathbf{X}^\top \mathbf{X})$, cela peut causer plusieurs ennuis; c'est long lorsque p est grand, impossible si $p > n$ ou si certains prédicteurs sont parfaitement corrélés.
- ▶ On peut généraliser le modèle pour différentes réponses Y
- ▶ On peut généraliser le modèle pour différents prédicteurs \mathbf{X}

Régression Linéaire : en pratique

- ▶ Comme il s'agit d'un modèle bien établi, plusieurs fonctions existent dans plusieurs langages comme R et Python
 - ▶ Peu de raison de ne pas commencer avec ce modèle.
 - ▶ Faire attention aux types de variables et s'assurer que ceux-ci concordent avec les options choisies
 - ▶ Porter particulière attention à la corrélation entre prédicteurs et la quantité de ceux-ci: $(\mathbf{X}^\top \mathbf{X})$ doit être inversible.

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 1 & 0 \end{bmatrix} \quad \mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 6 & -4 \\ 2 & -4 & 6 \end{bmatrix} \quad \mathbf{X}^\top \mathbf{X} + \mathbb{I} = \begin{bmatrix} 5 & 2 & 2 \\ 2 & 7 & -4 \\ 2 & -4 & 7 \end{bmatrix}$$

valeurs propres : $\{10, 6, 0\}$ $\{11, 7, 1\}$

Les valeurs propres de $\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}$ sont $\{10 + \lambda, 6 + \lambda, \lambda\}$

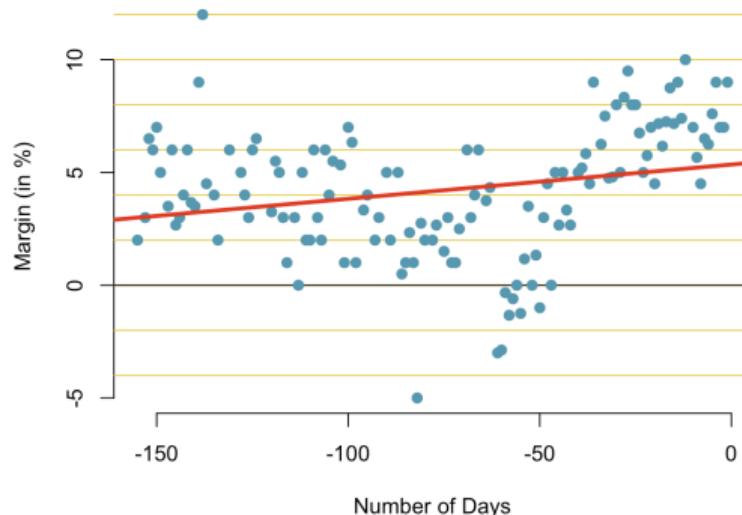
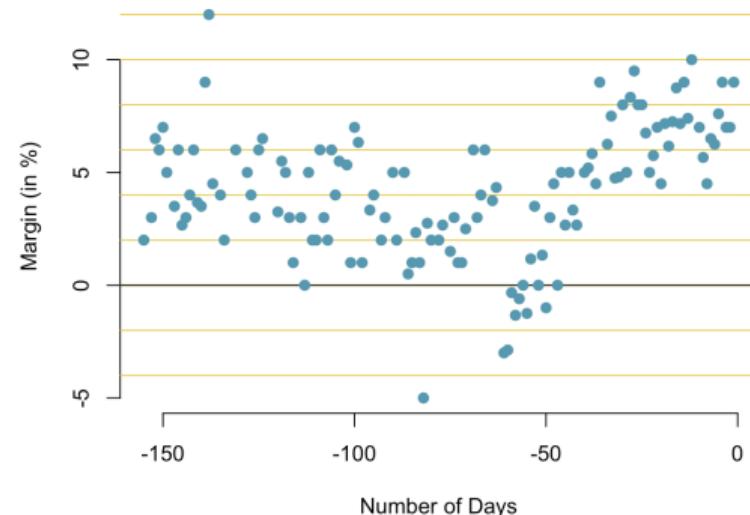
Stratégie ad-hoc: considérer $\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}$, pour $\lambda > 0$ (cf “ridge”).

