Today

- HW 4: due April 11

- Final Exam: April 11 2:00 – 5:00 pm SS 1085
- in the news
- semi-parametric regression

- March 28: §10.8; proportional hazards regression

In the News

Globe and Mail March 17

Home » **News** » **National**

How losing 18,000 people made Manitoba $100-million poorer

JOE FRIESEN

DEMOGRAPHICS REPORTER — The Globe and Mail

Published Monday, Mar. 17 2014, 6:00 AM EDT

Last updated Monday, Mar. 17 2014, 6:00 AM EDT

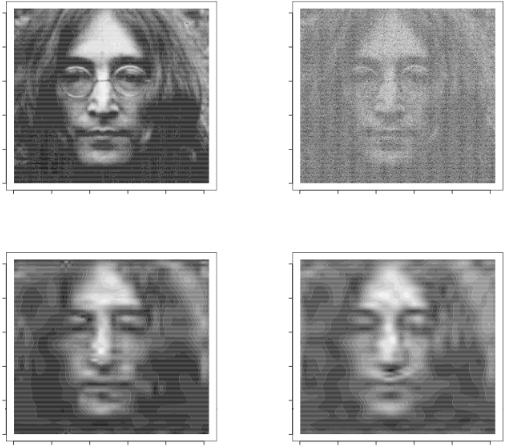
Smoothing regressions?

- 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

... smoothing regressions

- regression splines use a set of basis functions, and fit $E(y \mid x) = \sum_{m=1}^{M} \beta_m h_m(x)$
- 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)
  e.g. wavelet basis (good for extracting local behaviour)
- standard errors are computed by the usual methods for `lm` and `glm`

... wavelets



Vidaković and Mueller, ”Wavelets for kids (Part I)” 1994.

... 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":
- $$\min_g \sum_{j=1}^{n} \{y - g(t_j)\}^2 - \lambda \int_a^b \{g''(t)\}^2 dt, \quad , \lambda > 0$$
- has an explicit, finite-dimensional solution:
$$\min_{\underline{g}} (y - \underline{g})^T (y - \underline{g}) + \lambda \underline{g}^T K \underline{g}$$
- $\underline{g} = \{g(x_1), \ldots, g(x_n)\}$

... smoothing regressions

- `gam` in `library(gam)` fits cubic smoothing splines

  Hastie, Tibshirani & Friedman, Ch. 5
- `gam` in `library(mgcv)` fits penalized regression splines

  Wood, 2001
- see also help files for `gam(mvcv)`
- estimation of standard errors is more straightforward in `gam(mvcv)`
- excellent explanation in Appendix A of

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

... 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 $\sum \log f\{y_j; \eta_j\}$
- $\eta_j = g(x_j)$ 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 §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

Table 10.21 Spring barley data (Besag *et al.*, 1995). Spatial layout and plot yield at harvest y (standardized to have unit crude variance) in a final assessment trial of 75 varieties of spring barley. The varieties are sown in three blocks, with each variety replicated thrice in the design. The yield for variety 27 is missing in the third block.

