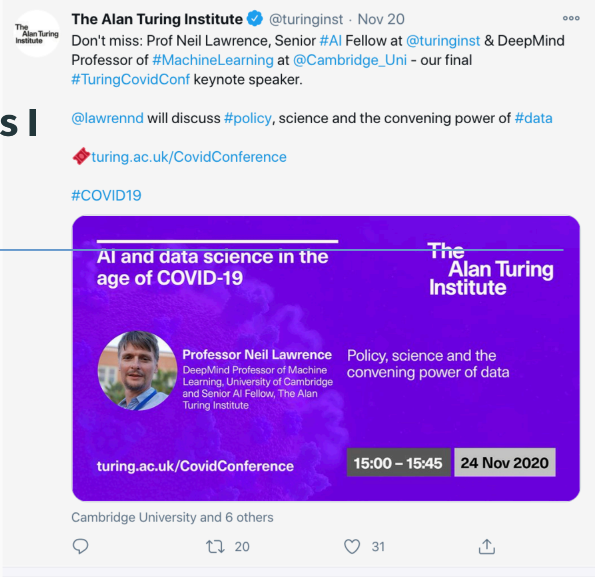
Methods of Applied Statistics I

STA2101H F LEC9101

Week 11

November 26 2020



The Alan Turing Institute @turinginst · Nov 20

Don't miss: Prof Neil Lawrence, Senior #AI Fellow at @turinginst & DeepMind Professor of #MachineLearning at @Cambridge_Uni - our final #TuringCovidConf keynote speaker.

@lawrennd will discuss #policy, science and the convening power of #data

🎟turing.ac.uk/CovidConference

#COVID19

AI and data science in the age of COVID-19

The Alan Turing Institute

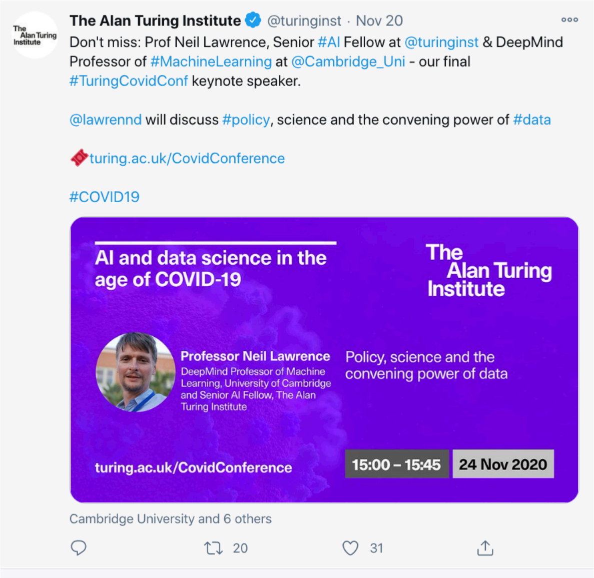
Professor Neil Lawrence
DeepMind Professor of Machine Learning, University of Cambridge and Senior AI Fellow, The Alan Turing Institute

Policy, science and the convening power of data

turing.ac.uk/CovidConference

15:00 – 15:45 | 24 Nov 2020

Cambridge University and 6 others



The Alan Turing Institute @turinginst · Nov 20

Don't miss: Prof Neil Lawrence, Senior #AI Fellow at @turinginst & DeepMind Professor of #MachineLearning at @Cambridge_Uni - our final #TuringCovidConf keynote speaker.

@lawrennd will discuss #policy, science and the convening power of #data

🎟️turing.ac.uk/CovidConference

#COVID19

AI and data science in the age of COVID-19

The Alan Turing Institute

Professor Neil Lawrence
DeepMind Professor of Machine Learning, University of Cambridge and Senior AI Fellow, The Alan Turing Institute

Policy, science and the convening power of data

turing.ac.uk/CovidConference

15:00 – 15:45 | 24 Nov 2020

Cambridge University and 6 others

20 31

“statisticians have a feel for the data that is crucial”

Today

Start Recording

1. HW3 due December 3
2. Nonparametric regression continued
3. Prediction and Explanation
4. Strategies for Modelling
5. In the News

- November 30 15.00 – 16.00 Tyler McCormick
- `https://canssiontario.utoronto.ca/?mec-events`
- “Identifying the latent space geometry of network models through analysis of curvature”

- Dec. 4 noon – Kathryn Roeder
- Dec. 7 – Margaret Roberts
- Dec. 14 – Kosuke Imai

Recap

- Generalized linear models theory
- flexible modelling via expected value and variance function
- inference using theory of likelihood functions

- nonparametric regression
- local averaging
- local polynomial regression

local least squares

- kernel function (for averaging)
- bandwidth (for local)

$$\mathbb{E}(y_i) \quad = \quad \mu_i; \qquad g(\mu_i) = x_i^T\beta; \qquad \text{Var}(y_i) = \phi_i V(\mu_i) \qquad \phi_i = a_i\phi$$

Recap 2: clarification re GLM fitting

$$\hat\beta = (X^T W^{-1} X) X^T W z$$

iteration

$$\begin{aligned} \hat\beta^{(t+1)} &= (X^T W^{(t)} X)^{-1} X^T W^{(t)} z^{(t)} \qquad z^{(t)} = X\hat\beta^{(t)} + W^{-1(t)} u^{(t)}; \\ W^{(t)} &= W(\hat\beta^{(t)}), \quad u^{(t)} = u(\hat\beta^{(t)}) \end{aligned}$$

