Methods of Applied Statistics I

STA2101H F LEC9101

Week 12

December 7 2022

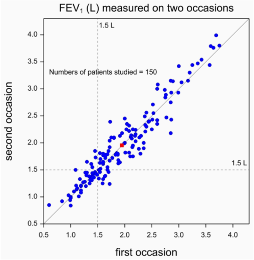


Figure 1. Data from a cross-over trial in asthma

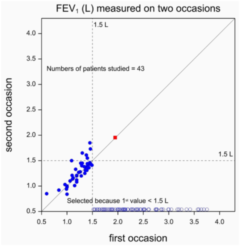


Figure 2. Data from a cross-over trial. Patients with poor values on the first occasion are measured on a second occasion.

Regression to the mean

link to Senn 2022

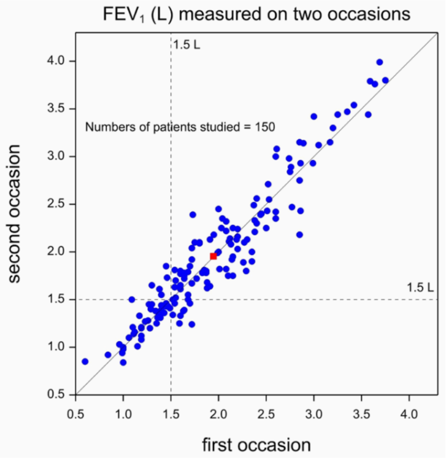


Figure 1. Data from a cross-over trial in asthma

Regression to the mean

link to Senn 2022

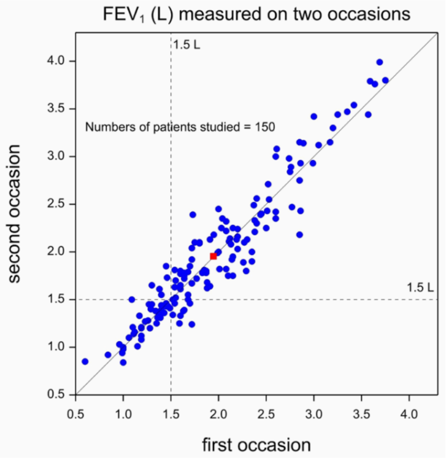


Figure 1. Data from a cross-over trial in asthma

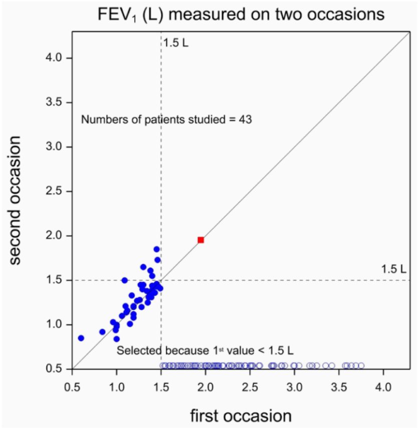


Figure 2. Data from a cross-over trial. Patients with poor values on the first occasion are measured on a second occasion.

"all 43 patients had FEV $<$ 1.5 in first period (extremely poor)"

6 patients have FEV $>$ 1.5 in second period improvement?

Today

1. Nonparametric regression
2. Bits and pieces
3. Project
4. Course evals Dec 13

Project due December 19 (11.59),
no extensions
So think of it as due on December 16 :)

Preliminary versions accepted
for feedback up to Dec 11

Project Guidelines STA 2101F: Methods of Applied Statistics I 2022

Outline

- Part I 3–5 pages, non-technical — 12 point type, 1.5 vertical spacing, thank you
  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

Project Marking

- 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 1 Non-parametric regression: smoothers

- model $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$ — $x_i \in \mathbb{R}$
- local polynomial fitting using kernel function and bandwidth — both to be specified
- local average, local linear, both very popular
- relatively easy to analyse properties of resulting estimator — linear regression
- robust regression can be used instead `loess` — `geom_smooth`
- smoothing parameter established using cross-validation
- degrees of freedom (for estimating σ^2) calculated from so-called smoothing matrix

  $\hat{f}_\lambda(x_0) = \Sigma S(x_0; x_i, \lambda) y_i$
- implemented in: `base::ksmooth`; `KernSmooth::locpoly`; `sm::sm.regression`
- `sm::sm.regression` computes bandwidth using cross-validation

