UNIVERSITÉ DE NEUCHÂTEL

La Statistique au Coeur de la Crise de la Science

Yves Tillé
Institut de Statistique
Université d Neuchâtel

24 février 2018

Table des matières

Table des matières

Le côté obscur de la science

- Montrer les côtés obscurs de la science.
- Montrer aussi le côté lumineux de la science.
- Montrer que la science contient son autocritique.

Wikipédia :

La méthode hypothético-déductive est une méthode scientifique qui consiste à formuler une hypothèse afin d'en déduire des conséquences observables futures (prédiction), mais également passées (rétrodiction), permettant d'en déterminer la validité.

Exemple 1 : Ce qu'il ne faut pas faire : "La vitamine delta"

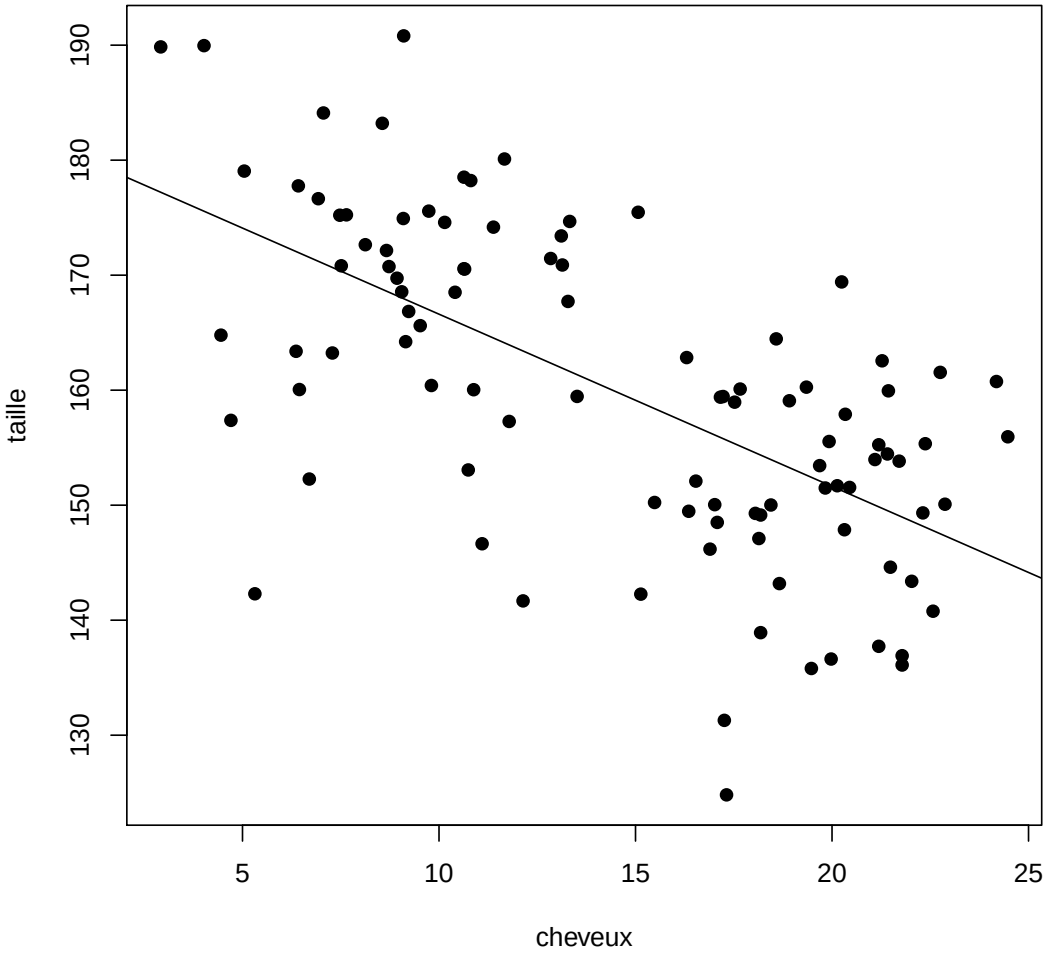
La vitamine delta

Hypothèses et projet de recherche :

- La vitamine delta est indispensable à la croissance.
- Les cheveux absorbent la vitamine delta.
- Le fait d'avoir des cheveux longs ralentit la croissance.

Expériences : Les personnes avec des cheveux longs devraient être plus petites.

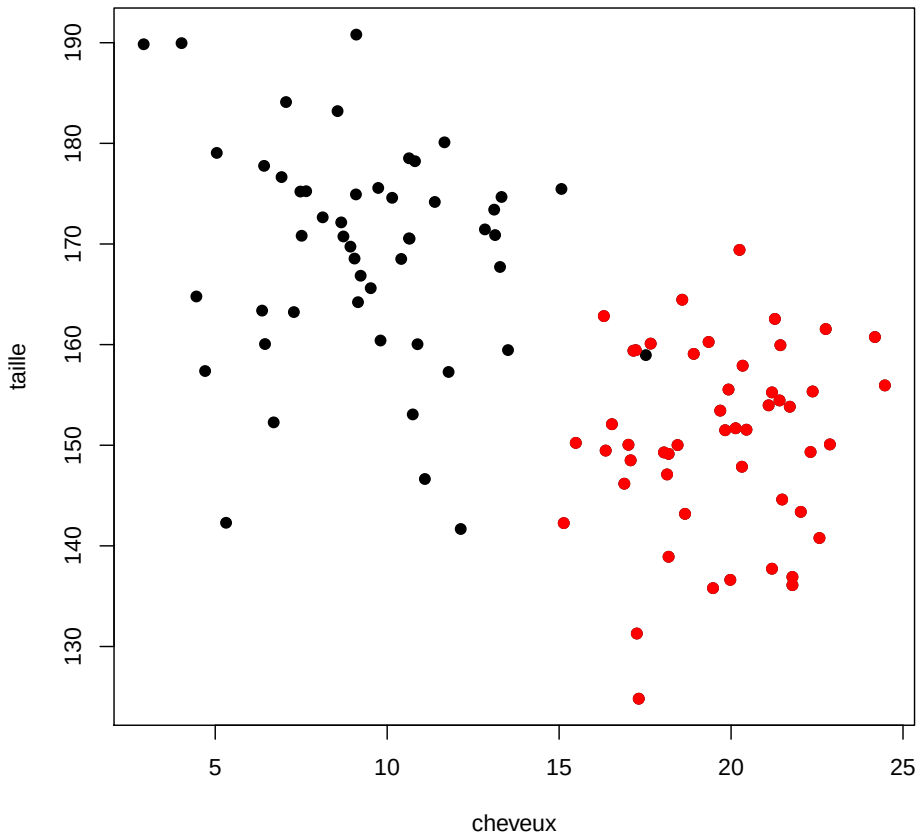
Exemple 1 : "La vitamine delta"



Exemple 1 : "La vitamine delta"