At convergence,

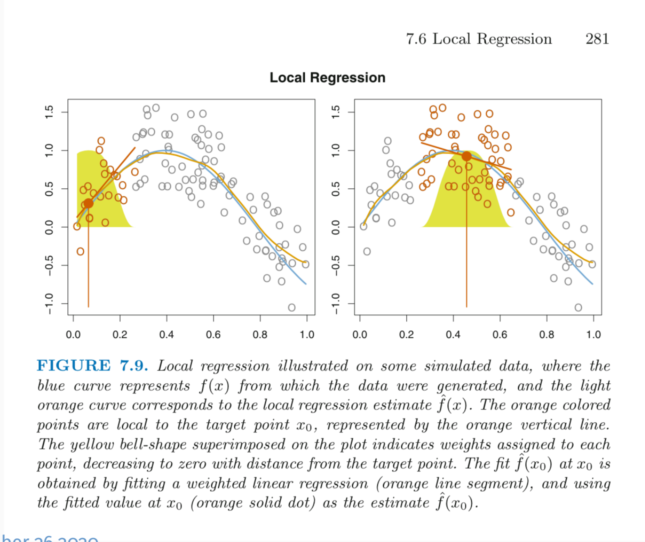
$$\begin{aligned} \hat\beta &= (X^T \hat W X)^{-1} X^T \hat z \\ \widehat{\text{Var}}(\hat\beta) &\doteq (X^T \hat W X)^{-1} \end{aligned}$$

W is diagonal

$$W_{ii} = \frac{1}{\phi a_i \{g'(\mu_i)\}^2 V(\mu_i)}, \qquad u_i = \frac{y_i - \mu_i}{\phi a_i g'(\mu_i) V(\mu_i)}$$

Local Polynomials

ISLR 7.9



7.6 Local Regression 281

FIGURE 7.9. *Local regression illustrated on some simulated data, where the blue curve represents $f(x)$ from which the data were generated, and the light orange curve corresponds to the local regression estimate $\hat{f}(x)$. The orange colored points are local to the target point x_0, represented by the orange vertical line. The yellow bell-shape superimposed on the plot indicates weights assigned to each point, decreasing to zero with distance from the target point. The fit $\hat{f}(x_0)$ at x_0 is obtained by fitting a weighted linear regression (orange line segment), and using the fitted value at x_0 (orange solid dot) as the estimate $\hat{f}(x_0)$.*

Local polynomials

SM §10.7

- $\hat\beta = (X^TWX)^{-1}X^TWy$ $\qquad W = \text{diag}(w_1, \dots, w_n)$

weights from kernel function Nov 19 Slide 9

- $\hat f_\lambda(x_0) = \hat\beta_0 = \sum_{i=1}^n S(x_0; x_i, \lambda)y_i$

fit a poly, use only the intercept

- $S(x_0; x_1, \lambda), \dots, S(x_0; x_n, \lambda)$ first row of ~~“hat”~~ matrix $(X^TWX)^{-1}X^TW$
- this makes it relatively easy to analyse the behaviour of local polynomial smoothers

SM §10.7

- and to simplify the expression for the cross-validation criterion $CV(\lambda)$
- fitting at each sample value gives

$$\hat f_\lambda(x_i) = \sum_{j=1}^n S(x_i; x_j, \lambda)y_j$$

smoothing matrix

Inference after fitting local polynomials

- model: $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n; \mathrm{E}(\epsilon_i) = 0; \mathrm{var}(\epsilon_i) = \sigma^2$
- $\hat{f}_\lambda(x_0) = \hat{\beta}_0 = \sum_{i=1}^n S(x_0; x_i, \lambda) y_i$
- $\mathrm{E}\{\hat{f}_\lambda(x_0)\} =$
- $\mathrm{var}\{\hat{f}_\lambda(x_0)\} =$
- how many parameters did we fit?
- by analogy with least squares, estimates of 'degrees of freedom' are $\nu_1 = \mathrm{tr}(S_\lambda)$, or $\nu_2 = \mathrm{tr}(S_\lambda^T S_\lambda)$

$$\tilde{\sigma}^2 = \frac{1}{n - 2\nu_1 + \nu_2} \sum \{y_i - \hat{f}_\lambda(x_i)\}^2$$

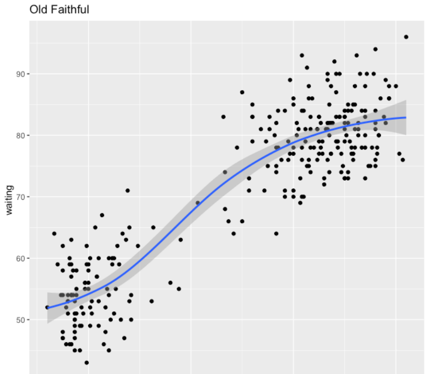
SM (10.39)

