

Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

Intro #2 (courte introduction épistémologique)

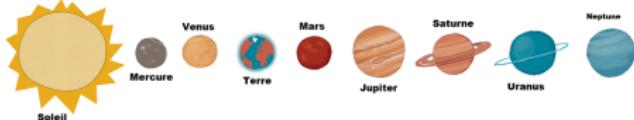
Modélisation

La modélisation est “la construction (intellectuelle) d'un modèle mathématique c'est-à-dire d'un réseau d'équation censé décrire la réalité”, Ivar Ekeland (1995)

- modèles physiques (lois de Kepler, de Galilée)
- modèles biologiques, écologiques ou démographique (loi de Gompertz, de Makeham)
- modèles économétriques (en sciences humaines)

“l'esprit humain reconnaissant l'impossibilité d'obtenir des notions absolues, renonce à chercher l'origine et la destination de l'univers, et à connaître les causes intimes des phénomènes, pour s'attacher uniquement à découvrir, par l'usage bien combiné du raisonnement et de l'observation, leurs lois effectives, c'est-à-dire leurs relations invariables de succession et de similitude ”, Auguste Comte (1830)

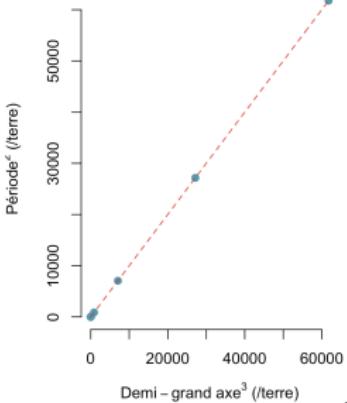
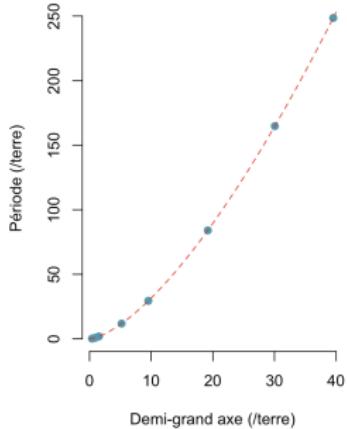
Les modèles en physique (1)



Troisième loi de Kepler (1609) “le carré de la période de révolution est proportionnel au cube du demi grand-axe de l'orbite.”

$$p^2 = \lambda a^3 \text{ ou } \frac{p_1^2}{p_2^2} = \frac{a_1^3}{a_2^3}$$

	planète	a	p
1	Mercure	0.39	0.24
2	Vénus	0.72	0.61
3	Terre	1.00	1.00
4	Mars	1.52	1.88
5	Jupiter	5.20	11.86
6	Saturne	9.54	29.46
7	Uranus	19.19	84.01
8	Neptune	30.06	164.79
10	Pluton	39.53	248.54



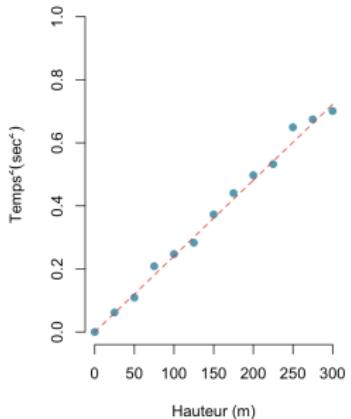
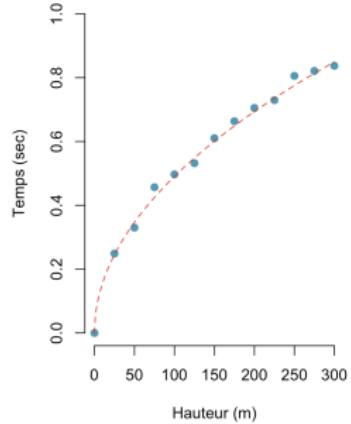
Les modèles en physique (2)

Loi de Galilée (1604) “les distances parcourues par un corps en chute libre sont proportionnelles au carré des temps”

$$d = \lambda t^2 \text{ ou } \frac{d_1}{d_2} = \frac{t_1^2}{t_2^2}$$

	d	t
2 1	25	0.249
3 2	50	0.330
4 3	75	0.457
5 4	100	0.497
6 5	125	0.532
7 6	150	0.610
8 7	175	0.664
9 8	200	0.705
10 9	225	0.730