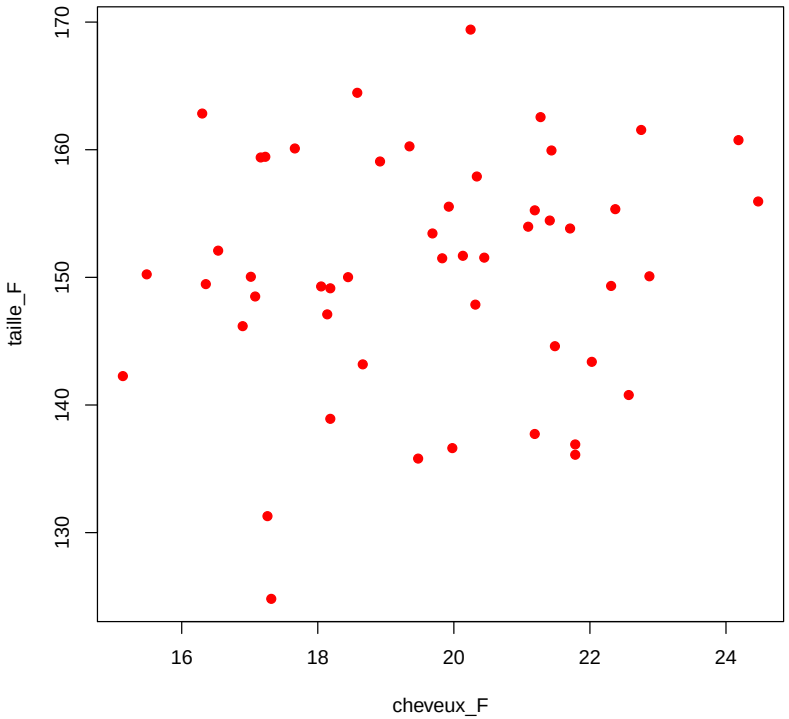
L'hypothèse a été prouvée.
Promotion : 50 francs la boîte de vitamine delta

Exemple 1 : "La vitamine delta"

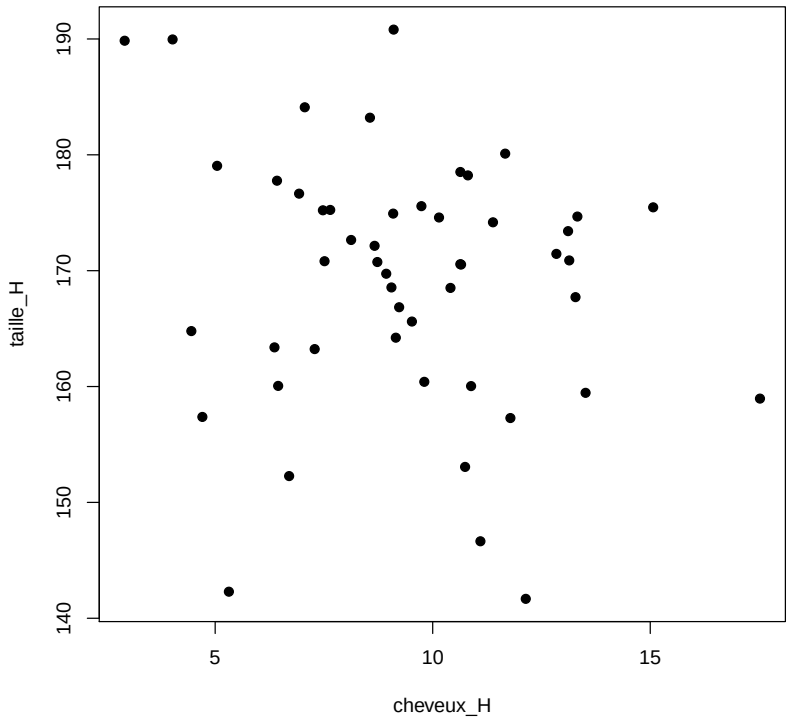


Les hommes sont en noir, les femmes en rouge

Exemple 1 : "La vitamine delta"



Femmes



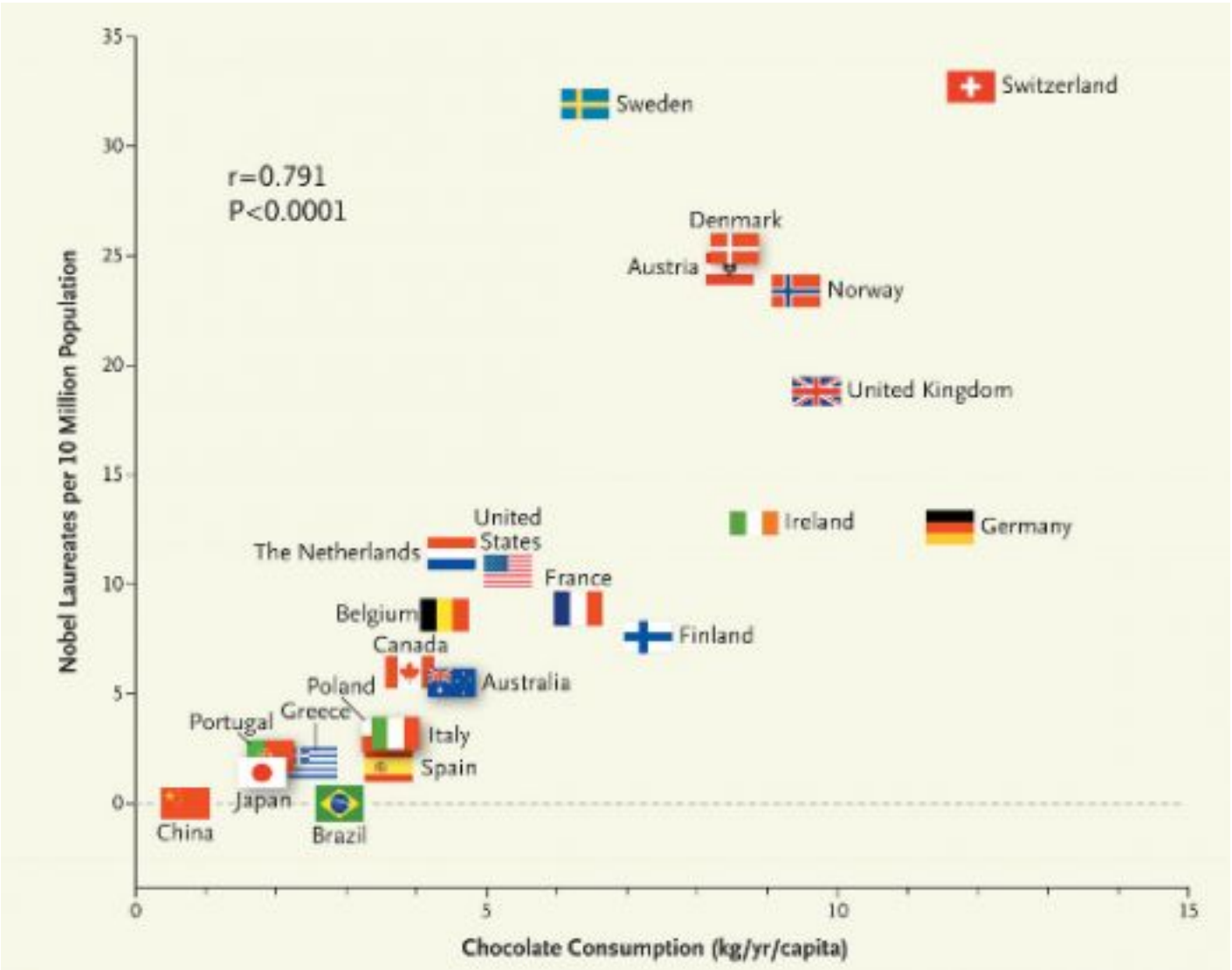
Hommes.