Natura non facit saltus

We want a continuous function... but probably not linear...

Data source: <http://www.pollster.com/08USPresGEMvO-2.html>

Vote difference between Obama and McCain (2008 US), last 150 days.

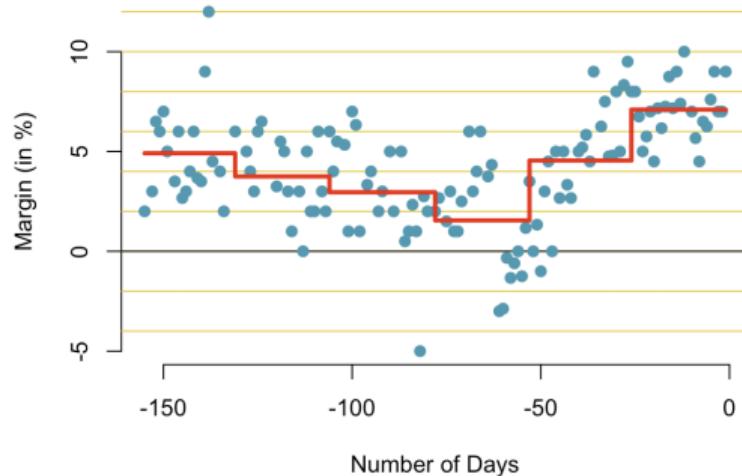
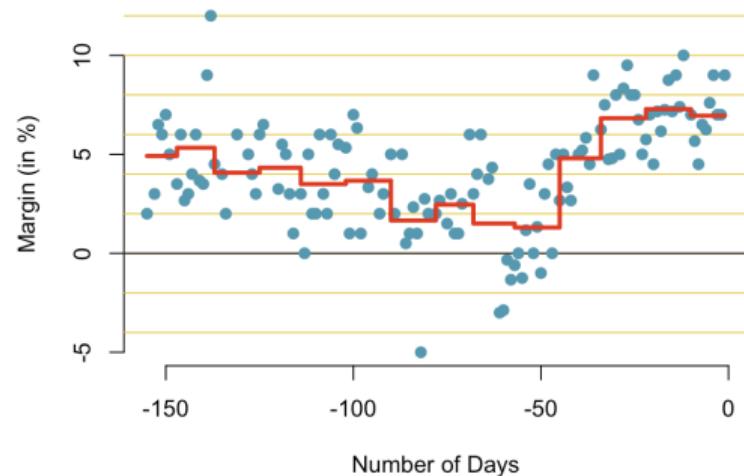


```
1 > library(dslabs)
2 > data("polls_2008")
3 > plot(polls_2008$day, polls_2008$margin*100)
```

Regressogram

From Tukey (1961), the regressogram is defined as

$$\hat{m}_a(x) = \frac{\sum_{i=1}^n \mathbf{1}(x_i \in [a_j, a_{j+1})) y_i}{\sum_{i=1}^n \mathbf{1}(x_i \in [a_j, a_{j+1}))}$$