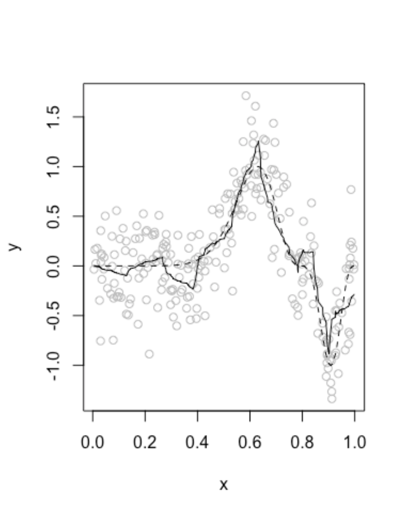
Recap 2 Nonparametric regression: regression splines

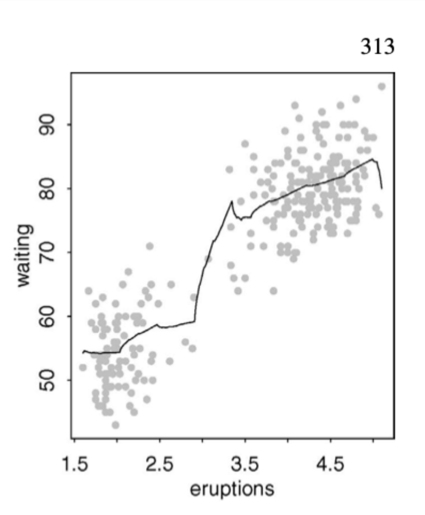
- model $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$ $x_i \in \mathbb{R}$
- allow $f(\cdot)$ to be "flexible" by expressing $f(x) = \sum_{m=1}^{M} \beta_m \phi_m(x)$ $\phi(\cdot)$ known
- fitting by least squares
- need to choose family $\{\phi_1(\cdot), \ldots, \phi_m(\cdot)\}$, and number of functions M
- B-spline basis and natural cubic spline basis are popular choices
  compromise between smoothness and flexibility
- splines are cubic polynomials on sub-intervals of x-space; smoothly joined at the endpoints of the intervals knots
- more knots means wigglier fits; fewer knots mean smoother fits knots $\leftrightarrow$ deg. freedom
- natural cubic splines have slightly better endpoint behaviour they are linear there
- `splines::bs` and `splines::ns` create the basis `ns(x,3)` for example
- other basis families include Fourier, wavelet

Aside: wavelets

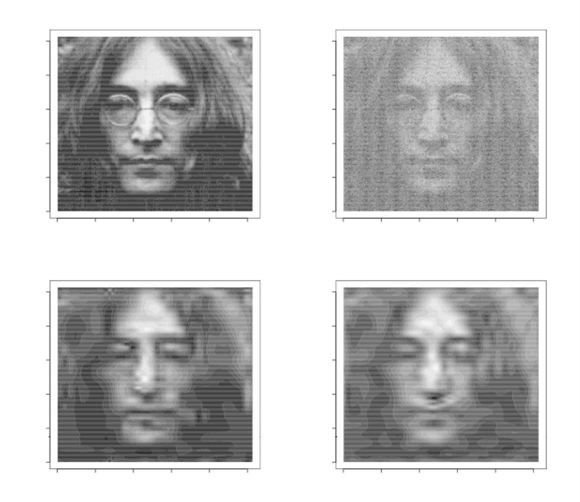
- $f(x) = \sum_{m=1}^{M} \beta_m \phi_m(x)$ regression spline with basis functions ϕ
- wavelet basis functions are orthogonal makes fitting easier
- also multi-resolution – able to track local wiggles better
- very useful for image processing, signal processing can find edges and short bursts
- `wavethresh`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

Aside: wavelets





… Aside: wavelets



Vidakovic and Mueller, "Wavelets for kids (Part I)" 1994.

Regularization and smoothing splines

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

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

Regularization and smoothing splines

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

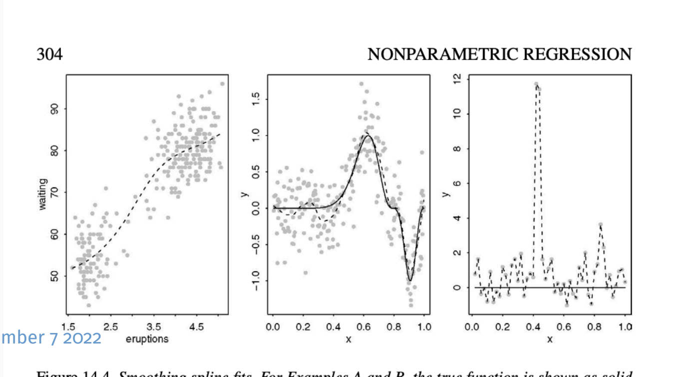
$$\hat{f} = \{\hat{f}(x_1), \ldots, \hat{f}(x_n)\} = (I + \lambda K)^{-1} y$$