La morale de l'histoire'

La morale de l'histoire

- Le fait que les données soient compatibles avec l'hypothèse ne confirme jamais l'hypothèse.
- Une autre hypothèse pourrait aussi être compatible avec les données.
- L'expérience ne permet que de réfuter une hypothèse, jamais de la valider.
  Principe de réfutabilité de Karl Popper (2005), *Logique de la découverte scientifique*.
- Gare aux variables confondantes (le sexe dans la relation taille/longueur des cheveux).

Chocolat et prix Nobel



Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error

http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

La p-valeur (ou valeur p)

Wikipédia : la p-valeur (ou valeur p)

"Dans un test statistique, la valeur-p (en anglais p-value), parfois aussi appelée p-valeur, est la probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) du test si l'hypothèse nulle était vraie.
Contrairement à ce qui est parfois écrit, la valeur-p n'est pas la probabilité que l'hypothèse nulle soit vraie."

La p-valeur (ou valeur p)

La p-valeur (ou valeur p)

- On considère deux populations : Neuchâtel et Appenzell.
- On se demande si les tailles moyennes des hommes de 18 ans et plus sont égales ou différentes.
- μ_1 taille moyenne des hommes à Neuchâtel.
- μ_2 taille moyenne des hommes à Appenzell.

La p-valeur (ou valeur p)

La p-valeur (ou valeur p)

- Théorie des tests d'hypothèses.
- On fait l'Hypothèse

$$H_0 : \mu_1 = \mu_2 \text{ contre l'hypothèse } H_1 : \mu_1 \neq \mu_2.$$

- On veut prendre une décision sur l'hypothèse H_0.
- On génère des données.

La p-valeur (ou valeur p)

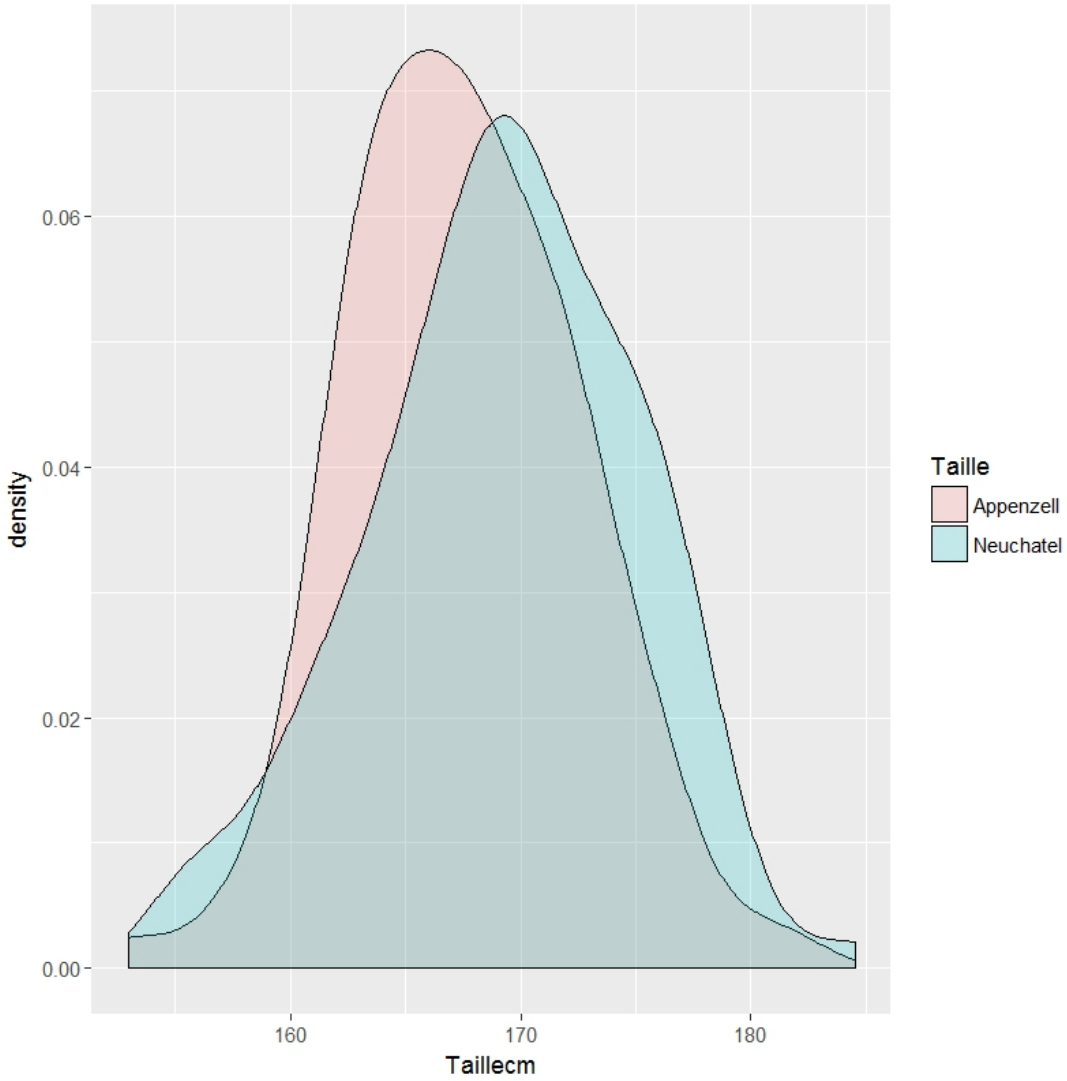
La p-valeur (ou valeur p)

- On sélectionne au hasard 100 Appenzellois et 100 Neuchâtelois.
- On calcule les deux moyennes de l'échantillon

$$\bar{x}_1 = 169.21 \qquad \bar{x}_2 = 167.55$$

- On ne connaît pas les moyennes dans la population μ_1 et μ_2.

Exemple 2 : la p-valeur (ou valeur p)



Exemple 2 : la p-valeur (ou valeur p)

La p-valeur (ou valeur p)

- On suppose que H_0 est vraie et que les moyennes sont égales dans la population.
- Sous H_0, on calcule la probabilité d'obtenir une différence au moins aussi importante que celle obtenue avec les données $169.21 - 167.55 = 1.66$.
- Cette probabilité est appelée la $p - valeur$ (ici 0.03039)
- Si la p valeur est inférieure à une quantité fixée petite (en général 5%), on rejette H_0.
- Ici, on rejette H_0 et on affirme que les tailles sont différentes dans les deux cantons.

