

The next weeks

Homework 3: due April 2, 5 pm

(5)

Final Test: April 17, 1 - 3 pm SS 2105

Office Hours:

Thursday, April 5, 1 - 3 pm

Friday, April 13, 10 - 12 am

Monday, April 16, 2 - 4 pm

email
if nec.
before 3

[HOME](#)[NEWS](#)[SPORT](#)[FINANCE](#)[COMMENT](#)[BLOGS](#)[CULTURE](#)[TRAVEL](#)[LIFESTYLE](#)[FASHION](#)[TECH](#)[Data](#)[UK](#)[World](#)[Politics](#)[Obituaries](#)[Education](#)[Earth](#)[Science](#)[Defence](#)[Health News](#)[Royal Family](#)[Celebrities](#)[Weird News](#)[University Education](#)[University Course Finder](#)[League Tables](#)[Primary](#)[Secondary](#)[Expat Education](#)[Make Britain Count](#)[HOME](#) » [EDUCATION](#) » [MAKE BRITAIN COUNT](#)

Numeracy campaign: 'It's never too late to brush up on your maths' says Joan Bakewell

Our numeracy campaign is now gaining support from readers, experts and celebrities. Here Joan Bakewell gives her support for learning maths at an early age .

[Share](#) [✉](#) [🖨](#)[Facebook](#) 6[Twitter](#) 55[LinkedIn](#) 9[+1](#) 0

[Make Britain Count](#)
[Education](#) »

SMART HOME MO
ALWAYS CONNE



[MAKE BRITAIN COUNT ON FACEBO](#)



[Numeracy and Ma](#)

“ I did enough statistics at Cambridge, though, to have remained suspicious of percentages ever since that time. I don't trust them for a minute because they can be distorted to prove anything. So when crime statistics are reported and they say there has been a 50 per cent drop in burglary in my area, I know that might easily mean that there have been two break-ins rather than three. ”

Thanks to the Forsooth! ⁴ column of the RSS News (April, 2012)

2 ↑ 3 50% inc.

Smoothing regressions?

- ▶ **kernel smoothers** fit locally weighted polynomials, using a kernel function as weights
- ▶ in R can use ksmooth (base) or `sm.regression` in `library(sm)`
- ▶ a more robust version is implemented in loess (base)
- ▶ kernel smoothing useful for graphical summaries, for exploring effect of bandwidth, for single explanatory variable
- ▶ refinements (in addition to loess), include adaptive bandwidth, running medians, running M -estimates

local linear regr.
downweight outliers

... smoothing regressions \leftarrow cubic splines \leftarrow wavelet basis

▶ regression splines use a set of basis functions, and fit $E(y | x) = \sum_{m=1}^M \beta_m h_m(x)$ \leftarrow linear model

- ▶ natural splines and B-splines are popular choices
- ▶ once the basis functions are chosen, fitting is by `lm` or `glm`
- ▶ you choose the number of basis functions for each explanatory variable
- ▶ implemented in R in `ns(x, df = 4)` and `bs(x, df = 4)`
- ▶ generalizations include different types of basis functions, e.g. Fourier basis (sine and cosine), or wavelet basis (good for extracting local behaviour)
- ▶ standard errors are computed by the usual methods for `lm` and `glm`



effect of an x , assessed by LRT (step \checkmark)

... smoothing regressions

- ▶ cubic smoothing splines put knots at each observations
- ▶ and shrink coefficients β_m by regularization
- ▶ popular because they provide smooth fits
- ▶ popular because they are “optimal”:



$$\min_{\underline{g}} \sum_{j=1}^n \{y - g(t_j)\}^2 - \lambda \int_a^b \{g''(t)\}^2 dt, \quad \lambda > 0 \quad \therefore$$

- ▶ has an explicit, finite-dimensional solution:

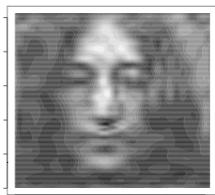
$$\min_{\underline{g}} (y - \underline{g})^T (y - \underline{g}) + \lambda \underline{g}^T K \underline{g}$$



- ▶ $\underline{g} = \{g(x_1), \dots, g(x_n)\}$

$$\hat{\underline{g}} = (\mathbf{I} + \lambda \mathbf{K})^{-1} \underline{y}$$

... wavelets



Vidaković and Müller, "Wavelets for kids (Part I)" 1994.
See also [Amara's wavelet page](#)