	Block 1		Block 2		Block 3	
Location t	Variety	Yield y	Variety	Yield y	Variety	Yield y
1	57	9.29	49	7.99	63	11.77
2	39	8.16	18	9.56	38	12.05
3	3	8.97	8	9.02	14	12.25
4	48	8.33	69	8.91	71	10.96
5	75	8.66	29	9.17	22	9.94
6	21	9.05	59	9.49	46	9.27
7	66	9.01	19	9.73	6	11.05
8	12	9.40	39	9.38	30	11.40
9	30	10.16	67	8.80	16	10.78
10	32	10.30	57	9.72	24	10.30
11	59	10.73	37	10.24	40	11.27
12	50	9.69	26	10.85	64	11.13
13	5	11.49	16	9.67	8	10.55
14	23	10.73	6	10.17	56	12.82
15	14	10.71	47	11.46	32	10.95
16	68	10.21	36	10.05	48	10.92
17	41	10.52	64	11.47	54	10.77
18	1	11.09	63	10.63	37	11.08
19	64	11.39	33	11.03	21	10.22
20	28	11.24	74	10.85	29	10.59
21	46	10.65	13	11.35	62	11.35
22	73	10.77	43	10.25	5	11.39
23	37	10.92	3	10.08	70	10.59
24	55	12.07	53	10.25	13	11.26
25	19	11.03	23	9.57	11	11.79
26	10	11.64	62	11.34	44	12.25
27	35	11.37	52	10.19	36	12.23
28	26	10.34	12	10.80	52	10.84
29	17	9.52	2	10.04	60	10.92
30	71	8.99	32	9.69	68	10.41
31	8	8.34	22	9.36	3	10.96
32	62	9.25	42	9.43	19	9.94
33	44	9.86	72	11.46	67	11.27
34	53	9.90	73	9.29	59	11.79
35	74	11.04	25	10.10	2	11.51
36	20	10.30	45	9.53	75	11.64
37	56	11.56	15	10.55	27	—
38	29	9.69	35	11.34	43	9.78
39	2	10.68	66	11.36	51	8.86
40	47	10.91	5	10.88	10	10.28
41	11	10.05	56	11.61	35	12.15
42	38	10.80	46	10.33	74	10.36
43	65	10.06	71	10.53	66	9.59
44	13	10.04	51	8.67	34	10.53
45	31	10.50	21	9.56	18	11.26
46	40	9.51	1	9.95	50	10.37
47	4	9.20	31	11.10	42	10.10
48	67	9.74	11	10.11	1	9.95
49	22	8.84	41	9.36	58	9.80
50	49	9.33	61	10.23	26	10.58
51	58	9.51	55	11.38	41	9.31
52	43	9.35	14	11.30	25	9.29

Table 10.21 (*cont.*)

	Block 1		Block 2		Block 3	
Location t	Variety	Yield y	Variety	Yield y	Variety	Yield y
53	7	9.01	44	10.90	33	10.03
54	25	10.58	34	10.97	9	9.49
55	61	11.03	54	12.22	17	11.52
56	16	9.89	24	10.10	57	12.24
57	52	11.39	4	11.22	65	11.64
58	70	11.24	65	10.01	49	10.74
59	34	12.18	75	10.29	73	10.29
60	42	10.21	38	10.95	7	10.25
61	24	11.08	17	9.66	23	11.39
62	33	11.05	68	9.31	72	13.34
63	51	10.29	7	8.84	55	12.73
64	60	10.57	27	10.64	31	12.62
65	69	10.42	58	9.45	39	10.19
66	15	10.49	48	9.66	47	11.61
67	6	10.00	28	9.85	15	10.52
68	63	9.23	60	9.24	20	9.07
69	54	10.57	30	10.11	61	10.76
70	18	10.27	70	9.63	28	9.91
71	45	8.86	20	9.04	53	10.17
72	72	9.45	9	8.43	69	8.68
73	9	8.03	40	10.97	45	8.74
74	36	9.22	50	8.98	12	9.15
75	27	8.70	10	9.88	4	9.39

10.7.3 More general models

We now consider how the discussion above should be modified when there are explanatory variables as well as a smooth variable, treating certain covariates nonparametrically and others not, and allowing the response to have a density other than the normal.

Let the data consist of independent triples $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$, with jth log likelihood contribution $\ell_j(\eta_j, \kappa)$, where $\eta_j = x_j^{\mathrm{T}}\beta + g(t_j)$; for now we suppress dependence on κ. Then the analogue of (10.47) is the *penalized log likelihood*

$$\ell_\lambda(\beta, g) = \sum_{j=1}^{n} \ell_j(\eta_j) - \frac{1}{2}\lambda \int_a^b [g''(t)]^2 \, dt, \quad \lambda > 0, \tag{10.49}$$

where a and b are chosen so that $a < t_1, \ldots, t_n < b$. If all the t_j are distinct and $\lambda = 0$, the maximum is obtained by choosing $g_j = g(t_j)$ to maximise the jth log likelihood contribution, but this is not useful because the resulting model has n parameters and is too rough. The integral in (10.49) penalizes roughness of $g(t)$, so λ has the same interpretation as before.