$$t = \beta \sqrt{d} + \varepsilon \quad (= \text{erreur de mesure})$$



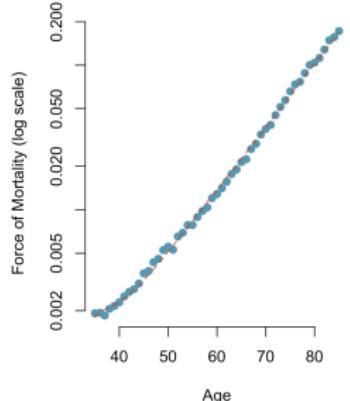
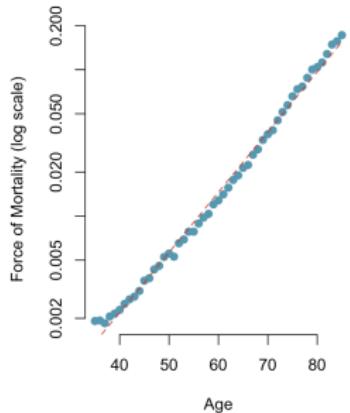
Les modèles en démographie (1)

Gompertz (1825), $\mu_x = bc^x$, "diminution exponentiellement rapide du nombre de vivants avec l'âge"

$$\log \frac{D_x}{E_x} = \beta + \gamma x$$

	x	Dx	Ex
1	35	637	332113
2	36	668	345573
3	37	638	344582
4	38	709	344146
5	39	742	344107
6	40	791	346281
7	41	863	343567
8	42	932	345820
9	43	952	338048

female population living in England & Wales in 1950, see also Halley (1693)



Les modèles en sciences humaines

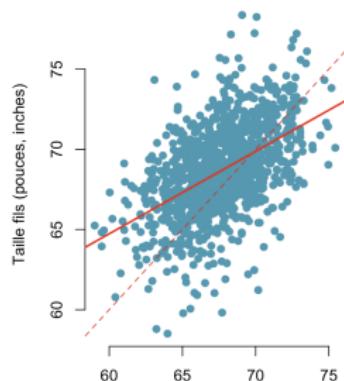
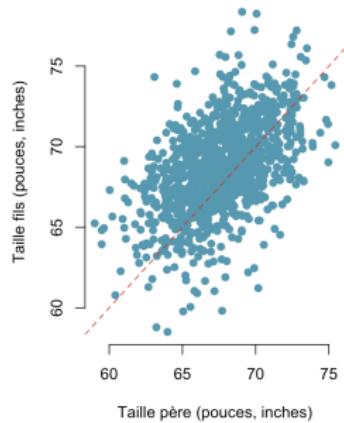
Galton (1886) et Pearson (1903),
taille du père et taille du fils

	father	son
1	65.04851	59.77827
2	63.25094	63.21404
3	64.95532	63.34242
4	65.75250	62.79238
5	61.13723	64.28113
6	63.02254	64.24221
7	65.37053	64.08231

$$\text{taille fils} = \text{taille père} + \varepsilon ?$$

$$\text{taille fils} = \underbrace{\alpha + \beta \text{ taille père}}_{\text{MODELE}} + \varepsilon$$

avec $\beta < 1$: "régression vers la moyenne"



Note: pour l'instant purement descriptif
(on ne cherche pas *pourquoi*)

$$\underbrace{\mathbf{x} = (x_1, \dots, x_p)}_{\text{input}} \xrightarrow{\text{MODELE}} y$$

Notion de classifier en informatique,
 \mathbf{x} est une image

- $y \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
- $y \in \{\text{chat, chien}\}$

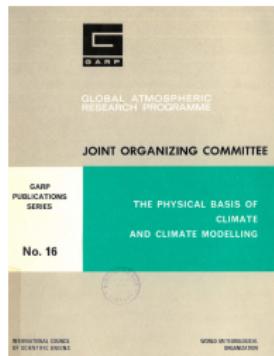
voir "interpretability" et "explainability"...



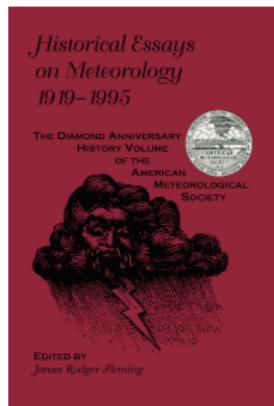
Modèles Climatiques et Météorologiques

4.1.3 Models

As has been briefly outlined in Chapter 3, one of the most promising methods of climate research lies in the development of “mathematical models” for purposes of experiment, hypothesis-testing and prediction. In this context, a “model” is a mathematical formulation of the physical principles that govern one, a few, or many interactive processes affecting the behaviour of the combined system of continents, oceans and atmosphere.



