Methods of Applied Statistics I

STA2101H F LEC9101

Week 12

December 3 2020

link re photo

STOCK PHOTO
#4993261 finish line by kikkerdirk

lineups

Today

Start Recording

1. Final HW due December 20 – extended
2. Lineups – Chenghui Zheng
3. Nonparametric regression overview
4. Strategies for Modelling
5. Course Overview
6. In the News

- December 7 15.00 – 16.00 Margaret Roberts
- `https://canssiontario.utoronto.ca/?mec-ev`
- "Resilience to online censorship"

- Dec. 4 noon – Kathryn Roeder
- Dec. 14 3 pm – Kosuke Imai

Recap

- Nonparametric regression — mis-named
- local polynomial regression — local least squares; kernel; bandwidth
- regression splines – you control the terms — `ns(var, df); bs(var, df)`
- smoothing splines $\min_f \sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt$
- `mgcv::s; mgcv::gam`

  "Smooth terms are specified in a gam formula using s, te, ti and t2 terms.... The smooths built into the mgcv package are all based one way or another on low rank versions of splines "
- `gam::gam; gam::s`

  "Built-in nonparametric smoothing terms are indicated by s for smoothing splines or lo for loess smooth terms. "

Recap 2

Inference after fitting smooth functions relies on their representation as:

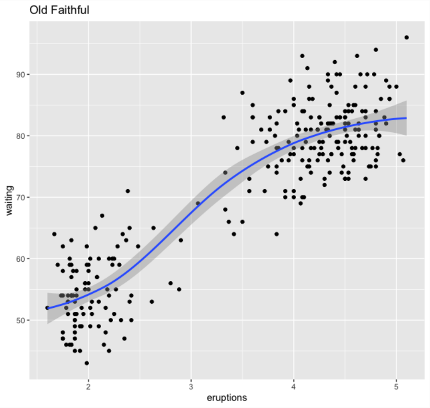
$$\begin{pmatrix} \hat{f}(x_1) \\ \hat{f}(x_2) \\ \vdots \\ \hat{f}(x_n) \end{pmatrix} = S_\lambda \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

where S_λ is an $n \times n$ smoothing matrix. The form of the matrix will depend on the smoothing method used. But the entries don't depend on y (unless λ was estimated using CV on the same data)

This enables study of the expected value and variance of the n-vector $\hat{\boldsymbol{f}}$, and confidence bands are usually constructed pointwise, as ± 1.96 times the estimated standard error of $\hat{f}(x_i), i = 1, \ldots, n$. Sometimes the confidence intervals are also constructed at non-data xs, to give a smoother curve.

The technical details are given in SM Ch 10.7 and in more detail in ESLR II Ch 5, 6.

not for the faint-hearted


Old Faithful
waiting
eruptions
90
80
70
60
50
2
3
4
5

When to use "non-parametric" fits?

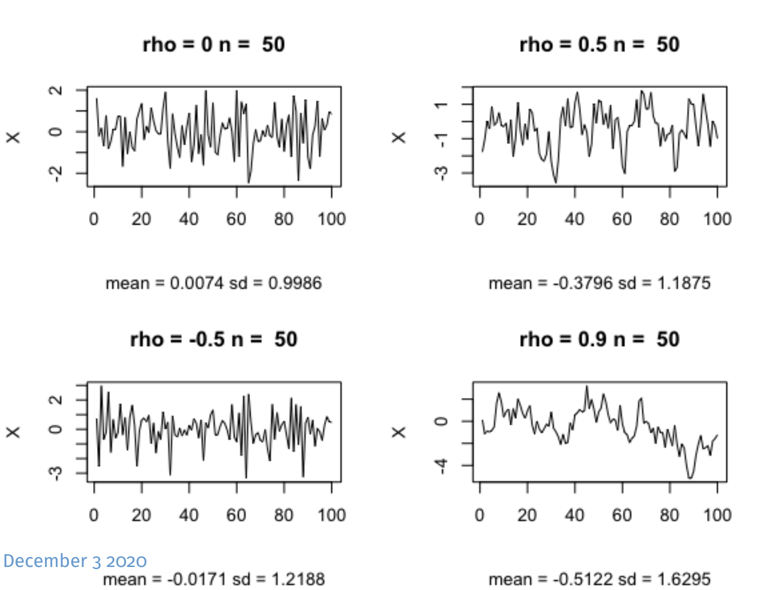
- depends on the problem
- some fields of science have their own conventions e.g. mortality and air pollution, NMMAPS
- may be useful for confounding variables
- may be useful for exploratory analyses
- Faraway suggests using smoothing methods when there is "not too much" noise in the data
- and parametric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

