Methods of Applied Statistics I

STA2101H F LEC9101

Week 12

December 8 2021

Today

Start Zoom

1. Upcoming events
2. Recap — HW 9,10
3. nonparametric regression
4. regularized regression
5. course summary

Upcoming

- Friday Dec 10 Toronto Data Workshop Zoom link

Nathan Taback U of T

Dec 20
5:15

Data Science Series

CANSSI-Ont.

- Thursday Dec 9 Statistics Seminar 10.00 am Zoom link

Soufiane Hayou, National University of Singapore

Short Bio

Soufiane Hayou obtained his PhD in statistics in 2021 from Oxford wher
Paris before joining Oxford.

During his PhD, he worked mainly on the theory of randomly initialized
the weights, activation function) and the architecture (fully-connected,
He is currently a visiting assistant professor of mathematics at the National University of Si

Some insights from doubly infinite neural networks

Project

- Part I 3–5 pages, non-technical

  12 point type, 1.5 vertical spacing, thank you

  .pdf

  1. a description of the scientific problem of interest
  2. how (and why) the data being analyzed was collected
  3. preliminary description of the data (plots and tables)
  4. non-technical summary for a non-statistician of the analysis and conclusions

- Part II 3–5 pages, technical

  LaTeX or R markdown; submit .Rmd and .pdf files

  1. models and analysis
  2. summary for a statistician of the analysis and conclusions

- Part III Appendix

  submit .Rmd and .pdf or .html files

  R script or .Rmd file; additional plots; additional analysis; References

- 40 points total

- Part I:
  description of data and scientific problem 5
  suitability of plots and tables 5 clear, non-technical, concise but thorough
  quality of the presentation 5

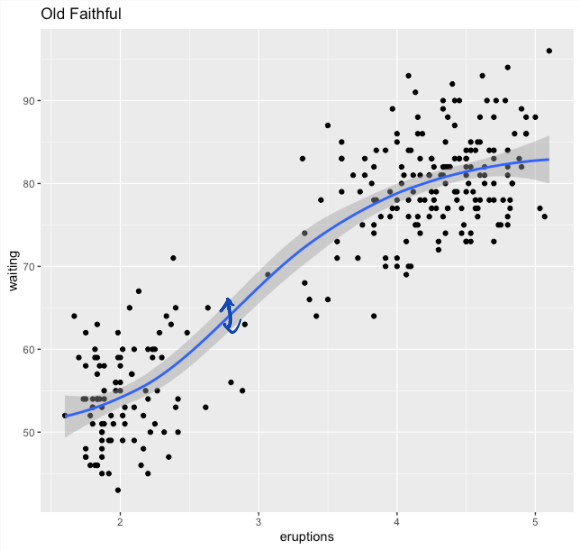
- Part II:
  summary of the modelling and methods 5 justification for choices
  suitability and thoroughness of the analysis 10 model checks, data checks

- Part III:
  relevance of additional material 5
  complete and reproducible submission 5

Recap: Nonparametric regression

- model $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \quad x_i$ scalar
- mean function $f(\cdot)$ assumed to be "smooth"
- local polynomial fit, using either $k = 0$ — Nadaraya-Watson or function estimate is the intercept $\hat{\beta}_0$
- $k = 1$ local linear regression
- $k = 3$ local cubic regression — odd k works better at the edges
- robustified version of local linear regression loess
- local polynomial fits easier to analyse $\hat{f}_\lambda(x_i) = \sum_{j=1}^n S(x_i, x_j; \lambda) y_j$ — weighted average
- $\mathrm{E}\{\hat{f}_\lambda(x_0)\} = \sum_{i=1}^n S(x_0; x_i, \lambda) f(x_i), \qquad \mathrm{var}\{\hat{f}_\lambda(x_0)\} = \sigma^2 \sum_{i=1}^n S^2(x_0; x_i, \lambda)$
- $\tilde{\sigma}^2 = \frac{1}{n - 2\nu_1 + \nu_2} \sum \{y_i - \hat{f}_\lambda(x_i)\}^2; \qquad \nu_1 = \mathrm{tr}(S_\lambda)$, or $\nu_2 = \mathrm{tr}(S_\lambda^T S_\lambda)$