... smoothing regressions

- ▶ generalized to several explanatory variables by smoothing each variable separately
- ▶ generalized to likelihood methods by replacing $\sum \{y_j - g(x_j)\}^2$ by $\sum \log f\{y_j; \eta_j\}$

*lm \rightarrow glm
 $\hat{\sigma}^2$
smaller*

- ▶ $\eta_j = g(x_j)$ or
 $\eta_j = g_1(x_{1j}) + g_2(x_{2j}) + \dots + g_p(x_{pj})$ or
 $\eta_j = \mathbf{x}_j^T \beta + g(t_j)$

$$y_{vb} = \alpha_b + \beta_v + \epsilon_{vb}$$

- ▶ last is used in §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
- more precise comp. of $\beta_v - \beta_v$*

Example: health effects of air pollution

Journal of the
Royal Statistical Society

SERIES A
Statistics
in Society

Model choice in time series studies of air pollution and mortality



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

Article first published online: 14 FEB 2006

DOI: 10.1111/j.1467-985X.2006.00410.x

Issue



Journal of the Royal
Statistical Society: Series A
(Statistics in Society)

Volume 169, Issue 2, pages
179–203, March 2006

Additional Information [\(Show All\)](#)

SEARCH

In this issue

Advanced > Saved

ARTICLE TOOLS

- Get PDF (648K)
- Save to My Profile
- E-mail Link to this Article
- Export Citation for this Article
- Get Citation Alerts
- Request Permission

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, ...
- ▶ Reference: Peng R., Dominici F., Louis T., (2006) JRSS A, 169, 179-203

... the NMMAPS studies

▶ $Y_t \sim \text{Poisson}(\mu_t)$

for each city: $PM_{10,t-1}$
↓ day of week

▶ $\log \mu_t =$ age specific intercepts + $\beta PM_t + \gamma DOW +$
 $g(t, df) + s(temp_t, 6) + s(temp_{t-1}, 6) + s(dewpoint_t, 3) +$
 $s(dewpoint_{t-1}, 3) + \cancel{s(dew_{t-2}, 3)} + \cancel{s(dew_{t-3}, 3)}$

▶ three ages categories; separate intercept for each (< 65 ,
 $65 - 74$, ≥ 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

$$\hat{\beta}_c, \text{se}(\hat{\beta}_c)$$
$$\hat{\beta} = \sum_c \hat{\beta}_c w_c$$

▶ very difficult to separate out weather and pollution effects

▶ also relevant: Crainiceanu, C., Dominici, F. and Parmigiani, G. (2006). Adjustment Uncertainty in Effect Estimation

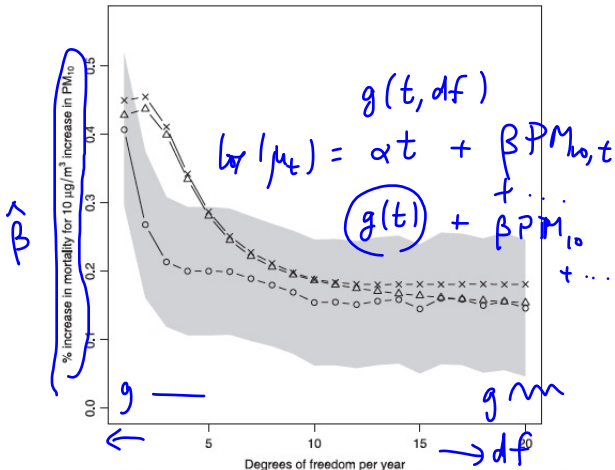


Fig. 3. Sensitivity analysis of the national average estimate of the percentage increase in mortality for an increase in PM_{10} of $10 \mu\text{g m}^{-3}$ at lag 1: city-specific estimates were obtained from 100 US cities using data for the years 1987–2000 and the estimates were combined by using a hierarchical normal model (○, GLM-NS; △, GAM-R; ×, GAM-S; ■, 95% posterior intervals for the estimates obtained by using GLM-NS)

Fitting generalized additive models

R package mgcv; functions gam and gamm

```
> dat = gamSim(1, n=400, dist="normal", scale=2)
```

```
> b = gam(y ~ s(x0) + s(x1)+s(x2)+s(x3), data = dat)
```

```
> plot(b, pages=1, seWithMean = T)
```