Aside: negative correlation

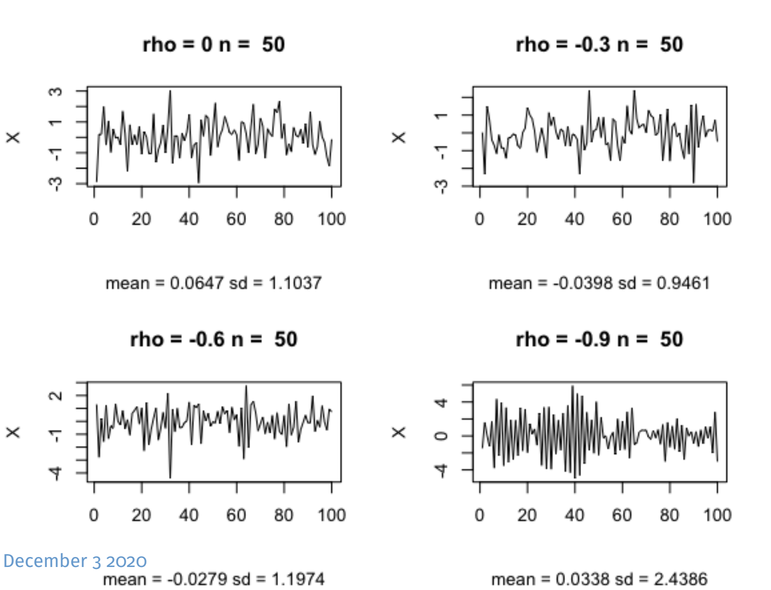
- last week we showed that positive correlation can increase the estimated variance of an estimate — e.g. sample mean
- $\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}(1 + k\rho)$, for example
- What if the correlation is negative? — actually can't be if $\mathrm{corr}(X_i, X_j) = \rho$ for all i, j

- Correlation works to our advantage, however, in paired experiments:
- data $(Y_i, X_i), i = 1, \ldots, n$ represent, say before and after measurements on subject i
- Differences are less variable: let $D_i = Y_i - X_i$ then $\mathrm{E}(\bar{D}) = \mu_Y - \mu_X$ and $\mathrm{var}(\bar{D}) = 2\sigma^2(1 - \rho)/n$