Polémique autour de la p valeur

La p-valeur (ou valeur p)

- Beaucoup de chercheurs croient à tort que la p-valeur est la probabilité que H_0 soit vrai.
- La p valeur est la probabilité d'obtenir un résultat au moins aussi extrême si H_0 était vrai.

Polémique autour de la p valeur : Faux positifs

Problème 1 : faux positifs

- Dans le meilleur des mondes, il y a des faux positifs.
- Si H_0 est vrai dans 5% des cas, on a un faux positif et on rejette H_0.

Polémique autour de la p valeur : HARK !!!!!

Problème 2 : HARK

- HARKing : Hypothesizing After the Results are Known.
- C'est dégoutant.
- Les chercheurs testent des tas d'hypothèses et puis ne publient que les p-valeurs significatives.
- La répétition et la sélection des tests amène à une augmentation des faux positifs.
- Il suffit de faire 20 essais cliniques pour avoir une bonne chance d'avoir un faux positif.
- Il suffit d'occulter les expériences non-concluantes pour pouvoir prouver n'importe quoi.
- Obligation de publier tous les essais cliniques (USA 2007).
- Importance des méta-analyses.
- Embarras de l'American Statistical Association.

Polémique autour de la p valeur : Biais de sélection des revues

Problème 3 : Biais de sélection des articles

- Les revues scientifiques aiment publier des résultats importants et novateurs.
- Si une expérience n'est pas probante, il est très difficile de publier les résultats.
- Or, montrer que quelque chose ne marche pas est aussi important que montrer que quelque chose marche.
- En science, on a une obligation de moyen, mais pas une obligation de résultat.
- Le résultat est souvent une question de chance.

Polémique autour de la p valeur : Biais de sélection des revues

Problème 4 : Style fallacieux des articles scientifiques

- Le style des articles scientifiques sont ultra-standardisés.
- On fait une hypothèse. Puis on génère des observations et on teste l'hypothèse.
- En réalité, une recherche ne se passe jamais de manière linéaire.
- Une partie de la démarche est occultée.

Polémique autour de la p valeur : Données massives

Problème 5 : Données massives

- En cas de données massives, tendance à sur-spécifier les modèles.
- Si on a beaucoup de variables potentielles, le fait de répéter des tests pour choisir les variables explicatives va produire des faux positifs.
- La théorie des tests est controversée.
- Objectif du traitement statistique (prédiction/analytique).
- Crise de la p-valeur.
- *Beyond the p-value*.

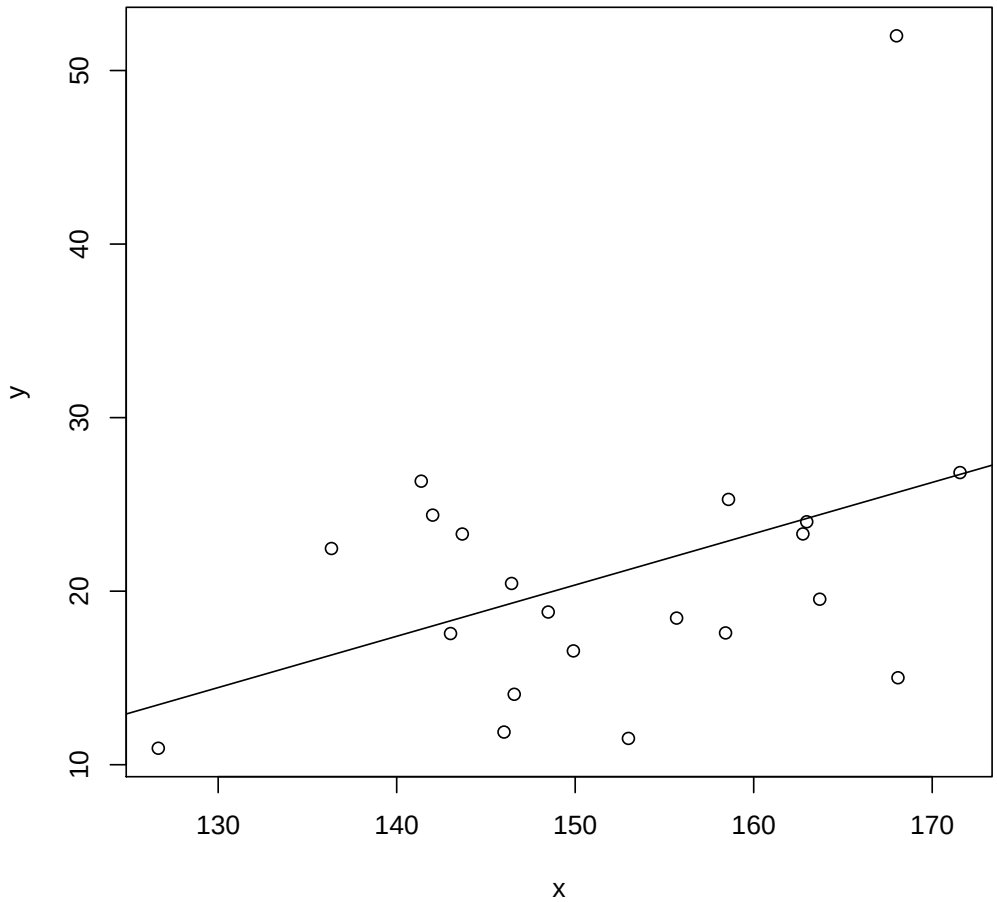
Nouveau paradigme

Nouveau problème : données massives

- Le défi des données massives a remis en question l'approche classique.
- Réflexion sur l'usage de la statistique.
  - Vision analytique : mettre en évidence des relations entre des variables.
  - Vision prédictive : faire un modèle pour pouvoir prédire.