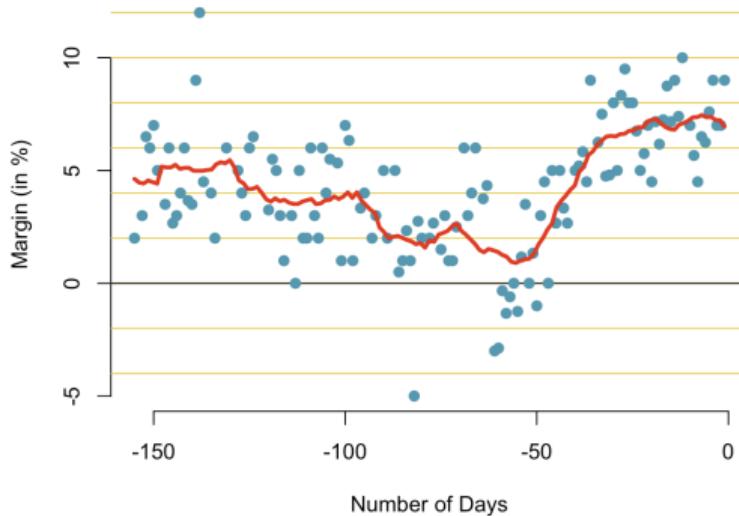
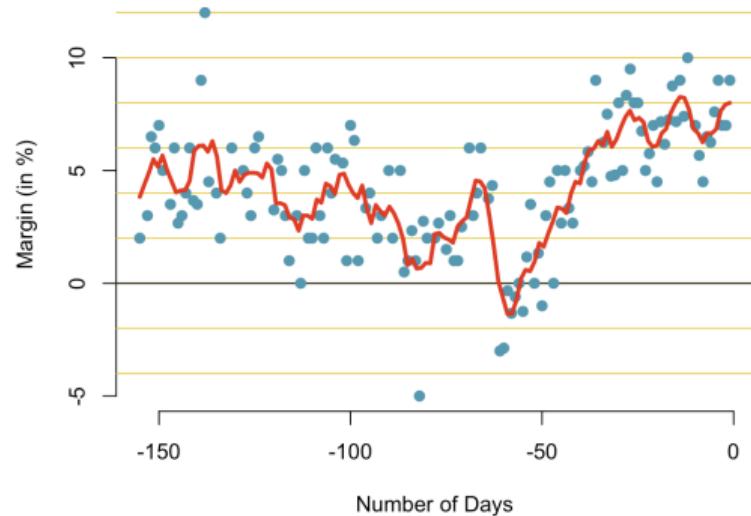


```
1 > reg=lm(margin~cut(day,seq(-160,0,length=15)),data=polls_2008)
```

Moving Regressogram

and the moving regressogram is

$$\hat{m}(x) = \frac{\sum_{i=1}^n \mathbf{1}(x_i \in [x \pm h_n]) y_i}{\sum_{i=1}^n \mathbf{1}(x_i \in [x \pm h_n])}$$



```
1 > with(polls_2008, ksmooth(day, margin, kernel = "box", bandwidth = 7))
```

with **bandwidth** h_n (size of the neighborhood around x)

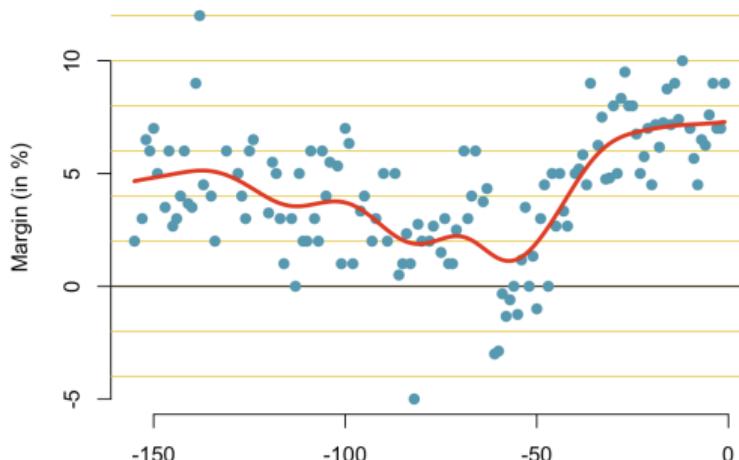
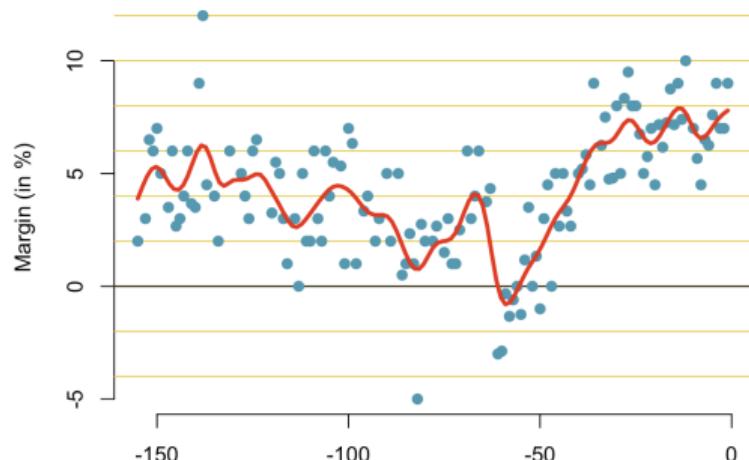
Local Regression