… inference after fitting local polynomials

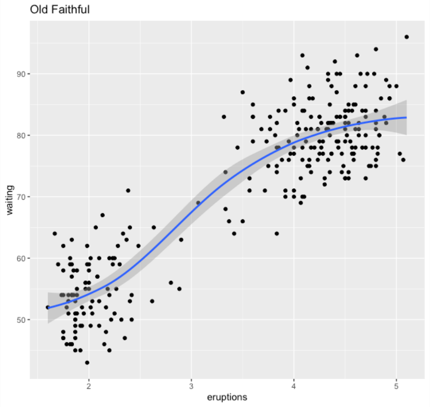
SM §10.7

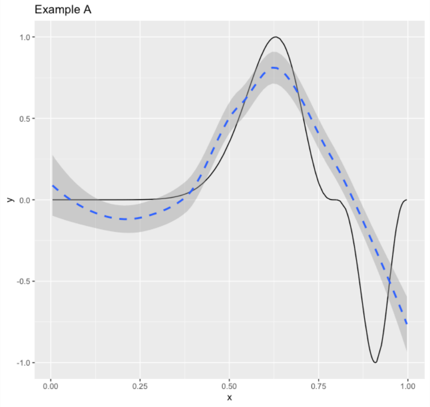
- $\mathrm{E}\{\hat{f}_\lambda(x_0)\} = \sum_{i=1}^n S(x_0; x_i, \lambda) f(x_i), \qquad \mathrm{var}\{\hat{f}_\lambda(x_0)\} = \sigma^2 \sum_{i=1}^n S^2(x_0; x_i, \lambda)$

$$\frac{\hat{f}_\lambda(x_0) - \mathrm{E}\{\hat{f}_\lambda(x_0)\}}{\widehat{\mathrm{var}}\{\hat{f}_\lambda(x_0)\}^{1/2}} \dot\sim N(0,1)$$



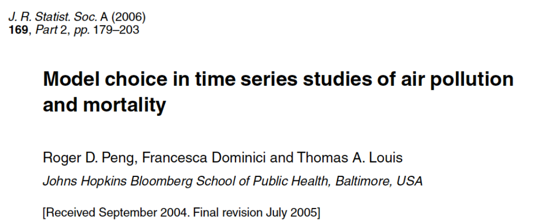
… inference after fitting local polynomials





Example: The NMMAPS studies

Peng, et al.(2006) JRSSA



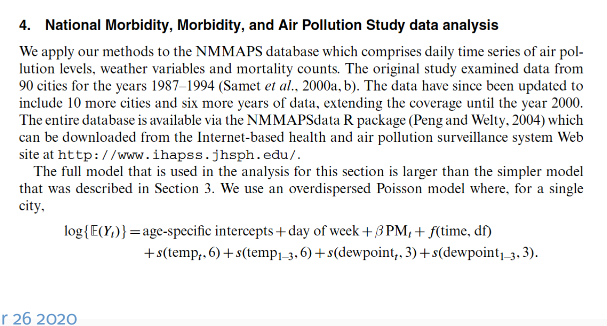
J. R. Statist. Soc. A (2006)
169, *Part* 2, *pp.* 179–203

Model choice in time series studies of air pollution and mortality

Roger D. Peng, Francesca Dominici and Thomas A. Louis

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

[Received September 2004. Final revision July 2005]



4. National Morbidity, Morbidity, and Air Pollution Study data analysis

We apply our methods to the NMMAPS database which comprises daily time series of air pollution levels, weather variables and mortality counts. The original study examined data from 90 cities for the years 1987–1994 (Samet *et al.*, 2000a, b). The data have since been updated to include 10 more cities and six more years of data, extending the coverage until the year 2000. The entire database is available via the NMMAPSdata R package (Peng and Welty, 2004) which can be downloaded from the Internet-based health and air pollution surveillance system Web site at `http://www.ihapss.jhsph.edu/`.

The full model that is used in the analysis for this section is larger than the simpler model that was described in Section 3. We use an overdispersed Poisson model where, for a single city,

$$\begin{aligned}\log\{\mathbb{E}(Y_t)\} = {} & \text{age-specific intercepts} + \text{day of week} + \beta\,\mathrm{PM}_t + f(\text{time, df})\\ & + s(\text{temp}_t, 6) + s(\text{temp}_{1\text{–}3}, 6) + s(\text{dewpoint}_t, 3) + s(\text{dewpoint}_{1\text{–}3}, 3).\end{aligned}$$

Example: The NMMAPS studies

Peng, et al.(2006) JRSSA

- 90 largest cities in US by population (US Census)
- daily mortality counts from National Center for Health Statistics 1987–1994
- hourly temperature and dewpoint data from National Climatic data Center
- data on pollutants PM_{10}, O_3, CO, SO_2, NO_2 from EPA

- response: Y_t number of deaths on day t
- explanatory variables: X_t pollution on day $t-1$, plus various confounders: age and size of population, weather, day of the week, time
- mortality rates change with season, weather, changes in health status, ...