Nouveau paradigme

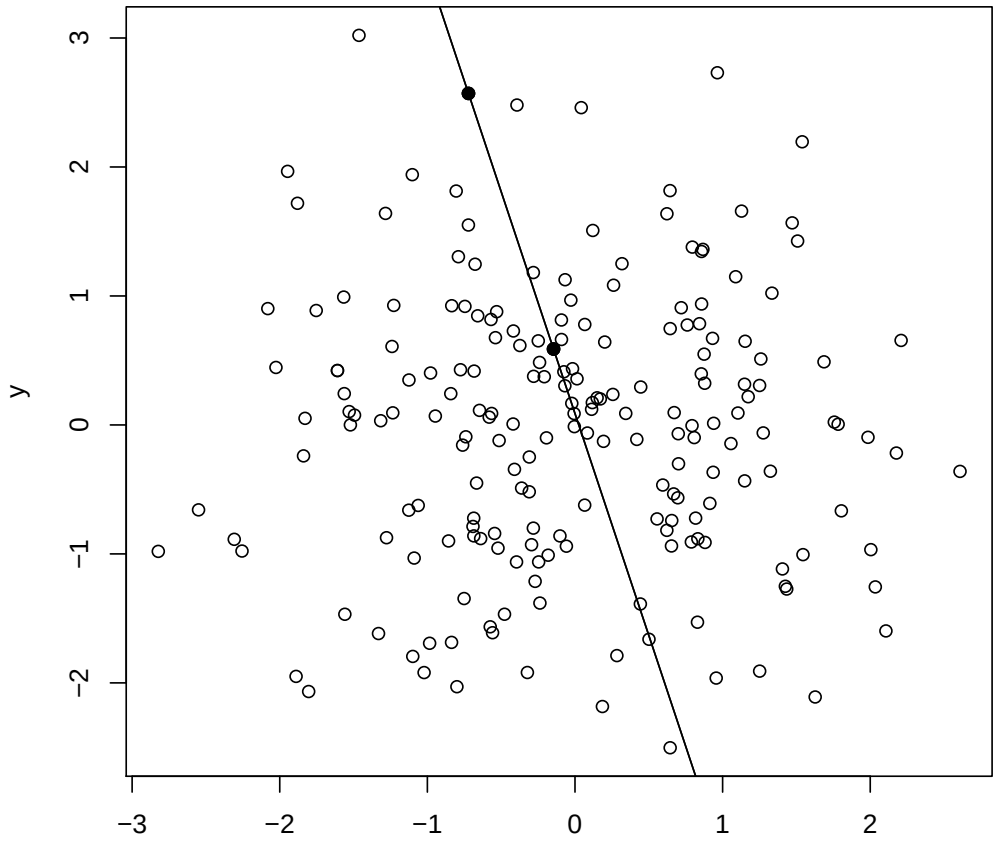
Nouveau problème : données bizarres : point levier



Les modèles s'ajustent trop bien

Example

Les modèles s'ajustent à l'échantillon sélectionné de manière trop optimiste.

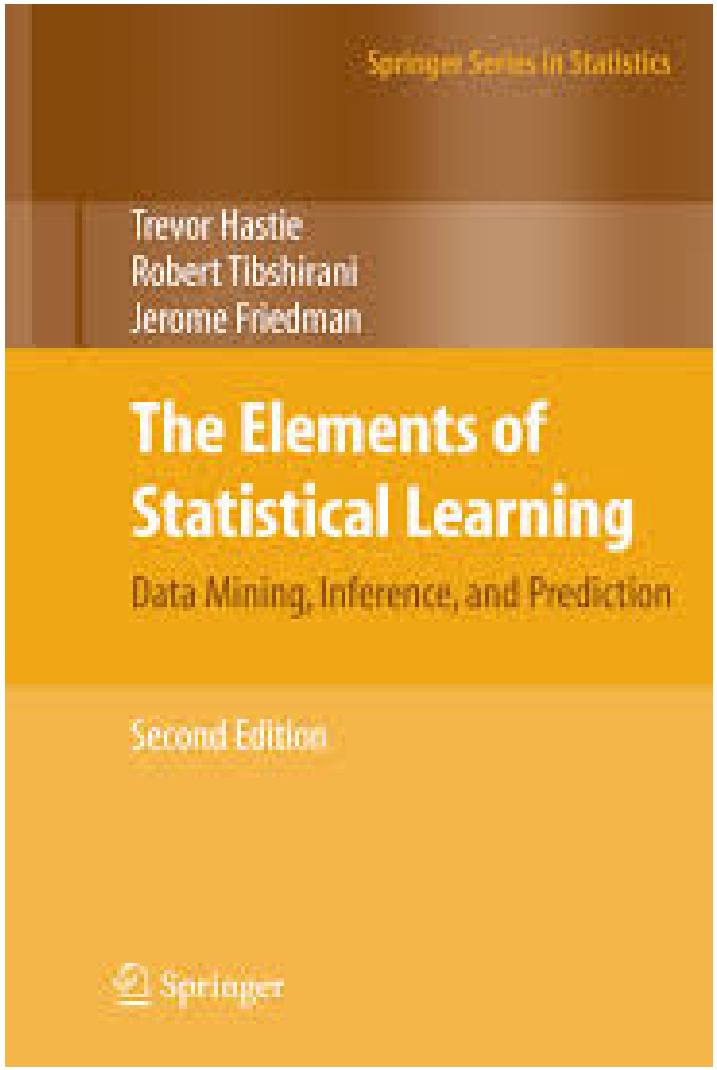


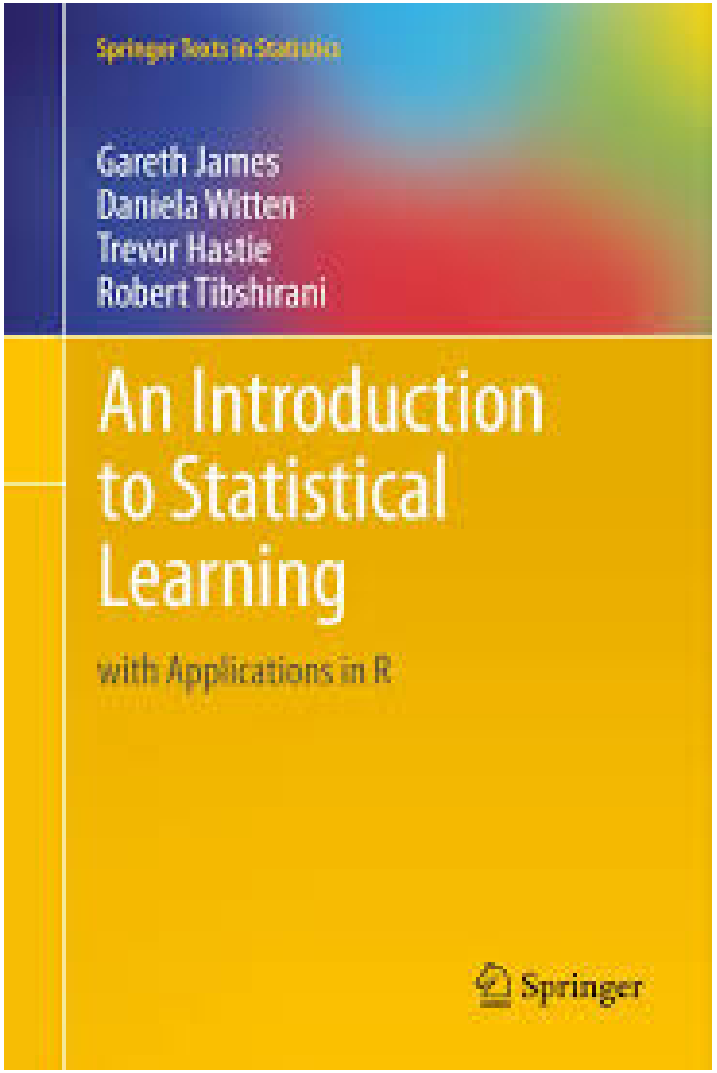
Nouveau paradigme

Validation croisée (cross-validation)

- On enlève une observation (un point).
- On estime le modèle sans cette observation.
- On prédit au moyen du modèle la valeur de l'observation qui n'a pas servi à son estimation.
- On répète cela pour toutes les observations.
- On évalue la qualité des prédictions.