Recap: Pointwise confidence intervals




```
ggplot(faithful) +
geom_point(aes(eruptions,waiting)) +
ggtitle("Old Faithful") +
geom_smooth(aes(eruptions,waiting),
se=T)
```

local k=1 linear reg.
don't use LS ← modified LS
downweights far

```
help(geom_smooth)
```

buyer beware

Recap: Kernel smoothers

- choose a bandwidth, λ to control smoothness of function
- larger bandwidth = more smoothing = increased bias, decreased variance

- choose a kernel function, $K(\cdot)$, controls smoothness and "local-ness"
- Faraway recommends Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2), |x| \leq 1$

- `ksmooth(base)` offers only uniform (box) or normal
- `bkde(KernSmooth)` offers `normal, box, epanech, biweight, triweight`
- `biweight`: $K(x) \propto (1 - |x|^2)^2, |x| \leq 1$ `triweight`: $K(x) \propto (1 - |x|^2)^3, |x| \leq 1$

Regression splines

- model $y_i = f(x_i) + \epsilon_i$ $\quad$ $f(\cdot)$ "flexible"
- above $f(\cdot)$ is estimated at several points using local constants or local linear regression

KernSmooth::locpoly

- another popular approach is to use some very flexible, but parametric form, for f
- for example, $f(x) = \sum_{m=1}^{M} \beta_m \phi_m(x)$

IJALM

Regression splines

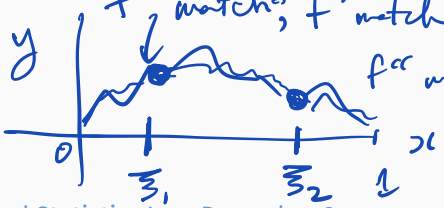
- model $y_i = f(x_i) + \epsilon_i$ $\quad$ $f(\cdot)$ "flexible"
- above $f(\cdot)$ is estimated at several points using local constants or local linear regression

KernSmooth::locpoly

- another popular approach is to use some very flexible, but parametric form, for f
- for example, $f(x) = \sum_{m=1}^{M} \beta_m \phi_m(x)$
- examples of ϕ_m: $1, x, x^2, x^3$; $\quad$ $1, \sin(x), \cos(x), \sin(2x), \cos(2x)$;
- piecewise polynomials: e.g. knots at $\xi_1, \xi_2 \in [0, 1]$

  basis functions $\phi(x) : 1, x, x^2, x^3, (x-\xi_1)^3_+, (x-\xi_2)^3_+$

IJALM



f, matches; f' matches
f'' matches

cubic spline, with knots ξ_1, ξ_2

columns

ns(x, df) natural cubic spl.

bs(x, df) b-splines (*)

Regression splines

- model $y_i = f(x_i) + \epsilon_i$ $\quad$ $f(\cdot)$ "flexible"
- above $f(\cdot)$ is estimated at several points using local constants or local linear regression

KernSmooth::locpoly

- another popular approach is to use some very flexible, but parametric form, for f
- for example, $f(x) = \sum_{m=1}^{M} \beta_m \phi_m(x)$

IJALM

- examples of ϕ_m: $1, x, x^2, x^3$; $\quad$ $1, \sin(x), \cos(x), \sin(2x), \cos(2x)$;
- piecewise polynomials: e.g. knots at $\xi_1, \xi_2 \in [0, 1]$
  basis functions $\phi(x)$: $1, x, x^2, = x^3, (x - \xi_1)^3_+, (x - \xi_2)^3_+$