NMMAPS: National Morbidity, Mortality and Air Pollution Study

… the NMMAPS studies

Peng, et al.(2006) JRSSA

- $Y_t \sim Poisson(\mu_t)$

- $\log(\mu_t) = \text{age specific intercepts } + \beta PM_t + \gamma DOW + g(t, df) + s(temp_t, 6) + s(temp_{t-1}, 6) + s(dewpoint_t, 3) + s(dewpoint_{t-1}, 3) + s_4(dew_0, 3) + s_5(dew_{1-3}, 3)$

- three ages categories; separate intercept for each ($< 65, 65 - 74, \geq 75$)
- dummy variables to record day of week
- $s(x, 7)$ a smoothing spline of variable x with 7 degrees of freedom
- estimate of β for each city; estimates pooled using Bayesian arguments for an overall estimate
- very difficult to separate out weather and pollution effects

see also: Crainiceanu, C., Dominici, F. and Parmigiani, G. (2008). *Biometrika* **95** 635–51

Nonparametric regression

Intro to Stat Learning; James et al.

- $y_i = f(x_i) + \epsilon_i$
- there are many different methods for estimating $f(\cdot)$
- local polynomial regression – `stats::loess`, `KernSmooth::locpoly` FELM 11.3; SM 10.7.1
- regression splines – `splines::bs` , `splines::ns` FELM 11.2b p 218ff
- smoothing splines – `stats:smooth.spline` FELM 11.2a; SM 10.7.2
- penalized splines – `pspline::smooth.Pspline` Peng et al. 2006
- wavelets – `wavethresh::wd` FELM 11.4
- and more... FELM 11.5; ISLR Ch.7

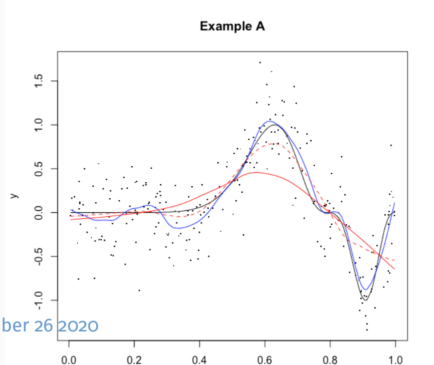
smoothing splines:

$$\min_f \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt$$

Elem Stat Learning, Hastie et al.

Example

```
with(exa, plot(x,y, cex = 2, main = "Example A", pch = "."))
lines(m ~ x, exa)
lines(exa.sm1$x, exa.sm1$y, col="blue")
lines(loess.smooth(exa$x, exa$y), col="red")
lines(loess.smooth(exa$x, exa$y, span = 1/3), col="red", lty = 2)
```



⟹ smooth.R

Extensions

FELM Ch.12

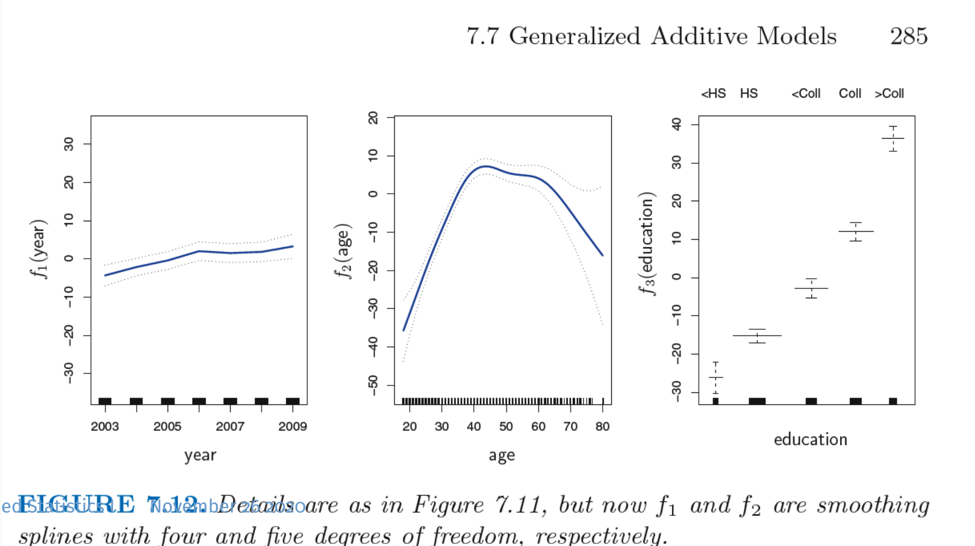
- more than 1 x
- local polynomials in two (or three) dimensions — difficult to view/fit
- thin-plate splines $d = 2, 3$
- additive models: $y_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi}) + \epsilon_i$ — FELM Ch.12

  `gam::gam` and `mgcv::gam`
- beyond least squares
- GAM: e.g. $\log\{p_i/(1 - p_i)\} = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi})$ — ISLR 7.7

- penalized generalized linear models
- e.g. linear predictor $\eta_i = x_i^T\beta + f(t_i)$ — SM 10.7.3