If the ordered distinct values of $t_1, \ldots, t_n$ are $s_1 < \cdots < s_q$ and if $g(t)$ is a natural cubic spline with knots at the s_i, then the integral in (10.49) may be written $g^{\mathrm{T}} K g$, where the $q \times 1$ vector g has ith element $g_i = g(s_i)$. Given a value of λ, our aim

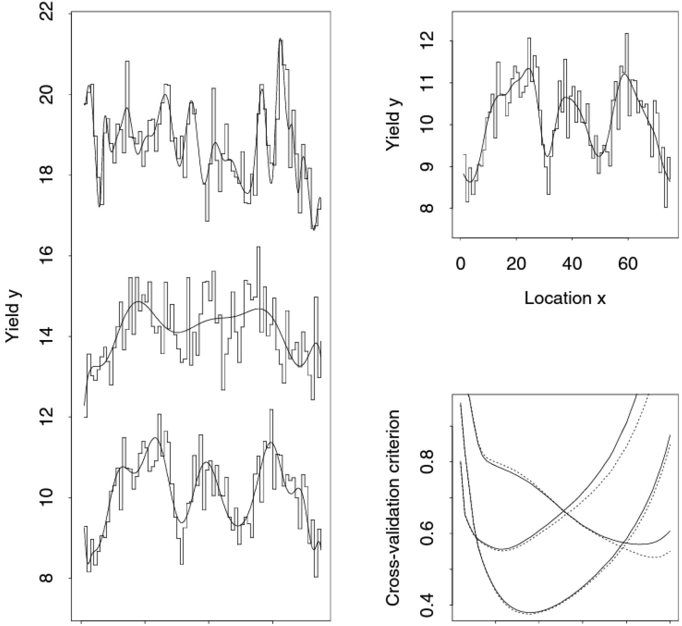


Figure 10.19 Spring barley data analysis. Left panel: yield y as a function of location x for the three blocks. Yields for blocks 2, 3 have been offset by adding 4, 8 respectively. The smooth solid lines are the fits of polynomials of degree 20, 10 and 40 to the data from blocks 1, 2 and 3. Upper right: yields for block 1, with smoothing spline fit with 18 degrees of freedom. Lower right: cross-validation (solid) and generalized cross-validation (dots) criteria for smoothing spline fits to blocks 1, 2 and 3, with minima at roughly 20, 10 and 40 equivalent degrees of freedom.

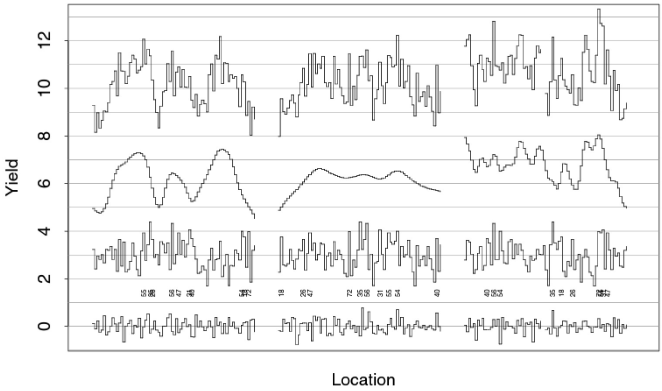


Figure 10.20 Spring barley data analysis. Block 1 is shown on the left and block 3 on the right. The panel shows, from the top, the original yields y, the fertility trend and variety effect estimates $\widehat{g}_b(t)$ and $\widehat{\beta}_v$, both offset for display, and the crude residuals. The varieties with the ten largest $\widehat{\beta}_v$ are marked.

Multidimensional splines

- so far we are considering 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, ...)`

Which smoothing method?

- basis functions: natural splines, Fourier, wavelet bases
- regularization via cubic smoothing splines
- kernel smoothers: locally constant/linear/polynomial
- Faraway (2006) Extending the Linear Model:
  - with very little noise, a small amount of local smoothing
  - with moderate amounts of noise, kernel and spline methods are effective
  - with large amounts of noise, parametric methods are more attractive
- "It is not reasonable to claim that any one smoother is better than the rest"
  - `loess` is robust to outliers, and provides smooth fits
  - spline smoothers are more efficient, but potentially sensitive to outliers
  - kernel smoothers are very sensitive to bandwidth

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

FREE

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 S

ARTICLE TOOLS

- Get PDF (648K)
- Save to My Profile
- E-mail Link to this A
- Export Citation for th
- 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, ...