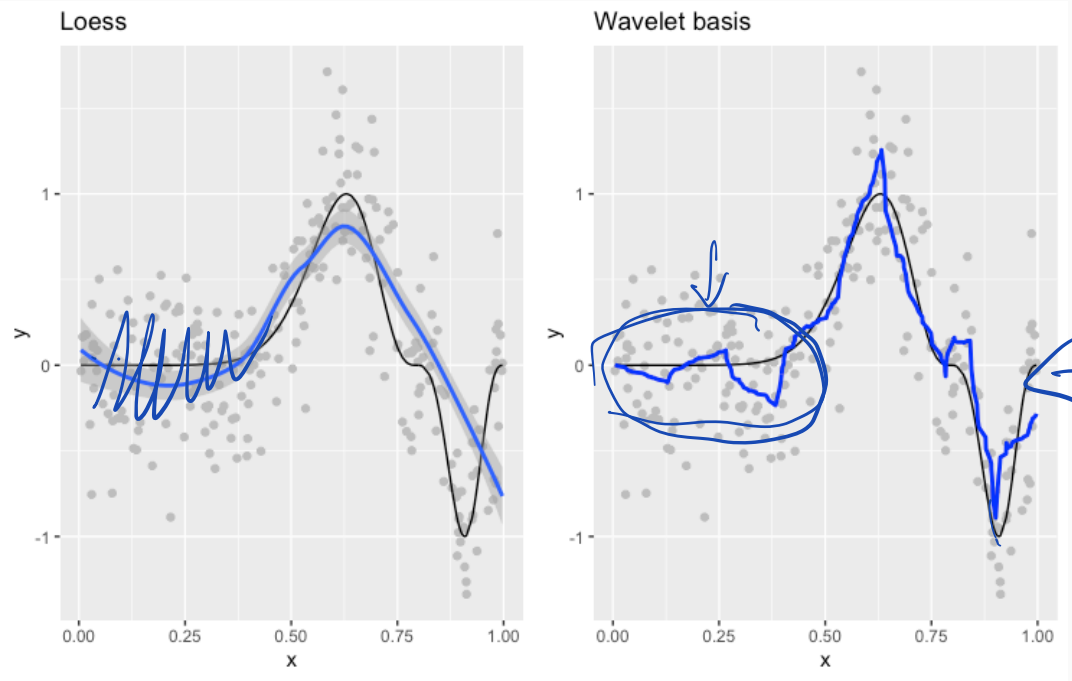
- ELM p.219 builds these "by hand"
- `splines::bs()` builds cubic splines automatically

Aside: wavelets

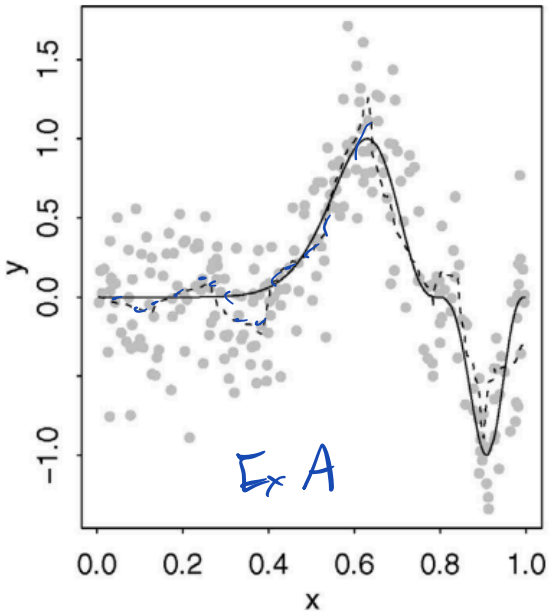
ELM §11.4

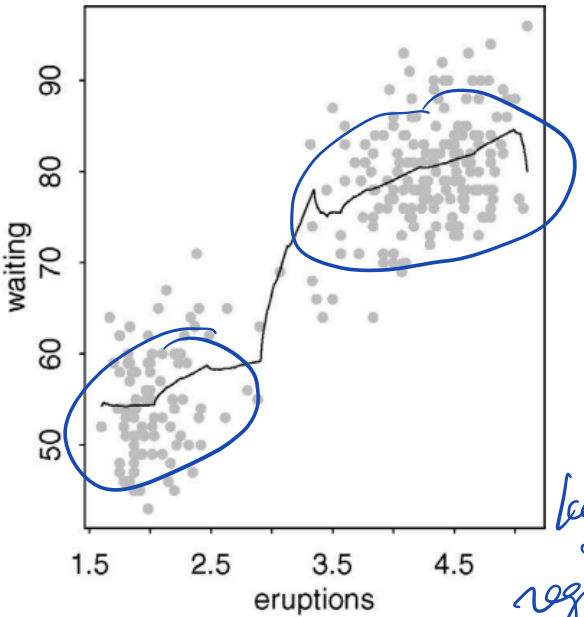
- $f(x) = \sum_{m=1}^{M} \beta_m \phi_m(x)$ — regression spline with basis functions ϕ
- wavelet basis functions are orthogonal — makes fitting easier
- also multi-resolution – able to track local wiggles better
- very useful for image processing, signal processing — can find edges and short bursts
- `wavethresh` package in `R`