Wage data

smooth.html



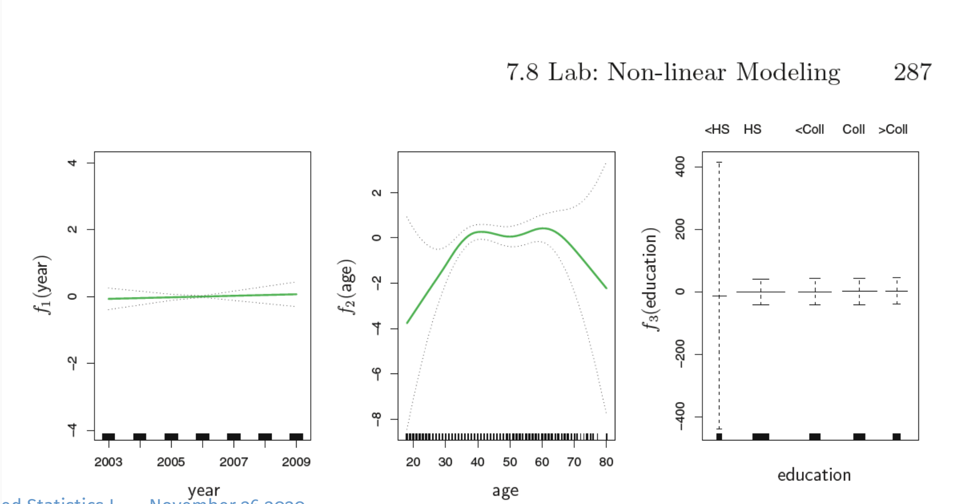
7.7 Generalized Additive Models 285

FIGURE 7.12. *Details are as in Figure 7.11, but now f_1 and f_2 are smoothing splines with four and five degrees of freedom, respectively.*

Wage data

ISLR, Ch.7

smooth.html



7.8 Lab: Non-linear Modeling 287

Aside: Explanation vs Prediction

- regression (and other) models may be fit in order to uncover some structural relationship between the response and one or more predictors
    - How do wages depend on education?
    - How does numeracy score affect probability of saying yes to vaccine?
- statistical analysis will focus on estimation and/or testing
- it is a remarkable fact that the data provides both an estimate of a model parameter and an estimate of uncertainty

- the focus might instead be on predicting responses for new values of x
- or classifying new observations on the basis of their x values
- the statistical analysis will focus on the accuracy and precision of the prediction/classification
- the data used to fit the model does not provide a good assessment of the prediction or classification error — motivates the division of data into training and test sets

How do we choose a model?

CD Ch. 6.5

- in many fields of study the models used as a basis for interpretation do not have a special subject-matter base
- rather represent broad patterns of haphazard variation quite widely seen
- this is typically combined with a specification of the systematic part of the variation
- which is often the primary focus
- modelling then often reduces to a choice of distributional form
- and of the independence structure of the random components

… empirical models

- functional form of the probability distribution sometimes critical, for example where an implicit assumption is involved of a relationship between variance and mean: geometric, Poisson, binomial
- the simple situations that give rise to binomial, Poisson, geometric, exponential, normal and log normal are some guide to empirical model choice in more complex situations

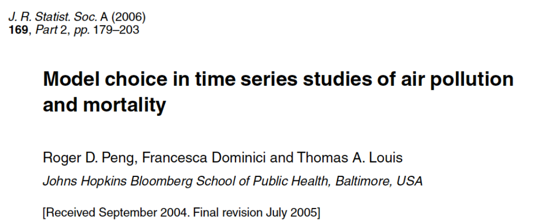
- In some specific contexts there is a tradition establishing the form of model
- illustration: financial time series – $Y(t) = \log\{P(t)/P(t-1)\}$ has a long-tailed distribution, small serial correlation, large serial correlation in $Y^2(t)$

stock price

- often have a long tail of large values; exponential distribution is a natural staring point
- extensions may be needed, including Weibull, gamma or log-normal

Example: The NMMAPS studies

Peng, et al.(2006) JRSSA



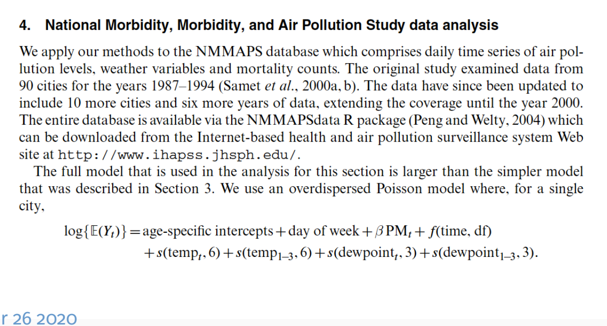
J. R. Statist. Soc. A (2006)
169, *Part* 2, *pp.* 179–203

Model choice in time series studies of air pollution and mortality

Roger D. Peng, Francesca Dominici and Thomas A. Louis

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

[Received September 2004. Final revision July 2005]



4. National Morbidity, Morbidity, and Air Pollution Study data analysis

We apply our methods to the NMMAPS database which comprises daily time series of air pollution levels, weather variables and mortality counts. The original study examined data from 90 cities for the years 1987–1994 (Samet *et al*., 2000a, b). The data have since been updated to include 10 more cities and six more years of data, extending the coverage until the year 2000. The entire database is available via the NMMAPSdata R package (Peng and Welty, 2004) which can be downloaded from the Internet-based health and air pollution surveillance system Web site at `http://www.ihapss.jhsph.edu/`.