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

... the NMMAPS studies

- $Y_t \sim Poisson(\mu_t)$
- $\log \mu_t =$ age specific intercepts $+ \beta PM_t + \gamma DOW + g(t, df) + s(temp_t, 6) + s(temp_{t-1}, 6) + s(dewpoint_t, 3) + s(dewpoint_{t-1}, 3) + s_4(dew_0, 3) + s_5(dew_{1-3}, 3)$
- three ages categories; separate intercept for each (< 65, $65 - 74$, ≥ 75)
- dummy variables to record day of week
- $s(x, 7)$ 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). Adjustment uncertainty in effect estimation. *Biometrika* **95** 635–51

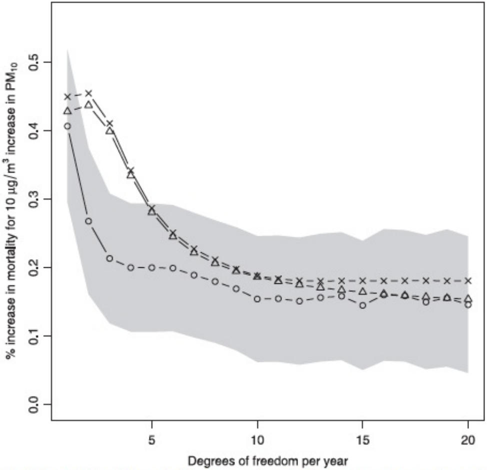


Fig. 3. Sensitivity analysis of the national average estimate of the percentage increase in mortality for an increase in PM_{10} of 10 $\mu g\, m^{-3}$ at lag 1: city-specific estimates were obtained from 100 US cities using data for the years 1987–2000 and the estimates were combined by using a hierarchical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