Nouveau paradigme





Robert Tibshirani

Grands éditeurs

Grands éditeurs

- Quelques grands éditeurs scientifiques publient presque tous les livres et revues scientifiques (Elsevier, Wiley, Oxford University Press, North Holland, Springer).
- Les auteurs, les arbitres et les rédacteurs travaillent presque tous sans rémunérations de ces éditeurs.
- Les éditeurs revendent aux universités des abonnements aux revues numériques.
- Le rôle des grands éditeurs scientifiques a été mis en cause.

Open Access

Science ouverte

- Développement de l'*open access.*
- Public Library of Science (PLOS Bilology 2003).
- Science ouverte. Mise en ligne automatique.
- Développement des *repository*.
- PLOS est une organisation sans buts lucratifs et est financé par les auteurs.
- Les auteurs paient quand l'article est accepté.
- Le système fonctionne.
- Open access à l'Université de Neuchâtel.

Open Access

Revues prédatrices

- Des milliers de revues ont été créées. Certaines ont comme unique objectif de gagner de l'argent.
- Mail or Spam or Phishing ou Hameçonnage :
  Exemple d'email
- Exemple d'éditeur prédateur

Jeaffrey Beall

Jeffrey Beall
Bibliothécaire,
puis Professeur associé
à l'Université du Colorado, Denver

Exemple d'hameçonnage

De : International Journal of Structural and Computational Biology [mailto:structuralbiology@symbiosisonline.org]
Envoyé : jeudi 19 janvier 2017 14:18
À : TILLE Yves
Objet : Request to submit manuscript for Inaugural issue: International Journal of Structural and Computational Biology
Importance : Haute
Critère de diffusion : Confidentiel

Dear Dr. Yves Tillé,

Greeting from IJSCB!

Hope you doing well!

We are privileged to introduce International Journal of Structural and Computational Biology (IJSCB) by **Symbiosis** which aims to provide a unique platform for publishing high quality research work. The journal aims to frame up an outstanding issue with scholarly articles focusing on current research. Our motto is to provide easily accessible information in the field of science & technology across the world.

We had a glance at your published article **"Quasisystematic sampling from a continuous population"**

We found your article very innovative, insightful & interesting; we really value your outstanding contribution towards Scientific Community. The time and attention you devoted in presenting a realistic and pragmatic article was really appreciated.

Being impressed by your quality work, we are contacting you to know if you can associate with us by submitting your upcoming research.

We accept any article **(Research Paper, Review Articles, Short Communications, Case Reports, Mini- Review, Opinions, and Letter to Editors etc.)** for publication as inaugural issue of the journal.

Being aware of your expertise and research in the concemed field, we request you to contribute towards our mission of *open access.*

Please submit your manuscript Here or e-mail it to computationalbiology@symbiosisonline.org

IJSCB requests you, to provide your manuscript on or before *February 10th, 2017.*

Note: If need arises, we will extend the date of submission as per your convenience.

Your diligent work and speedy submission will expedite the release of the inaugural issue.

Please do not hesitate to contact us for any further queries.

Hoping for a favorable response and everlasting scientific association with you!

Best Regards,
Hermione Ella,
International Journal of Structural and Computational Biology,
Symbiosis Group

Publications

- Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, 489(7415):179–179
- Beall, J. (2015). Criteria for determining predatory open-access publishers. Scholarly open access https://web.archive.org/web/20161130184313/https://scholarlyoa.files.wordpress.com/2015/01/criteria-2015.pdf,(accessed 2015-02-14)
- Mehrpour, S. and Khajavi, Y. (2014). How to spot fake open access journals. *Learned Publishing*, 27(4):269–274

Un bon business

- Les auteurs sont principalement des chercheurs jeunes et inexpérimentés de pays en développement (Xia et al., 2015).
- En 2014, environ 420'000 articles ont été publiés dans plusieurs milliers de revues prédatrices (Shen and Björk, 2015).
- En moyenne, 178 USD pour publier un article.
- Chiffre d'affaires 75 millions de dollars annuel.

Liste de Beall et de Thomson Reuter

Liste de revues

- Liste de Beall, 1293 journaux.
- Liste de Thomson Reuter, 8761 journaux.
- Peut-on prévoir à quelle liste appartient un journal sur base de son nom ? (Modélisation prédiction).

Données

Partie du tableau :

Nom	y	esp	acta	advanced	advances	american	and
zoosystema	0	0	0	0	0	0	0
zoosystematics and evolution	0	0	0	0	0	0	1
zootaxa	0	0	0	0	0	0	0
zuchtungskunde	0	0	0	0	0	0	0
zygote	0	0	0	0	0	0	0
academic exchange quarterly	1	0	0	0	0	0	0
academic research reviews	1	0	0	0	0	0	0
academy of contemporary research journal	1	0	0	0	0	0	0
acme intellects	1	0	0	0	0	0	0
acta de gerencia ciencia	1	0	1	0	0	0	0
acta advances in agricultural sciences	1	0	1	0	1	0	0
acta kinesiologica	1	0	1	0	0	0	0
acta medica international	1	0	1	0	0	0	0
acta scientiae et intellectus	1	0	1	0	0	0	0
acta velit	1	0	1	0	0	0	0
actual problems of economics	1	0	0	0	0	0	0

Les données

Table – Variables explicatives : Tableau des occurrences des mots

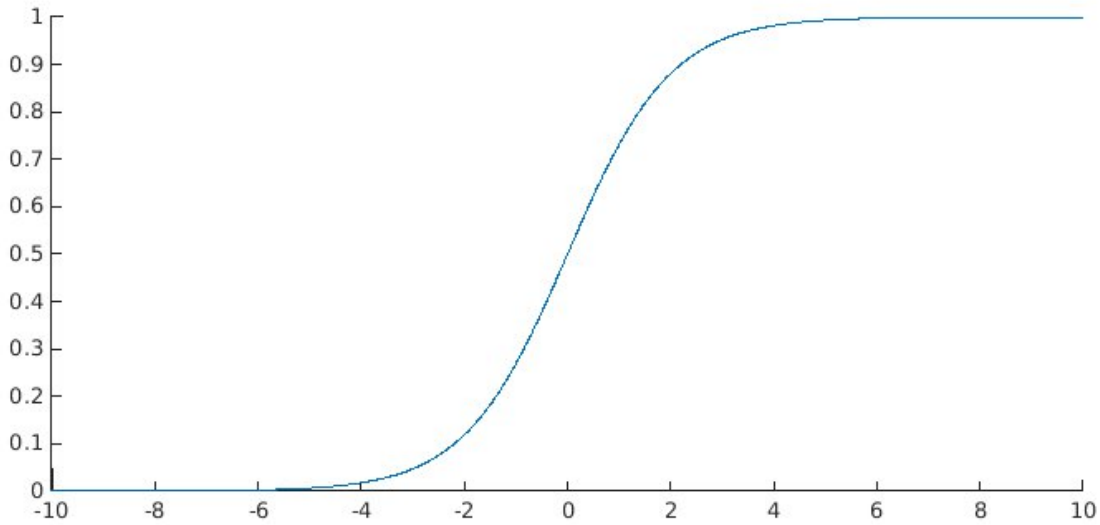
	Mot 1	Mot 2	Mot 3	Mot 4	$\cdots$
Journal 1	0	1	0	0	$\cdots$
Journal 2	0	0	0	0	$\cdots$
Journal 3	0	0	2	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

10054 Revues, 41671 mots (ou formes).

Régression logistique

Régression logistique

- On ne prédit pas une quantité, mais à quelle liste appartient une revue.
- On va prédire une probabilité, autrement dit, une valeur entre 0 et 1.
- Truc : Fonction logistique $F(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$.