The full model that is used in the analysis for this section is larger than the simpler model that was described in Section 3. We use an overdispersed Poisson model where, for a single city,

$$\begin{aligned}\log\{\mathbb{E}(Y_t)\} = {} & \text{age-specific intercepts} + \text{day of week} + \beta\,\mathrm{PM}_t + f(\text{time}, \text{df}) \\ & + s(\text{temp}_t, 6) + s(\text{temp}_{1\text{–}3}, 6) + s(\text{dewpoint}_t, 3) + s(\text{dewpoint}_{1\text{–}3}, 3).\end{aligned}$$

... empirical models

- often helpful to develop random and systematic parts of the model separately
- models should obey natural or known constraints, even if these lie outside the range of the data
- example $P(Y=1) = \alpha + \beta x \Longrightarrow \log \dfrac{P(Y=1)}{P(Y=0)} = \alpha' + \beta' x$
- however, β measures the change in probability per unit change in x

β' does not

- when relationship between y and several variables $x_1, \ldots x_p$ is of interest
  - unlikely that the system is wholly linear
  - impractical to study nonlinear systems of unknown form
  - therefore reasonable to begin with a linear model
  - and seek isolated nonlinearities

… empirical models

CD Ch. 6.5

- often helpful to develop random and systematic parts of the model separately
- naive approach: one random variable per study individual
- values for different individuals independent

- more realistic: possibility of structure in the random variation
- dependence in time or space, or a hierarchical structure corresponding to levels of aggregation
- ignoring these complications may give misleading assessments of precision, or bias the conclusions

... empirical models

CD Ch. 6.5

- example: standard error of mean $\sigma/\sqrt{n}$
- but, under mutual correlation, becomes $(\sigma/\sqrt{n})(1 + \Sigma\rho_{ij})^{1/2}$
- if each observation correlated with k others, at same level, $(\sigma/\sqrt{n})(1 + k\rho)^{1/2}$

```
0.1  0.2  0.4  0.8
--------------------------
1.14 1.26 1.48 1.84
1.18 1.34 1.61 2.05
1.22 1.41 1.73 2.24
1.26 1.48 1.84 2.41
1.30 1.55 1.95 2.57
1.34 1.61 2.05 2.72
```

$k = 3 : 8$

… empirical models

- important to be explicit about the unit of analysis
- has a bearing on independence assumptions involved in model formulation
- example: if all patients in the same clinic receive the same treatment
- then the clinic is the unit of analysis

- in some contexts there may be a clear hierarchy
- assessment of precision comes primarily from comparisons between units
- modelling of variation within units is necessary only if of intrinsic interest

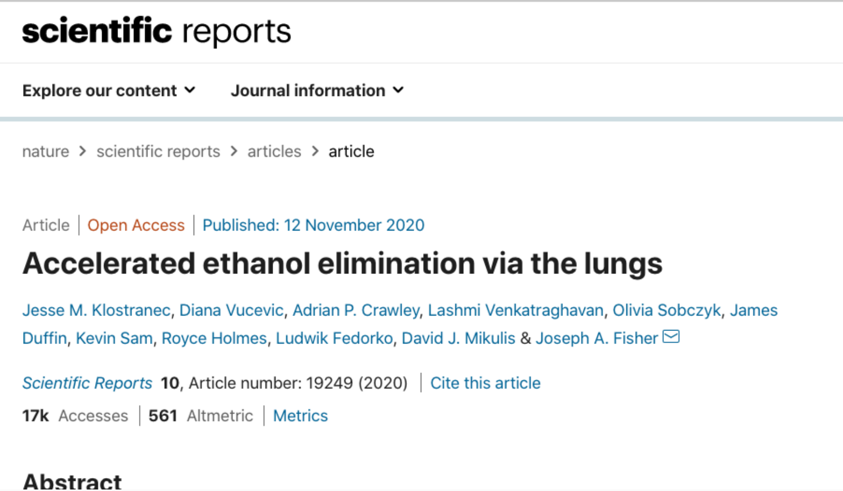
- when relatively complex responses are collected on each individual, the simplest way of condensing these is through a number of summary descriptive measures
- in other situations it may be necessary to represent explicitly the different hierarchies of variation

In the News

- “Toronto researchers have found a new way to speed up the body’s ability to rid itself of alcohol” Globe & Mail, Nov 12
- “Williams prof disavows own finding of mishandled GOP ballots” Twitter C. Bergstrom
- “The association between early career informal mentorship in academic collaborations and junior author performance” Gelman’s blog
- “What We Know About AstraZeneca’s Head-Scratching Vaccine Results” NY Times Nov 24
- “A very, very bad look for remdesivir” Science, Nov 6

Alcohol

Klostranec et al., Nature Comm.



Article | Open Access | Published: 12 November 2020

Accelerated ethanol elimination via the lungs

Jesse M. Klostranec, Diana Vucevic, Adrian P. Crawley, Lashmi Venkatraghavan, Olivia Sobczyk, James Duffin, Kevin Sam, Royce Holmes, Ludwik Fedorko, David J. Mikulis & Joseph A. Fisher ✉

Scientific Reports **10**, Article number: 19249 (2020) | Cite this article

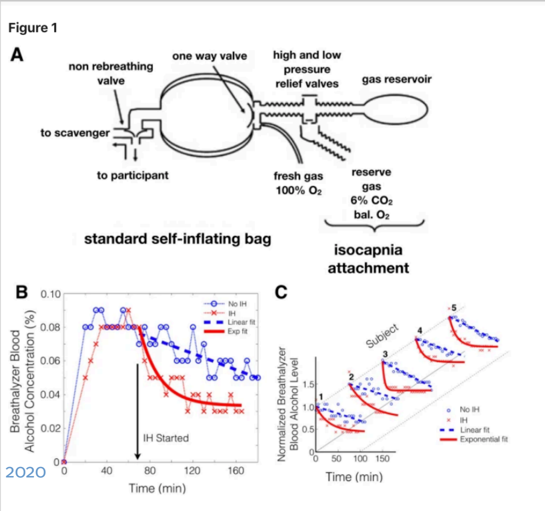
17k Accesses | **561** Altmetric | Metrics