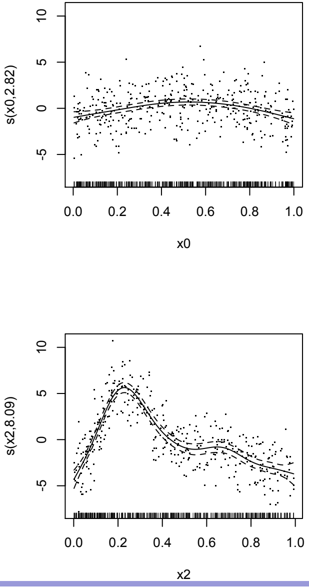
Fitting generalized additive models

`R` package `mgcv`; functions `gam` and `gamm`

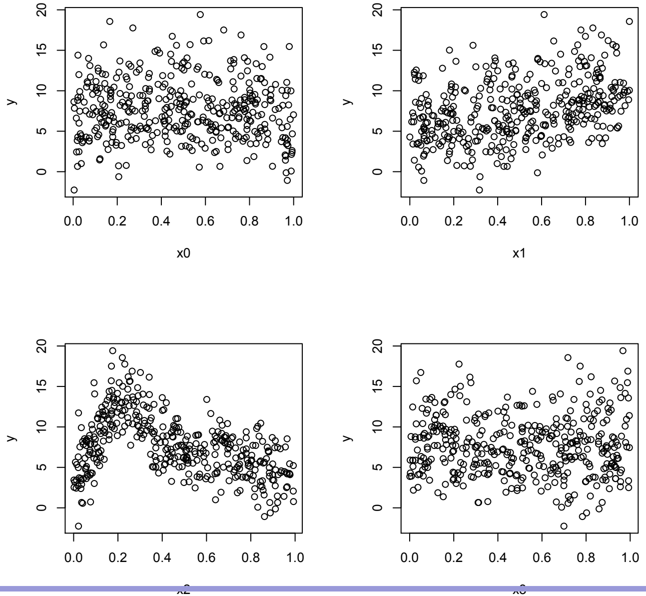
```
> dat = gamSim(1,n=400,dist="normal",scale=2)

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

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

$$y = 2\sin(\pi x_0) + \exp(2x_1) + poly(x_3, degree = 11) + \epsilon$$

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



s(x0,2.82)
x0
s(x1,2.33)
x1
s(x2,8.09)
x2
s(x3,1)
x3


y
x0
x1
0.0 0.2 0.4 0.6 0.8 1.0
0 5 10 15 20

Shrinkage Methods

HTF §3.4

- Ridge regression
-

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y$$

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

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

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

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

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

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

... ridge regression

- $$\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||^2\}$$
- λ is a tuning parameter: $\lambda = 0$ gives $\hat{\beta}_{LS}$, $\lambda \to \infty$
- 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

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

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

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

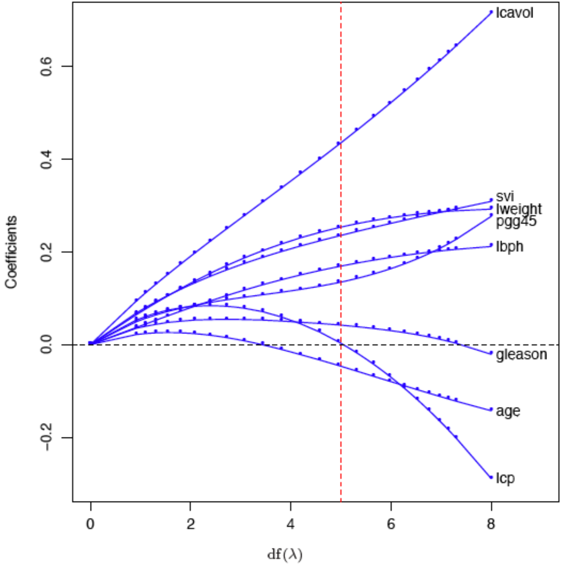


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter* λ *is varied. Coefficients are plotted versus* $\mathrm{df}(\lambda)$*, the effective degrees of freedom. A vertical line is drawn at* $\mathrm{df} = 5.0$*, the value chosen by cross-validation.*

Lasso

- $$\min_{\beta} \left(\Sigma\{y_i - \beta_0 - \Sigma_{j=1}^p x_{ij}\beta_j\}^2 + \lambda \Sigma_{j=1}^p |\beta_j| \right)$$
- $$\min_{\beta} \Sigma\{y_i - \beta_0 - \Sigma_{j=1}^p x_{ij}\beta_j\}^2 \quad \text{s.t. } \Sigma|\beta_j| \le t$$
- quadratic programming problem
- $\hat{\beta}^{lasso}$ is nonlinear function of y
- Tibshirani (1996), JRSS B and (2011), JRSS B
- `http://http://www-stat.stanford.edu/~tibs/lasso.html`

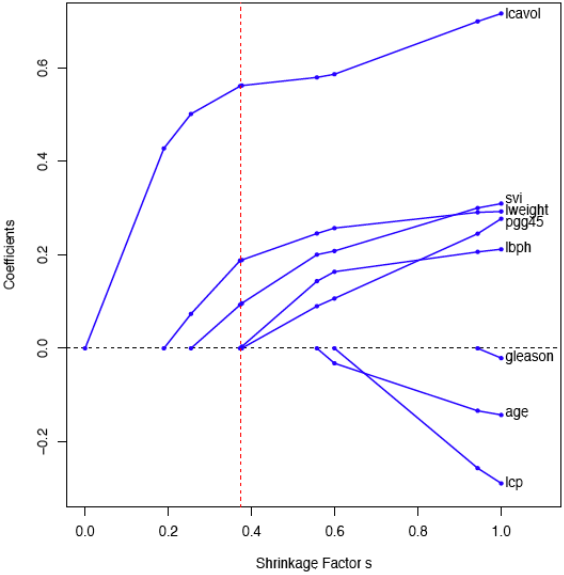


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso*

... shrinkage

- ridge regression gives "proportional shrinkage"
- subset selection gives "hard thresholding" (some $\beta_j \to 0$)
- lasso gives "soft thresholding": blend of shrinkage and zeroing

- elastic net combines lasso and ridge regression

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

- implemented in `R` in `library(glmnet)`

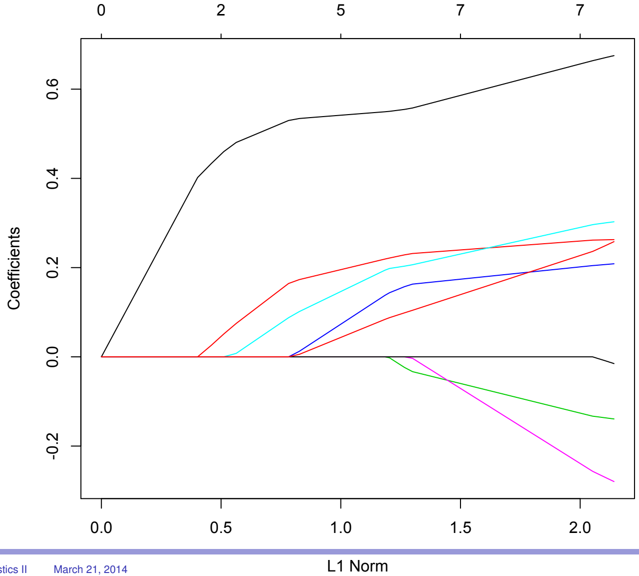
- estimates of coefficients are biased (but may have small mean-squared error)
- Lasso is now used as a variable selection method
- improvements in algorithms allow fast computation even for $p > n$

Prostate data

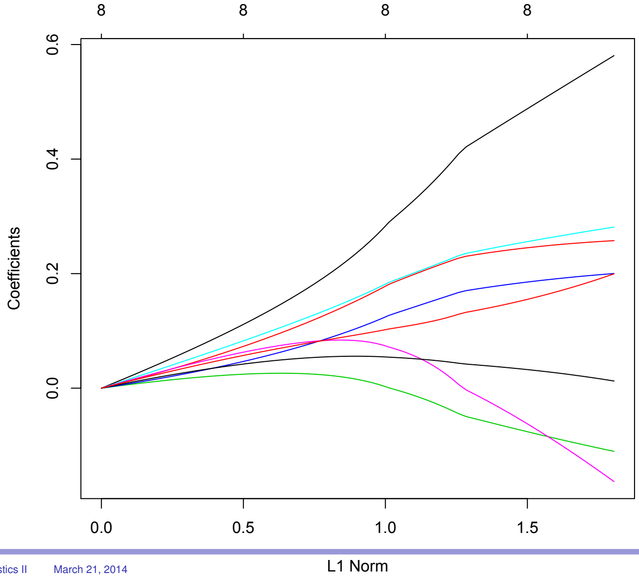
Ch.3, HTF