Régression logistique

Applications

- On calcule la probabilité d'être dans la liste de Beall.
- $P($ Beall $|$ mots du titre $) =$
  $F(\alpha + \beta_1(\text{mot 1 est présent}) + \cdots + \beta_p(\text{mot } p \text{ est présent}))$.
- Estimation des paramètres par la méthode du maximum de vraisemblance.
- On cherche les valeurs des β qui maximise la probabilité d'obtenir l'échantillon que l'on a observé.

Méthode du maximum de vraisemblance

Ronald Aylmer Fisher, 1890-1962

Crise de la *p*-valeur : Rappel

Crise de la *p*-valeur : Rappel

- En cas de données massives, tendance à sur-spécifier les modèles.
- Si on a beaucoup de variables potentielles, le fait de répéter des tests pour choisir les variables explicatives va produire des faux positifs.
- La théorie des tests est controversée.
- Objectif du traitement statistique (prédiction/analytique).
- Crise de la *p*-valeur.
- *Beyond the p-value*.

Régression Lasso

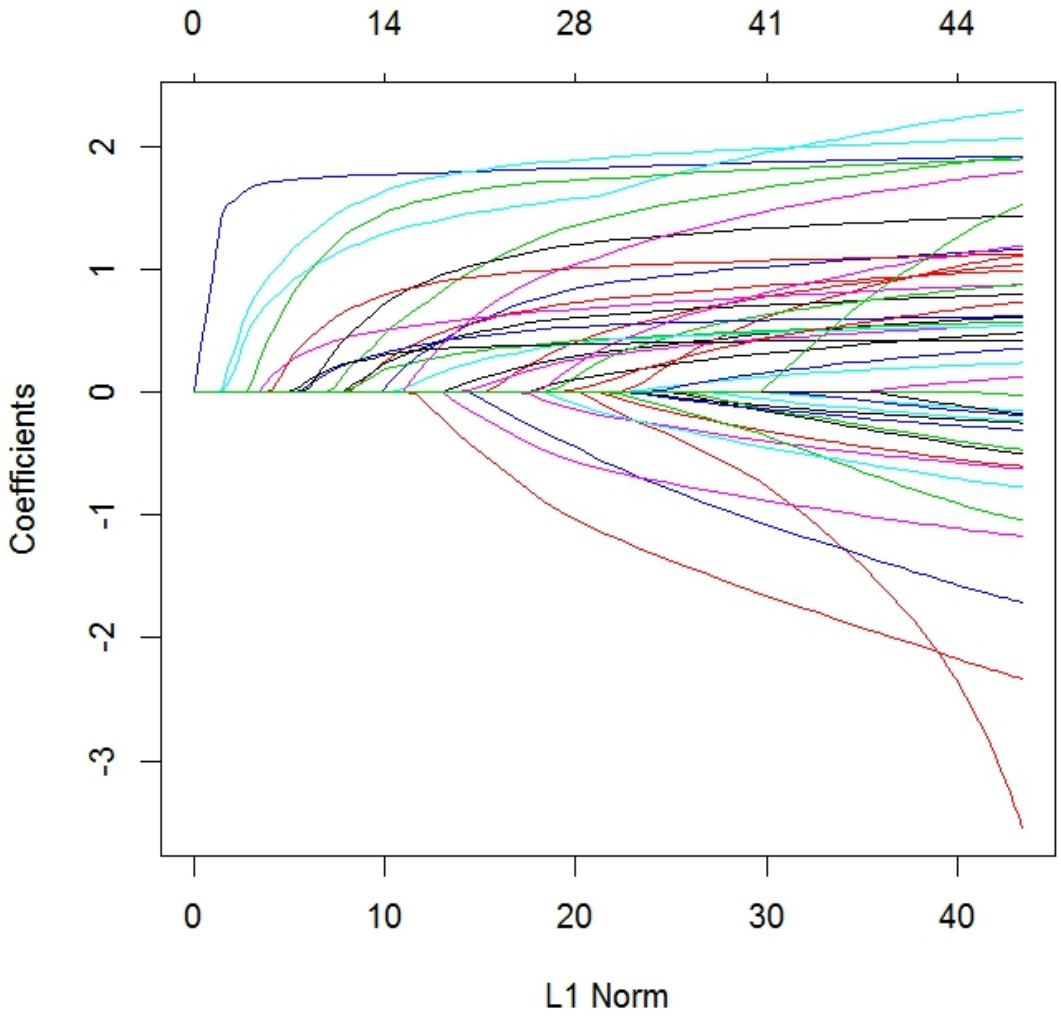
Régression Lasso

- Principe de pénalisation.
- On veut maximiser la vraisemblance mais on veut aussi que les coefficients de régression restent le plus petit possible (en valeur absolue).
- On considère toutes les variables.
- Le problème :
  Maximiser la vraisemblance, sous la contrainte que

$$|\beta_1| + |\beta_2| + \cdots + |\beta_p| \leq \lambda.$$

- $\lambda > 0$ est un paramètre de réglage (tuning).
- Les variables sont standardisées.
- Les variables sont les mots (formes).

Coefficients



Plus λ est petit, plus les coefficients sont proches de 0.
Plus λ est petit, plus le nombre de coefficients nuls augmente.

Intérêt de la méthode

Intérêt de la méthode

- La méthode permet de traiter des très grands tableaux (big data).
- La méthode permet d'identifier les variables.
- Pénaliser protège contre une sur-spécification du modèle.
- Choix du meilleur λ par validation croisée.
- Pas de tests d'hypothèse, pas de p-valeur.