Abstract

Alcohol

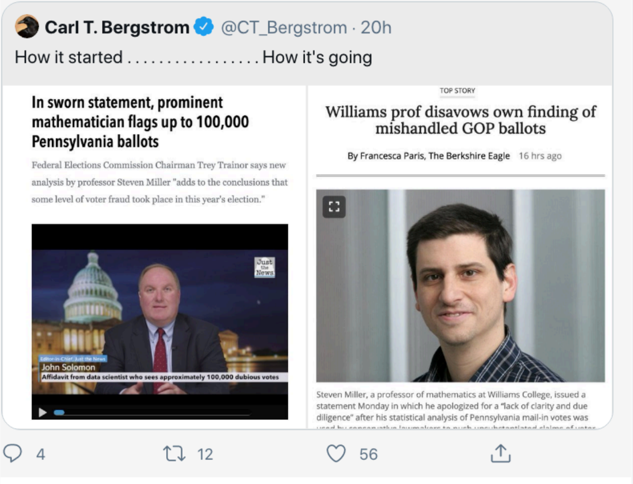
Klostranec et al., Nature Comm.

Link



Voting error

Link



Carl T. Bergstrom @CT_Bergstrom · 20h

How it started . . . . How it's going

In sworn statement, prominent mathematician flags up to 100,000 Pennsylvania ballots

Federal Elections Commission Chairman Trey Trainor says new analysis by professor Steven Miller "adds to the conclusions that some level of voter fraud took place in this year's election."

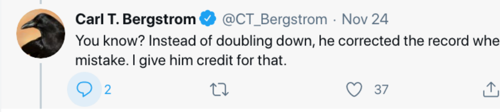
Just the News
Editor-in-Chief, Just the News
John Solomon
Affidavit from data scientist who sees approximately 100,000 dubious votes

TOP STORY

Williams prof disavows own finding of mishandled GOP ballots

By Francesca Paris, The Berkshire Eagle 16 hrs ago

Steven Miller, a professor of mathematics at Williams College, issued a statement Monday in which he apologized for a "lack of clarity and due diligence" after his statistical analysis of Pennsylvania mail-in votes was used by conservative lawmakers to push unsubstantiated claims of voter



Carl T. Bergstrom @CT_Bergstrom · Nov 24

You know? Instead of doubling down, he corrected the record whe mistake. I give him credit for that.

Voting error

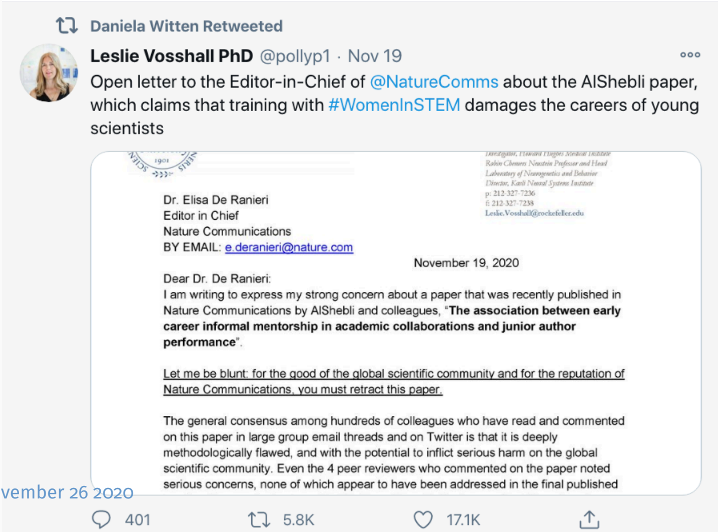
Berkshire Eagle

- “In a signed affidavit, Miller claimed to show that more than 89,000 ballots requested by Pennsylvania Republicans either were not counted by the state or requested by someone other than the registered Republican Link
- “He used data provided by former Donald Trump campaign staffer Matt Braynard.
- “Election officials have called the voting process secure
- “Miller told The Eagle that he made a mistake separating his analysis of the data from questions about the reliability of the data itself.
- “Braynard collected the data by contracting call centers to get in touch with Republican voters across six swing state
- “In his analysis, Miller wrote that the group called 20,000 Republican voters in Pennsylvania who, according to state records, had requested but not returned ballots. In all, 2,684 agreed to answer questions
- “Of the respondents, 463 reported that they actually had mailed in a ballot and 556 reported that they had not requested a ballot ... Miller extrapolated from those numbers
- ‘To apply nave statistical formulas to biased data and publish this is both irresponsible and unethical,’ De Veaux wrote in a statement

Mentoring

Nature Comm. Nov.17

Link



Daniela Witten Retweeted

Leslie Vosshall PhD @pollyp1 · Nov 19

Open letter to the Editor-in-Chief of @NatureComms about the AlShebli paper, which claims that training with #WomenInSTEM damages the careers of young scientists

Investigator, Howard Hughes Medical Institute
Robin Chemers Neustein Professor and Head
Laboratory of Neurogenetics and Behavior
Director, Kavli Neural Systems Institute
p: 212-327-7236
f: 212-327-7238
Leslie.Vosshall@rockefeller.edu

Dr. Elisa De Ranieri
Editor in Chief
Nature Communications
BY EMAIL: e.deranieri@nature.com

November 19, 2020

Dear Dr. De Ranieri:
I am writing to express my strong concern about a paper that was recently published in Nature Communications by AlShebli and colleagues, "**The association between early career informal mentorship in academic collaborations and junior author performance**".