Example A



Old Faithful



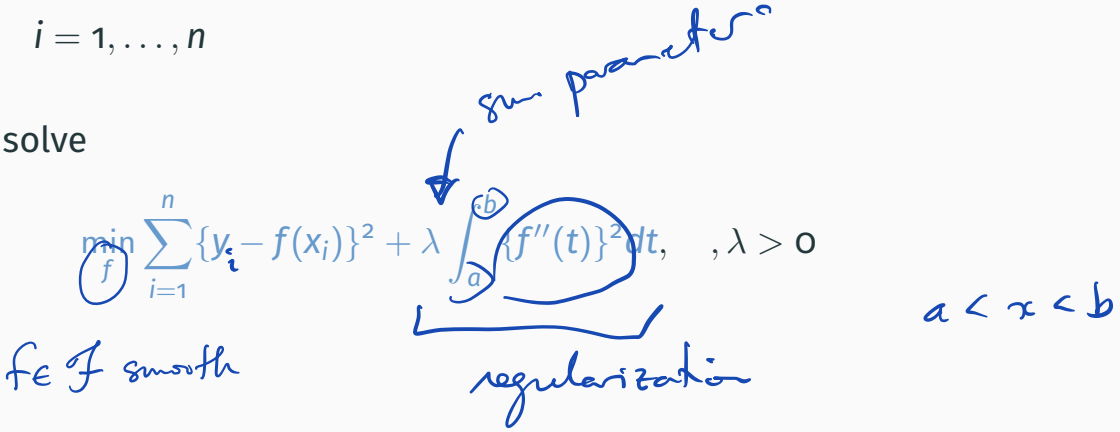


kernel sm.
reg Spln
↑IJALM

Figure 14.13 *Daubechies wavelet N=2 thresholded fit to the Example A data shown on the left. Irregular wavelet fit to the Old Faithful data is shown on the right.*

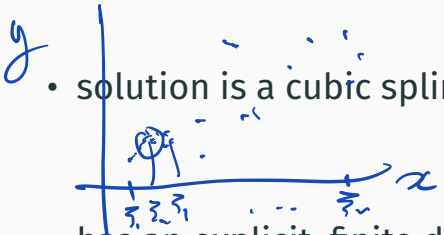
- $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$
- choose $f(\cdot)$ to solve

$$\min_f \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 dt, \quad , \lambda > 0$$



- $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$
- choose $f(\cdot)$ to solve

$$\min_f \sum_{i=1}^{n} \{y - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 dt, \quad , \lambda > 0$$



- solution is a cubic spline, with knots at each observed x_i value

see SM Figure 10.18 for a non-regularized solution

- has an explicit, finite dimensional solution

- $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$
- choose $f(\cdot)$ to solve

$$\min_f \sum_{i=1}^{n} \{y - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 dt, \quad , \lambda > 0$$

- solution is a cubic spline, with knots at each observed x_i value

see SM Figure 10.18 for a non-regularized solution

- has an explicit, finite dimensional solution
- $\hat{f} = \{\hat{f}(x_1), \ldots, \hat{f}(x_n)\} = (I + \lambda K)^{-1} y$.
- K is a symmetric $n \times n$ matrix of rank $n - 2$

- $y_i = f(x_i) + \epsilon_i$

$y_i = x_i^T\beta + \varepsilon_i \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- local polynomial regression – `stats::loess, KernSmooth::locpoly` ELM 11.3; SM 10.7.1
- regression splines – `splines::bs , splines::ns` ELM 11.2b p 218ff
- smoothing splines – `stats:smooth.spline` ELM 11.2a; SM 10.7.2

- penalized splines – `pspline::smooth.Pspline` Peng et al. 2006
- wavelets – `wavethresh::wd` ELM 11.4
- and more... ELM 11.5; ISLR Ch.7

- same ideas can be applied to generalized linear models
- replace linear predictor $\eta_i = \beta_0 + \beta_1 x_i$ with $f(x_i)$
- use local poly, reg splines, etc. SM Ex. 10.32 logistic regression