More generally, as moving from the histogram to kernel estimate

$$\tilde{m}(x) = \frac{\sum_{i=1}^n y_i \kappa_h(x - x_i)}{\sum_{i=1}^n \kappa_h(x - x_i)}$$

Observe that this regression estimator is a weighted average

$$\tilde{m}(x) = \sum_{i=1}^n \omega_i(x) y_i \text{ with } \omega_i(x) = \frac{\kappa_h(x - x_i)}{\sum_{i=1}^n \kappa_h(x - x_i)}$$



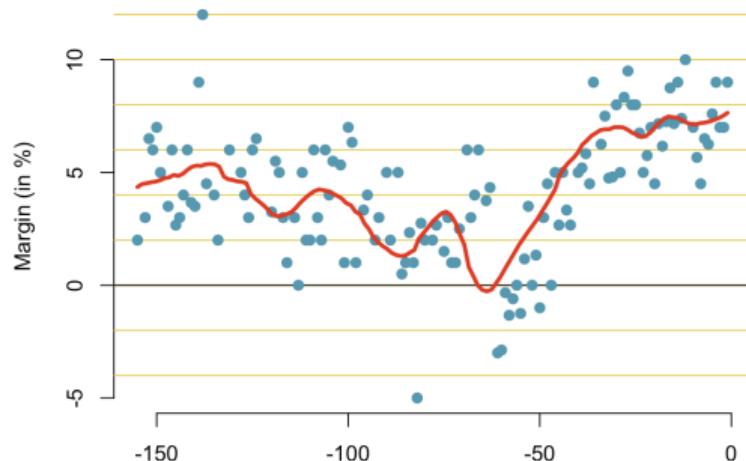
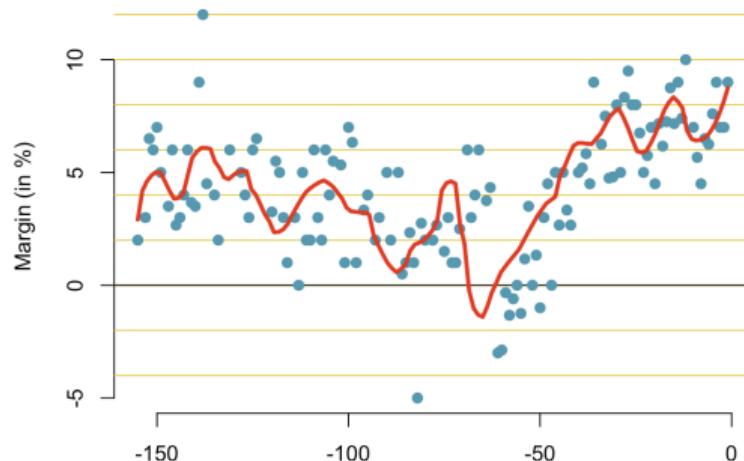
k -Nearest Neighbors