K is a symmetric $n \times n$ matrix of rank $n - 2$

… Regularization and smoothing splines

ELM 14.2, SM 10.7.2

- smoothing splines available in `base::smooth.spline`
- amount of smoothing can be specified; if not, will be automatically computed
- `predict.smooth.spline` can also predict derivatives
- for generalized linear models, `mgcv::gam` or `gam::gam` allow linear predictor to smooth functions of one or more covariates



Example: logistic regression

SM Ex.10.29 and 10.32

516 *10 · Nonlinear Regression Models*

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).

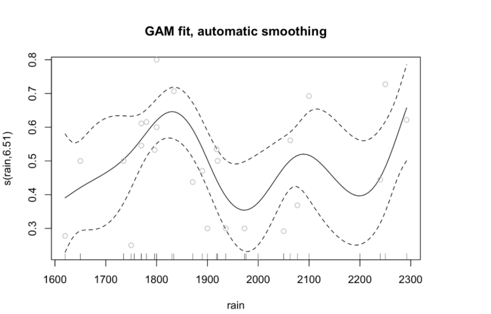
Terms	df	Deviance
Constant	33	74.21
Linear	32	74.09
Quadratic	31	74.09
Cubic	30	62.63

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

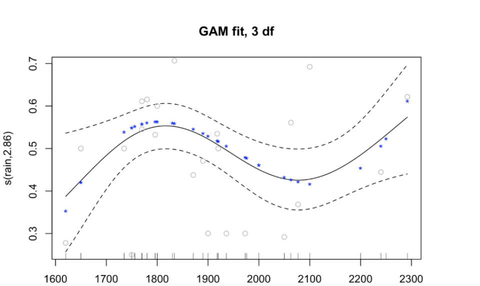
Example: logistic regression

See: `toxoplasmosis.Rmd` for several versions

```
plot(toxo.gam, seWithMean=TRUE, trans= ilogit, main = "GAM fit, automatic smoothing")
with(toxo, points(rain, r/m, col="gray"))
```




```
plot(toxo.gam2, seWithMean=TRUE, trans=ilogit, main="GAM fit, 3 df")
with(toxo, points(rain, r/m, col="gray"))
points(toxo$rain, toxo.glm5$fitted.values, pch="*", col="blue" )
```



Example: The NMMAPS studies

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,

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

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

Example: The NMMAPS studies

- 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

- $Y_t \sim Poisson(\mu_t)$ generalized additive model `gam`

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

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

... the NMMAPS studies

- $Y_t \sim Poisson(\mu_t)$ generalized additive model `gam`
- $\log(\mu_t) =$ age specific intercepts $+ \beta PM_t + \gamma DOW + s(t, 7) + s(temp_t, 6) + s(temp_{t-1}, 6) + s(dewpoint_t, 3) + s(dewpoint_{t-1}, 3) + s(dew_0, 3) + s(dew_{1-3}, 3)$
- three ages categories; separate intercept for each ($< 65, 65 - 74, \geq 75$)
- 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

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

Generalizations

- generalized to several explanatory variables by smoothing each variable separately
- generalized to likelihood methods by replacing $\sum\{y_j - g(x_j)\}^2$ by $\sum \log f\{y_j; \eta_j\}$
- $\eta_j = g(x_j)$ or
  $\eta_j = g_1(x_{1j}) + g_2(x_{2j}) + \cdots + g_p(x_{pj})$ or
  $\eta_j = x_j^T\beta + g(t_j)$
- last is used in SM 10.7.3 for spring barley data:

$$y_{vb} = g_b(t_{vb}) + \beta_v + \epsilon_{vb}$$

- allow block effects to depend on location (t_{vb}) in a 'smooth' way

Multidimensional splines

- so far we have considered just 1 X at a time
- for regression splines we replace each X by the new columns of the basis matrix
- for smoothing splines we get a univariate regression

- it is possible to construct smoothing splines for two or more inputs simultaneously, but computational
  difficulty increases rapidly
- these are called thin plate splines

- implemented in `gam(mgcv)` as `bs = "tp"` in `s(x1,x2, ...)`

When to use "non-parametric" fits?