Example: logistic regression

SM Ex.10.29 and 10.32

City	Rain	r/m	City	Rain	r/m	City	Rain	r/m	City	Rain	r/m
1	1735	2/4	11	2050	7/24	21	1756	2/12	31	1780	8/13
2	1936	3/10	12	1830	0/1	22	1650	0/1	32	1900	3/10
3	2000	1/5	13	1650	15/30	23	2250	8/11	33	1976	1/6
4	1973	3/10	14	2200	4/22	24	1796	41/77	34	2292	23/37
5	1750	2/2	15	2000	0/1	25	1890	24/51			
6	1800	3/5	16	1770	6/11	26	1871	7/16			
7	1750	2/8	17	1920	0/1	27	2063	46/82			
8	2077	7/19	18	1770	33/54	28	2100	9/13			
9	1920	3/6	19	2240	4/9	29	1918	23/43			
10	1800	8/10	20	1620	5/18	30	1834	53/75			

Table 10.19 Toxoplamosis data: rainfall (mm) and the numbers of people testing positive for toxoplasmosis, r, our of m people tested, for 34 cities in El Salvador (Efron, 1986).

$y = \frac{r}{m} \sim$ family = binom

Terms	df	Deviance
Constant	33	74.21
Linear	32	74.09
Quadratic	31	74.09
Cubic	30	62.63

Table 10.20 Analysis of deviance for polynomial logistic models fitted to the toxoplasmosis data.

cbind $(r, m-r) \sim$ rain

Example: logistic regression

SM Ex.10.29 and 10.32

$\longrightarrow$ toxoplasmosis.Rmd

Example: logistic regression

SM Ex.10.29 and 10.32

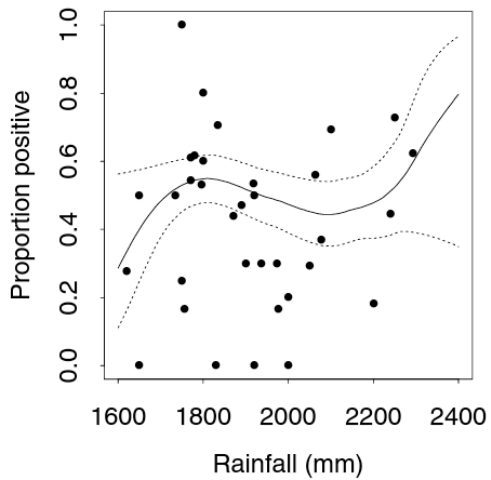
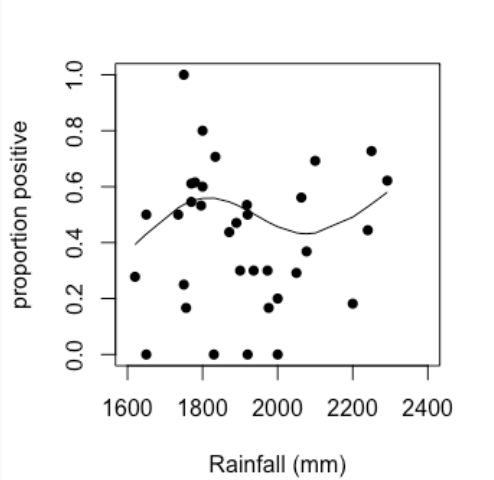


Figure 10.17 Local fits to the toxoplasmosis data. The left panel shows fitted probabilities $\widehat{\pi}(x)$, with the fit of local linear logistic model with $h = 400$ (solid) and 0.95 pointwise confidence bands (dots). Also shown is the local linear fit with $h = 300$ (dashes). The right panel shows the local quadratic fit with $h = 400$ and its 0.95 confidence band. Note the increased variability due to the quadratic fit, and its stronger curvature at the boundaries.



- more than 1 x
- local polynomials in two (or three) dimensions — difficult to view/fit
- thin-plate splines $d = 2, 3$
- additive models: $y_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi}) + \epsilon_i$ — ELM 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
- smooth generalized linear models
- e.g. linear predictor $\eta_i = x_i^T \beta + f(t_i)$ — SM 10.7.3

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,

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

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~~ 2000
- hourly temperature and dewpoint data from National Climatic data Center
- data on pollutants PM_{10}, O_3, CO, SO_2, NO_2 from EPA

1987 - 2000

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

- $Y_t \sim Poisson(\mu_t)$ w overdispersion

generalized additive model `gam`

- $\log(\mu_t) = \text{age specific intercepts} + \beta PM_t + \gamma DOW + s(t, 7) + s(temp_t, 6) + s(temp_{t-1}, 6) + s(dewpoint_t, 3) + s(dewpoint_{t-1}, 3) + s_4(dew_0, 3) + s_5(dew_{1-3}, 3)$