Validation croisée : rappel

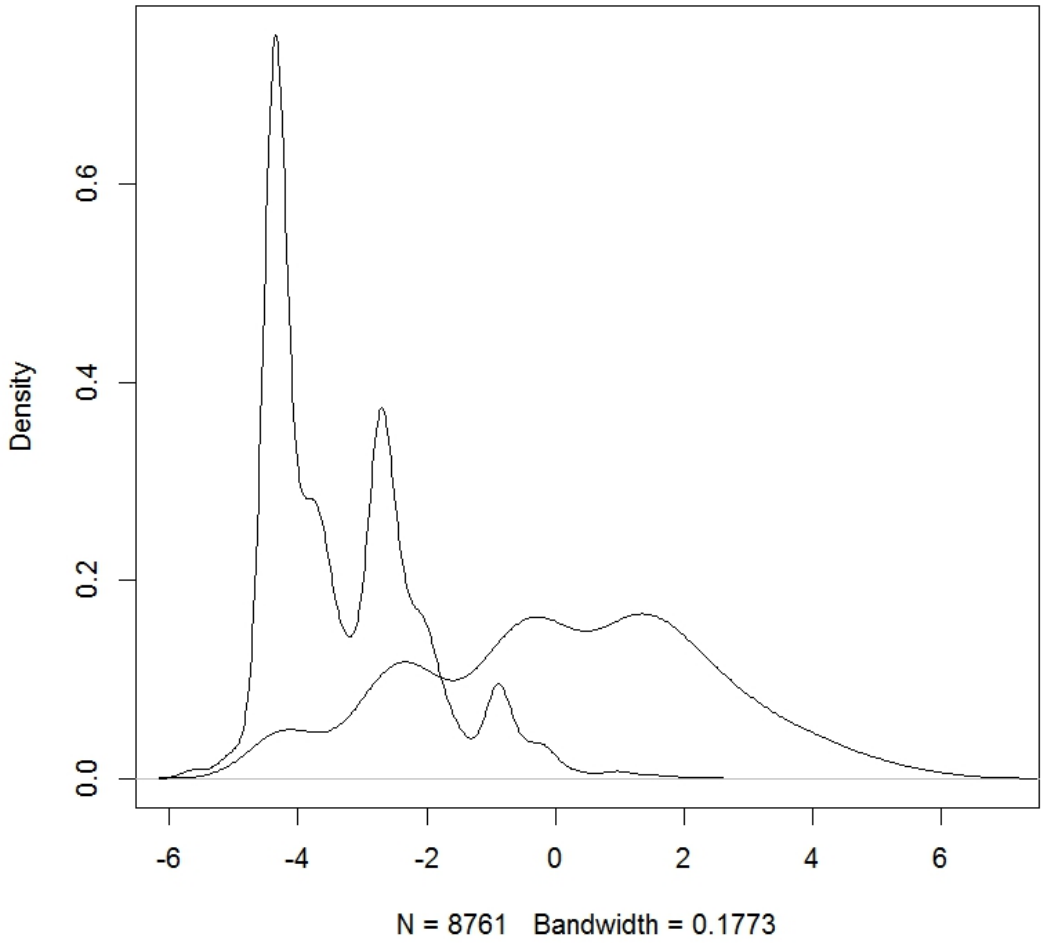
- On estime le modèle en enlevant une observation (ici un titre) des données).
- On prédit l'observation enlevée et on compare avec la vraie valeur.
- On répète cela pour toutes les observations.
- On répète pour toutes les valeurs de λ.
- On choisit la valeur de λ qui donne la meilleur capacité prédictive.
- Pas de p-valeur.

Résultats et coefficients

Table – Table des coefficients de régression de la régression logistique

Mots	coef.	Mots	coef.	Mots	coef.	Mots	coef.
intercept	-4.35						
materials	-1.27	american	0.02	letters	0.49	advanced	1.49
society	-0.70	acta	0.07	engineering	0.56	journal	1.71
systems	-0.70	the	0.19	current	0.58	sciences	1.76
ieee	-0.25	applications	0.20	computer	0.65	international	1.84
clinical	-0.24	european	0.34	and	0.72	research	1.93
surgery	-0.24	&	0.37	applied	0.79		
of	-0.10	de	0.39	advances	0.93		
on	-0.06	technology	0.40	science	1.04		
physics	-0.01	medical	0.44	review	1.22		
health	-0.003	in	0.45	management	1.26		

Résultats et coefficients



Densités des scores pour les deux listes.

Faux positifs, faux négatifs

Table – Table with the number of journals in both lists with respect to the predicted values.

	Web of Science	Beall's list	Total
Predicted Web of Science	8667 (93.4%)	616 (6.6%)	9283 (100%)
Predicted Beall's list	94 (12.2%)	677 (87.8%)	771 (100%)

Probabilités les plus élevées

Table – List of the 10 journals with the largest probability of belonging to Beall's list.

prob.	Journal names
0.99435	international journal of advanced research in engineering & management
0.99435	shiv rudraksha international journal of advanced research in engineering & management
0.99490	international journal of management and social science research review
0.99501	international journal of advanced research in engineering and science
0.99568	international journal of advanced research in computer science & technology
0.99666	international journal of advanced research in science engineering and technology
0.99697	international advanced research journal in science engineering and technology
0.99733	international journal of advanced research in applied science and technology
0.99740	international journal of advanced research in computer science and electronics engineering
0.99740	international journal of advanced research in computer science and software engineering
0.99753	international journal of research and development in technology & management sciences kailash

Faux positifs

Table – Faux positifs

Probabilités	Noms
0.8296	international journal of food sciences and nutrition
0.8296	international journal of nonlinear sciences and numerical simulation
0.8296	international journal of rock mechanics and mining sciences
0.85207	international journal of environmental research and public health
0.85248	international journal of peptide research and therapeutics
0.85248	journal of iron and steel research international
0.8642	journal of research in medical sciences
0.86427	international journal for vitamin and nutrition research
0.87556	space weather-the international journal of research and applications
0.9069	international journal of applied research in veterinary medicine
0.90869	international journal of applied mathematics and computer science

Le pire nom

Le pire nom

International Journal of Research and Review in Advanced Management Sciences and Advances in Applied Computer Engineering (IJRRAMSAACE)

Conclusions

Conclusions

- Les noms des revues prédatrices sont tellement caricaturaux qu'on peut en identifier une grande partie sur base de leur nom.
- La méthode peut servir d'outil de détection, mais nécessite une vérification manuelle.

Conclusions générales

Conclusions

- La science est face à des dilemmes et à des défis.
- Il existe une fake science comme il existe des fake news.
- Les scientifiques réfléchissent.
- Chevassus-au Louis, N. (2016). *Malscience. De la fraude dans les labos.*
  Le Seuil, Paris

Bibliography I

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, 489(7415) :179–179.

Beall, J. (2015). Criteria for determining predatory open-access publishers. Scholarly open access https://web.archive.org/web/20161130184313/https://scholarlyoa.files.wordpress.com/2015/01/criteria-2015.pdf,(accessed 2015-02-14).

Chevassus-au Louis, N. (2016). *Malscience. De la fraude dans les labos*. Le Seuil, Paris.

Mehrpour, S. and Khajavi, Y. (2014). How to spot fake open access journals. *Learned Publishing*, 27(4) :269–274.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Shen, C. and Björk, B.-C. (2015). 'predatory'open access : a longitudinal study of article volumes and market characteristics. *BMC medicine*, 13(1) :1.

Xia, J., Harmon, J. L., Connolly, K. G., Donnelly, R. M., Anderson, M. R., and Howard, H. A. (2015). Who publishes in "predatory" journals ? *Journal of the Association for Information Science and Technology*, 66(7) :1406–1417.