An alternative is to consider

$$\tilde{m}_k(x) = \frac{1}{n} \sum_{i=1}^n \omega_{i,k}(x) y_i$$

where $\omega_{i,k}(x) = \frac{n}{k}$ if $i \in \mathcal{I}_x^k$ with

$$\mathcal{I}_x^k = \{i : x_i \text{ one of the } k \text{ nearest observations to } x\}$$



Local Regression & k -NN

```
1 > fit = with(polls_2008, ksmooth(day, margin, kernel = "normal",
2   bandwidth = span))
2 > lines(fit$x, fit$y)
```

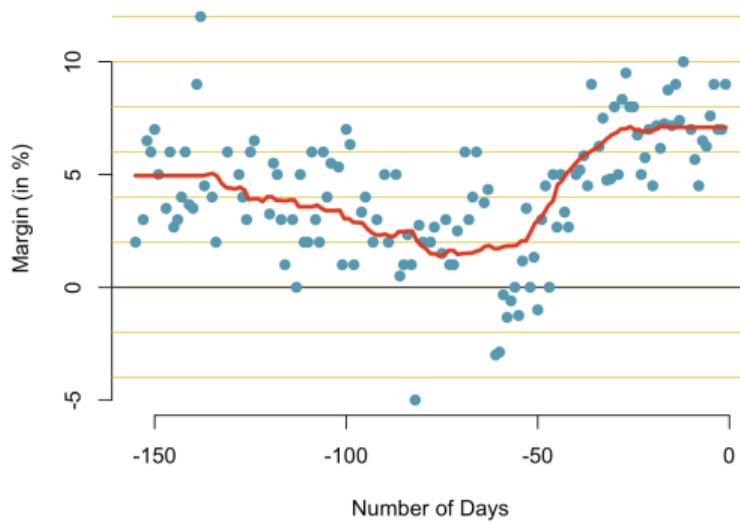
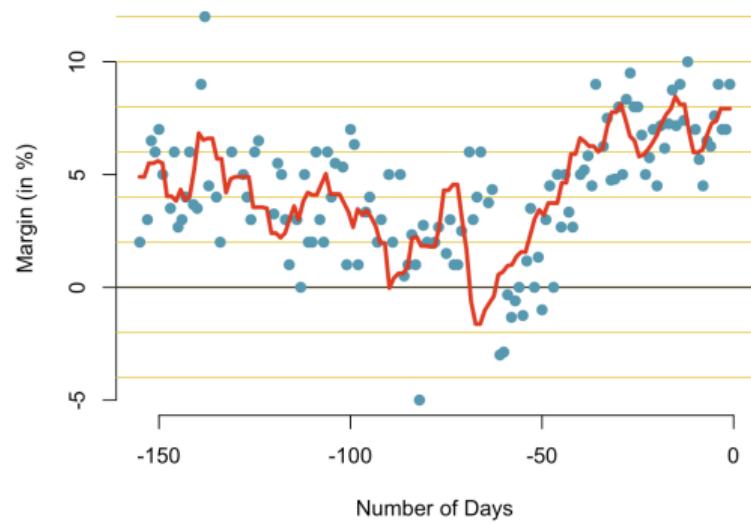
or

```
1 > library(FNN)
2 > p2=knn.reg(train = polls_2008, test = polls_2008, y =
3   polls_2008$margin, k = 25)
3 > lines(polls_2008$day, p2$pred)
```

LOESS (locally weighted polynomial)

Solve

$$\tilde{m}(x) = \operatorname{argmin} \left\{ \sum_{i=1}^n \omega_i(x)(y_i - \alpha - \beta x_i)^2 \right\}, \quad \omega_i(x) = \frac{\kappa_h(x - x_i)}{\sum_{i=1}^n \kappa_h(x - x_i)}$$