Let me be blunt: for the good of the global scientific community and for the reputation of Nature Communications, you must retract this paper.

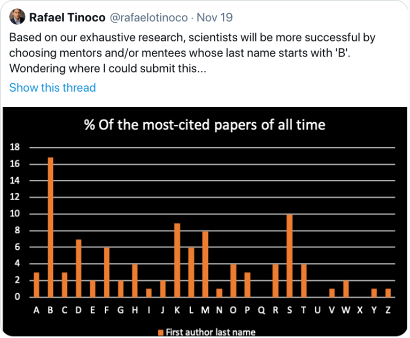
The general consensus among hundreds of colleagues who have read and commented on this paper in large group email threads and on Twitter is that it is deeply methodologically flawed, and with the potential to inflict serious harm on the global scientific community. Even the 4 peer reviewers who commented on the paper noted serious concerns, none of which appear to have been addressed in the final published

Mentoring

Nature Comm. Nov.17

Link to Gelman’s blog post

Thanks to Emma, here’s a link to a re-analysis of the data that points out several problems with it. Link



Rafael Tinoco @rafaelotinoco · Nov 19

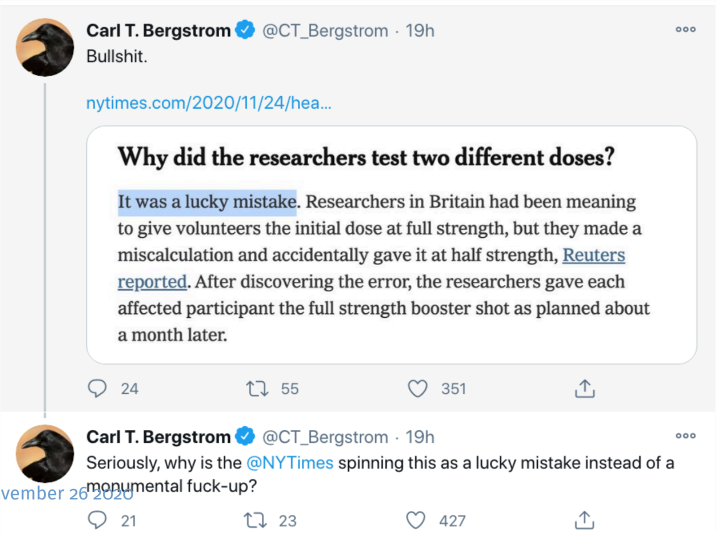
Based on our exhaustive research, scientists will be more successful by choosing mentors and/or mentees whose last name starts with 'B'. Wondering where I could submit this...

Show this thread

% Of the most-cited papers of all time
18
16
14
12
10
8
6
4
2
0
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
First author last name

Oxford vaccine

Link



Carl T. Bergstrom @CT_Bergstrom · 19h

Bullshit.

nytimes.com/2020/11/24/hea...

Why did the researchers test two different doses?

It was a lucky mistake. Researchers in Britain had been meaning to give volunteers the initial dose at full strength, but they made a miscalculation and accidentally gave it at half strength, Reuters reported. After discovering the error, the researchers gave each affected participant the full strength booster shot as planned about a month later.

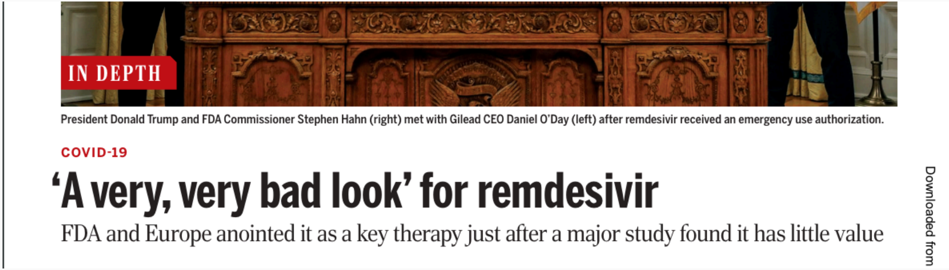
Carl T. Bergstrom @CT_Bergstrom · 19h

Seriously, why is the @NYTimes spinning this as a lucky mistake instead of a monumental fuck-up?

Remdesivir

Science, Nov.6

Link to Science article Link to BMJ article



IN DEPTH

President Donald Trump and FDA Commissioner Stephen Hahn (right) met with Gilead CEO Daniel O'Day (left) after remdesivir received an emergency use authorization.

COVID-19

'A very, very bad look' for remdesivir

FDA and Europe anointed it as a key therapy just after a major study found it has little value

Downloaded from