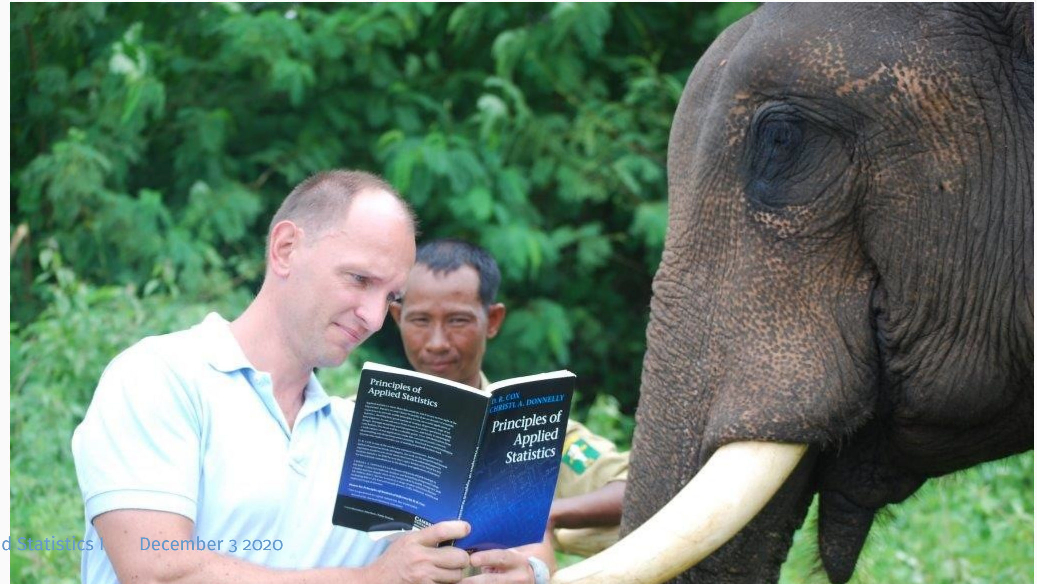
Some time series plots



Some time series plots



How do we choose a model?



How do we ~~choose~~ formulate a model? CD Ch. 6

1. preliminaries — the role of models, level of detail
2. nature of probability models — what is random; what is fixed; what parameters are of interest; when to use nonparametric
3. types of models — substantive vs empirical
4. interpretation of probability — generalizability, extrapolation
5. empirical models — Nov 26

How do we choose a model?

- criteria for parameters — CD §7.1
- non-specific effects — CD §7.2
- choice of a specific model — CD §7.3

Model Choice

- often this will involve at least two levels of choice, first between distinct separate families and then between specific models within a chosen family
- of course all choices are to some extent provisional
- example: survival data – gamma or weibull model both extend the exponential
- example: linear regression $E(Y) = \beta_0 + \beta_1 x$, or nonlinear regression $E(Y) = \gamma_0/(1 + \gamma_1 x)$
- neither, one, or both may be adequate

... choice of a specific model

- comparisons between models are sometimes made using Bayes factors, ... however, misleading if neither model is adequate
- for dependencies of y on x that are curved, a low-degree polynomial might be adequate
- but subject-matter may suggest an asymptote, in which case $E(Y) = \alpha + \gamma e^{-\delta x}$ may be preferred

… model choice with a natural hierarchy

- polynomials provide a flexible family of smooth relationships, although poor for extrapolation
- it will typically be wise to measure the x_i from a meaningful origin near the centre of the data

- example: $E(Y) = \beta_{00} + \beta_{10}x_1 + \beta_{01}x_2 + \beta_{20}x_1^2 + \beta_{11}x_1x_2 + \beta_{02}x_2^2$
- it would not normally be sensible to include β_{11}, and not β_{20}, β_{02}

- with qualitative (categorical) x's, this means models with interaction terms should include the corresponding main effects

… model choice

- example: $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$
- example: time series $AR(p)$
  $y_t = \mu + \rho_1(y_{t-1} - \mu) + \cdots + \rho_p(y_{t-p} - \mu) + \epsilon_t$
- for a single set of data choose the smallest order compatible with the data, using standard tests
- for several sets of data, usually would choose the same order for each set

… choosing among explanatory variables

- response y, potential explanatory variables $x_1, \ldots, x_p$
- suppose interest focusses on the role of a particular variable or set of variables, x^*
- the value, standard error, and interpretation of the coefficient of x^* depends on which other variables are included
- variables prior to x^* in the generating process should be included in the model unless…
- unless these variables are conditionally independent of y, given x^* (and other variables in the model)
- OR unless they are conditionally independent of x^*, given other variables in the model
- variables intermediate between x^* and y are omitted in initial assessment of the effect of x^*
- but may be needed later to study the pathways of dependence

… choosing among explanatory variables

- relatively mechanical methods of choosing may be helpful in preliminary exploration, but are insecure as a basis for final interpretation
- explanatory variables not of direct interest, but known to have a substantial effect, should be included
- several different models may be equally effective
- if there are several potential explanatory variables on an equal footing, interpretation is particularly difficult

- A two-phase approach:
- First search among a large number of possibilities for a base for interpretation
- Second check the adequacy of that base