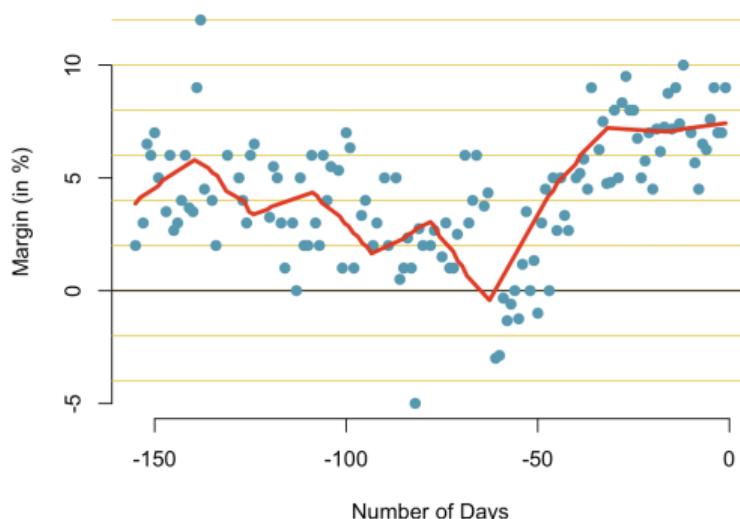
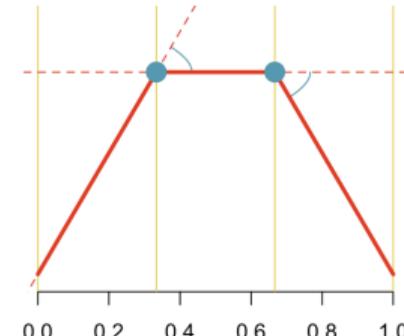
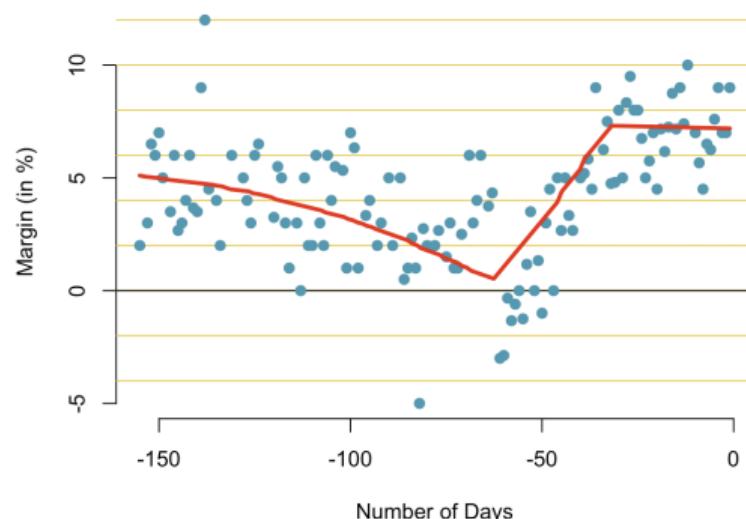
```
1 > fitL = loess(margin ~ day, degree=1, span = 7, data=polls_2008, se=TRUE)
```