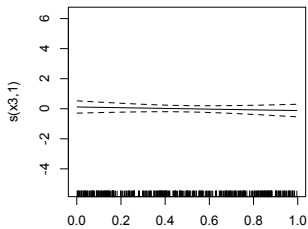
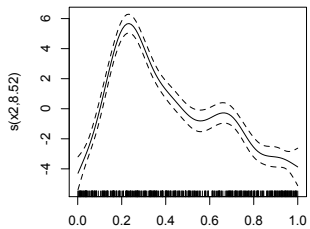
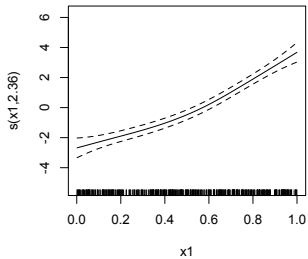
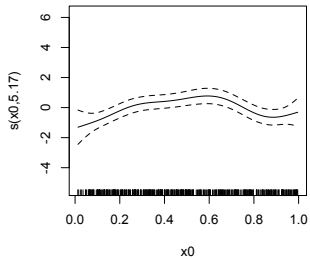
dat ↗

$$y = 2 \sin(\pi x_0) + \exp(2x_1) + \text{poly}(x_2, \text{degree} = 11) + \epsilon$$

↑ β_1 ↑ β_2 ↑ 2

Reference: Wood (2006) *Generalized Additive Models: An Introduction with R*.

Faraway Extending the linear model



Shrinkage Methods (HTF §3.4)

- ▶ Ridge regression



$$\begin{aligned}\hat{\beta}_{LS} &= (X^T X)^{-1} X^T y \\ \hat{\beta}_{ridge} &= (X^T X + \lambda I)^{-1} X^T y\end{aligned}$$

- ▶ can show that $\hat{\beta}_{ridge}$ satisfies

$$\begin{aligned}\min_{\beta} \left(\sum \{y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j\}^2 + \lambda \sum_{j=1}^p \beta_j^2 \right) \\ \min_{\beta} \sum \{y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j\}^2 \quad \text{s.t. } \sum \beta_j^2 \leq t\end{aligned}$$

- ▶ Assume x_j 's are centered and put these in matrix X (with no column of 1's):

$$\min_{\beta} (y - X\beta)^T (y - X\beta) \quad \text{s.t. } \|\beta\|^2 \leq t$$

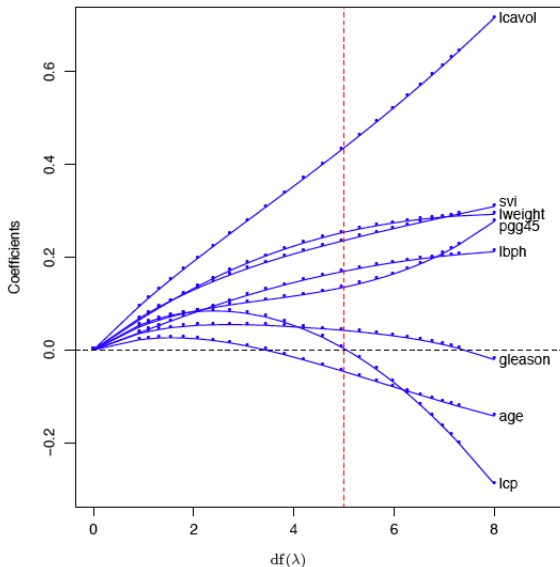
... ridge regression

- ▶ $\min_{\beta} \{(y - X\beta)^T (y - X\beta) + \lambda \|\beta\|^2\}$
- ▶ λ is a tuning parameter: $\lambda = 0$ gives $\hat{\beta}_{LS}$, $\lambda \rightarrow \infty$
- ▶ in R the `library MASS library(MASS)` has a ridge regression version of `lm` called `lm.ridge`
- ▶ if columns of X are nearly linearly dependent (multicollinearity), $\hat{\beta}$'s for these columns should be shrunk towards 0.
- ▶ essential that the predictors are all scaled to the same units
- ▶ this is difficult for interpretation of the coefficients

$$\begin{aligned}
X\hat{\beta}_{ridge} &= X(X^T X + \lambda I)^{-1} X^T y \\
&= UDV^T (VD^2 V^T + \lambda I)^{-1} VDU^T y \\
&= UDV^T (VD^2 V^T + \lambda VV^T)^{-1} VDU^T y \\
&= UD(D^2 + \lambda I)^{-1} DU^T y \\
&= \sum_{j=1}^p u_j \left(\frac{d_j^2}{d_j^2 + \lambda} \right) u_j^T y
\end{aligned}$$

$$df(\lambda) = \text{tr}[X(X^T X + \lambda I)^{-1} X^T] = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}$$

$df(\lambda)$ called **effective number of parameters**



$$df(x) \propto \frac{1}{\lambda}$$

FIGURE 3.8. Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $df(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $df = 5.0$, the value chosen by cross-validation.