- 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

- linear regression: interpretation of β as partial derivatives, inference conditional on X, $\mathrm{E}(Y)$ is linear in β; very flexible (IJALM); decomposition of variance; testing sets of coefficients; factor variables; orthogonality; model selection; model building; hierarchical structure; transformation of y and/or x; Lasso and ridge regression

Summary

- linear regression: interpretation of β as partial derivatives, inference conditional on X, $\mathrm{E}(Y)$ is linear in β; very flexible (IJALM); decomposition of variance; testing sets of coefficients; factor variables; orthogonality; model selection; model building; hierarchical structure; transformation of y and/or x; Lasso and ridge regression

- generalized linear models: binary and binomial responses, logistic regression, residual deviance; likelihood-based inference; Poisson regression; overdispersion; link function, mean function, variance function; Gamma regression; iteratively re-weighted least squares; Pearson's chi-squared

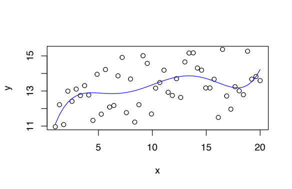
Summary

- linear regression: interpretation of β as partial derivatives, inference conditional on X, $\mathrm{E}(Y)$ is linear in β; very flexible (IJALM); decomposition of variance; testing sets of coefficients; factor variables; orthogonality; model selection; model building; hierarchical structure; transformation of y and/or x; Lasso and ridge regression

- generalized linear models: binary and binomial responses, logistic regression, residual deviance; likelihood-based inference; Poisson regression; overdispersion; link function, mean function, variance function; Gamma regression; iteratively re-weighted least squares; Pearson's chi-squared

- non-parametric regression: kernel smoothing, basis functions, smoothing splines, cross-validation, generalized additive models

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$ 1st column of X?
- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 \epsilon_i$
- $y_i = \beta_0 \pm \beta_1 + \epsilon_i$
- $y_i = \beta_0 + \beta_1 \sin(x_i) + \beta_2 \cos(x_i) + \epsilon_i$
- $y_i = \gamma_0 x_{1i}^{\gamma_1} x_{2i}^{\gamma_2} \eta_i, \quad \eta_i \sim$ positive r.v.
- $y_i = \varphi_0 + \sum_{k=1}^{K} \varphi_k s_k(x_i) + \epsilon_i$ e.g smoothing splines



... Summary

- principles: components of investigations, workflow, experiments and observational studies, design of studies, unit of study and unit of analysis, ecological bias, causality, support for causality in observational studies, case-control studies, measures of risk; meta-analysis; prospective/retrospective sampling

... Summary

- principles: components of investigations, workflow, experiments and observational studies, design of studies, unit of study and unit of analysis, ecological bias, causality, support for causality in observational studies, case-control studies, measures of risk; meta-analysis; prospective/retrospective sampling
- survival data: hazard function, parametric models, nonparametric estimation, proportional hazards regression (Cox model); censoring; partial likelihood

... Summary

- principles: components of investigations, workflow, experiments and observational studies, design of studies, unit of study and unit of analysis, ecological bias, causality, support for causality in observational studies, case-control studies, measures of risk; meta-analysis; prospective/retrospective sampling

- survival data: hazard function, parametric models, nonparametric estimation, proportional hazards regression (Cox model); censoring; partial likelihood

- mixed and random effects: components of variance, random factors, nested and crossed factors; multi-level data; expected mean squares

Examples

- hydroxychloroquine NEJM ; citation impacts of humour; health benefits of tea — Sep 21
- peer review biased by author prominence — Sep 28
- well-being, religiosity, and SES (PNAS) — Oct 12
- Challenger O-ring failure — Oct 19
- decline of Shopify — Oct 26
- sleep and video gaming (PNAS) — Nov 2
- anxiety and exam performance — Nov 16
- ANDROMEDA trial of treatment for septic shock — Nov 16
- mask use in schools – natural experiment — Nov 23

Homework Notes

- HW 1 – ridge regression; extensive notes from TA in solutions
- HW 2 – orthogonal polynomials (simulation)
- HW 3 – Box & Cox transformation; residual plots and smoothers
- HW 4 – errors in covariates; delta method
- HW 5 – glm diagnostics
- HW 6 – math education and brain development hmmmm
- HW 7 – negative binomial (Poisson with Gamma prior)
- HW 8 – solution to ML equations for GLMs, variance modelling, multinomial as conditional Poisson
- HW 9 – parametrization issues
- HW 10 – biomarkers and all-cause mortality