- three ages categories; separate intercept for each (< 65, $65 - 74$, ≥ 75)
- dummy variables to record day of week

smooth function of

$y = f(x_i) + \varepsilon_i$

$x_i \in \mathbb{R}$

... the NMMAPS studies

Peng, et al.(2006) JRSSA

- $Y_t \sim Poisson(\mu_t)$

dlnm ← Chicago NMMAPS

generalized additive model `gam`

- $\log(\mu_t) =$ age specific intercepts $+ \beta PM_t + \gamma DOW + s(t, 7) + s(temp_t, 6) + s(temp_{t-1}, 6) + s(dewpoint_t, 3) + s(dewpoint_{t-1}, 3) + s_4(dew_0, 3) + s_5(dew_{1-3}, 3)$
- three ages categories; separate intercept for each (< 65, $65 - 74$, ≥ 75)
- dummy variables to record day of week
- $s(t, 7)$ a smoothing spline of variable t 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

initial published 1.0070 ← est.

gam ← glm + nonp

RR ≈ ~~1.035~~ .2% 1.0035

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

Regularized regression

LM 11.3, 11.4

- parametric fit, but using some regularization, as in smoothing splines
- useful with high-dimensional data, i.e. large p — many explanatory variables
- two popular versions for least squares:

Regularized regression

- parametric fit, but using some regularization, as in smoothing splines
- useful with high-dimensional data, i.e. large p many explanatory variables
- two popular versions for least squares:
- ridge regression: $\min_\beta \{\sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2\}$ x's scaled to have mean 0 and sd 1
- resulting estimates are biased, but might have smaller MSE
- explicit solution available: $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$

Regularized regression

- parametric fit, but using some regularization, as in smoothing splines
- useful with high-dimensional data, i.e. large p — many explanatory variables
- two popular versions for least squares:

- ridge regression: $\min_\beta\{\sum_{i=1}^n(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^p \beta_j^2\}$ — x's scaled to have mean 0 and sd 1
- resulting estimates are biased, but might have smaller MSE
- explicit solution available: $\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^Ty$

- lasso regression $\min_\beta\{\sum_{i=1}^n(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^p |\beta_j|\}$ — x's scaled to have mean 0 and sd 1
- resulting estimates are biased, but are also sparse — several $\hat{\beta}_j$'s set to 0
- thus serves as a model selection method as well as an estimation method
- no explicit solution available, and little theory about the distribution of $\hat{\beta}_{Lasso}$

glmnet ← summary coef

- both regularization methods require a choice of λ
- works like the smoothing parameters in nonparametric regression – trades off variance and bias ← in prediction on a test set
- usually chosen by some version of cross-validation
- `library(glmnet)`

- both regularization methods require a choice of λ
- works like the smoothing parameters in nonparametric regression – trades off variance and bias
- usually chosen by some version of cross-validation
- `library(glmnet)`

$$\min_\beta \sum_i (y_i - x_i^T\beta)^2 + \lambda\Sigma|\beta_j|$$

- regularized regression can be generalized
- e.g. $\max_\beta\{\ell(\beta) - \lambda\sum_{j=1}^n |\beta_j|\}$; equivalently $\min_\beta\{-\ell(\beta) + \lambda\sum_{j=1}^n |\beta_j|\}$

$\log f(y|x;\beta)$

- theoretical properties of estimators poorly understood
- difficult to get, e.g. estimated standard errors for $\hat{\beta}_{Lasso}$

When to use “non-parametric” fits?

ELM, 11.6

- depends on the problem
- some fields of science have their own conventions e.g. mortality and air pollution, NMMAPS
- may be useful for confounding variables
- may be useful for exploratory analyses
- Faraway suggests using smoothing methods when there is “not too much” noise in the data
- suggests using parametric models when there are larger amounts of noise in the data

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 socio-economic status affect probability of severe covid?
- statistical analysis will focus on estimation and/or testing
- the data provides both an estimate of a model parameter and an estimate of uncertainty

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 socio-economic status affect probability of severe covid?
- statistical analysis will focus on estimation and/or testing
- 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

Summary Fall 2021

web page: This course will focus on principles and methods of applied statistical science. It is designed for MSc and PhD students in Statistics, and is required for the Applied Paper of the PhD comprehensive exams. The topics covered include: planning of studies, review of linear models, analysis of random and mixed effects models, model building and model selection, theory and methods for generalized linear models, and an introduction to nonparametric regression. Additional topics will be introduced as needed in the context of case studies in data analysis.