Lasso



$$\min_{\beta} \left(\sum \{y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j\}^2 + \lambda \sum_{j=1}^p |\beta_j| \right)$$



$$\min_{\beta} \sum \{y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j\}^2 \quad \text{s.t.} \quad \sum |\beta_j| \leq t$$

- ▶ quadratic programming problem

- ▶ $\hat{\beta}^{lasso}$ is nonlinear function of y

- ▶ Tibshirani (1996), JRSS B and (2011), JRSS B

- ▶ [http://http:](http://http://www-stat.stanford.edu/~tibs/lasso.html)

[//www-stat.stanford.edu/~tibs/lasso.html](http://www-stat.stanford.edu/~tibs/lasso.html)

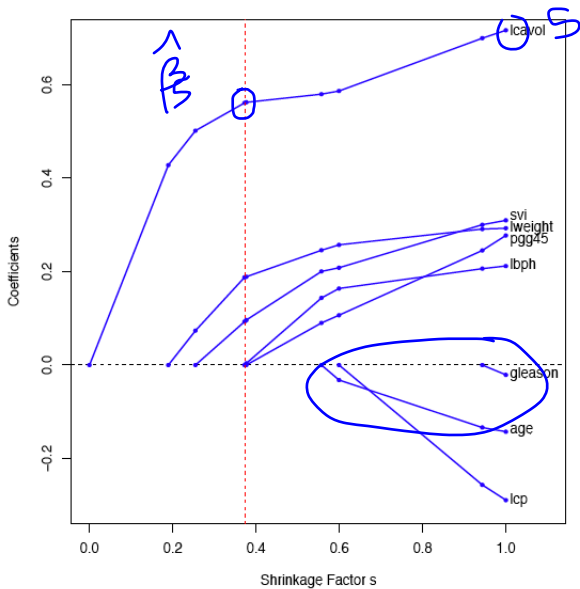


FIGURE 3.10. Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t / \sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso

... smoothing

- ▶ ridge regression gives “proportional shrinkage”
- ▶ subset selection gives “hard thresholding” (some $\beta_j \rightarrow 0$)
- ▶ lasso gives “soft thresholding”: blend of shrinkage and zeroing
- ▶ elastic net combines lasso and ridge regression


$$\min_{\beta} \left(\sum \{y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j\}^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right)$$

- ▶ implemented in R in `library(glmnet)`
- ▶ estimates of coefficients are biased (but may have small mean-squared error)
- ▶ Lasso is now used as a variable selection method
- ▶ improvements in algorithms allow fast computation even for $p > n$
- ▶ used in McShane et al (2010) to predict the historical temperature record from proxy data

Coverage for test

- ▶ 4 questions
 - ▶ one theory question ← math
 - ▶ one applied question ← interp. of output
 - ▶ one question from HW ←
 - ▶ one "study" question ← read & discover etc.
 - ▶ one question with computer output
-
- ▶ Cox and Donnelly, Chapters 1 (Some general concepts) and 2 (Design of Studies)
 - ▶ Davison: 9.1, 9.2, 9.3.1, 9.4 (omit split unit experiments)
 - ▶ Davison: 10.3, 10.4, 10.5.2 (omit Marginal Models), 10.6 (omit quasi-likelihood)
 - ▶ Davison: 10.7: omit inference (p. 525), omit computation of bias and variance (p. 529); combination of HW 3 question and material from slides will suffice

Topics and concepts

- ▶ experiments and observational studies, randomization, causality, unit-treatment additivity, random and systematic error, unit of study and analysis
- ▶ randomized block designs, two-way analysis of variance
- ▶ components of variance, fixed and random effects, non-specific effects and 'stable treatment effect' 
- ▶ generalized linear models; iteratively reweighted least squares, deviance, deviance and Pearson residuals, link functions, scale parameter, interpretation of parameters, overdispersion
- ▶ special classes: normal, binomial, gamma, Poisson; regression with binary and with count data
- ▶ generalized estimating equations and quasi-likelihood, generalized linear mixed models **not on test**
- ▶ semiparametric regression: kernel, regression splines, smoothing splines
- ▶ penalized methods: ridge regression, lasso, elastic net