“pour pouvoir traduire en langage mathématique les phénomènes de la nature, il est toujours nécessaire d’admettre des simplifications et de simplifier certaines influences et irrégularités”, Milankovitch (1920)



“It is a good thing we did not have supercomputers at the beginning. There was too much to learn before we could have been ready for them”, George Cressman, in Fleming (1996)

Parcimonie

On veut un **bon** modèle mais aussi un modèle **simple**

Pluralitas non est ponenda sine necessitate

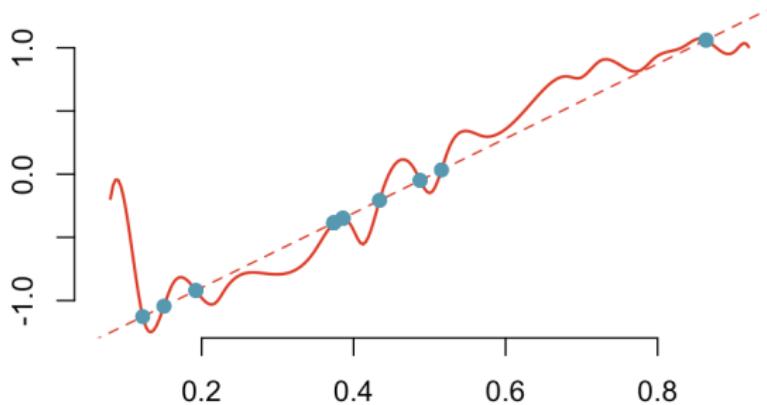
Entia non sunt multiplicanda praeter necessitatem, Guillaume d'Ockham, XIV^e siècle

"la parcimonie est un principe consistant à n'utiliser que le minimum de causes élémentaires pour expliquer un phénomène"



OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."



<http://phdcomics.com>

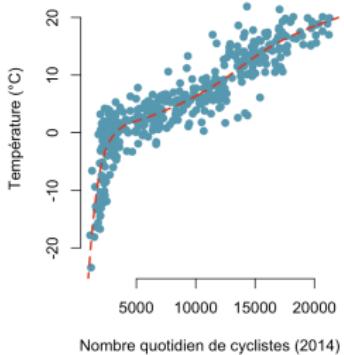
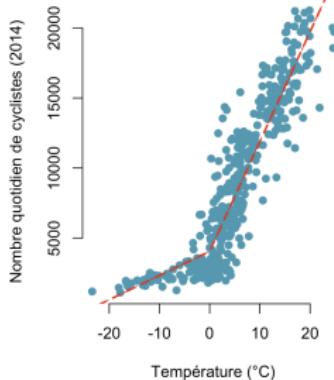
cf aussi **overfit**
(sur-apprentissage)

Modèles, prévision & explication (causale)

Loi de Charpentier (2020) “le nombre de cyclistes sur les routes augmente proportionnellement à la température” (ou presque)

$$\frac{\partial n}{\partial t} = \beta \text{ ou } n = \alpha + \beta t$$

	date	cyclists	meanTemp
1	2014-01-02	3087	0.8
2	2014-01-03	3132	1.4
3	2014-01-13	2352	-10.9
4	2014-01-14	1980	-14.4
5	2014-01-16	2512	-11.9
6	2014-01-17	1639	-13.3
7	2014-01-21	2042	-11.0



“la température augmente proportionnellement avec le nombre de cyclistes sur les routes” ?

$$\frac{\partial t}{\partial n} = b \text{ ou } t = a + bn$$

Modèles Linéaires ?

On va construire des modèles ayant une **forme algébrique** simple,

$$\underbrace{\text{poids}}_y = \beta_0 + \beta_1 \underbrace{\text{taille}}_{x_1} + \beta_2 \underbrace{\mathbf{1}(\text{genre} = \text{"homme"})}_{x_2} + \underbrace{\text{"erreur}}_\varepsilon$$

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon$$

où (y, \mathbf{x}) sont donnés, $(\boldsymbol{\beta}, \varepsilon)$ sont inconnus.

Modèle paramétrique, $\boldsymbol{\beta}$ est le paramètre,
et le modèle est linéaire en $\boldsymbol{\beta}$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 \log(x_2) + \varepsilon$$

sera un modèle "linéaire".

Modèles en épidémiologie

(pour aller plus loin...)

Kermack & McKendrick (1927), 3 groupes de proportion (population de taille constante)
 S_t (susceptible) I_t (infected) R_t (recovered)

$$\frac{dS_t}{dt} = -\beta I_t S_t,$$

$$\frac{dI_t}{dt} = \beta I_t S_t - \gamma I_t,$$

$$\frac{dR_t}{dt} = \gamma I_t,$$

Il faut estimer β et γ , cf $R_0 = \frac{\beta}{\gamma}$ (taux de reproduction)

voir <https://www.cdc.gov/coronavirus/2019-ncov/>