```
> prostate <- read.csv(file="prostate.data",sep="\t")
> rm(try)
> head(prostate)
  X     lcavol  lweight age      lbph svi       lcp gleason pgg45
1 1 -0.5798185 2.769459  50 -1.386294   0 -1.386294       6     0
2 2 -0.9942523 3.319626  58 -1.386294   0 -1.386294       6     0
3 3 -0.5108256 2.691243  74 -1.386294   0 -1.386294       7    20
4 4 -1.2039728 3.282789  58 -1.386294   0 -1.386294       6     0
5 5  0.7514161 3.432373  62 -1.386294   0 -1.386294       6     0
6 6 -1.0498221 3.228826  50 -1.386294   0 -1.386294       6     0
        lpsa train
1 -0.4307829  TRUE
2 -0.1625189  TRUE
3 -0.1625189  TRUE
4 -0.1625189  TRUE
5  0.3715636  TRUE
6  0.7654678  TRUE
> xp <- scale(prostate[,2:9])
> y <- prostate[,10]
> train <- prostate[,11]
## standardize data; y is the response (log psa); extract training data
##
> library(glmnet)
> pr.lasso <- glmnet(xp[train,],y[train])
> plot(pr.lasso)
```

... prostate data



... prostate data



stuff

International Day for the Elimination of Racial Discrimination