Topics

- linear regression: interpretation of coefficients, estimation, Wald test/t-test, comparing models, likelihood ratio test/F-test, model checking, residual and diagnostic plots, collinearity, prediction, model selection, shrinkage

Topics

- linear regression: interpretation of coefficients, estimation, Wald test/t-test, comparing models, likelihood ratio test/F-test, model checking, residual and diagnostic plots, collinearity, prediction, model selection, shrinkage

- designed experiments: factors, anova, blocking, randomized blocks, components of variance, randomization, causality

Topics

- linear regression: interpretation of coefficients, estimation, Wald test/t-test, comparing models, likelihood ratio test/F-test, model checking, residual and diagnostic plots, collinearity, prediction, model selection, shrinkage

- designed experiments: factors, anova, blocking, randomized blocks, components of variance, randomization, causality

- observational studies: retrospective/prospective, case-control

Topics

- linear regression: interpretation of coefficients, estimation, Wald test/t-test, comparing models, likelihood ratio test/F-test, model checking, residual and diagnostic plots, collinearity, prediction, model selection, shrinkage
- designed experiments: factors, anova, blocking, randomized blocks, components of variance, randomization, causality
- observational studies: retrospective/prospective, case-control
- logistic regression: binary and binomial response, logit transform, linear predictor, likelihood inference, Wald test, likelihood ratio test, residual deviance as model check, analysis of deviance, overdispersion, prediction, diagnostics and residuals

Topics cont'd

- principles: statistical science/data science "workflow", types of studies, design of studies, explanation and prediction, measures of risk, model choice, model selection

Topics cont'd

- principles: statistical science/data science "workflow", types of studies, design of studies, explanation and prediction, measures of risk, model choice, model selection

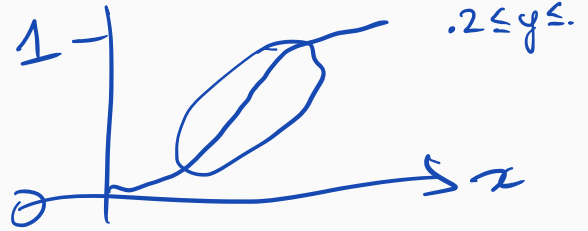
- generalized linear models: density, link function, dispersion parameter, normal/gamma/inverse Gaussian, binomial/Poisson/negative binomial, quasi-likelihood, over-dispersion, residuals, estimation, iteratively re-weighted LS

Topics cont'd

- principles: statistical science/data science "workflow", types of studies, design of studies, explanation and prediction, measures of risk, model choice, model selection
- generalized linear models: density, link function, dispersion parameter, normal/gamma/inverse Gaussian, binomial/Poisson/negative binomial, quasi-likelihood, over-dispersion, residuals, estimation, iteratively re-weighted LS
- non-parametric regression: kernel smoothers, local polynomial regression, regression splines, smoothing splines, cross-validation, inference
- regularization: lasso and ridge penalties on coefficients

In the News

- excess deaths, covid Sep 15
- hydroxychloroquine and ivermectin treatments Sep 22
- risk matrices HW2
- fragile states and paternalism Sep 29
- computational advertising Oct 6
- religiosity and economic and mental well-being Oct 13
- excess deaths again Oct 20
- math education and brain development HW6
- Ivermectin again; replicability; long-covid Oct 29
- statistics communication Nov 17
- replicability, vaccination/hospitalization Nov 24



n —— n_1 ← lasso etc. --> x1 x4 x7

n_2

$y \sim$ lm( )

response score $\in [0,1]$ ← quality of apt.

total $\sim \sum m_i$ (# of items)

$y \le M$

multiple comparisons

$\bar{y}_1, \ldots, \bar{y}_{11}$ s.e. $(\bar{y}_j - \bar{y}_{j'})$

$x \quad x^2$

$x_1 \quad x_2 \quad x_1 x_2$

$\bar{y} \uparrow$ until $9 \longrightarrow$

0 ↑ 1 ↑

$M_1 \le y_i \le M_2$

$X\beta \in \mathbb{R}^p$

$-\infty \le x_i^T \beta < \infty$