(Linear) Spline Regression

Select some knots $\{s_1, \dots, s_k\}$, then with $s_0 = 0$

$$\tilde{m}(x) = \alpha + \sum_{j=0}^k \beta_j (x - s_k)_+$$

where $(x - s)_+ = (x - s)$ if $x > s$, 0 otherwise

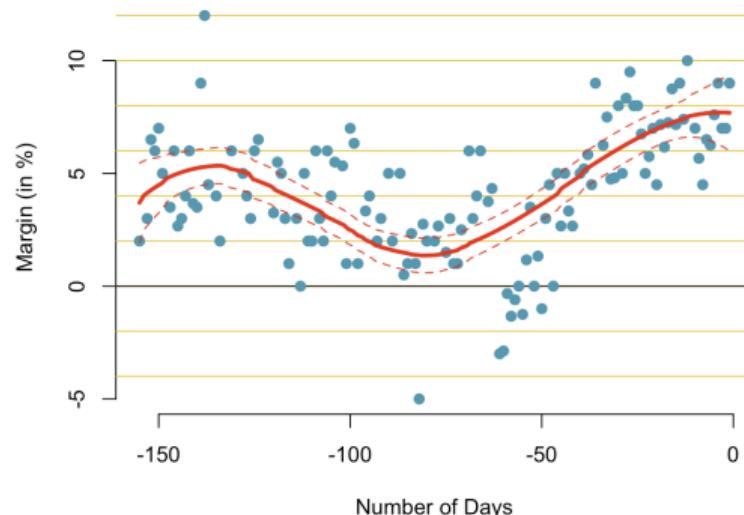
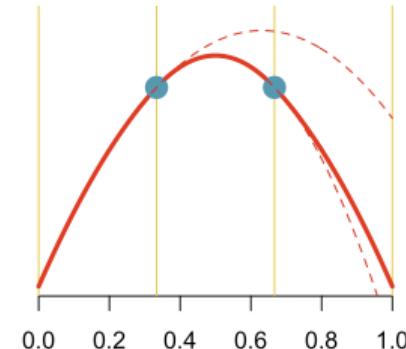
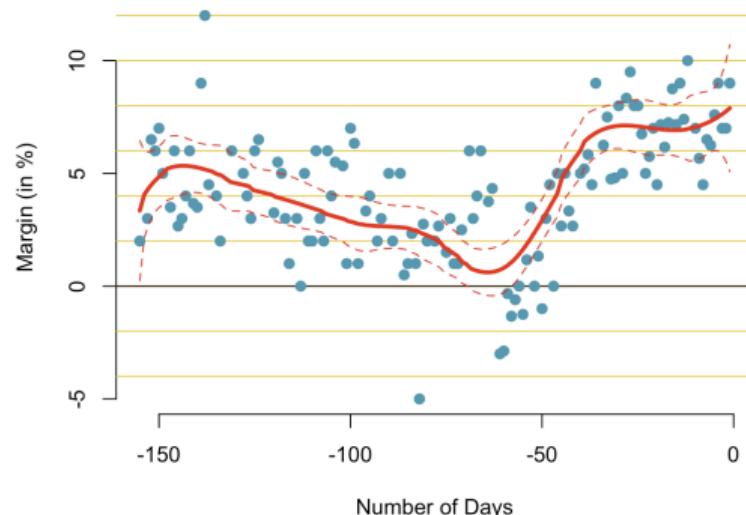


(Quadratic) Spline Regression

Select some knots $\{s_1, \dots, s_k\}$, then with $s_0 = 0$

$$\tilde{m}(x) = \alpha + \gamma x + \sum_{j=0}^k \beta_j (x - s_k)_+^2$$

where $(x - s)_+^2 = (x - s)^2$ if $x > s$, 0 otherwise



Pour le prochain cours

- ▶ Lecture: Chapitre 2 du manuel de référence ([James et al. \(2013\)](#)).
- ▶ Lab: 2.3 (en entier), 3.6.1, 3.6.2, 3.6.3, 3.6.4
- ▶ Exercices 2.4.1, 2.4.2, 2.4.8
- ▶ Préparation: Chapitres 3 et 4 du manuel de référence ([James et al. \(2013\)](#)).

Références

- Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.
- Charpentier, A. (2023). *Insurance: biases, discrimination and fairness*. Springer Verlag.
- Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142.
- James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.
- Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.
- Tukey, J. W. (1961). Curves as parameters, and touch estimation. In Neyman, J., editor, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 681–694. University of California Press.
- Tukey, J. W. (1962). The future of data analysis. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 408–452. Springer.
- Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.