First phase: a broad strategy

- x^*, required explanatory variables; $\tilde{x}$ some potential further explanatory variables
- $\tilde{x}$ conceptually prior to x^*

- fit a reduced model with x^* only $\mathcal{M}_{\text{red}}$
- fit, if possible, a full model with x^* and $\tilde{x}$ $\quad \mathcal{M}_{\text{full}}$
- compare the estimated standard errors of the coefficients for x^* under the two models

- if these are of the same order, then $\mathcal{M}_{\text{full}}$ is safer
- if precision improvement under $\mathcal{M}_{\text{red}}$ seems substantial, then explore eliminating some of $\tilde{x}$
- for example with backwards elimination

- with emphasis on the effect of x^*

Second phase: adequacy of the model

- add back selected components of the omitted variables $\tilde{x}$
- to check that conclusions are not changed
- or to report on the differences if they are
- if the model to date has been linear, may be important now to check some curvature terms, for continuous xs, and interaction terms for categorical xs
- these provide a 'warning system', but not usually direct interpretation

- interpretation of coefficients, especially in observational studies, needs care
- example: x includes several measurements of smoking behaviour: yes/no; years since quitting; no. of cigarettes smoked; pipe/cigar; etc.
- role of these depends on the goal of the study – confounder? primary exposure?

Nuclear plant data

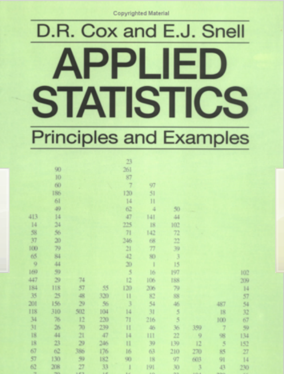
SM §8.7

- data on p. 401
- SM analysis 1: full, backward, forward

FLM calls these hierarchical

- SM analysis 2: *AIC* and AIC_c

FLM calls these criterion-based



$\longrightarrow$ nuclear.R

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 +$$

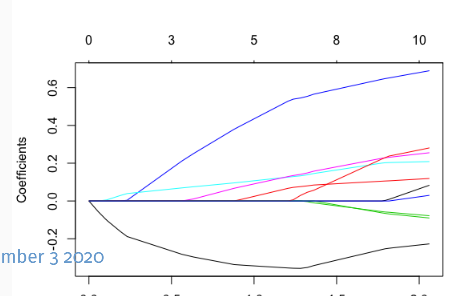
equivalent to

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

several other shrinkage methods:

ridge, PCA, PLS

high-dimensional inference ISLR §§6.4



Lasso

```
lasso.coef <-
predict(glmnet(x,y,alpha=1,lambda = grid ), type="coefficients", s = bestlam)
> lasso.coef
12 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept) -11.94255
(Intercept)   .
pt           -0.27248
ct            0.10113
log(cum.n)   -0.04792
log(cap)      0.62993
date          0.19049
ne            0.21351
log(t1)       .
log(t2)       0.19188
pr           -0.05141
bw            .
```

Summary Fall 2020

web page, timetable: Advanced topics in statistics and data analysis with emphasis on applications. Diagnostics and residuals in linear models, introduction to generalized linear models, graphical methods, additional topics such as random effects models, split plot designs, analysis of censored data, introduced as needed in the context of case studies.

web page, course descriptions: 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
- designed experiments: factors, anova, blocking, randomized blocks, components of variance, randomization, causality
- observational studies: retrospective/prospective, Bradford-Hill criteria, 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, principles of measurement, 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
- visualization

In the News

- wildfire Sep 10
- covid testing Sep 17, 24, Oct 1, 8
- covid prediction Sep 24
- hydroxychloroquine Sep 24
- A-level grades (UK) Oct 8
- polls Oct 8, 15. Nov 5, 12
- covid misinformation Oct 15, HW 2
- four cardinal rules Oct 22
- pollution and mortality Nov 26
- alcohol, ballots, mentoring, remdesivir (C19) Nov 26